



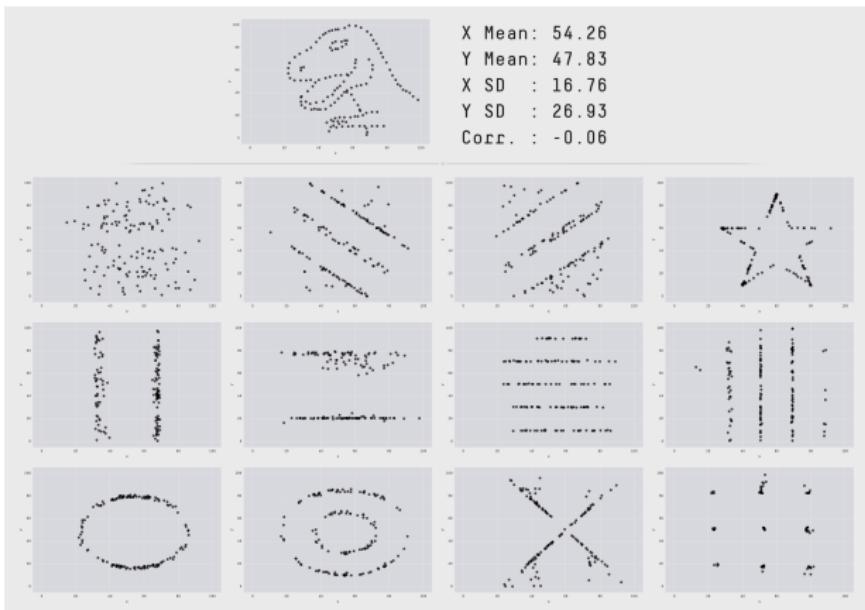
BUENAS PRÁCTICAS DE VISUALIZACIÓN DE DATOS EN INVESTIGACIÓN CON R

Lic. Joaquín Ferreyra - Lic. Natalia Labadie

Departamento de Matemática y Estadística. Facultad de Ciencias Bioquímicas y Farmacéuticas (Universidad Nacional de Rosario) - Instituto de Química Rosario
(UNR - CONICET)

¿Por qué visualizar? El Datasaurus Dozen¹

El cálculo de **estadísticas** no nos permite contar la historia completa de los datos con los que trabajamos.



¹Matejka, J.; Fitzmaurice, G. (2017). **Same Stats, Different Graphs: Generating Datasets with Varied Appearance and Identical Statistics through Simulated Annealing.**

¿Por qué visualizar?

- Las herramientas de visualización ayudan a poner en evidencia información relevante sobre un dataset, facilitando la detección de **tendencias, patrones, valores atípicos y correlaciones entre variables**.
- La visualización de datos es más que la representación gráfica de información. Es una **forma de comunicación visual** orientada a **generar conocimiento** acerca de los datos.

¿Qué hace que un gráfico sea un mal gráfico?²

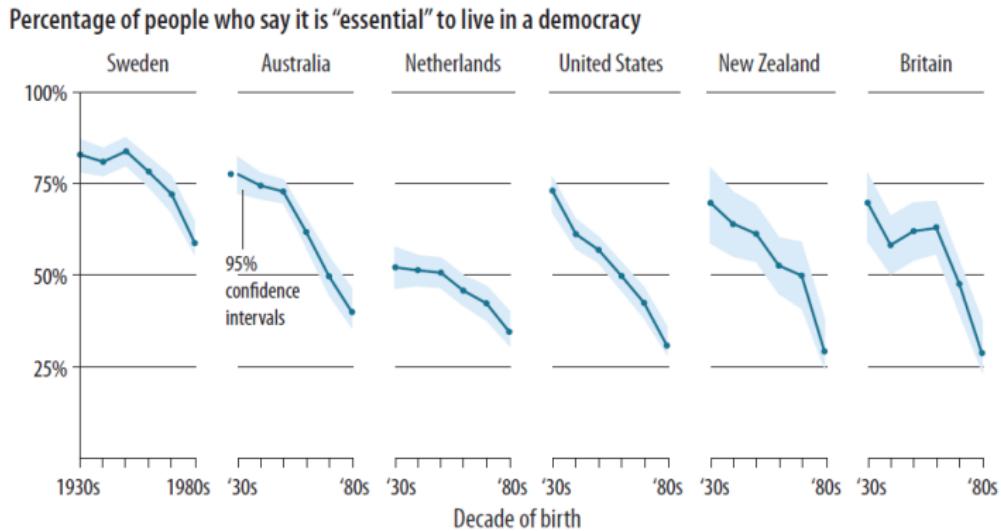
PROBLEMAS ESTÉTICOS: una combinación de mal gusto y un diseños pobres e inconsistentes.



²Ideas extraídas de: Healy, K. (2019). **Data Visualization. A practical introduction.** Princeton University Press.

¿Qué hace que un gráfico sea un mal gráfico?

