

Minería de Datos para el Análisis de Big Data

Por: Carlos Carreño
ccarrenovi@gmail.com

Abril, 2021

Acerca del Instructor

- Carlos Carreño, *ccarrenovi@Gmail.com*
- Ingeniero de Sistemas, Físico Matemático, ***Data Scientist (UAH)***.
- Certificaciones: OCP, ScrumMaster, RHCJA, RHC BPM, RHCB RMS, otros
- Oracle WDP Instructor
- Red Hat Certified Instructor Latam/Spain
- NIIT India Instructor
- Red Hat Consulting Arquitecto de Soluciones
 - RHPAM, Nginx, Kafka, MS Azure, AWS
 - Red Hat Openshift, Kubernetes
 - Proyectos de Desarrollo e Implementación



Modulo 1 Introducción a la Minería de Datos

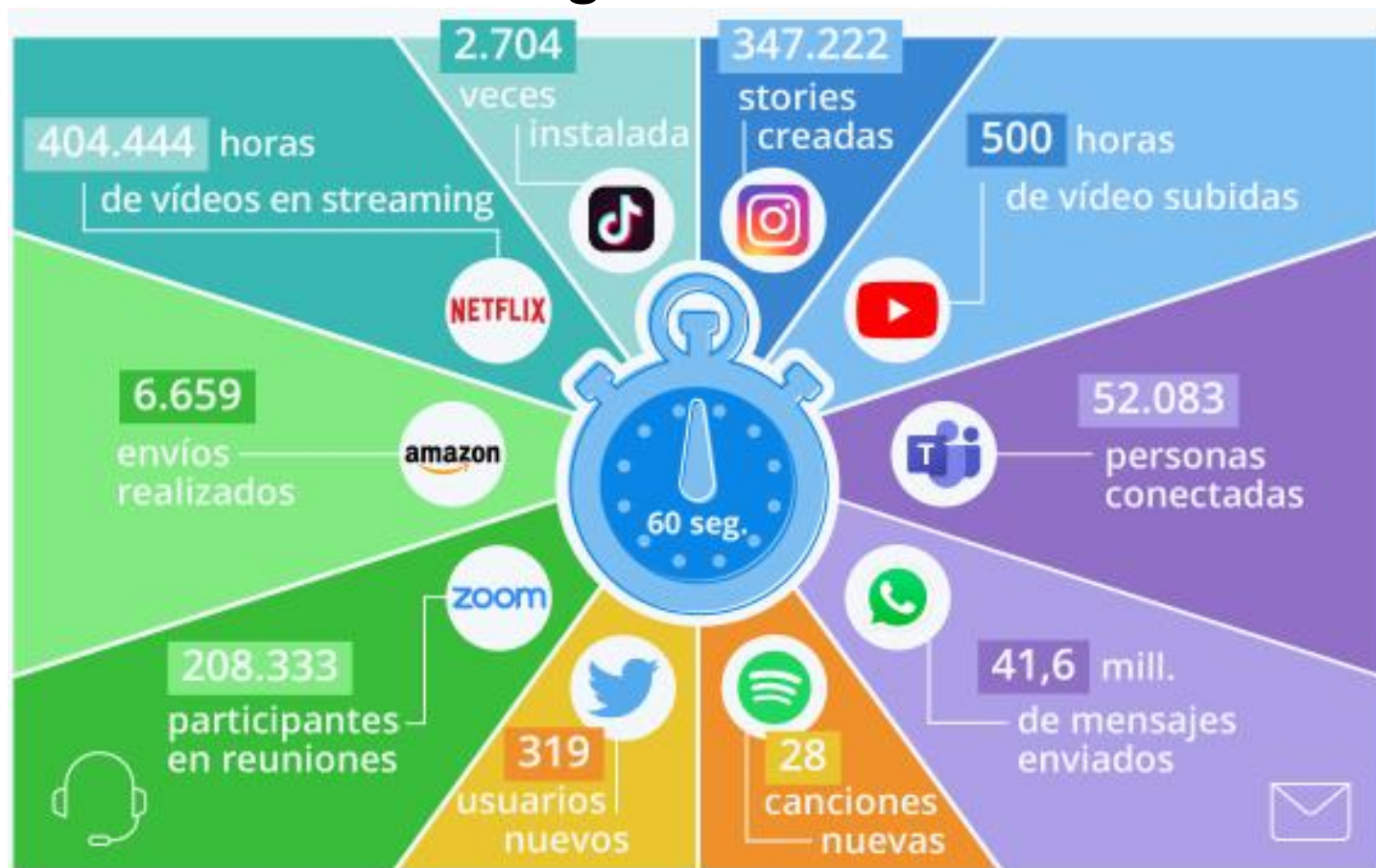
- Fundamentos
- Modelos de predicción: clasificación, regresión y series temporales
- Preparación de datos
- Modelos de agrupamiento o segmentación
- Modelos de asociación
- Modelos avanzados de minería de datos
- Big data

Fundamentos

- Tsunami de Datos
- Estructura de Datos
- Escenarios y Datos
- Que es la minería de datos?
- Como es la información obtenida del Data Mining
- Proceso KDD
- Data Mining y Big Data

Tsunami de Datos

- Cantidad de datos generados en internet en 1 minuto, 2020

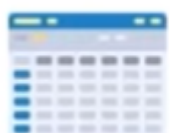
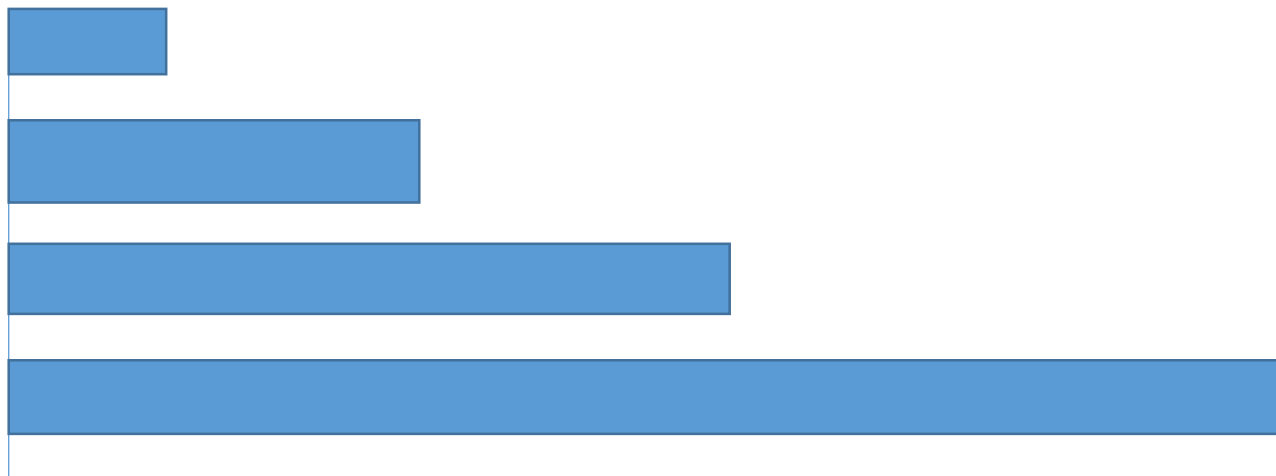


