DATA ANALYTICS Y 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 Science vs Analytics

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?)

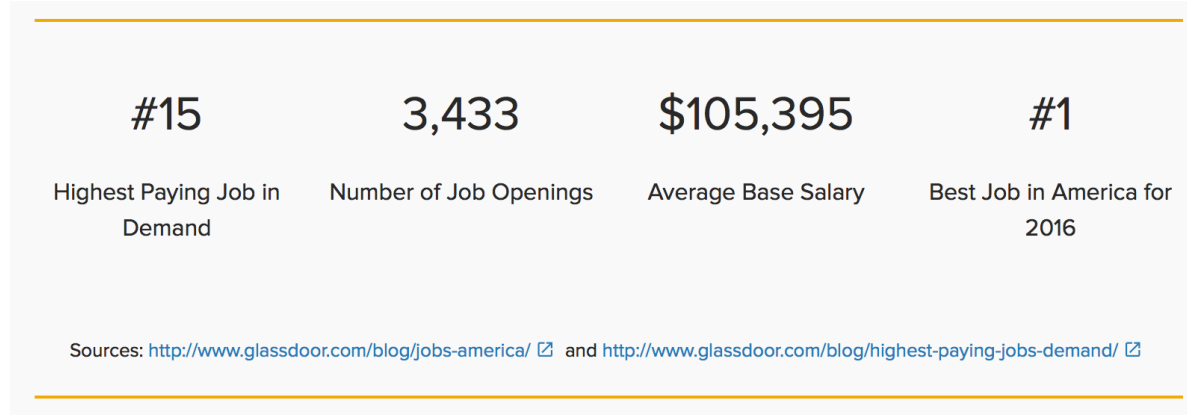
Data Mining
Proceso de descubrimiento de patrones y tendencias útiles en grandes conjuntos de datos.

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

I. Problema organizacional
II. Transormación 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

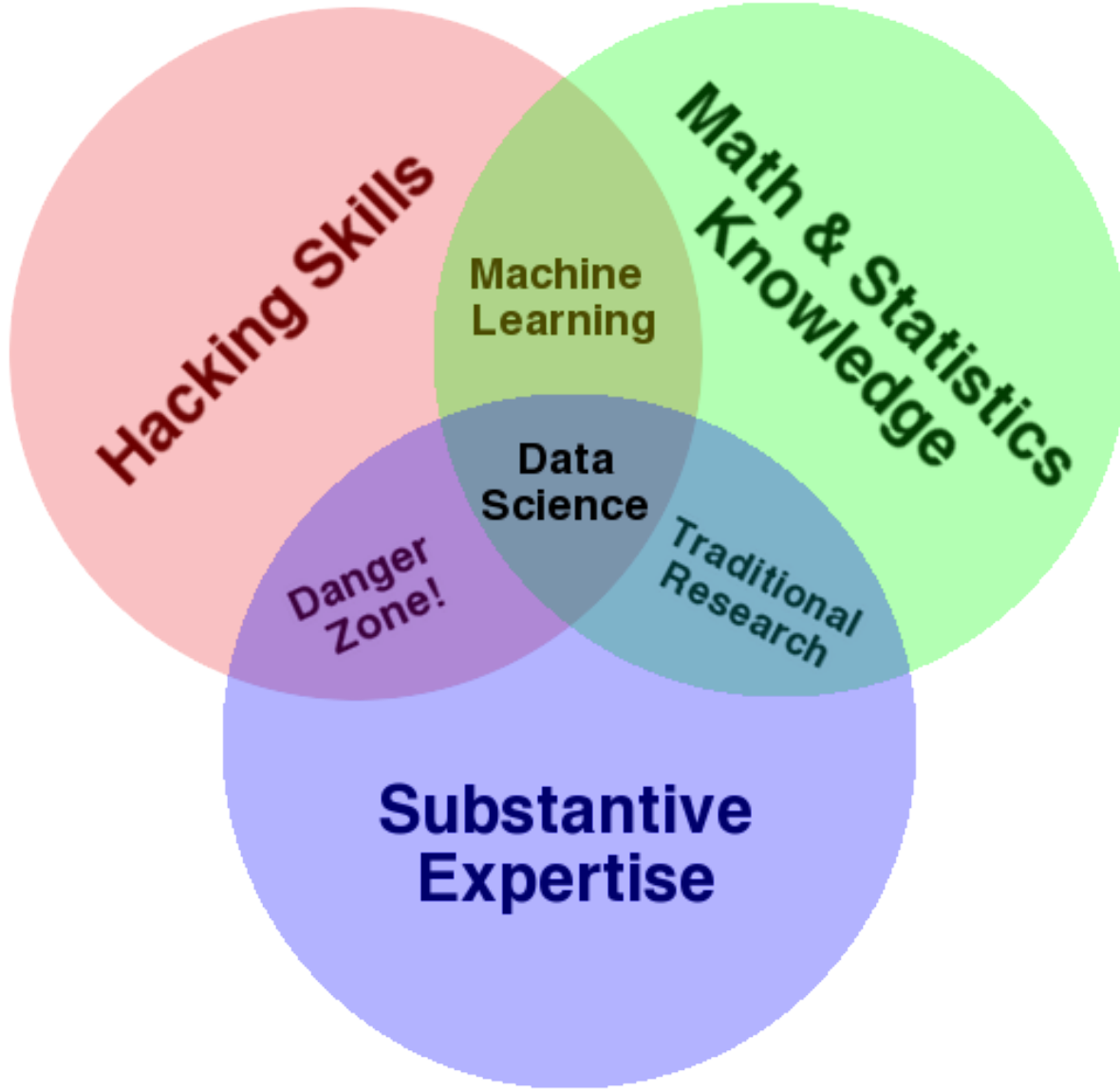
Perspectiva

> 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."



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

Diagrama de Ven explicando qué es Data Science



Data Science and Data Scientists: What's in a Name?
Saunders, 2013

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

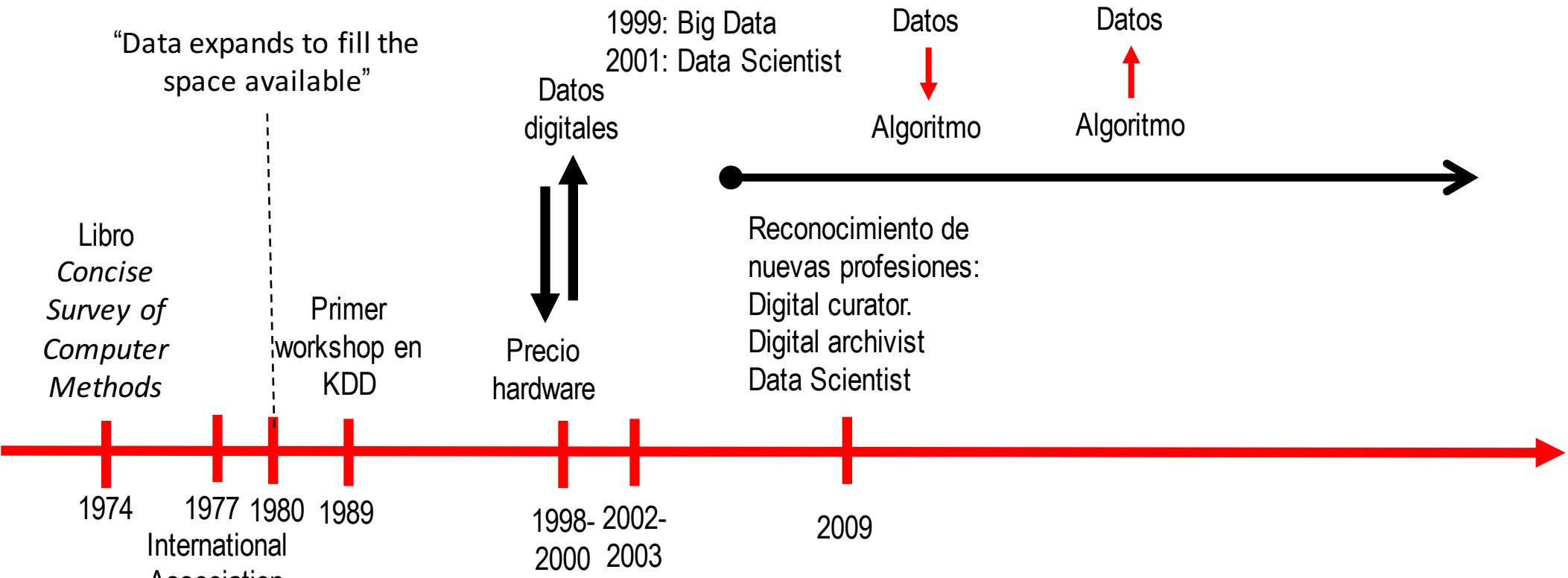
Habilidades en la programación de computadores para manejo de datos y modelado predictivo (estadística, aprendizaje de máquinas, minería de datos, etc.).

Analytics

Data Science + Optimización + Simulación

DISCIPLINE	TECHNOLOGIES	SKILLS	FOCUS
BUSINESS INTELLIGENCE	• ETL Tools / SQL • RDBMS • Reporting • Visualization	• Programming • Data Analysis • Data Modeling • Report Development • Basic Statistics • Technical Architecture • Business Analysis & Strategy • Presentation	• Information Delivery and Reporting • Data Visualization • Descriptive Statistics • Data Integration and Consolidation
DATA ANALYSIS	• Data Modeling Software • Diagraming Software • Documentation Software • SQL • Data Profiling Software	• Data Modeling • Business Analysis • Data Manipulation • Basic statistics	• Business Rules • Data Definitions and Lineage • Data Entity Relationships • Data Attributes • Data Structures • Sources and Targets of Data • Data Quality
DATA SCIENCE	• Statistics Software • Columnar Data • Map-Reduce • NoSQL • Programming Languages • Graphing/Charting Software	• Advanced Statistics • Programming • Business Analysis • Modern Data Management Technologies and Architectures	• Predictive Modeling • Advanced Statistical Analysis • Data Mining • Unstructured Data Management • Large data volumes • Research

Línea de tiempo. Data Science / Big Data



Data Science Journal: management of data and databases in Science and Technology. The scope of the Journal includes descriptions of data systems, their publication on the internet, applications and legal issues

Journal of Data Science: Data Science' we mean almost everything that has something to do with data: Collecting, analyzing, modeling…… yet the most important part is its applications — all sorts of applications. This journal is devoted to applications of statistical methods at large…. The Journal of Data Science will provide a platform for all data workers to present their views and exchange ideas.

