

OPSO79-1-UCSH2021

Ciencia abierta y profundización  
en RMarkdown.

12 de noviembre 2021

# Ciencia abierta

Introducción y desafíos en sociología

# Introducción

*"La información es poder. Pero como con todo poder, hay quienes lo quieren mantener para sí mismos. La herencia científica y cultural del mundo completa, publicada durante siglos en libros y journals, está siendo digitalizada y apresada en forma creciente por un manjo de corporaciones privadas (...)".*

*"(...) Es tiempo de salir a la luz y en la gran tradición de la desobediencia civil, declarar nuestra oposición a este robo privado de la cultura pública (...) Necesitamos tomar la información, donde sea que esté guardada, hacer nuestras copias y compartirlas con el mundo"*

(Swartz, 2008)



# Introducción

Ciencia abierta como acceso concepción más cercana.

Inspiró a much@s. A una de ellas probablemente la han escuchado (y sí no es así deberían).

La científica Alexandra Elbakyan ha  
creado un [repositorio y página web \(sci-hub\)](#) pirata de más de 87 millones de  
artículos académicos y libros.

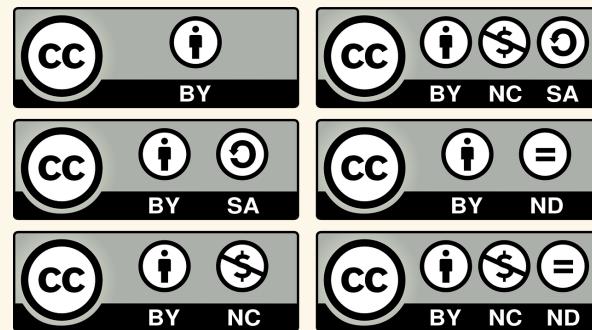
A la "fuerza" consigue un acceso abierto a  
las publicaciones científicas.

Por esto se ha ganado varias demandas.  
En algunos países se encuentra bloqueada  
(e.g. Alemania)



# Introducción

En paralelo a la opción "guerrillera", otras líneas institucionales han avanzado.



Pero la ciencia abierta no es solo el acceso libre a artículos y libros.

Ciencia abierta es un concepto "paraguas", que cubre diferentes usos.

Reflexiones pueden agruparse a lo menos en cinco "escuelas" ([Fecher and Friesike, 2014](#)) o cuatro "dimensiones" ([Breznau, 2021](#)).

# El paraguas



**What do we mean when we talk about Open Science?**

Image courtesy of Robin Champieux

# Dimensiones ciencia abierta

## 1. Acceso abierto

Los resultados de la investigación científica deben ser accesibles a todos/as, removiendo las barreras de pago de las revistas.

Es justo por el uso de fondos públicos.

Si al hacer ciencia nos paramos sobre hombres de gigantes, ¿cómo construir investigación de calidad sin revisar lo producido por otros/as?

**Posibles salidas:** Publicar en revistas open access, liberar prepints, difundir y cooperar en expandir el repositorio sci-hub/lig-gen, sabotaje a grandes revistas, redireccionar el uso de fondos de las universidades, etc.

# Dimensiones ciencia abierta

## 2. Productividad y calidad de las ciencias

Permanente obligación por publicar puede sacrificar la calidad de la misma ciencia

Presión sobre académicos por publicar (para mantener y subir posiciones).

Universidades mutan a instituciones productoras de *papers* que nadie lee.

Solo importa el **cuanto** y el **donde** se publica.

El donde se publica no es garantía de nada:

- importa más el poder simbólico que el argumento
- acceso a publicar relacionado a capital social que académicos construyen
- Activistas que publicaron premeditadamente "**basura**" en prestigiosas revistas.

# Dimensiones ciencia abierta

La presión por publicar:

- incentiva el plagio
- promueve la reiteración de lo ya dicho
- el *p-hacking* (iterar modelos hasta que da significativo)
- invento de data y uso de trabajo ajeno (ayudantes)
- Primacía de los matices no concluyentes

La ciencia puede aportar en otros aspectos, ¿Como medir la calidad de la investigación?

- Relevancia pública de la temática investigada
- Uso de formatos de publicación alternativos (blogs, libros, documentales)
- Vinculación externa y con otros investigadores

# Dimensiones ciencia abierta

## 3. Transparencia del proceso investigativo

Se deben dar a conocer todos los métodos, códigos, data y conflictos de interés antes y después de que la investigación sea realizada.