PROBLEMAS CON LOS DATOS REPRESENTADOS: un mal uso de los datos con los que se cuenta.

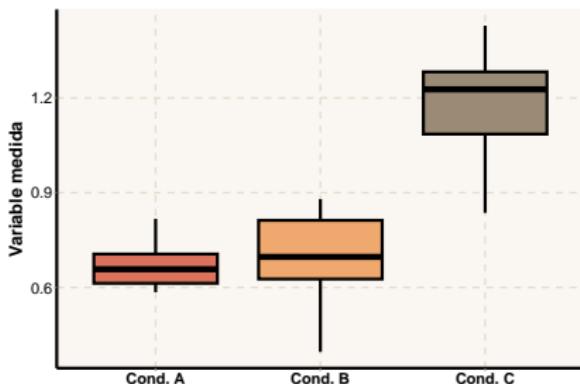
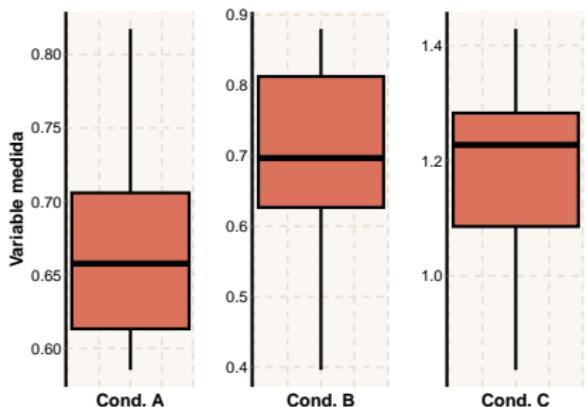


“How Stable Are Democracies? Warning Signs Are Flashing Red”
(New York Times)

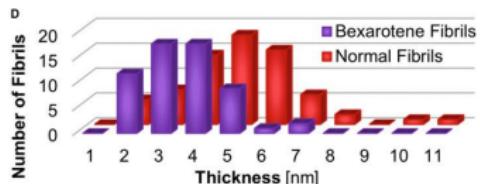
¿Qué hace que un gráfico sea un mal gráfico?

PROBLEMAS PERCEPTUALES: una mala utilización de herramientas visuales que produce gráficos confusos y engañosos.

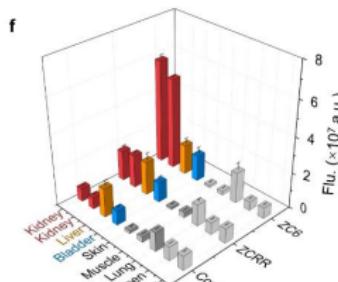
A pesar de estar construidos con los mismos datos, estos dos gráficos no parecen mostrar lo mismo:



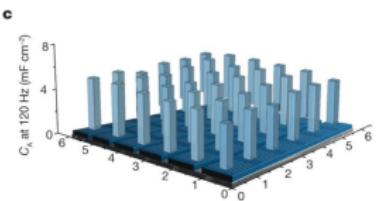
Pero... ¿hay malos gráficos en los papers?



J. Biol. Chem. 298 (12), 102662 (2022).

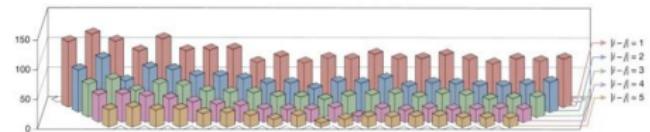


Angew. Chem. Int. Ed. 2023, e202315457.



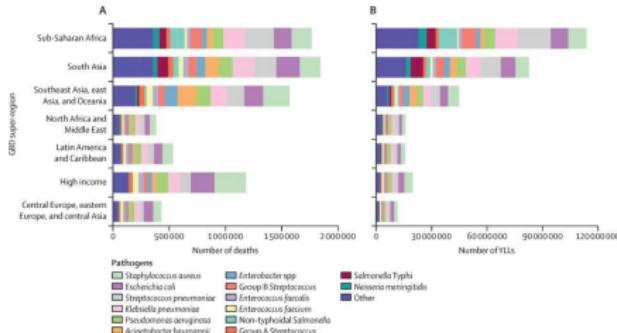
Nature (2023).

<https://doi.org/10.1038/s41586-023-06712-2>

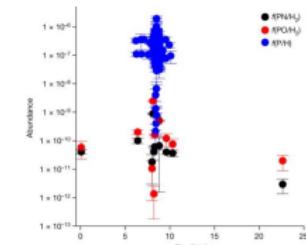
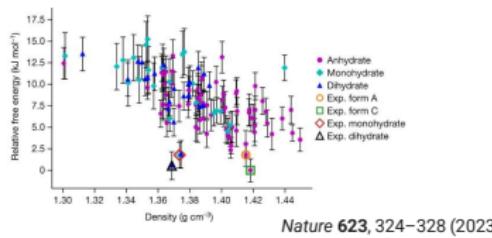


Nature 623, 713–717 (2023)

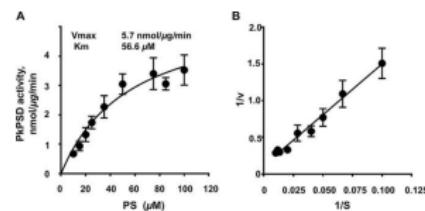
Pero... ¿hay malos gráficos en los papers?



