INTRODUCCIÓN AL LENGUAJE ESTADÍSTICO Y PROGRAMACIÓN

RSTUDIO

¿QUÉ ES RSTUDIO?

- Es un entorno y lenguaje programación con un enfoque al análisis estadístico. R nació como una reimplementación de software libre, de lenguaje estadístico.

- Dialecto de Software S.

- Desde los años 70's el análisis estadístico se realizaba por medio de subrutinas de Fortran (un lenguaje de programación alto nivel desarrollado por IBM en 1957, especialmente adaptado al cálculo numérico y a la computación científica), esto era muy tedioso y tardado en realizar el análisis.

```
PROGRAM TRIVIAL
    INTEGER I
    I=2
    IF(I .GE. 2) CALL PRINTIT
    STOP
END
SUBROUTINE PRINTIT
    PRINT *,'Hola Mundo'
    RETURN
END
```

¿QUÉ ES RSTUDIO?

- Por esta razón 1976, un equipo dirigido por John Chambers, Rick Becker y Allan Wilks, pertenecientes a los laboratorios Bells, desarrollaron S que implementaba librerías de macros Fortran. Lo llamaron S por Statistical, porque en esa época era común nombrar a los lenguajes de programación con una sola letra (por ejemplo C).

- En 1988, se reescribió completamente S, a la versión 3, S3 en código C, sustituyendo las **MACROS** por funciones y modificando la sintaxis para hacerla más consistente, también se añadieron funciones de modelado estadístico, ausentes anteriormente. John Chambers público el libro "Statistical models in S" para documentar este proceso, por su importancia el libro se conoce como el libro blanco.

¿QUÉ ES RSTUDIO?

StatSci

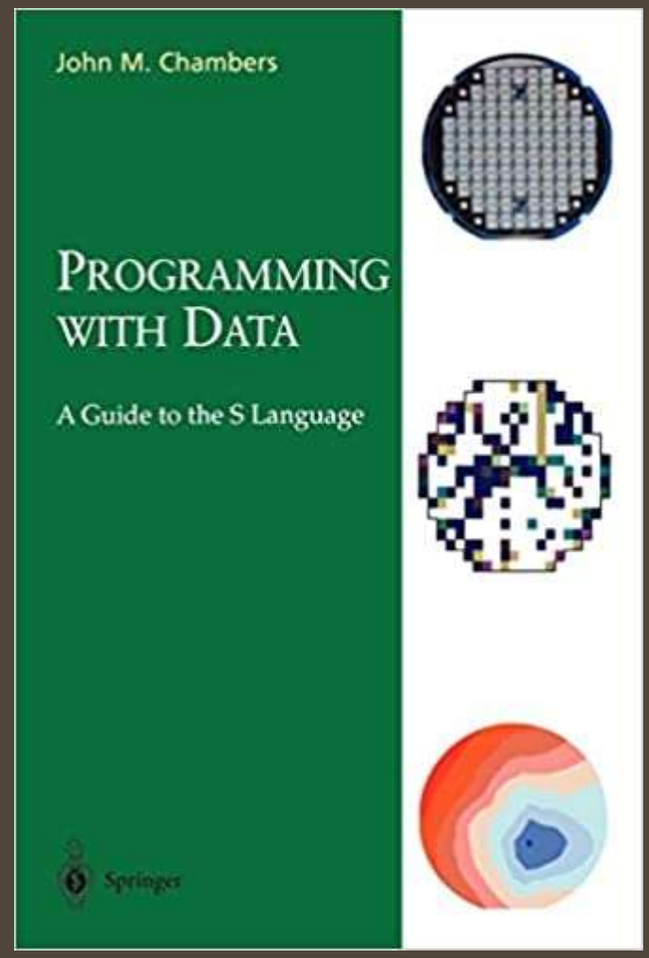
mathsoft

Data Analysis Products Division

S-PLUS

- En 1993, los laboratorios Bells venden S a la empresa StatSci la licencia para explotarla comercialmente, Statci se fusiona con MathSoft y pasan a denominarse Data Analysis Products Division y sacan la versión comercial S-Plus, con la mejora de ser una interfaz gráfica.

- En 1998, se libera la cuarta versión de S, S4, con características orientadas a objetos mucho más avanzadas. Chambers documenta esta versión en el libro "Programming with data", llamado libro verde. En este años S gana el premio "Association for Computing Machinery's Software System Award".



¿QUÉ ES RSTUDIO?

- En el 2001 Data Analysis Products Division cambia de nombre a Insightful Corporation, en el 2004 adquiere la totalidad del lenguaje S y en el 2008 TIBCO adquiere Insightful Corporation por 25 millones de dólares. Tras todos estos cambios de compañía, los fundamentos de S no a cambiando desde la versión S4 de 1998.

¿QUÉ ES RSTUDIO?

- Mientras S cambiaba de dueño, en 1991 en el Departamento de Estadística de la Universidad de Auckland en Nueva Zelanda, Ross Ihaka y Robert Gentleman crean R, creado con la intención de ser un lenguaje didáctico para ser utilizado en el curso de "Introducción a la Estadística" de dicha universidad, como un subdialecto de S e implementado su propio dialecto. El nombre de R es acuñado en forma de broma, debido a la primera letra de sus creadores Ross y Robert. Anunciándolo en 1993 al público y en 1995 Martin Machler convence a Ross y Robert de usar la Licencia Publica General GNU haciendo a R como software libre. Esto supone una revolución paradigmática.

¿QUÉ ES RSTUDIO?

Gracias por TANTO

Martin Machler.

GPL v3

Free Software

Free as in Freedom

¿QUÉ ES RSTUDIO?

Aquí inicia el gran camino de R

- En 1996 sale la versión 0.16, es la última versión alfa desarrollada por Ross y Robert, que incluye gran parte de las características del libro blanco.
- En 1997 sale versión 0.49, la versión más antigua a la que se conserva el código, que todavía compila en algunas plataformas de UNIX. También arranco CRAN (red integral de archivos en R) que albergaba 12 paquetes, y poco después aparecen las versiones para Windows y Mac. En este mismo año sale la versión 0.60 que se incluye oficialmente en el proyecto GNU.
- En 2000 sale la versión 1.0.0, Se considera suficientemente estable para su uso en producción.
- En 2001 la versión 1.4.0, introduce los métodos S4 y aparece la primera versión para Mac OS X.
- En 2004 versión 2.0.0, Implementa el método lazy loanding, permite una carga rápida de datos con coste de memoria mínima
- En 2013 versión 3.0.0, se incluyen mejoras en la interfaz de usuario, las funciones de gráficos, en la gestión y rendimiento de la memoria.
- En 2015 versión 3.2.1, se incluyen mejoras en el rendimiento y fiabilidad, avances en manejo de datos grandes en la memoria y en el sistema de paquetes

¿QUÉ ES RSTUDIO?

Como vimos R es reciente, no tiene más de 28 años desde su creación y es debido a la licencia GNU, que hoy en día sea el lenguaje más utilizado en investigación por la comunidad estadística, siendo además muy popular en el campo de la investigación biomédica, la bioinformática y la economía.