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.

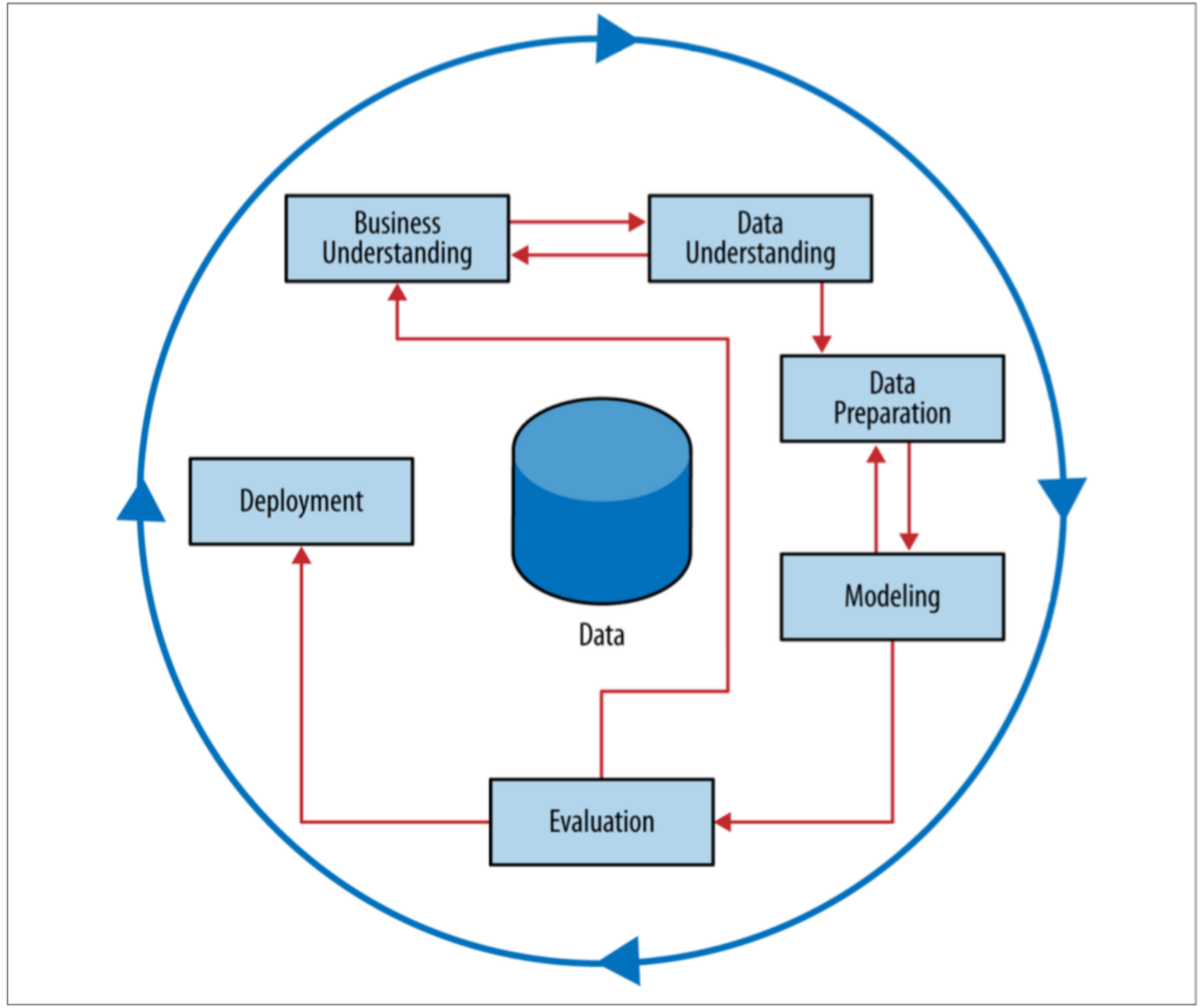
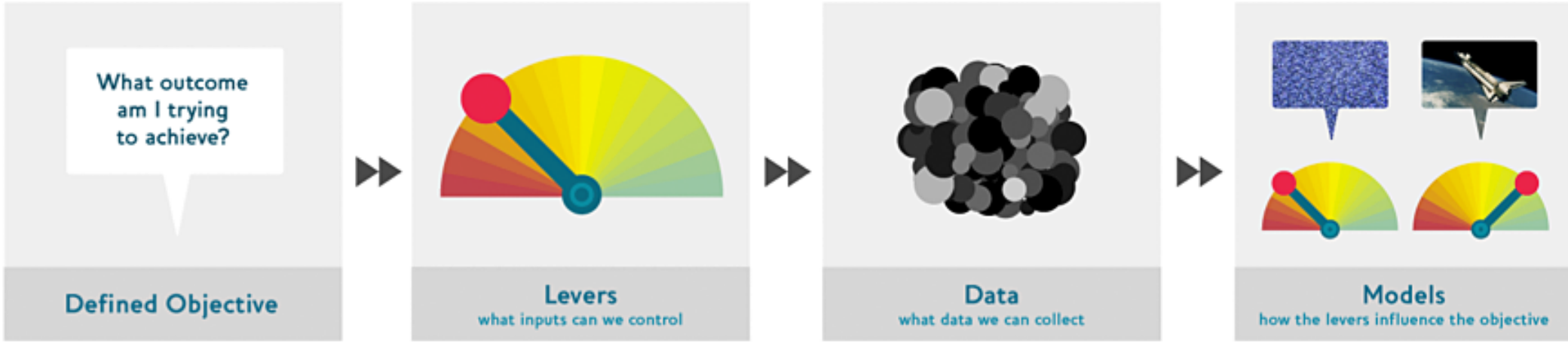
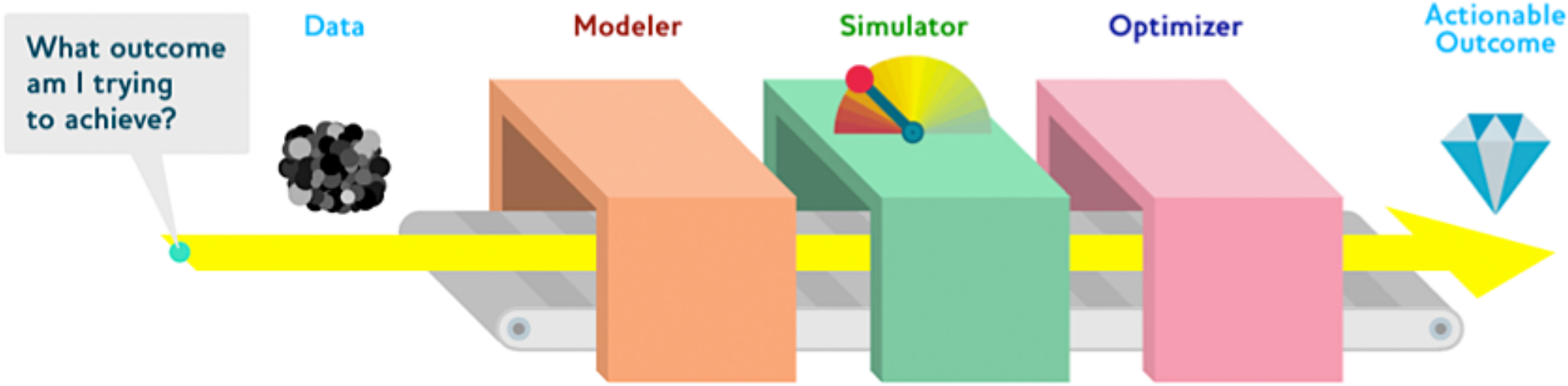


Figure 2-2. The CRISP data mining process.