Siempre que el hacerlo no dañe seres humanos ni viole leyes (*e.g. anonimato informantes*).

Permite la sinergia entre investigadores, previniendo la duplicación de recolecciones y análisis de datos.

Permite discusión más transparente y continuar o expandir proyectos de investigación de terceros.

La creación de conocimiento puede ser más eficiente si los científicos trabajan en conjunto (Escuela pragmática). Problemas complejos requieren esfuerzos combinados.

**Posibles salidas:** Preregistro de investigaciones (*OSF*), publicación de códigos y data (*github y otros repositorios*), Investigación Reproducible.

# Dimensiones ciencia abierta

## 4. Código abierto (Open source)

Todos los programas, apps, algoritmos, herramientas y scripts deben ser transparentes y usables por otros.

Cuando un científico desarrolla una nueva tecnología, cualquier tecnología de otro puede interactuar con esta.

Cualquiera puede modificar la tecnología para adoptarlas mejor a sus propias necesidades

R y sus paquetes van en esta línea.

# Dimensiones ciencia abierta

## 5. Escuela Pública (Open academia)

Todos/as pueden participar de la academia. Las desigualdades el mundo social deben ser eliminadas del mundo académico.

Incluso, las cs. podrían tener el objetivo de eliminar las desigualdades del mundo social (e.g. sociología pública de ([Burawoy, 2005](#)):

- Intervenir en discusiones públicas o elaboración de políticas
- Apoyar la organización de movimientos sociales.

Debiese importar el argumento y la evidencia, no el poder o posición de los actores.

Desafío de la comunicación de los resultados de la investigación, ¿basta con liberar el *paper*?.

Líneas innovadoras de involucrar a la ciudadanía ([citizen science](#)).

# ¿Abrir la sociología?

Principal camino para evitar su fracaso como disciplina científica ([Brenzau, 2021](#)).

Permite resolver la actual crisis de legitimidad en la que se encuentran las ciencias sociales, la **poca confianza** desde el público y de quienes hacen las políticas públicas

Las pretensiones de la ciencia abierta están en las concepciones de ciencia de algunos de los principales referentes de la sociología, como [Merton \(1973\)](#).

- **Universalismo:** pretensiones de verdad, cualquiera sea la fuente, deben ser evaluadas con criterios impersonales preestablecidos.
- **Comunismo:** Los hallazgos sustantivos de la ciencia son producto de la colaboración social. La propiedad intelectual debe limitarse al reconocimiento y estima.
- **Desinterés:** altruismo en la búsqueda de la verdad. Rechazo a la acumulación de prestigio y a la subordinación a grupos de interés fuera del campo.
- **E scepticismo organizado:** escrutinio periódico de las creencias en términos lógicos y empíricos (no distinción entre lo sagrado y lo profano)

# Desafíos de la apertura

- Acción directa de recuperación de literatura es ilegal.
- La apertura nunca puede ser total: dilemas éticos y políticos, y restricciones legales.
- Aparente limitación a los enfoques cuantitativos de las ciencias sociales (*misconception*).
- ¿Es posible transparentar todo el proceso investigativo?, ¿El explicitar las decisiones no razonadas y las rationalidades pasadas a llevar no podría acrecentar el problema?
- Ayuda a la calidad, legitimidad e impacto de la sociología.

## El movimiento de la ciencia abierta

*"No debe ser un movimiento positivista. Es un movimiento para abrir la caja negra que rodea lo que hacen los sociólogos, sea lo que sea, para crear una comunidad de control de calidad y diálogo."* (Breznau, 2021).

# ¿Ciencia abierta en la sociología chilena?

Dificultad de dar una única respuesta

Multidimensionalidad del concepto de cs. abierta

Desarticulación de la triple vocación de la sociología ([Garretón, 2015](#)).

- la científica (academia, docencia)
- la crítica (militantes, opinión pública, intelectuales orgánicos)
- la profesional (Estado, empresas, ONGs)

Hiperespecialización en subcampos (educación, trabajo, salud, gestión cultural, etc.)

Dualismo cuantitativo / cualitativo

No hay investigación sobre el tema

# Cs. abierta en soc. chilena

## 1. Acceso

Mayor parte de las revistas con acceso abierto.

Investigadores e instituciones comparten artículos y documentos en plataformas

Con libros hay mayores restricciones. Vía biblioteca universidades se accede.

Resultados de instituciones públicas y privadas liberados.

## 2. Impacto / calidad

Academia neoliberal. Publica o perece.

La centralidad está en indexación de la revista.

