Laboratorio de Datos
Verano 2023

Andrés Farall afarall@gmail.com
Jemina García jeminagarcia@gmail.com

Andrés Farall

Hola !!!!

Jemina García

Hola !!!!.

Objetivos de la Materia

Brindar una introducción al **Análisis Exploratorio** de Datos (EDA) y al **Modelado** de Datos, utilizando elementos básicos de matemáticas y de programación, sin el uso de nociones de Probabilidad y Estadística.

Generar una serie de **preguntas** que pueden hacerse sobre un conjunto de datos, que finalmente serán respondidas mediante modelos estadísticos o algoritmos de *machine learning*.

Introducir algunos **conceptos fundamentales** de la Ciencia de Datos, como ser: Descripción-Predicción-Explicación, significatividad estadística, sobreajuste, bondad de ajuste, funciones de pérdida, asociación entre variables, análisis supervisado vs. no supervisado, modelos paramétricos vs. no paramétricos, etc.

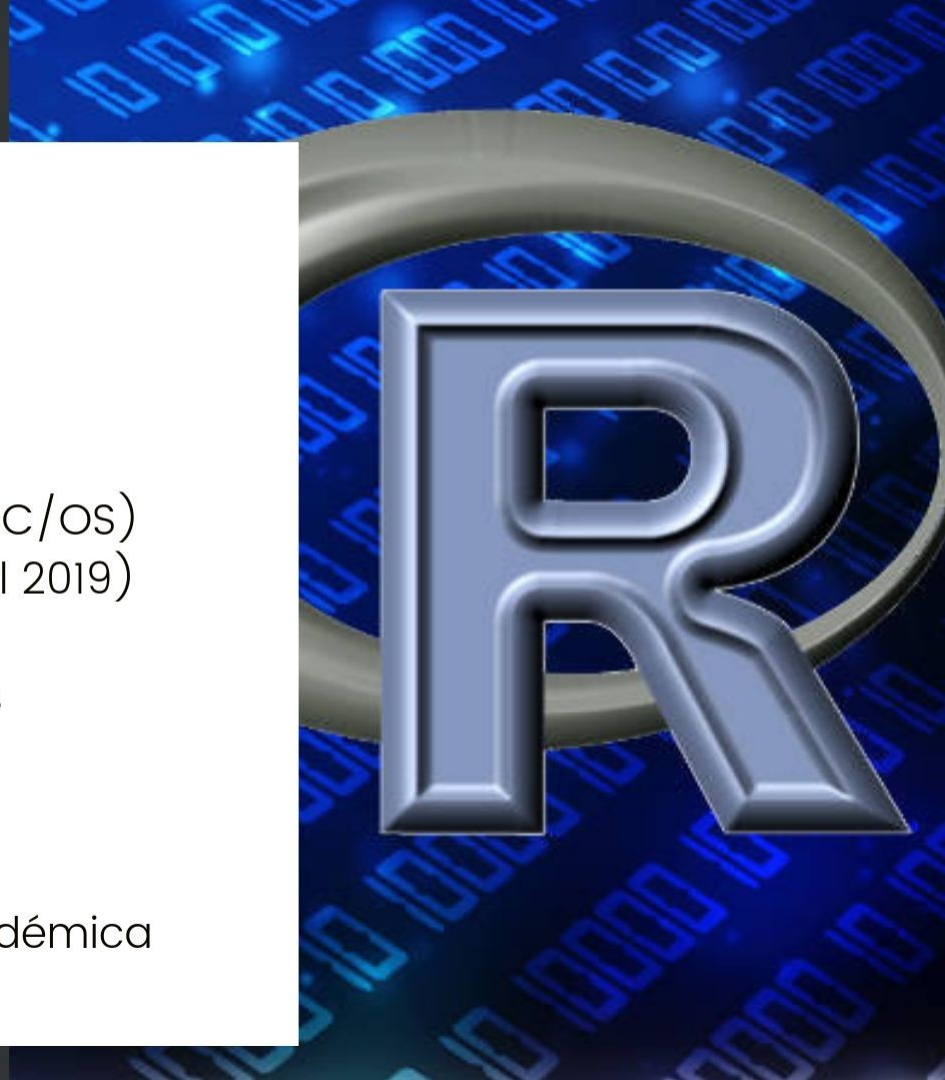
Programa

1. Obtención y organización de datos. Datos estructurados y no estructurados.
2. Visualización de datos como herramienta exploratoria antes del desarrollo de modelos y aprendizaje estadísticos. Análisis exploratorio de datos.
3. Introducción al modelado. Regresión Lineal Múltiple y Vecinos más Cercanos. Modelos predictivos vs modelos explicativos. Distinción entre modelos univariados y multivariados, y modelos paramétricos y no-paramétricos.
4. Herramientas de validación de un modelo. Muestras de testeo y entrenamiento. Métricas y métodos para la evaluación de algoritmos y modelos estadísticos.
5. Análisis Supervisado: Regresión y Clasificación
6. Análisis No Supervisado: Clustering y Reducción de Dimensión

La Herramienta más difundida

#11 conceptos

- Código Abierto (GNU-GPL V 3)
- Gratuito (GNU-GPL V 3)
- Multiplataforma (Windows, Linux, MAC/OS)
- Comunitario (>2.000.000 usuarios al 2019)
- Orientado a objetos
- Especializado en el análisis de datos
- Potentes gráficos
- Flexible (interprete)
- Alto nivel de expresión
- Fuerte aceptación/intervención académica
- Facil integración vertical

