



Introducción al lenguaje R

Tutorial de Congreso Argentino de Agroinformática 2019

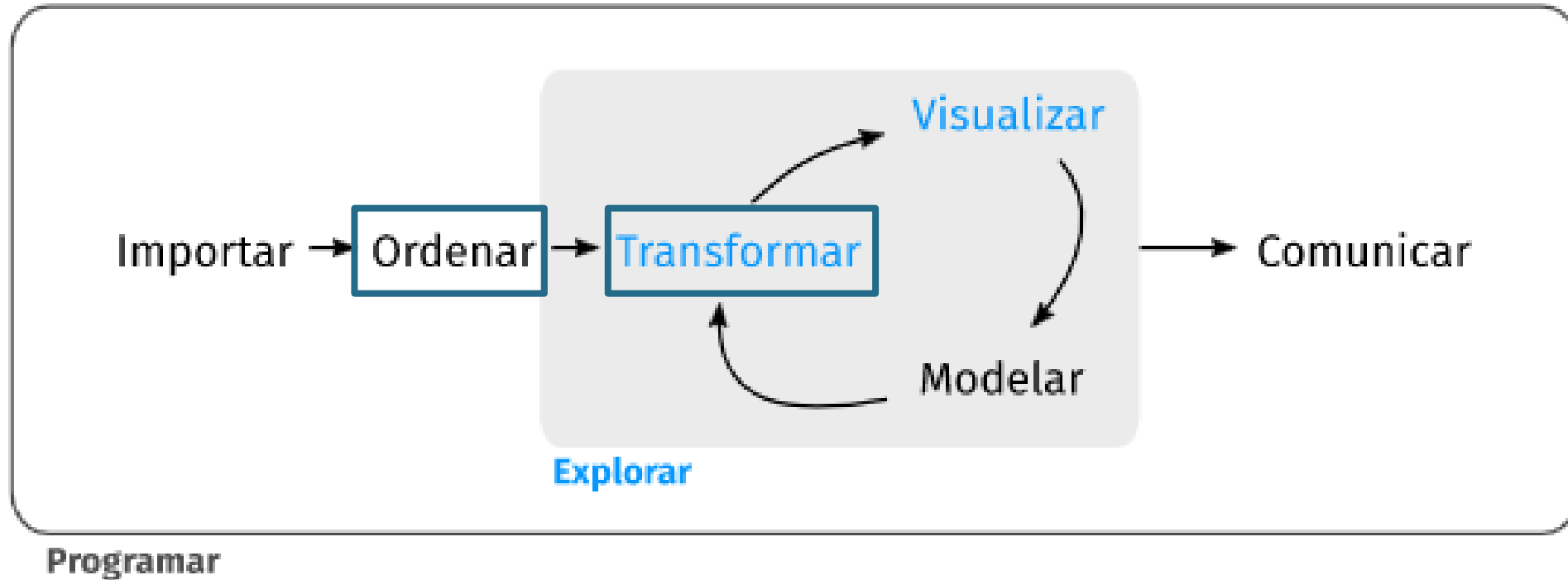
Mg. Yanina Bellini Saibene - INTA Anguil

Dra. María Florencia D'Andrea - IRB - CNIA

Datos Ordenados

Un lenguaje para ciencia de datos

Ordenar datos



Principios de Tidy Data

CULTIVAR	Días a floración	Altura (cm)	Vuelco (%)	Densidad (pl/ha)	Humedad de grano	Rendimiento de granos (kg/ha)	Aceite (%)
ACA 203 CL	85	181	0	48554	6.1	2719	43.6
ACA 861	85	166	0	47521	6.1	2319	51.8
ACA 869	87	189	3	45455	6.0	2300	54.0

Observación

Variable ó Atributo

1. Cada **variable** es una **columna**.
2. Cada **observación** es una **fila**.
3. Cada **tipo de unidad de observación** forma una **tabla**.

Síntomas comunes de datos desordenados

- ▶ Los encabezados de **columna** son **valores**, no nombres de variables.
- ▶ Múltiples **variables** se almacenan en una **columna**.
- ▶ Las **variables** se almacenan **tanto en filas como en columnas**.
- ▶ Múltiples tipos de **unidades de observación** se almacenan en la misma **tabla**.
- ▶ Una sola **unidad de observación** se almacena en **varias tablas**.

Todo bien, pero ¿y R?

Data Wrangling with dplyr and tidyr

Cheat Sheet



Syntax - Helpful conventions for wrangling

dplyr::tbl_df(iris)

Converts data to tbl class. tbl's are easier to examine than data frames. R displays only the data that fits onscreen:

```
Source: local data frame [150 x 5]
  Sepal.Length Sepal.Width Petal.Length
1           5.1           3.5           1.4
2           4.9           3.0           1.4
3           4.7           3.2           1.3
4           4.6           3.1           1.5
5           5.0           3.6           1.4
..
Variables not shown: Petal.Width (dbl),
Species (fctr)
```

dplyr::glimpse(iris)

Information dense summary of tbl data.

utils::View(iris)

View data set in spreadsheet-like display (note capital V).

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3.0	1.4	0.2	setosa
3	4.7	3.2	1.3	0.2	setosa
4	4.6	3.1	1.5	0.2	setosa
5	5.0	3.6	1.4	0.2	setosa
6	5.4	3.9	1.7	0.4	setosa
7	4.6	3.4	1.4	0.3	setosa
8	5.0	3.4	1.5	0.2	setosa

dplyr::%>%

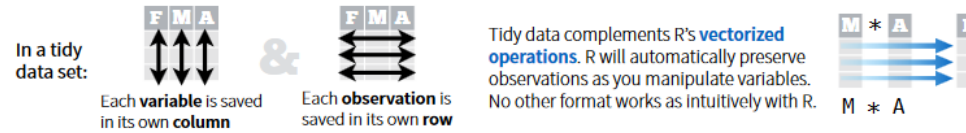
Passes object on left hand side as first argument (or argument) of function on righthand side.

$x \%>\% f(y)$ is the same as $f(x, y)$
 $y \%>\% f(x, , z)$ is the same as $f(x, y, z)$

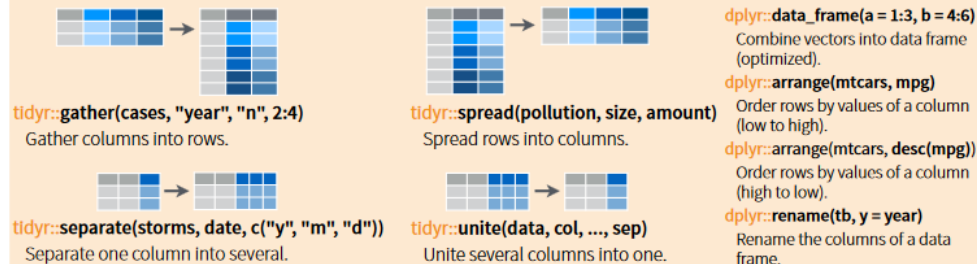
"Piping" with `%>%` makes code more readable, e.g.

```
iris %>%
  group_by(Species) %>%
  summarise(avg = mean(Sepal.Width)) %>%
  arrange(avg)
```

Tidy Data - A foundation for wrangling in R



Reshaping Data - Change the layout of a data set



Subset Observations (Rows)



dplyr::filter(iris, Sepal.Length > 7)

Extract rows that meet logical criteria.

dplyr::distinct(iris)

Remove duplicate rows.

dplyr::sample_frac(iris, 0.5, replace = TRUE)

Randomly select fraction of rows.

dplyr::sample_n(iris, 10, replace = TRUE)

Randomly select n rows.

dplyr::slice(iris, 10:15)

Select rows by position.

dplyr::top_n(storms, 2, date)

Select and order top n entries (by group if grouped data).

