

Usando R para la Ciencia de Datos

Mayo 2018

Edgar Ruiz



@theotheredgar

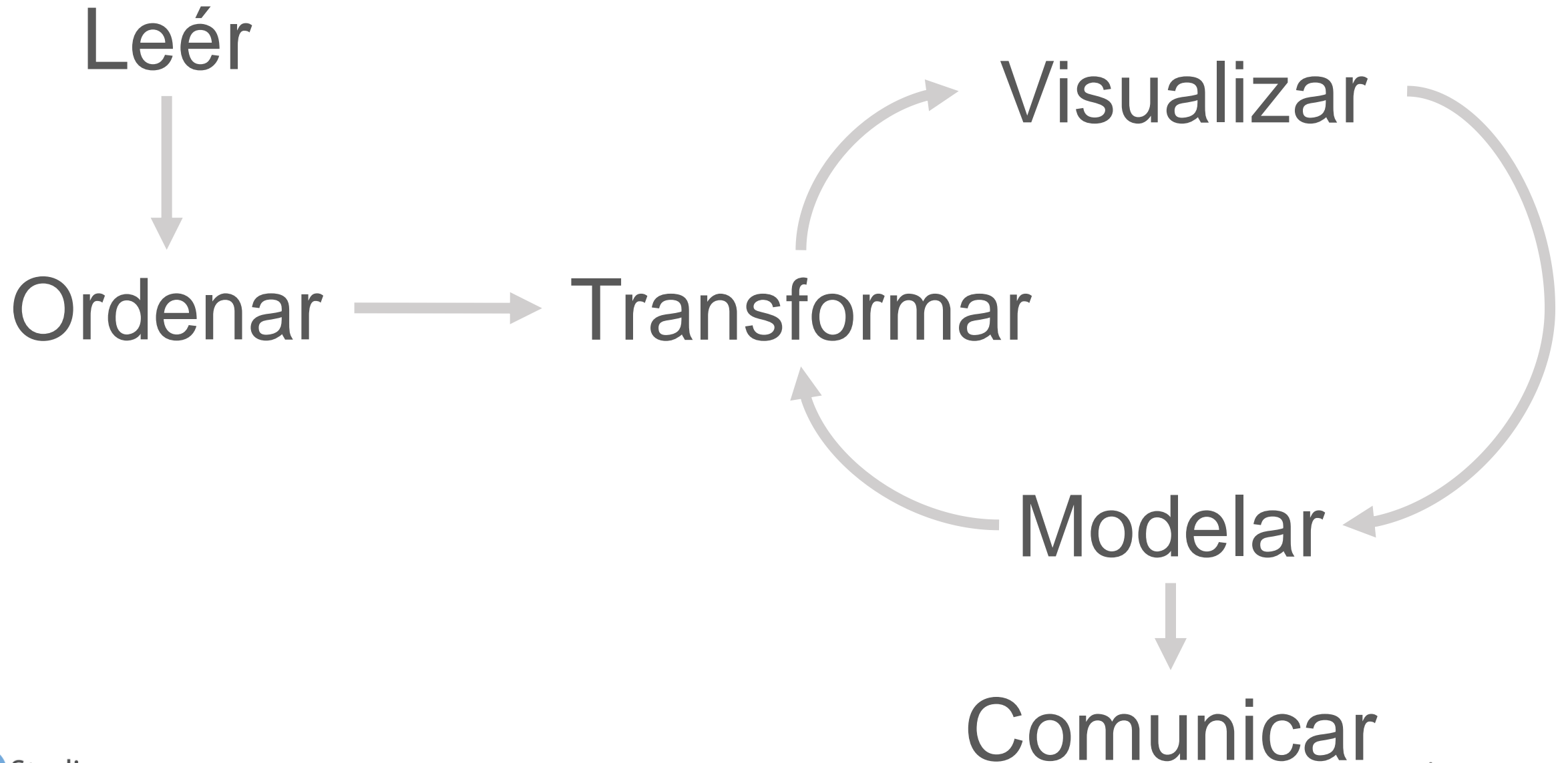


linkedin.com/in/edgararuiz



github.com/edgararuiz

