

# Sesión 3

## Curso: R Aplicado a los Proyectos de Investigación

---

**Percy Soto-Becerra, M.D., M.Sc(c)**

**InkaStats Data Science Solutions | Medical Branch**

**2022-10-05**

 <https://github.com/psotob91>



# Agenda

1. **Importación de datos**
2. Más verbos de dplyr para manejo de datos
3. Otros verbos útiles para manejo de datos
4. Uso de helpers
5. Análisis Exploratorio de Datos versus Análisis Inicial de Datos
6. Pasos para un buen AID / AED

# Importar datos de fuentes externas a R

---

El paquete `{rio}` es la **navaja suiza** de la importación de datos en R.



Su función **import** permite importar prácticamente cualquier formato.

# Importando datos con {rio}

Instalar {rio}

import

Plano

MS Excel

Stata (\*.dta)

SPSS (\*.sav)

Otros

- {rio} debe instalarse de una manera especial
- Pasos:
  1. Primero se instala como siempre, usando `install.packages`
  2. Adicionalmente, se sugiere correr `install_formats`. inmediatamente después de la primera instalación.
  3. Se carga el paquete usando `library`.

```
1 install.packages("rio")
2 rio::install_formats()
3 library(rio)
```

- Más detalles sobre el paquete y sus funciones pueden encontrarse en la página web del paquete: <https://thomasleeper.com/rio/index.html>

# ¿Cómo importar metadatos de Stata o SPSS?

¿Metadatos?

Stata (\*.dta)

SPSS (\*.sav)

- Son los datos de los datos.
- En bases de datos para análisis estadístico, dos metadatos bastante usados son:
  - Etiquetas de la variable
  - Etiquetas de los valores de la variable

**STATA**

Editor de Datos (Edición) - [Sin\_título]

Archivo Edición Ver Datos Herramientas

sexo[1] 0

	peso	sexo
1	56.07676	Femenino
2	53.12789	Femenino
3	50.89314	Masculino
4	44.83361	Masculino
5	47.8124	Masculino
6	52.58548	Femenino
7	49.28194	Masculino
8	51.54965	Femenino
9	44.56395	Masculino
10	46.95463	Masculino
11	53.2861	Femenino
12	53.99795	Masculino

*Handwritten notes:*

- Valor | Etiqueta
- 0 | Femenino
- 1 | Masculino
- Etiqueta de la variable (pointing to 'Femenino' in the first row)
- Etiqueta de valor (pointing to 'Femenino' in the first row)
- valor de la variable (pointing to the '0' in the first row)

**Variables**

Nombre	Etiqueta	Tipo	Formato	Etiquetas de val.
peso	Peso, kg	float	%9.0g	
sexo	Sexo	float	%9.0g	sexo

*Handwritten note:* Etiqueta de la variable (pointing to 'Sexo' in the table)

**Propiedades**

Propiedad	Valor
Nombre	sexo
Etiqueta	Sexo
Tipo	float
Formato	%9.0g
Etiqueta de valor	sexo

# Exportando datos con {rio}

export()

Ejemplos:

- Se usa `export()` para exportar un objeto `data frame` o `data tibble` a otro formato:

```
1 export(datos_para_importa, file = "datos_exportados.fmt")
```

- Los formatos pueden ser:

Formato	Extensión típica
Comma-separated data	.csv
Pipe-separated data	.psv
Tab-separated data	.tsv
CSVY (CSV + YAML metadata header)	.csvy
SAS	.sas7bdat

Formato	Extensión típica
SPSS	.sav
SPSS (compressed)	.zsav
Stata	.dta
SAS XPORT	.xpt
SPSS Portable	.por
Excel	.xls
Excel	.xlsx
R syntax	.R
Saved R objects	.RData, .rda
Serialized R objects	.rds
Epiinfo	.rec

# Nuestro turno

---

- Descargue la carpeta denominada **taller03** disponible en la carpeta compartida.
- Abra el proyecto denominado **taller03.Rproj**
- Complete y ejecute el código faltante en los chunk de código de la PRIMERA PARTE.
- Una vez culmine todo el proceso, renderice el archivo .qmd.