Subset Variables (Columns)



dplyr::select(iris, Sepal.Width, Petal.Length, Species)

Select columns by name or helper function.

Helper functions for select - ?select

select(iris, contains(" "))
Select columns whose name contains a character string.

select(iris, ends_with("Length"))
Select columns whose name ends with a character string.

select(iris, everything())
Select every column.

select(iris, matches("t"))
Select columns whose name matches a regular expression.

select(iris, num_range("x", 1:5))
Select columns named x1, x2, x3, x4, x5.

select(iris, one_of(c("Species", "Genus")))
Select columns whose names are in a group of names.

select(iris, starts_with("Sepal"))
Select columns whose name starts with a character string.

select(iris, Sepal.Length:Petal.Width)
Select all columns between Sepal.Length and Petal.Width (inclusive).

select(iris, -Species)
Select all columns except Species.

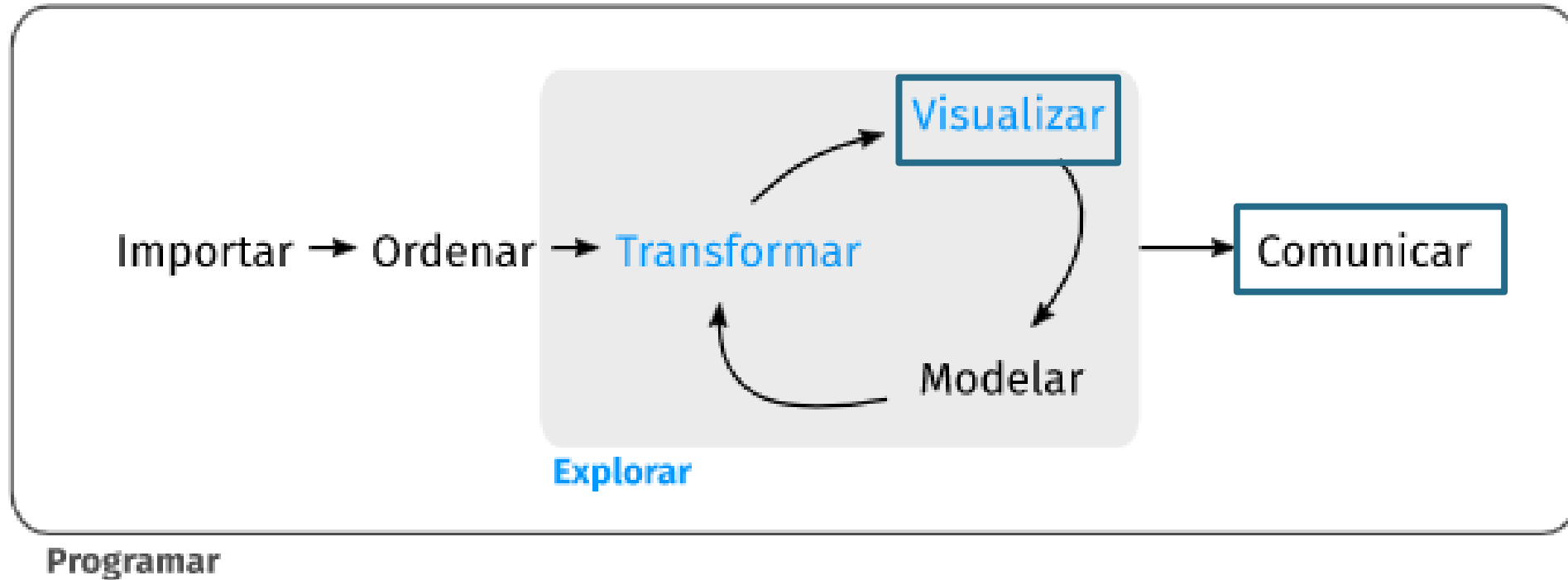
Fechas: **Lubridate**

<https://cran.r-project.org/web/packages/lubridate/vignettes/lubridate.html>

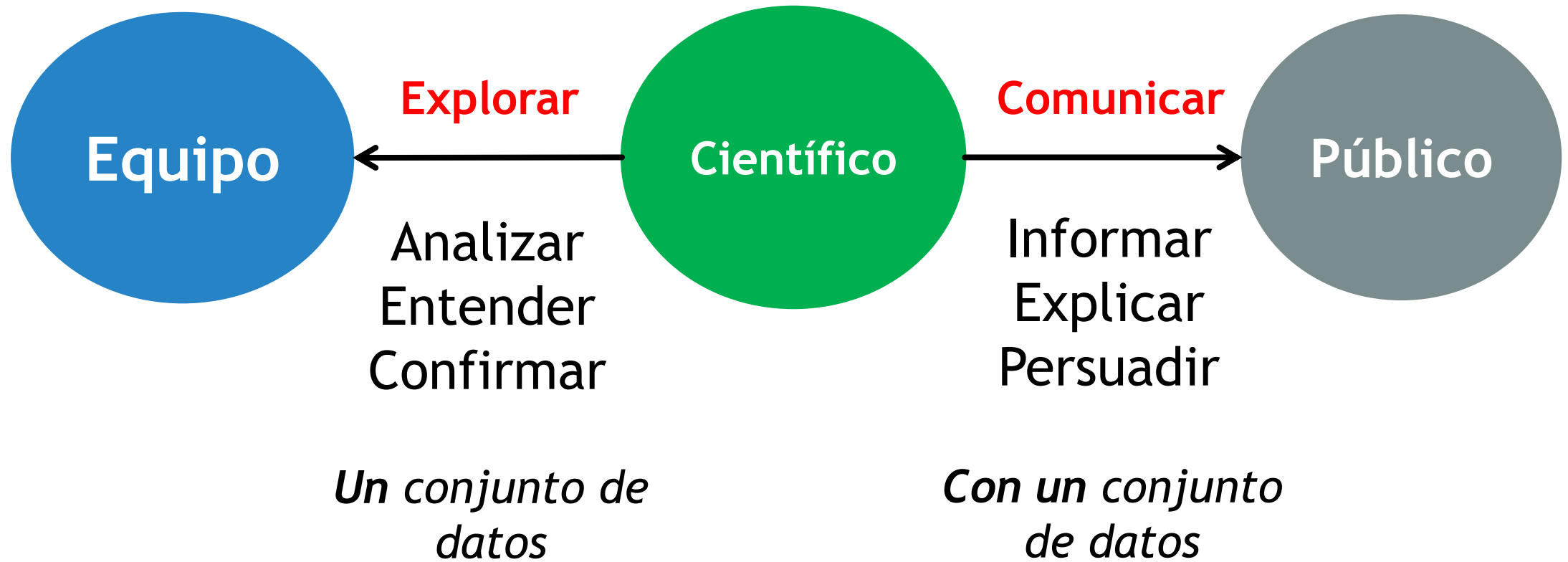
Datos Ordenados

Un lenguaje para ciencia de datos

Ordenar datos



Explorar vs Comunicar



Explorar



Taha Yasseri ✓

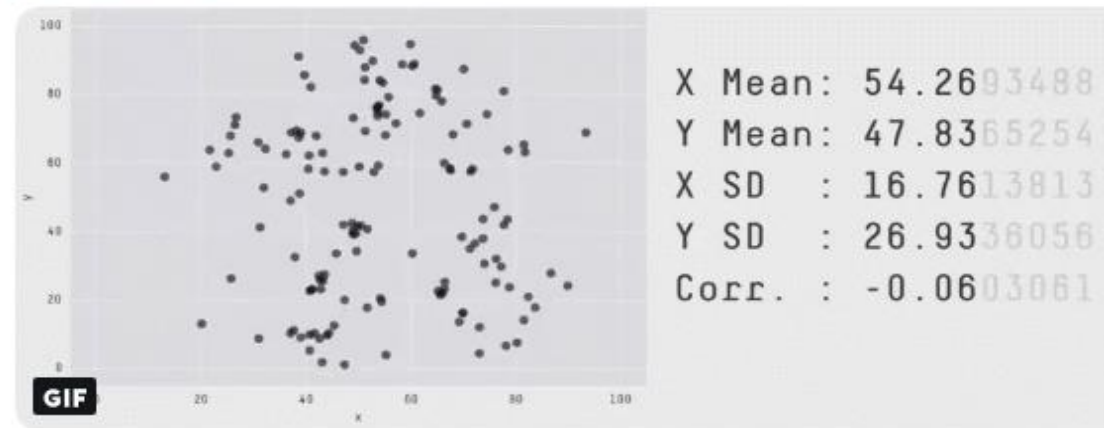
@TahaYasseri

Seguir



A great demonstration of why we need to plot the data and never trust statistics tables!
autodeskresearch.com/publications/s...