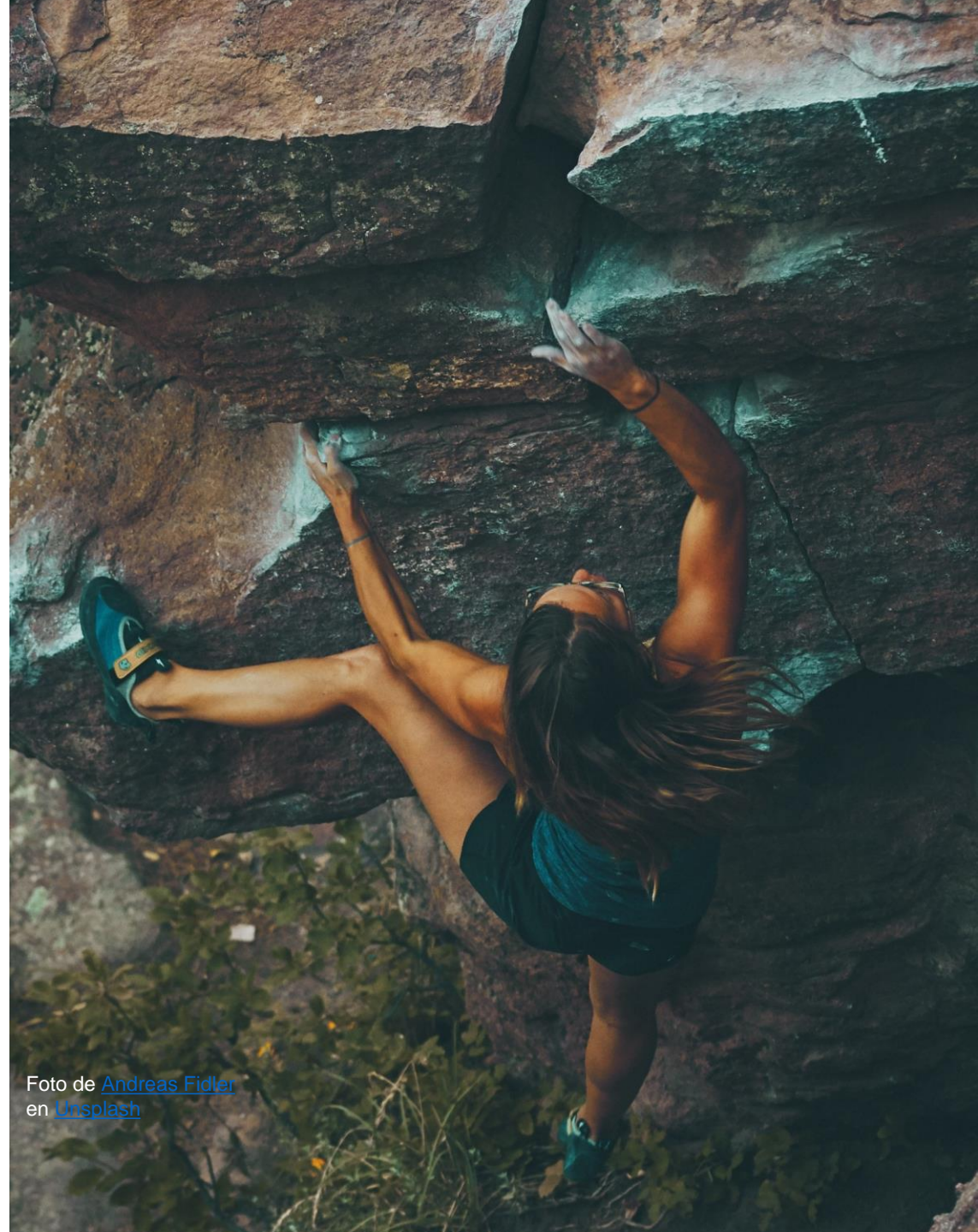
Ciencia de Datos



La fosa del éxito

Para llegar a la cima de una montaña, o atravesar un desierto, es necesario sobrevivir sorpresas y tribulaciones.