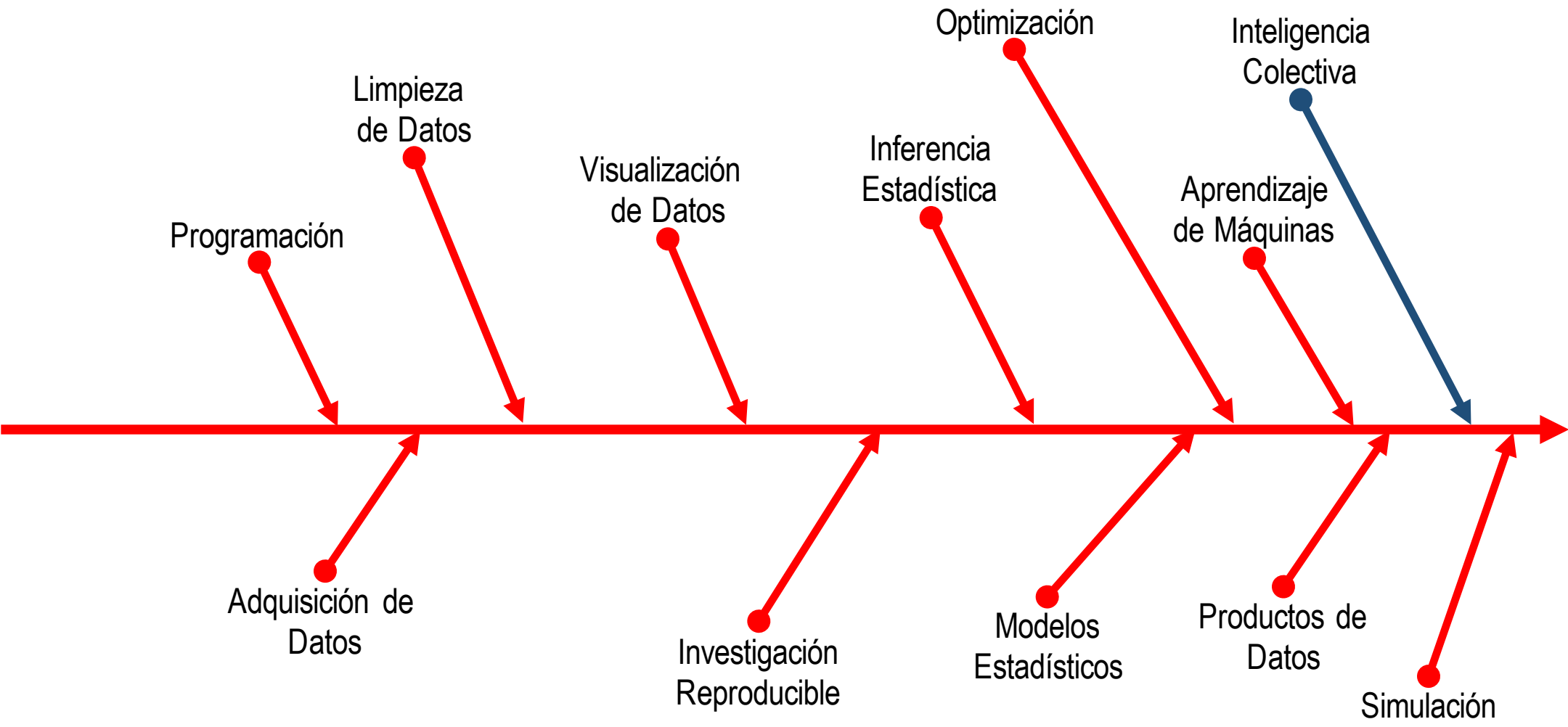
Designing great data products

http://radar.oreilly.com/2012/03/drivetrain-approach-data-products.html





Data Science / Analytics



Data-driven decision making!

¿Usted sabe programar … / Es capaz de …?

¿Ordenar un vector de números?

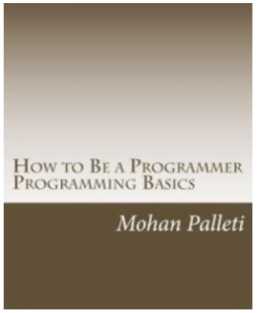
¿Extraer la tercera línea de texto de un conjunto de archivos?