Traducir Tweet



13:37 - 1 may. 2017

9.757 Retweets **11.221** Me gusta



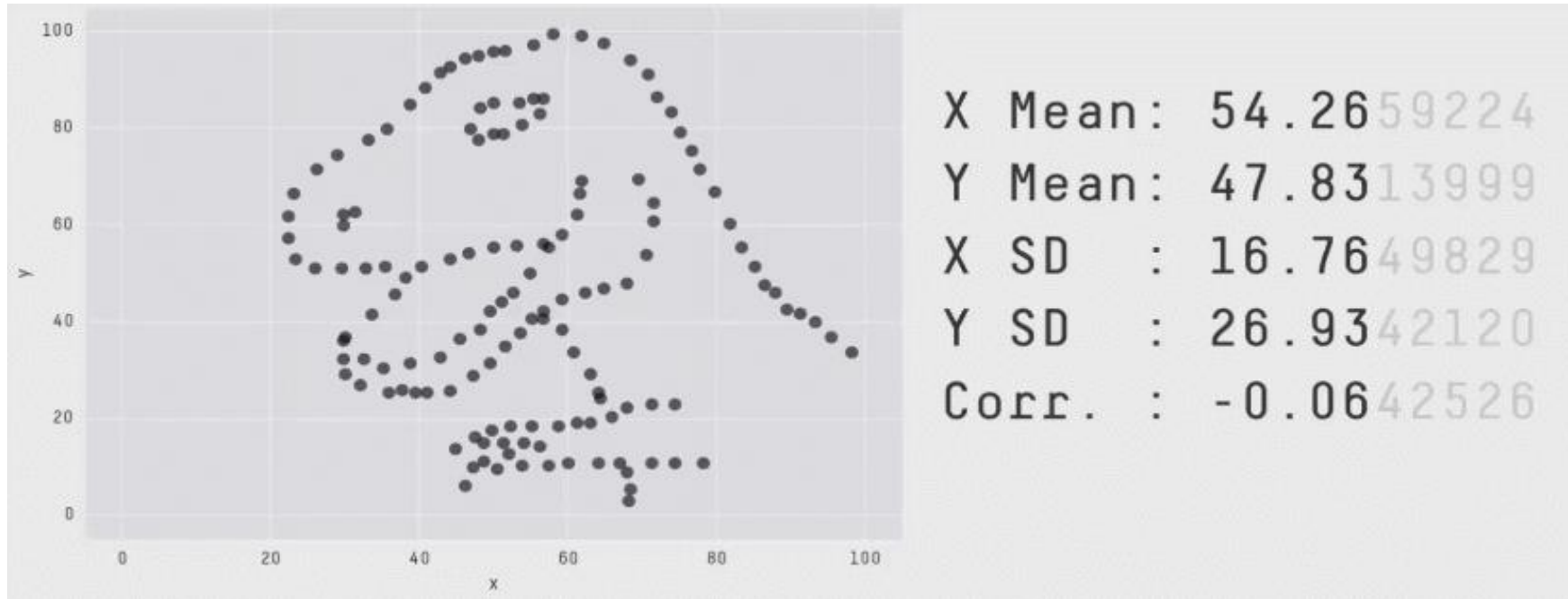
96

9,8K

11K

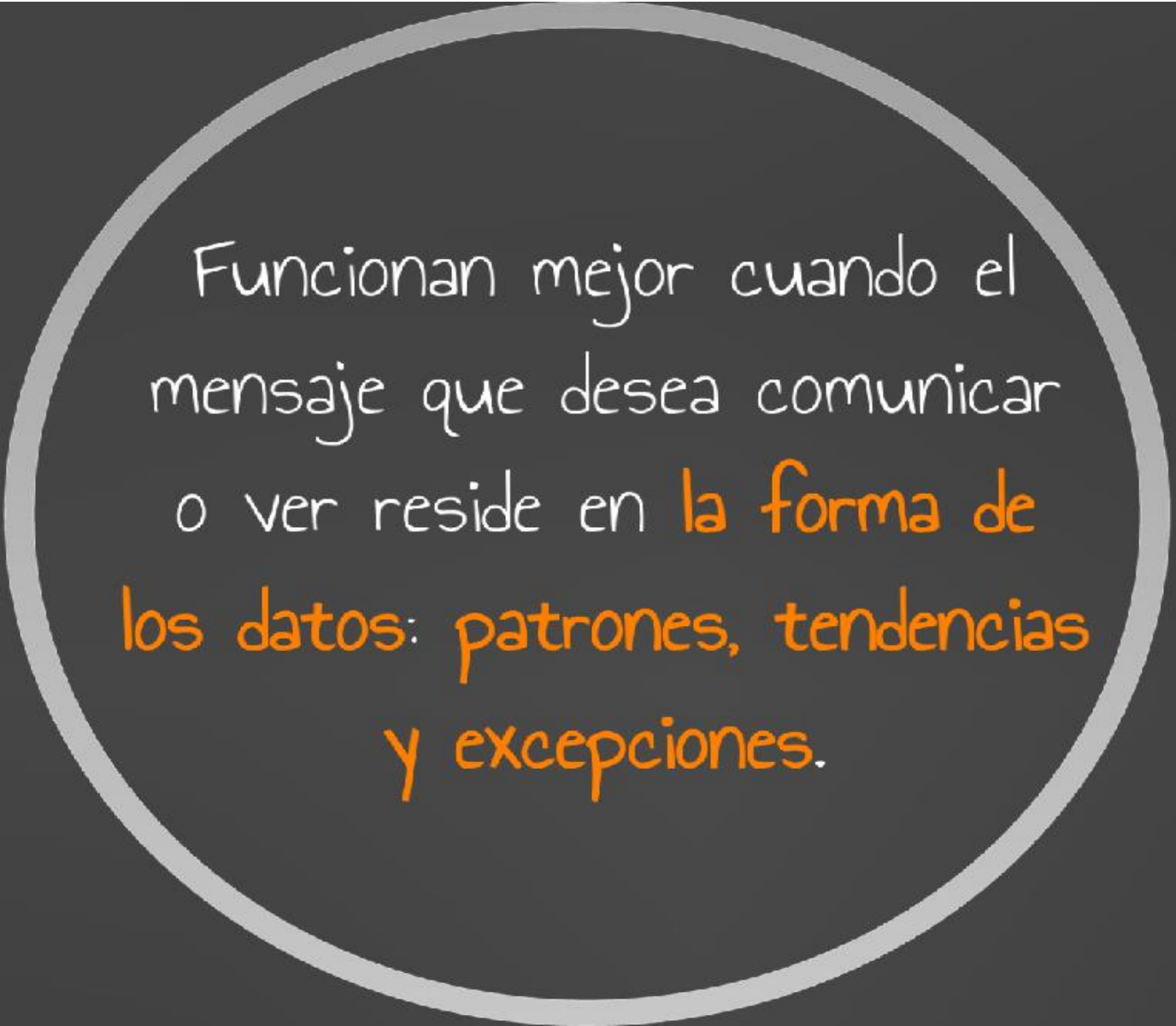


Explorar



La importancia de la visualización

Los gráficos....



Funcionan mejor cuando el
mensaje que desea comunicar
o ver reside en **la forma de
los datos: patrones, tendencias
y excepciones.**

Armando el gráfico correcto

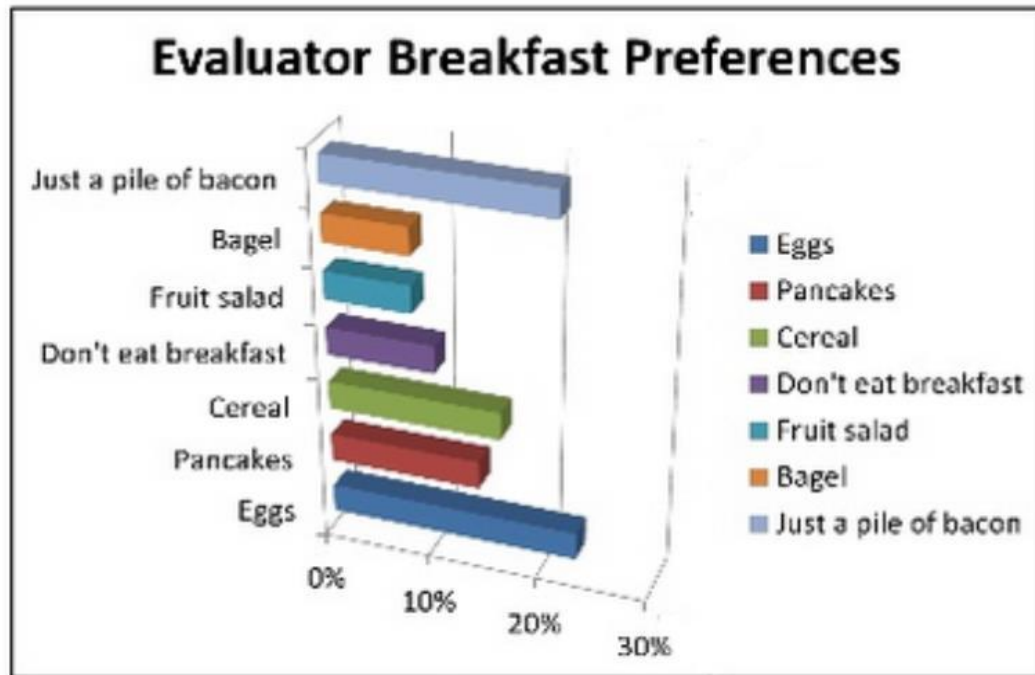
Paso 1: Identificar qué queremos comunicar

