

ANALYTICS & DATA SCIENCE

Una introducción corta

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>

Data Mining vs Data Science vs Analytics

Data Mining

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

Data Science (¿Data Analytics?):

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

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."

#15	3,433	\$105,395	#1
-----	-------	-----------	----

Highest Paying Job in
Demand

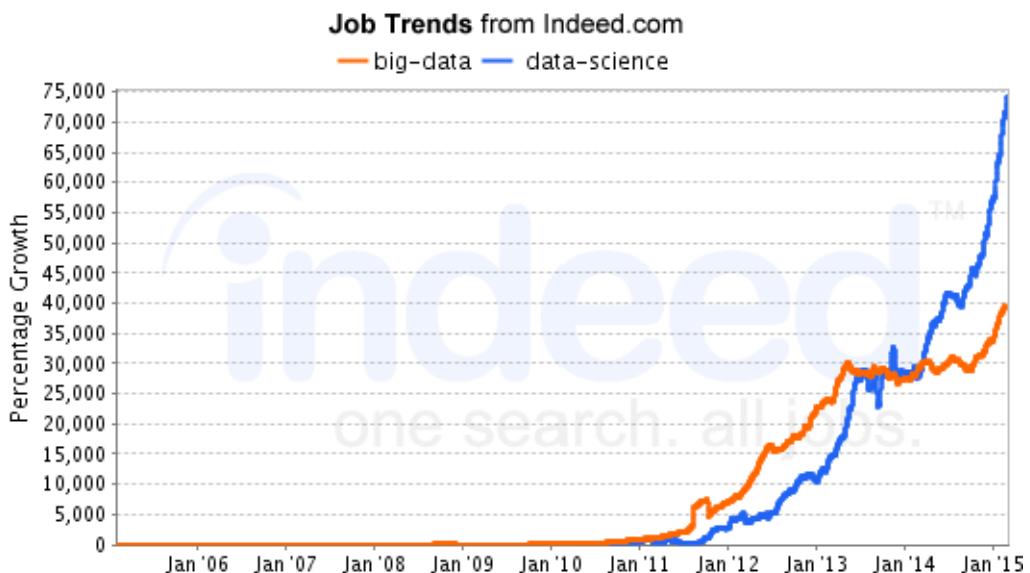
Number of Job Openings

Average Base Salary

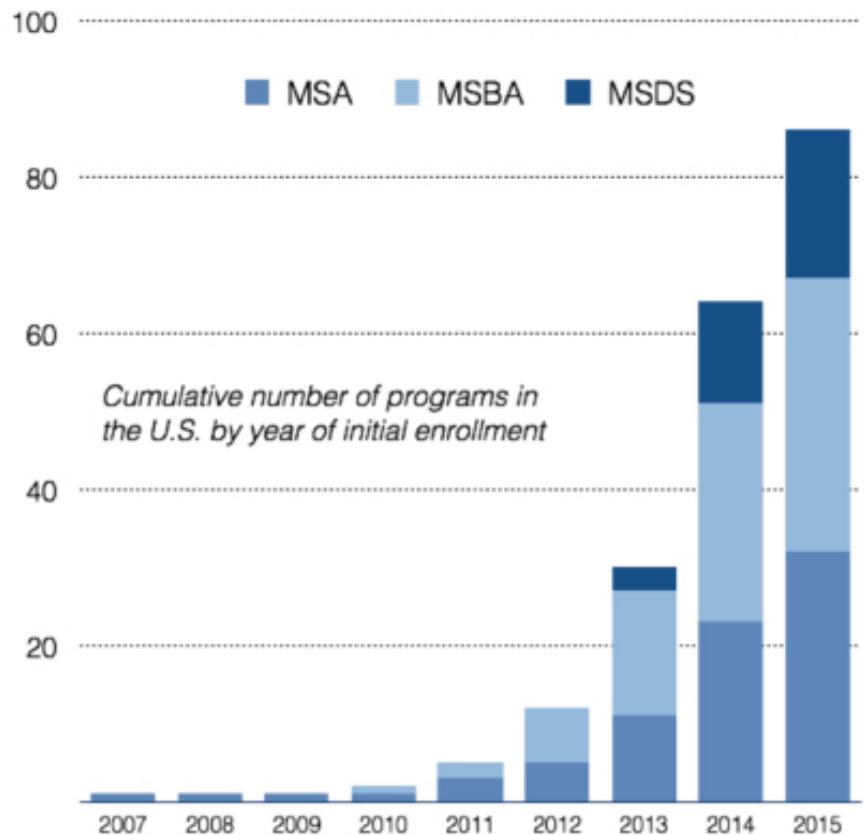
Best Job in America for
2016

Sources: <http://www.glassdoor.com/blog/jobs-america/> and <http://www.glassdoor.com/blog/highest-paying-jobs-demand/>

Evolución de empleos/educación en Big Data & Data Science

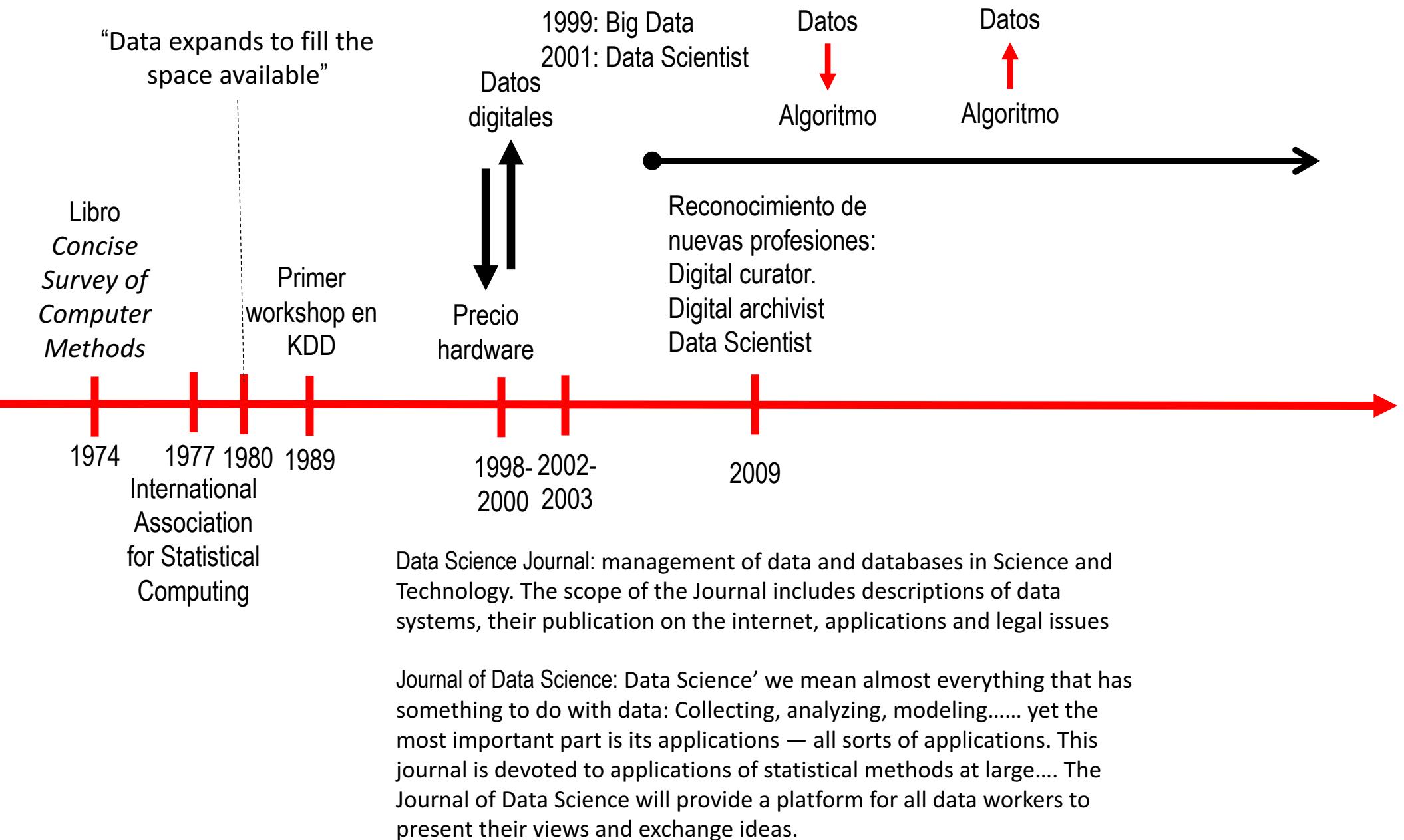


GROWTH OF MASTER'S DEGREE PROGRAMS IN ANALYTICS AND DATA SCIENCE



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

Línea de tiempo. Data Science / Big Data



Elementos constitutivos

- Datos
- Algoritmos, técnicas y metodologías
- Infraestructura computacional

Infraestructura Computacional

Infraestructura computacional

Máquinas Locales

- Servidores + red + estaciones de trabajo.

Computación en la nube

- **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.