Acumulación de producciones irrelevantes, poco concluyentes y repletas de matices.

Fenómeno empuja a publicar en el Norte Global (acceso limitado).

# Cs. abierta en soc. chilena

## 3. Apertura proceso investigación

Academia: caja negra del proceso (solo apartado metodológico) y lejos de ser reproducible

Limitado acceso a datos, incluso con financiamiento público

Excepción: [https://dataverse.harvard.edu/dataverse/coes\\_data\\_repository](https://dataverse.harvard.edu/dataverse/coes_data_repository)

Excusa del anonimato para no liberar datos.

Avance en transparencia desde el Estado. No a los procesos, pero sí a los datos.

Hoy en día el estándar es publicar las bases de datos de encuestas.

Consultoras/empresas no reguladas. Proceso y datos ocultos. Resultados resumidos públicos (ppt)

# Cs. abierta en soc. chilena

## 4. Apertura al mundo social, ¿investigar para quien?

Sociología crítica-intelectual la con mayor impacto

Irrelevancia e incomprensibilidad de la producción académica para los sujetos estudiados

Sin duda hay excepciones, sobre todo en estudios de caso.

Actividades de extensión separadas del proceso investigativo. Parte de la carta gantt para cumplir.

Mayor interés por producciones de organismos estatales y consultoras

- Intención de voto
- Niveles de empleo

# Investigación reproducible

Apertura de los procesos de investigación social

# Investigación reproducible

Que otra persona (o mi yo del futuro) pueda ejecutar mi código con los mismos datos llegando al mismo resultado.

		Data	
		Same	Different
Analysis	Same	Reproducible	Replicable
	Different	Robust	Generalisable

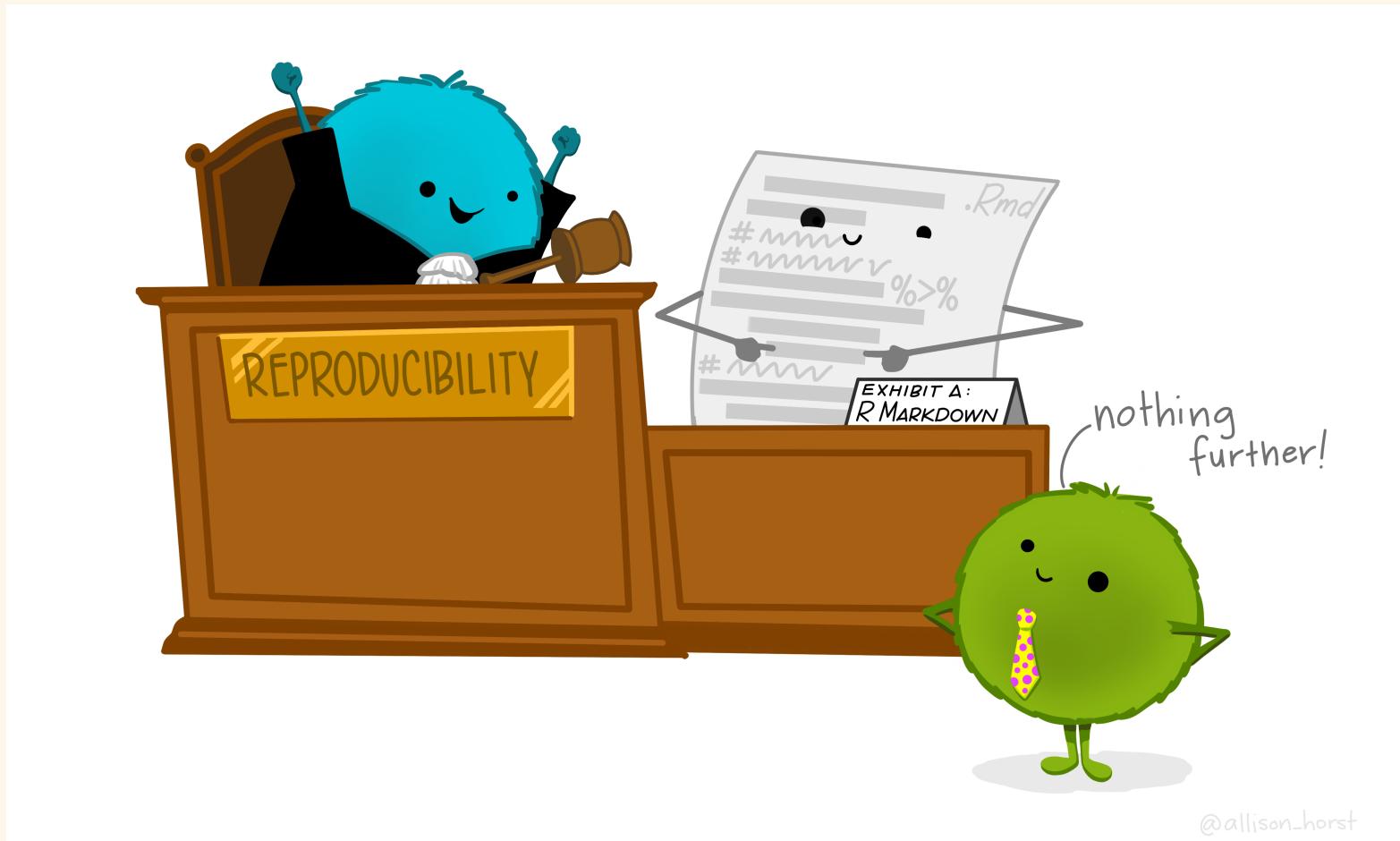
Fig. 3 How the Turing Way defines reproducible research