¿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…

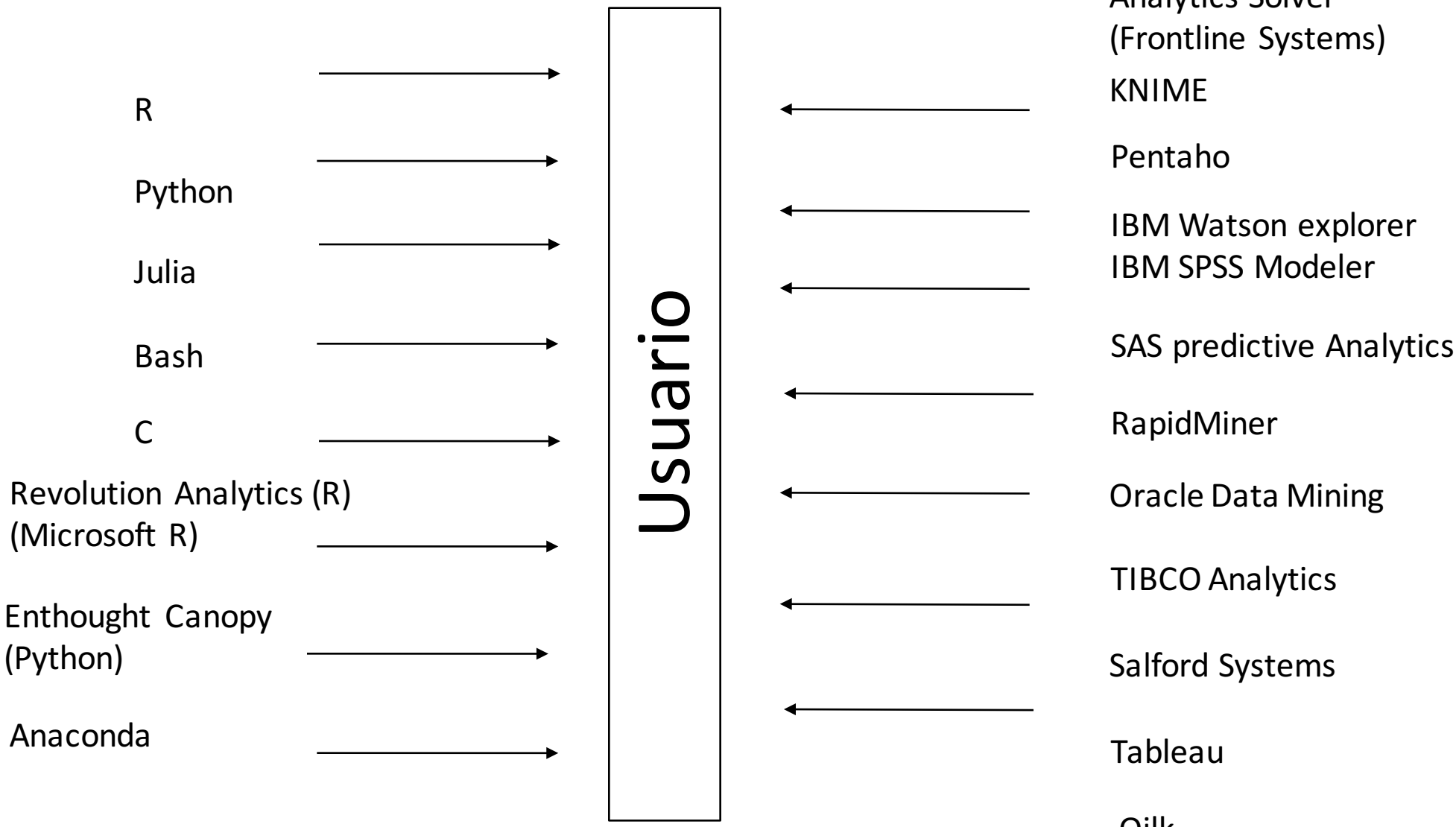


How to Be a Programmer: Programming Basics

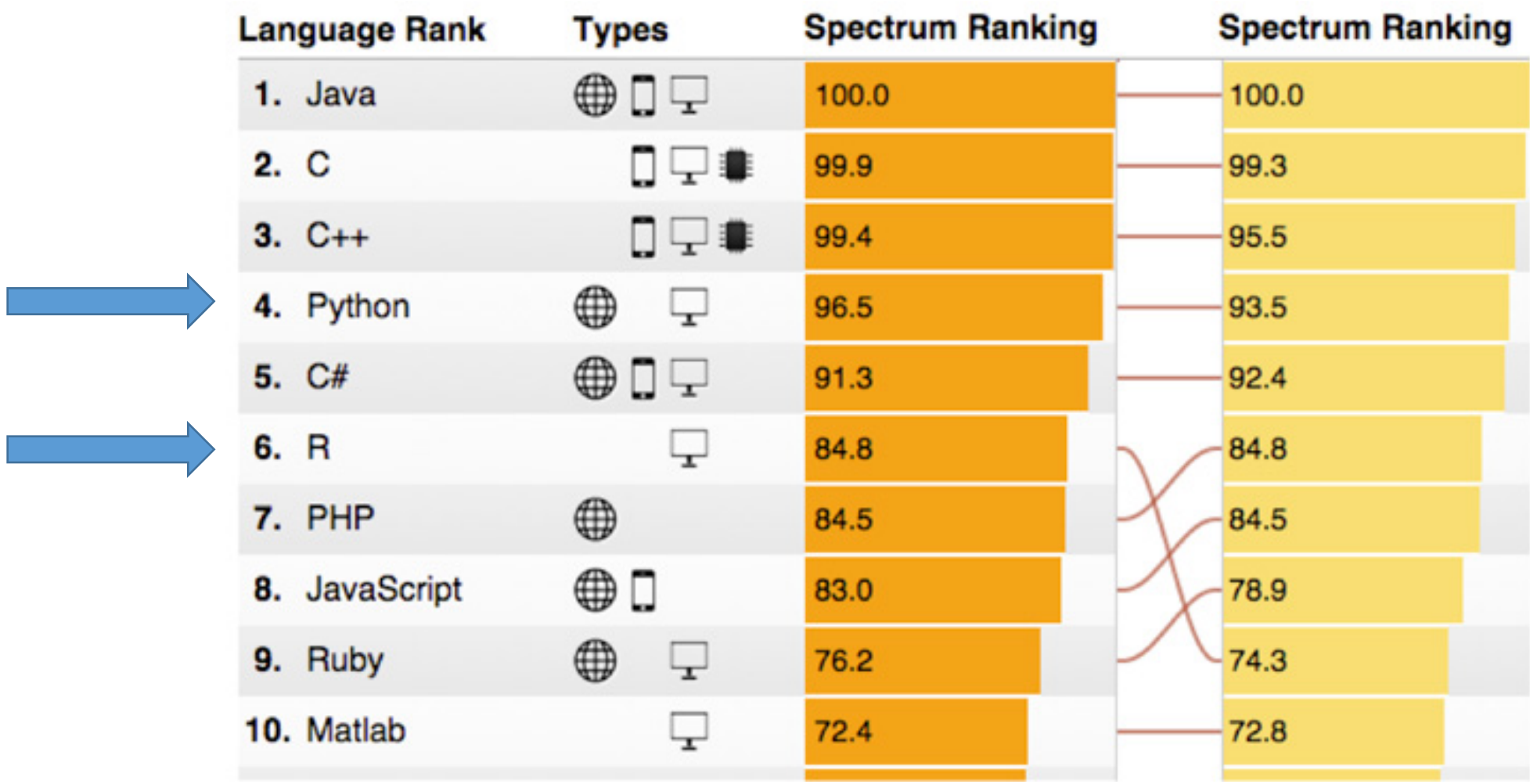
By: Mohan Palleti

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

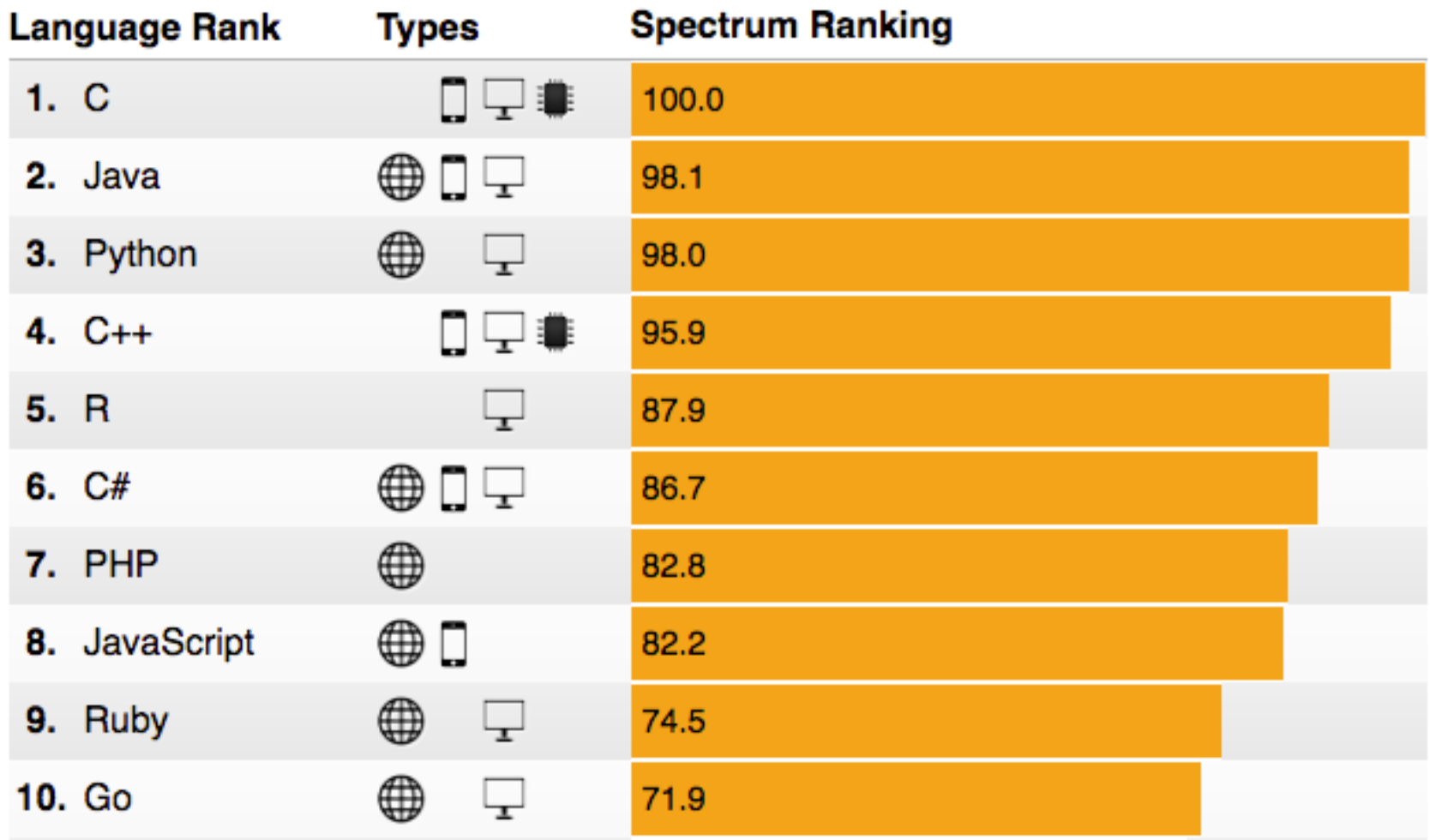


The **2015** Top Ten Programming Languages (IEEE Spectrum)



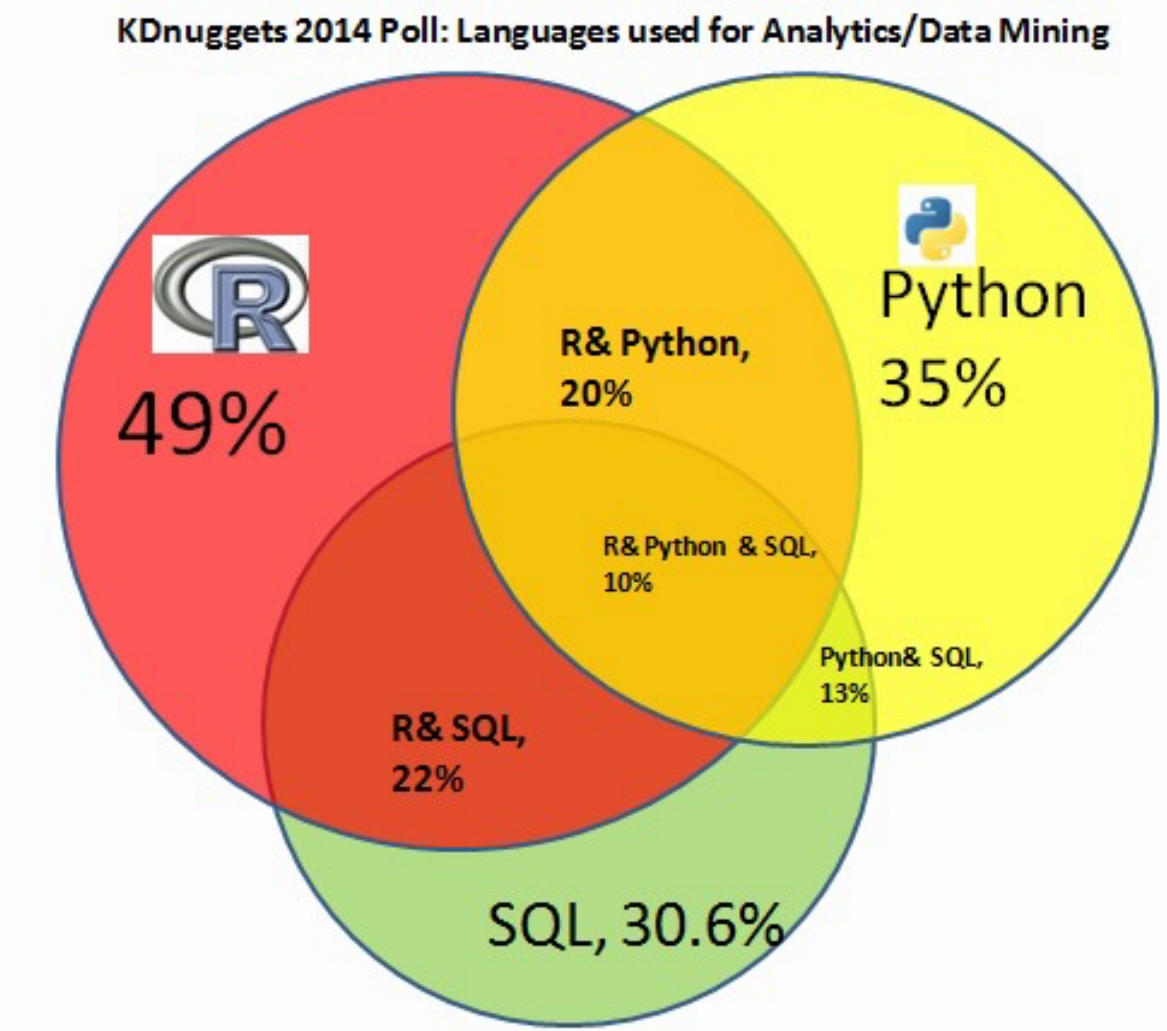
http://spectrum.ieee.org/computing/software/the-2015-top-ten-programming-languages

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



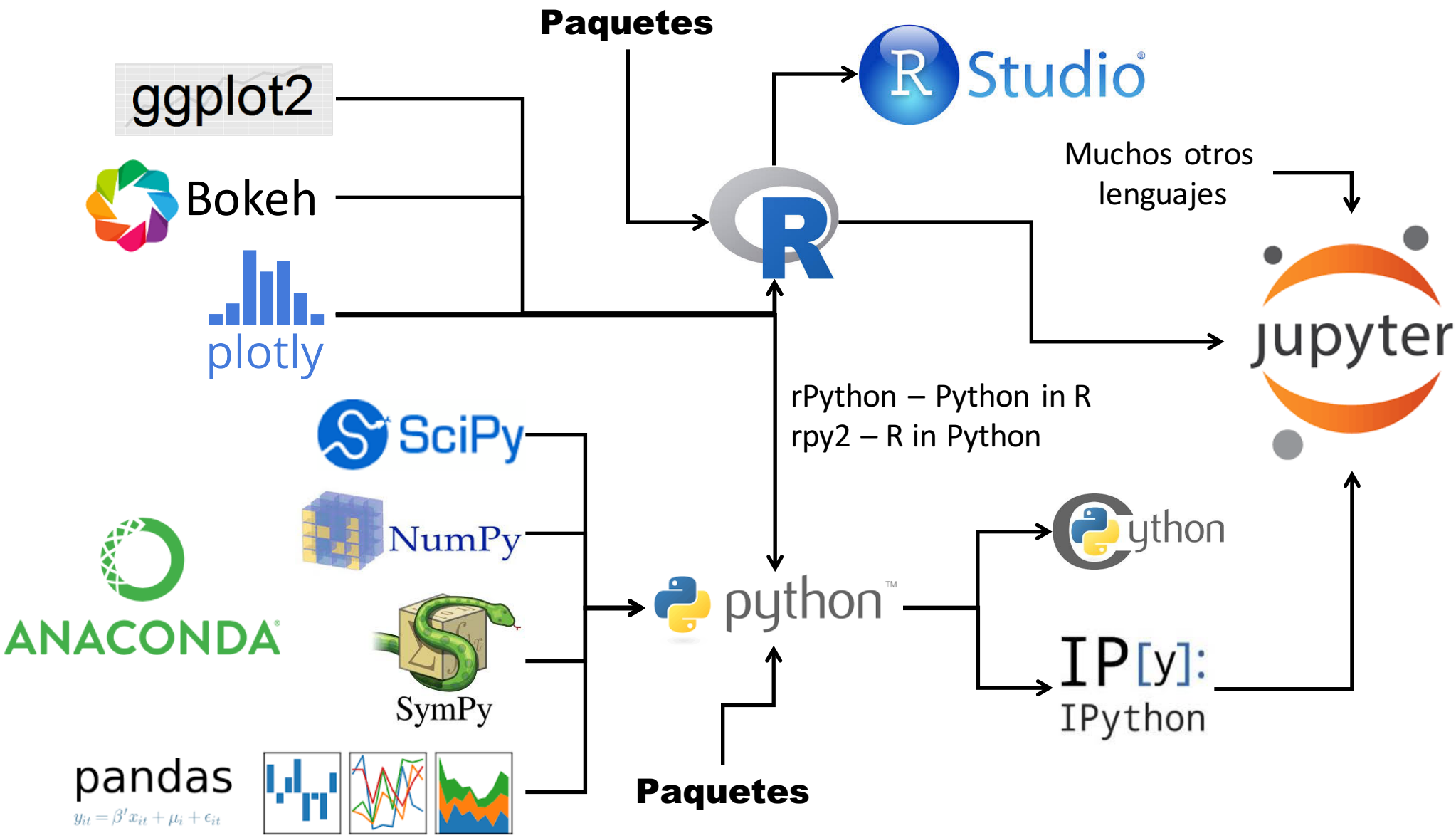
http://spectrum.ieee.org/computing/software/the-2016-top-programming-languages