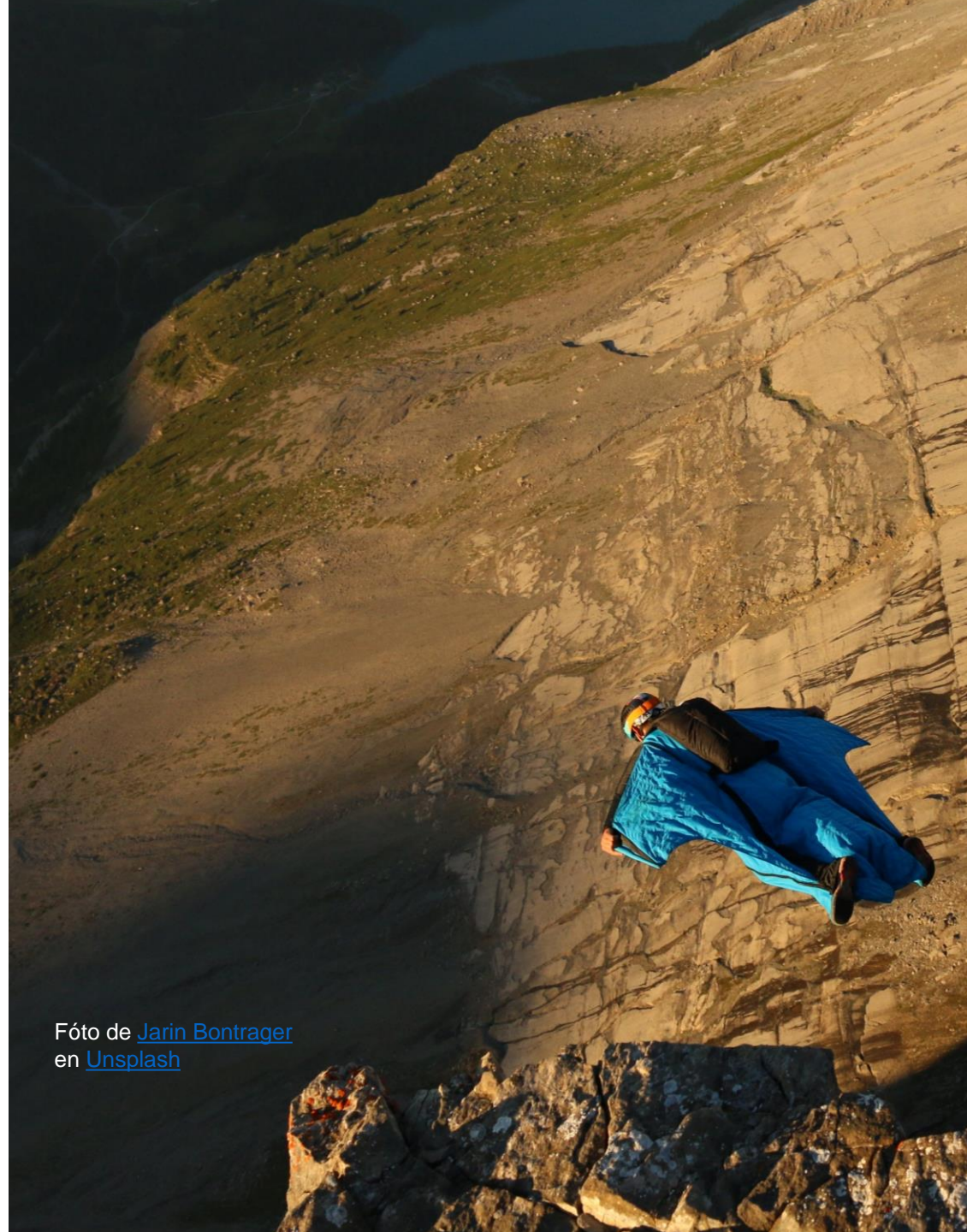
<https://blog.codinghorror.com/falling-into-the-pit-of-success/>



La fosa del éxito

Cuando usamos herramientas que no solo son **eficaces**, pero también **bastante fáciles de usar**, es como que “caemos” en una fosa...del éxito,

<https://blog.codinghorror.com/falling-into-the-pit-of-success/>



El
“tidyverse”



¿Que es el “tidyverse”?