Aplicaciones
Datos
Sistema Operativo
Virtualización
Servidores
Almacenamiento
Red

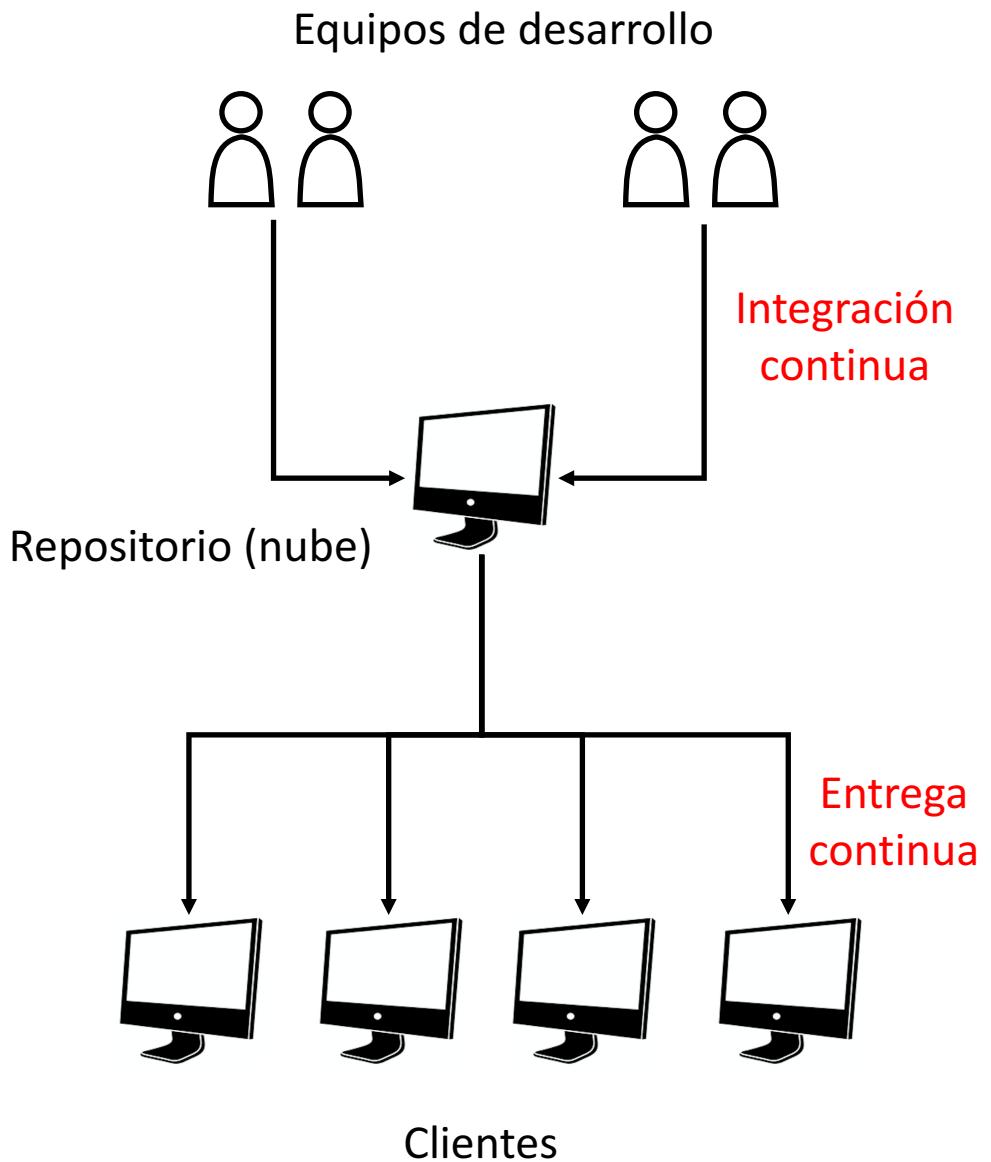
DevOps (Development Operations)

Beneficios

- Velocidad
- Entrega rápida
- Fiabilidad
- Escalado
- Colaboración mejorada
- Seguridad

Prácticas

- Integración continua
- Entrega continua
- Microservicios
- Infraestructura como código
- Monitorización y registro
- Comunicación y colaboración



Datos

Almacenamiento de Datos y Bases de Datos

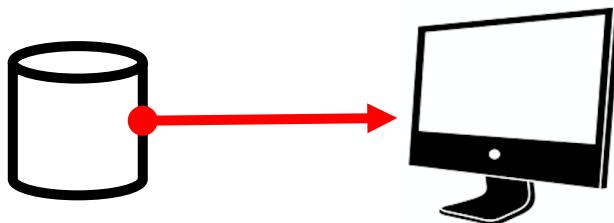
Archivos de datos

- Delimited Text Files
- XML Files
- Log Files
- Archivos específicos de cada aplicación (Excel, Access,...)

Databases

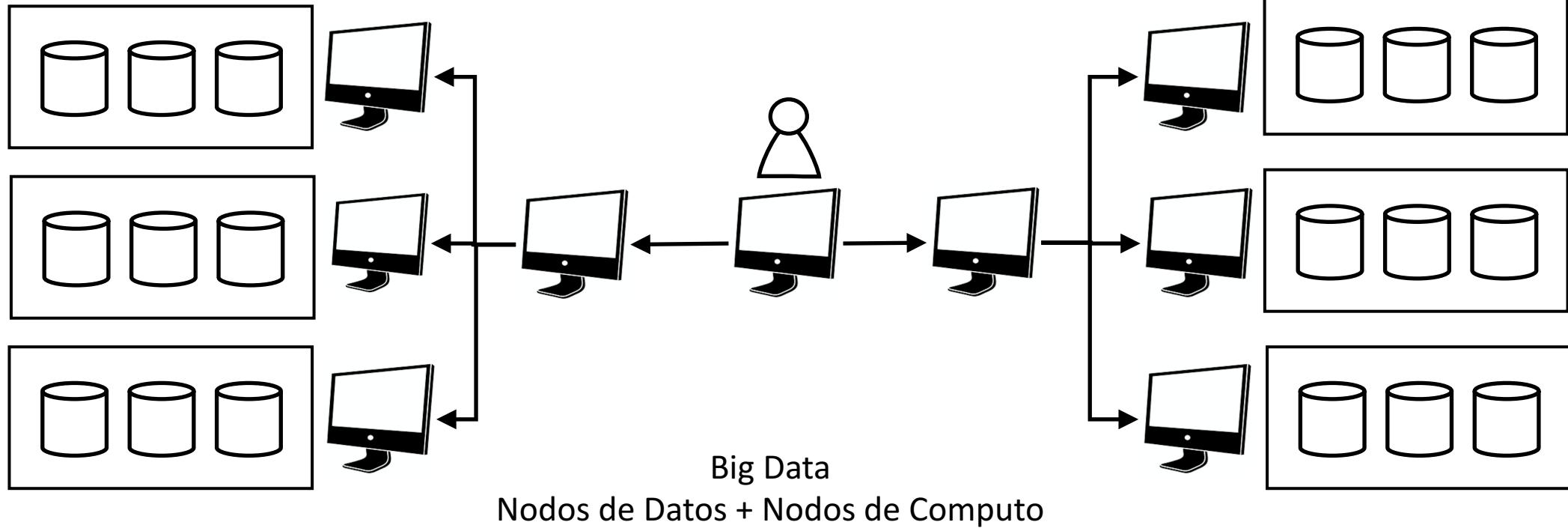
- Bases de datos relacionales
- Bases de datos gráficos
- Almacenes de documentos
- Bases de datos columnares
- Diccionarios (clave – valor)