Popularidad de los lenguajes

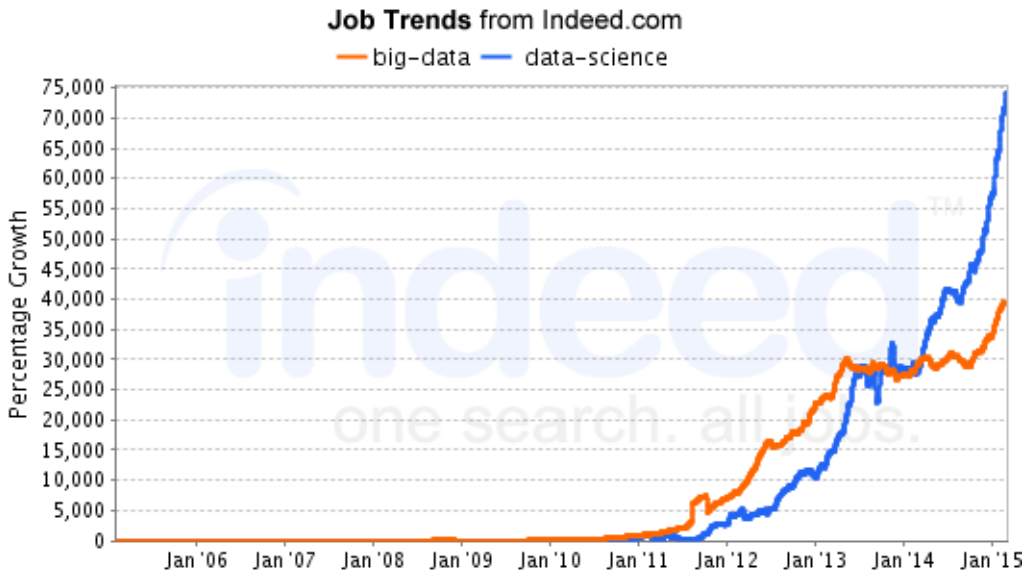


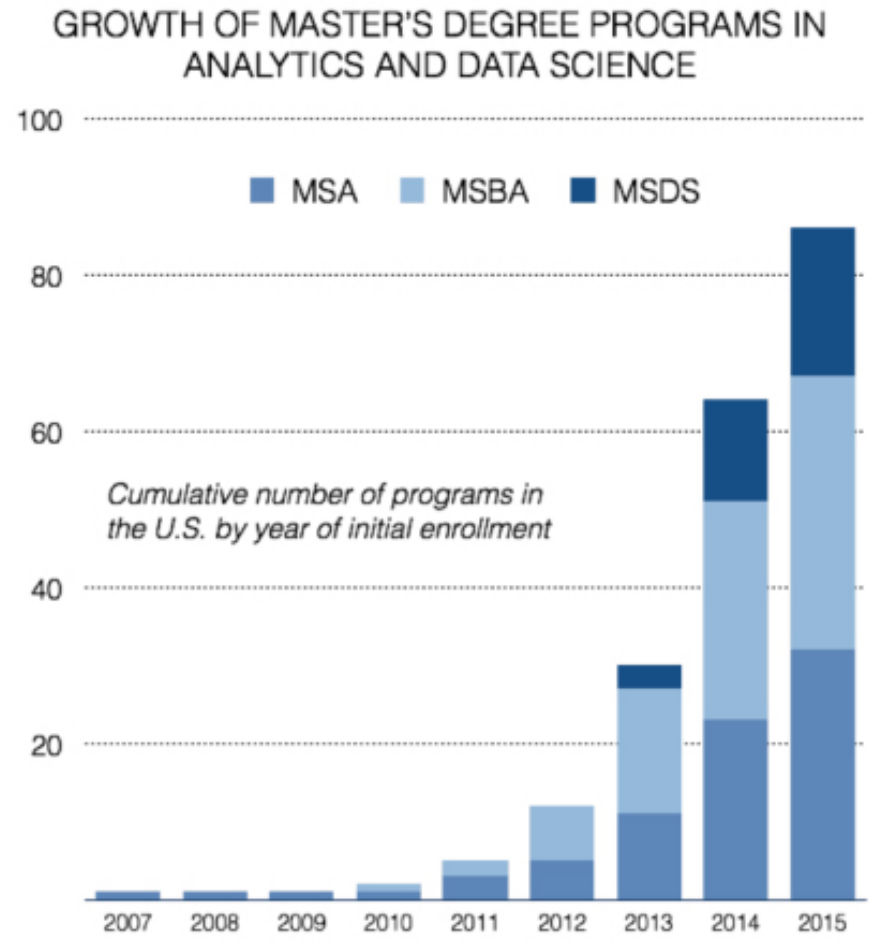
http://www.kdnuggets.com/2014/08/four-main-languages-analytics-data-mining-data-science.html

Ecosistema de computación científica: Python y R



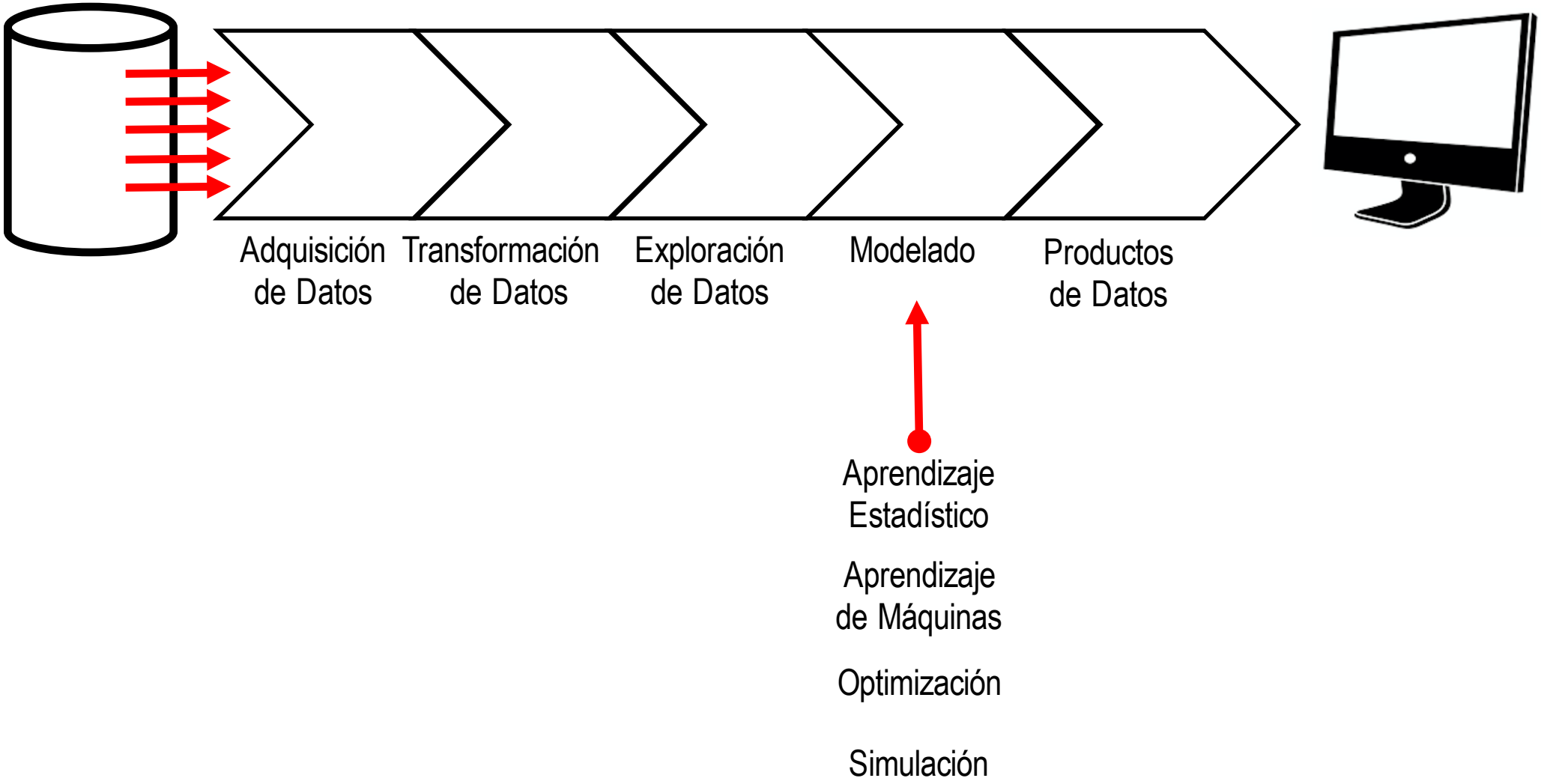
Evolución de empleos/educación en Big Data & Data Science