Fuente: https://www.r-project.org/

VENTAJAS RSTUDIO

- Lenguaje Sencillo.
- Herramientas Optimizadas para el Análisis de Grandes Volúmenes de Datos.
- Visualización de Información de Manera Sencilla.
- Generador de Informes.
- Motor apto para la Automatización.
- Diseñado para ser interactivo.
- Usuario/Desarrollador.
- Simulaciones
- Modelación

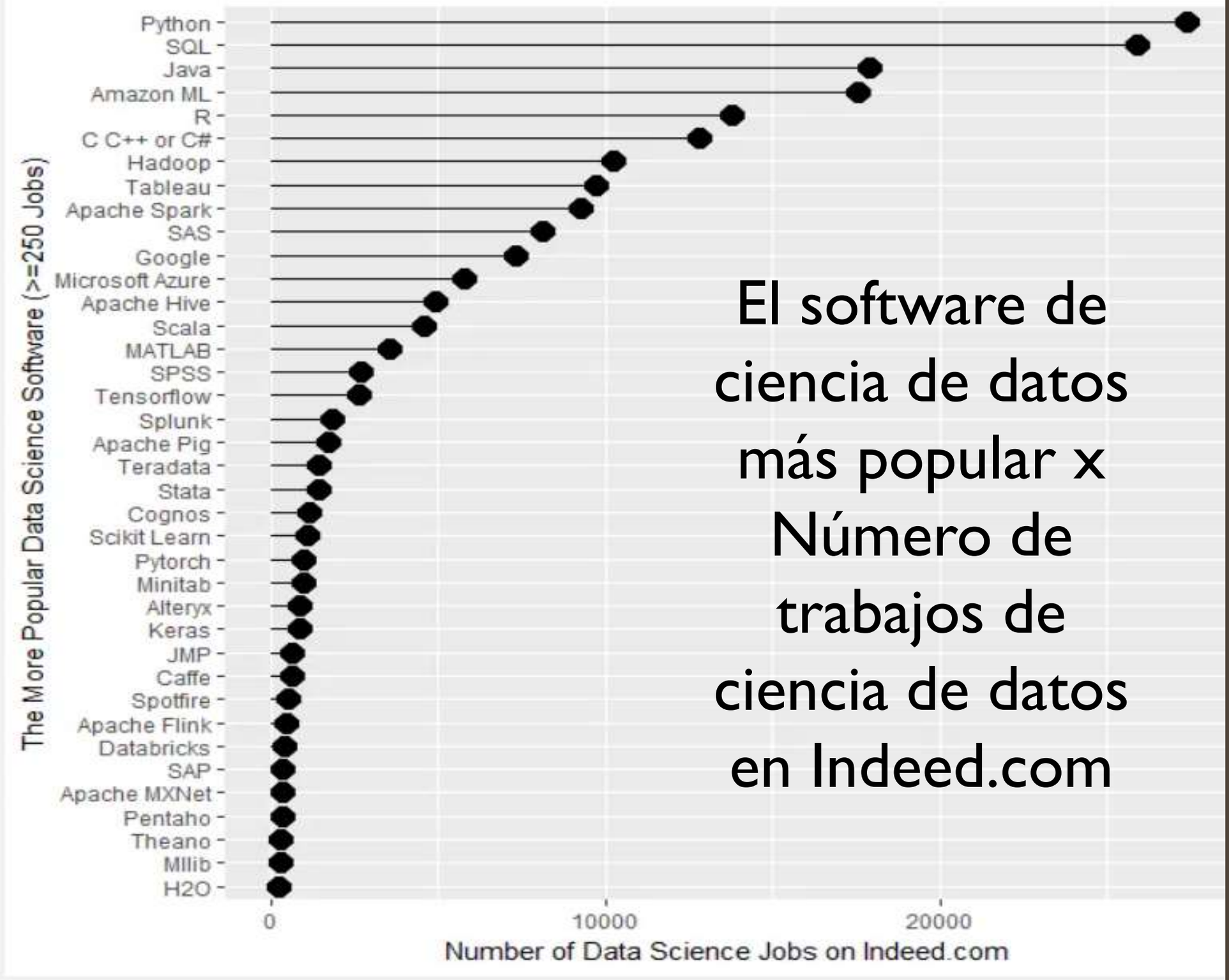
VENTAJAS RSTUDIO

- Predicciones
- Aprendizaje de máquinas
- Minería de texto
- Test de hipótesis
- Shiny (Librería para aplicaciones web)
- RStudio
- Notebooks
- ...

VENTAJAS RSTUDIO

- Ejecución línea a línea
- Notebooks (como Jupyter)
- Integración con Python. No viceversa.
- Gratis
- Gran soporte comunitario
- Ideal para trabajo interactivo

COMPARACION



El software de ciencia de datos más popular x Número de trabajos de ciencia de datos en Indeed.com

Figure 1a. Number of data science jobs for the more popular software.

