Manejo Habil de Datos y Ciencia Reproducible

2018-03-08

Preambulo

Quien soy?

Gabriel Muñoz

- Biólogo
- Ecólogo Computacional

Consultor en Datos *Biodiversidad y Geoespaciales*

Coordinador General

Charla Introductoria

Talleres participativos

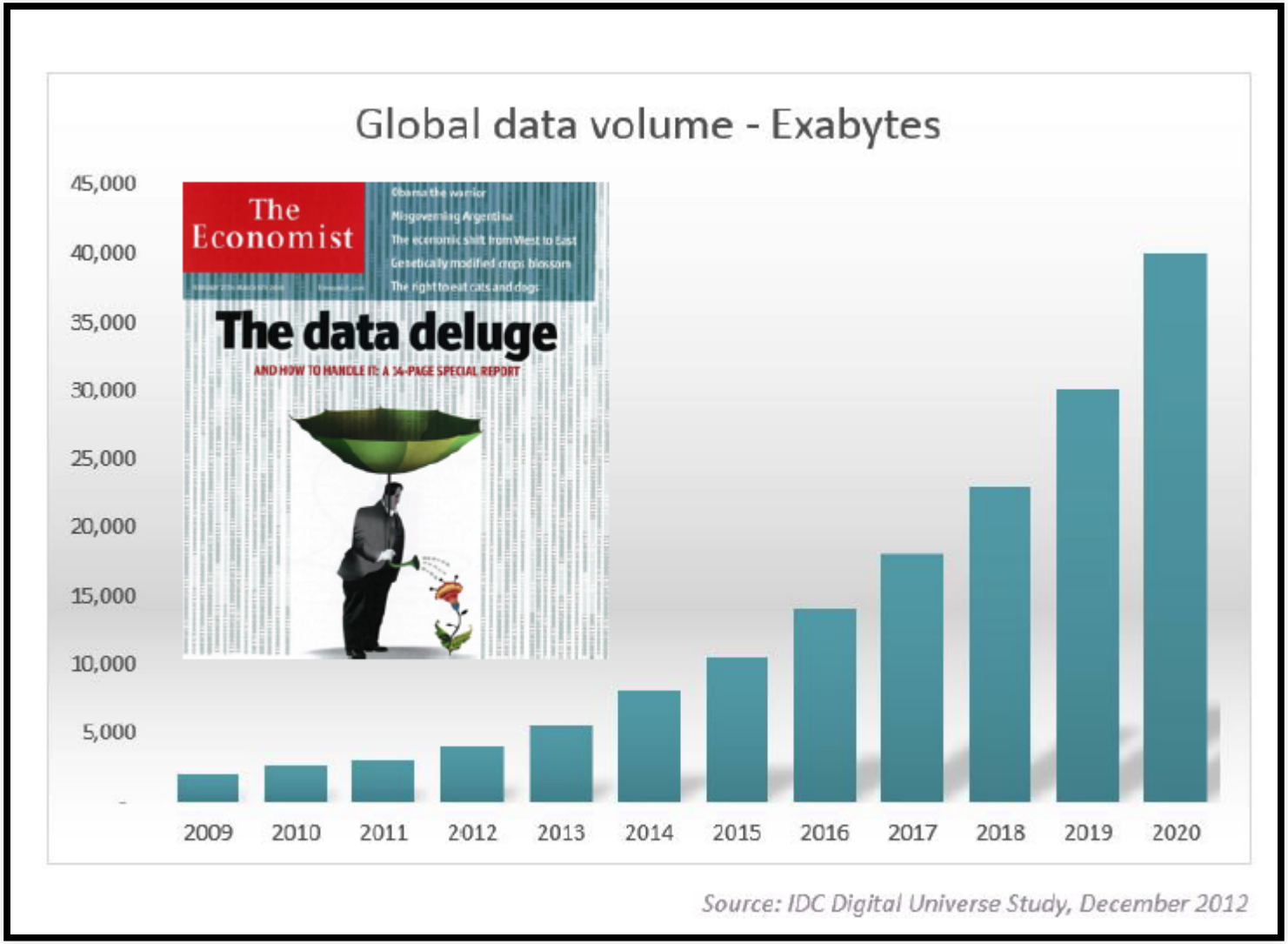
Aprender herramientas de manera práctica

22,23 de Marzo.

MediaLAB

Datos

Mundo de Datos



Mundo de Datos

- Contar Historias
- Comunicar Ideas
- Entender el Mundo

Datos tienen Valor

- Inversión e
- Inversión e

(

- Organizació
  de Datos He

Hum
Machi

H

Manejo Hábil de Datos

De que se trata?