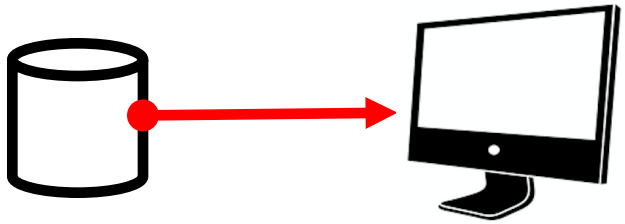


http://analytics.ncsu.edu/?page_id=4184

Fases en Data Science / Analytics

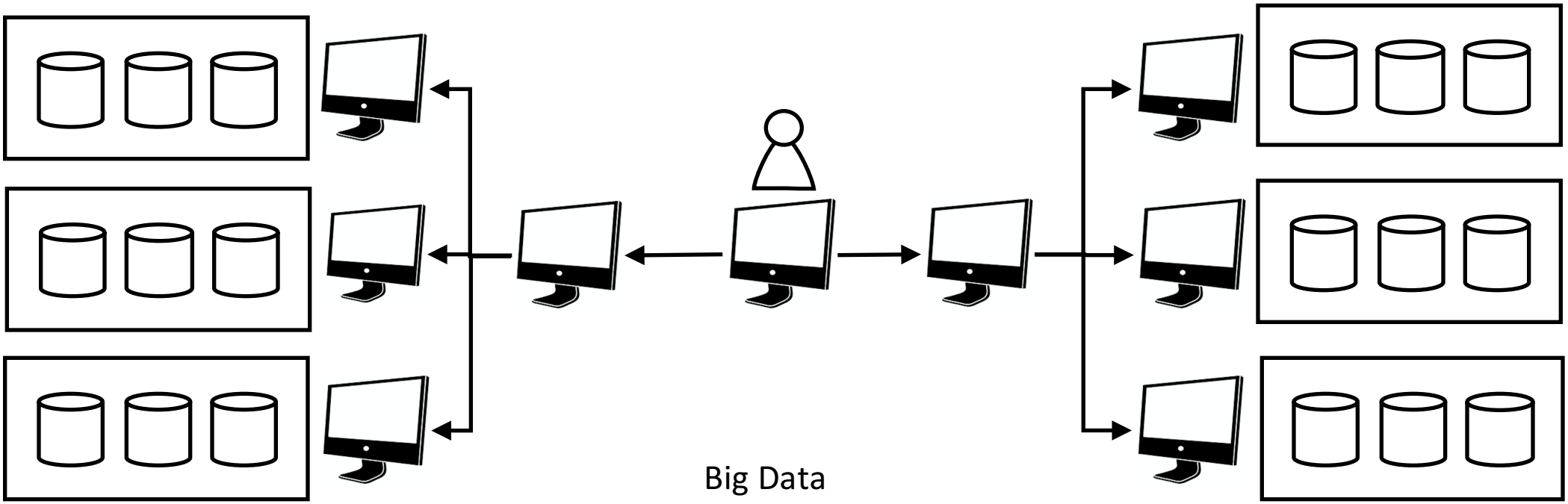


Aproximación tradicional vs Big Data



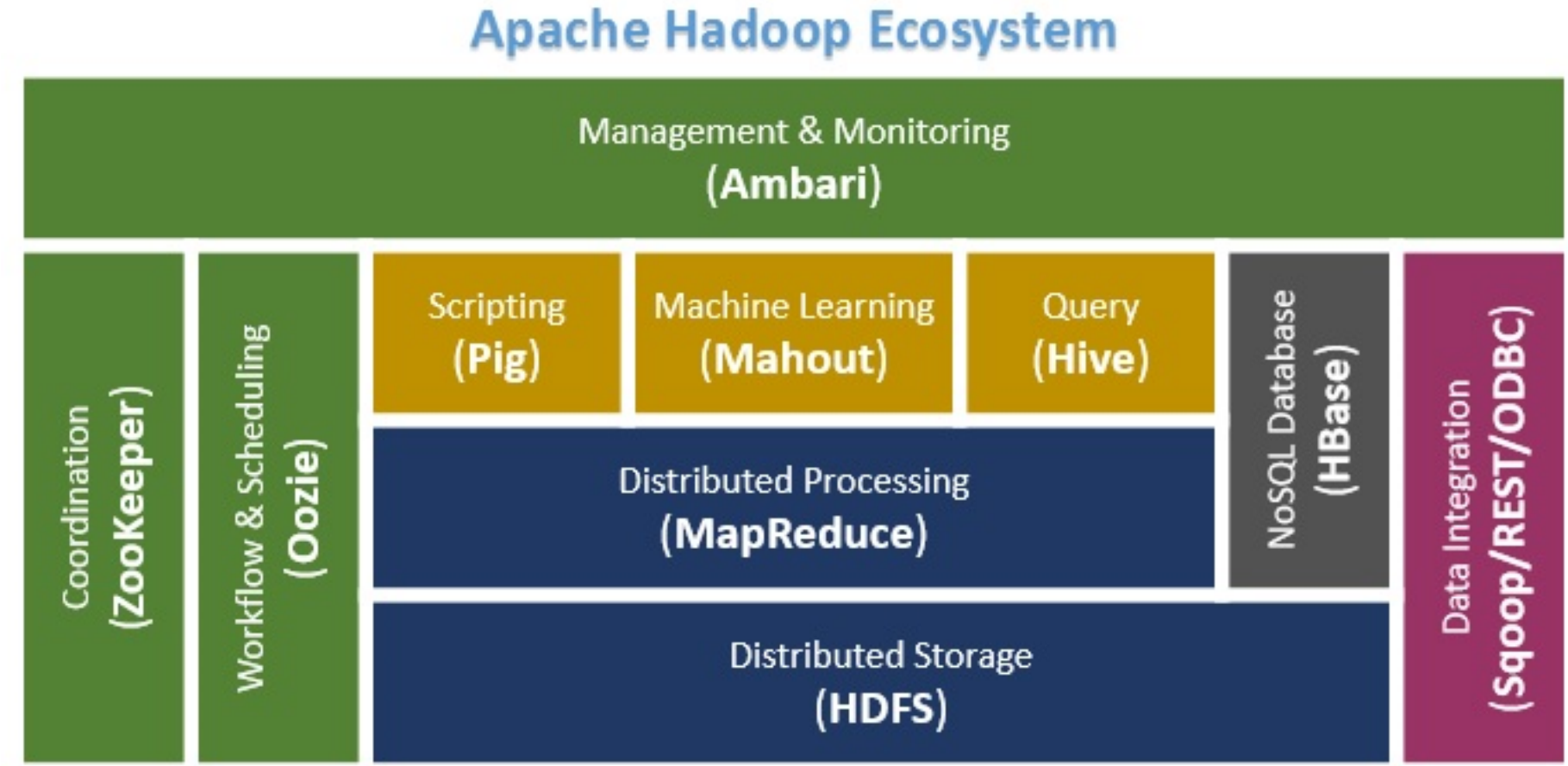
Programación Tradicional

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



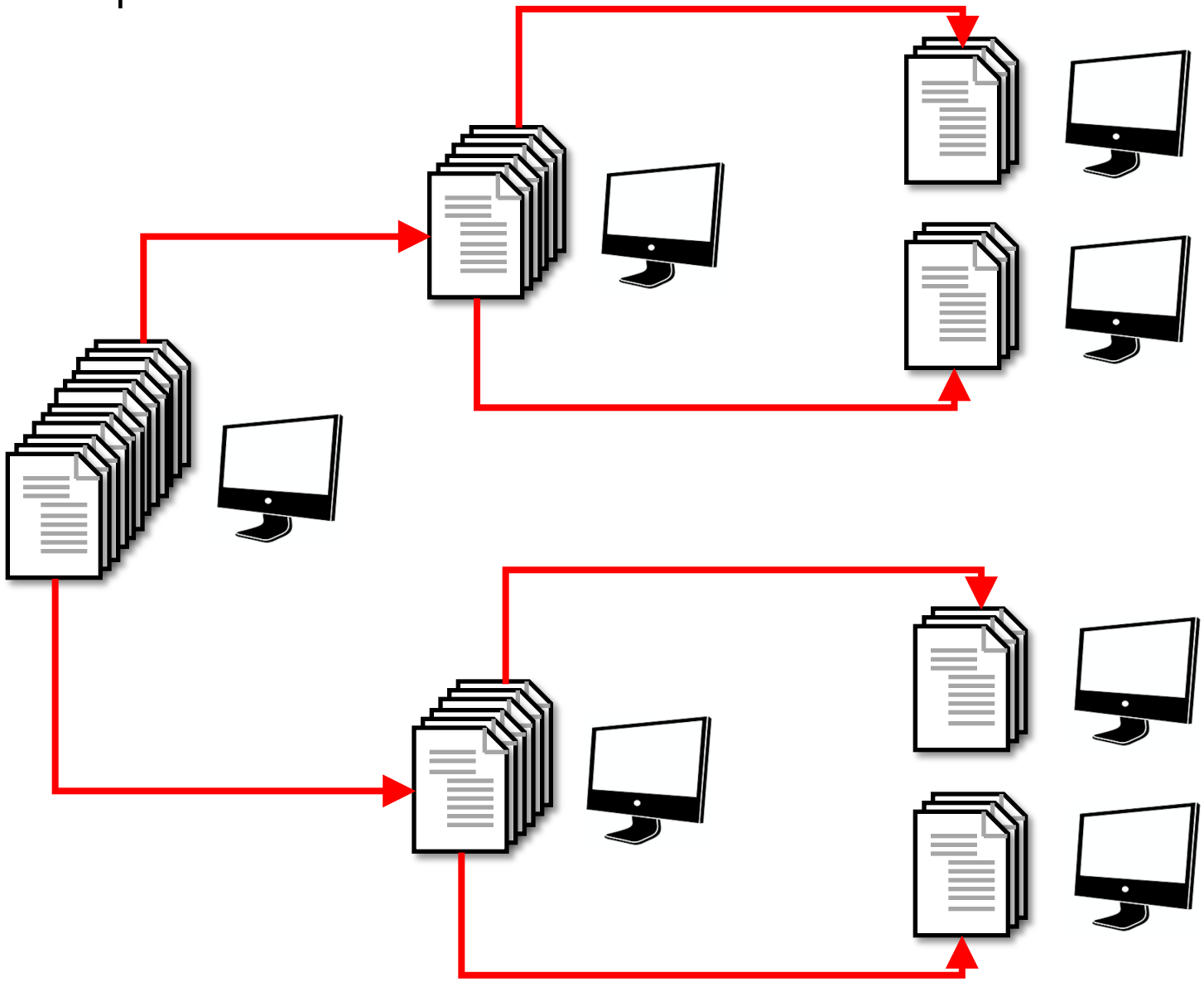
Big Data
Nodos de Datos + Nodos de Computo

Big Data – Apache Hadoop



https://www.dezyre.com/article/hadoop-components-and-architecture-big-data-and-hadoop-training/114

MapReduce



Adquisición y Limpieza de Datos

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

Páginas web (HTML) y Google Groups.