Una colección de paquetes de R que son diseñados para la Ciencia de Datos.

Todos los paquetes utilizan la misma filosofía de diseño, gramática y estructuras de datos.

Paquetes del “tidyverse”

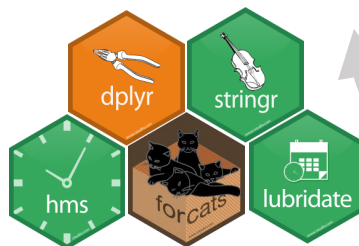
Leér



Ordenar



Transformar



Visualizar



Modelar



La meta del “tidyverse”



Fóto de [Kelly Sikkema](#) en [Unsplash](#)

Resolver problemas complicados mediante la combinación de diferentes piezas que son consistentes unas con otras

El equipo “tidyverse”



@hadleywickham



@jimhester_



@JennyBryan



@_lionelhenry



@krlmlr



@dataandme



@romain_francois



@drob



@thomasp85



@GaborCsardi



@LucyStats



@topepos

Los principios del “tidyverse”

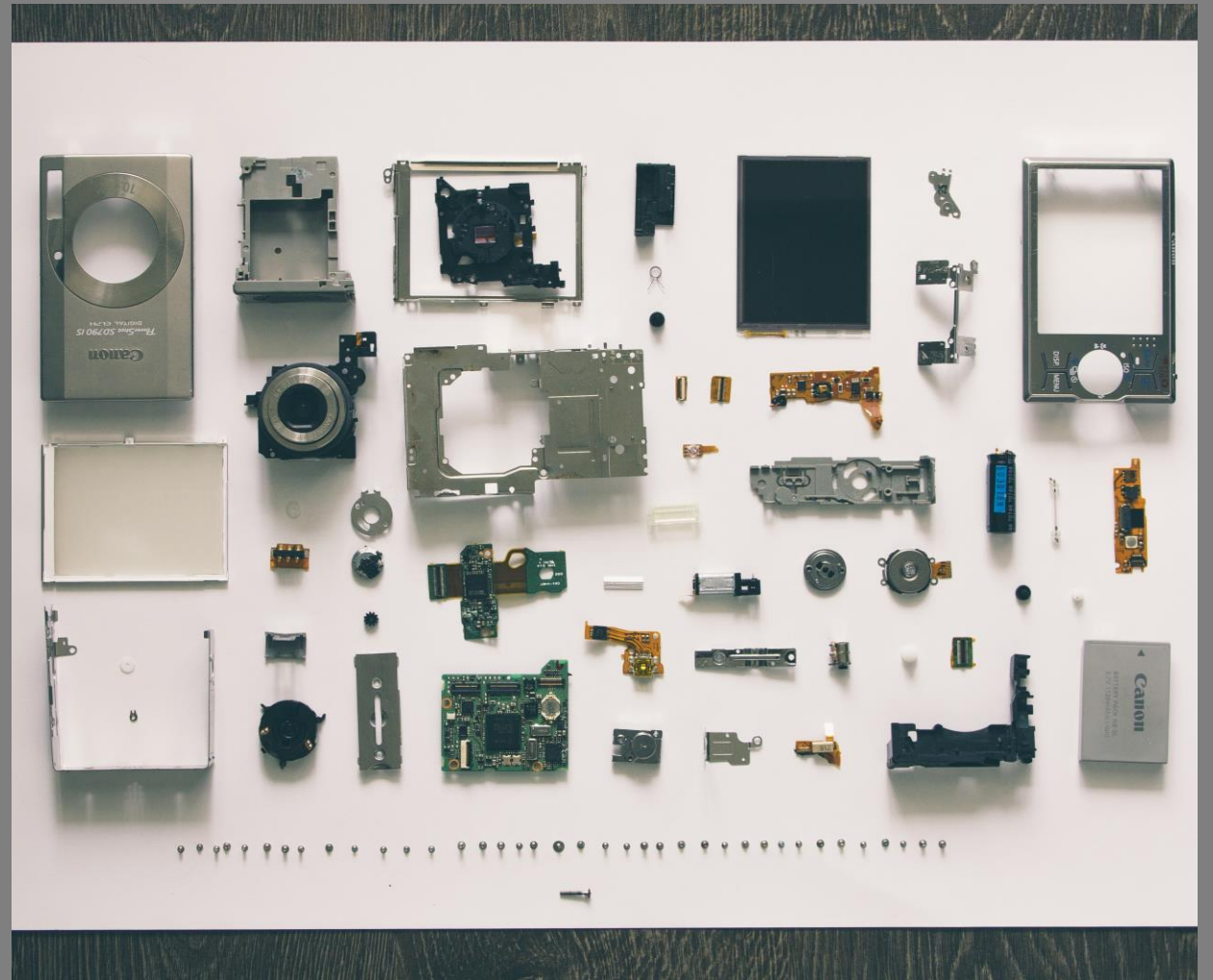


Foto de [Vadim Sherbakov](#) en [Unsplash](#)

Principio #1 – Use datos ordenados

1. Cada línea es una observación
2. Cada columna es una variable

Fecha	Nombre	Mate	Ingles
1-11-2015	Hernandez, Rodrigo	90	60



mes	año	primer	apellido	materia	puntos
11	2015	Rodrigo	Hernandez	mate	90
11	2015	Rodrigo	Hernandez	ingles	60



Principio #2 – Cada función es un paso

“Cual es el promedio actual de cada estudiante en la materia de Matemática”

- | | | |
|---|---|--|
| 1. Se queda con solo las calificaciones de Matemática | → | <code>matematica <- filter(datos, materia == "matematica")</code> |
| 2. Agrupa por cada primer nombre y apellido | → | <code>estudiante <- group_by(matematica, primer, apellido)</code> |
| 3. Calcule el promedio de las calificaciones | → | <code>promedio <- summarise(estudiante, promedio = mean(puntos))</code> |
| 4. Imprime los resultados | → | <code>promedio</code> |

Problemas con el código

```
matematica <- filter(datos,  
  materia == "matematica")  
  
estudiante <-  
  group_by(matematica, primer,  
    apellido)  
  
promedio <-  
  summarise(estudiante,  
    promedio = mean(puntos))  
  