*“Aquí está todo lo que necesitas para reproducir mi trabajo”*

*“Lee el archivo README para las instrucciones”*

(Christensen, Freese, and Miguel, 2019).

# Investigación reproducible



@allison\_horst

# Investigación reproducible

¿Que cosas pueden dificultar la reproducción del análisis?

- No sé si estoy usando los mismos datos (*problemas de acceso, diferentes versiones*)
- Rutas a archivos que son incorrectas
- Aleatoriedad no reproducible
- Comentarios escuetos en el código (*no se entienden las decisiones*)
- Variables no localizadas (no estaban creadas en la rutina)
- Diferentes versiones del *software* o de los paquetes

# Investigación reproducible

## Posibles soluciones

- No sé si estoy usando los mismos datos

### ☞ Script, Control de versiones

- Rutas a archivos que son incorrectas

### ☞ RProject, RMarkdown

- Aleatoriedad no reproducible

### ☞ Fijar semillas con `set.seed()`

- Comentarios escuetos en el código

### ☞ Orden y estilo de código

- Variables no localizadas

### ☞ RMarkdown

- Diferentes versiones del *software* o de los paquetes

### ☞ `sessionInfo()` y paquetes

# ¿Por qué reproducibilidad?

- Para recordar como se hizo lo que se hizo (yo del futuro y supervisión)
- Facilita la detección de errores antes de publicar (si el código no corre es porque hay problemas)
- Permite el control de versiones (volver al pasado)
- Facilita la continuación del trabajo ante el recambio de personal
- Permite la colaboración y transparencia del proyecto
- Credibilidad de lo realizado
- Evita malos entendidos, confusiones frente a revisión o solicitudes innecesarias (cualquier duda consulta el código)
- Código puede reutilizarse o complejizarse para mejorar un procedimiento o una investigación

Estos y más argumentos en [video](#) de D. Ballari.

# ¿Por qué reproducibilidad?

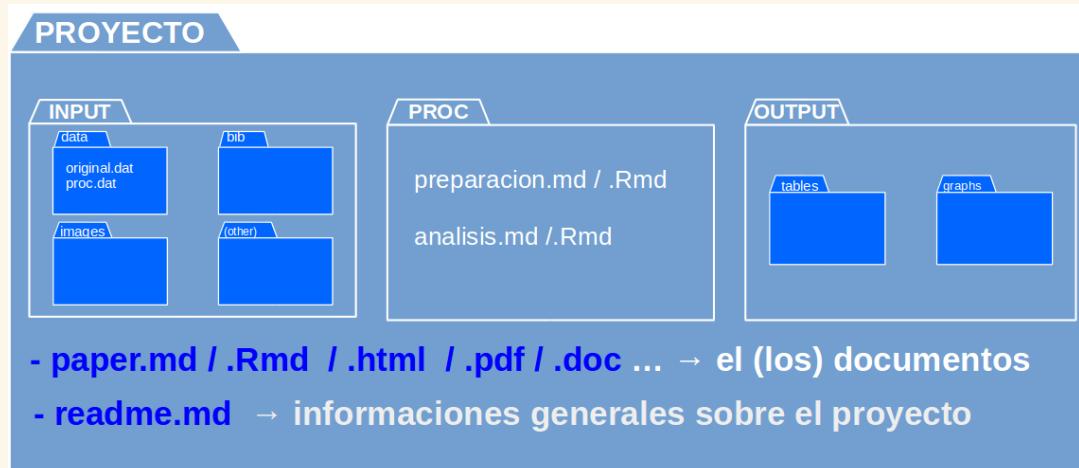
Argumentos ([Rodriguez-Sanchez, PÁrez-Luque, Bartomeus, and Varela, 2016](#)) para el mundo académico