10:00



 Hagamos una pausa

---

Tomemos un descanso de 5 minutos...

Estire las piernas ...

Deje de ver las pantallas ...

... cualquier , las del celular  también

05:00

# Agenda

1. Importación de datos
2. **Más verbos de dplyr para manejo de datos**
3. Otros verbos útiles para manejo de datos
4. Uso de helpers
5. Análisis Exploratorio de Datos versus Análisis Inicial de Datos
6. Pasos para un buen AID / AED

# Creación de nuevas columnas con `mutate()`

- La función `mutate()` crea columnas nuevas o modifica existentes.

## Larga

```
1 mutate(.data = DATA, ...)
```

## Abreviada

```
1 mutate(DATA, ...)
```

## Se estila usar `%>%`

```
1 DATA %>%
2   mutate(...)
```

Argumento	Descripción
<code>.data</code>	Data frame o extensión de data frame (por ejemplo, tibble).
<code>...</code>	Columnas existente para modificar o columnas nuevas para crear.

# mutate() en acción

Crear una nueva variable

Crear varias nuevas variables

Reemplazar variable existente

- Si queremos crear la variable índice de masa corporal:

```
1 datos_fase1 %>%
2   select(id, weight, height) %>% # Nos quedamos con peso y talla
3   mutate(imc = weight / height ^ 2) # Creamos IMC en base a peso y talla
```

# A tibble: 106 × 4

	id	weight	height	imc
	<dbl>	<dbl>	<dbl>	<dbl>
1	1	59	1.4	30.1
2	1	59.9	1.3	35.4
3	2	62	1.5	27.6
4	2	62.1	1.6	24.3
5	3	62	1.6	24.2
6	3	60	1.6	23.4
7	4	60.9	1.5	27.1
8	4	61.4	1.5	27.3
9	5	64	1.5	28.4
10	5	58.1	1.6	22.7

# ... with 96 more rows

# Agenda

1. Importación de datos
2. Más verbos de dplyr para manejo de datos
3. **Otros verbos útiles para manejo de datos**
4. Uso de helpers
5. Análisis Exploratorio de Datos versus Análisis Inicial de Datos
6. Pasos para un buen AID / AED

# Renombrar variables con rename()

- Cambia los nombres de variables individuales.

## Larga

```
1 rename(.data = DATA, ...)
```

## Abreviada

```
1 rename(DATA, ...)
```

## Se estila usar %>%

```
1 DATA %>%
2   rename(...)
```

Argumento	Descripción
.data	Data frame o extensión de data frame (por ejemplo, tibble).
...	nuevo_nombre = viejo_nombre

# rename() en acción

- Cambiar el nombre de **married** por **casado1** y el de **age** por **edad**.

```
1 datos_fase1 %>%
2   rename(
3     casado1 = married,
4     edad = age
5   )
```

# A tibble: 106 × 14

	id	time	treat	edad	race	casado1	marri... <sup>1</sup>	proce... <sup>2</sup>	weight	height	e2
	<dbl>	<fct>	<fct>	<dbl>	<chr>	<fct>	<fct>	<chr>	<dbl>	<dbl>	<dbl>
1	1	Baseline	Place...	33	Mest...	Single	Withou...	Callao	59	1.4	87.3
2	1	3 months	Place...	32	Mest...	Single	Withou...	Callao	59.9	1.3	210.
3	2	Baseline	Dosis...	27	Mest...	Single	Withou...	Santa ...	62	1.5	169.
4	2	3 months	Dosis...	27	Mest...	Single	Withou...	Santa ...	62.1	1.6	99.9
5	3	Baseline	Dosis...	25	Mest...	Single	Withou...	Callao	62	1.6	78.8
6	3	3 months	Dosis...	25	Mest...	Single	Withou...	Callao	60	1.6	155.
7	4	Baseline	Dosis...	37	Mest...	Divorc...	Withou...	Callao	60.9	1.5	41.0
8	4	3 months	Dosis...	38	Mest...	Divorc...	Withou...	Callao	61.4	1.5	109.
9	5	Baseline	Place...	31	Mest...	Single	Withou...	La Mol...	64	1.5	43.0
10	5	3 months	Place...	32	Mest...	Single	Withou...	La Mol...	58.1	1.6	56.0