Aproximación tradicional vs Big Data

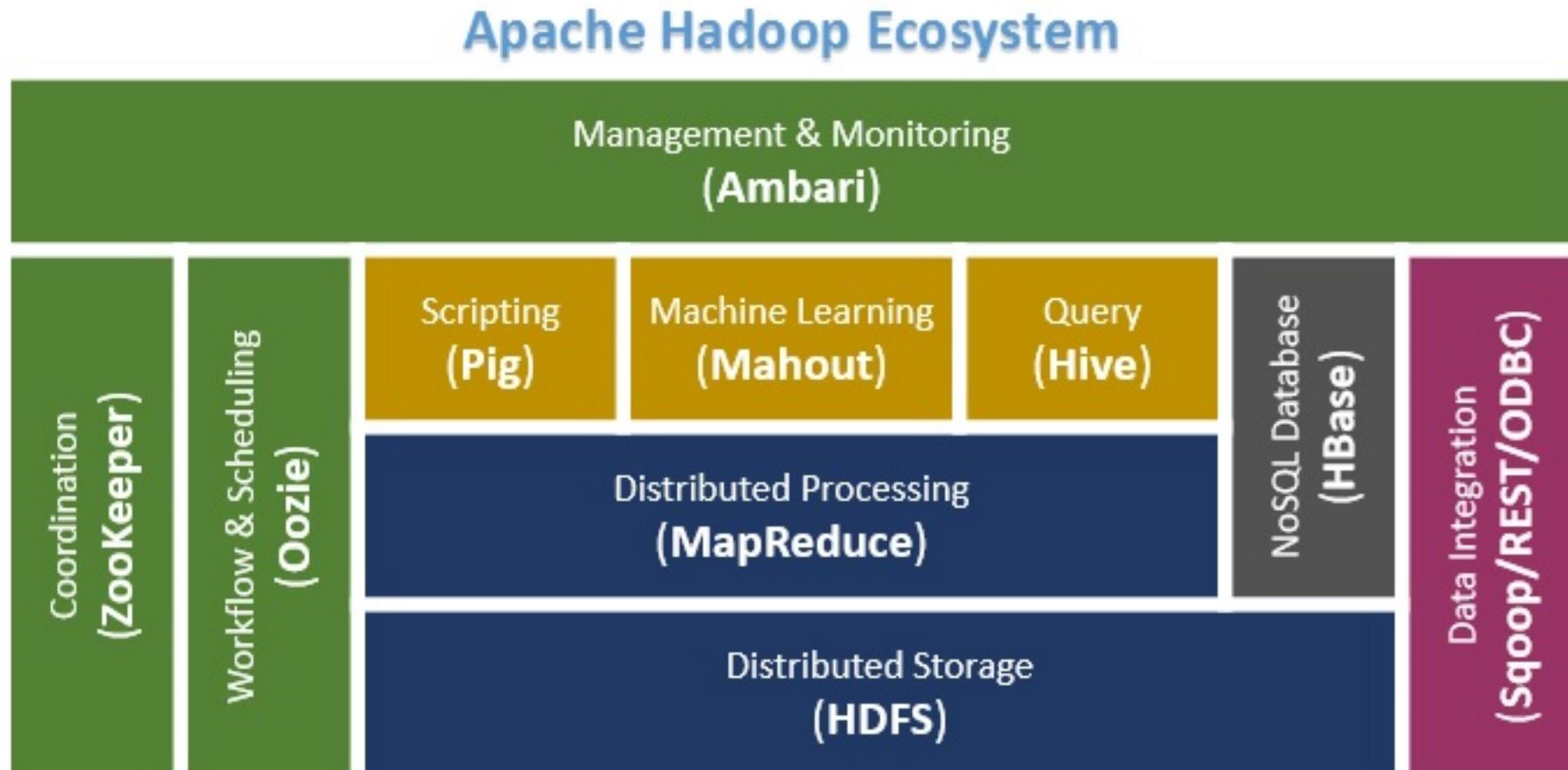


Programación Tradicional

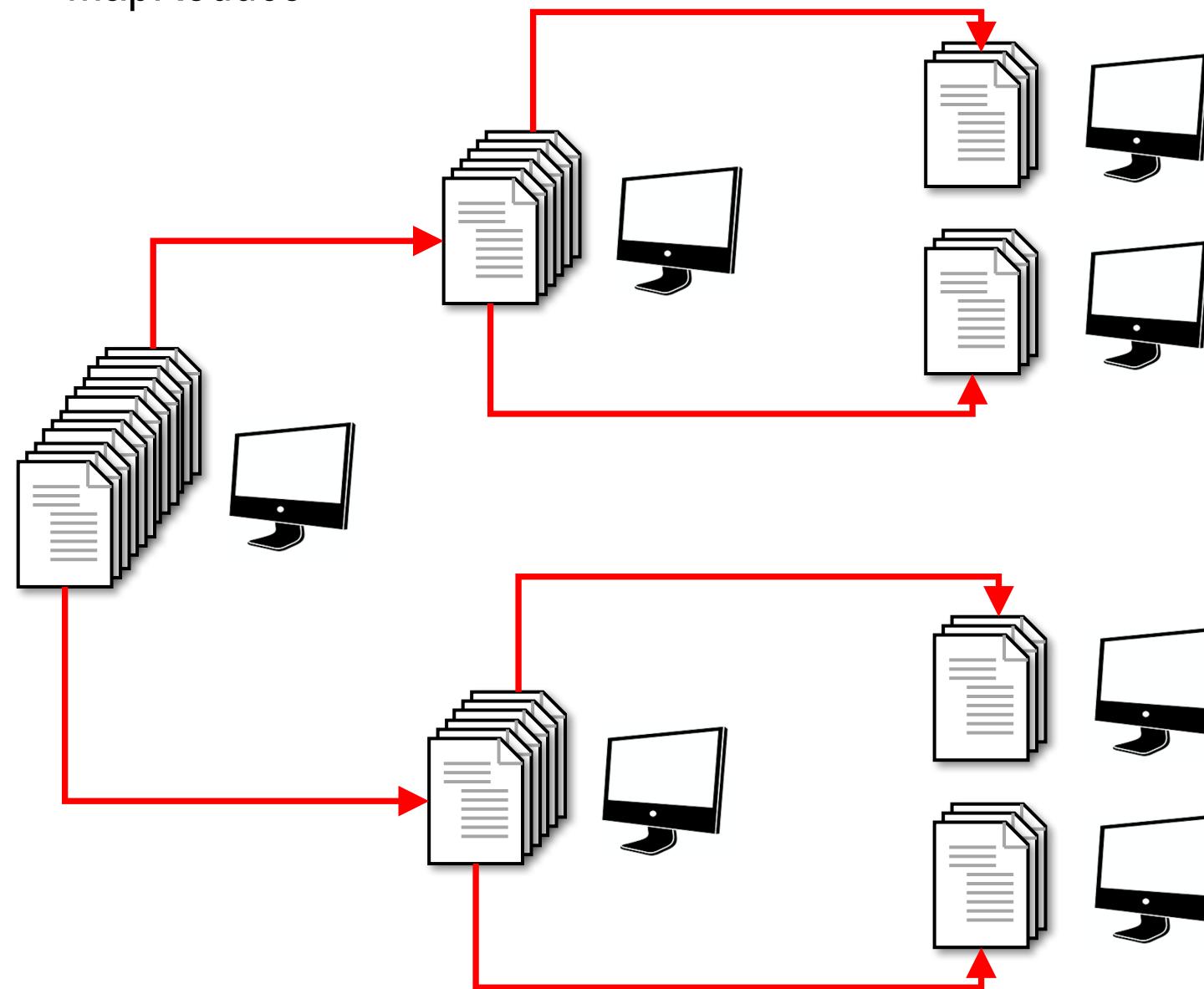
¿Cómo ejecutar algoritmos tradicionales que son voraces en recursos computacionales?



Big Data – Apache Hadoop

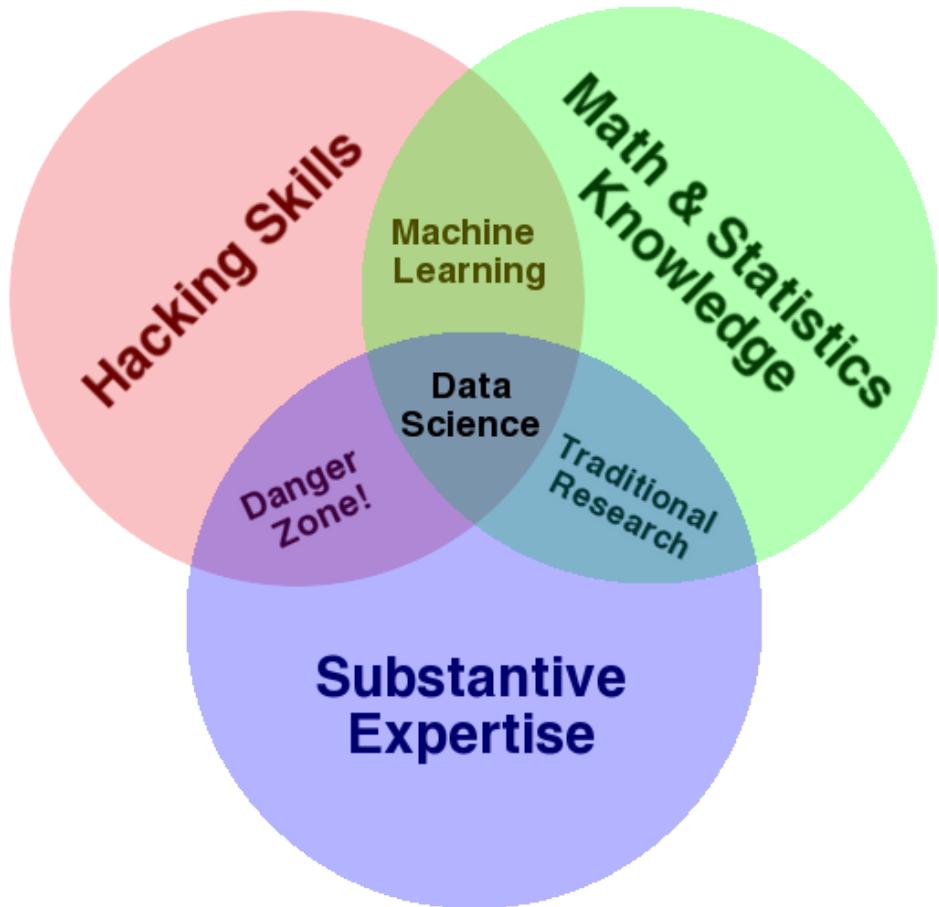


MapReduce



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

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	<ul style="list-style-type: none"> • ETL/SQL • RDBMS • Reportes • Visualización 	<ul style="list-style-type: none"> • Programación • Análisis de datos • Modelado de datos • Desarrollo de reportes • Estadística Básica • Análisis del negocio & Estrategia • Presentación oral 	<ul style="list-style-type: none"> • Suministro de información y reporte • Visualización de datos • Estadísticos descriptivos • Integración de datos y consolidación
Análisis de datos	<ul style="list-style-type: none"> • Software para modelado de datos • Software para diagramación • Software para documentación • SQL • Software para perfilado de datos 	<ul style="list-style-type: none"> • Modelado de datos • Análisis del negocio • Manipulación de datos • Estadística básica 	<ul style="list-style-type: none"> • 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)	<ul style="list-style-type: none"> • Software estadístico • Datos columnares • Map-Reduce • NoSQL • Lenguajes de programación • Software para graficación • Software para optimización, simulación, predicción y análisis de decisiones 	<ul style="list-style-type: none"> • 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 	<ul style="list-style-type: none"> • 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.