- Los datos están en todas partes y crecen de forma exponencial
- Los datos se están generando desde múltiples fuentes y en múltiples formatos

Ref. <https://es.statista.com/grafico/17539/datos-creados-online-en-un-minuto/>

Estructura de los Datos

- Datos estructurados
- Datos Semi Estructurados
- Datos Cuasi Estructurados
- Datos no Estructurados



Escenarios y Datos

- Tengo datos de finanzas, necesito encontrar si algunas transacciones son fraudulentas.
- Necesito saber que correos son spam
- Tengo los datos de las ventas en línea, que productos podrían venderse mejor juntos.



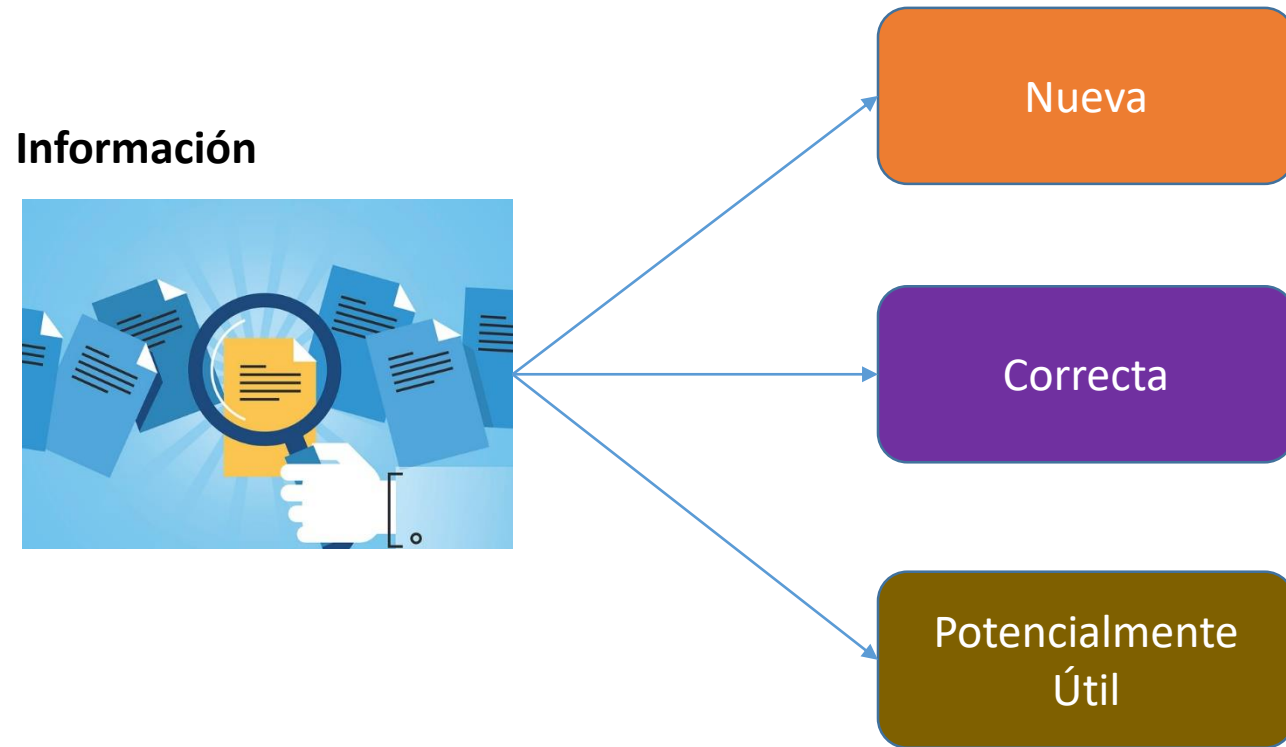
Que es Data Mining?

- Que es la Minería de Datos o Data Mining ?

Data Mining es el proceso mediante el cual podemos descubrir patrones en grandes bases de datos, involucra métodos de la intersección de **machine learning, estadística y de sistemas de base de datos**

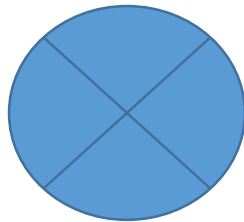
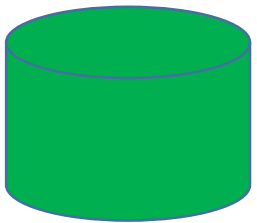


Como debería ser la información de obtenida del Data Mining

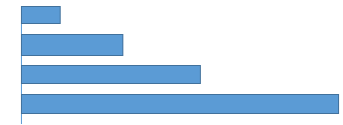
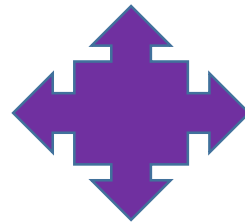