Desarrollar habilidades

y aprender herramientas…

Ciclo de vida de los Datos

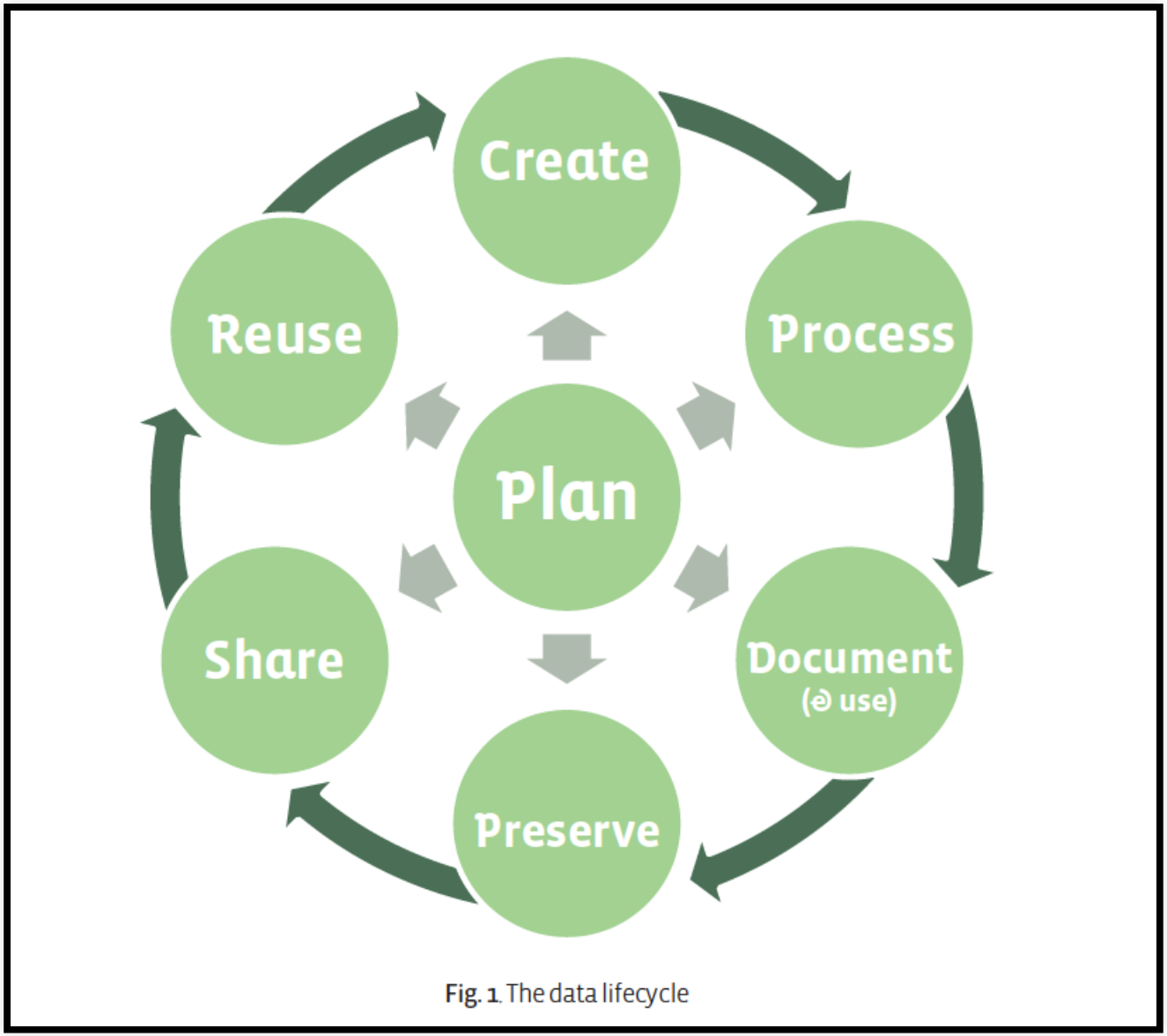


Fig. 1. The data lifecycle

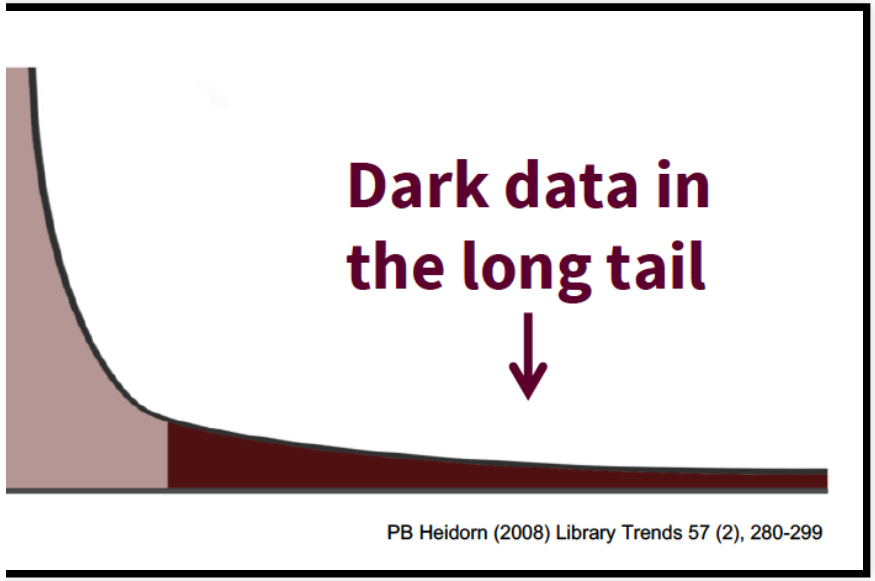
Creación (*Obtención y Descubrimiento*)

que ya existe?

Open Databases

ope
DataVe
Cifras- re3
EOL- Zenoc

Creación (*Obtención y Descubrimiento*)

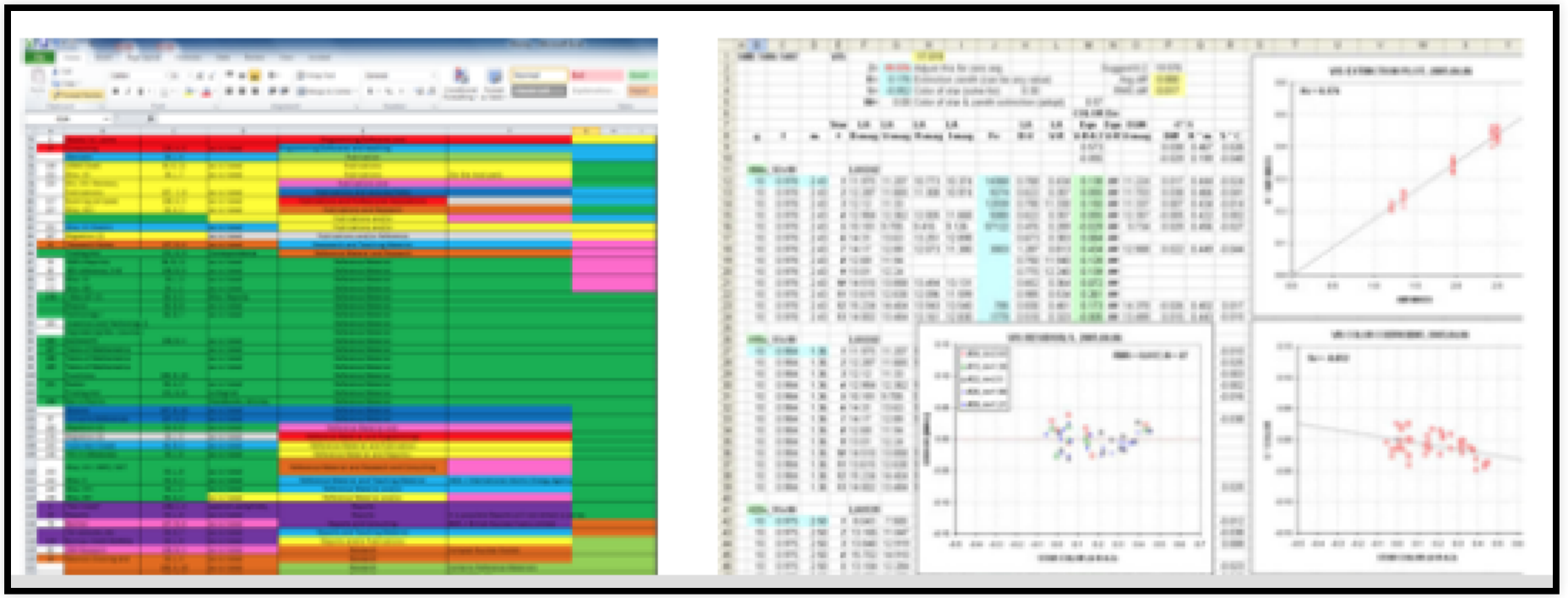


Datos específi
de en

Procesamiento

Datos heterogeneos, desordenados (“Messy Data”)

Hojas de calculo



Procesamiento

Datos heterogeneos, desordenados (“Messy Data”)