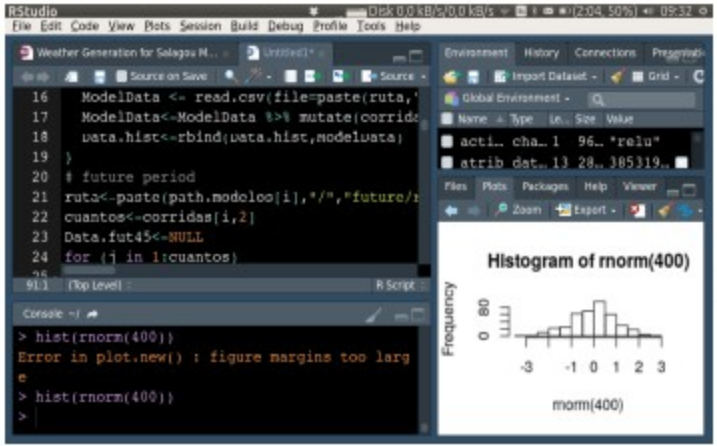


Cómo se trabaja en R

Datos

Análisis

Preguntas



Resultado

Informes

APIs

Prototipos

Datos

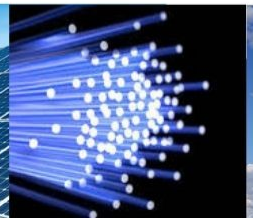
Métodos

Algoritmos

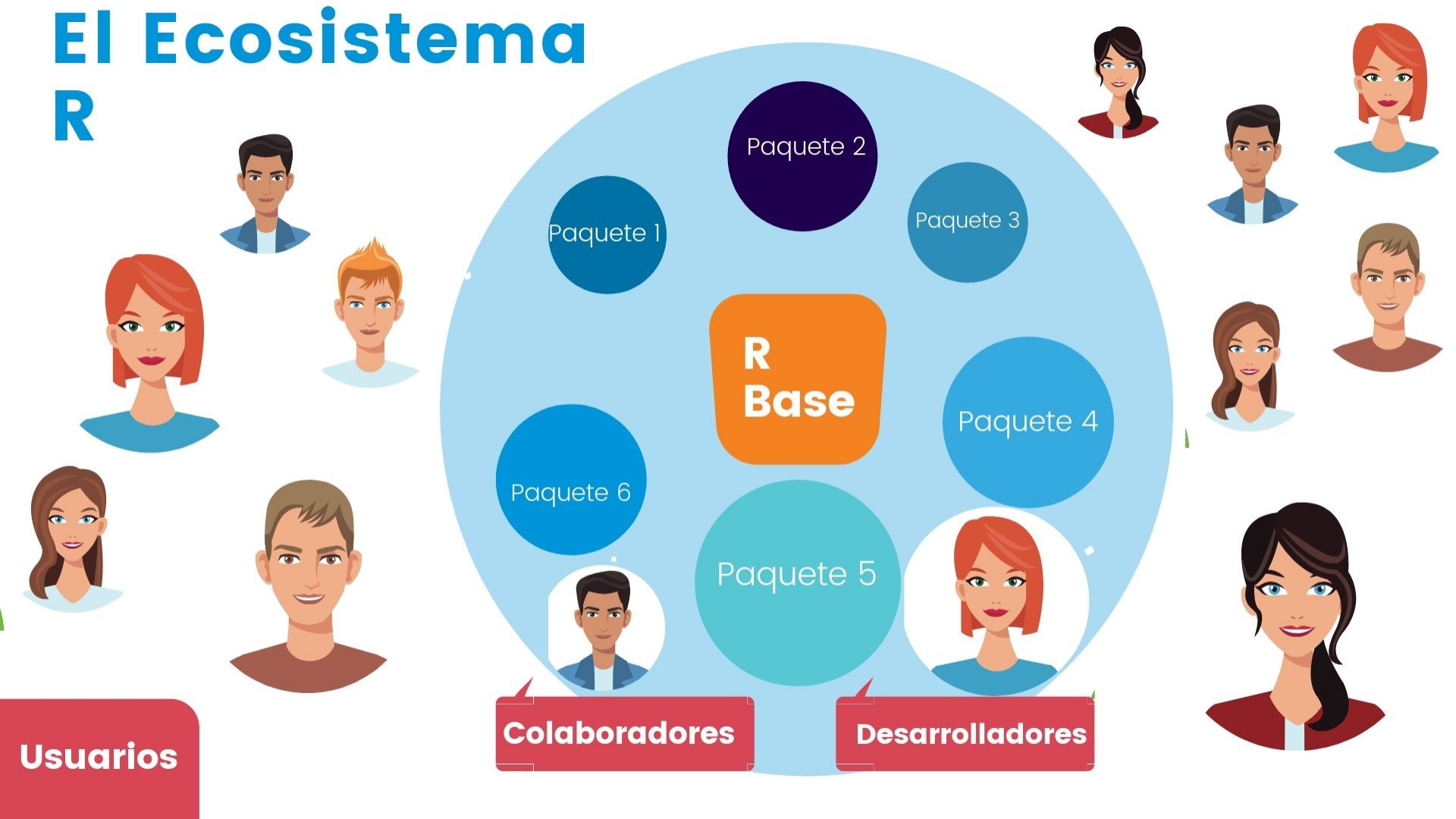
UNIX

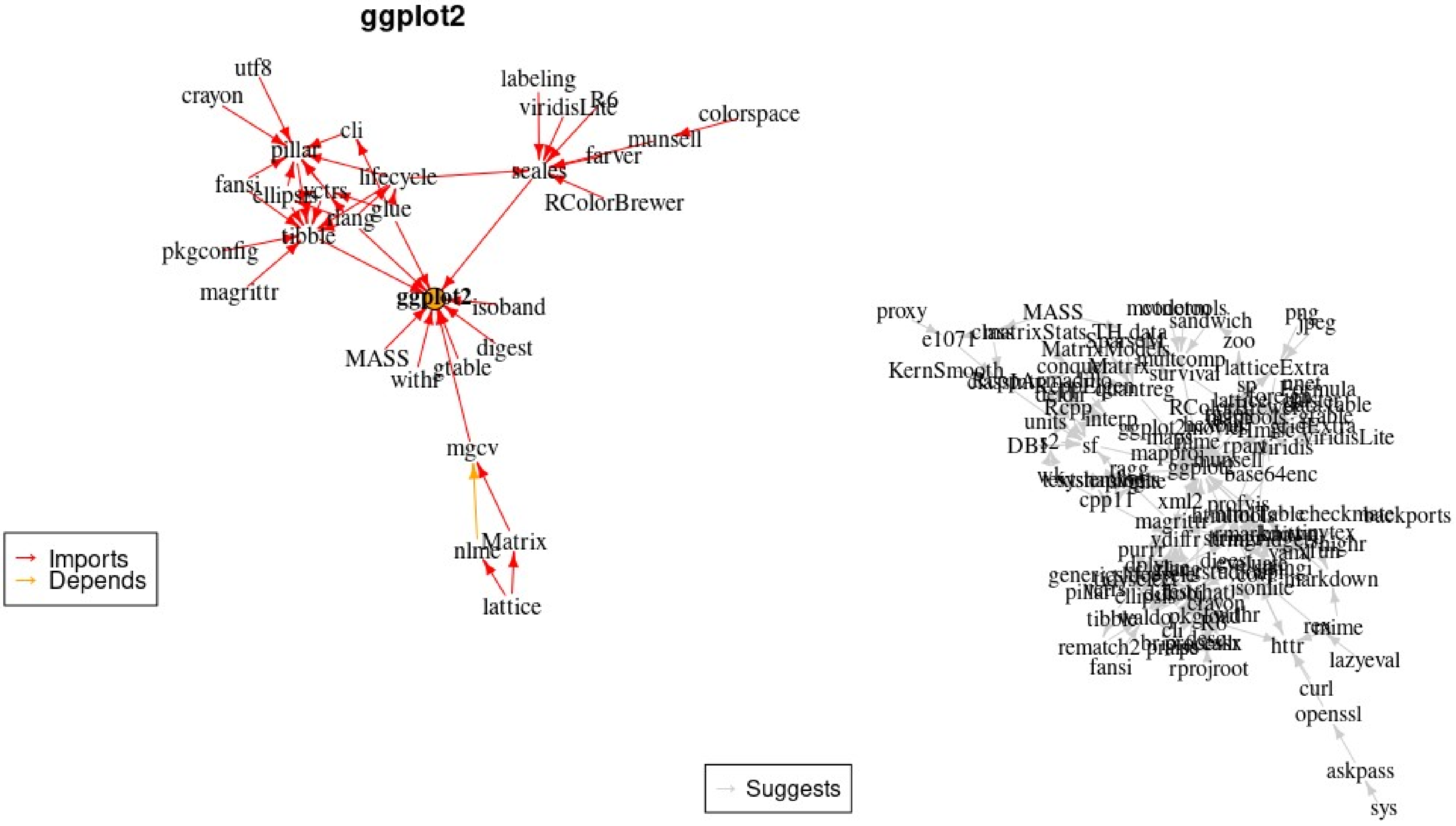
El Origen de R...