Fuente: https://www.r4stats.com

COMPARACION

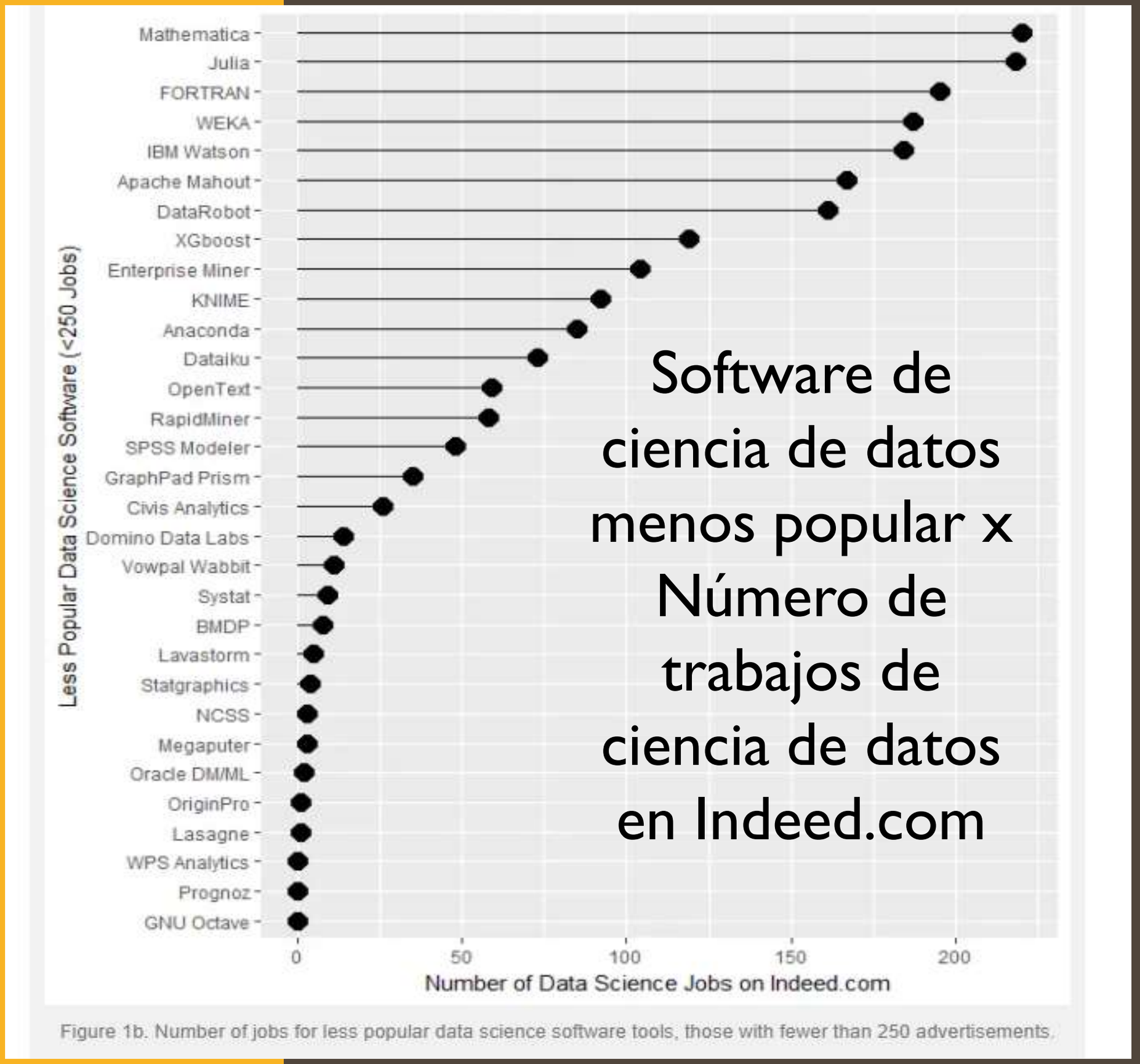
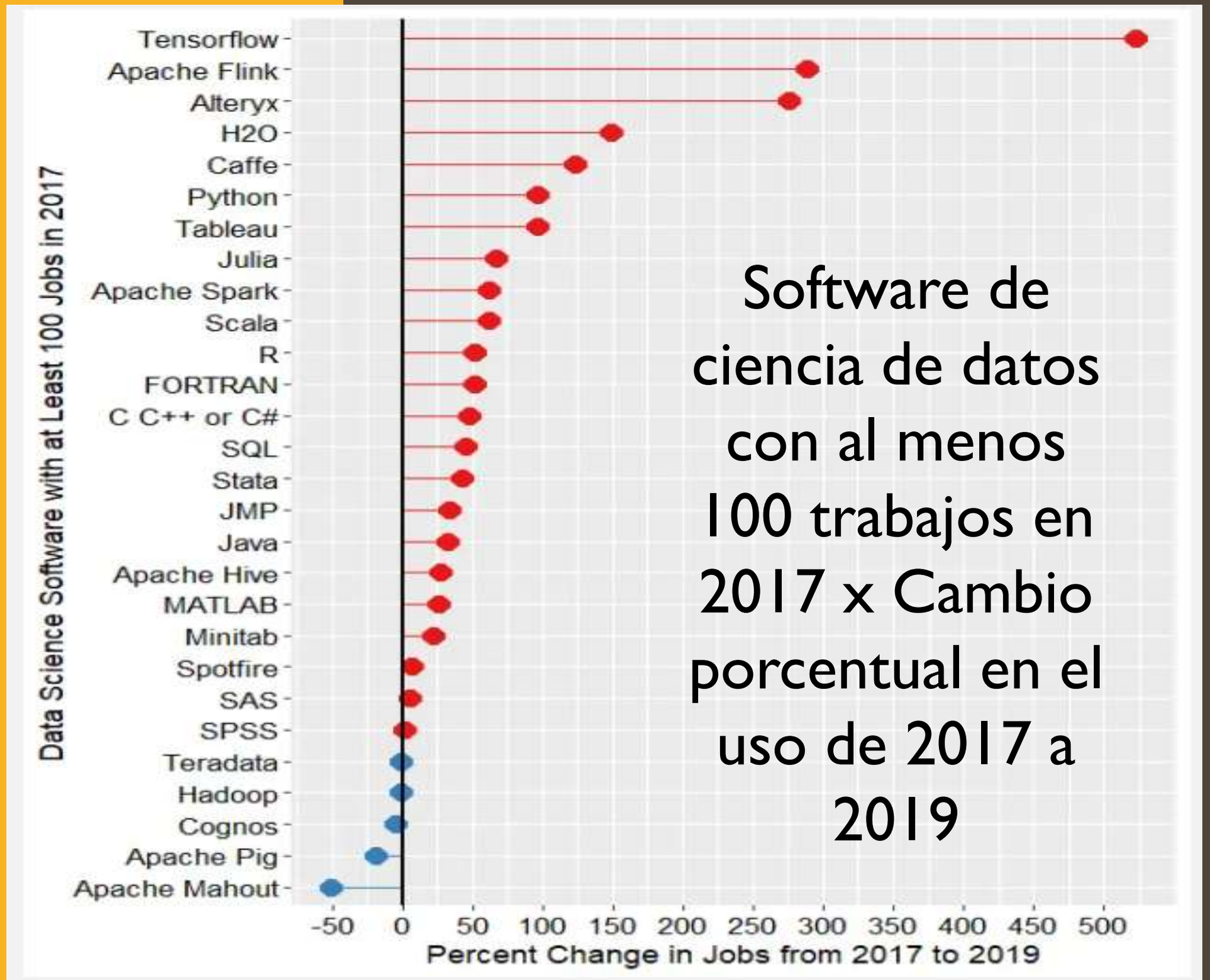


Figure 1b. Number of jobs for less popular data science software tools, those with fewer than 250 advertisements.

Software de ciencia de datos menos popular x Número de trabajos de ciencia de datos en Indeed.com