Proceso KDD

- Proceso Knowledge Discovery Database

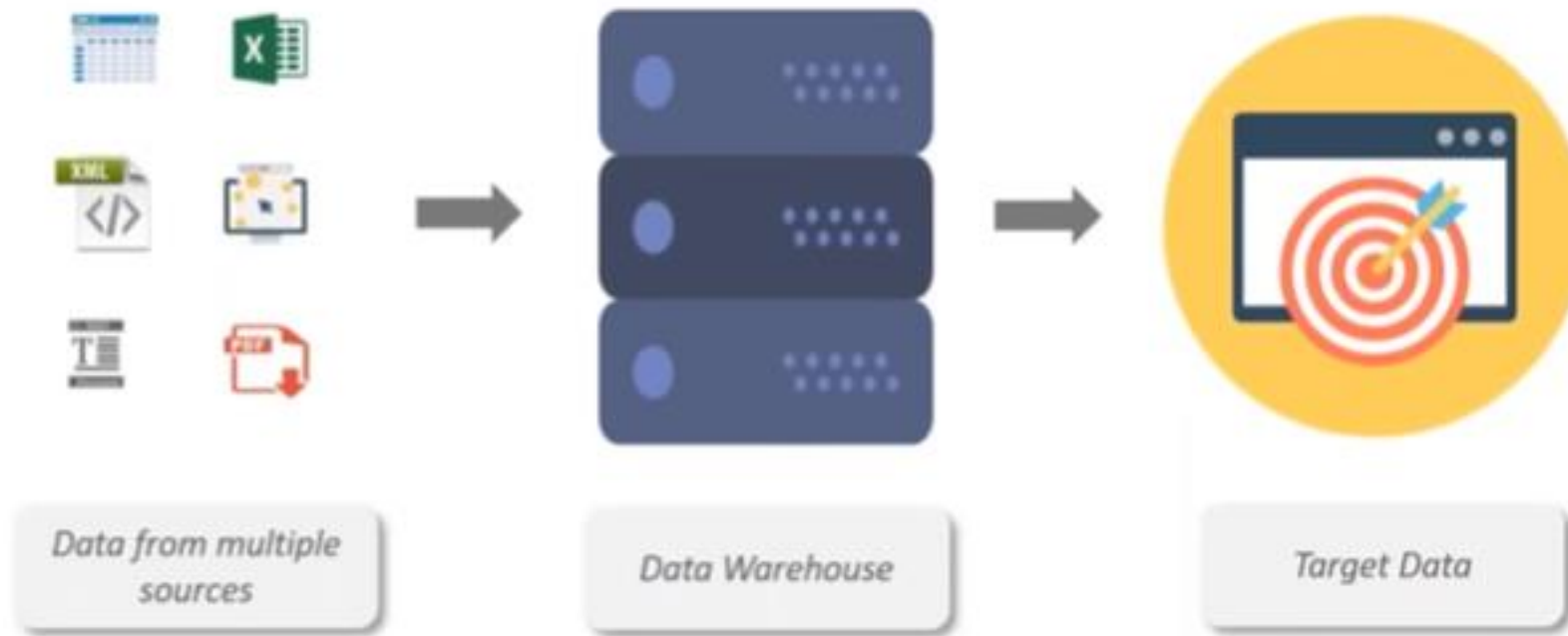


transformaciones



Seleccionar Datos

- Los datos están en múltiples fuentes y en múltiples formatos.
- Desde el DW seleccionamos el “dataset” para nuestro análisis.



Procesamiento de Datos

- Entendimiento de la estructura de los datos.
 - Correlación entre variables
 - Tipos de Variables
- Operaciones de sumarización, agregación, normalización pueden ser realizadas para transformar y consolidar los datos para la minería.



Data Mining

- Este es el paso mas importante del KDD
- Operaciones inteligentes tales como clustering, clasificación, regresión son aplicadas para extraer los patrones



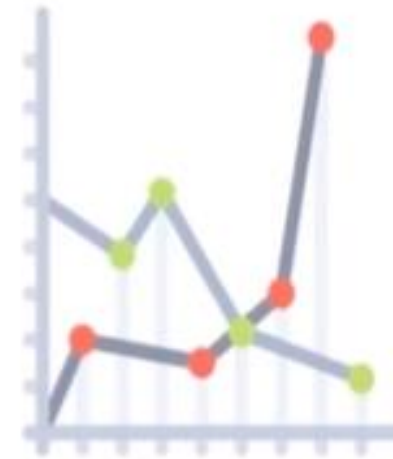
Evaluación de Patrones

- Una vez que las técnicas de Data Mining han sido aplicadas, los resultados obtenidos necesitan ser evaluados para validar su precisión.
- La Información obtenida es:
 - *Nueva*
 - *Correcta*
 - *Potencialmente Útil*



Representación del Conocimiento

- Los patrones identificados deben ser representados de forma simple usando por ejemplo gráficos.



Modelos de predicción: clasificación, regresión y series temporales

- Modelos predictivos
 - Clasificación
 - Regresión
 - Series temporales

Clasificación

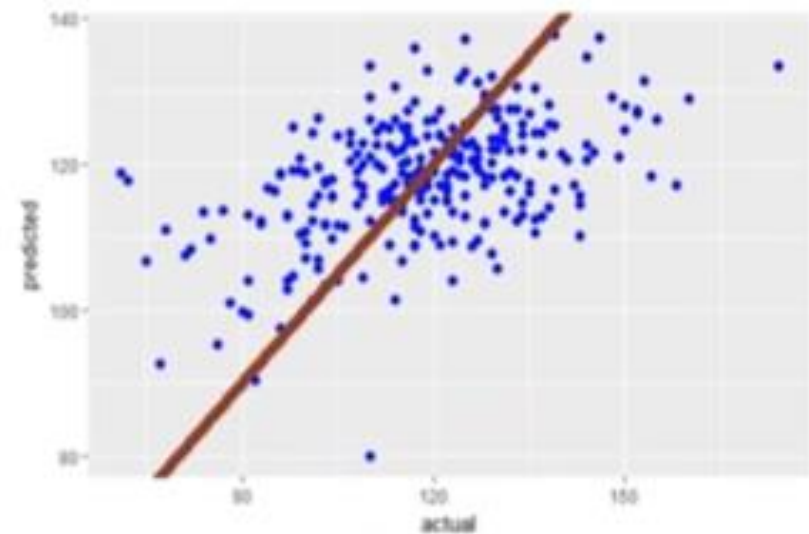
- La clasificación es el proceso de identificar a que categoría pertenece una observación dada.