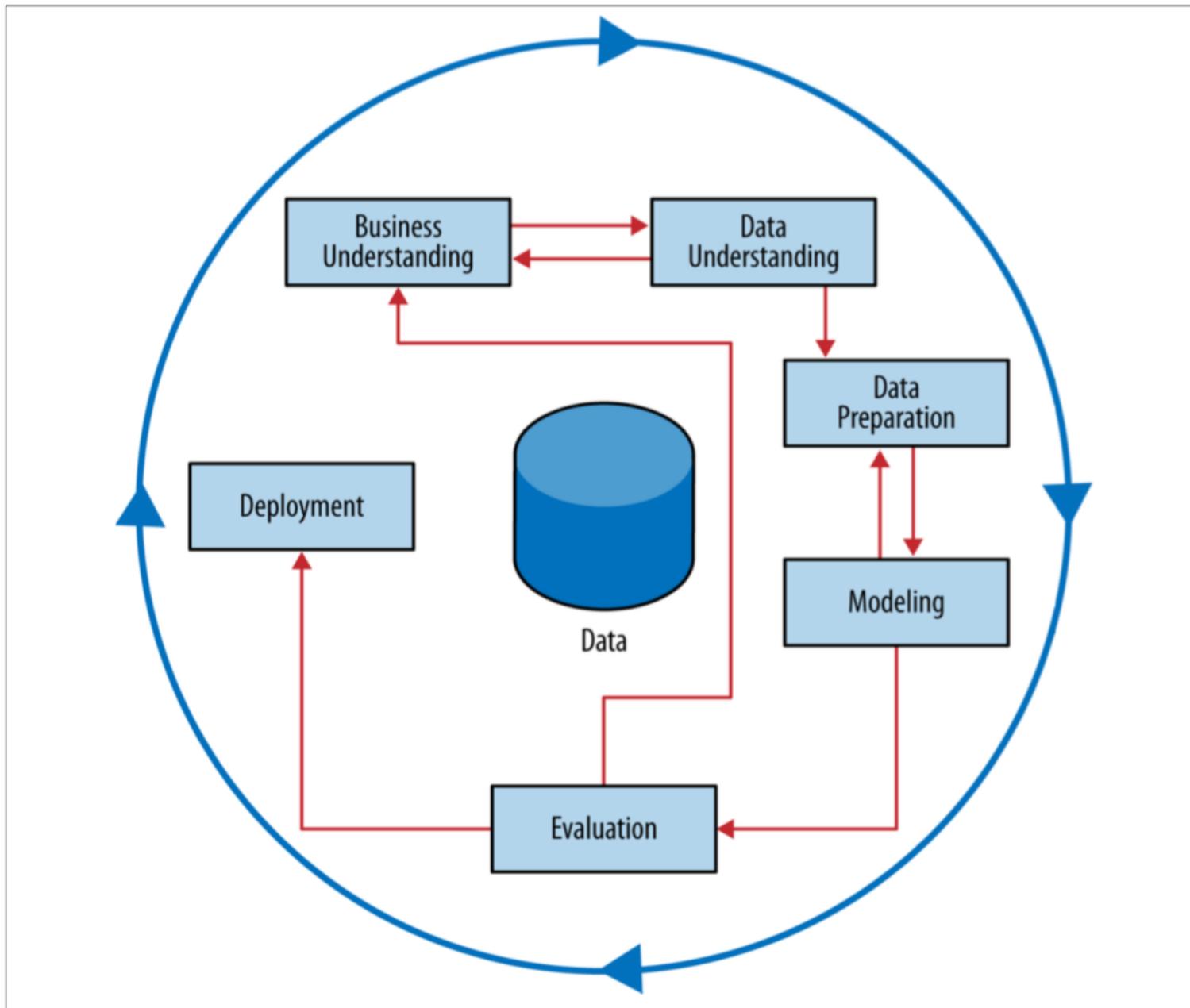
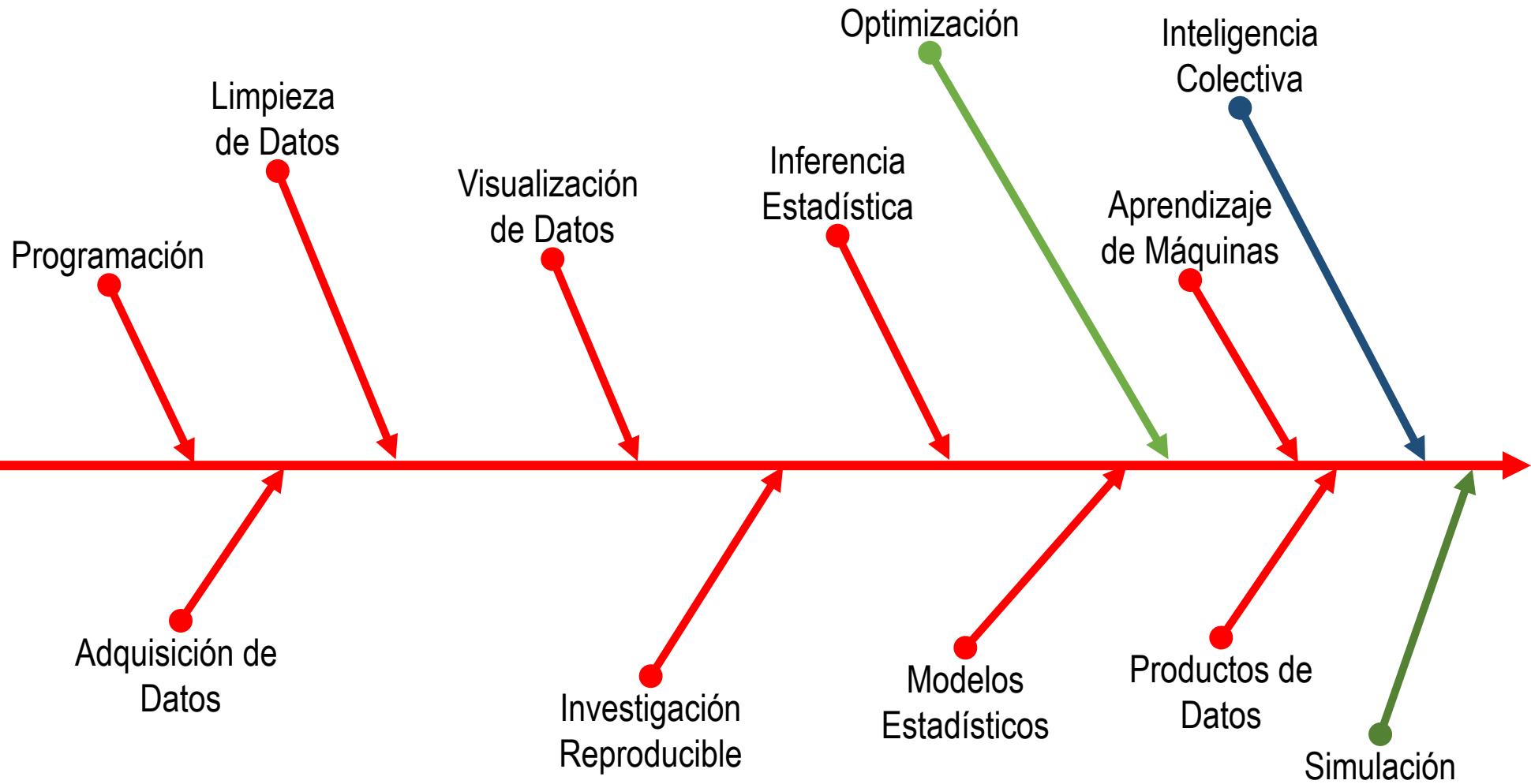


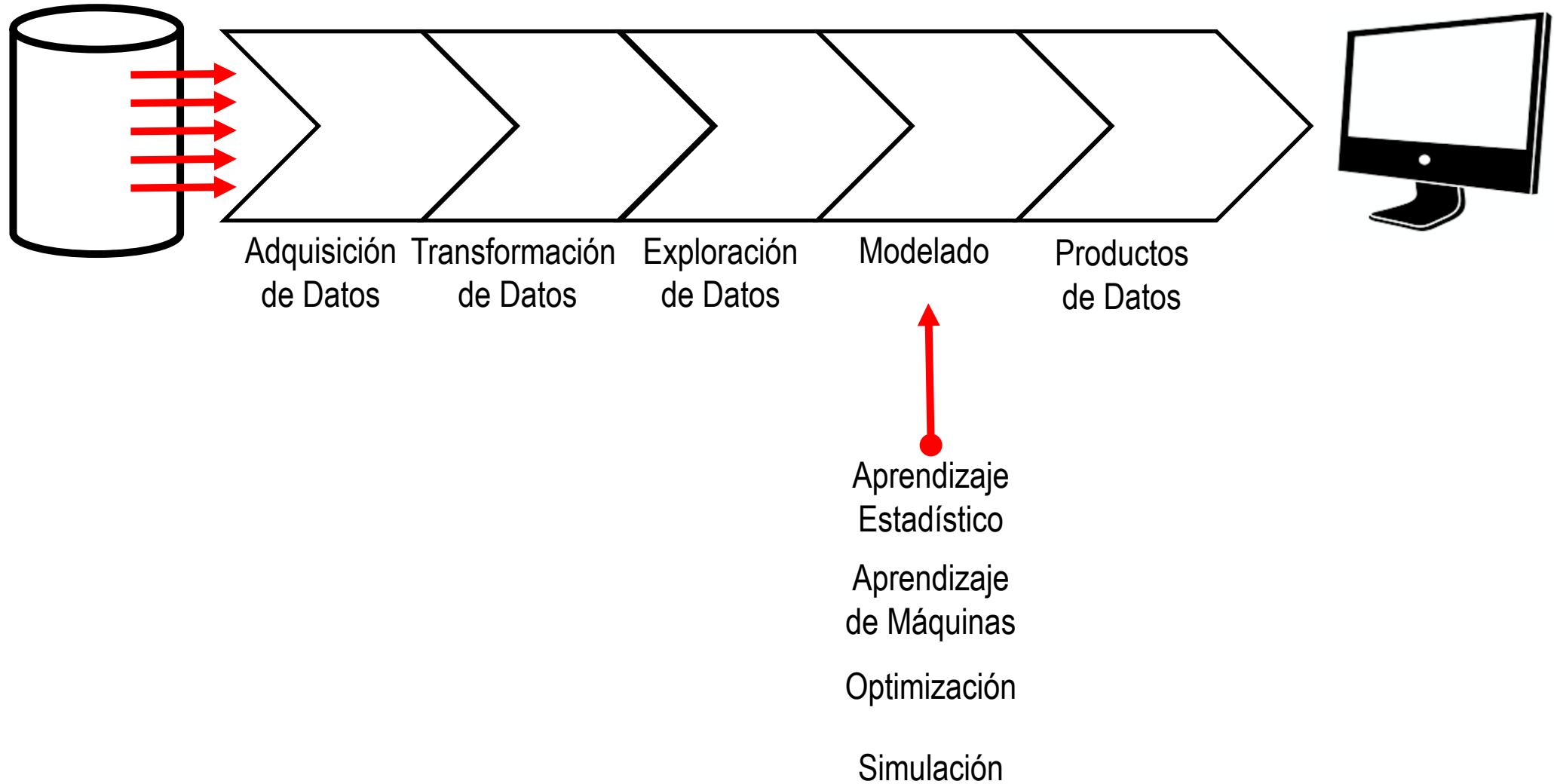
Figure 2-2. The CRISP data mining process.

Componentes de Data Science / Analytics



Data-driven decision making!

Fases en Data Science / Analytics

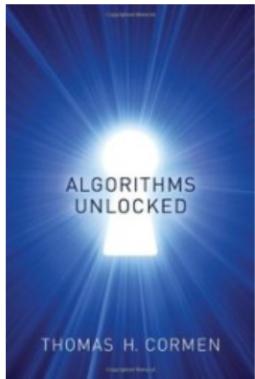


Programación -- ¿Usted sabe programar ... / Es capaz de ...?

¿Ordenar un vector de números?

¿Extraer la tercera línea de texto de un conjunto de archivos?

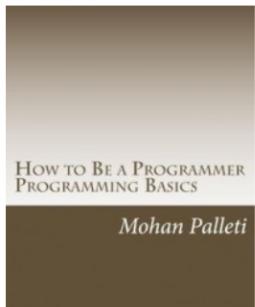
¿Calcular la suma de los primeros 20 números primos?



Algorithms Unlocked

By: Thomas H. Cormen

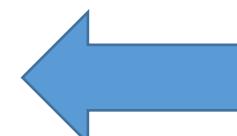
Have you ever wondered how your GPS can find the fastest way to your destination, selecting one route from seemingly countless possibilities in mere seconds? How your credit card account number is protected when you make a purchase over the Internet? The answer is algorithms. And how do...