Paso 2: acomodar los datos de acuerdo al mensaje

Paso 3: preparar el gráfico.

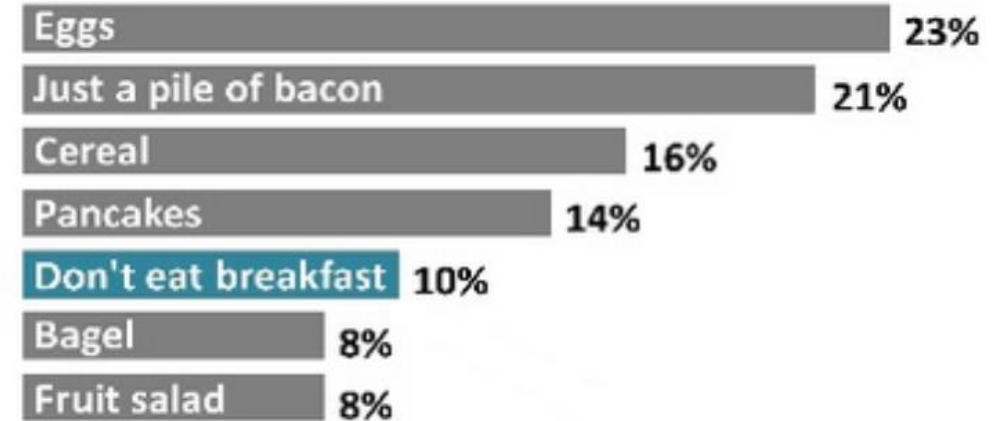
Paso 4: dar un formato que no distraiga.

Gráfico correcto, interpretación más rápida



Breakfast preferences focus on protein.

But 1 in 10 fellow evaluators do not consume adequate energy for their first meal of the day.



What kind of data do you have? Pick the main type using the buttons below. Then let the decision tree guide you toward your graphic possibilities.

