CLASE 02 - PROGRAMACIÓN CON R

Diplomado en Análisis de Datos con R e Investigación reproducible para Biociencias.

Dr. José Gallardo Matus | https://genomics.pucv.cl/

Pontificia Universidad Católica de Valparaíso

01 September 2022

PLAN DE CLASE

1. **Introducción**

- ¿Qué es R y Rstudio?
- ¿Por qué usar R para el análisis de datos en biociencias?
- Qué es la investigación reproducible y por qué es importante en Biociencias?.

2. **Práctica con R y Rstudio (cloud)**

- Elaborar un script para el análisis de datos con R.
- Familiarizarse con manipulación de objetos de R y datos de biociencias.

¿QUÉ ES R?

1. **R** es un lenguaje y entorno de programación de código abierto o libre creado por Ross Ihaka y Robert Gentleman en 1993 (University of Auckland) para realizar análisis estadísticos y gráficos.
2. Los usuarios de R tienen la libertad de ejecutar, copiar, distribuir, estudiar, modificar y mejorar el ***software***.
3. Utilizar **R** supone un ahorro económico para los estudiantes, las instituciones educativas o incluso las empresas que decidan usarlo.

¿POR QUÉ USAR "R"?

1. Aprender a usar **R** te da ***independencia digital***, te permite ***cooperar con otros*** y ***beneficiarte de la ayuda de otros.***
2. Actualmente existen cerca de **17.000 librerías o apps** disponibles de forma gratuita para trabajar con R en ámbitos tan diferentes como las ciencias sociales, la economía, la astronomía, la ingeniería y por su puesto las biociencias.
3. **R** permite entonces difundir el conocimiento a toda la sociedad y no solo a los que pueden pagar por ella.