Regresión

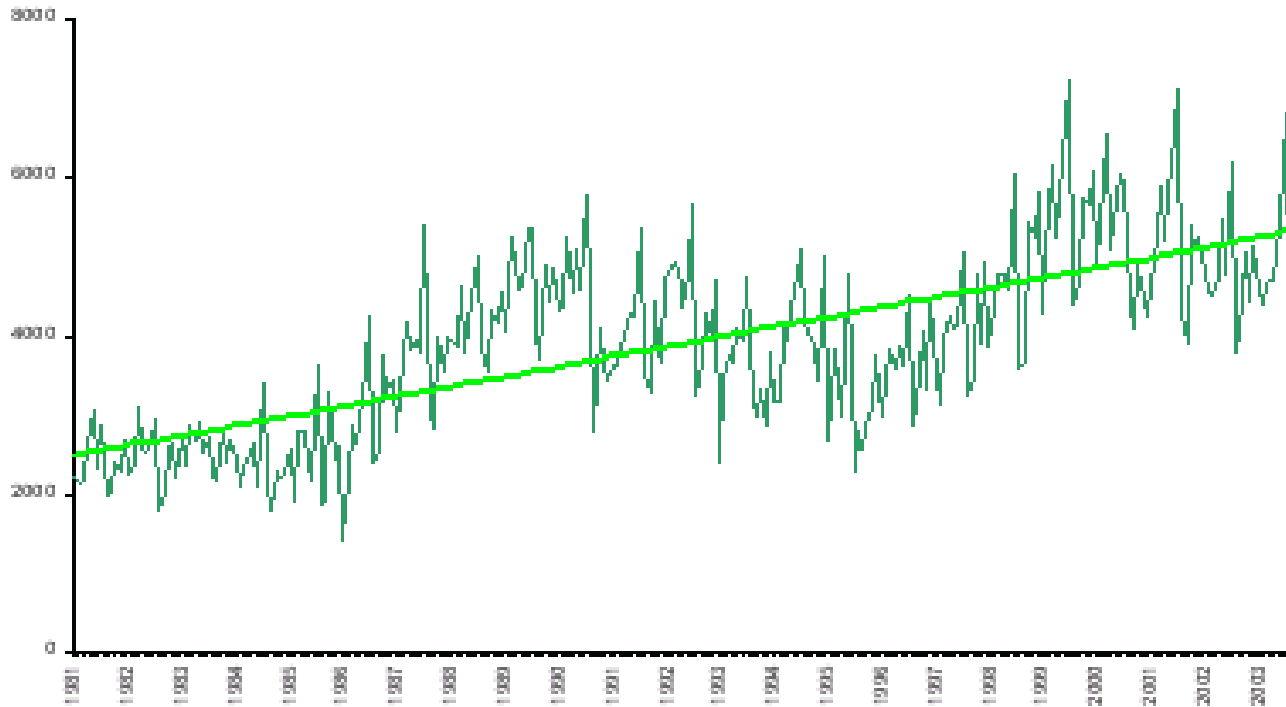
- Con la regresión podemos identificar la extensión de la relación entre variables
- Entender como varia la variable dependiente respecto a la variación de la variable independiente.

Regresión lineal



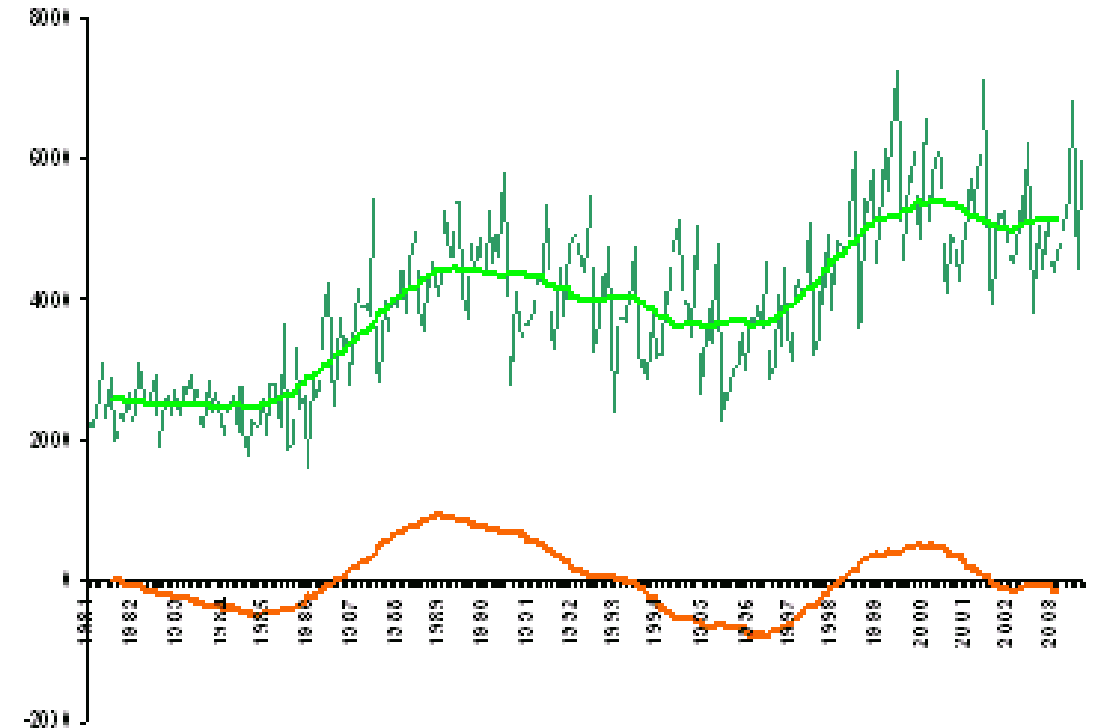
Series temporales

- Una Serie Temporal es una variable estadística cuyas observaciones están ordenadas temporalmente (años, meses, días, horas, minutos, etc.).