Numeric

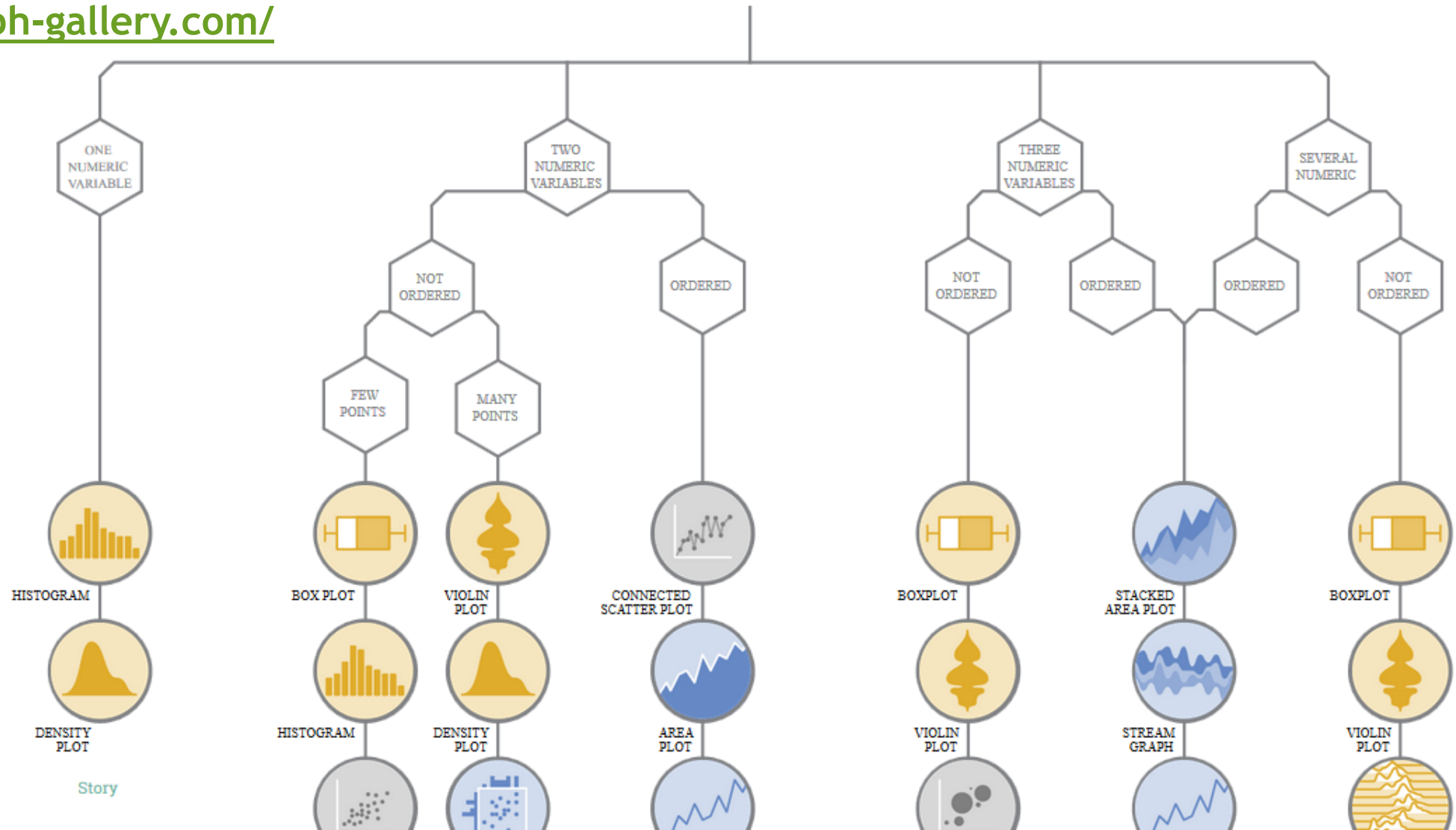
Categoric

Num & Cat

Maps

Network

Time series



La gramática de los gráficos

define un conjunto de reglas para construir gráficos estadísticos combinando diferentes tipos de capas.

La gramática nos dice que:



A statistical graphic is a **mapping** of **data** variables to **aesthetic** attributes of **geometric** objects.

Un gráfico estadístico es un **mapeo** de variables de **datos** a atributos **estéticos** de objetos **geométricos**.

La gramática nos dice que:

Específicamente, podemos **dividir un gráfico** en los siguientes **tres componentes esenciales**:

- 1. Datos (data):** el conjunto de datos compuesto por variables que mapeamos.
- 2. Geometría (geom):** el objeto geométrico en cuestión. Se refiere al tipo de objeto que compone el gráfico, por ejemplo: puntos, líneas y barras.
- 3. Estética (aes):** atributos estéticos del objeto geométrico. Por ejemplo, posición x / y, color, forma y tamaño. Cada atributo estético asignado se puede asignar a una variable en nuestro conjunto de datos.

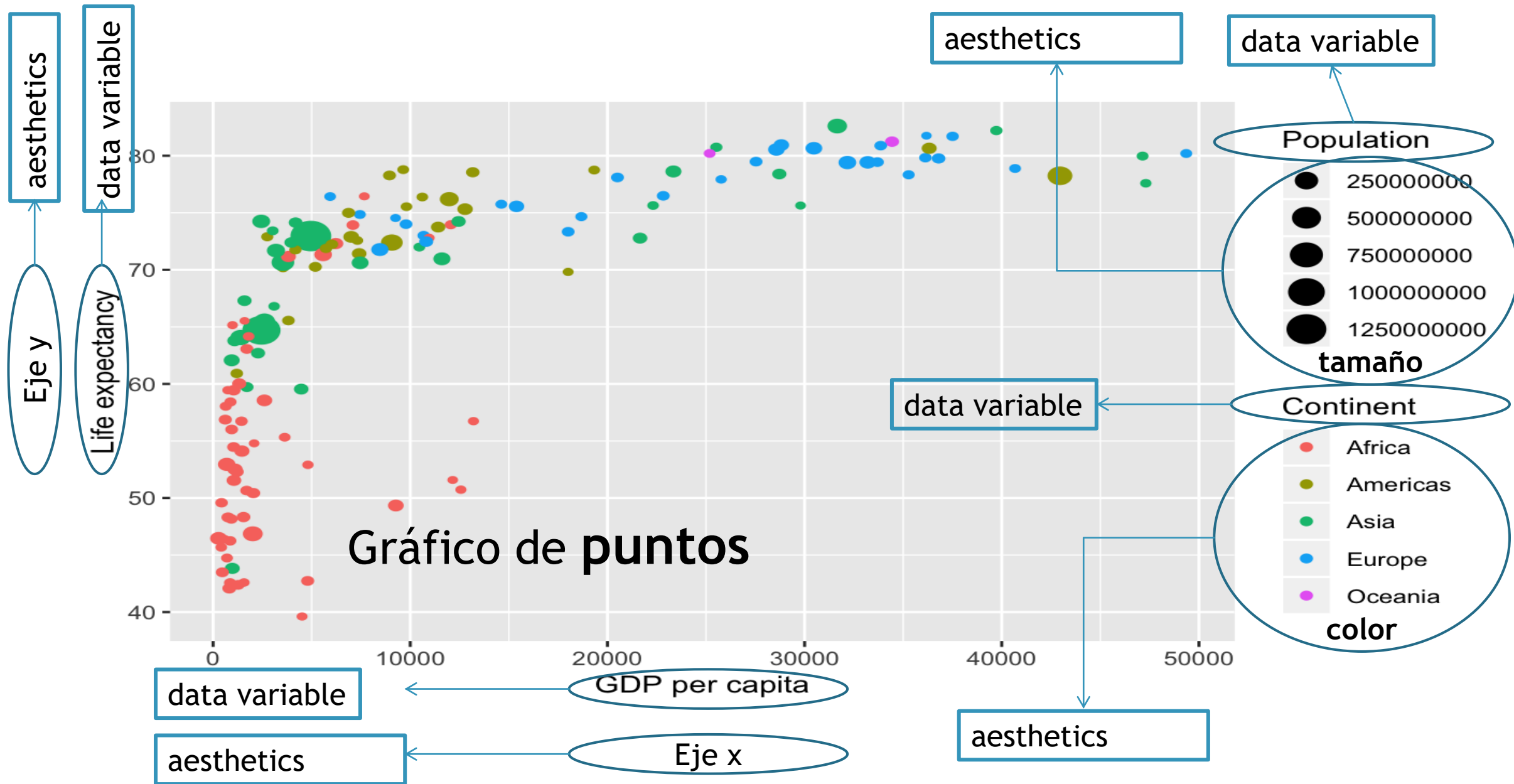
Ggplot es como un SIG, se pueden ir agregando capas a cada gráfico