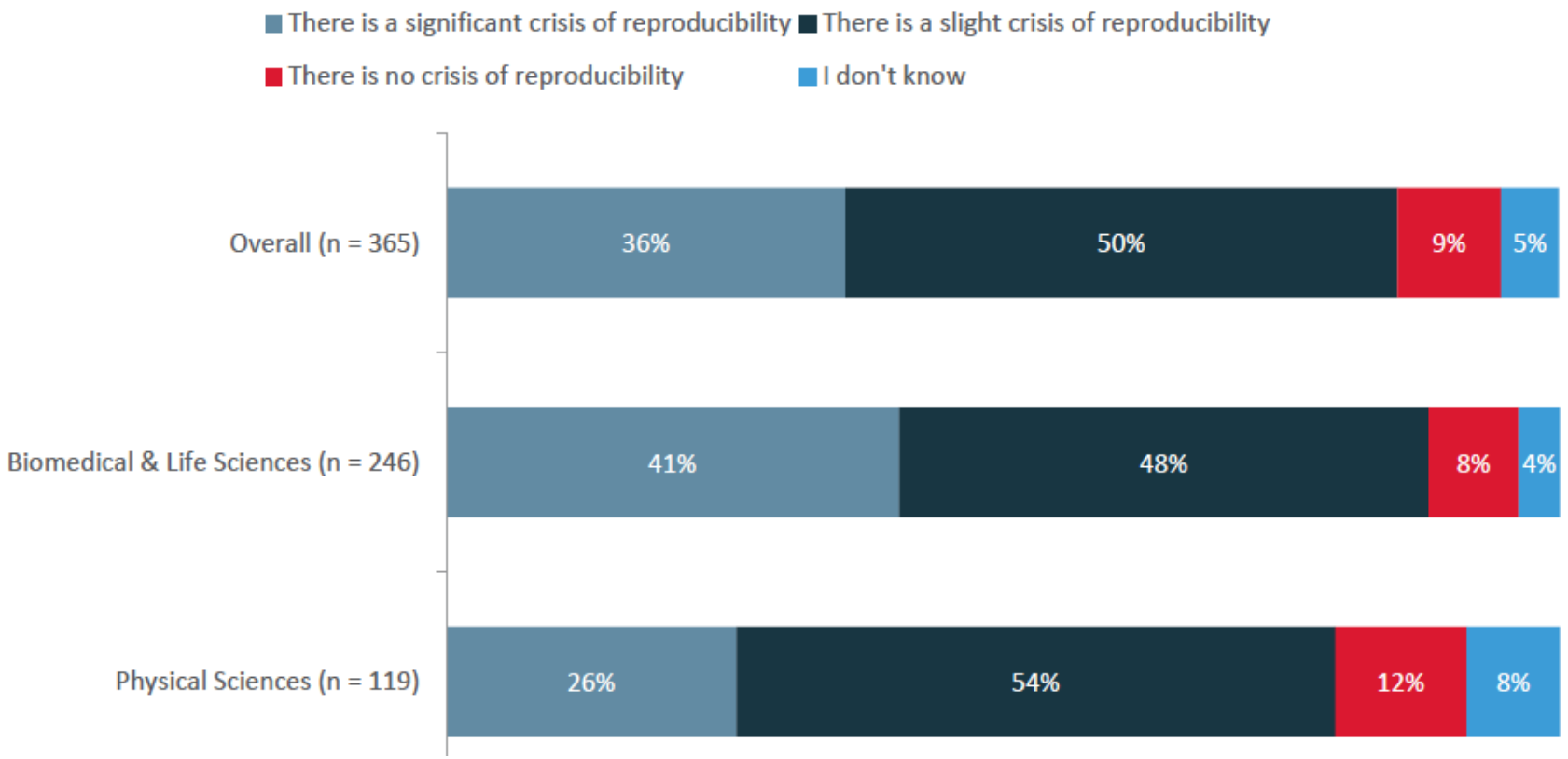
INVESTIGACIÓN REPRODUCIBLE

La investigación reproducible nace de la idea de que cualquier investigador pueda **reproducir los resultados de un estudio** al analizar los datos con los que fueron generados.

Peng. 2011

CRISIS DE REPRODUCIBILIDAD

70 % (1103/1,576) de los investigadores declaran que quisieron pero no pudieron reproducir un experimento de otro científico.



Baker. 2016

ALGUNOS CRITERIOS DE REPRODUCIBILIDAD

- Los datos están almacenados en formato abierto (texto).
- **Todo el análisis y manejo de datos se hace mediante código.**
- El código genera las tablas y figuras finales.
- **Los datos brutos están separados de los datos derivados.**
- Existe un ***'script'*** maestro que ejecuta todos los pasos del análisis ordenadamente.
- **Existe un documento README que explica los objetivos y organización del proyecto.**
- Tanto el reporte, como los datos y código son públicos.

Sánchez et al. 2016

BENEFICIOS EN BIOCIENCIAS

- **Permite la ejecución de tareas de análisis repetitivo sin esfuerzo.**
- Muy fácil corregir y regenerar resultados, tablas y figuras.
- **Reducción drástica del riesgo de errores.**
- Facilita la colaboración.
- **Mayor facilidad para escribir reportes y publicaciones.**
- Facilita el proceso de revisión por pares.
- **Ahorro de tiempo y esfuerzo al reutilizar código en diferentes proyectos.**

RUTA DEL ANÁLISIS DE DATOS REPRODUCIBLE CON R

1. **Toma de datos.**
   Es importante estandarizar y mantener estructura.
2. **Manipulación de datos.**
   Es importante cuidar los datos originales.
   Trabajaremos con R + Rstudio
3. **Análisis datos integrado con texto.**
   Facilita la elaboración automática de reportes.
   Trabajaremos con RMarkdown.
4. **Control de versiones.**
   Permite respaldar y recuperar versiones de un proyecto. Muy util para supervisar el trabajo de los analistas. Trabajaremos Github.
5. **Publicar resultados.** Es importante comunicar de forma efectiva. Daremos recomendaciones clave.

