

# **CLASE 01 - PROGRAMACIÓN CON R**

## **DBT 845 - Investigación reproducible y análisis de datos biotecnológicos con R.**

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

Pontificia Universidad Católica de Valparaíso

19 March 2022

# PLAN DE CLASE

## 1. Introducción

- ▶ ¿Qué es R y Rstudio?
- ▶ ¿Por qué usar R para el análisis de datos biotecnológicos?
- ▶ Investigación reproducible.

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

- ▶ Iniciar un proyecto de análisis de datos biotecnológicos con R.
- ▶ Familiarizarse con manipulación de objetos de R y datos biotecnológicos.

# ¿QUÉ ES R?

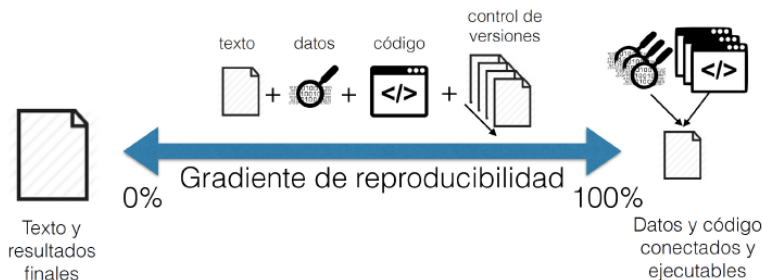
1. **R** es un lenguaje y entorno de programación de código abierto o libre creado por Ross Ihaka y Robert Gentleman en 1993 (University of Auckland) para realizar análisis estadísticos y gráficos.
2. Los usuarios de R tienen la libertad de ejecutar, copiar, distribuir, estudiar, modificar y mejorar 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 la biotecnología.
3. **R** permite entonces difundir el conocimiento a toda la sociedad y no solo a los que pueden pagar por ella.

# INVESTIGACIÓN REPRODUCIBLE

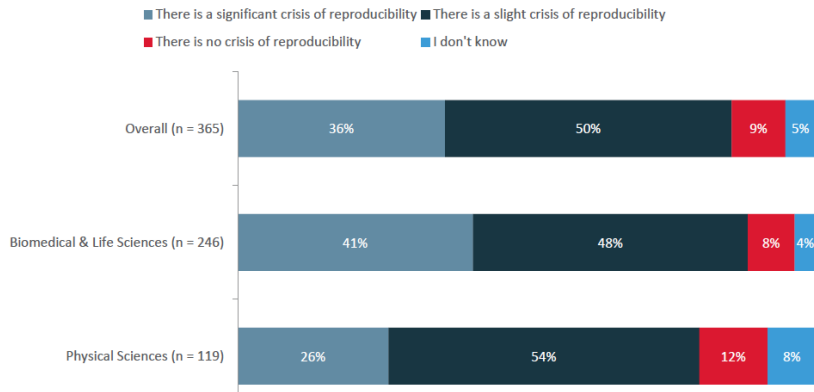
La investigación reproducible implica que desde los mismos datos y códigos se generarán los mismos resultados.



Peng. 2011

# CRISIS DE REPRODUCIBILIDAD

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



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 BIOTECNOLOGÍA

- ▶ **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 colaboración.

Trabajaremos con RMarkdown.

## 4. Publicar resultados y control de versiones.

Es importante comunicar de forma efectiva.

Trabajaremos Github.

# 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).



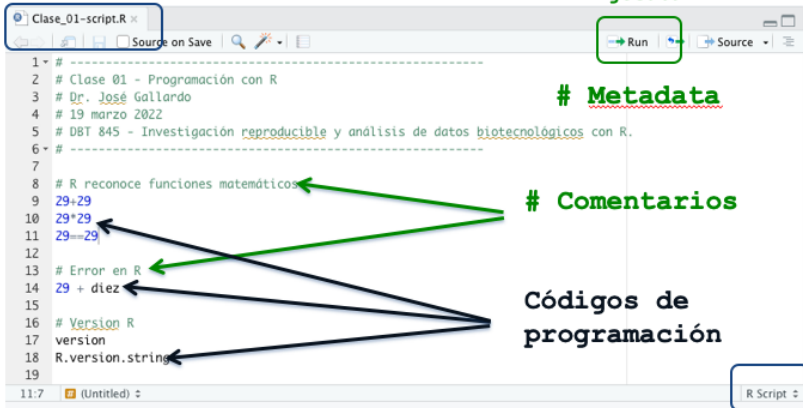
# ¿QUÉ ES UN SCRIPT?