Themes
Coordinates
Statistics
Facets
Geometries
Aesthetics
Data



Incluso varias capas del mismo tipo, por ejemplo dos capas **geometries**

<div>Nombre del país</div> <div>Nombre del continente</div>		<div>Expectativa De vida</div>	<div>Población</div>	<div>Producto Bruto Interno</div>
Country	Continent	Life Expectancy	Population	GDP per Capita
Afghanistan	Asia	43.8	31889923	975
Albania	Europe	76.4	3600523	5937
Algeria	Africa	72.3	33333216	6223
Angola	Africa	42.7	12420476	4797
Argentina	Americas	75.3	40301927	12779
Australia	Oceania	81.2	20434176	34435



Gramática del gráfico

data variable	aes	geom
GDP per Capita	x	point
Life Expectancy	y	point
Population	size	point
Continent	color	point

```
ggplot(data = gapminder,  
       mapping = aes(x = gdpPercap, y = lifeExp,  
                     color=continent, size=pop))+  
  geom_point()
```

Opciones en las capas de un gráfico

Elemento	Posibles valores				
Data	Variables de interés				
Aesthetics	Eje x Eje y	colour Fill	size labels	alpha shape	with type
Geometries	point	histogram	line	bar	boxplot
Facets	columns	rows			
Statistics	binning	smoothing	descriptive	inferential	
Coordinates	cartesian	polar	fixed	limits	
Themes	Configuración de diversos aspectos del gráfico				

¡Manos a la obra!

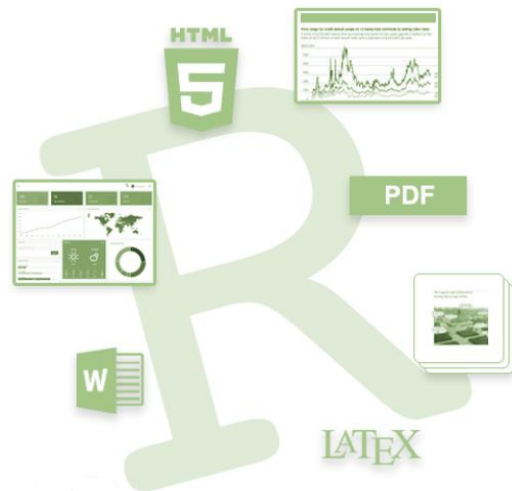
R_inta_LC3_2019.R

- ▶ Gráfico de barras de los institutos
- ▶ Gráfico de barras de asistencia al taller
- ▶ Gráfico de barras apiladas de uso de herramientas
- ▶ Uso de estilo BBC.

¿Qué es RMarkdown (Rmd)?

Lenguaje de marcado que integra texto, código R y resultados.

RMarkdown permite generación de informes, presentaciones, páginas web, tesis, libros, poster....



RMarkdown

Un documento -> Varias salidas



Partes de un Archivo .Rmd

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

YAML

texto

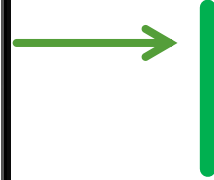
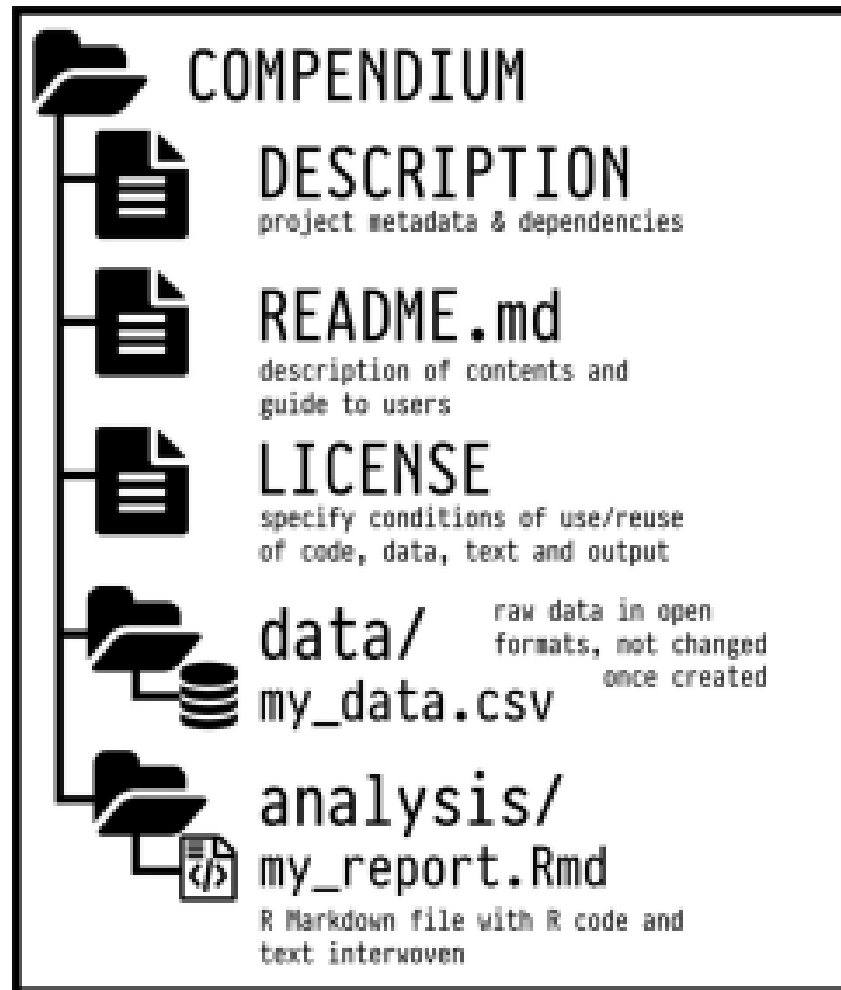
Código (chunk)

¡Manos a la obra!

- ▶ Gráfico en ggplot
- ▶ Obtener un reporte con RMarkdown

Compendio

forma de **organizar los materiales digitales** de un proyecto para permitir que otros reproduzcan y extiendan la investigación



Se obtiene un DOI y se
Se puede citar.



Search



Upload

Communities

Log in

Sign up

Recent uploads

August 8, 2019 (vv3)

Dataset

Open Access

View

Analysis of correlation-based biomolecular networks from different omics data by fitting stochastic block models

Baum, Katharina; Rajapakse, Jagath C.; Azuaje, Francisco

Baum_et_al_2019_Supplementary_Figures.pdf: Supplementary Figures S1-S4. Legends are included under each figure.
sbm-for-correlation-based-networks-master.zip: Archived source code of R and Python functions for the analyses and example workflow description at time of publication. Files are...

Uploaded on August 8, 2019

2 more version(s) exist for this record

August 6, 2019 (v3.1)

Software

Open Access

View

OpenScienceMOOC/Module-5-Open-Research-Software-and-Open-Source: 3.1

Jon Tennant; Julien Colomb; Lisa Matthias; Simon Worthington; Florian Kohrt; irrubio; Tania Allard; Philipp Zumstein; Daniel S. Katz; Alexander Morley; Tobias Steiner; Stephan Druskat; Zoran Pandovski; Arfon Smith; Gabriele Orlandi; Rutger Vos; José Raúl Canay Pazos; Paul Griffiths; Nithiya Streethran; Hollie Marshall; Luke W Johnston; Luis Camacho; Konrad Förstner; Heidi Seibold; Eric Wilhelm; Esmeralda Martínez Álvarez; Brandon Delmon; Alessandro Carretta; Alberto...