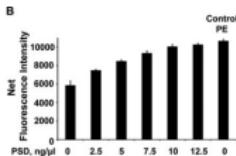
The Lancet 400 (10369), 2221-2248 (2022).



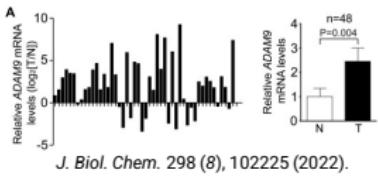
Nature 623, 292–295 (2023)



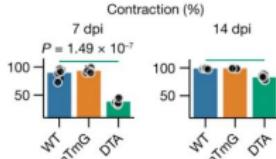
Pero... ¿hay malos gráficos en los papers?



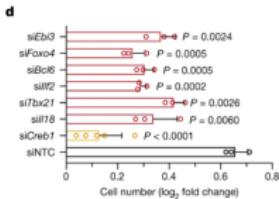
J. Biol. Chem. 293 (5), 1493–1503 (2018).



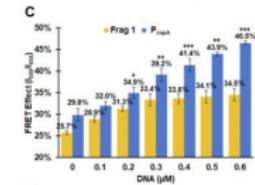
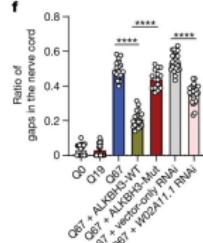
J. Biol. Chem. 298 (8), 102225 (2022).



Nature 623, 792–802 (2023)



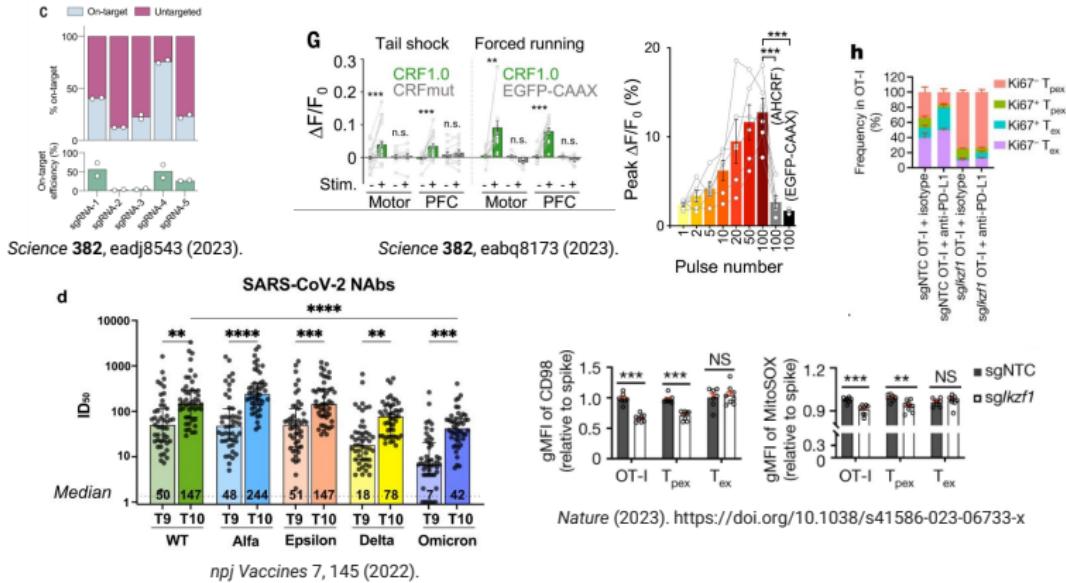
Nature (2023). <https://doi.org/10.1038/s41586-023-06749-3>



J. Biol. Chem. 298 (5), 102225 (2022).

Nature 623, 580–587 (2023).

Pero... ¿hay malos gráficos en los papers?



Barplot with error bars - a.k.a. gráfico dinamita

Gráfico... ¿dinamita?



Barplot with error bars - a.k.a. gráfico dinamita

Gráfico... ¿dinamita?



Barplot with error bars - a.k.a. gráfico dinamita

Gráfico... ¿dinamita?