Fuente: https://www.r4stats.com

COMPARACION

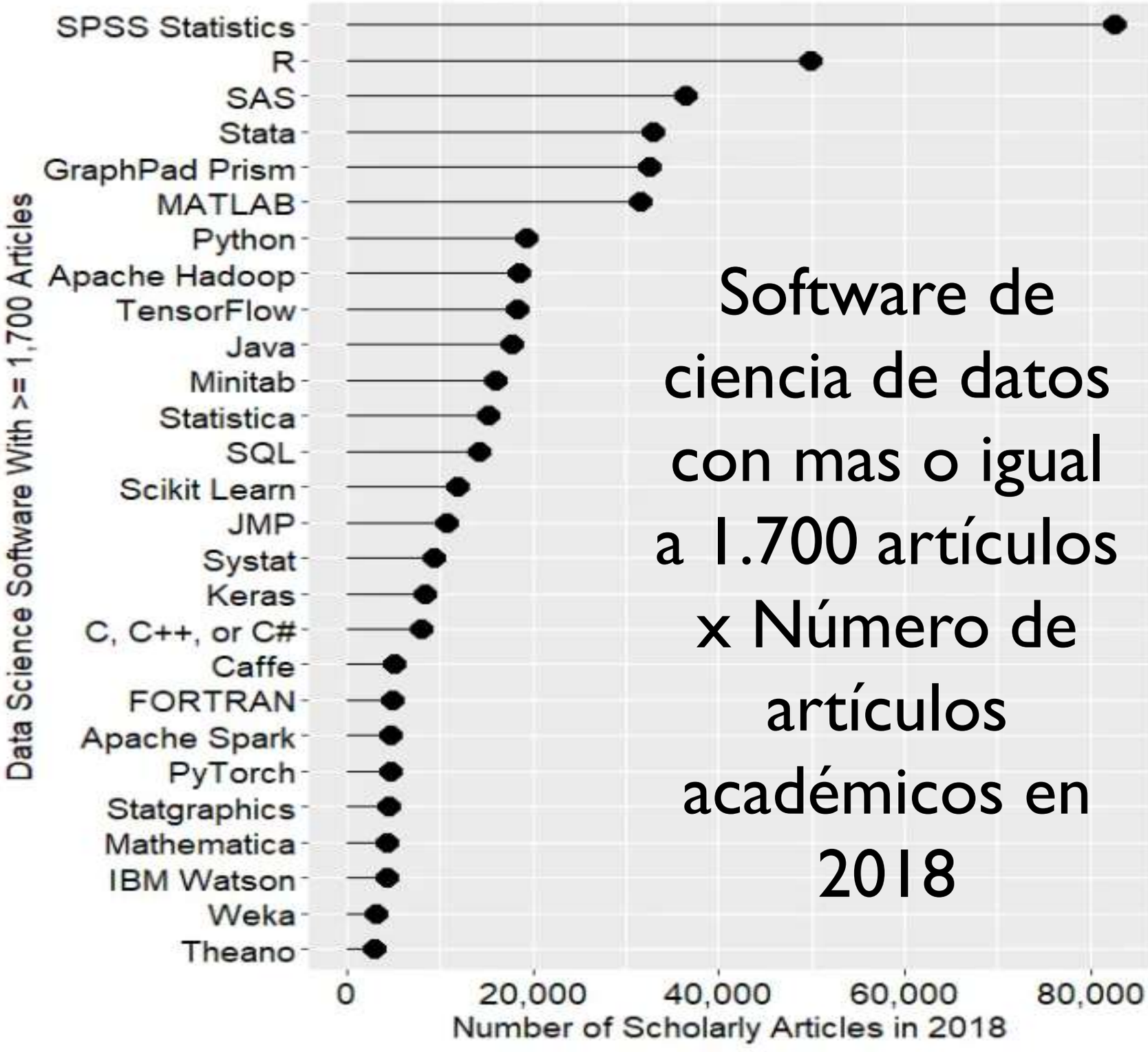


Software de ciencia de datos con al menos 100 trabajos en 2017 x Cambio porcentual en el uso de 2017 a 2019

Figura 1c. Cambio porcentual en las listas de trabajos de 2017 a 2019. Solo se muestra el software que tuvo al menos 100 trabajos en 2017.

Fuente: https://www.r4stats.com

COMPARACION

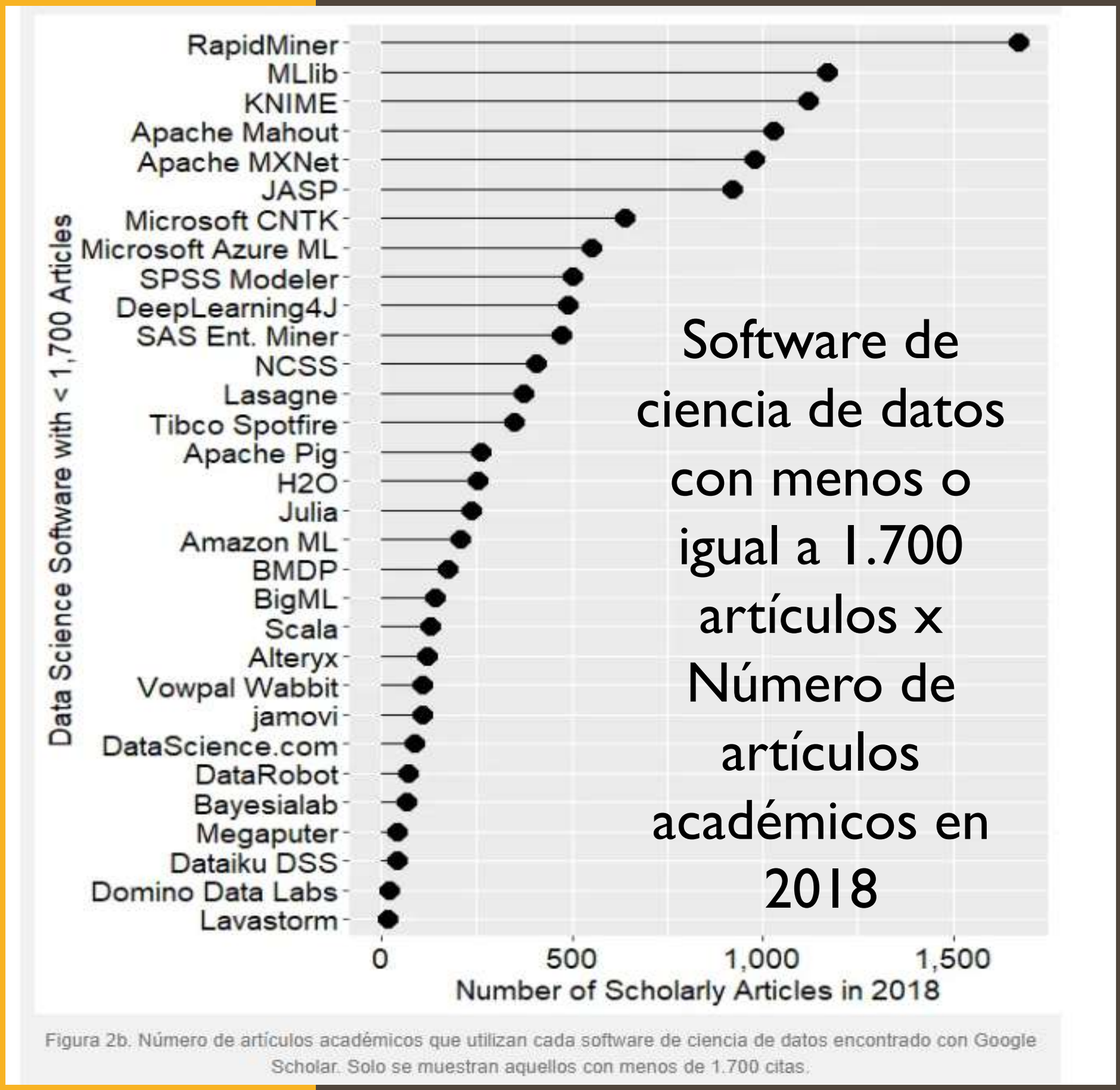


Software de ciencia de datos con mas o igual a 1.700 artículos x Número de artículos académicos en 2018

Figura 2a. La cantidad de artículos académicos encontrados en Google Scholar, para software de ciencia de datos. Solo se muestran aquellos con más de 1.700 citas.