Zenodo now supports usage statistics!



[Read more](#) about it, in our newest blog post.

Using GitHub?



Just [Log in](#) with your GitHub account and [click here](#) to start preserving your repositories.

Zenodo in a nutshell

- **Research. Shared.** — all research outputs from across all fields of research are welcome! Sciences and Humanities, really!
- **Citeable. Discoverable.** — uploads get a

Comunidades

Links en página web https://flor14.github.io/cai_2019

R-Ladies

¡Sumate a nuestra comunidad!



/RLadiesBA



R Ladies
Buenos Aires



@RLadiesBA



/RLadies-
Buenos-Aires



/RLadies-BA



129

R-Ladies groups on meetup.com

40

R-Ladies Countries

129

R-Ladies Cities

30904

R-Ladies members on meetup.com



<https://gqueiroz.shinyapps.io/rshinylady/>

Ggplot2
Ciencia de Datos
Taller de
Casos

R en Buenos Aires

<https://renbares.github.io/>



R-Spatial ES

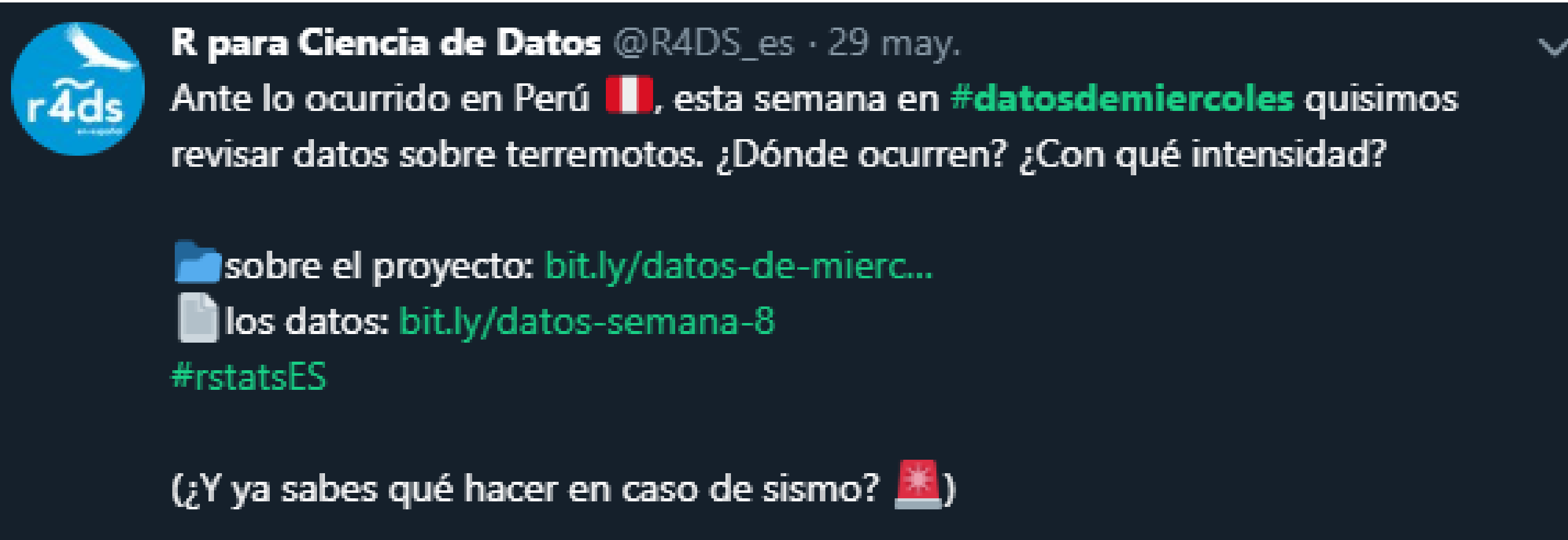
- ▶ Datos espaciales con R

R4DS_ES

- ▶ Traducción del libro R4DS al español
- ▶ #datosdemiercoles



#datosdemiercoles



R para Ciencia de Datos @R4DS_es · 29 may.

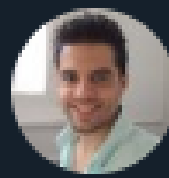
Ante lo ocurrido en Perú 🇵🇪, esta semana en **#datosdemiercoles** quisimos revisar datos sobre terremotos. ¿Dónde ocurren? ¿Con qué intensidad?

📁 sobre el proyecto: bit.ly/datos-de-mierc...

📄 los datos: bit.ly/datos-semana-8

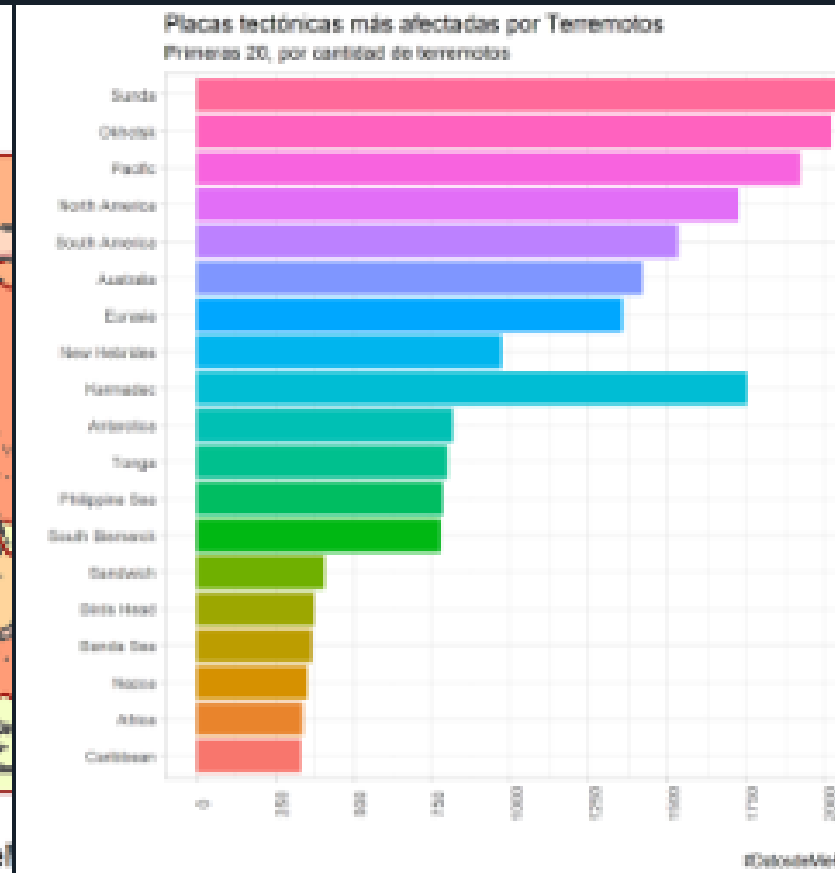
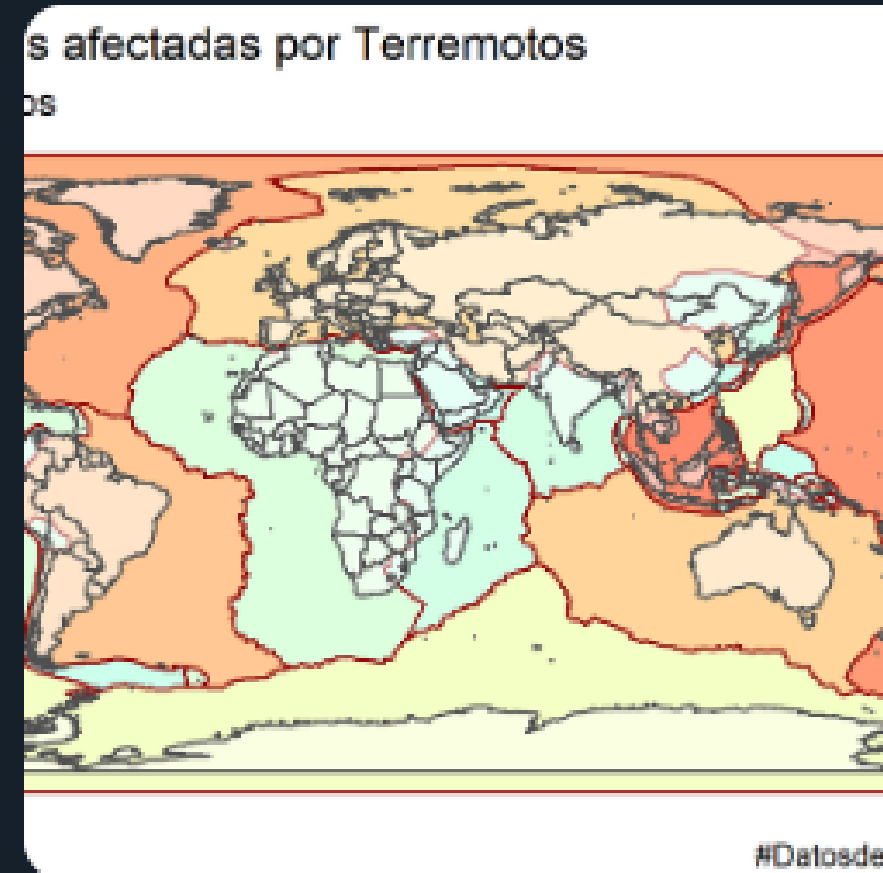
#rstatsES