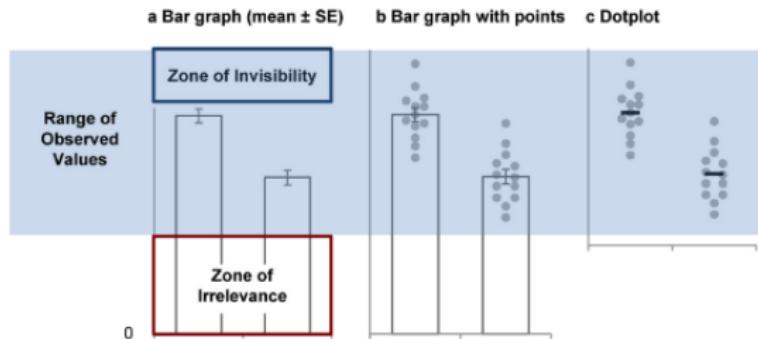


Redflags del gráfico dinamita

Redflag #1 Los gráficos de barras están diseñados para representar **conteos y proporciones** relacionados al trabajo con **variables categóricas**. No obstante, continúan siendo una estrategia ampliamente aceptada para presentar información sobre variables cuantitativas.

Redflags del gráfico dinamita

Redflag #2 En un gráfico de dinamita se reconocen **dos regiones** que, potencialmente, pueden generar confusión y malas interpretaciones: una **región de irrelevancia** y una **región de invisibilidad**.



Adaptado de Weissberger *et al.* (2017)³.

³Weissberger, T. L., Savic, M. *et al.* (2017) **Data visualization, bar naked: A free tool for creating interactive graphics.** *J. Biol. Chem.* 292(50) 20592-20598.

Redflags del gráfico dinamita

Redflag #3 Si nos quedamos únicamente con una representación gráfica de la media y el desvío estándar de cada conjunto, nos estamos perdiendo **muchísima** información valiosa sobre nuestros datos y la posibilidad de evaluarlos críticamente.

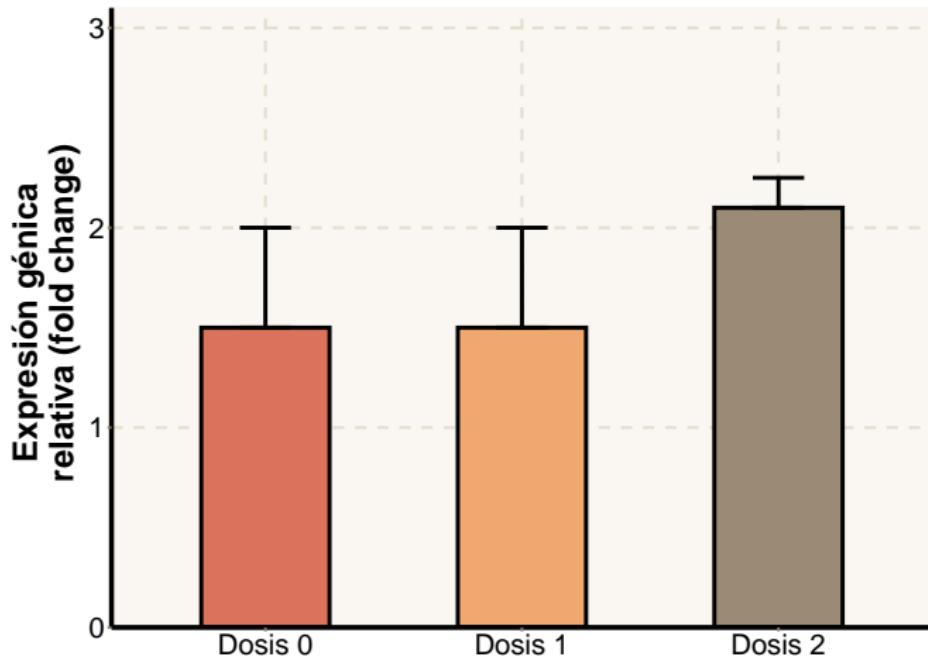
Primer ejemplo - Conjuntos de pocos datos

Tomando como ejemplo un trabajo cuyo objetivo sea analizar el efecto de tres dosis de una droga sobre el **nivel de expresión génica** de un gen de interés, se modelaron tres conjuntos de 5 observaciones:

Dosis 0	Dosis 1	Dosis 2
1.97	1.99	1.97
1.43	1.63	2.32
0.94	0.68	2.16
2.05	1.46	1.96
1.10	1.74	2.09