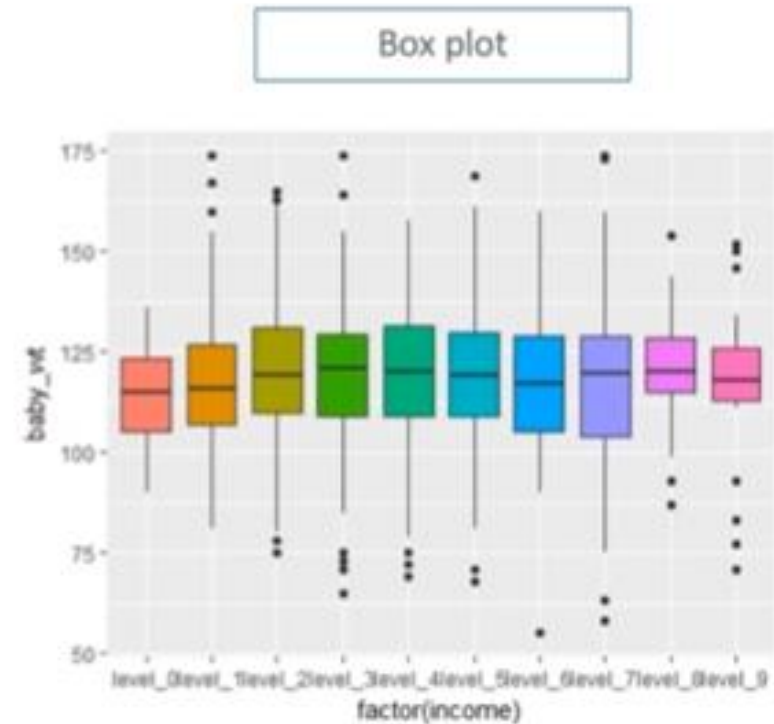
Componentes de una Serie Temporal

- Componente de Tendencia
- Componente Cíclica
- Componente Estacional
- Componente irregular o ruido



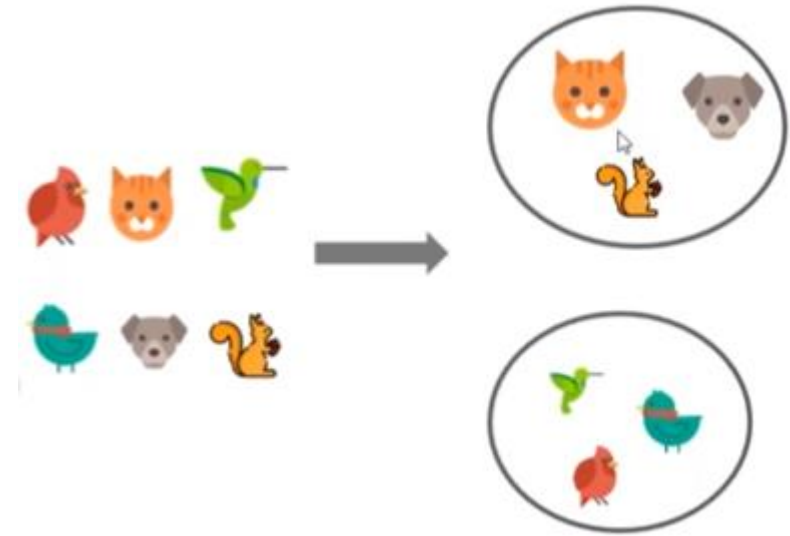
Preparación de Datos

- Limpieza de los datos
- Normalización de los datos
- Detección de anomalías
 - Patrones inusuales
 - Detección de outliers



Modelos de agrupamiento o segmentación – “clustering”

- Identifica grupos/clases en los datos los cuales son similares entre ellos
- La similaridad dentro del clúster es alta y entre los clústeres es baja



Modelos de Asociación

- Los modelos de asociación son utilizados para descubrir patrones interesantes de asociación entre las variables



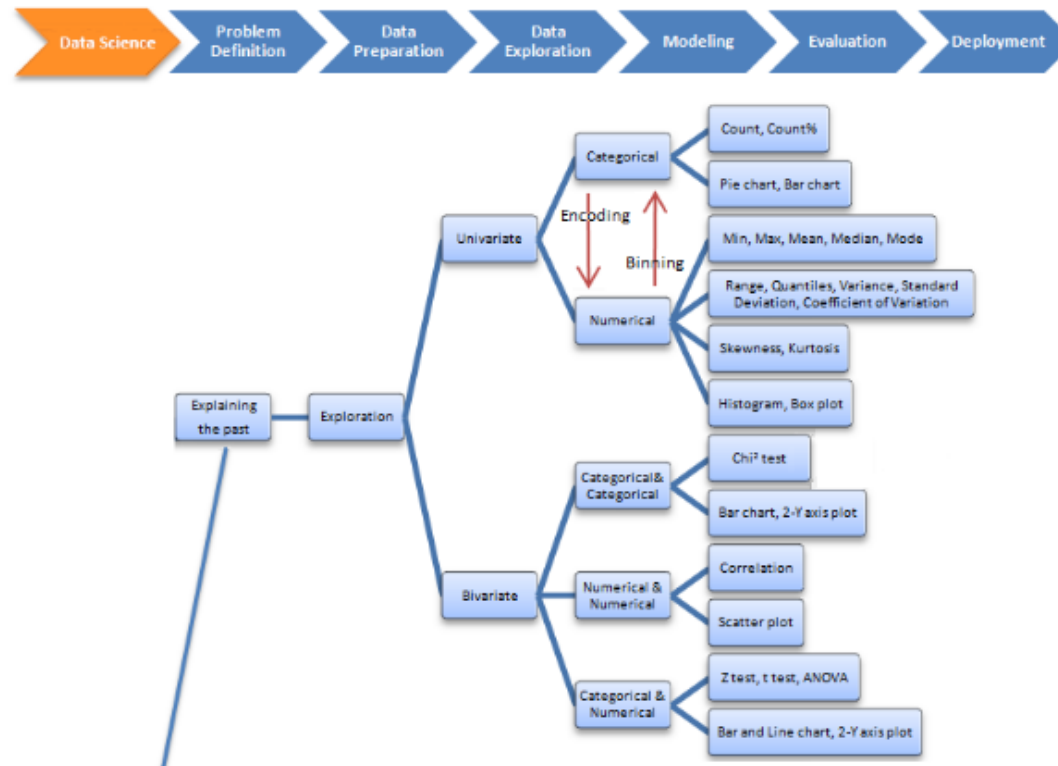
“En un supermercado se descubrió que los Jueves los hombres que compraban pañales también compraban cerveza”