Bases de datos relacionales.

Lenguaje Natural.

Imágenes (Captcha)

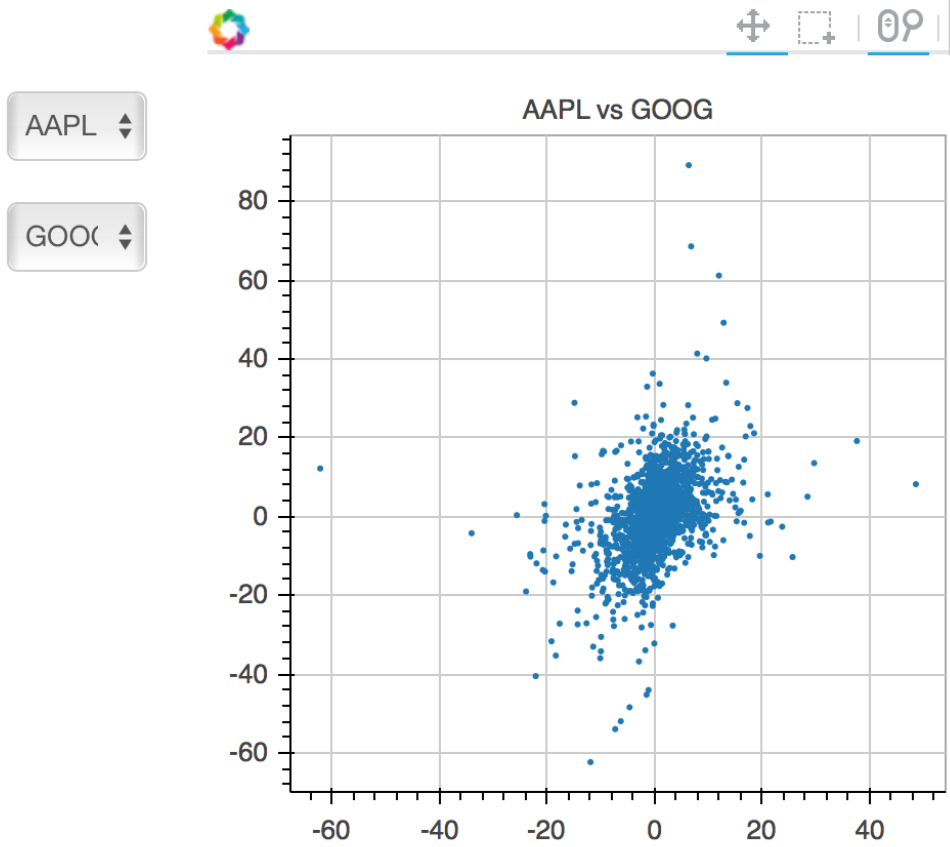
Manipulación de texto

Conversión de un formato a otro

Detección de datos faltantes, datos nulos, datos inconsistentes

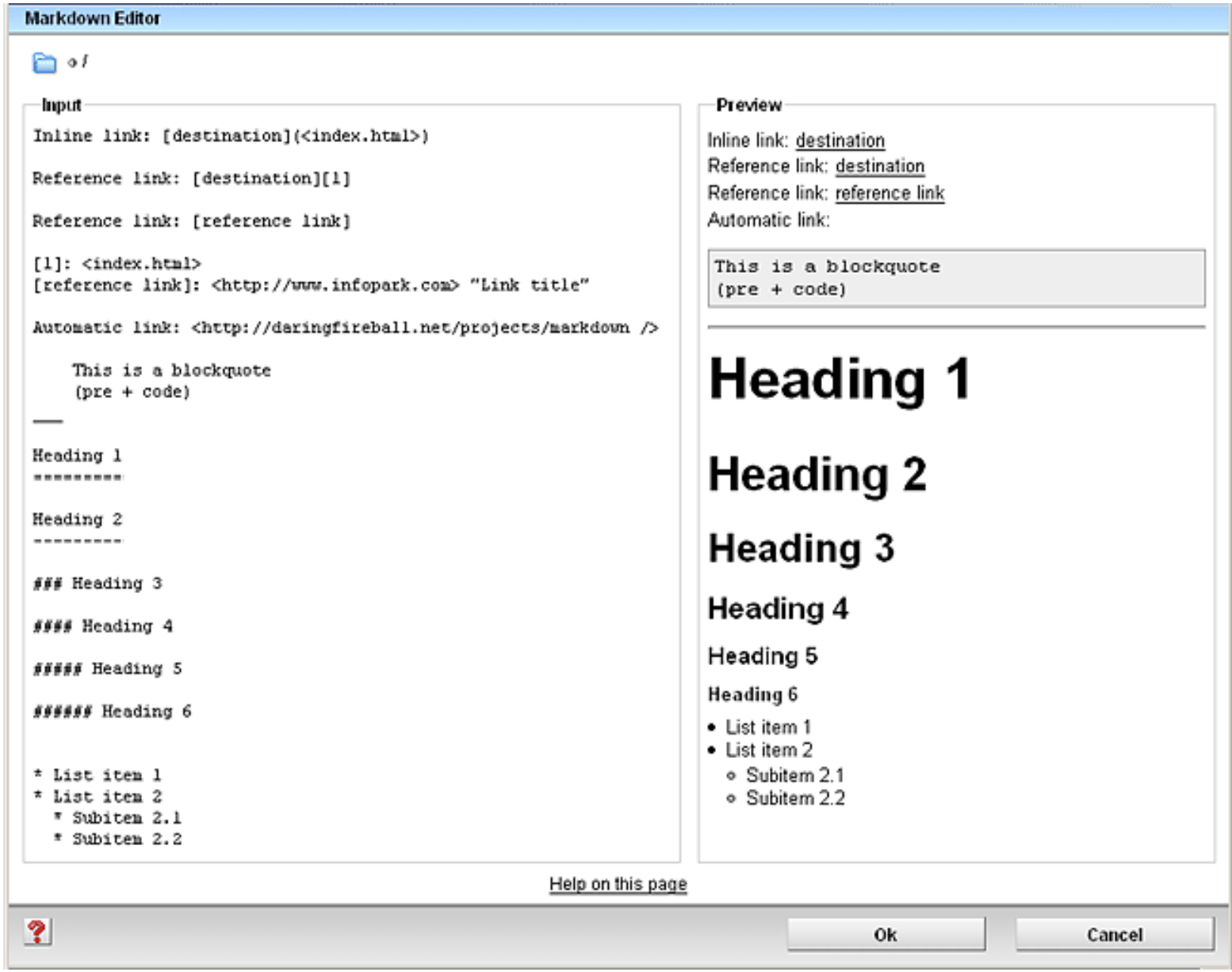
Visualización de Datos

link to this

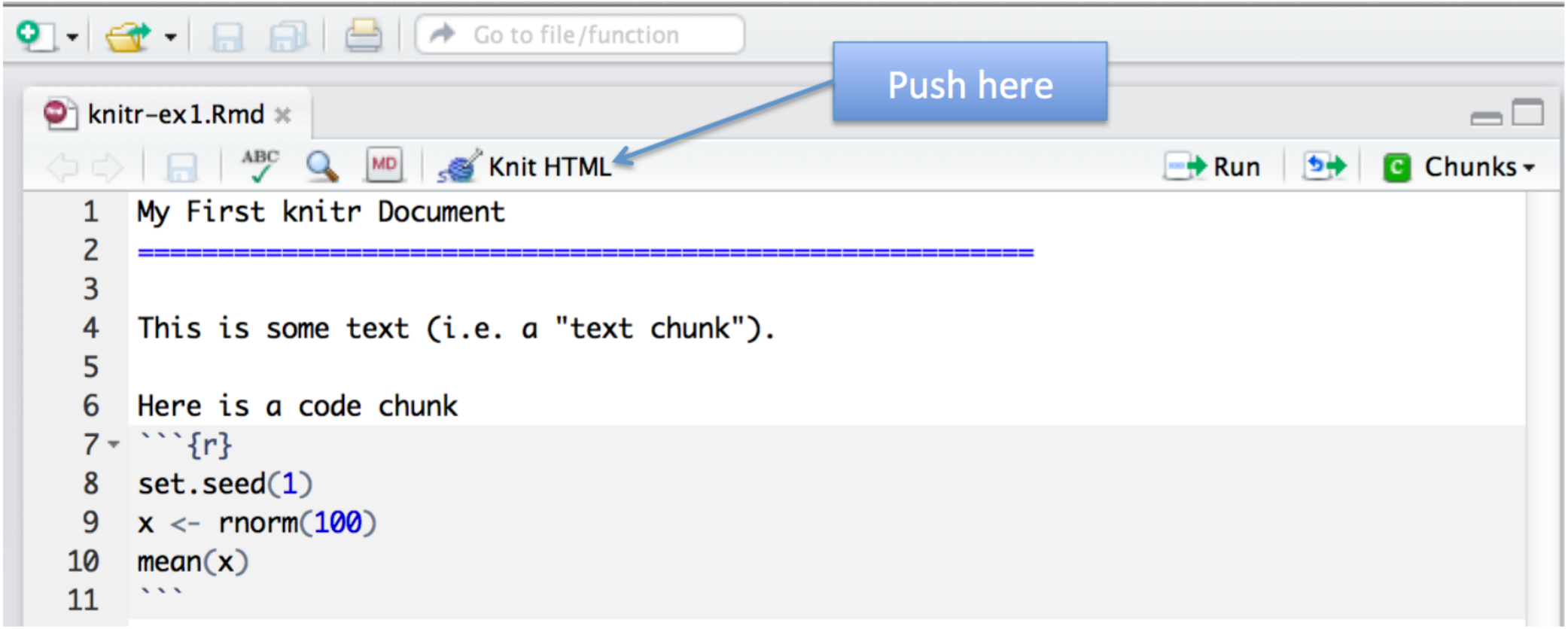


	AAPL	AAPL_returns	GOOG	GOOG_returns
nt	2258.000000	2258.000000	2258.000000	2258.000000
n	218.713048	0.194594	494.682700	0.346306
	171.674172	5.239248	164.446873	9.411532
	15.108500	-62.237000	99.950000	-62.400000
	80.502825	-1.631675	393.127500	-4.095000
	163.155500	0.146000	499.785000	0.240000
	338.082500	2.155750	595.465000	4.957500
	685.662000	48.618000	923.670000	89.310000