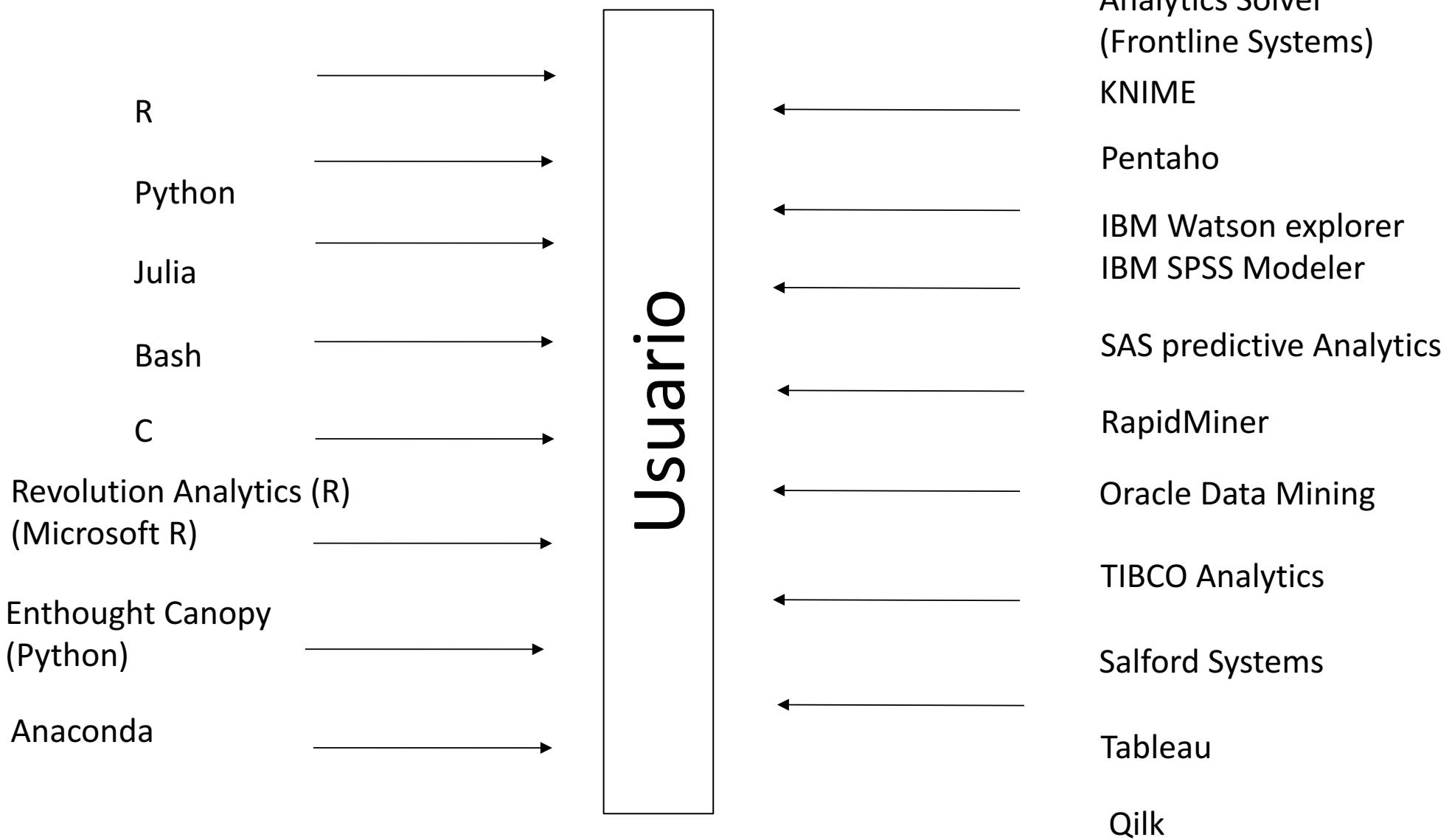
How to Be a Programmer: Programming Basics

By: Mohan Palleti

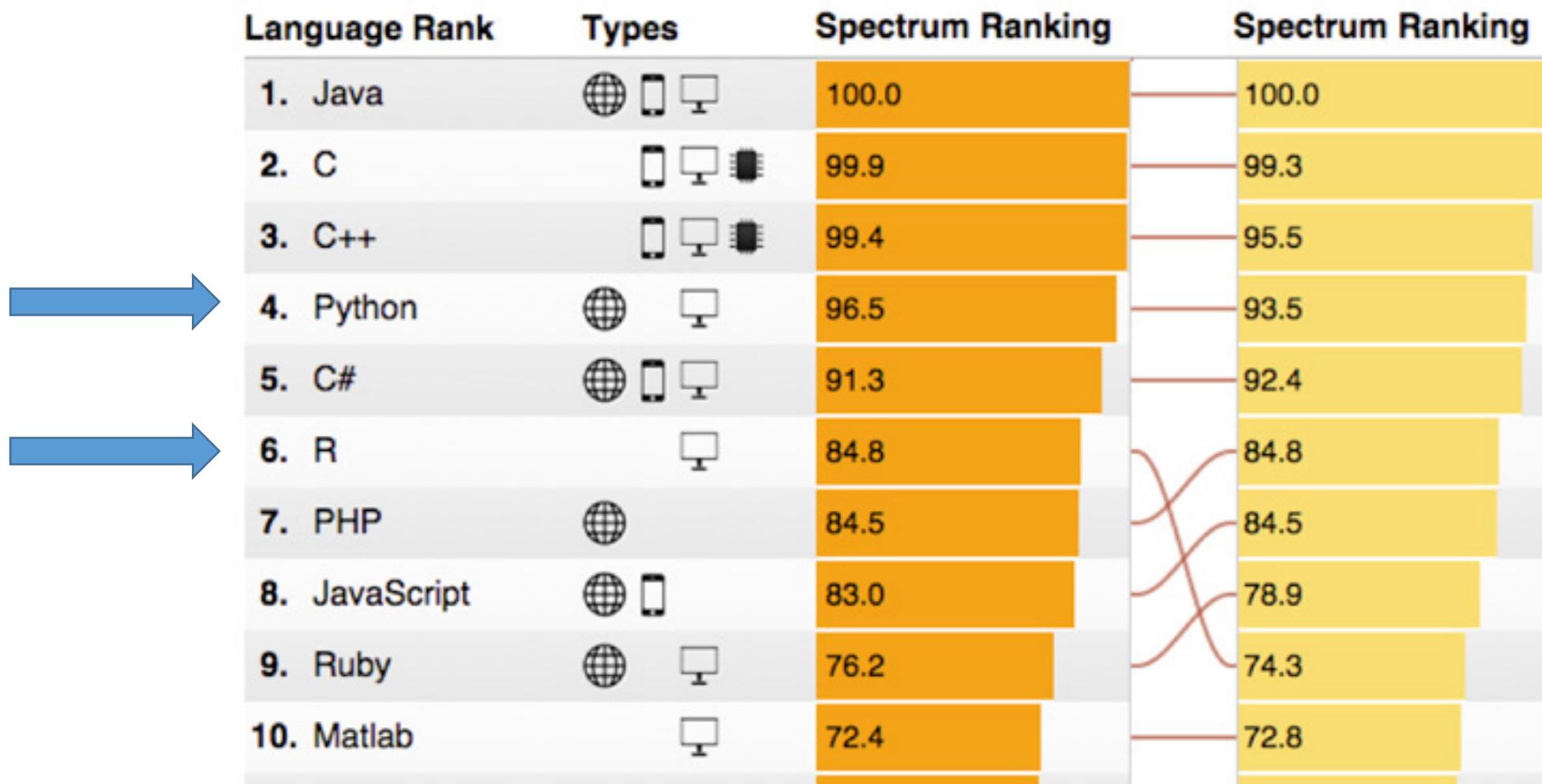
A Self-help 97 pages book to learn the basics of programming using Microsoft Excel's VBA tools. Ideal resource for school teachers and educators wanting to teach programming basics.



Programación vs Aplicaciones de usuario final



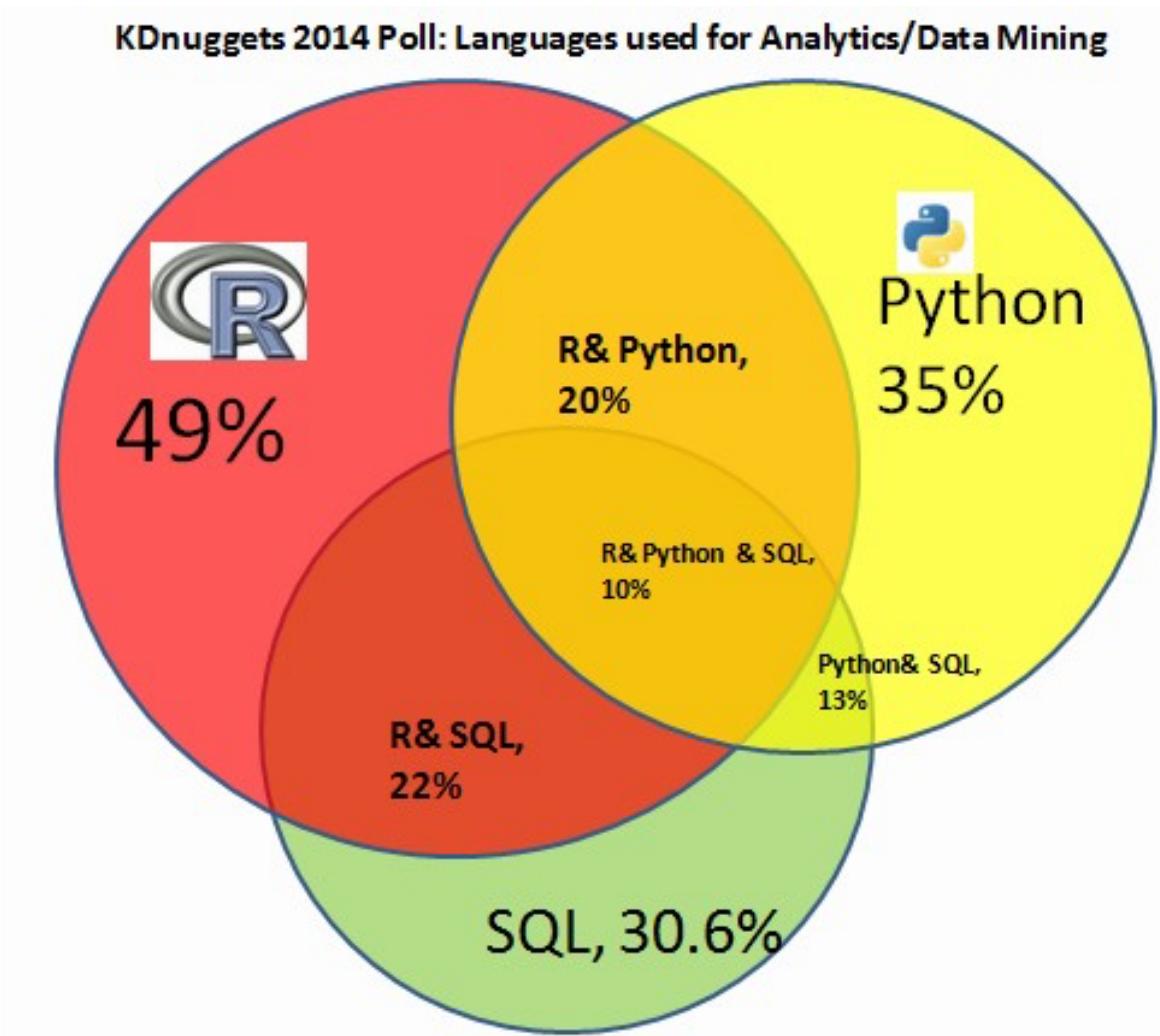
The 2015 Top Ten Programming Languages (IEEE Spectrum)