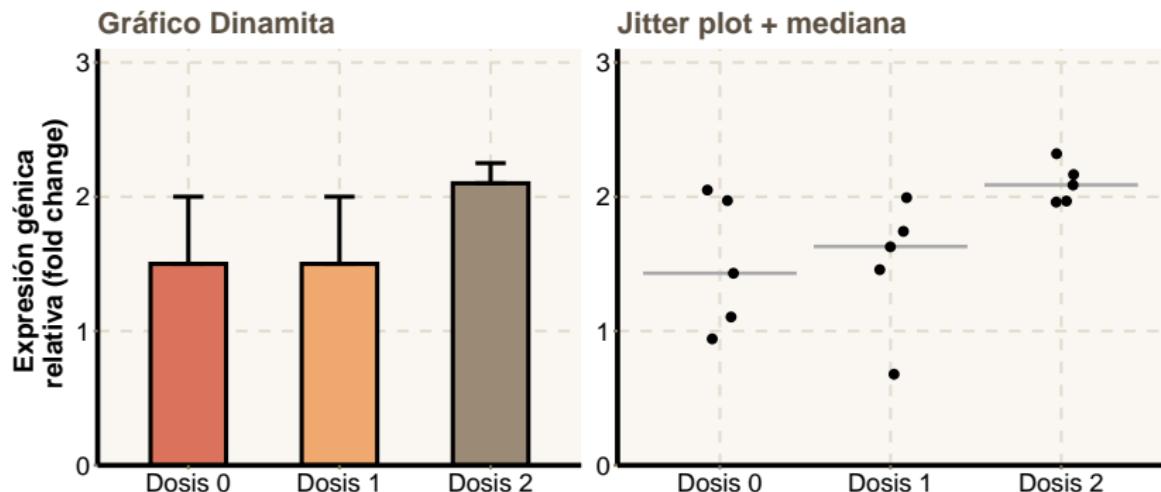
Primer ejemplo - Conjuntos de pocos datos

Supongamos que decidimos representar estos datos a través de un gráfico dinamita. **¿Qué información nos aportaría?**



Primer ejemplo - Conjuntos de pocos datos

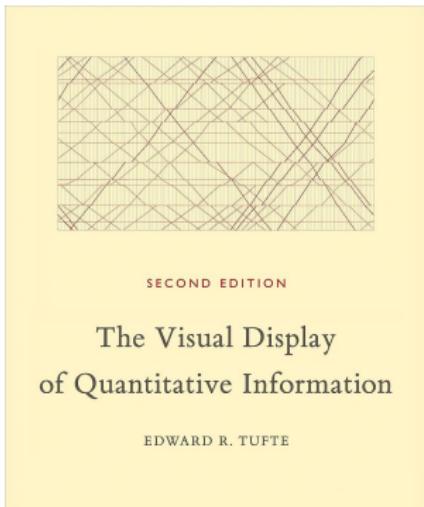
Cuando se tienen pocos datos, una alternativa al gráfico dinamita es representar las observaciones individuales (jitter plot) y agregar la mediana en forma de línea horizontal (o la media según se considere más adecuado).



Reglas de Tufte

“Un gráfico excelente es aquel que da al espectador el mayor número de ideas en el menor tiempo con la menor tinta en el espacio más pequeño.” (Edward Tufte)

- Ante todo, **mostrar los datos**.
- Maximizar la **proporción datos/tinta** (proporción de tinta de datos vs. tinta total de la gráfica).
- Minimizar el uso de decoración gráfica innecesaria (*chartjunk*).
- Evitar la distorsión de los datos.
- Suele ser mejor tener una mayor **densidad de datos** en el gráfico.



Primer ejemplo - Conjuntos de pocos datos

Los datasets se pueden generar de forma sencilla con funciones del paquete base de R.

En primer lugar, construimos algunas funciones para facilitar la adaptación a múltiples ejemplos:

```
# scaling(): Escalar el set de datos para que tenga la media
# y el desvío estándar deseados
scaling <- function(set, desired.mean, desired.sd) {
  desired.sd * (set - mean(set)) / sd(set) + desired.mean
}

# outlier(): Genera datos provenientes de una distribución normal
# estándar y les agrega 1 outlier.
outlier <- function(n) {
  d <- rnorm(n)
  k <- sample(c(1, -1), 1)*runif(1, 3, 6)
  d[sample(1:n, 1)] <- 0 + k*1

  return(d)
}
```

Primer ejemplo - Conjuntos de pocos datos

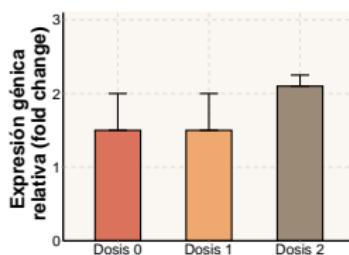
Luego utilizamos las funciones para simular los tres sets de datos:

```
# Simulación de los 3 sets de datos
set.seed(2019) # fijar semilla
setA <- scaling(rnorm(5), desired.mean = 1.5, desired.sd = 0.5)
setB <- scaling(outlier(5), desired.mean = 1.5, desired.sd = 0.5)
setC <- scaling(rnorm(5), desired.mean = 2.1, desired.sd = 0.15)

# Creación y procesamiento del dataframe
datos1 <- data.frame(setA, setB, setC) %>%
  mutate(id = 1:5) %>%
  pivot_longer(cols = starts_with("set"),
               names_to = "set",
               names_prefix = "set",
               values_to = "fold_change") %>%
  mutate(set = factor(set, levels = c("A", "B", "C")),
         labels = c("Dosis 0", "Dosis 1", "Dosis 2"))
```