# ... with 96 more rows, 3 more variables: lh <dbl>, fsh <dbl>, prog <dbl>, and  
# abbreviated variable names <sup>1</sup>married2, <sup>2</sup>procedence

# Recodificar valores de variables con recode()

- Cambia los nombres de las etiquetas de valores de las variables.

## Larga

```
1 recode(.data = DATA, ...)
```

## Abreviada

```
1 recode(DATA, ...)
```

## Se estila usar %>%

```
1 DATA %>%
2   recode(...)
```

recode es el raro del grupo: primeo viejo nombre y luego nuevo nombre

Argumento	Descripción
.data	Data frame o extensión de data frame (por ejemplo, tibble).
...	viejo_nombre = nuevo_nombre



# recode() en acción

- Cambiar los valores de married2 del inglés al español:

```
1 datos_fase1 %>%
2   mutate(married2 = recode(married2,
3     "Without couple" = "Sin pareja",
4     "With couple" = "Con pareja"))
```

# A tibble: 106 × 14

	id	time	treat	age	race	married	marri... <sup>1</sup>	proce... <sup>2</sup>	weight	height	e2
	<dbl>	<fct>	<fct>	<dbl>	<chr>	<fct>	<fct>	<chr>	<dbl>	<dbl>	<dbl>
1	1	Baseline	Place...	33	Mest...	Single	Sin pa...	Callao	59	1.4	87.3
2	1	3 months	Place...	32	Mest...	Single	Sin pa...	Callao	59.9	1.3	210.
3	2	Baseline	Dosis...	27	Mest...	Single	Sin pa...	Santa ...	62	1.5	169.
4	2	3 months	Dosis...	27	Mest...	Single	Sin pa...	Santa ...	62.1	1.6	99.9
5	3	Baseline	Dosis...	25	Mest...	Single	Sin pa...	Callao	62	1.6	78.8
6	3	3 months	Dosis...	25	Mest...	Single	Sin pa...	Callao	60	1.6	155.
7	4	Baseline	Dosis...	37	Mest...	Divorc...	Sin pa...	Callao	60.9	1.5	41.0
8	4	3 months	Dosis...	38	Mest...	Divorc...	Sin pa...	Callao	61.4	1.5	109.
9	5	Baseline	Place...	31	Mest...	Single	Sin pa...	La Mol...	64	1.5	43.0
10	5	3 months	Place...	32	Mest...	Single	Sin pa...	La Mol...	58.1	1.6	56.0

# ... with 96 more rows, 3 more variables: lh <dbl>, fsh <dbl>, prog <dbl>, and  
# abbreviated variable names <sup>1</sup>married2, <sup>2</sup>procedence

# Nuestro turno

---

- Descargue la carpeta denominada **taller03** disponible en la carpeta compartida.
- Abra el proyecto denominado **taller03.Rproj**
- Complete y ejecute el código faltante en los chunk de código de la SEGUNDA PARTE.
- Una vez culmine todo el proceso, renderice el archivo .qmd.