promedio
```

1. Creamos variables que usamos solo una vez
2. Es difícil de leer, y entender, todo los pasos rápidamente.

Principio #3 – Use para combinar

Antes

```
matematica <- filter(datos,  
  materia == "matematica")  
  
estudiante <-  
  group_by(matematica, primer,  
    apellido)  
  
promedio <-  
  summarise(estudiante, promedio  
    = mean(puntos))  
  
promedio
```

Después

```
datos %>%  
  filter(materia == "matematica") %>%  
  group_by(primer, apellido) %>%  
  summarise(promedio = mean(puntos))
```

Explorar es fácil!

*“Cual es el promedio actual de cada estudiante en la materia de **Ingles**”*

Antes

“**promedio**” ya no tiene la información más reciente. Se tiene que correr la mayoría del código de nuevo. El nombre de algunas variables ya no son apropiadas.

```
matematica <- filter(datos,  
  materia == "ingles")  
  
estudiante <- group_by(matematica,  
  primer, apellido)  
  
promedio <- summarise(estudiante,  
  promedio = mean(puntos))  
  
promedio
```

Después

%>% hace que todo el código sea considerado un solo comando

```
datos %>%  
  filter(materia == "ingles") %>%  
  group_by(primer, apellido) %>%  
  summarise(promedio = mean(puntos))
```

Principio #4 – Comando o consulta

Consultas

`filter()`

`mutate()`

`summarise()`

Comandos

```
# R, asigne a la variable df
df <- datos %>%
  filter(materia == "ingles") %>%
  group_by(primer, apellido) %>%
  summarise(promedio = mean(puntos))

datos %>%
  filter(materia == "ingles") %>%
  group_by(primer, apellido) %>%
  summarise(promedio = mean(puntos))%>%
  print() # R, imprima los resultados.
# tidyverse lo agrega "virtualmente",
# sin que se tenga que escribir
```

Los 4 principios del “tidyverse”

1. Las estructuras principales son datos ordenados
2. Cada función representa un paso
3. Las funciones se combinan con `%>%`
4. Cada paso es una consulta o un comando