Primer ejemplo - Conjuntos de pocos datos

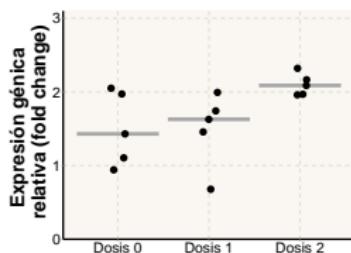
Código de R para la construcción del **gráfico dinamita** empleando funciones de **ggplot2**:



```
dinamite_plot <- datos1 %>%
  group_by(set) %>%
  summarise(mean = mean(fold_change),
            sd = sd(fold_change)) %>%
  ggplot(aes(y = mean, x = set, fill = set)) +
  geom_bar(stat = 'identity',
           position = position_dodge(.9)) +
  geom_errorbar(aes(ymin = mean,
                    ymax = mean + sd),
                width = .2,
                position = position_dodge(.9))
```

Primer ejemplo - Conjuntos de pocos datos

Código de R para la construcción del **Jitter plot + mediana** empleando funciones de ggplot2:

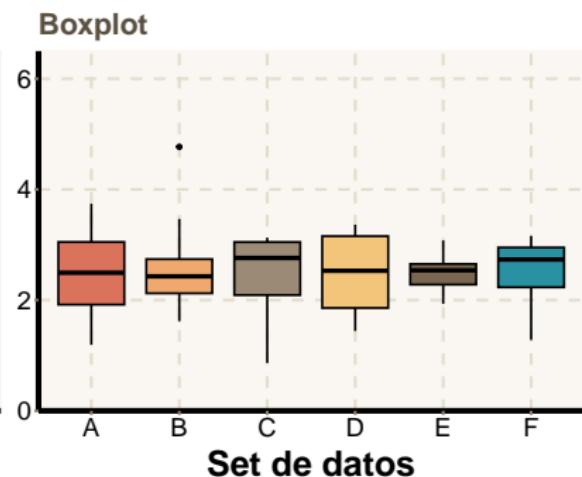
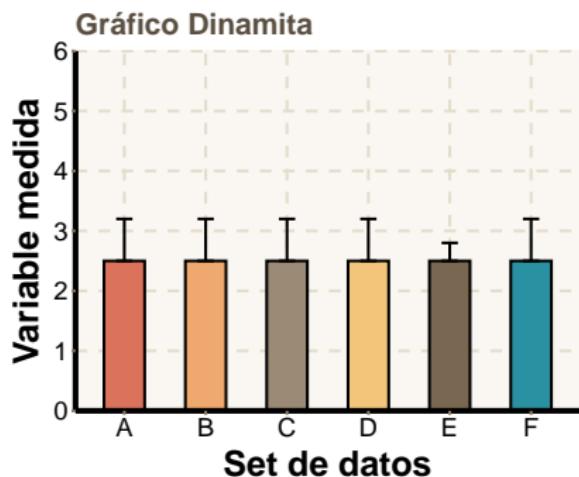


```
jitter_plot <- datos1 %>%
  ggplot(aes(x = set, y = fold_change)) +
  stat_summary(fun = median,
              show.legend = FALSE,
              geom = "crossbar",
              color = "darkgray") +
  geom_jitter(shape = 16,
              position = position_jitter(0.1),
              size = 3.5,
              colour = "black")
```

Segundo ejemplo - Datasets de mayor extensión

Se simularon seis conjuntos de datos considerando el caso de contar con observaciones de una variable cuantitativa para seis niveles de un factor de interés.

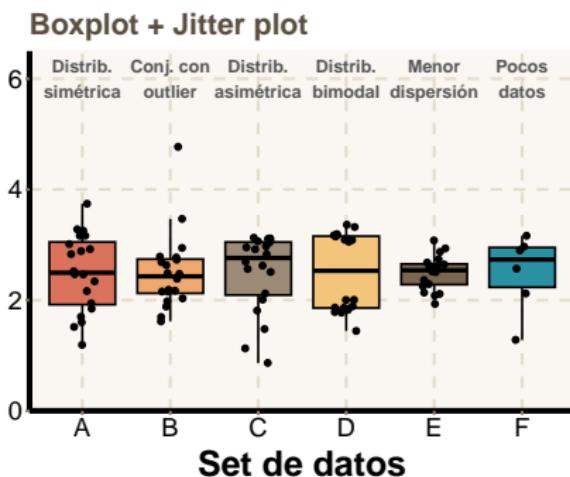
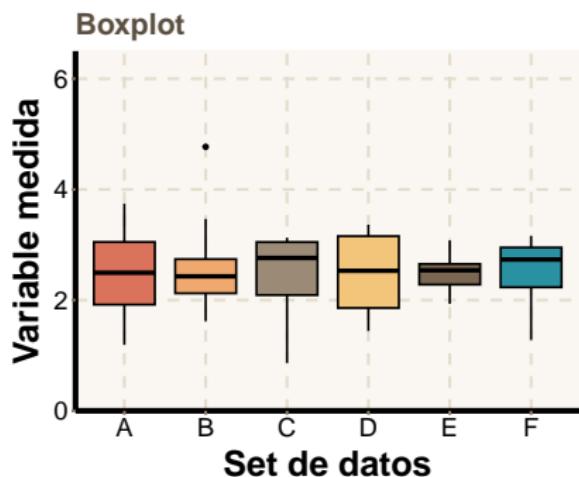
¿Cuál es el gráfico más apropiado para representar estos datos?