Investigación Reproducible (Markdown)



Investigación Reproducible (Markdown + R)



Investigación Reproducible (Markdown + R)

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)

localhost

```
import scipy
import sys

# make nice plots
import plt_fmt
```

```
Populating the interactive namespace from numpy and matplotlib
```

"m" key denotes a markdown cell

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

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

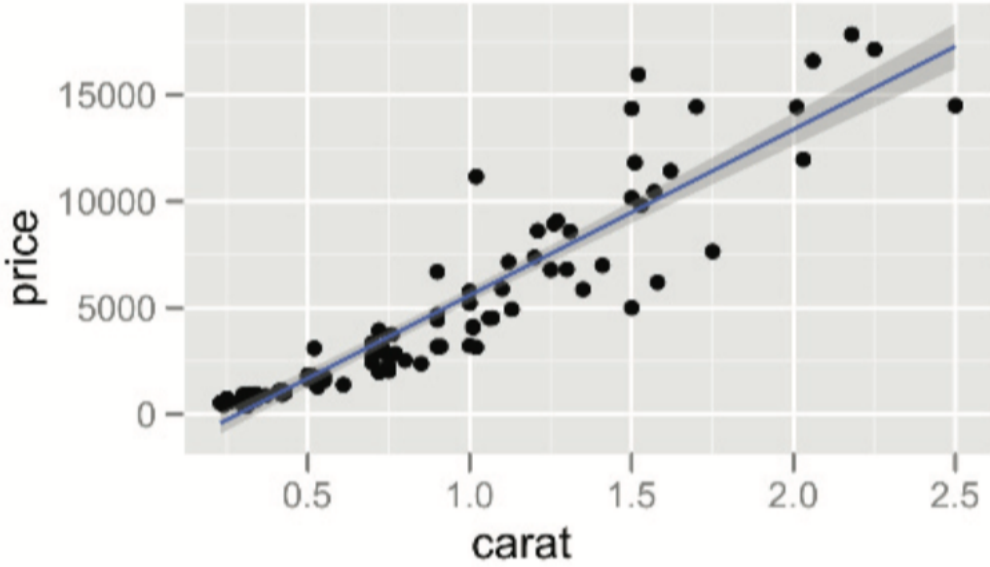
```
[ 0.20757795  0.01992547]
0.207577947999
0.019925471486
```

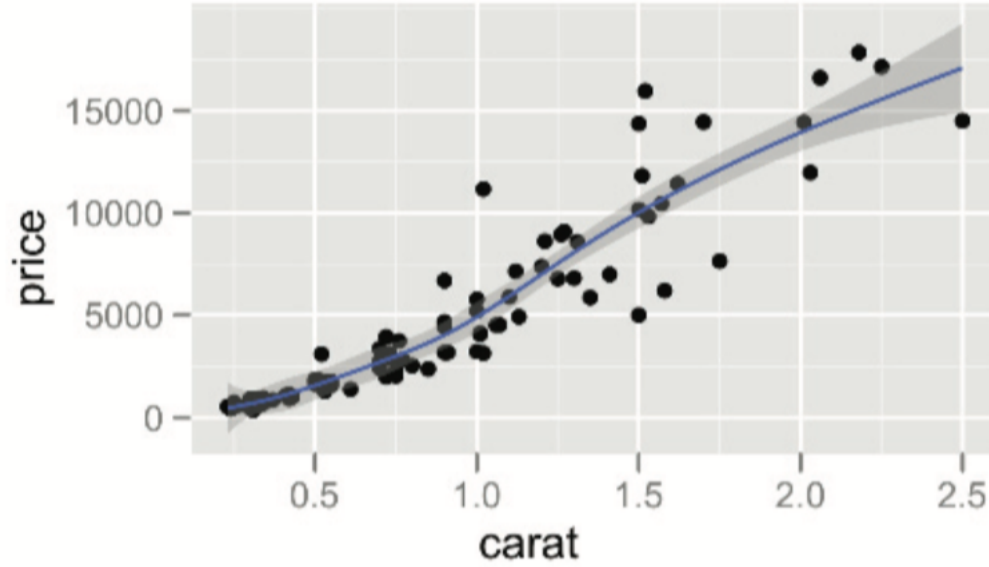
In [4]:
```
def vfield(n,time, param):
    """
    param is an Nx2 matrix specifying the parameters for
    the dynamical system
    """

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

Out[4]:
```
[<matplotlib.lines.Line2D at 0x106695048>]
```

Modelos Estadísticos

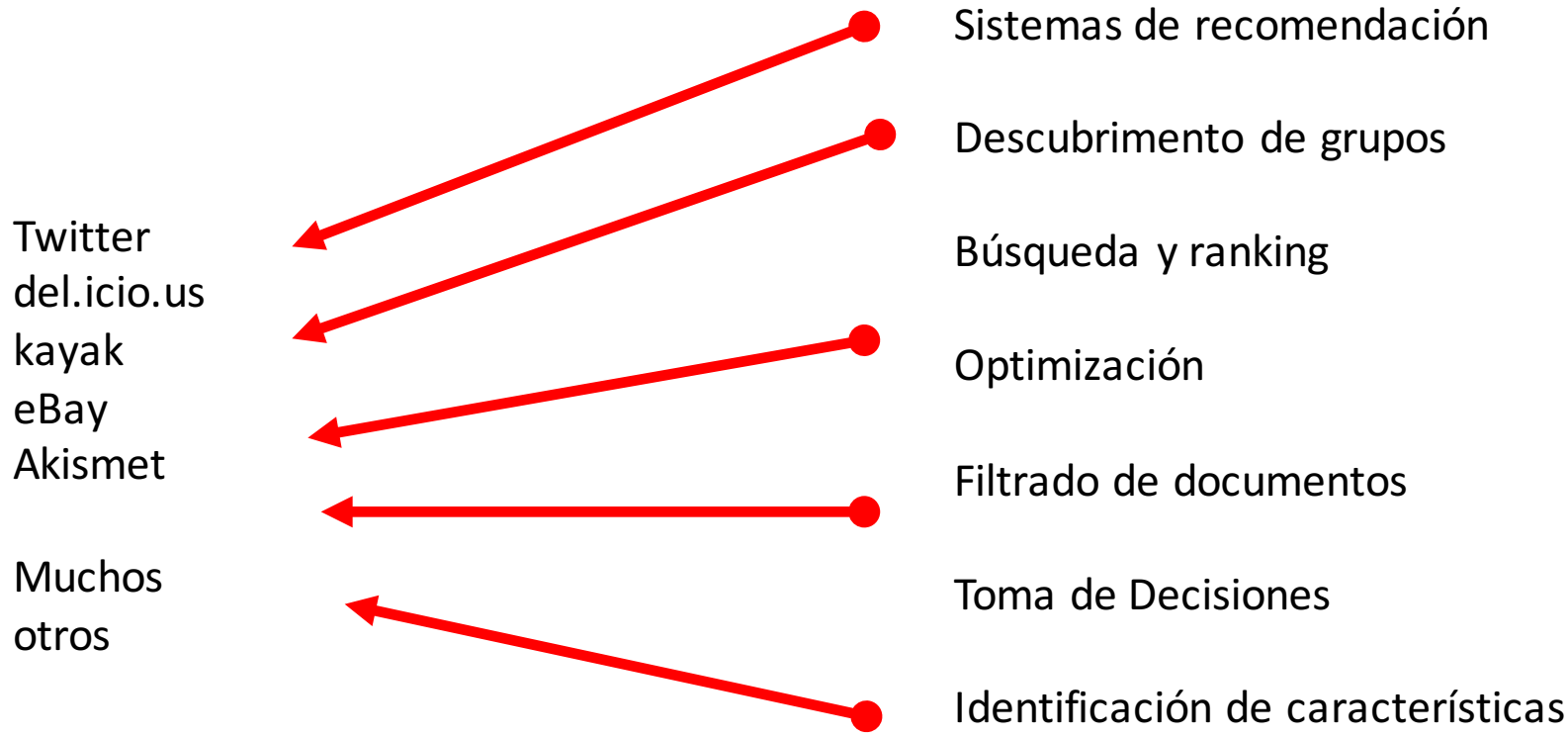
Aplicación clásica





¿Y si hay 10 millones de datos?

Inteligencia Colectiva



Especialización en Analítica

Ciencia de los datos aplicada.

Decisiones bajo incertidumbre en las organizaciones.

Sistemas de bases de datos masivos.

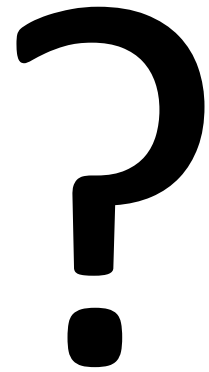
Aprendizaje de máquinas para datos masivos.

Optimización y simulación.

Modelado predictivo y series de tiempo.

Inteligencia de negocios.

En resumen….



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