Modelos Avanzados de Minería de Datos

- Técnicas avanzadas: <http://saedsayad.com/>

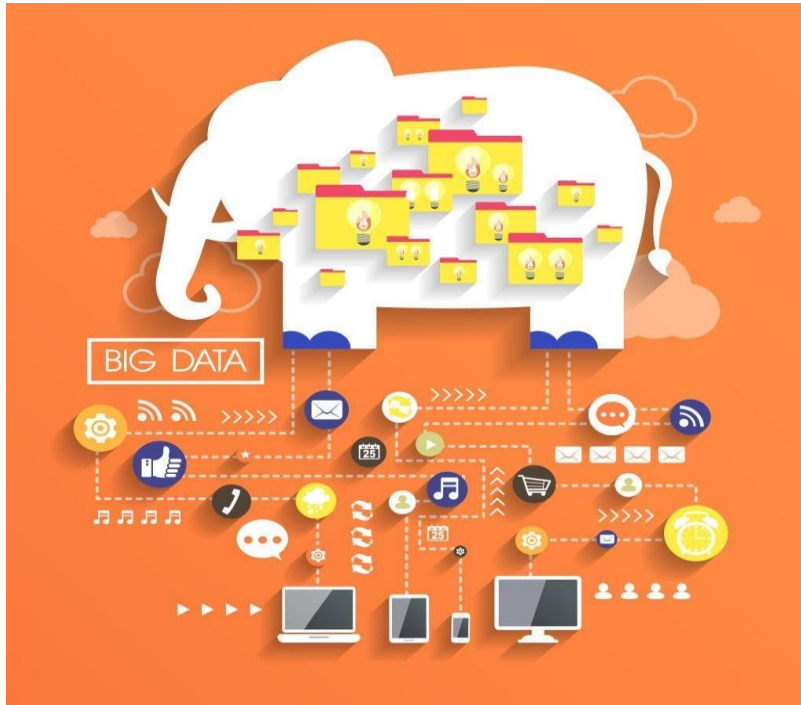
Copyright © 2010-2021, [Dr. Saed Sayad](http://saedsayad.com/)

An Introduction to Data Science




Big Data

- Data Mining vs Big Data



Lenguajes de Programación para Data Mining



Que lenguajes
necesito conocer
para hacer Data
Mining?

- R
- Python
- Julia
- SAS

Porque R?

- R es un lenguaje de programación usado por muchos modelos estadísticos y tareas de ciencia de los datos.
- Es un lenguaje de programación dinámicamente tipado
- Provee mas de 10,000 paquetes libres
- Es fácil de integrar con otros software como Ms Excel, SQL, Tableau



Obteniendo R

- <https://www.r-project.org/>



[\[Home\]](#)

Download

[CRAN](#)

R Project

[About R](#)

[Logo](#)

[Contributors](#)

[What's New?](#)

[Reporting Bugs](#)

[Conferences](#)

[Search](#)

[Get Involved: Mailing Lists](#)

[Developer Pages](#)

[R Blog](#)

R Foundation

[Foundation](#)

[Board](#)

[Members](#)

~

The R Project for Statistical Computing

Getting Started

R is a free software environment for statistical computing and graphics. It compiles and runs on a wide variety of UNIX platforms, Windows and MacOS. To [download R](#), please choose your preferred [CRAN mirror](#).

If you have questions about R like how to download and install the software, or what the license terms are, please read our [answers to frequently asked questions](#) before you send an email.

News

- [R version 4.0.5 \(Shake and Throw\)](#) has been released on 2021-03-31.
- Thanks to the organisers of useR! 2020 for a successful online conference. Recorded tutorials and talks from the conference are available on the [R Consortium YouTube channel](#).
- [R version 3.6.3 \(Holding the Windsock\)](#) was released on 2020-02-29.
- You can support the R Foundation with a renewable subscription as a [supporting member](#)