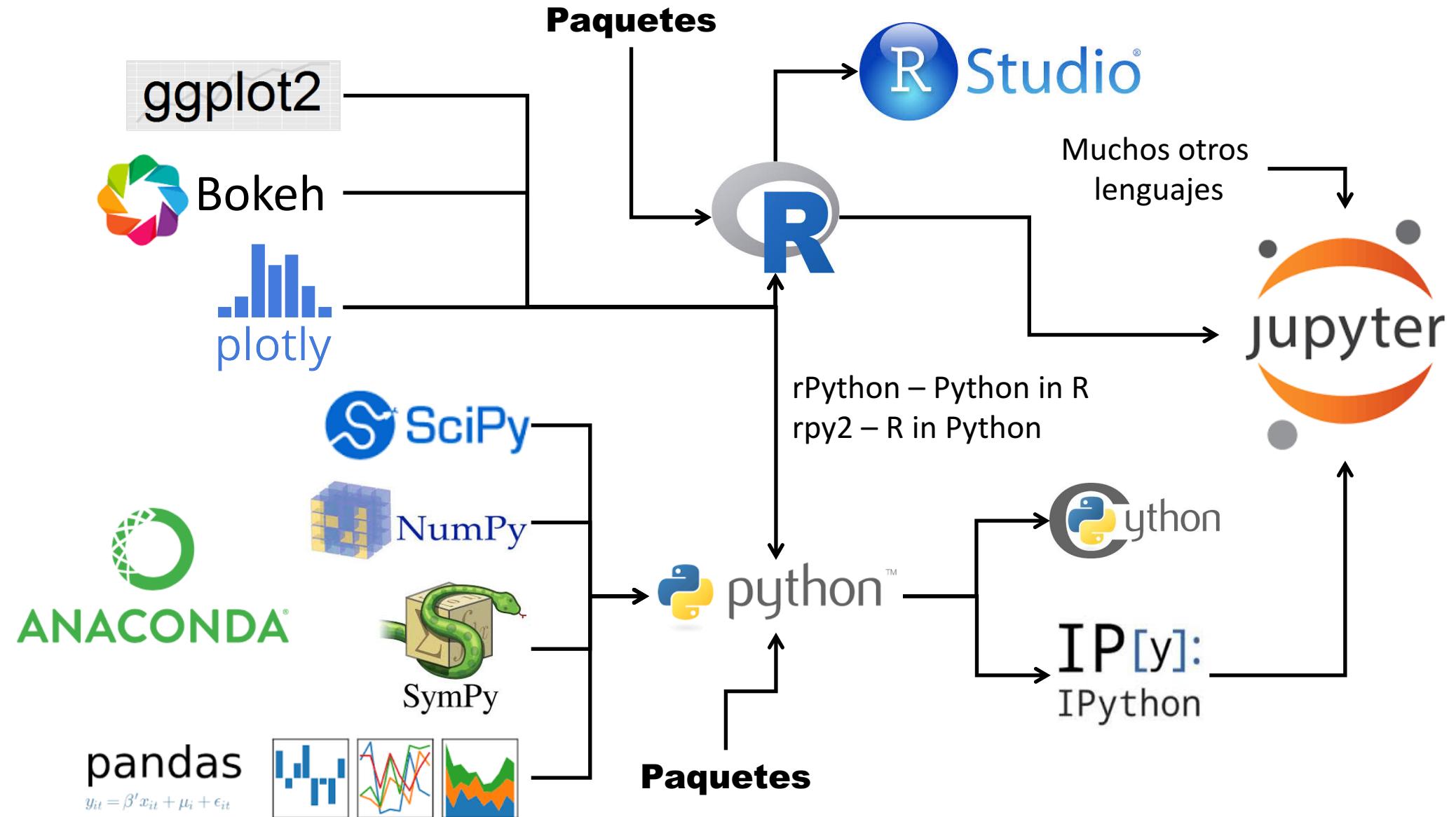
The 2016 Top Ten Programming Languages (IEEE Spectrum)

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

Popularidad de los lenguajes



Ecosistema de computación científica: Python y R



Adquisición y Limpieza de Datos

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

Páginas web (HTML) y Google Groups.

Bases de datos relacionales.

Lenguaje Natural.

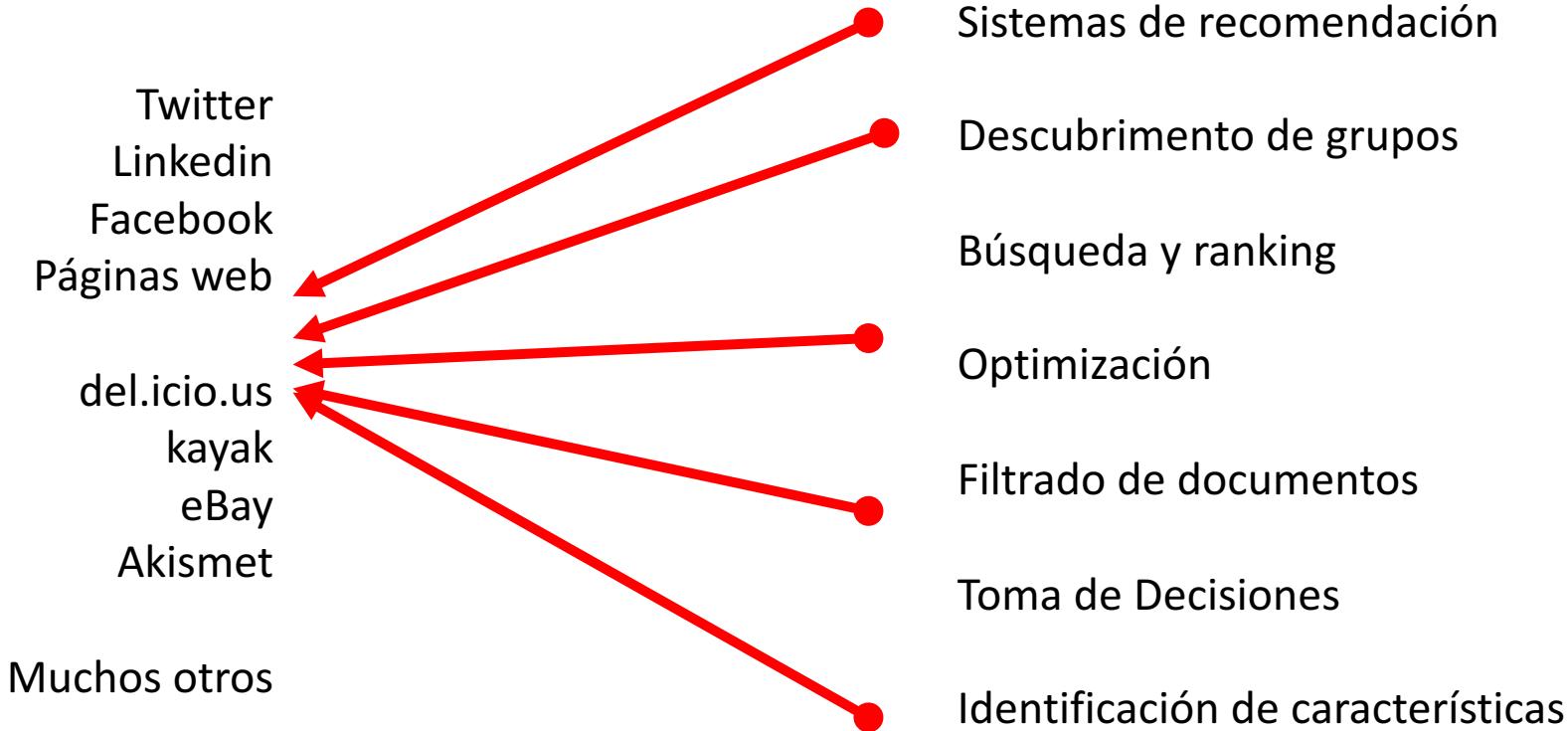
Imágenes (Captcha)