CONCEPTOS BÁSICOS DE PROGRAMACIÓN

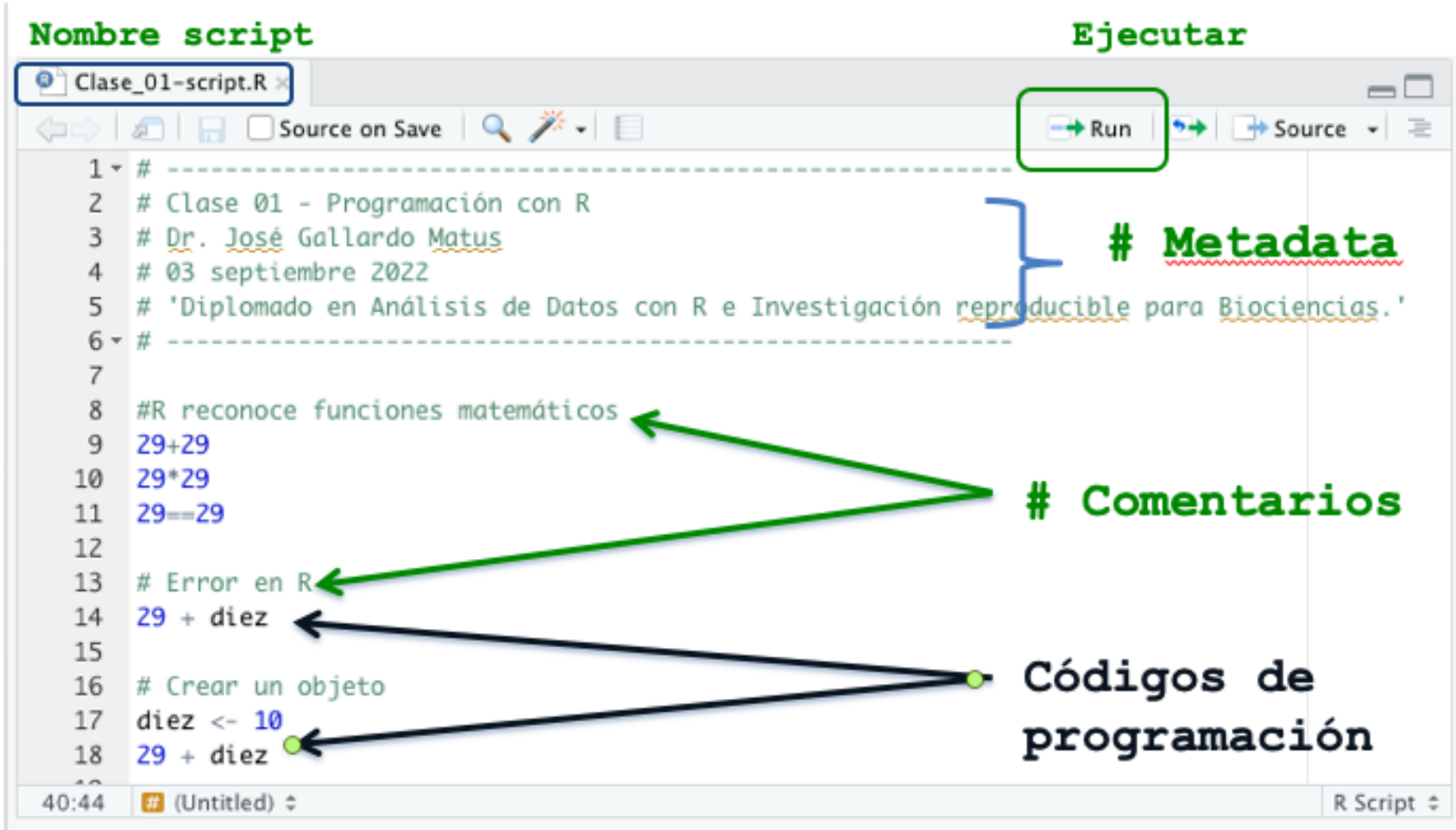
Metáfora de la maquina expendedora de bebidas

1. La máquina tiene una función específica.
2. Los productos son objetos almacenados de forma ordenada.
3. Los objetos tienen características (Nombre, precio, ubicación).
4. Para comprar debo seguir una secuencia de pasos (similar a un programa = códigos en secuencia).

¿QUÉ ES UN SCRIPT?

1. Los scripts son documentos de texto con una secuencia de comandos que permiten ejecutar programas.
2. Estos archivos son iguales a cualquier documentos de texto, pero R puede leer y ejecutar el código que contienen.
3. Los códigos de R están contenidos en librerías o packages o aplicaciones.
4. Algunos script que usaremos en este curso tienen extensión de archivo .R, por ejemplo mi_script.R.

EJEMPLO R SCRIPT



R ES UN LENGUAJE ORIENTADO A OBJETOS

Tipos de objetos para trabajar con R

Vector

- 1 column or row of data
- 1 type (numeric or text)

Matrix

- multiple columns and/or rows of data
- 1 type (numeric or text)