News via Twitter



The R Foundation Retweeted



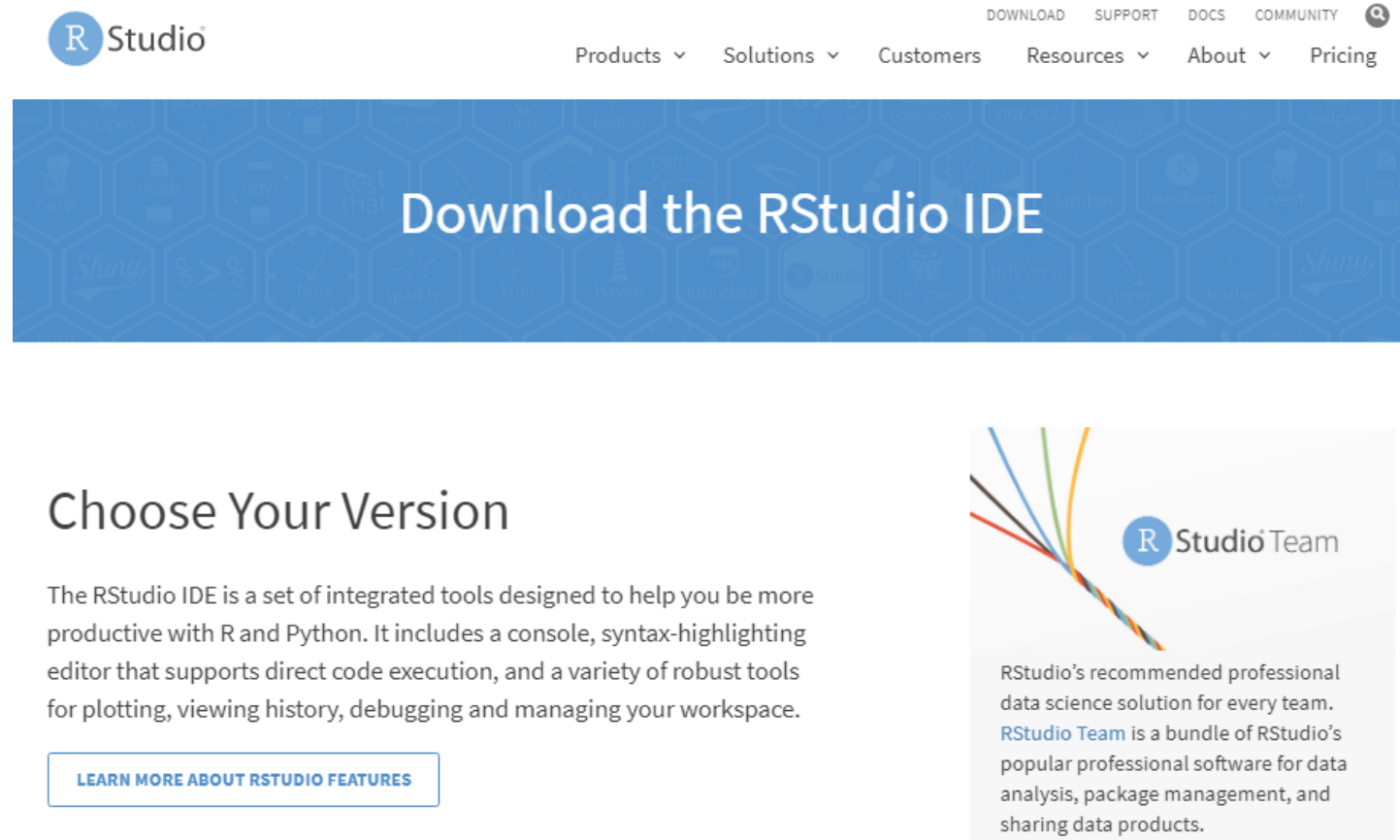
Peter Dalgaard

@pdalgd

[#rstats](#) 4.0.5 CRAN issues should be resolved now. In

Obteniendo R Studio IDE

- <https://www.rstudio.com/products/rstudio/download/>



The screenshot shows the RStudio website's download page. At the top, the RStudio logo is on the left, and navigation links for DOWNLOAD, SUPPORT, DOCS, and COMMUNITY are on the right. Below these, a secondary navigation bar includes Products, Solutions, Customers, Resources, About, and Pricing. A large blue banner with a hexagonal pattern contains the text "Download the RStudio IDE". Below the banner, the section "Choose Your Version" is visible, followed by a paragraph describing the RStudio IDE as a set of integrated tools for R and Python. A button labeled "LEARN MORE ABOUT RSTUDIO FEATURES" is present. To the right, there is a section for "RStudio Team" with a graphic of colorful lines and text describing RStudio's recommended professional data science solution.

RStudio

DOWNLOAD SUPPORT DOCS COMMUNITY

Products Solutions Customers Resources About Pricing

Download the RStudio IDE

Choose Your Version

The RStudio IDE is a set of integrated tools designed to help you be more productive with R and Python. It includes a console, syntax-highlighting editor that supports direct code execution, and a variety of robust tools for plotting, viewing history, debugging and managing your workspace.


[LEARN MORE ABOUT RSTUDIO FEATURES](#)

RStudio Team


RStudio's recommended professional data science solution for every team. RStudio Team is a bundle of RStudio's popular professional software for data analysis, package management, and sharing data products.

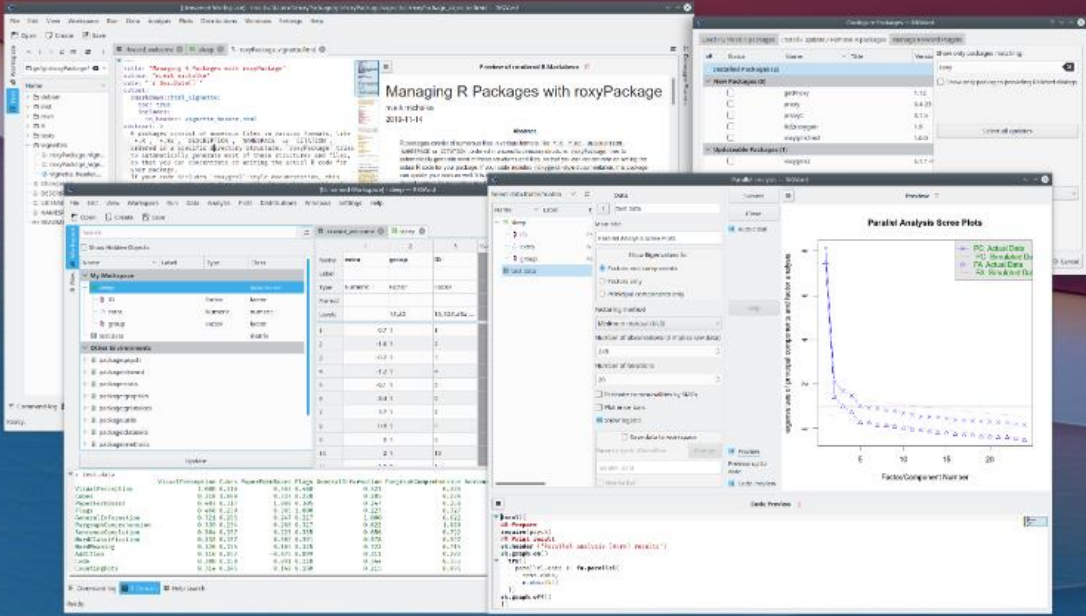
Obteniendo Rkward

- <https://rkward.kde.org/>

 **RKward**

[Learn](#) [Download](#) [Get support](#) [Get Involved](#) [Donate](#)

 Made by KDE



The screenshot displays the RKward interface with several windows open. The 'Managing R Packages with roxyPackage' window is prominent, showing a list of installed and available packages. Other windows include 'Parallel Analysis Scree Plots' showing a scree plot with eigenvalues and cumulative variance, and a 'Data' window showing a table of data. The interface is designed to be user-friendly for R package management and data analysis.

RKward

RKward is an easy to use and easily extensible IDE/GUI for R.