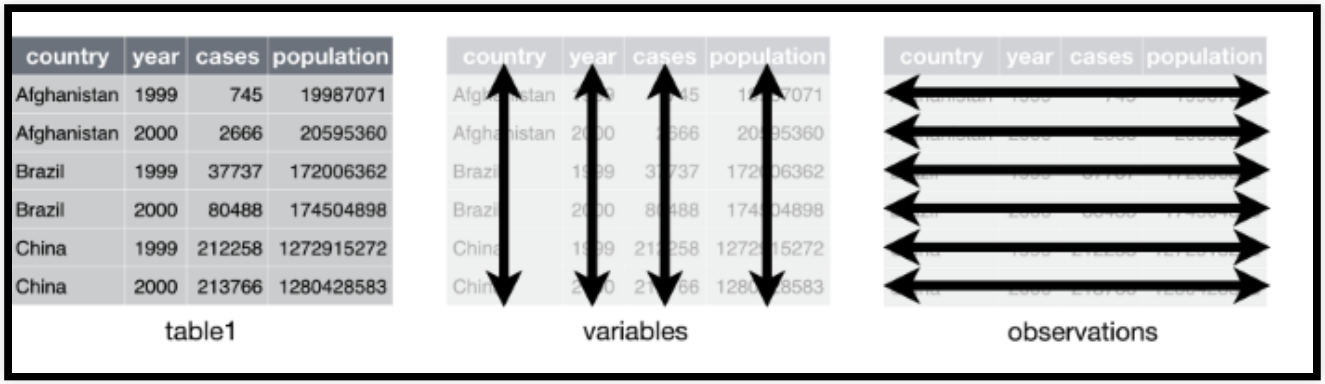
Procesamiento

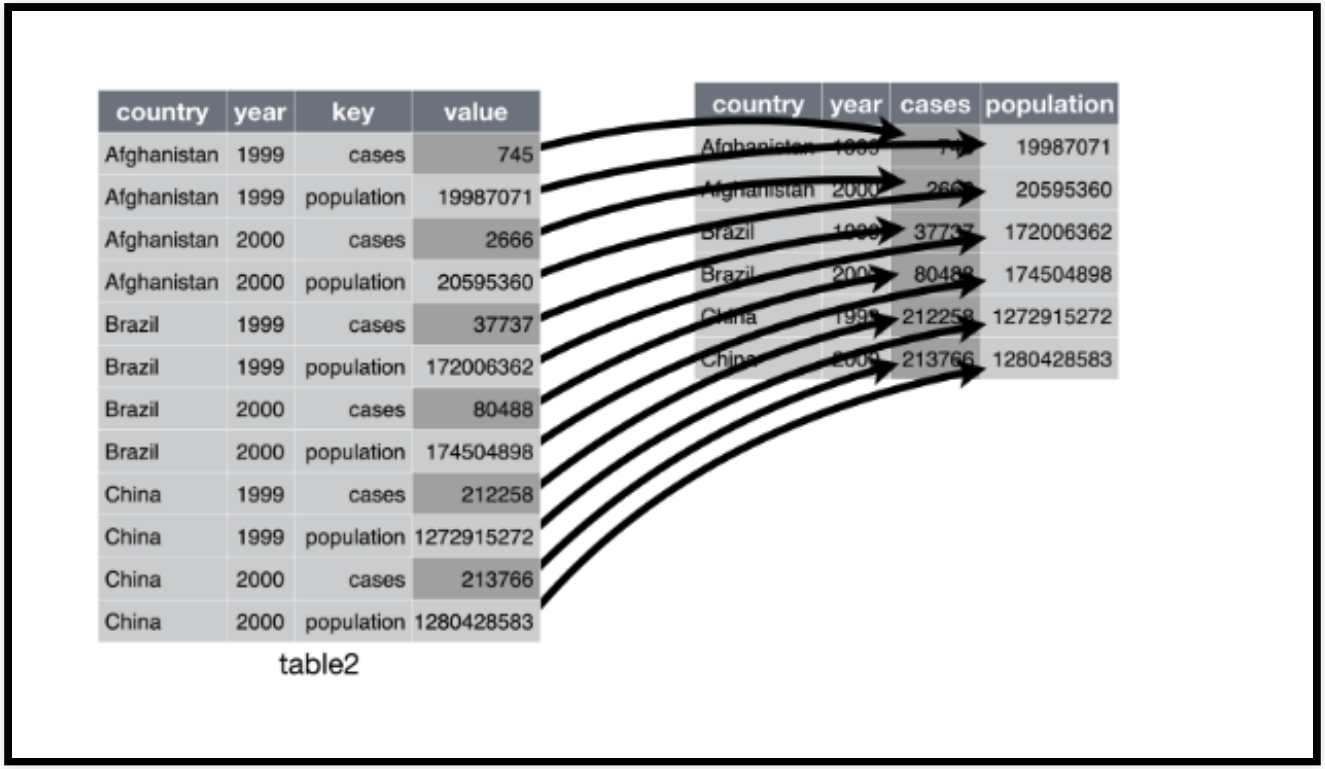
Procesamiento

Una Variable, Una Columna!!!!

Procesamiento

Una Variable, Una Columna!!!!




country year key value
Afghanistan 1999 cases 745
Afghanistan 1999 population 19987071
Afghanistan 2000 cases 2666
Afghanistan 2000 population 20595360
Brazil 1999 cases 37737
Brazil 1999 population 172006362
Brazil 2000 cases 80488
Brazil 2000 population 174504898
China 1999 cases 212258
China 1999 population 1272915272
China 2000 cases 213766
China 2000 population 1280428583
table2
country year cases population
19987071
20595360
172006362
174504898
1272915272
1280428583

Procesamiento

Pequeños problemas, grandes dolores de cabeza!

- Errores de tipeo
- Discrepancias
- Fueras de lugar
- Errores de formato
- Irregularidades
- Datos faltantes

Procesamiento

Pequeños problemas, grandes dolores de cabeza!

- Contradicciones
- Duplicaciones
- Fueras de rango
- Incongruencias
- Multiples valores

Procesamiento (Soluciones)