(¿Y ya sabes qué hacer en caso de sismo? 🚨)



Julio Spairani @jspairani · 2 jun.

2019-05-28 #DatosdeMieRcoles #rstats_ES Desafío Terremotos! inspirado en el post de placas de @violetrzn , Quise ver que placas eran las que más terremotos tienen asociados. Va barchart complementario para ver los nombres. 😊
para contexto: youtube.com/watch?v=T2WqVj...



2



1



15





Julio Spairani @jspairani · 2 jun.

algunos recursos: 📖

- para datos de placas use estos que ya traian las placas como polígonos y no como líneas: github.com/fraxen/tectoni...
- para ver puntos en poligonos: spatialEco, referido de aca: stackoverflow.com/questions/3647...
- paleta de colores inspirada en:



Summer Sunset at the lake Color Palette

color-hex.com



violeta ❤️ @violetrzn · 2 jun.

En respuesta a [@jspairani](#)

copado! me gustó

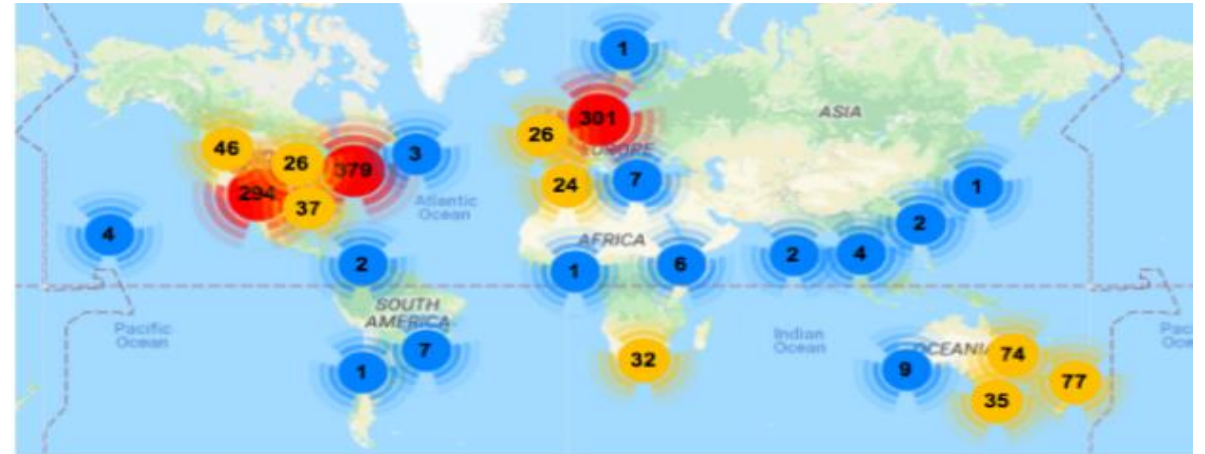




We teach foundational coding and data science skills to researchers worldwide.

Somos una **comunidad global** que enseña habilidades básicas de computación y ciencia de datos a investigadores en

- el mundo académico,
- la industria
- y el gobierno.



<https://twitter.com/thecarpentries>

<https://carpentries.org>

Organización sin fines de lucro

- > Hay gran **demanda de entrenamiento** en habilidades básicas de programación
- > Los libros de texto de ingeniería de software no son apropiados para **enseñar a programar** a la mayoría de los científicos.

Software Carpentry



Los materiales de aprendizaje se encuentran abiertos y disponibles

<http://swcarpentry.github.io/r-novice-gapminder/>



Teaching basic lab skills
for research computing

R para Análisis Científicos Reproducibles

El objetivo de esta lección es enseñar a las programadoras principiantes a escribir códigos modulares y adoptar buenas prácticas en el uso de R para el análisis de datos. R nos provee un conjunto de paquetes desarrollados por terceros que se usan comúnmente en diversas disciplinas científicas para el análisis estadístico. Encontramos que muchos científicos que asisten a los talleres de Software Carpentry utilizan R y quieren aprender más. Nuestros materiales son relevantes ya que proporcionan a los asistentes una base sólida en los fundamentos de R y enseñan las mejores prácticas del cómputo científico: desglose del análisis en módulos, automatización tareas y encapsulamiento.

Ten en cuenta que este taller se enfoca en los fundamentos del lenguaje de programación R y no en el análisis estadístico.

A lo largo de este taller se utilizan una variedad de paquetes desarrollados por terceros, los cuales no son necesariamente los mejores ni se encuentran explicadas todas sus funcionalidades, pero son paquetes que consideramos útiles y han sido elegidos principalmente por su facilidad de uso.

Data Carpentry

<https://datacarpentry.org/lessons/>



Curriculum materials

- [Ecology curriculum](#)
- [Genomics curriculum](#)
- [Social Sciences curriculum](#)
- [Geospatial data curriculum](#)

Eventos



<http://latin-r.com/>

2da Conferencia Latinoamericana sobre Uso de R en Investigación + Desarrollo



2018 – Buenos Aires / 2019 – Santiago de Chile



Conferencia Latinoamericana sobre
el Uso de R en Investigación + Desarrollo

25 - 27 DE SEPTIEMBRE | 2019
SANTIAGO DE CHILE

Keynote Speakers



Mine Çetinkaya-Rundel



Hadley Wickham



¡Muchas gracias!