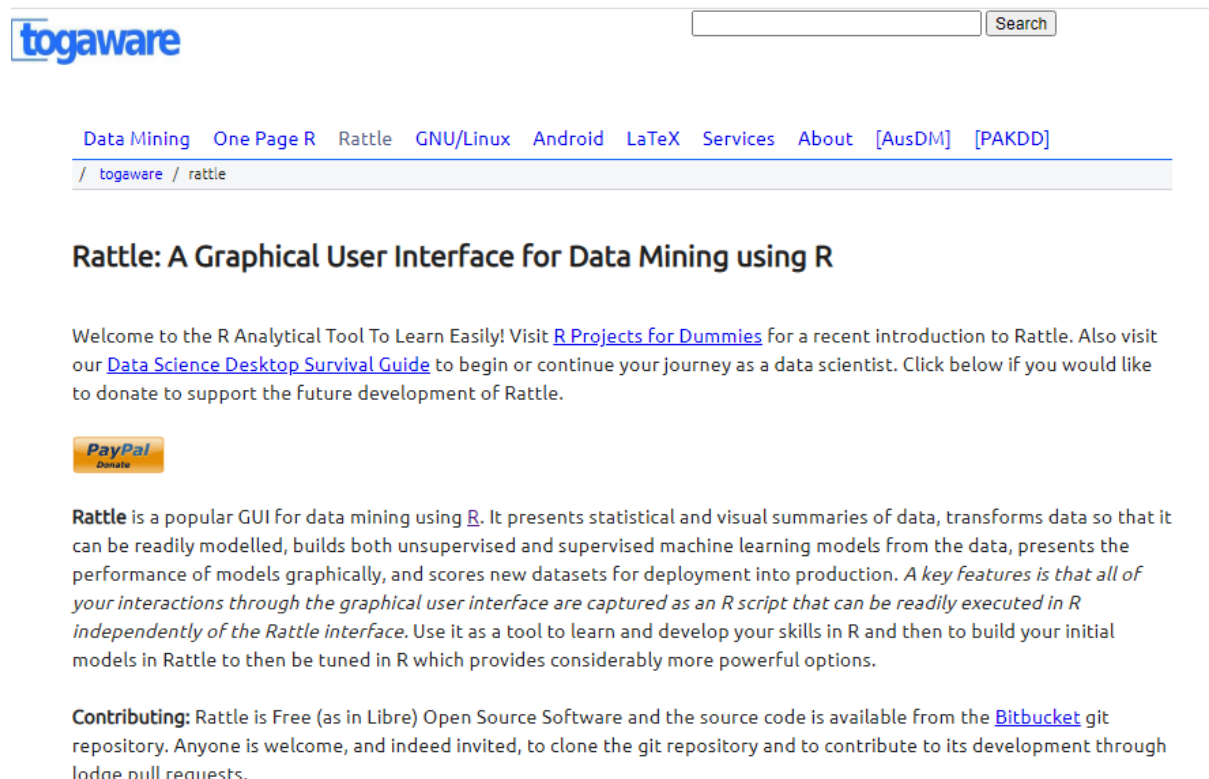
[Learn more](#)

Download RKward 0.7.2

for
[macOS](#) | [Windows](#) | [GNU/Linux](#)

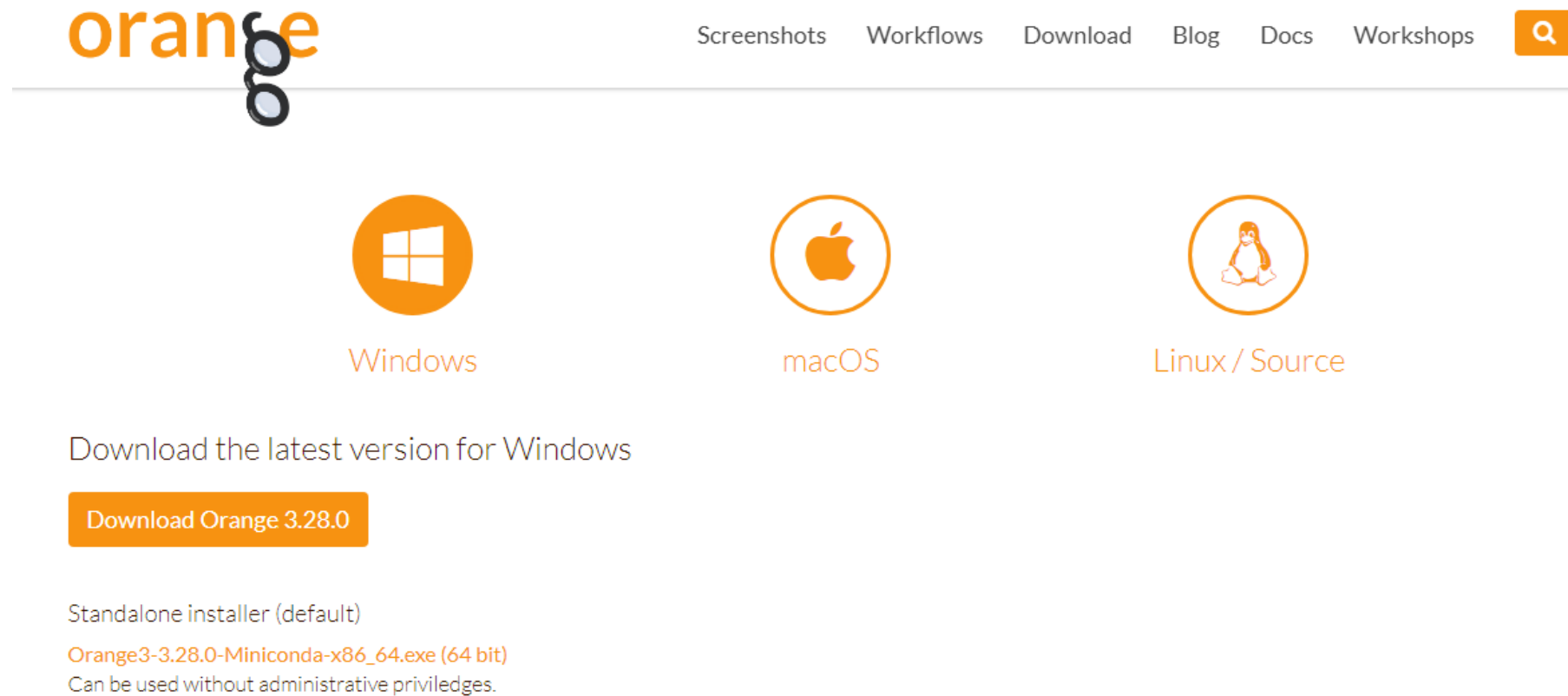
Rattle

- Es una Herramienta Analítica para R
- Se puede obtener Rattle desde: <https://rattle.togaware.com/>



Orange

- <https://orangedatamining.com/download/#windows>



Preguntas

- Alguna pregunta?



Demo

- Instalación de R y R Studio IDE

RStudio Desktop 1.4.1106 - [Release Notes](#)

1. Install R. RStudio requires R 3.0.1+.
2. Download RStudio Desktop. Recommended for your system:



Requires Windows 10 (64-bit)