- Aprender a usar codigo R tidyr, dplyr
- tidyverse
- GoogleOpenRefine

Ciclo de vida de los Datos

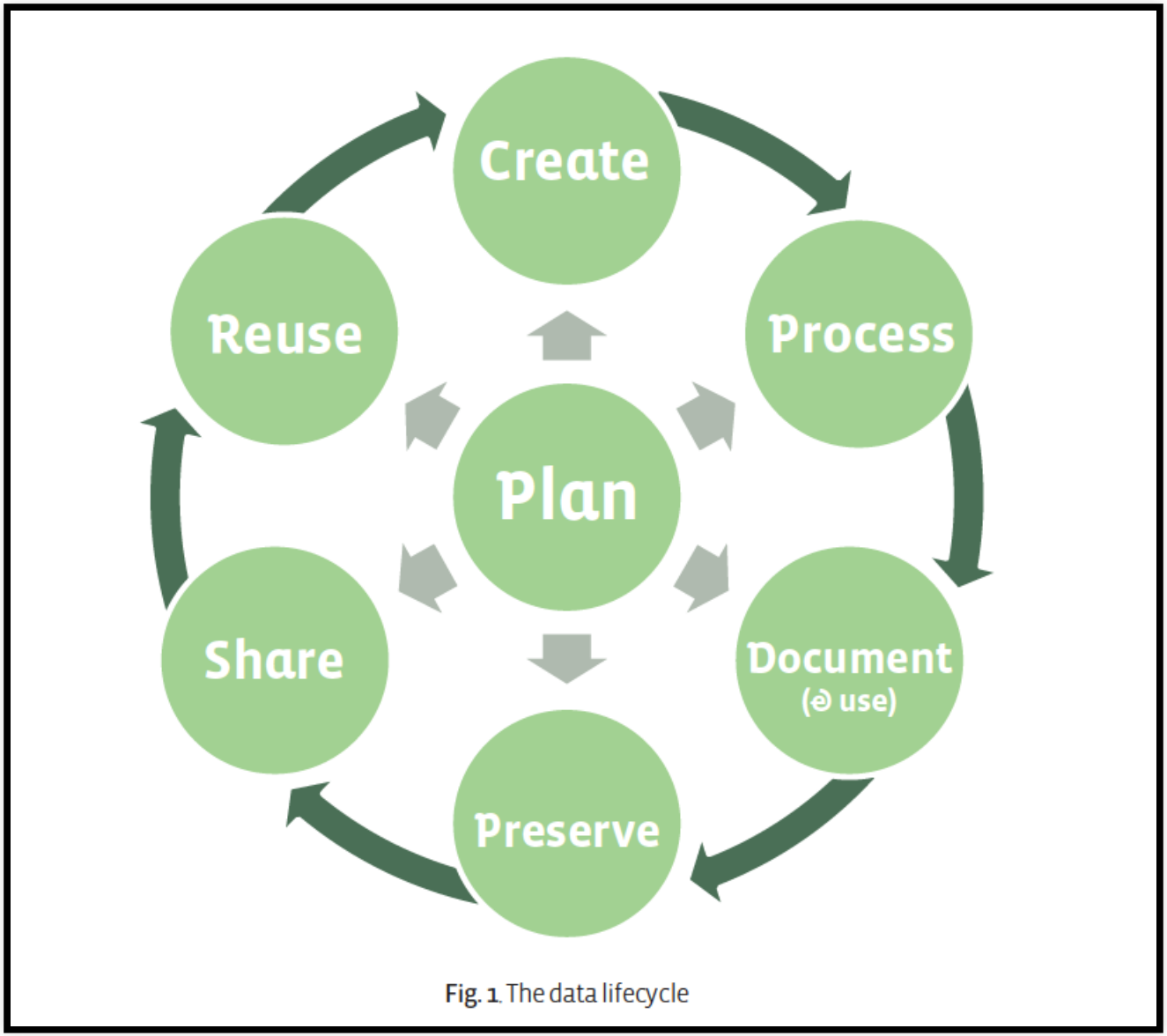


Fig. 1. The data lifecycle

Documentacion

Metadata a nivel del proyecto

- Objectivos
- Personal
- Standards
- Metodos de colecta de datos
- Estructura y organización de los datos
- Software usado
- Como citar los datos
- Propiedades intelectuales y licencias

Documentacion

Metadata a nivel de variables

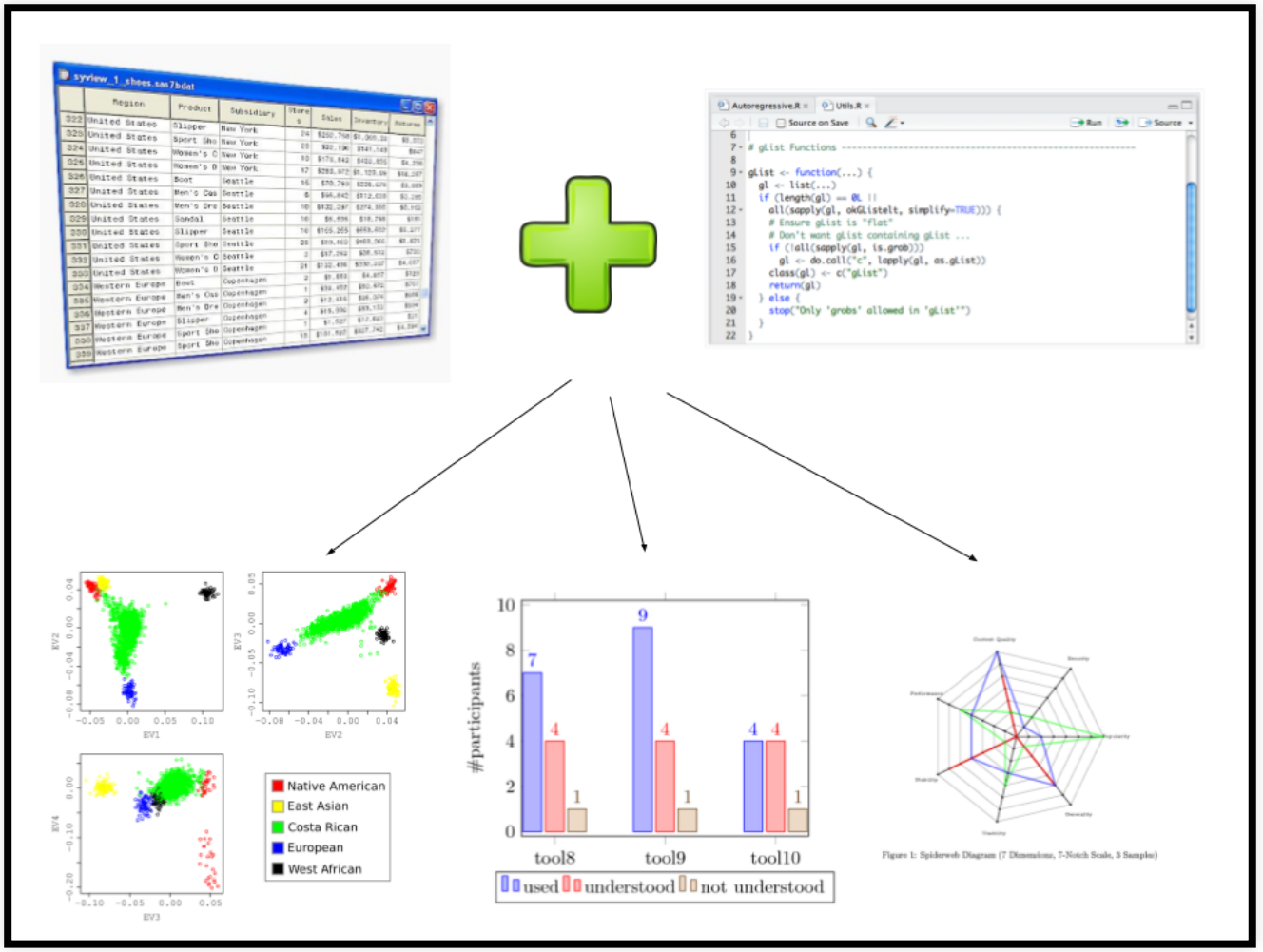
- Nombres, etiquetas y descripción
- Codigos, explicación de cada uno
- Acrónimos
- Valores faltantes? Razones
- Datos derivados del raw file

Documentacion

Uso de software como R, mantiene la documentación de datos durante el proceso de análisis

Documentación del análisis

Usa codigo (Scripts)