Fuente: https://www.r4stats.com

COMPARACION

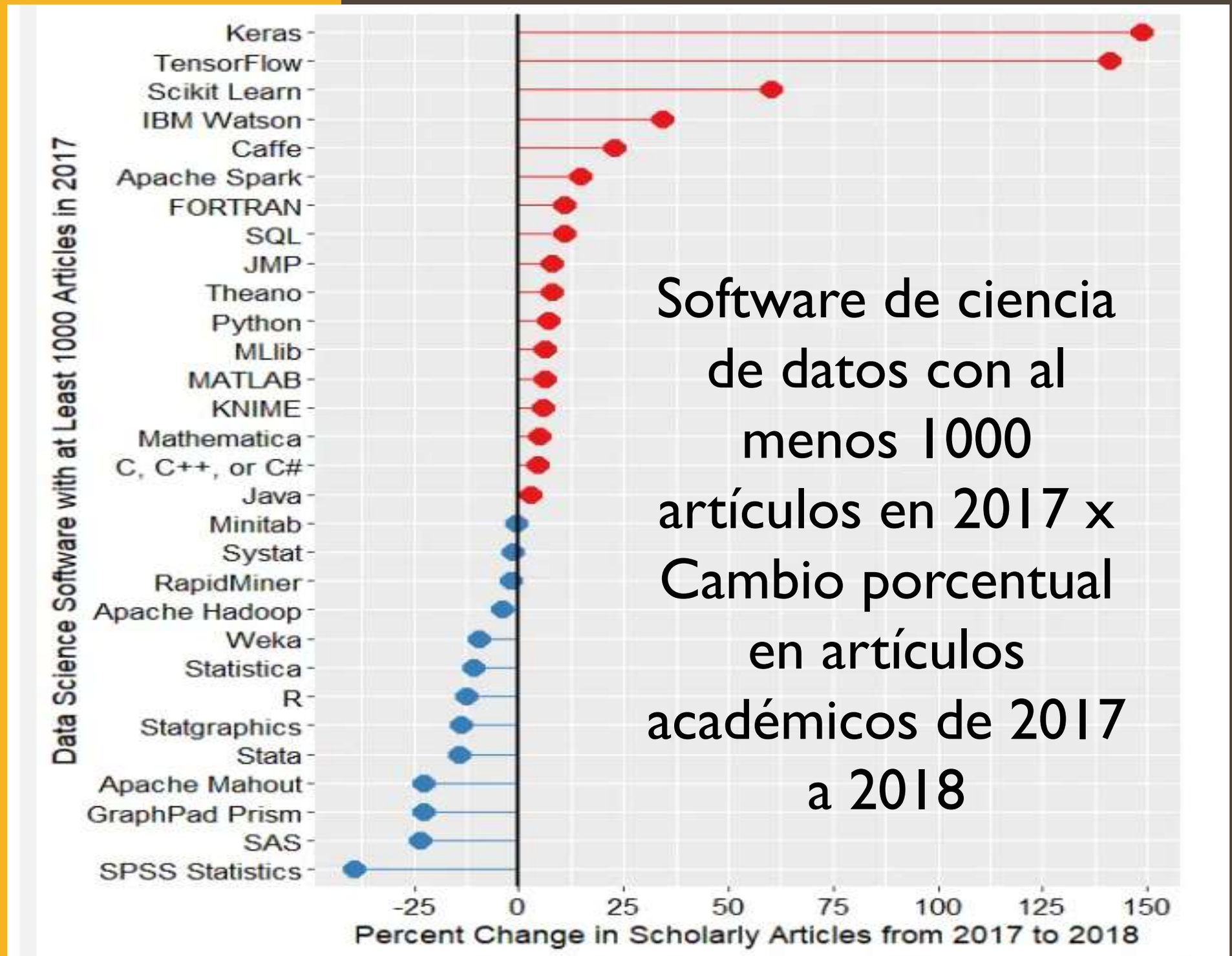


Software de ciencia de datos con menos o igual a 1.700 artículos x Número de artículos académicos en 2018

Figura 2b. Número de artículos académicos que utilizan cada software de ciencia de datos encontrado con Google Scholar. Solo se muestran aquellos con menos de 1.700 citas.

Fuente: https://www.r4stats.com

COMPARACION



Software de ciencia de datos con al menos 1000 artículos en 2017 x Cambio porcentual en artículos académicos de 2017 a 2018

Figura 2c. Cambio en la tasa de citas de Google Académico en los dos años completos más recientes, 2017 y 2018.

Fuente: https://www.r4stats.com

COMPARACION

Número de visitas académicas de Google

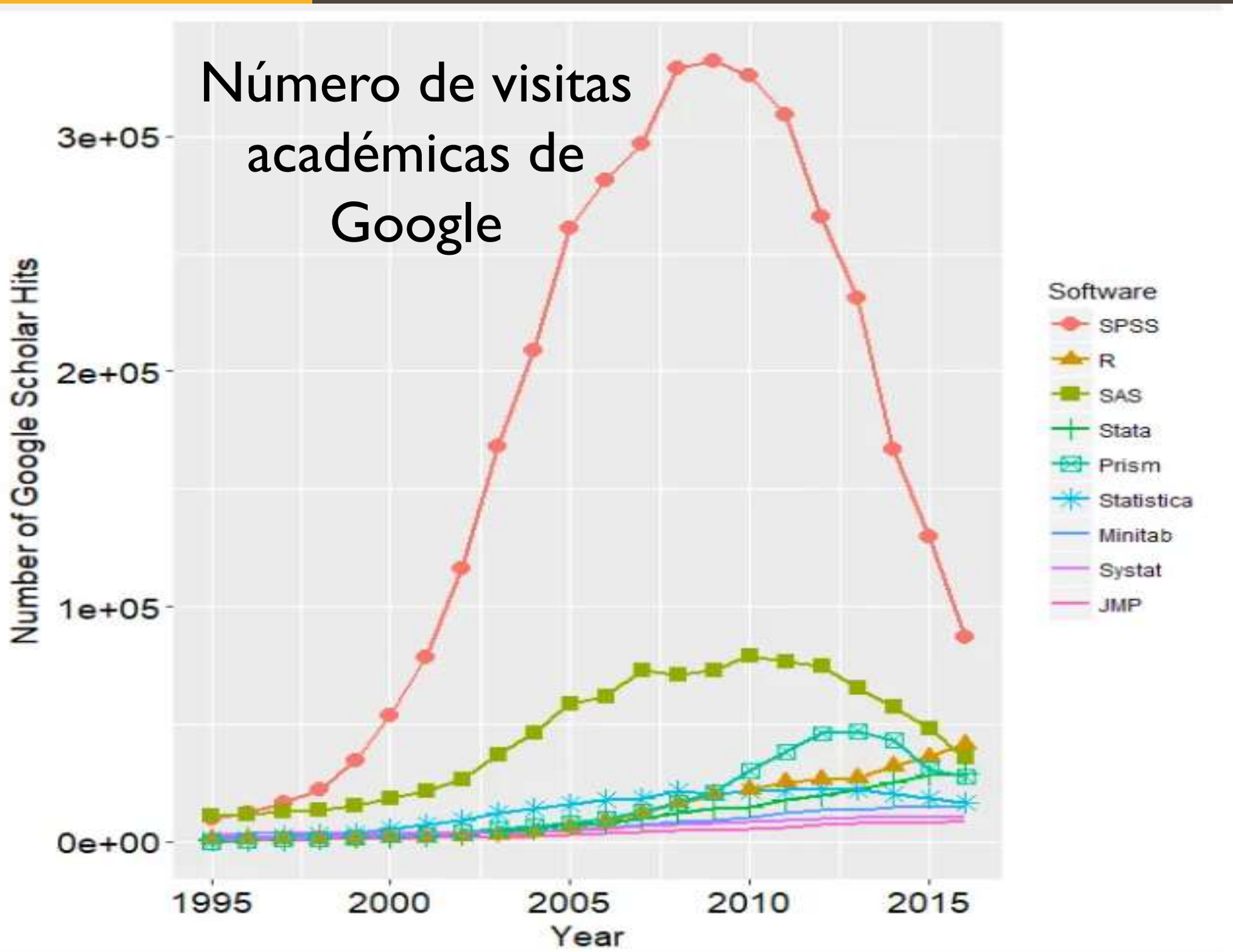


Figura 2d. La cantidad de citas de Google Académico para cada paquete de estadísticas clásicas por año desde 1995 hasta 2016.