10:00

# Agenda

1. Importación de datos
2. Más verbos de dplyr para manejo de datos
3. Otros verbos útiles para manejo de datos
- 4. Uso de helpers**
5. Análisis Exploratorio de Datos versus Análisis Inicial de Datos
6. Pasos para un buen AID / AED

# Selectores tidy

- Son funciones que ayudan a especificar un grupo específico de columnas.
- Ejemplos típicos de funciones tidyselect son:
  - `starts_with()`
  - `end_with()`
  - `contains()`
  - `matches()`
  - `last_col()`
  - `num_range()`
  - `where()`
- Más información sobre tidyselect: <https://dplyr.tidyverse.org/reference/select.html>

# Nuestro turno

---

- Descargue la carpeta denominada **taller03** disponible en la carpeta compartida.
- Abra el proyecto denominado **taller03.Rproj**
- Complete y ejecute el código faltante en los chunk de código de la TERCERA PARTE.
- Una vez culmine todo el proceso, renderice el archivo .qmd.

10:00

# Agenda

1. Importación de datos
2. Más verbos de dplyr para manejo de datos
3. Otros verbos útiles para manejo de datos
4. Uso de helpers
- 5. Análisis Exploratorio de Datos versus Análisis Inicial de Datos**
6. Pasos para un buen AID / AED

El análisis inicial de datos  
y el análisis exploratorio  
de datos son dos cosas  
diferentes!!

# Análisis Exploratorio de Datos

---



## Análisis Exploratorio de Datos

- El AED es un enfoque de análisis de conjunto de datos para identificar patrones y formular nuevas hipótesis.
- Se trata de ver qué nos dice los datos más allá de ideas pre-concebidas.
- Las nuevas hipótesis luego se confirman en otros nuevos estudios rigurosos.
- Su versión moderna: Minería de Datos (*Data Mining*)

## Análisis Inicial de Datos

- El AID, a menudo, se confunde erróneamente con el AED:
  - Ambos son dos enfoques totalmente diferentes que comparten herramientas comunes.
- Objetivo del AID:

*“(...) garantizar principalmente la transparencia y la integridad de las condiciones previas para realizar análisis estadísticos apropiados de manera responsable para responder preguntas de investigación predefinidas.”*

*Baillie M, et al. [PLoS Comput Biol, 2022]  
(<https://doi.org/10.1371/journal.pcbi.1009819>)*

# Análisis Inicial de Datos vs. Análisis Exploratorio de Datos

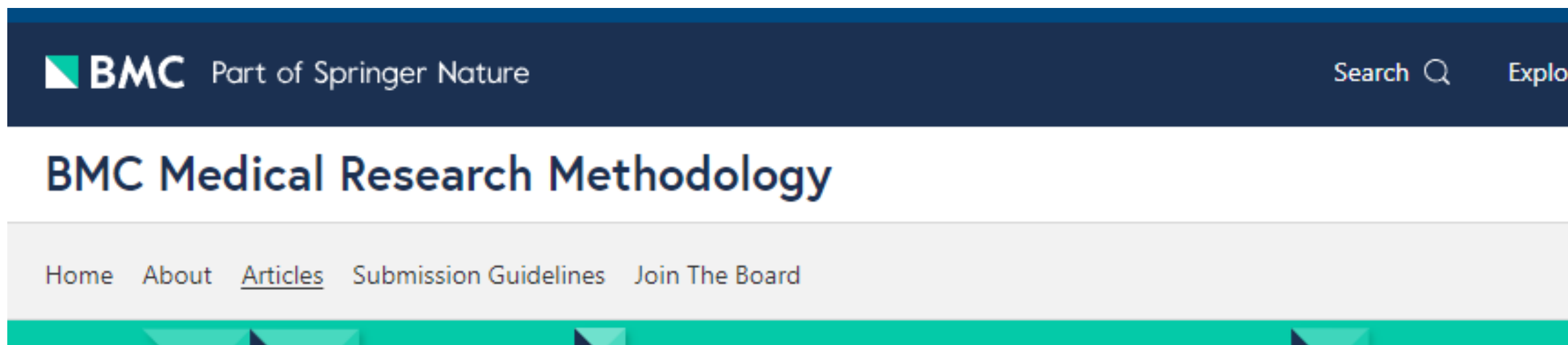
AID	AED
1. AID es el paso inicial del proceso de confirmación de hipótesis pre-definidas.	1. AED busca generar hipótesis nuevas.
2. En investigación clínico-epidemiológica, a menudo queremos y deberíamos hacer AID.	2. Rara vez queremos AED (p. ej., enfermedades nuevas, fenómenos muy poco conocidos)
3. Proceso clave para garantizar responder adecuadamente objetivos pre-planeados de estudio.	3. Proceso con alto riesgo de contaminar respuesta a objetivos pre-planeados de estudio.