## Beneficios de la ciencia reproducible para el investigador

- La utilización de código permite la automatización: ejecución de tareas repetitivas sin esfuerzo
- Muy fácil corregir y regenerar resultados, tablas y figuras
- Reducción drástica del riesgo de errores
- Los flujos de trabajo reproducibles facilitan la colaboración
- Mayor facilidad para escribir artículos al tener registro exhaustivo de todo el proceso de análisis
- La publicación del código ayuda a detectar errores antes de la publicación definitiva
- La publicación del código facilita el proceso de revisión
- La publicación del código facilita la comprensión del artículo y evita malinterpretaciones
- La reproducibilidad es un sello de calidad y aumenta la probabilidad de aceptación (cuando no es simplemente requerida)
- La reproducibilidad aumenta el impacto de las publicaciones (citas, reconocimiento, reutilización, coautorías)
- Ahorro de tiempo y esfuerzo al reutilizar código en otros proyectos

# ¿Cómo organizar un proyecto reproducible?

Una opción es el protocolo IPO



# RMarkdown

Articulación de código y texto para la reproducibilidad

# Breve repaso de RMarkdown



Articulación de lenguaje R y texto plano.

Los documentos creados con R Markdown son completamente reproducibles.

Sin importar el .RProject abierto. RMarkdown fija directorio de trabajo desde donde está el archivo .rmd

Si intentamos renderizar un documento y el proceso falla, no hay reproducción:

- Se requieren objetos (data frames, e.g.) que no han sido creados todavía
- Se intenta importar data desde rutas que no existen
- Problemas en la escritura de código (confusión entre mayusc y minusc, guiones bajos)
- Problemas en el orden del código (instrucción X antes de Y, cuando X requiere de Y)

# Breve repaso de RMarkdown

Si el código se reproduce, tenemos un indicador de la calidad de una investigación social (primer filtro).

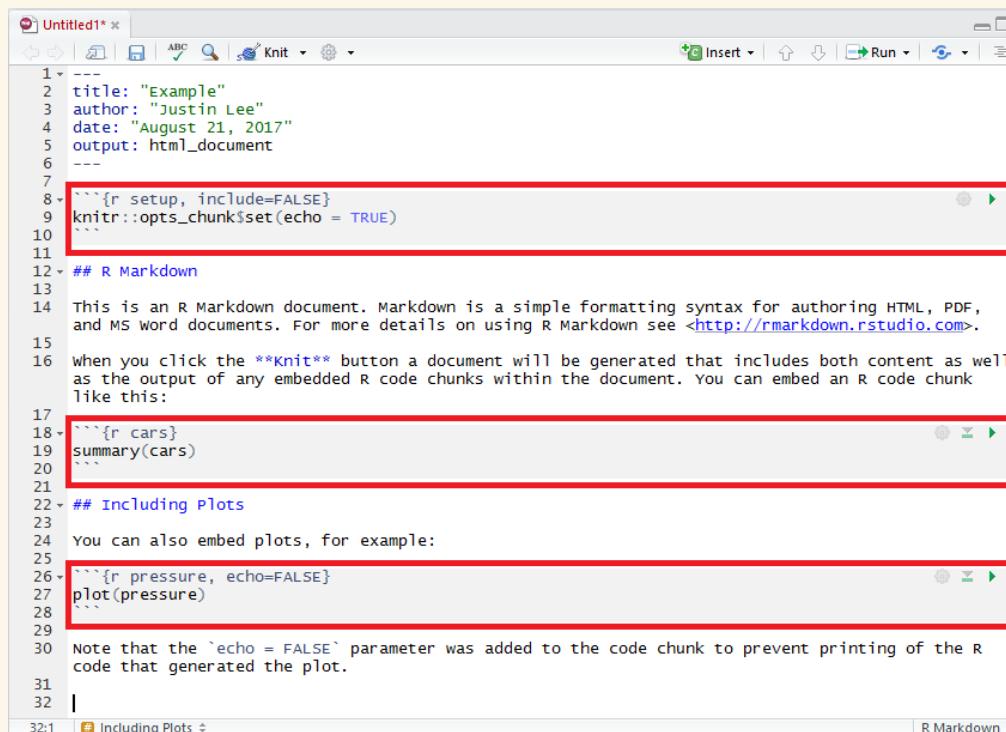
Otras cuestiones sustantivas (buenas ideas, revisión exhaustiva de antecedentes, validez de las mediciones, justificación de decisiones, gramática, etc.)