Manipulación de texto

Conversión de un formato a otro

Detección de datos faltantes, datos nulos, datos inconsistentes

Adquisición de datos -- Inteligencia colectiva

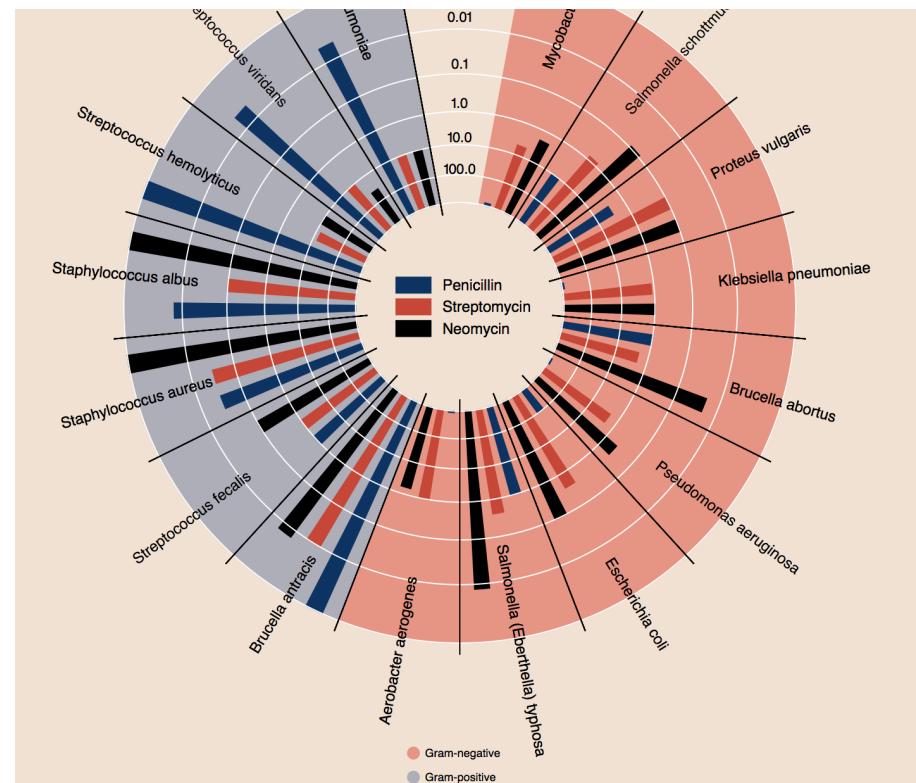
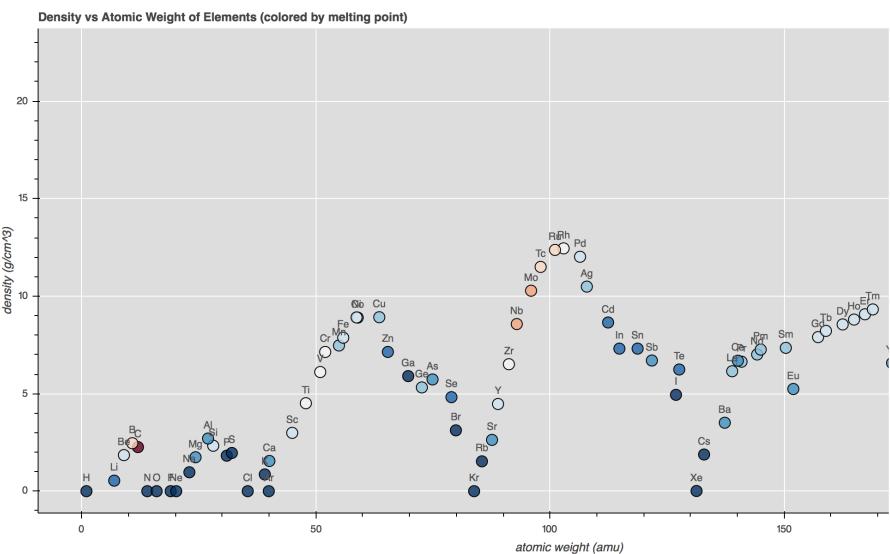
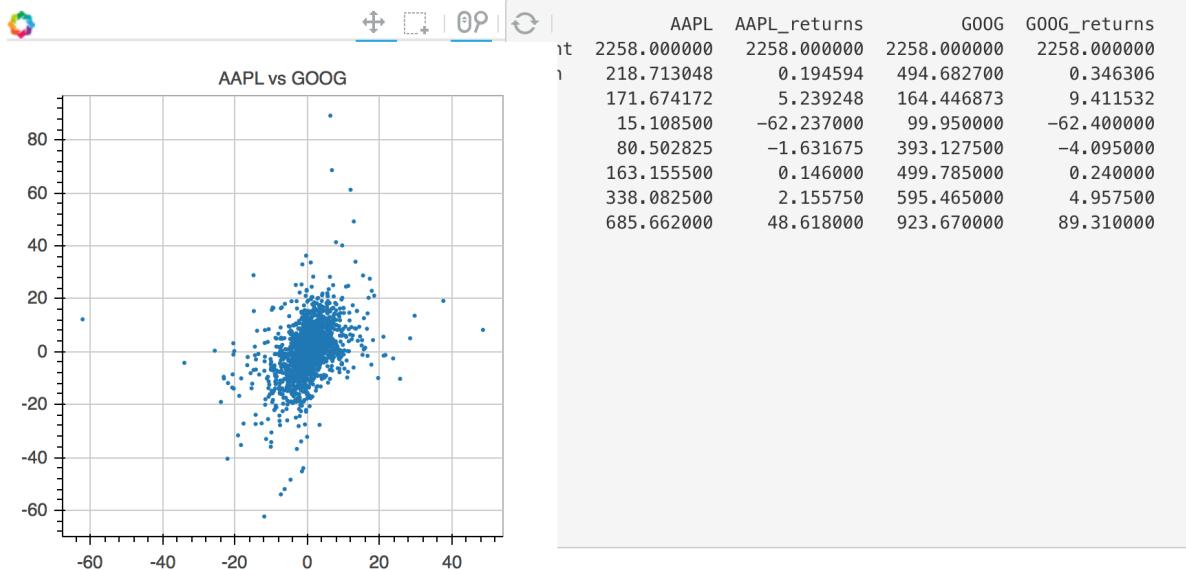


link to this

Visualización de Datos

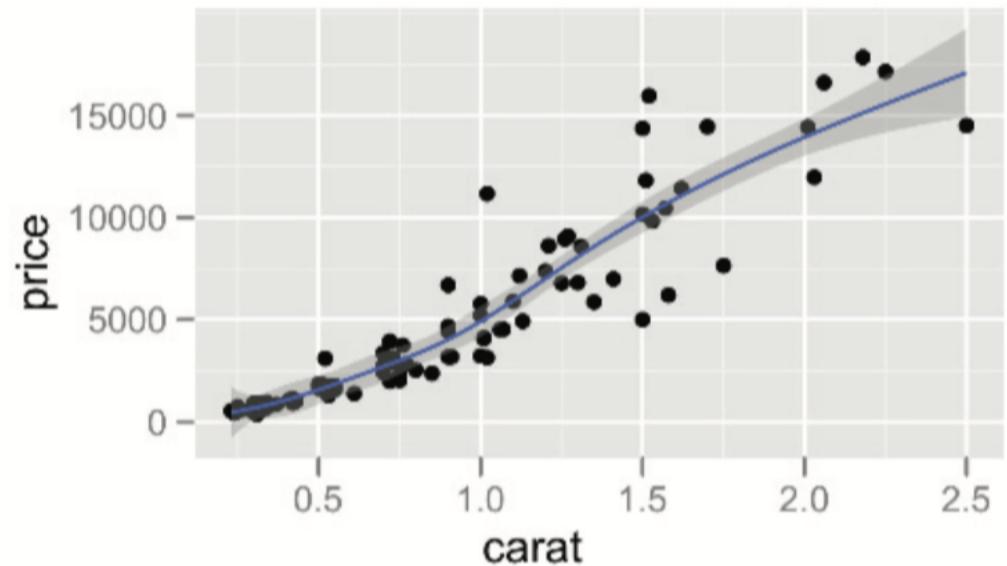
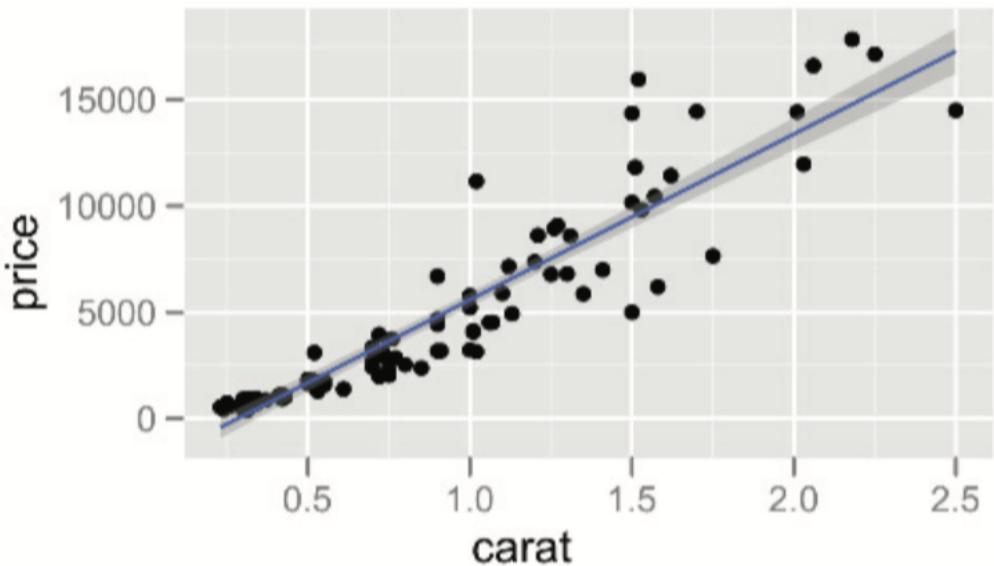
- Bokeh
- Matplotlib
- R
-

AAPL
GOOG



Modelado Estadístico y Aprendizaje de Máquinas

Aplicación clásica



¿Y si hay 10 millones de datos?

Investigación 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)

The screenshot shows the RStudio interface with a document titled "knitr-ex1.Rmd". A blue arrow points from a button labeled "Push here" in a blue box to the "Knit HTML" button in the toolbar. The code editor contains the following R Markdown code:

```
1 My First knitr Document
2 -----
3
4 This is some text (i.e. a "text chunk").
5
6 Here is a code chunk
7 ```{r}
8 set.seed(1)
9 x <- rnorm(100)
10 mean(x)
11 ```
```

The resulting HTML output is displayed below, with a "Code input" box and a "Numerical output" box highlighting specific sections of the rendered content.

My First knitr Document

This is some text (i.e. a “text chunk”).