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

Nombre script

Ejecutar



The screenshot shows an R script editor window titled 'Clase\_01-script.R'. The script content is as follows:

```
1 # -----  
2 # Clase 01 - Programación con R  
3 # Dr. José Gallardo  
4 # 19 marzo 2022  
5 # DBT 845 - Investigación reproducible y análisis de datos biotecnológicos con R.  
6 # -----  
7  
8 # R reconoce funciones matemáticas  
9 29+29  
10 29*29  
11 29==29  
12  
13 # Error en R  
14 29 + diez  
15  
16 # Version R  
17 version  
18 R.version.string  
19
```

Annotations on the image include:

- A green box around the title bar 'Clase\_01-script.R' with the label 'Nombre script'.
- A green box around the 'Run' button with the label 'Ejecutar'.
- Green arrows pointing from the text '# Metadata' to lines 2 through 6.
- Green arrows pointing from the text '# Comentarios' to lines 8, 10, 11, 13, and 14.
- Black arrows pointing from the text 'Códigos de programación' to lines 9, 10, 11, 17, and 18.

The status bar at the bottom shows '11:7' and '(Untitled)'.

# R ES UN LENGUAJE ORIENTADO A OBJETOS

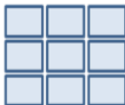
## Tipos de objetos para trabajar con R

### 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



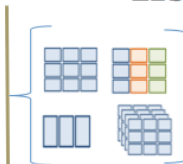
- multiple columns and/or rows of data
- multiple types

### 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

## Barra de herramientas

## R version

The screenshot shows the RStudio Cloud interface with several red annotations. A red box highlights the top menu bar (File, Edit, Code, View, Plots, Session, Build, Debug, Profile, Tools, Help). Another red box highlights the R version dropdown menu (R 4.1.3). The main editor area is labeled 'Script' and contains R code. The console area is labeled 'R Console' and displays the R startup message. The 'Object and history' panel is labeled 'Object and history'. The 'Files, plots, Package' panel is labeled 'Files, plots, Package' and shows a file list.

**Script**

```
1 # -----  
2 # Clase 01 - Programación con R  
3 # Dr. José Gallardo  
4 # 19 marzo 2022  
5 # DBT 845 - Investigación reproducible y análisis de datos biotecnológicos con R  
6 # -----  
7  
8 #R reconoce funciones matemáticas  
9 29+29  
10 29*29  
11 29==29  
12  
1.1 (Untitled) R Script
```

**R Console**

R 4.1.3 - /cloud/project/

R is free software and comes with ABSOLUTELY NO WARRANTY.  
You are welcome to redistribute it under certain conditions.  
Type 'license()' or 'licence()' for distribution details.

R is a collaborative project with many contributors.  
Type 'contributors()' for more information and  
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or  
'help.start()' for an HTML browser interface to help.  
Type 'q()' to quit R.

**Object and history**

Environment is empty

**Files, plots, Package**

Name	Size	Modified
..		
..rhistory	0 B	Mar 19, 2022, 11:36 AM
project.Rproj	205 B	Mar 19, 2022, 1:22 PM
Clase_01-script.R	2.5 KB	Mar 19, 2022, 1:22 PM



# PRÁCTICA CON R Y RSTUDIO.CLOUD

Guía de trabajo programación con R en Rstudio.cloud.



**0. RUN**



**1. STUDY**



**3. SHARE**



**4. IMPROVE**

# 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 código de programación de **R** con **Rstudio cloud**.
- ▶ Nos familiarizamos con la manipulación de objetos y datos de biotecnología: vector, matriz, `dta.frame`.