# Mala práctica: ¡Hacer AED en vez de AID!

---

- A menudo, investigadores no realizan AID de manera sistemática.
- Mezclan actividades de AID con tareas posteriores de análisis de datos, como generación o exploración de hipótesis, análisis formal e interpretación de conclusiones.
- Como se hacen “informalmente”, no se reportan en detalle generándose análisis ocultos.
- Estos análisis ocultos generan problemas en la reproducibilidad de los estudios.
- Generan muchos grados de libertad adicionales ocasionando problemas serios de validez de los análisis: *p-hacking*, *post-selection inference*, *double-dipping*, *overfitting*, etc.
- Iniciativa STRATOS ha dado pautas para realizar AID apropiados.

# El problema de los análisis ocultos



Research article | [Open Access](#) | [Published: 13 March 2020](#)

## Hidden analyses: a review of reporting practice and recommendations for more transparent reporting of initial data analyses

[Marianne Huebner](#) , [Werner Vach](#), [Saskia le Cessie](#), [Carsten Oliver Schmidt](#) & [Lara Lusa](#) [on behalf of the Topic Group "Initial Data Analysis" of the STRATOS Initiative \(STRengthening Analytical Thinking for Observational Studies, <http://www.stratos-initiative.org>\)](#)


[BMC Medical Research Methodology](#) **20**, Article number: 61 (2020) | [Cite this article](#)

**1613** Accesses | **3** Citations | **6** Altmetric | [Metrics](#)

*BMC Med Res Methodol* 20, 61 (2020)



# Recomendaciones de STRATOS para hacer un buen AID

## PLOS COMPUTATIONAL BIOLOGY


 OPEN ACCESS

EDITORIAL

### Ten simple rules for initial data analysis

Mark Baillie, Saskia le Cessie, Carsten Oliver Schmidt, Lara Lusa, Marianne Huebner ,  
for the Topic Group “Initial Data Analysis” of the STRATOS Initiative 

Published: February 24, 2022 • <https://doi.org/10.1371/journal.pcbi.1009819>

Article	Authors	Metrics	Comments	Media Coverage
				
<a href="#">Introduction</a> <a href="#">Conclusions</a> <a href="#">Acknowledgments</a> <a href="#">References</a>	<p><b>Citation:</b> Baillie M, le Cessie S, Schmidt CO, Lusa L, Huebner M, for the Topic Group “Initial Data Analysis” of the STRATOS Initiative (2022) Ten simple rules for initial data analysis. PLoS Comput Biol 18(2): e1009819. <a href="https://doi.org/10.1371/journal.pcbi.1009819">https://doi.org/10.1371/journal.pcbi.1009819</a></p>			