Preserva, Comparte y Reusa

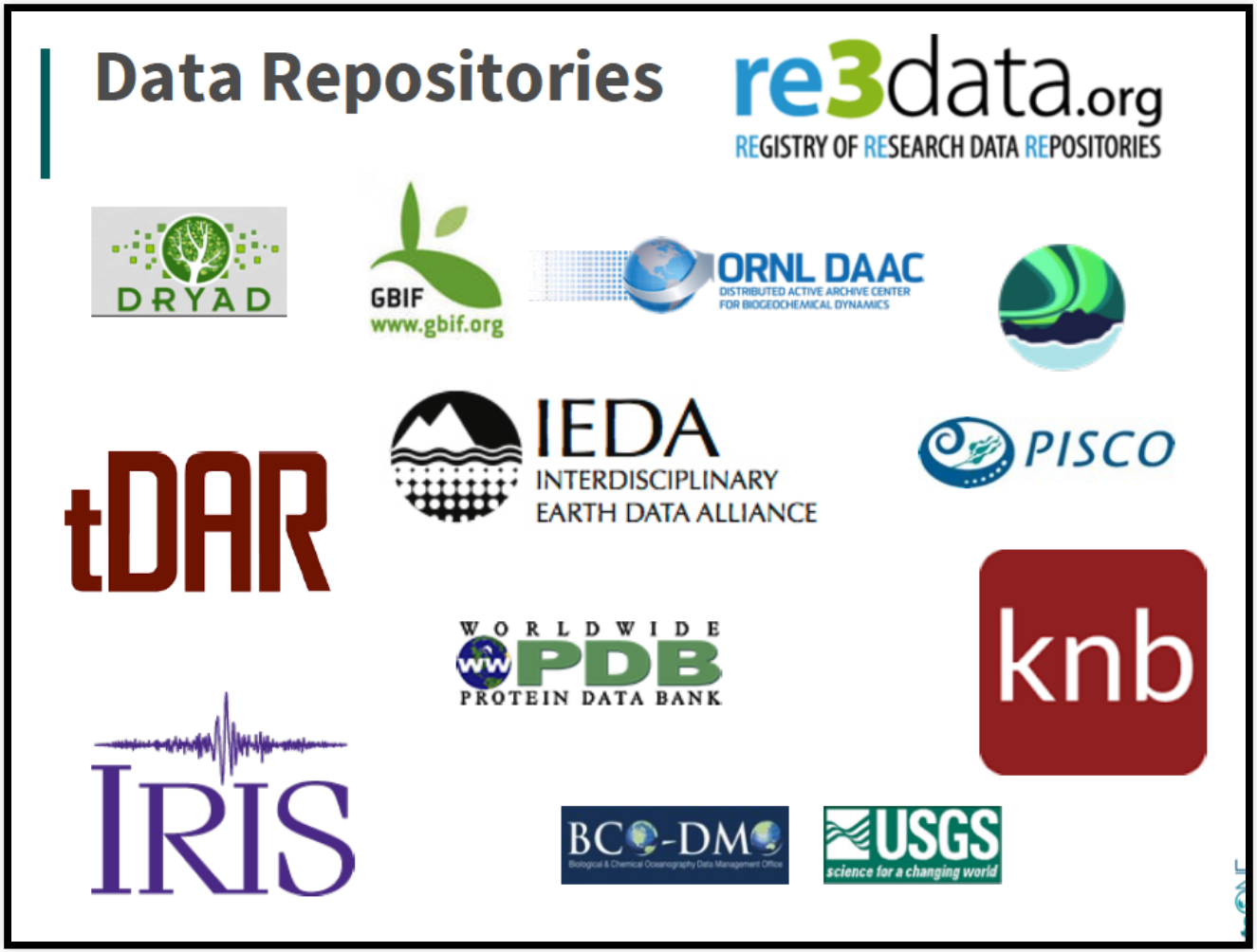
- Registros similares en un dataset
- Seguir standards internacionales en unidades y formatos (e.g. yyyy-mm-dd)
- Se consistente (e.g. Mts, (m), metros)
- Preserva en formatos estables (.csv, .txt, TIFF)

Preserva, Comparte y Reusa

- Especificar coordenadas espaciales y temporales (e.g. UTM 17S, Geograficas)
- RawData as “ReadOnly” y comparte tus scripts
- Asegura la calidad de tus datos
- Provee Documentación

Preserva, Comparte y Reusa

- Protege tus datos (Original, Online, Offline)



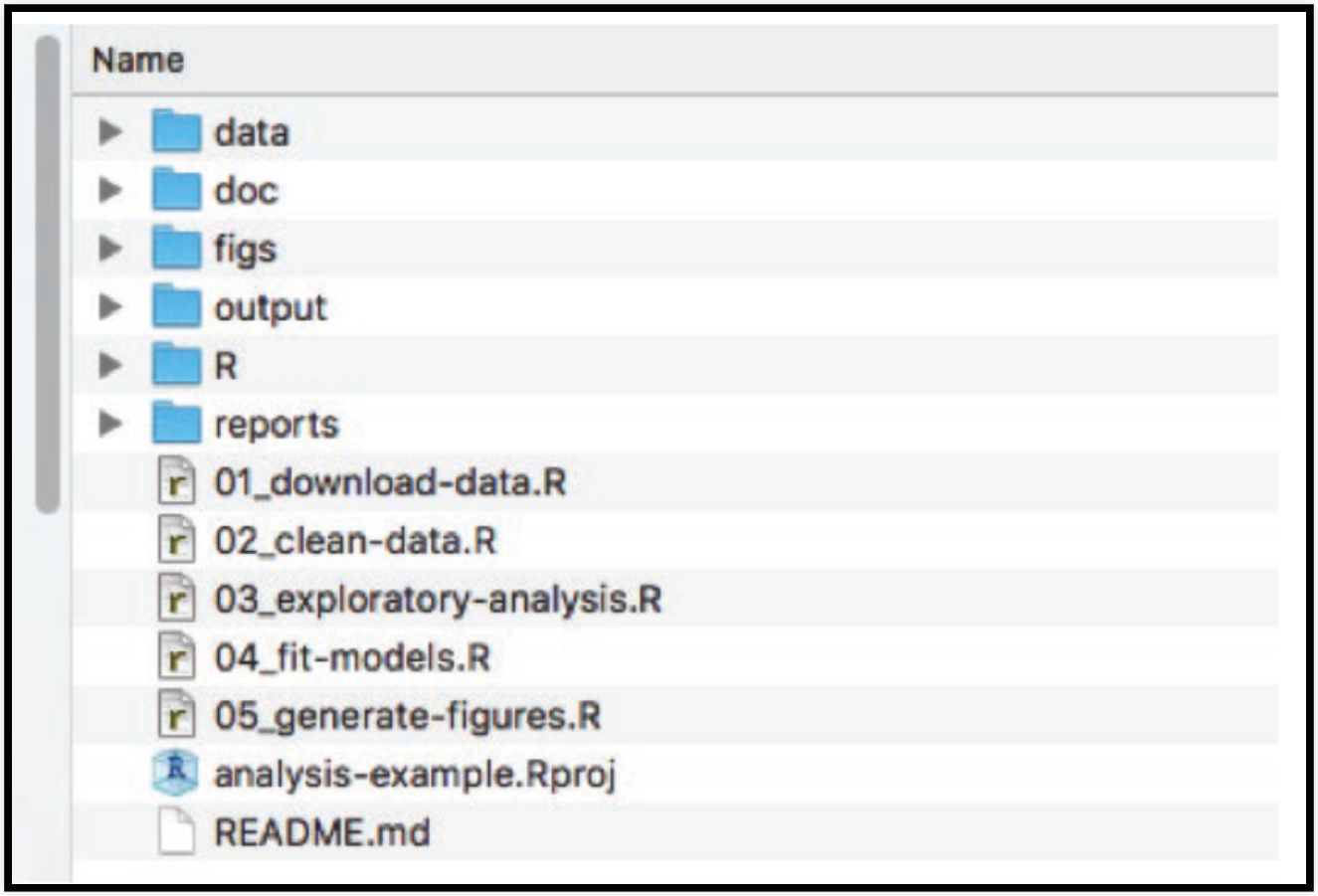
Preserva, Comparte y Reusa

- ZENODO
- GitHub *Personal favorite*
- DataVerse

Ciencia Reproducible

Un proyecto, una carpeta

No te metas con el “raw data”