Recursos disponibles



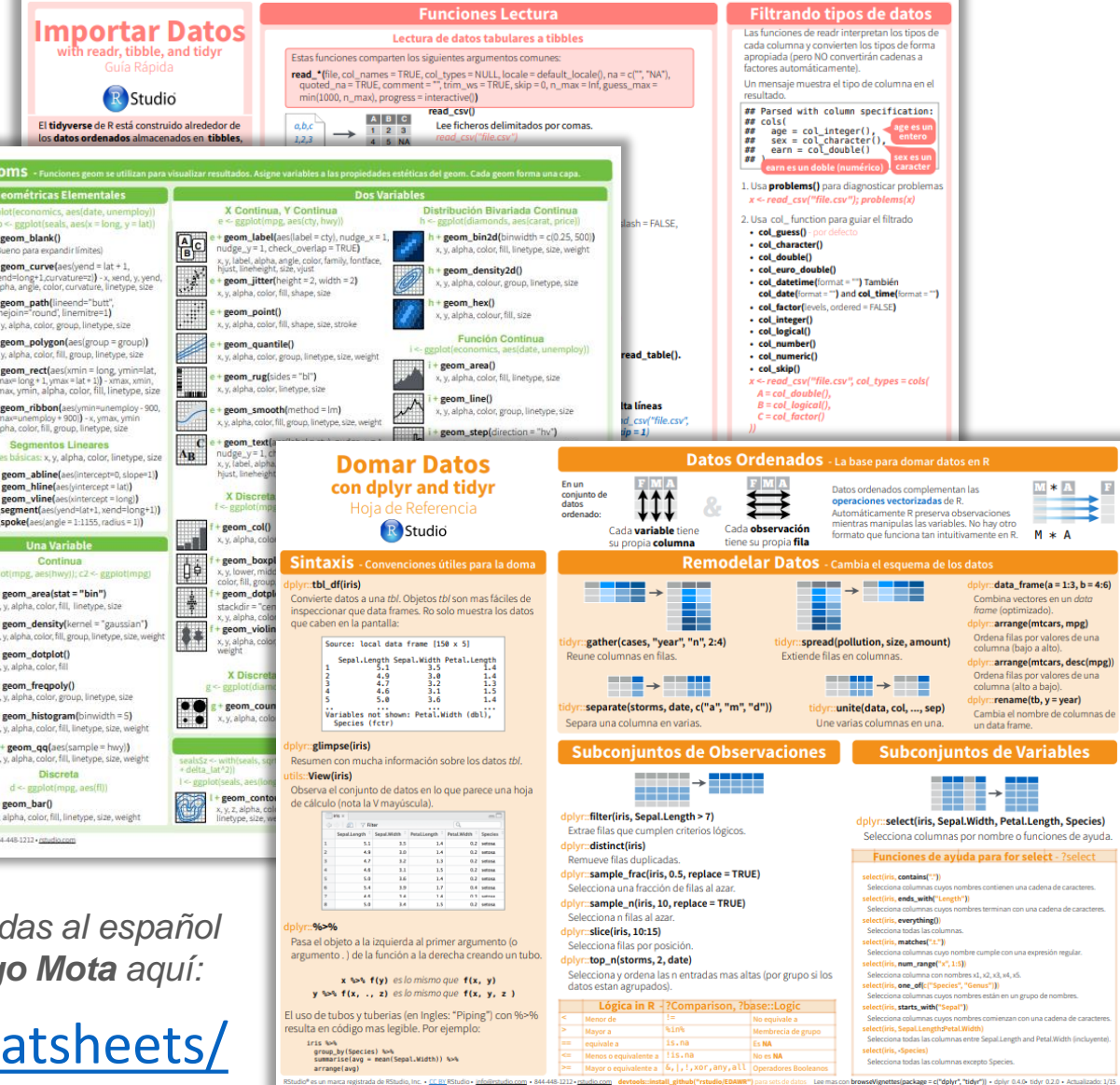
Photo by [Jonathan Simcoe](#) on [Unsplash](#)