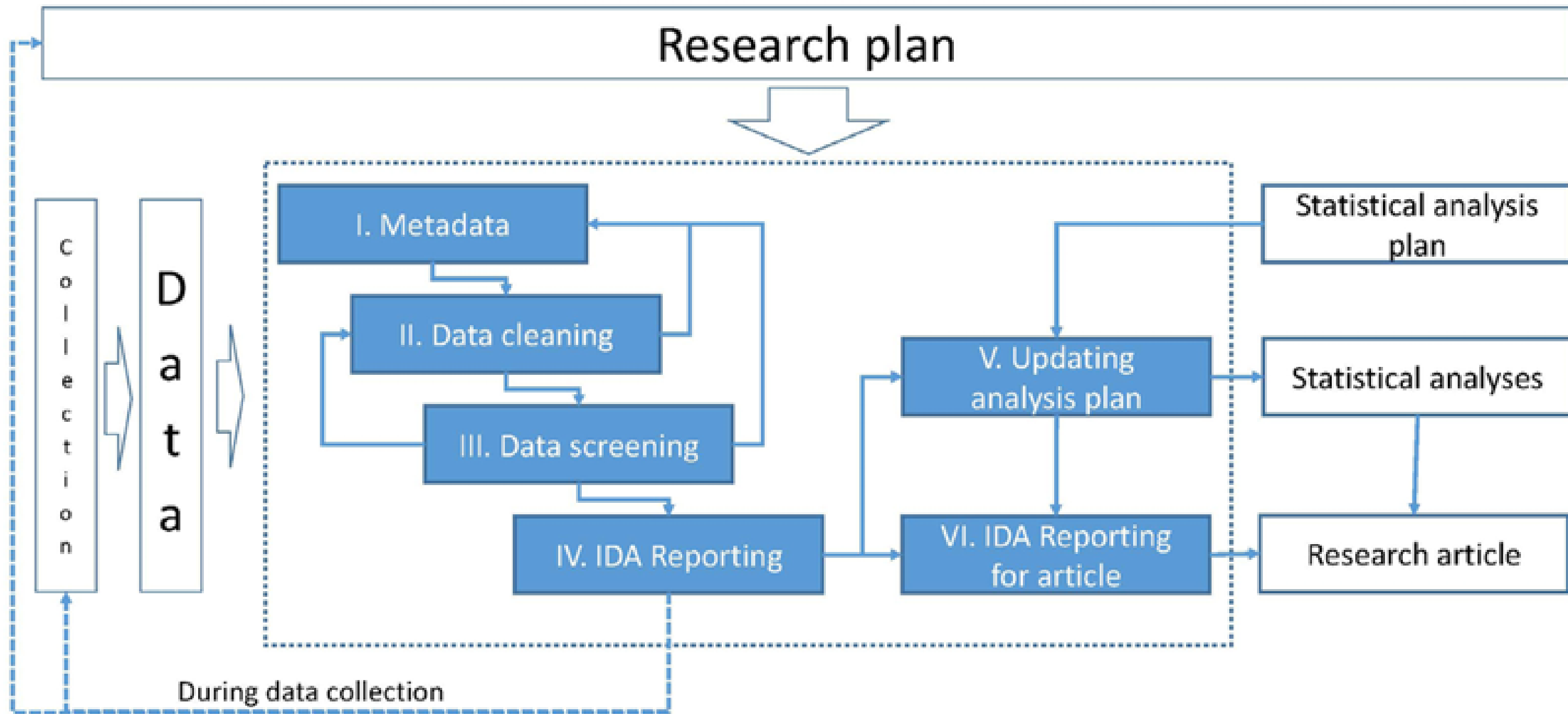
*PLoS Comput Biol* 18(2): e1009819

# AID es un proceso iterativo

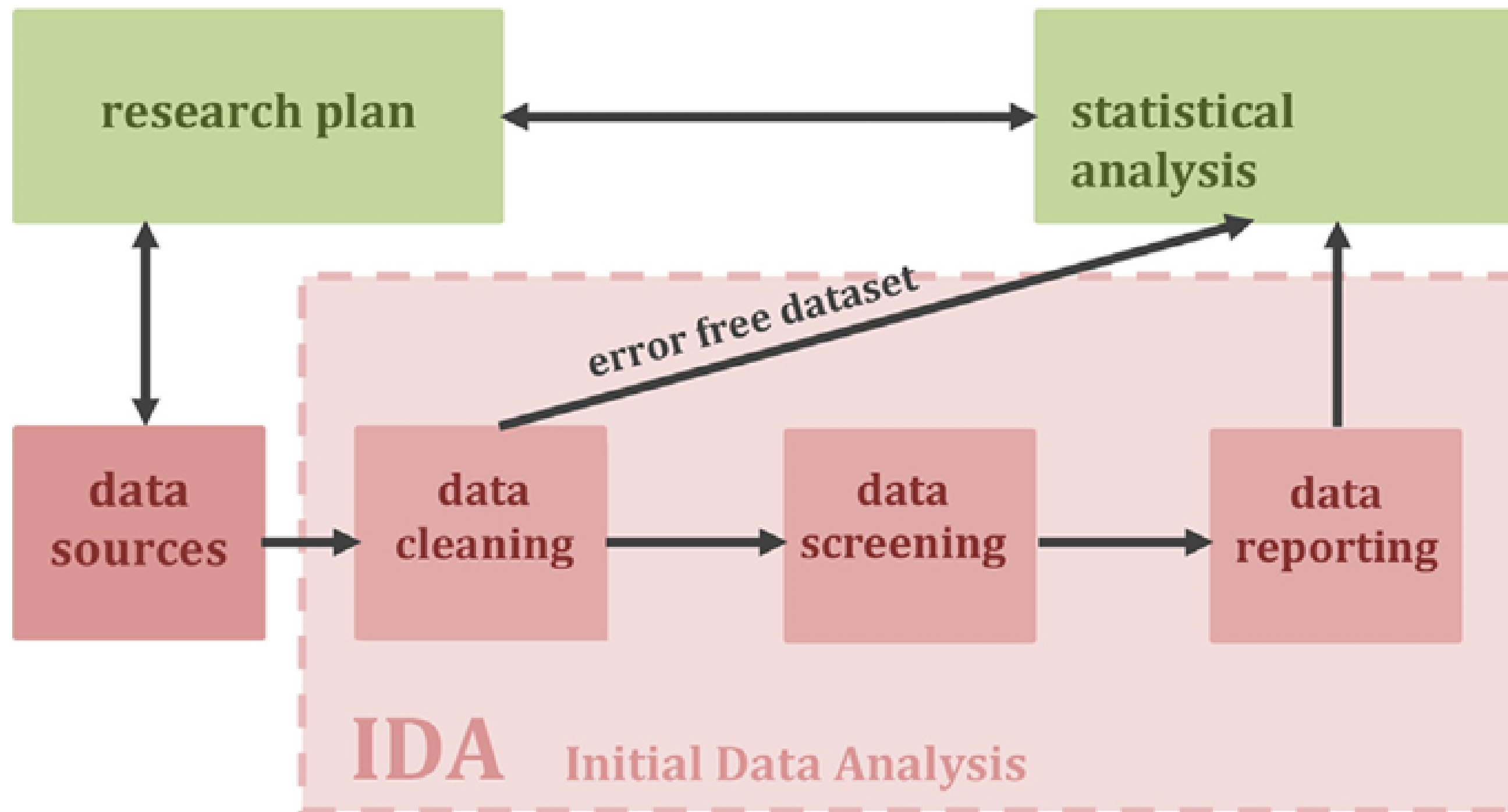
---

- Proceso no lineal, al contrario, requiere muchas iteraciones.
- Riesgo: Puede influir en análisis e inducir conclusiones erróneas.
  - Mayor riesgo de resultados falsos positivos.
- Ser cuidados para:
  - Evitar alterar la pregunta de investigación.
  - Proveer documentación completa del proceso.