Elige bien tus nombres

Bueno:

- rawDatasetAgo2017.csv
- 01_figurasIntro.R
- 02_figurasResultados.R
- gastoPublicoVicepresciденciaGlass2016.csv
- datos_corrupción_Ecuador.csv

Elige bien tus nombres

Malo:

- 1.csv
- Archivo-Corregido.R
- asnbx.csv
- 1-4.csv
- download.csv
- No.separes.con.puntos.csv

Elige bien tus nombres

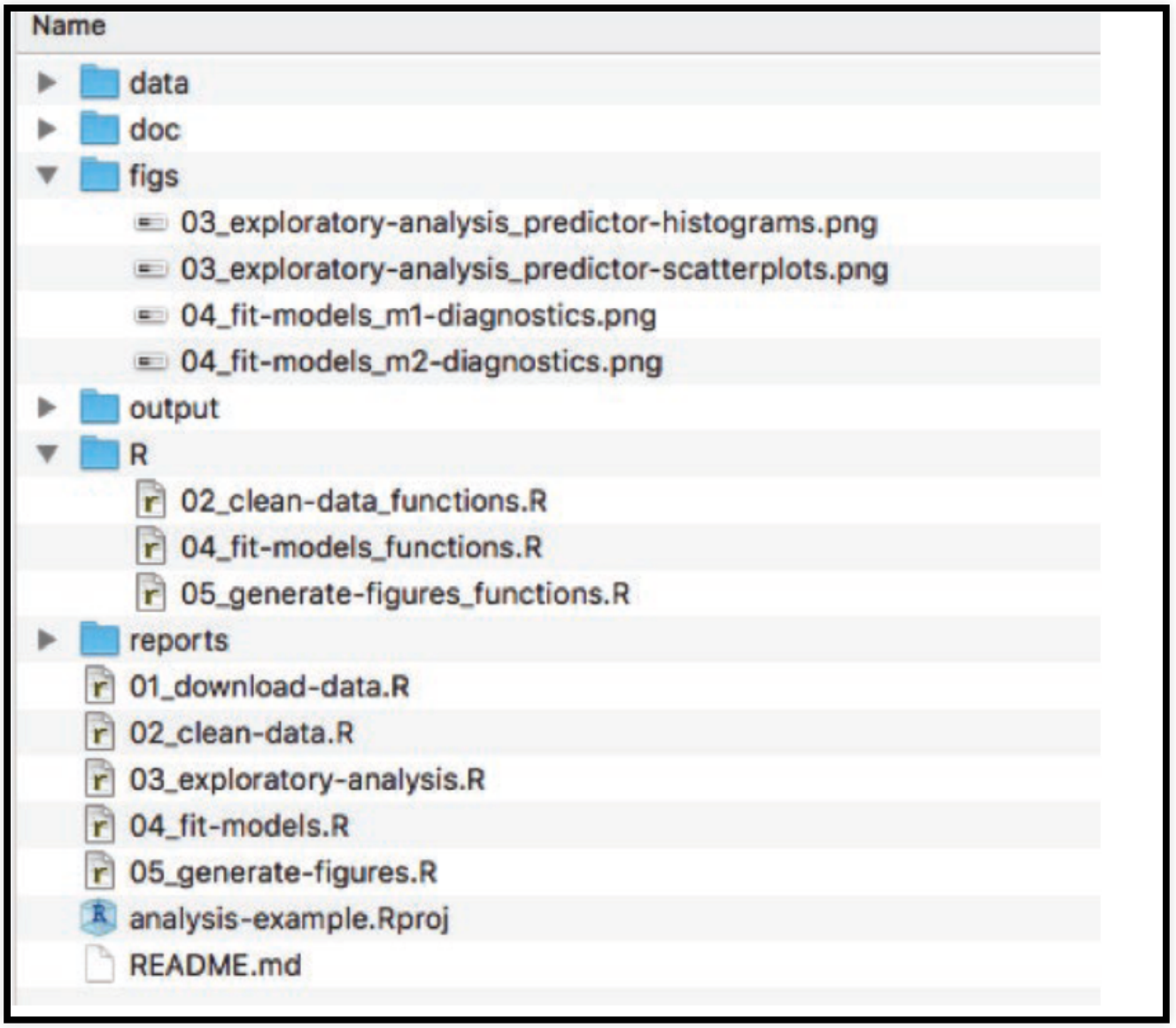
Empieza listas con un 0

- 01
- 02
- 03
- …
- 10

de lo contrario

- 10
- 1
- 2
- 3
- ..

Un proyecto, una carpeta :)



Ahora si, a escribir código!

Herramientas

- Abiertas
- Grátis!
- Poderosas
- Gran Comunidad Mundial
- Actualizables
- Desarrollador –> Usuario

Ahora si, a escribir código!

Lenguages de programación

- R
- Python
- SQL
- Java
- Scala
- Julia
- Perl
- Ruby
- C++
- Matlab

R y R Studio

- OpenSource
- Gratis

https://www.rstudio.com/products/rstudio/features/

Buenas prácticas de programación

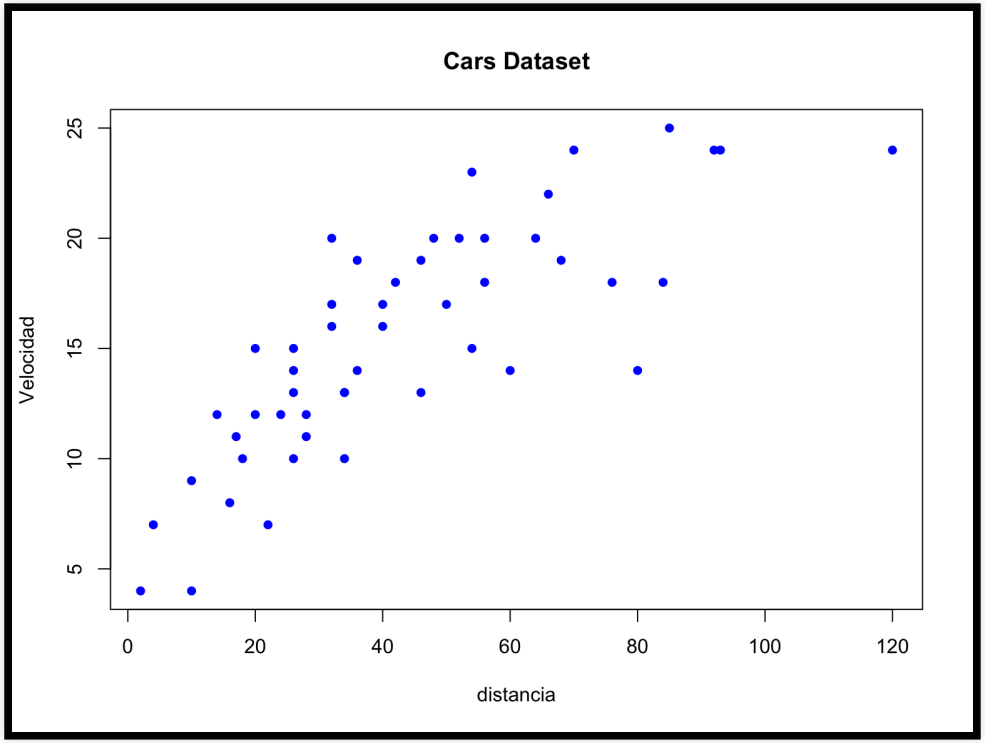
Summary

Good code is like a good joke:

It needs no explanation

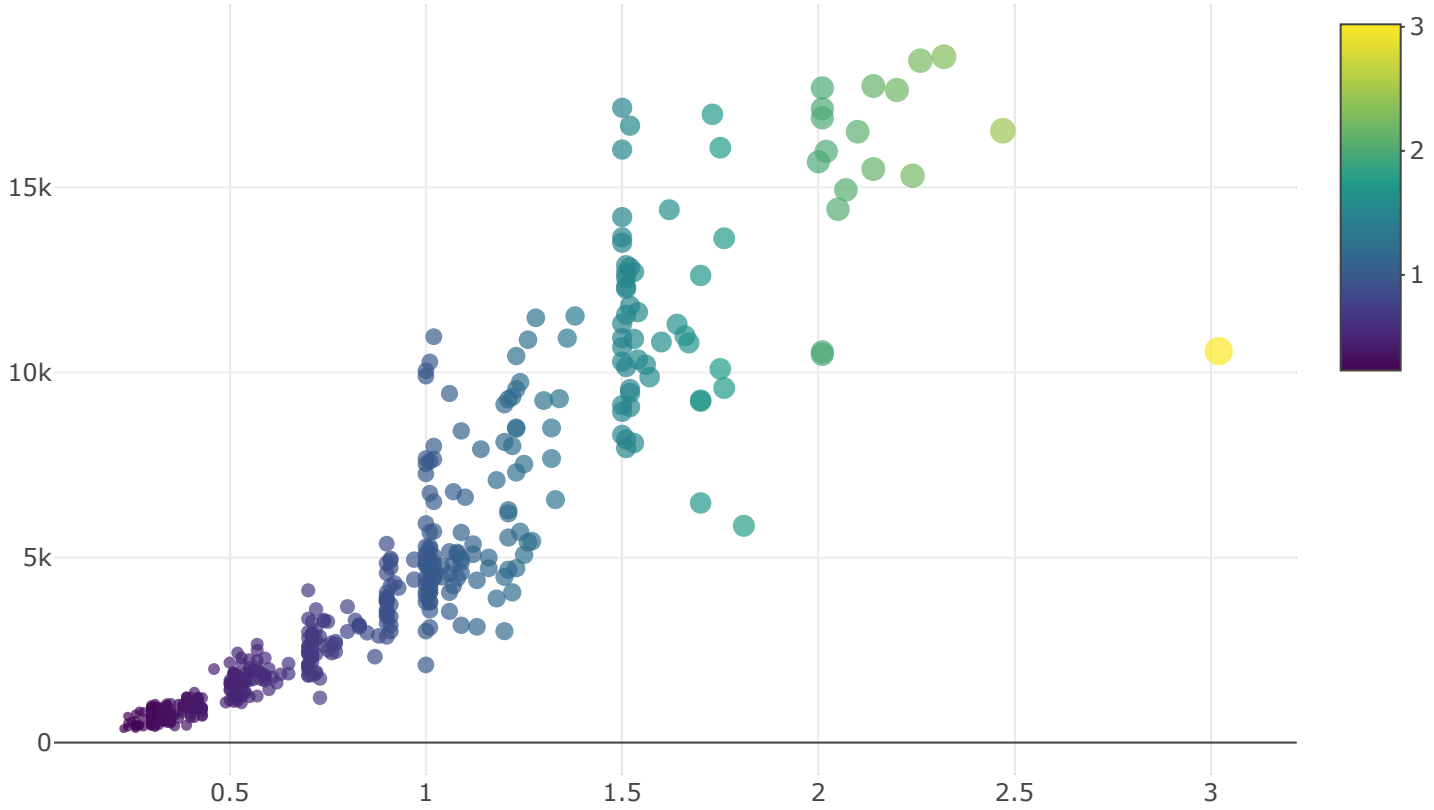
Visualización estática

```
plot(speed~dist, cars,
     xlab = "distancia",
     ylab = "Velocidad",
     col = "blue",
     pch = 16,
     main = "Cars Dataset")
```



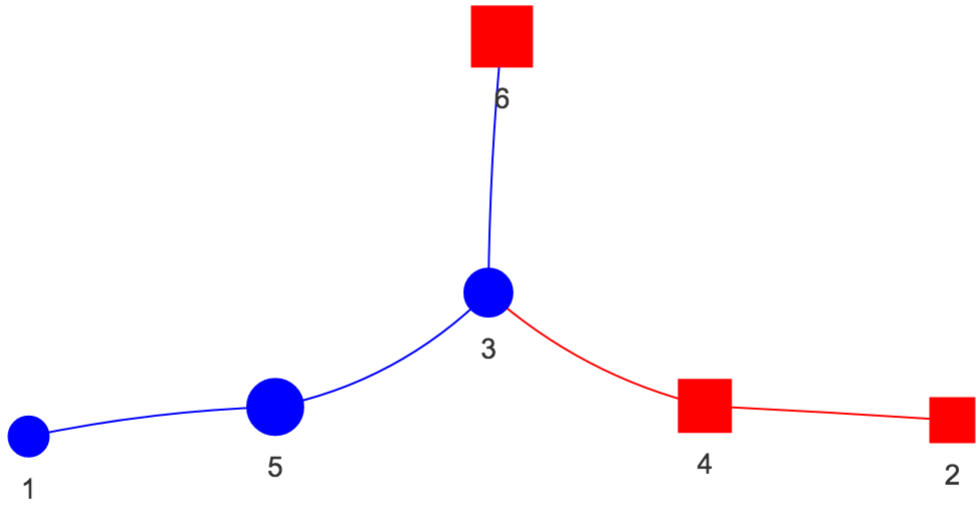
Visualización dinámica

```
library(plotly)
d <- diamonds[sample(nrow(diamonds), 500), ]
plot_ly(d, x = d$carat, y = d$price,
        text = paste("Clarity: ", d$clarity),
        mode = "markers", color = d$carat, size = d$carat)
```