Fuente: https://www.r4stats.com

COMPARACION

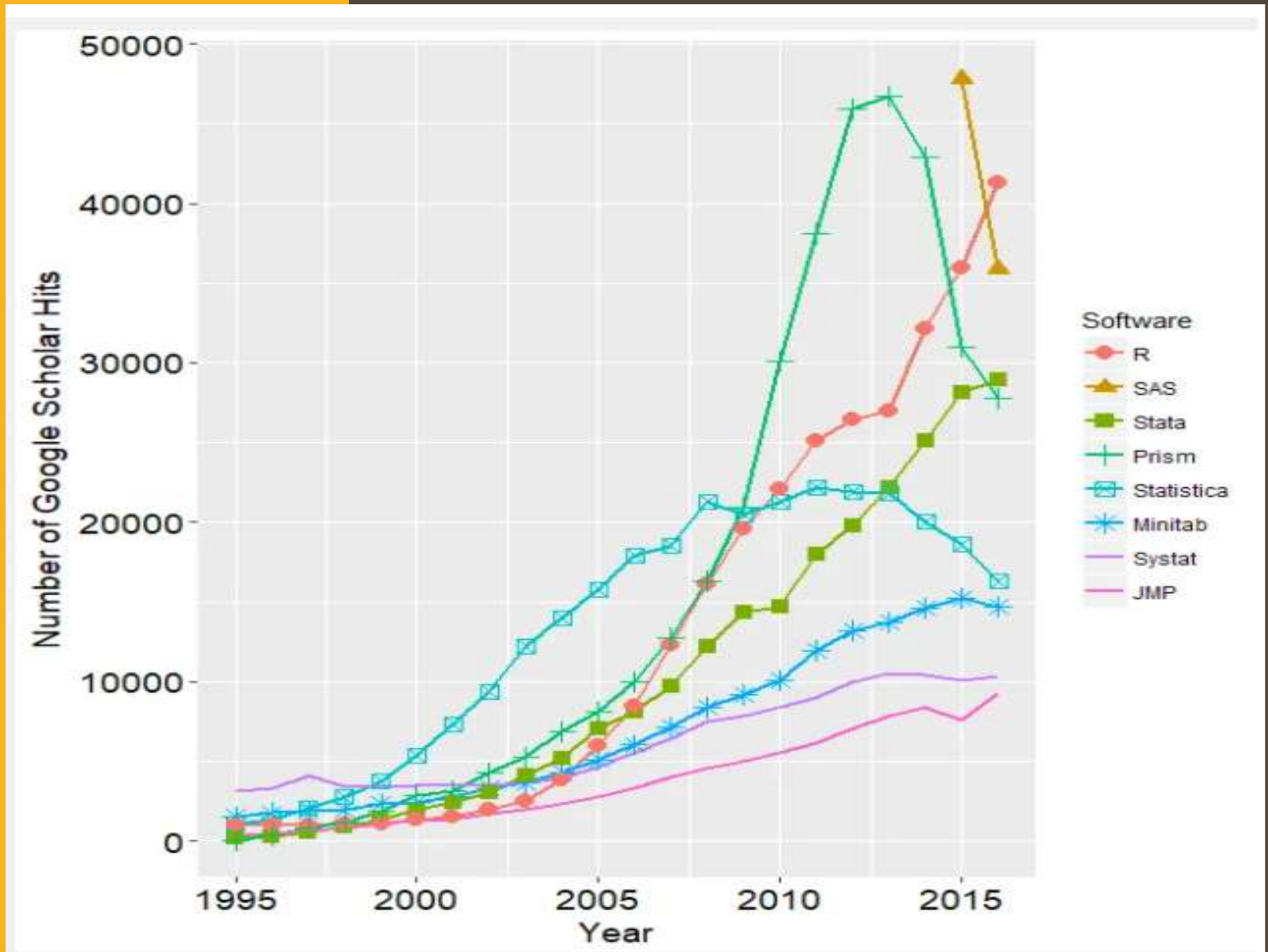


Figura 2e. El número de citas de Google Scholar para cada paquete de estadísticas clásico desde 1995 hasta 2016, esta vez con SPSS eliminado y SAS incluido solo en 2014 y 2015. La eliminación de la escala ampliada de SPSS y SAS hace que sea más fácil ver el rápido crecimiento de los menos populares. paquetes.

Fuente: https://www.r4stats.com

COMPARACION

Figura 3a. Cuadrante mágico de Gartner para plataformas de ciencia de datos y aprendizaje automático de su informe de 2019 (gráfico realizado en noviembre de 2018, informe publicado en 2019).

Fuente: https://www.r4stats.com

COMPARACION

R (azul) vs SAS (naranjo)

Fuente: https://www.r4stats.com

COMPARACION

R (azul) vs Python (naranjo)

Fuente: https://www.r4stats.com

VENTAJAS RSTUDIO

"The popularity of R at universities could threaten SAS Institute, the privately held business software company that specializes in data analysis software."

–New York Times, 2009

"La popularidad de R en las universidades podría amenazar a SAS Institute, la empresa privada de software empresarial que se especializa en software de análisis de datos"

VENTAJAS RSTUDIO

"I think it addresses a niche market for high-end data analysts (...). We have customers who build engines for aircraft. I am happy they are not using freeware when I get on a jet"

–Anne H. Milley, directora de marketing en SAS

"Creo que se dirige a un nicho de mercado para analistas de datos de alto nivel (...). Tenemos clientes que fabrican motores para aviones. Me alegro de que no estén usando software gratuito cuando subo a un jet ".

VENTAJAS RSTUDIO

¿QUIÉNES NO ESTÁN DE ACUERDO?

- Google
- Microsoft
- Facebook
- Twitter
- Ford
- Uber
- McKinsey
- IBM
- HP
- Airbnb
- Roche
- New York Times
- Mahindra
- Tata
- American Express
- Bank of America
- Citibank
- JP Morgan
- HSBC
- Wells Fargo
- Lloyds Banking
- Oracle
- Infosys
- ...

Fuente: http://www.listendata.com/2016/12/companies-using-r.html

COMPARACION

redeszone.net/noticias/seguridad/vulnerabilidad-critica-python-miles-programas/

D-LINK

RZ redes zone

WiFi | Evitar firewall | Picos de latencia | Mejorar WiFi | Configuracion OpenVPN | Mejores switches | Mejores AP | Mejores Routers | Mejores Wi-Fi

Portada > Noticias > Seguridad > Un fallo critico al validar la IP de Python afecta a miles de programas

Seguridad

Un fallo crítico al validar la IP de Python afecta a miles de programas

Javier Jiménez | Publicado el 03 de mayo, 2021 · 09:22

En muchas ocasiones surgen fallos de seguridad que ponen en riesgo los programas y dispositivos que usamos en nuestro día a día. En este caso nos hacemos eco de una importante vulnerabilidad que afecta a la **validación de direcciones IP en Python**, algo que afecta a miles de programas. Vamos a explicar en qué consiste este problema y cómo puede afectar a los usuarios.

Fuente: https://www.redeszone.net/noticias/seguridad/vulnerabilidad-critica-python-miles-programas/

COMPARACION

Una vulnerabilidad en la validación de IP afecta a Python

Se trata de una vulnerabilidad crítica que afecta a la **biblioteca estándar de Python**. Afecta a la validación de direcciones IP, algo que ya ha ocurrido en la biblioteca netmask hace unos meses. El fallo ha sido registrado como CVE-2021-29921.