# AID como parte del plan de investigación



# AID como parte del plan de investigación





# 10 recomendaciones de STRATOS para un buen AID

---

- **Regla 1:** Desarrolle un plan de AID que respalde el objetivo de la investigación.
- **Regla 2:** AID toma tiempo y recursos.
- **Regla 3:** AID debe ser reproducible.
- **Regla 4:** El contexto importa, conoce tus datos.
- **Regla 5:** Evite los adelantos, AID no toca la pregunta de investigación.
- **Regla 6:** Visualiza tus datos.
- **Regla 7:** Compruebe lo que falte.
- **Regla 8:** Comunicar los hallazgos y considerar las consecuencias.
- **Regla 9:** Reporte los hallazgos del AID en trabajos de investigación (¡adjunte anexos!)
- **Regla 10:** Sea proactivo y riguroso.

# 10 recomendaciones de STRATOS para un buen AID

---

- **Regla 1:** Desarrolle un plan de AID que respalde el objetivo de la investigación → *¡Escríballo en el proyecto o anexe un plan de análisis estadístico detallado!*
- **Regla 2:** AID toma tiempo y recursos. → *Presupueste RRHH y tiempo razonable*
- **Regla 3:** AID debe ser reproducible. → *Use programas que generen código*
- **Regla 4:** El contexto importa, conoce tus datos. → *{dplyr} en R*
- **Regla 5:** Evite los adelantos, AID no toca la pregunta de investigación. → *¡No haga 'análisis preliminar' sin antes inspeccionar y limpiar bien los datos!*
- **Regla 6:** Visualiza tus datos. → *{ggplot2} en R*
- **Regla 7:** Compruebe lo que falte. → *{tidyverse} para queries en R*
- **Regla 8:** Comunicar los hallazgos y considerar las consecuencias. → *Quarto para programación literaria en R*
- **Regla 9:** Reporte los hallazgos del AID en trabajos de investigación (¡adjunte anexos!). → *Ídem*
- **Regla 10:** Sea proactivo y riguroso. → *¡Los datos son como sus pacientes, use las mejores técnicas y herramientas disponibles!*

# Regla 4: El contexto importa, conoce tus datos

---

- Dé una primera mirada global a los datos
- Diseñe una lista de validaciones a realizar desde el proyecto.
- Valida tus datos:
  - Identifique duplicados y detecte inconsistencias
  - Valores extremos no plausibles
  - Identifique valores perdidos
- En R, use los verbos básicos de {dplyr} para hacer consultas (“queries”) a sus datos: `filter()`, `select()`, `mutate()`, `arrange()` y `summarise()`.

# Agenda

1. Importación de datos
2. Más verbos de dplyr para manejo de datos
3. Otros verbos útiles para manejo de datos
4. Uso de helpers
5. Análisis Exploratorio de Datos versus Análisis Inicial de Datos
6. **Pasos para un buen AID / AED**

## Paso 1: Resumen global de los datos

¿Qué debo inspeccionar de manera global?

`glimpse()`   `skim()`   `describe()`

- Dimensiones: columnas y filas
- Variables y tipos
- Datos completos y faltantes
- Variables numéricas: Mínimos, máximos y valores extremos
- Variables categóricas: Valores o categorías muy poco frecuentes y datos perdidos encubiertos

# Nuestro turno

---

- Descargue la carpeta denominada **taller03** disponible en la carpeta compartida.
- Abra el proyecto denominado **taller03.Rproj**
- Complete y ejecute el código faltante en los chunk de código de la CUARTA PARTE.
- Una vez culmine todo el proceso, renderice el archivo .qmd.

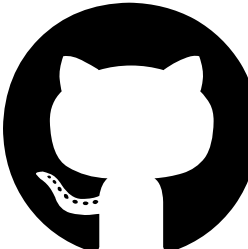
10:00

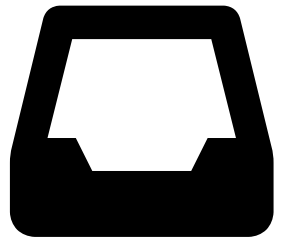
# ¡Gracias!

# ¿Preguntas?



 @psotob91

 <https://github.com/psotob91>

 percys1991@gmail.com