1976: John Chambers crea el Lenguaje S en Bell Labs.
1984: AT & T (Bell Labs) vende la Licencia de S.
1991: Ross Ihaka y Robert Gentleman lanzan el Proyecto R basado en el lenguaje S.
2001: La versión 1.0.0 de R es lanzada.
2019: R cuenta con,
-> 2.000.000 de usuarios
-> 10.000 colaboradores
-> 100 grandes empresas lo utilizan

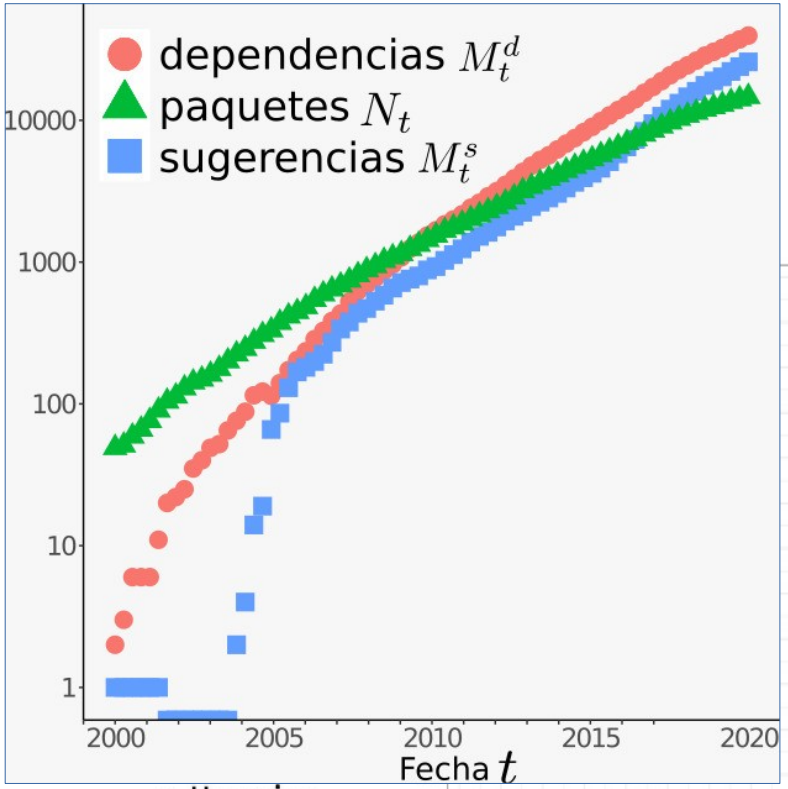


El Ecosistema R

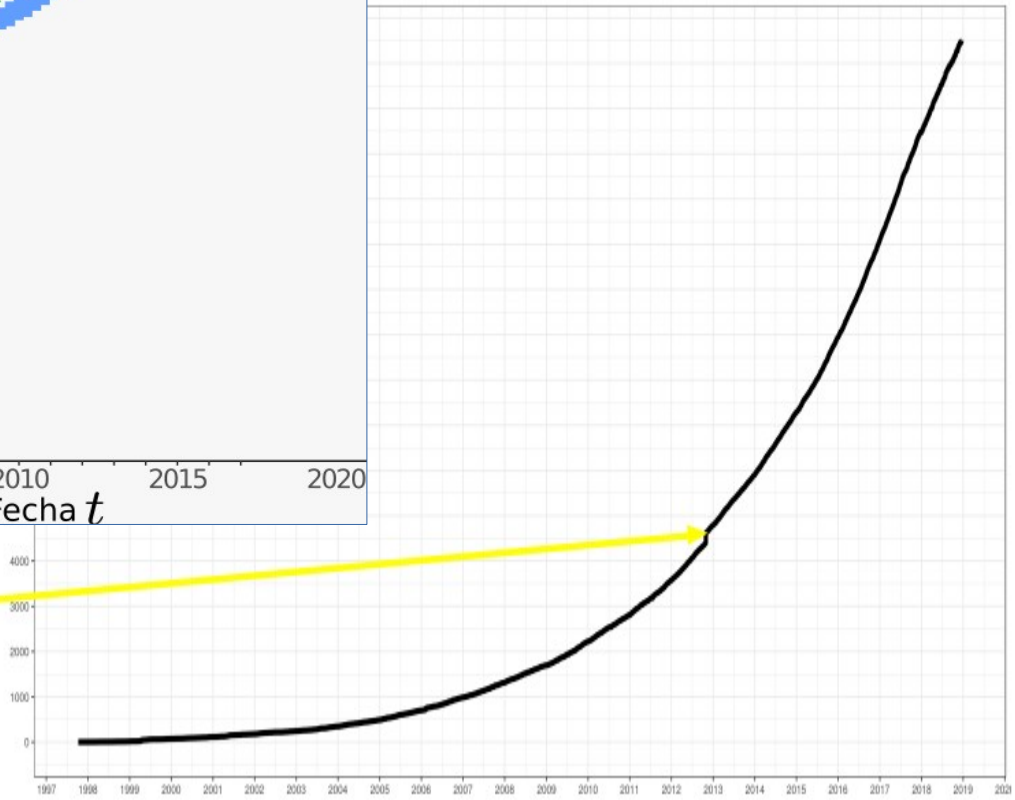



ggplot2
utf8
crayon
labeling
R6
viridisLite
colorspace
munsell
cli
pillar
farver
scales
lifecycle
fansi
ellipsis
vctrs
glue
rlang
tibble
RColorBrewer
pkgconfig
magrittr
ggplot2
isoband
digest
MASS
gtable
withr
mgcv
Matrix
nlme
lattice
Imports
Depends
Suggests
proxy
e1071
KernSmooth
MASS
sandwich
zoo
png
jpeg
latticeExtra
multcomp
survival
Rcpp
units
interp
DBI
sf
mapproj
base64enc
cpp11
xml2
profvis
checkmate
backports
purrr
markdown
mime
httr
lazyeval
rematch2
fansi
rprojroot
curl
openssl
askpass
sys

Crecimiento Exponencial de R



- Usuarios
- Paquetes
- Colaboradores
- Conexiones



¿Qué Explica el Surgimiento del Paradigma FOS?

- Compartir un desarrollo digital tiene un costo de oportunidad negativo. Uva vez resuelto un problema propio conviene compartirlo !
- Subsidios cruzados: profesionales y científicos de países desarrollados generan y mantienen los proyectos financiados por organizaciones altamente lucrativas
- La existencia de un entorno tecnológico global interconectado (Internet)
- Bajo costo de generación de proyectos, pero alto impacto

¿Qué Explica el Crecimiento del Paradigma FOS?