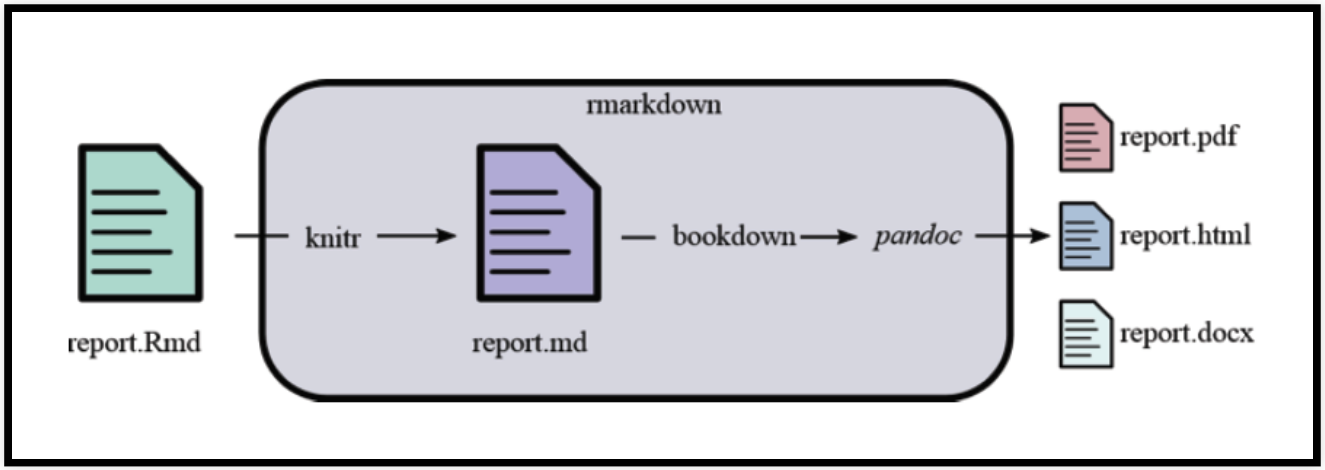


Mas ejemplos

Select by id



Markdown y reportes reproducibles



Maarkdown es un lenguaje simple, pero poderoso para crear publicaciones. Facil sintaxis que permite el formato y la inclusión de links, imagenes, referencias, en el texto.

Ejemplos:

Primer example

Una ecuación: `$A = \sin(x)^2*log(a)+25^{2}$`

Esto es código `this is code`

$$A = \sin(x)^2 * log(a) + 25^2$$

Segundo ejemplo

Una imagen: `!(figs/copypaste.jpg)`

No máss copy - paste!

Tercer ejemplo:

hypervinculo: `[GBIF](http://www.gbif.org/)`

Este el es link a la página del GBIF.

Sintax general

Pandoc's Markdown

Write with syntax on the left to create effect on right (after render)

```
Plain text

End a line with two spaces
to start a new paragraph.

*italics* and **bold**

`verbatim code`

sub/superscript^2^~2~

~~strikethrough~~

escaped: \* \_ \\

endash: --, emdash: ---

equation: $A = \pi*r^{2}$

equation block:

$$E = mc^{2}$$

> block quote

# Header1 {#anchor}
```

Plain text

End a line with two spaces
to start a new paragraph.

italics and **bold**

`verbatim code`

sub/superscript$^{2}_{2}$

~~strikethrough~~

escaped: * _ \

endash: –, emdash: —

equation: $A = \pi * r^2$

equation block:

$$E = mc^2$$

> block quote

```
## Header 2  {#css_id}

### Header 3  {.css_class}

#### Header 4

##### Header 5

###### Header 6

<!--Text comment-->

\textbf{Tex ignored in HTML}
<em>HTML ignored in pdfs</em>

<http://www.rstudio.com>
[link](www.rstudio.com)
Jump to [Header 1](#anchor)
image:

![Caption](smallorb.png)

* unordered list
    + sub-item 1
    + sub-item 2
        - sub-sub-item 1

* item 2

    Continued (indent 4 spaces)
```

Header1

Header 2

Header 3

Header 4

Header 5

Header 6

HTML ignored in pdfs

http://www.rstudio.com
link
Jump to Header 1
image:

Caption

- unordered list
    - sub-item 1
    - sub-item 2
        - sub-sub-item 1
- item 2

    Continued (indent 4 spaces)

```
1. ordered list
2. item 2
        i) sub-item 1
          A.  sub-sub-item 1

(@)  A list whose numbering
continues after

(@)  an interruption

Term 1

:   Definition 1

| Right | Left | Default | Center |
|------:|:-----|---------|:------:|
|    12 |  12  |    12   |    12  |
|   123 |  123 |   123   |   123  |
|     1 |    1 |     1   |     1  |

- slide bullet 1
- slide bullet 2

(>- to have bullets appear on click)

horizontal rule/slide break:

***

A footnote [^1]
```

1. ordered list
2. item 2
    i. sub-item 1
        A. sub-sub-item 1

1. A list whose numbering

continues after

2. an interruption

Term 1

Definition 1

Right	Left	Default	Center
12	12	12	12
123	123	123	123
1	1	1	1

- slide bullet 1
- slide bullet 2

(>- to have bullets appear on click)

horizontal rule/slide break:

A footnote [1]

```
[^1]: Here is the footnote.
```

1. Here is the footnote.↩

Por que usar Rmarkdown?

R Markdown te permite integrar tu codigo y flujo de analisis con texto escrito en sintaxis markdown. Esto asegura que tus análisis sean reproducibles, interactivos, compartibles y agradables de visualizar al momento de reportar. Al mismo tiempo reduce el tiempo ocupado en escribir y **Formatear** reporters. Tareas manuales como enumerar figuras, bibliografía, formatos de tablas son ahora automatizadas. Empiezas a escribir el reporte al tiempo que haces tus análisis.

Esta presentación fue hecha en Markdown

Por que usar Rmarkdown?

Markdown soporta no solo R, pero tambien otros lenguajes

Python:

```
print 'Hello, world. This is Python:'
import sys
print(sys.version) # Python version
```

```
## Hello, world. This is Python:
## 2.7.10 (default, Feb  7 2017, 00:08:15)
## [GCC 4.2.1 Compatible Apple LLVM 8.0.0 (clang-800.0.34)]
```

Por que usar Rmarkdown?

R

```
print("Hello, world. This is R")
```

```
## [1] "Hello, world. This is R"
```

```
version
```

```
##                _
## platform       x86_64-apple-darwin15.6.0
## arch           x86_64
## os             darwin15.6.0
## system         x86_64, darwin15.6.0
## status
## major          3
## minor          4.3
## year           2017
## month          11
## day            30
## svn rev        73796
## language       R
## version.string R version 3.4.3 (2017-11-30)
```

```
## nickname       Kite-Eating Tree
```

En resumen:

$$Markdown + R + Bookdown = Reports(Dynamic + Reliable)$$

1

Shiny: Une todos los pasos en una aplicación web!

https://shiny.rstudio.com/

Shiny: Une todos los pasos en una aplicación web!

https://shiny.rstudio.com/gallery/movie-explorer.html

http://datasociety.co/kitamba-the-opportunity-project/

https://shiny.rstudio.com/gallery/widget-gallery.html

Gracias!

Espero verlos en los talleres

Contacto

Gabriel Muñoz

- fgabriel1891@gmail.com
- nasua.research@gmail.com
- 0960809080
- /fgabriel1891 (GitHub)

Repositorio de esta presentación:

https://github.com/fgabriel1891/ManejoHabilDatosMediaLab

1. Reliable in the sense the code is properly written.↩