Recursos en español

Para aprender como usar los paquetes en práctica, las Hojas de Referencia, o *Cheatsheets*, son los mejores recursos, no importa el idioma

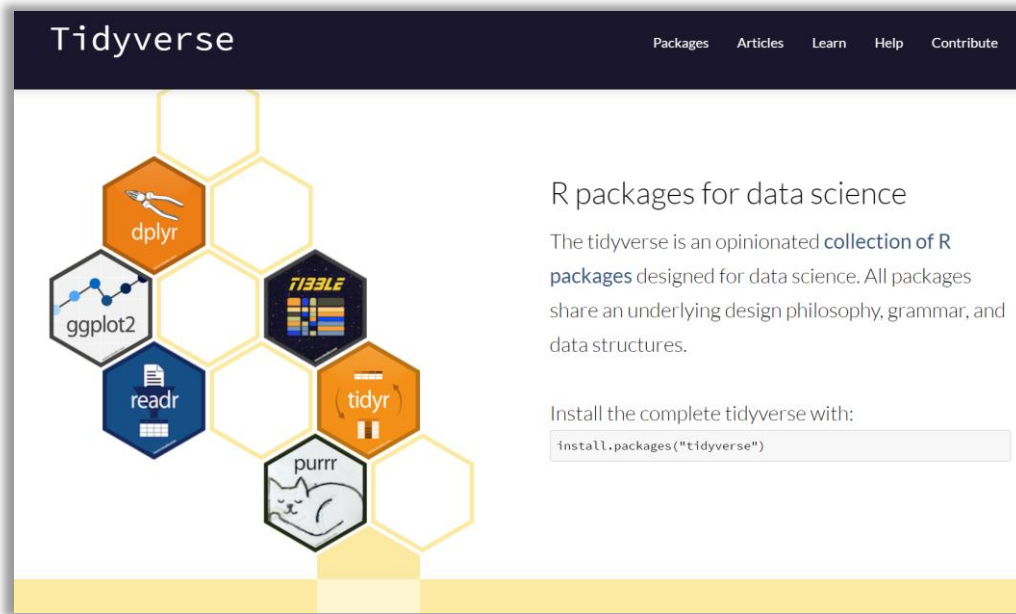
Tenemos disponibles varias Hojas de Referencia traducidas al español cortesía de **Frans van Dunné, Carlos Ortega y Santiago Mota** aquí:

rstudio.com/resources/cheatsheets/



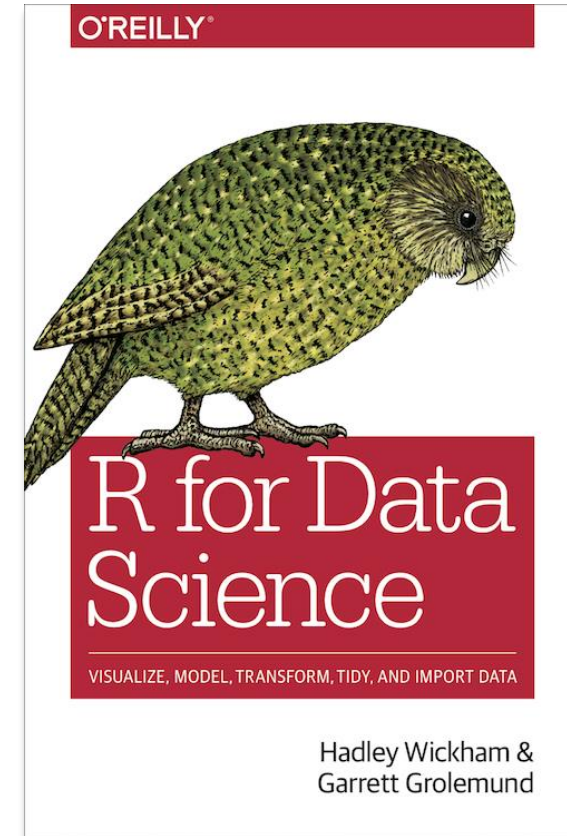
Recursos en ingles

Para las últimas noticias



tidyverse.org

Para aprender más a fondo



[R for Data Science](https://r4ds.had.co.nz)

Próximamente...

¡Traducción del libro al español!