- Una vez creada la herramienta FOS, las ventajas comparativas son INMENSAS

- Precios inferiores a la competencia (gratis)
- Mayor adaptabilidad a las necesidades de la demanda
- Facilidad de difusión, ya que el costo de adquisición es 0
- Un Desarrollador debiera preferir SIEMPRE una herramienta FOS

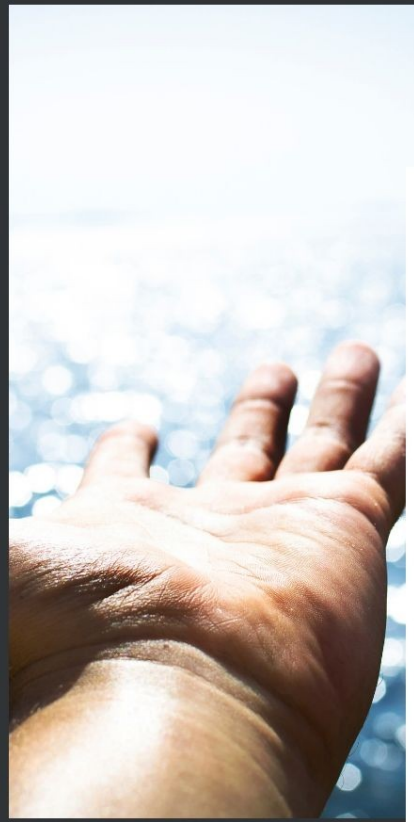
Las Claves del Exito

Libre y Gratuito FREE

El precio es 0 (cero)!
En un mercado de competencia perfecta, el precio de los productos debe tender a su costo marginal. Los bienes digitales tiene un costo marginal de ..... 0 (cero)!

Cualquier herramienta paga debiera ser desplazada por una gratuita, si las prestaciones son similares.

Fácil difusión de la herramienta, no requiere:
-Costosos acuerdos corporativos
-Distribución ilegal (piratería)
-Complejos esquemas de promoción gratuita a universidades, escuelas, fundaciones, ministerios, etc.



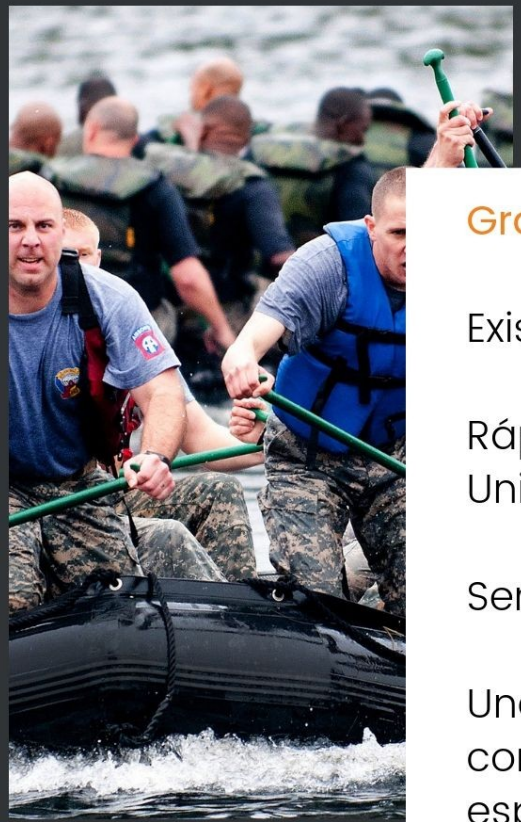
Las Claves del Exito
Código Abierto OPEN SOURCE

Transparencia.
Cualquiera puede saber que está haciendo la herramienta, en cualquier situación.

Expandible y Mejorable.
Se la puede adaptar a requerimientos específicos.

Se la puede integrar en desarrollos comerciales, reduciendo los costos de operación y aumentando el beneficio.

Facilita la integración en otros sistemas informáticos.



Las Claves del Exito
Comunitario

Gran red de contribuidores al proyecto.

Existencia de foros de ayuda y discusión.

Rápidamente asimilable en ámbitos públicos (Gobiernos y Universidades.

Sentimiento de pertenencia de los colaboradores.

Una comunidad diversa y extendida asegura la contribución de herramientas útiles en nichos pequeños y específicos que NO pueden ser atendidos por software propietario.

Algunos Ejemplos

Y, Entonces...
¿Donde esta el negocio?