Segundo ejemplo - Datasets de mayor extensión

Se simularon seis conjuntos de datos considerando el caso de contar con observaciones de una variable cuantitativa para seis niveles de un factor de interés.

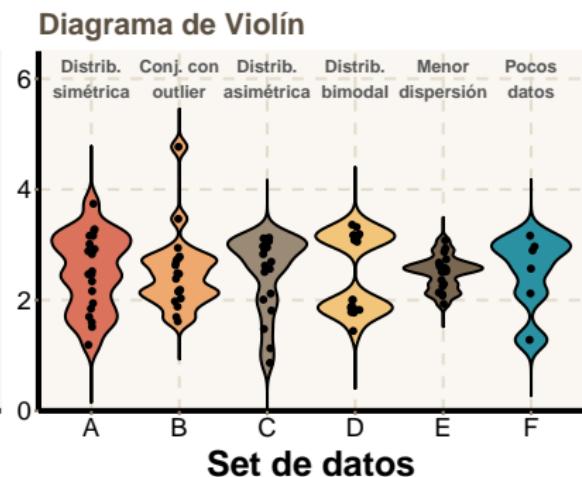
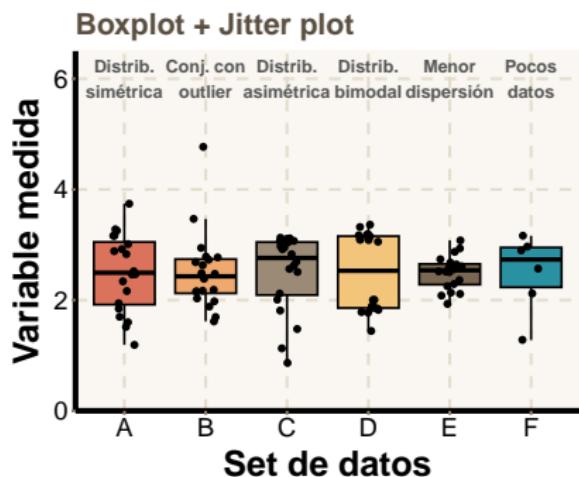
¿Cuál es el gráfico más apropiado para representar estos datos?



Segundo ejemplo - Datasets de mayor extensión

Se simularon seis conjuntos de datos considerando el caso de contar con observaciones de una variable cuantitativa para seis niveles de un factor de interés.

¿Cuál es el gráfico más apropiado para representar estos datos?



Segundo ejemplo - Datasets de mayor extensión

En la generación de los datasets se emplearon funciones de R base para simular datos con distribución normal, así como funciones del paquete `truncnorm` para modelar datos con distribuciones normales truncadas.

Segundo ejemplo - Datasets de mayor extensión

En primer lugar, construimos algunas funciones para facilitar la adaptación a múltiples ejemplos:

```
# rightskewed(): Genera un set de datos con distribución
# asimétrica hacia la derecha, provenientes de una distribución Beta.
rightskewed <- function(n) {
  d <- rbeta(n, 1, 5)
  return(d)
}

# leftskewed(): Genera un set de datos con distribución
# asimétrica hacia la izquierda, provenientes de una distribución Beta.
leftskewed <- function(n) {
  d <- rbeta(n, 5, 1)
  return(d)
}

# bimodal(): Genera datos con una distribución bimodal,
# construida a partir de dos distribuciones normales truncadas.
bimodal <- function(n) {
  c(rtruncnorm(n/2, a=0, b=2, mean=1, sd=.3),
    rtruncnorm(n/2, a=2, b=4, mean=3, sd=.3))
}
```