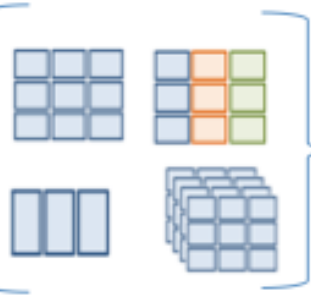
Data Frame

- multiple columns and/or rows of data
- multiple types

Array

-3 dimensiones
- 1 tipo: numérico
- o caracter

Listas



- Conjunto de objetos diversos

OBJETO: DATA.FRAME

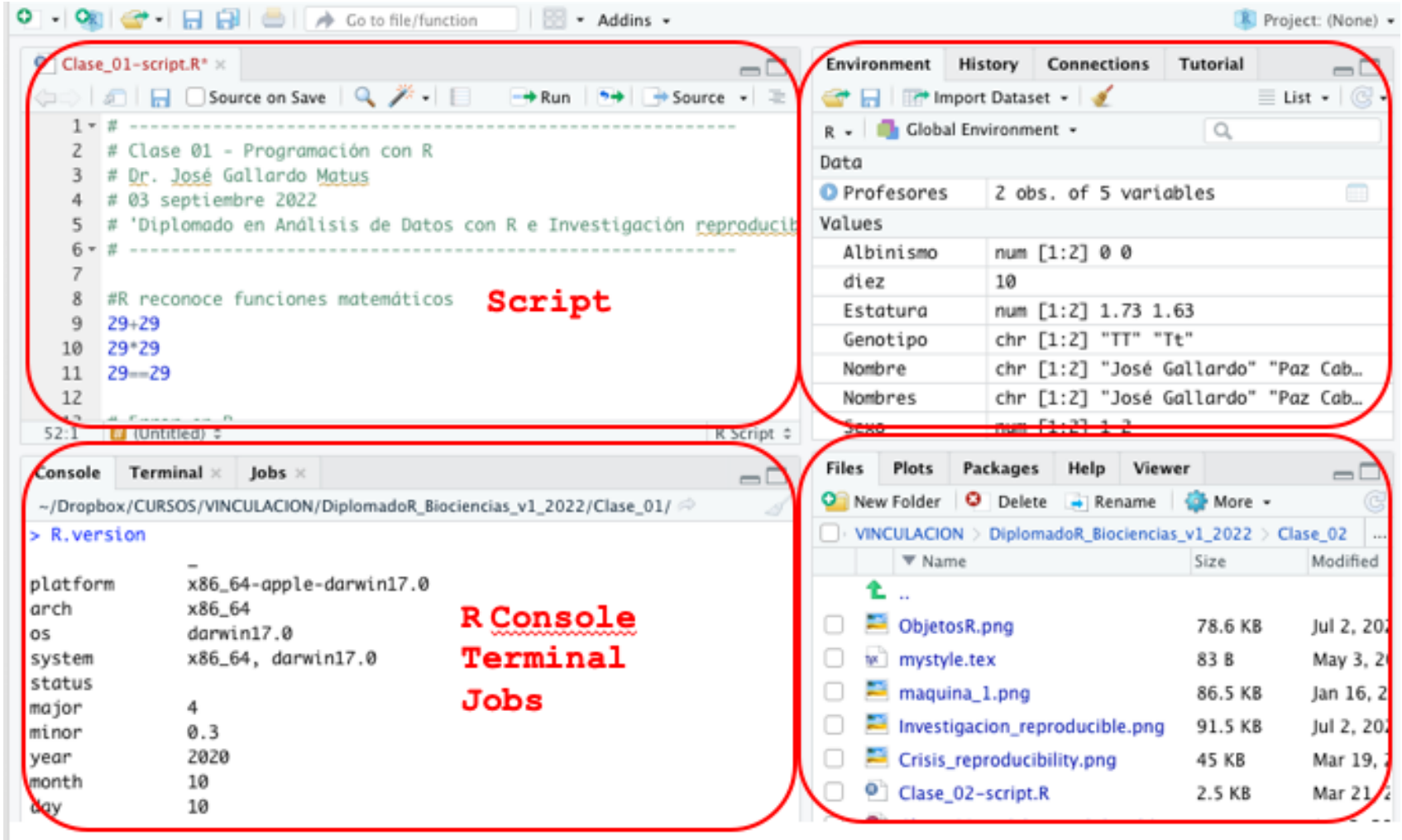
Principales características.

- Objeto similar a una tabla de datos.
- Almacenan texto o números.
- Primera fila contiene el nombre de las variables.
- Puedo unir con otro **data.frame**.
- Puedo aplicar funciones para calcular estadísticos.
- Pero, no tiene atributos de una matriz, ni de un vector, no es una serie de tiempo.

¿QUÉ ES R STUDIO?