Los archivos de R Markdown en general tienen 3 partes:

(1). Un encabezado que permite configurar inicialmente el documento que vamos a escribir (`yaml`)

# Breve repaso de RMarkdown

(2). "Pedazos de código" (o en inglés, *chunk codes*, shortcut **ctrl+alt+i**).



The screenshot shows an RStudio interface with an untitled R Markdown file. The code is as follows:

```
1 ---  
2 title: "Example"  
3 author: "Justin Lee"  
4 date: "August 21, 2017"  
5 output: html_document  
6 ---  
7 ```{r setup, include=FALSE}  
knitr::opts_chunk$set(echo = TRUE)  
```  
8 ## R Markdown  
9  
10 This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF,  
11 and MS Word documents. For more details on using R Markdown see <http://rmarkdown.rstudio.com>.  
12  
13 when you click the **Knit** button a document will be generated that includes both content as well  
14 as the output of any embedded R code chunks within the document. You can embed an R code chunk  
15 like this:  
16  
17 ```{r cars}  
summary(cars)  
```  
18  
19 ## Including Plots  
20  
21 You can also embed plots, for example:  
22  
23 ```{r pressure, echo=FALSE}  
plot(pressure)  
```  
24  
25 Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R  
26 code that generated the plot.  
27  
28  
29  
30  
31  
32 |
```

The code blocks are highlighted with red boxes. The first code block (lines 7-10) is for setup. The second (lines 17-20) is for a summary of the 'cars' dataset. The third (lines 26-29) is for a plot of 'pressure'.

(3). Texto plano, donde escribimos como en cualquier otro procesador de texto.

# Rmd como editor de texto

RStudio lo podemos ocupar como simple editor de texto (libre).

A diferencia de word, no necesitamos licencia para escribir y somos dueños de nuestros textos.

¿Han abierto un documento word desde bloc de notas?



Toda una discusión al respecto en ([Castillo, 2021](#)). Ver acá

# Rmd como editor de texto

Para que nos sirva como editor de texto es fundamental que nos corrija la ortografía (*spell check*)

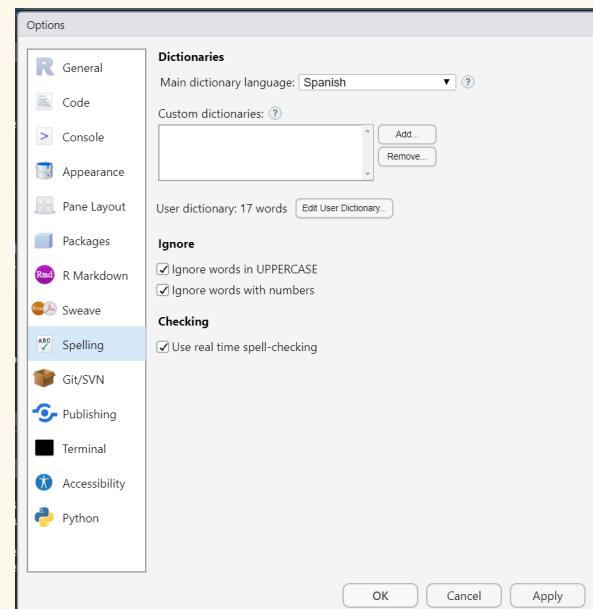
Tools -> Global Options -> Spelling

Seleccionar idioma español y dar aceptar

Si no está el español apretar "update dictionaries"

Funciona como word. Palabras mal escritas o que no reconoce son subrayadas.

Botón derecho sugiere alternativas o incluir palabra a diccionario.



Todo el detalle [acá](#).

# RMarkdown. Nuevas cosas.

Los chunks tienen muchos argumentos que permiten moldear el output que deseamos.

Existe un chunk llamado por defecto "setup" (general para todos los chunks del documento).