Here is a code chunk

```
set.seed(1)
x <- rnorm(100)
mean(x)
```

Code input

```
## [1] 0.1089
```

Numerical output

Investigación Reproducible (Jupyter Notebook)

The screenshot shows a Jupyter Notebook interface with the following content:

```
import scipy
import sys

# make nice plots
import plt_fmt

Populating the interactive namespace from numpy and matplotlib
```

"m" key denotes a markdown cell

```
In [8]: kk = rand(5,2)

(r1,r2) = kk[1][:]
print (kk[1][:])
print (r1)
print (r2)

[ 0.20757795  0.01992547]
0.207577947999
0.019925471486
```

```
In [4]: def vfield(n,time, param):
    """
        param is an Nx2 matrix specifying the parameters for
        the dynamical system
    """

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

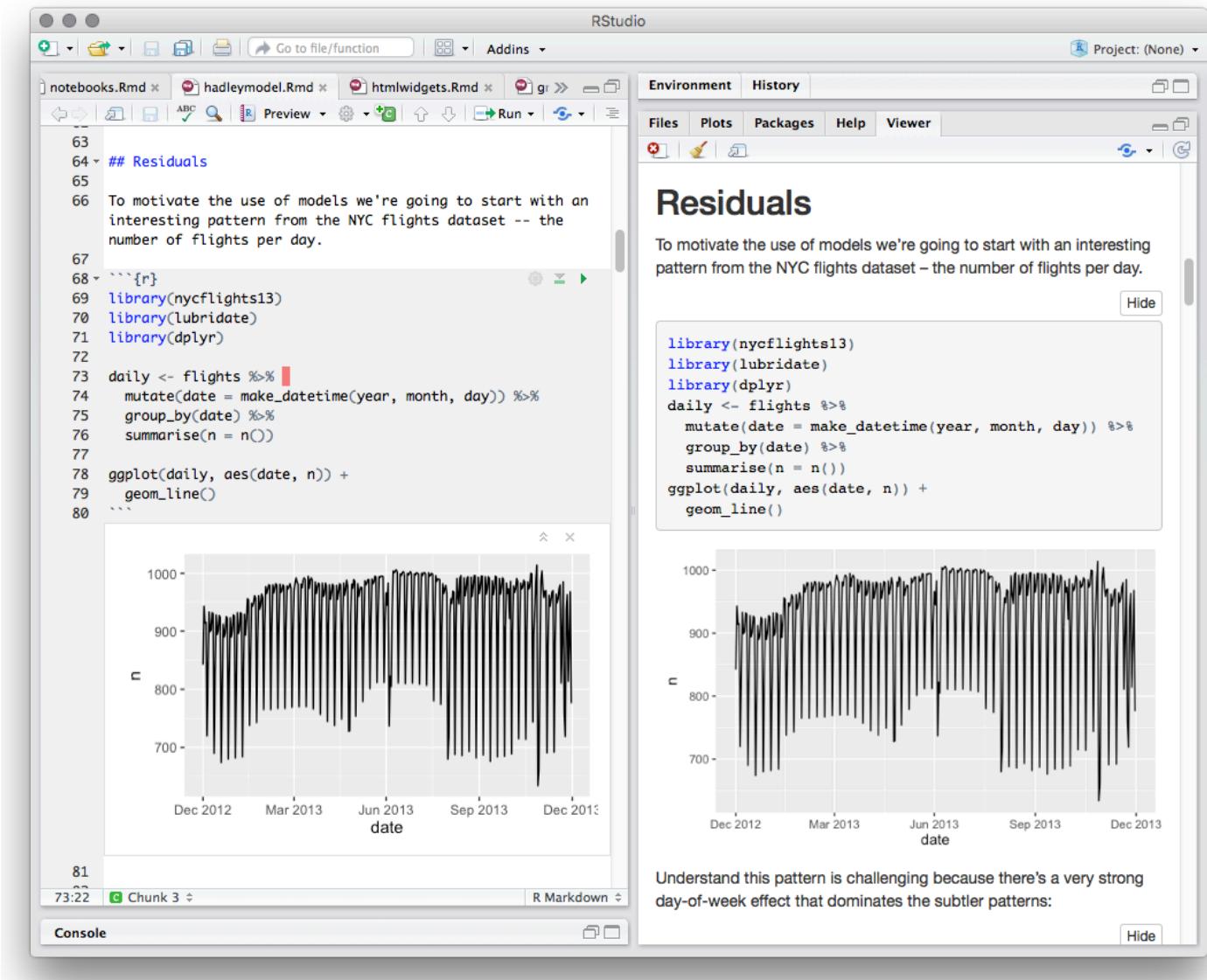
```
Out[4]: [

A plot window is visible at the bottom, showing a blue line segment starting near the bottom right corner of a unit square and extending upwards towards the top left.


```

Productos de datos

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

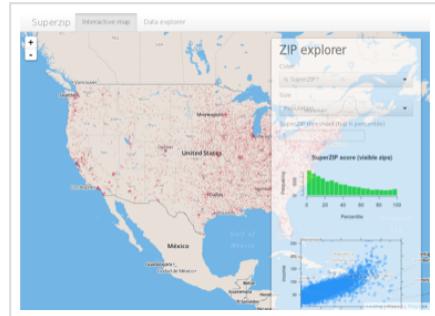


Gallery

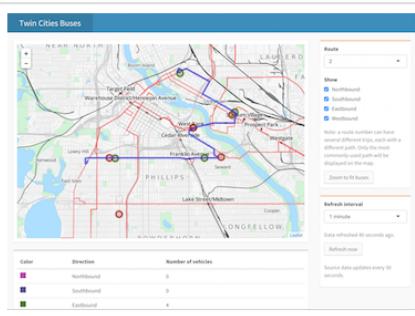
This gallery contains useful examples to learn from. Visit the [Shiny User Showcase](#) to see an inspiring set of sophisticated apps.

Interactive visualizations

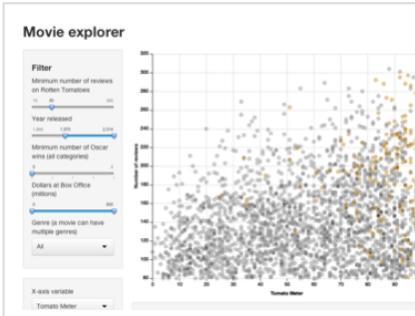
Shiny is designed for fully interactive visualization, using JavaScript libraries like [d3](#), [Leaflet](#), and [Google Charts](#).



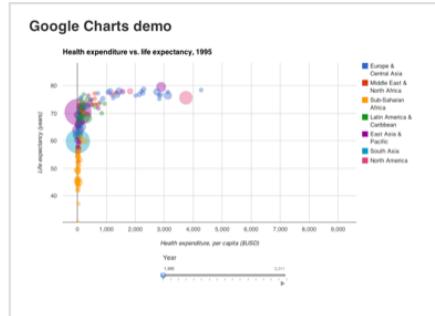
[SuperZip example](#)



[Bus dashboard](#)



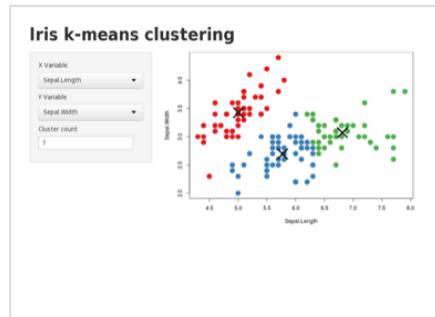
[Movie explorer](#)



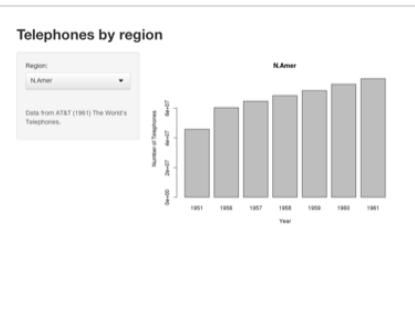
[Google Charts](#)

Start simple

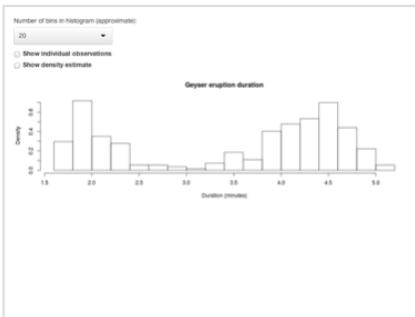
If you're new to Shiny, these simple but complete applications are designed for you to study.



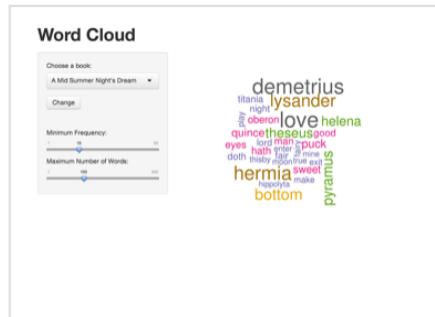
[Kmeans example](#)



[Telephones by region](#)



[Faithful](#)



[Word cloud](#)

R Dashboards

My Dashboard

Search...

Dashboard

Widgets new

Charts

Source code for app

Distribution

A histogram titled "Distribution" showing the frequency of data points across a range from -3 to 3. The x-axis is labeled "data" and the y-axis is labeled "Frequency". The distribution is roughly symmetric and centered around 0.

View 1 View 2

A scatterplot titled "View 1" and "View 2" showing a positive linear relationship between two variables, X and Y. The x-axis ranges from -4 to 3, and the y-axis ranges from 10 to 40. The data points form a dense cloud along a diagonal line.

Histogram control

Count

Range: 100 to 500

Step: 10

Value: 120

Appearance

Fill:

- None
- Blue
- Black
- red

Scatterplot control

Spread

Value: 60%

Twin Cities Buses

Route 2

Show

- Northbound
- Southbound
- Eastbound
- Westbound

Note: a route number can have several different trips, each with a different path. Only the most commonly-used path will be displayed on the map.

Zoom to fit buses

Color	Direction	Number of vehicles
■	Northbound	0
■	Southbound	0
■	Eastbound	5
■	Westbound	4
	Total	9

Refresh interval 1 minute

Data refreshed 0 seconds ago.

Refresh now

Source data updates every 30 seconds.

Especialización en Analítica

Ciencia de los datos aplicada.

Decisiones bajo incertidumbre en las organizaciones.

Sistemas de bases de datos masivos.

Aprendizaje de máquinas para datos masivos.

Optimización y simulación.

Modelado predictivo y series de tiempo.

Inteligencia de negocios.

Gracias por su atención

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>