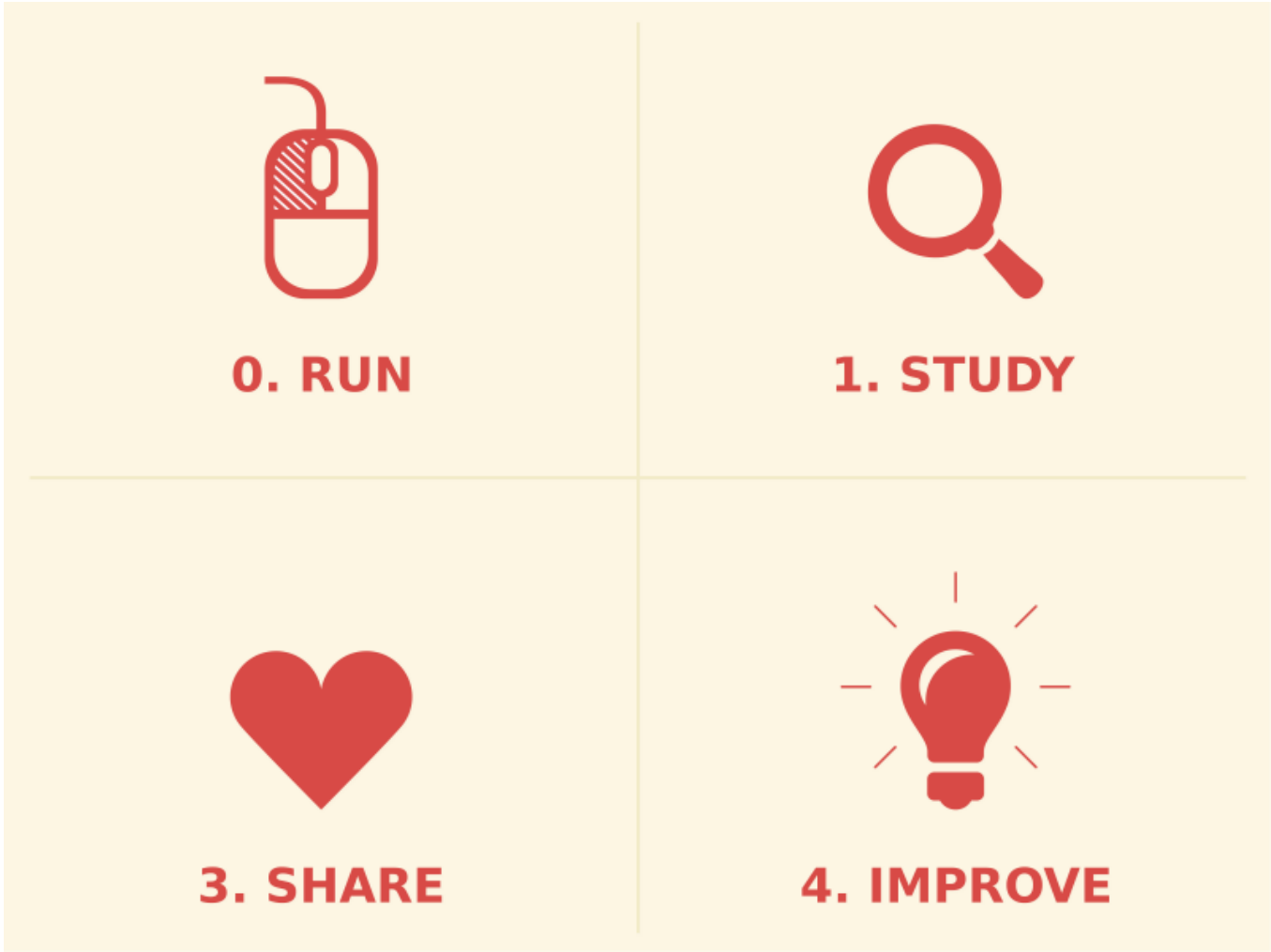
1. **Rstudio** es el más popular entorno de desarrollo integrado (integrated development environment, IDE) para trabajar con **R**.
2. **Rstudio** es un ***software*** libre y de código abierto creado por **Joseph J. Allaire en 2009** para la ciencia de datos, la investigación científica y la comunicación técnica.
3. Actualmente es mantenido por la Corporación de Beneficio Público **Rstudio PCB**, la que ha creado otros software como Rmarkdown.

EJEMPLO RSTUDIO - VERSION CLOUD



PRÁCTICA PROGRAMACIÓN CON R

Guía de trabajo programación con R en Rstudio.cloud.



RESUMEN DE LA CLASE

- Investigación reproducible.
- Ruta del análisis de datos reproducible con **R**.
- Iniciamos un proyecto de análisis de datos con **R**.
- Escribimos un script o código de programación de **R** con **Rstudio**.
- Nos familiarizamos con la manipulación de objetos y datos de R: vector, matriz, data.frame.