Concretamente afecta al módulo **ipaddress de Python 3.x** y se debe a un cambio que realizaron en 2019. Como decimos, hace unos meses vimos una serie de vulnerabilidades que afectaban a la máscara de subred. Ahora afecta también a esta biblioteca estándar de Python.

Esta vulnerabilidad provoca un **análisis incorrecto de las direcciones IP** por parte de la biblioteca estándar ipaddress. Este módulo se encarga de que los desarrolladores puedan crear fácilmente direcciones IP, redes e interfaces. Hay que indicar que una dirección IPv4 puede aparecer en formato decimal, entero, octal o hexadecimal, aunque lo más normal es que aparezca en el primer formato. Podemos ver todos los **datos que pertenecen a una IP**.

Aquí llega el problema con los ceros a la izquierda. Es lo mismo que ocurría con la biblioteca netmask. La manera en la que gestiona esa IP cambia al agregar un valor cero antes de la IP en formato decimal. Lo que ocurre con la vulnerabilidad de Python es que los ceros a la izquierda los descartaría.

Esta vulnerabilidad permite ataques remotos

Fuente: https://www.redeszone.net/noticias/seguridad/vulnerabilidad-critica-python-miles-programas/

COMPARACION

En resumen

- SAS
  - Caro
  - Buen soporte comercial
  - Futuro incierto
- Python
  - Gratis
  - Versátil
  - Escalable
- R
  - Gratis
  - Específico

Fuente: https://www.r4stats.com

COMPARACION

	Costo	Open/Closed Source	Versión gratis	Popularidad
Python	$0	Open	Por defecto	Alta y subiendo
R	$0	Open	Por defecto	Alta y subiendo
SAS	$8700	Closed	Solo en EEUU	Alta y bajando

Fuente: https://www.r4stats.com

TAREA 1

NASA y SPACEX.

ENTREVISTA CON EL DIVULGADOR CIENTÍFICO

Steven Johnson: "El análisis de los datos es nuestra mejor defensa"

El escritor Steven Johnson. / CAPITÁN SWING

9 Se lee en minutos

La mayor epidemia hasta entonces en la historia de Londres. Un asesino con nombre de bacteria siembra las calles de cadáveres. Un médico tenaz le sigue la pista, con la ayuda indispensable de un párroco descreído

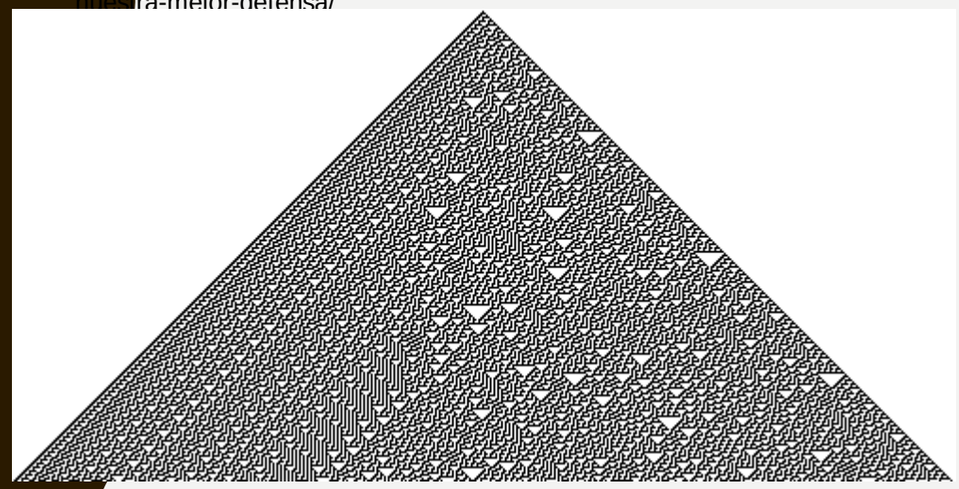
Un artículo de Ricardo Mir de Francia

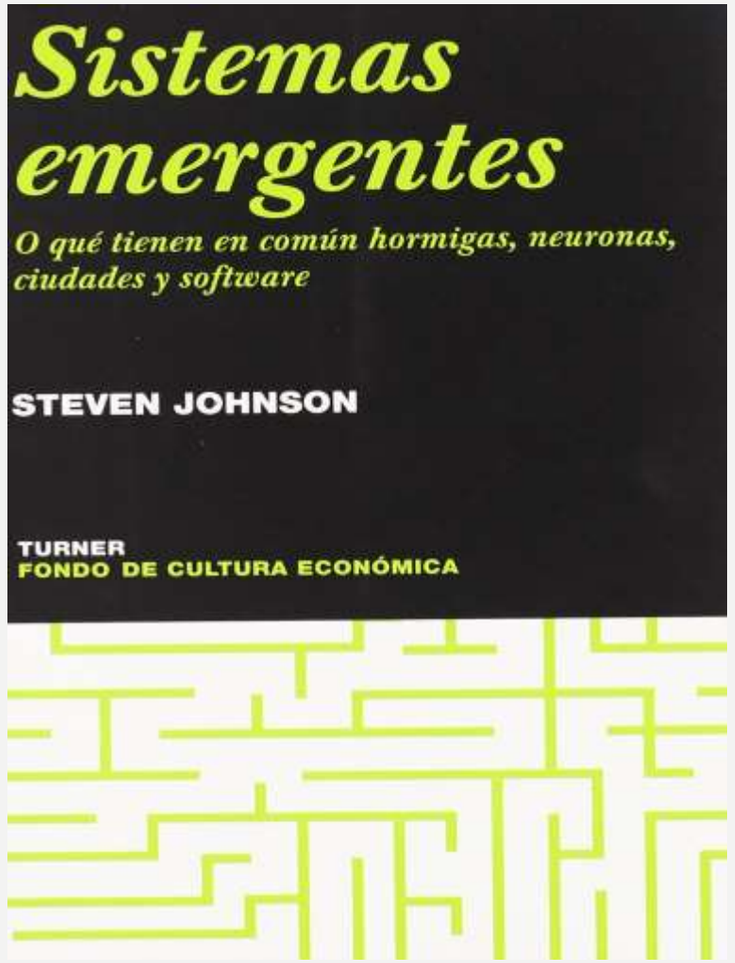
13 de junio del 2020, 08:00

Fuentes:

https://www.elperiodico.com/es/cuaderno/20200613/steven-johnson-entrevista-mapa-fantasma-coronavirus-colera-7996164

http://espiasdecocina.com/steven-johnson-el-anlisis-de-los-datos-es-nuestra-mejor-defensa/