```
{r setup, include=FALSE}  
knitr::opts_chunk$set(echo = TRUE)
```

Podemos agregar opciones ([detalle acá](#)):

- include = FALSE/TRUE
- echo = FALSE/TRUE
- message = FALSE/TRUE
- warning = FALSE/TRUE
- error = FALSE/TRUE

O especificar opciones para cada chunk:

```
{r message=FALSE, warning=FALSE}  
library(dplyr)
```

# RMarkdown. Nuevas cosas.

Podemos crear títulos (#), énfasis con **negrita** o *cursiva*, listar de elementos, etc:

```
# titulo 1  
  
# titulo 2  
  
## titulo de segundo nivel  
  
+ elemento listado 1  
  
+ elemento listado 2  
  
![ ](link.png)  
  
[palabra](link.com)
```

Imagenes con mayor control:

```
{r echo=FALSE, fig.align='center', out.width = "65%"}  
knitr:::include_graphics("link.png")
```

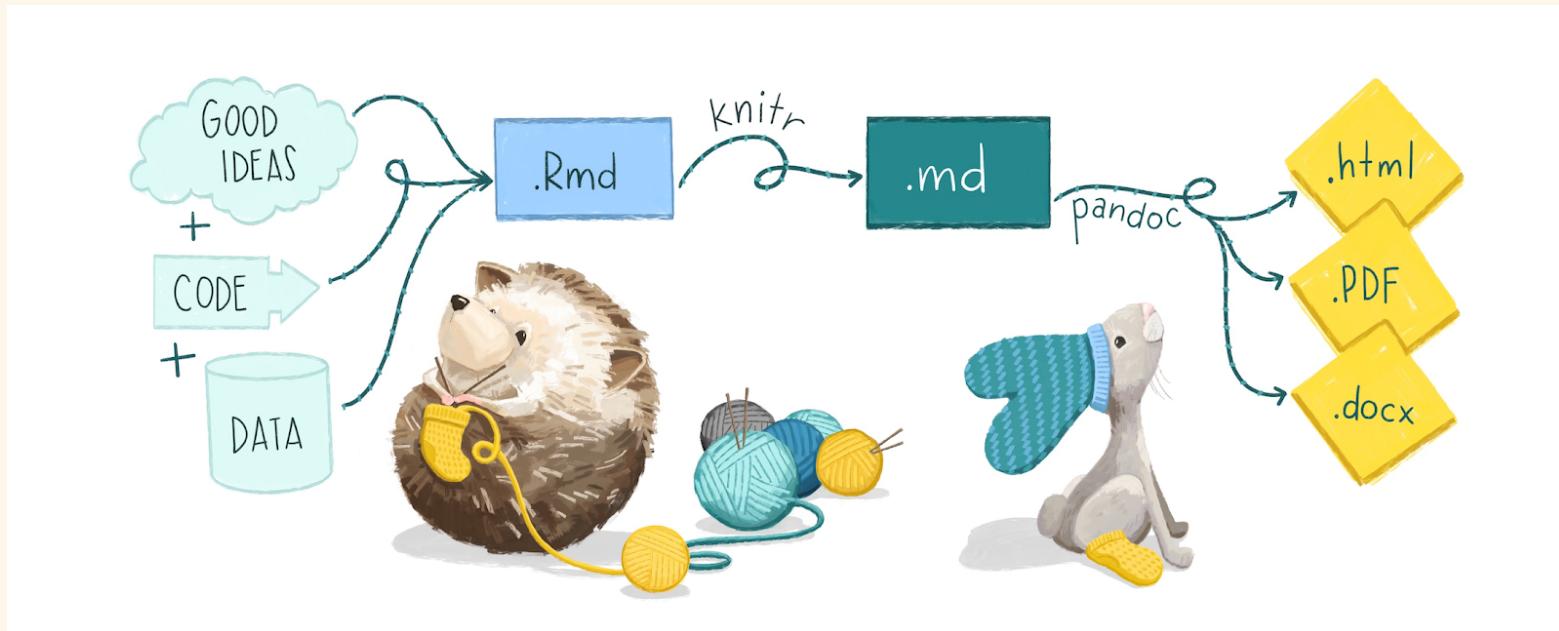
# RMarkdown. Nuevas cosas.

Para generar el documento se puede utilizar el comando `render` o bien utilizar el botón Knit (Tejer).



O podemos ir ejecutando líneas de código en específico, chunks particulares o Run All, lo que nos permite detectar errores cuando el código no corre.

# Exportar documento