Modelo Freemium (No es el caso de R ni de Linux)

Potencial de Monetización:

- Servicios de Implementación e Integración
- Soporte a Empresas
- Reducción del Riesgo
- Capacitación y Difusión
- Desarrollos a medida
- Creación de Fundaciones sin fines de lucro

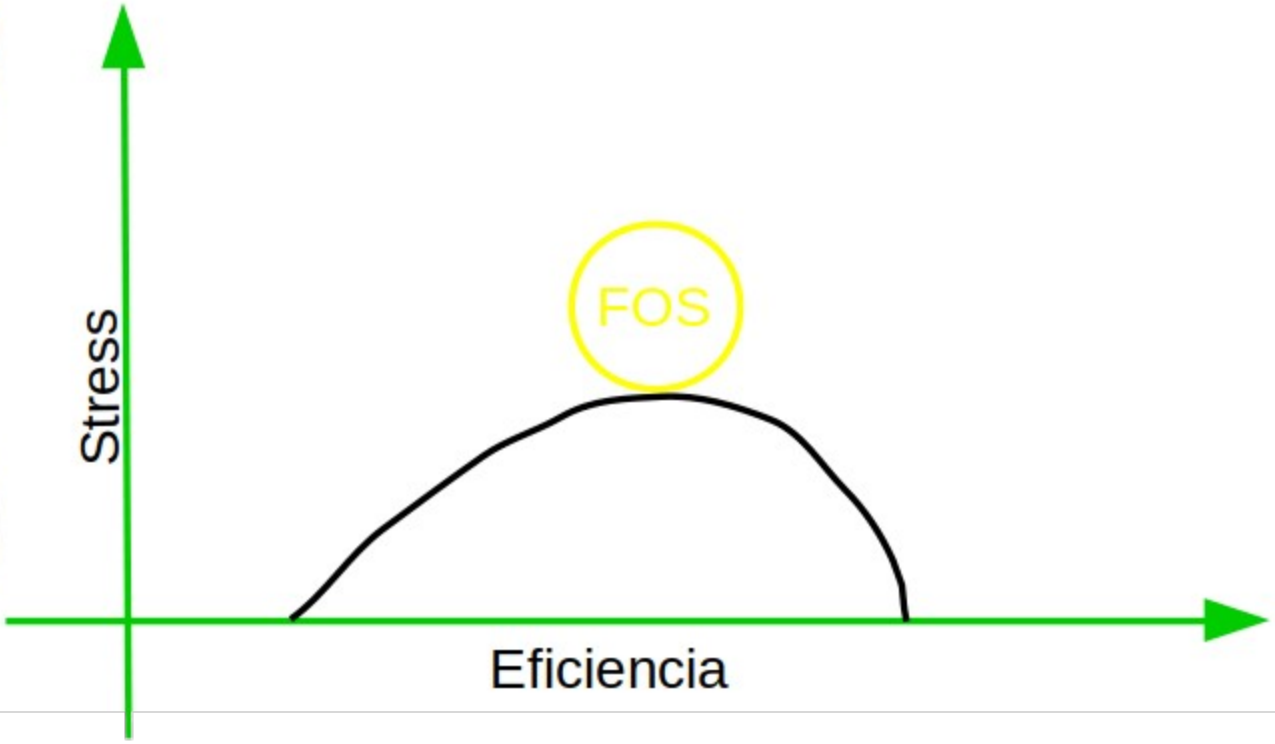
¿Es suficiente?

Apoyo de corporaciones que utilizan y aprovechan la herramienta.

¿Que podemos esperar en el futuro?

Expertos trabajando grátis
Empresas perdiendo mercados

Expertos remunerados
Empresas conservando mercados



¿A dónde podemos llegar con esta materia?

1

Análisis de las Películas de los Últimos 50 Años

Internet Movie Database

IMDb

Información general	
Dominio	IMDb (en inglés)
Tipo	Base de datos cinematográfica Compilador de reseñas Base de datos de videojuegos Catalogación social Television series database Directorio de podcasts
Comercial	Sí
Registro	El registro es opcional para miembros para participar en discusiones, comentarios, calificaciones y votaciones, incluyendo acceso a listados de películas, catálogos y horarios[1]
Idiomas disponibles	Inglés
En español	No
Estado actual	Activo
Gestión	
Desarrollador	Col Needham
Propietario	IMDb.com, Inc.
Lanzamiento	17 de octubre de 1990
Estadísticas	
Ranking Alexa	▲ 63° (8 de marzo de 2021)[3]