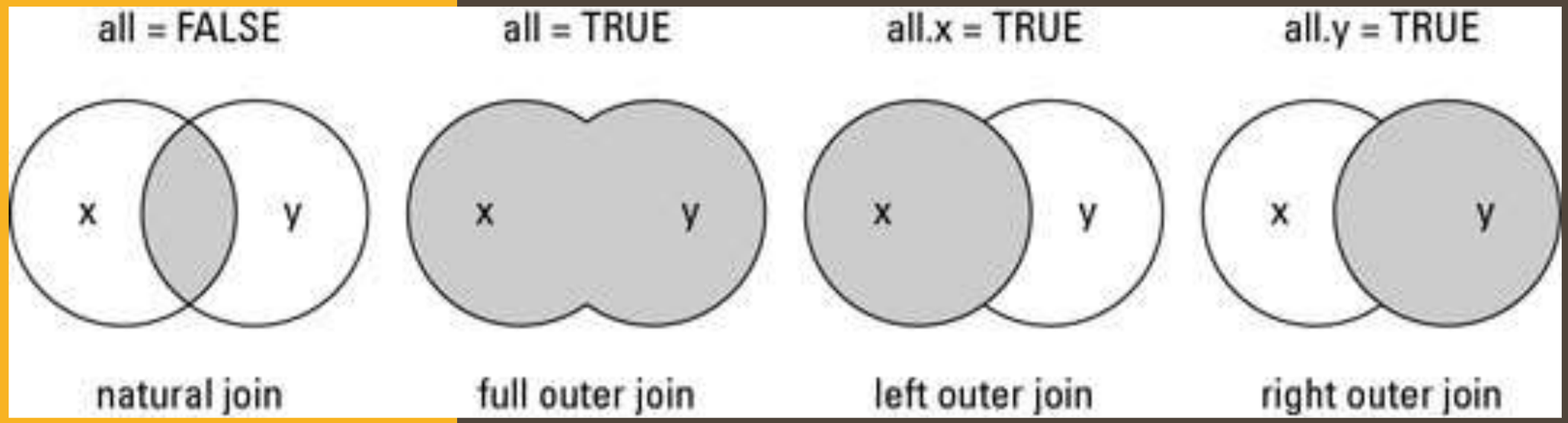
”Stephen Johnson, propone que la organización espontánea y sin leyes explícitas que ocurre en las colonias de hormigas, en el cerebro humano o en las ciudades, se debe a las reglas de la emergencia según las cuales los agentes de un nivel inferior adoptan comportamientos de un nivel superior. Para demostrarlo, nos lleva en un recorrido por algunas aplicaciones de su teoría que incluyen la formación, en el futuro, de una World Wide Web inteligente.” (https://books.google.cl/)

“Sistemas emergentes son sistemas complejos de adaptación que despliegan comportamientos emergentes. Se caracterizan por resolver problemas, al menos en apariencia, espontáneamente; es decir, sin recurrir a una inteligencia de tipo centralizado o jerarquizado (descendente), sino de forma ascendente, desde la base, a partir de masas de elementos relativamente no inteligentes. El comportamiento separado, individual, de cada uno de los agentes, al aumentar la escala comienza a producir un comportamiento colectivo propio de un nivel de organización superior, a pesar de la aparente carencia de organización en forma de leyes o instrucciones provenientes de una autoridad superior.

Ejemplos de estos sistemas de autoorganización, sorprendentemente parecidos entre sí, se están analizando en las ciencias naturales y sociales desde finales del siglo XX: las colonias del moho del fango (Dictyostelium discoideum) estudiadas por Evelyn Fox Keller y Lee Segel (biomatemáticos inspirados en Alan Turing), los barrios urbanos estudiados por Jane Jacobs2 o las redes del cerebro humano estudiadas por Marvin Minsky. El software y las redes sociales de reciente creación se desarrollaron siguiendo los mismos patrones. Sistemas emergentes; qué tienen en común hormigas, neuronas, ciudades y software (Steven Johnson)

COMPARACION

FUNCIÓN MERGE RSTUDIO



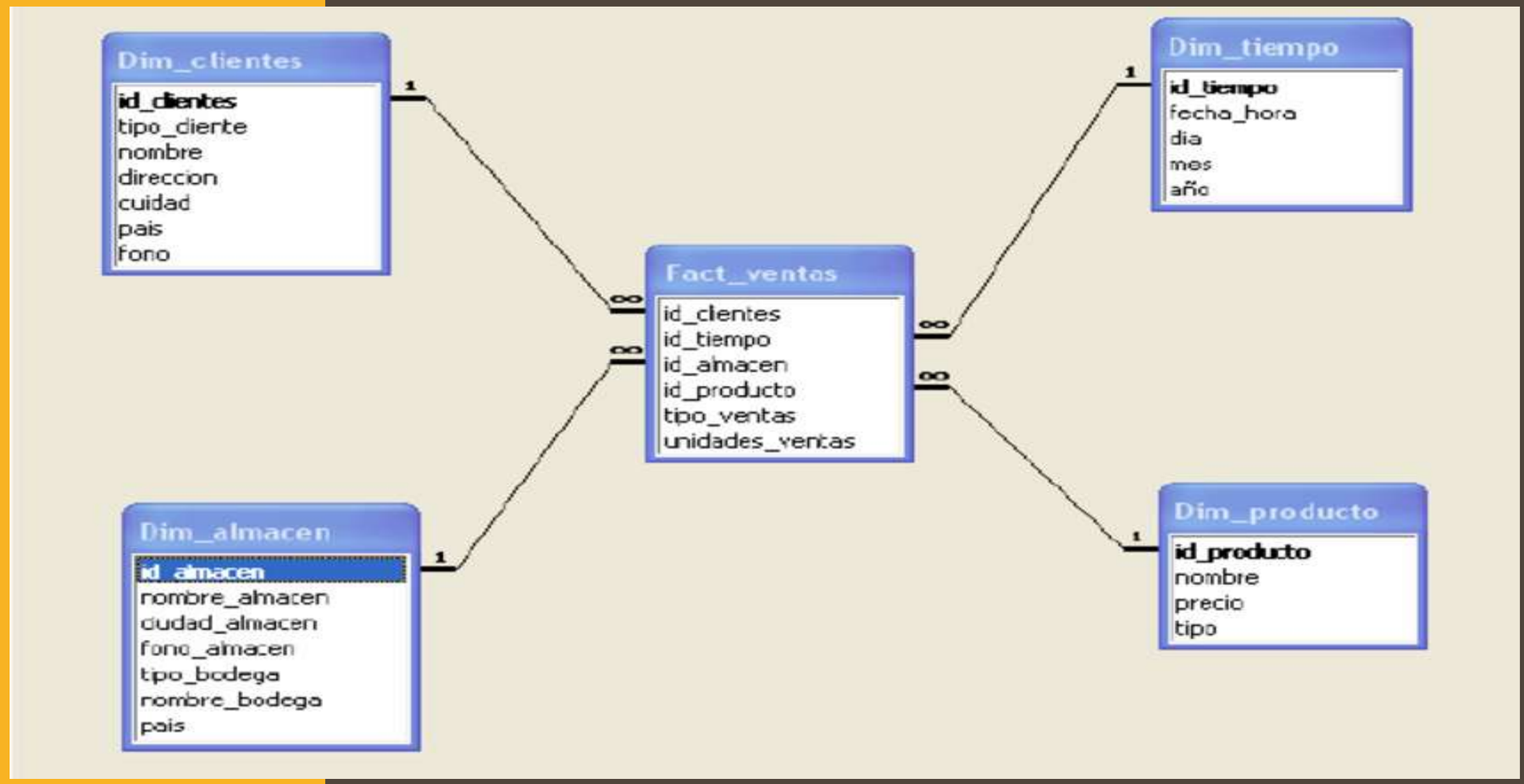
Fuente: https://www.r4stats.com

- Pensemos sobre la aleatoriedad

¿Existe REALMENTE LA ALEATORIEDAD EN LENGUAJES DE PROGRAMACIÓN?

COMPARACION

MODELO RELACIONAL TIPO ESTRELLA



Fuente: https://www.r4stats.com