```
---
```

```
title: "Título"
author: "Autor"
output:
  html_document:
---
```

```
---
```

```
title: "Título"
author: "Autor"
output:
  pdf_document:
---
```

# Documentos en formato html

Descargar [formato html](#) estilizado.

# Documentos en formato pdf

## Formato tareas UCSH

[Descargar formato pdf](#) y proyecto que lo genera.

Para utilizarlo se requiere instalar algún distribuidor de latex:  
`install.packages('tinytex')` ([más información acá](#))

## Formatos de tesis

Plantilla para hacer tesis: [Universidad de Chile](#)

Incluso hay formatos para hacer Currículums

Acá dos paquetes con varias plantillas o templates (hay muchos más):

```
install.packages("stevetemplates")
install.packages("vitae")
```

# Documentos formato word

Descargar [formato word](#). No utilizar en este curso.

# Para recordar y no fallar

- La ruta del RProject abierto no es considerada. Se utiliza la ruta del archivo . rmd
- Las librerías que necesiten deben estar cargadas en el script de Rmarkdown (**error habitual**).
- Tanto los microdatos como las variables a utilizar deben estar llamadas dentro el script de Rmarkdown.
- Knitr() incluye la impresión de los resultados, por lo que no es necesario usar print().
- No usar View() en RMarkdown, porque en el output no mostrará nada.
- Cada chunk puede tener o no tener un nombre. Si los deciden nombrar, los nombres NO DEBEN REPETIRSE (otra fuente de error)

# Código en línea

Existe el código R "en línea" (*in line*), lo que nos permite hacer reportes automáticos:

El día de hoy es 2021-11-11 y el valor de 5 por 2 es 50

```
El día de hoy es `r Sys.Date()` y el valor de 5 por 2 es `r 5*10`
```

# Dar formato a las tablas

Función `kable()` y paquete `kableExtra()`.

```
Orange %>% head(4)
```

```
##   Tree  age circumference
## 1    1 118              30
## 2    1 484              58
## 3    1 664              87
## 4    1 1004             115
```

```
library(knitr)
Orange %>% head(3) %>% kable()
```

| Tree | age | circumference |
|------|-----|---------------|
| 1    | 118 | 30            |
| 1    | 484 | 58            |
| 1    | 664 | 87            |

# Dar formato a las tablas

```
library(kableExtra)
Orange %>% head(3) %>%
  kable(caption = "Título de tabla") %>%
  kable_styling(bootstrap_options=c("striped", "hover", "condensed", "responsive"))
```

| Título de tabla |     |               |
|-----------------|-----|---------------|
| Tree            | age | circumference |
| 1               | 118 | 30            |
| 1               | 484 | 58            |
| 1               | 664 | 87            |

Para más información sobre distintos estilos revisar [Create Awesome HTML Table with knitr::kable and kableExtra](#)

# Herramientas de gestión bibliográfica

Presentación de Zotero y Bibtex, 12 de noviembre solo si alcanzamos (de lo contrario 19 de noviembre)

Contenidos:

- Buscar bibliografía
- Acceder a bibliografías (romper barreras de pago)
- Gestionar eficientemente las bibliografías (Zotero)
- Archivos bibtex
- Citar bibliografías desde RMarkdown

# Material de texto y audiovisual de interés

Curso "Ciencia Social Abierta" de Juan Carlos Castillo (en español).

[¿Por qué es importante la reproducibilidad computacional?](#) De Daniela Ballari. En español

[Como escribir manuscritos reproducibles.](#) De Francisco Rodríguez-Sánchez. En español

Documental [Paradojas del nihilismo, La academia.](#)

# Bibliografía

- Breznau, N. (2021). "Does Sociology Need Open Science?" In: *Societies* 11.1, p. 9.
- Burawoy, M. (2005). "Por Una Sociología Pública". In: *Política y sociedad* 42.1, pp. 197-225.
- Castillo, J. C. (2021). "Capítulo 1 Lenguaje Reproducible". In: *Investigación Social Abierta*.
- Christensen, G., J. Freese, and E. Miguel (2019). *Transparent and Reproducible Social Science Research: How to Do Open Science*. University of California Press.
- Fecher, B. and S. Friesike (2014). "Open Science: One Term, Five Schools of Thought". In: *Opening science*, pp. 17-47.
- Garretón, M. A. (2015). "La Recomposición de La Triple Vocación de La Ciencia Social En América Latina". In: *Polis. Revista Latinoamericana*.
- Rodríguez-Sánchez, F., A. J. PÁREZ-LUQUE, I. Bartomeus, et al. (2016). "Ciencia Reproducible: Qué, Por Qué, Cómo". In: *Ecosistemas* 25.2, pp. 83-92.
- Swartz, A. (2008). "Guerilla Open Access Manifesto". In: *Aaron Swartz [Internet]*.