Segundo ejemplo - Datasets de mayor extensión

Posteriormente, las utilizamos para simular los distintos datasets:

```
set.seed(2019)
nn = 20

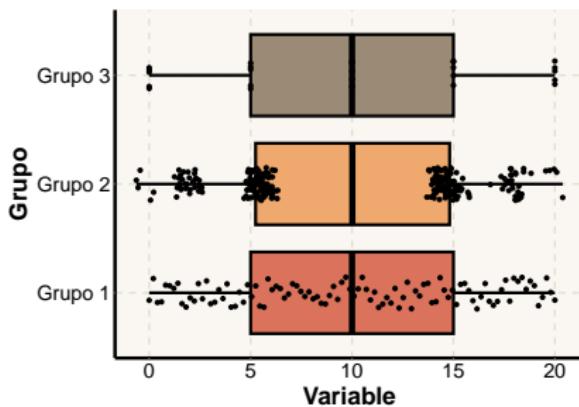
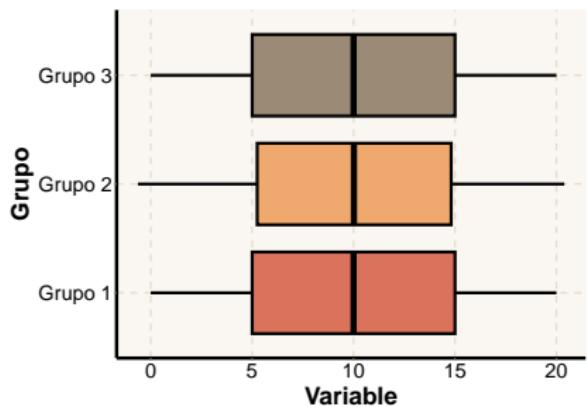
distribuciones <- c("rnorm", "outlier", "leftskewed", "bimodal")
datos2A <- lapply(1:4,
                  function(x) {data.frame(
                    var = scaling(get(distribuciones[x])(nn),
                                  desired.mean = 2.5,
                                  desired.sd = 0.7),
                    set = rep(LETTERS[x], each = nn))
                  })
datos2A <- bind_rows(datos2A)

datos2B <- data.frame(var = c(
  scaling(rnorm(nn), desired.mean = 2.5, desired.sd = 0.3),
  scaling(rnorm(6), desired.mean = 2.5, desired.sd = 0.7)),
  set = rep(c("E", "F"), c(nn, 6)))
)

datos2 <- full_join(datos2A, datos2B)
```

No hay gráfico que cuente la historia completa⁴

A pesar de que los boxplots para los tres grupos lucen idénticos, superponer un Jitter plot con las observaciones individuales nos revela la existencia de tres patrones radicalmente diferentes en los datos que los originan:



⁴Scherer, C. (2021). **Visualizing Distributions with Raincloud Plots (and How to Create Them with ggplot2)**. Recuperado de: <https://www.cedricscherer.com/2021/06/06/visualizing-distributions-with-raincloud-plots-and-how-to-create-them-with-ggplot2/>

Hay muchas alternativas . . .

Cédric Scherer 🐦 ➡️ 🐻 @CedScherer@... ⋮
That moment when you review a journal submission and you see dynamite plots: rage and joy at the same time!

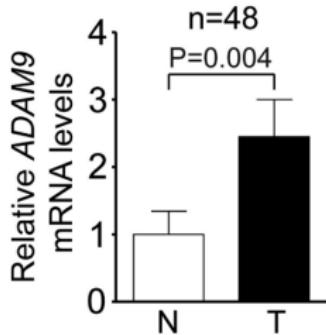
#dataviz #datavisualization
#barbarplot #DoBetter

Graphs: Cédric Scherer (@CedScherer)

Tercer ejemplo - Datos correlacionados

Un buen gráfico cuenta una historia sobre los datos.

¿Qué nos lleva a interpretar este gráfico en relación al tipo de muestras y a los resultados del análisis estadístico realizado?



J. Biol. Chem. 298 (8), 102225 (2022).

Tercer ejemplo - Datos correlacionados

¡Los datos provienen de muestras pareadas!

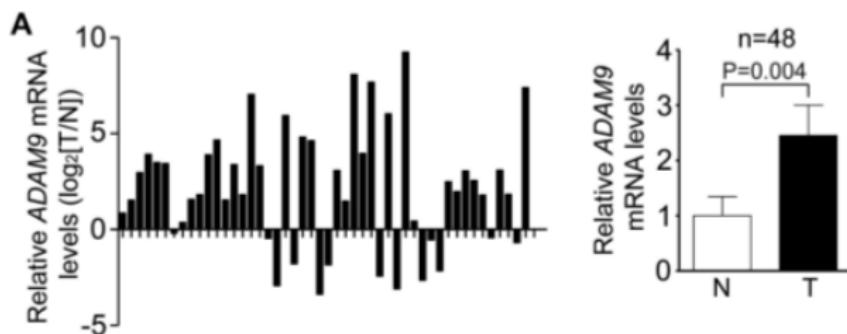


Figure 2. ***ADAM9* mRNA is upregulated in human CRC samples and promotes CRC cell migration and invasion *in vitro*.** A, comparison of *ADAM9* transcripts in CRC tissues (T) with matched adjacent normal tissues (N) from individual patients, as measured by RT-qPCR analyses (left). Results of 48 samples are summarized on the right, and Wilcoxon matched-pairs signed rank test was performed.

J. Biol. Chem. 298 (8), 102225 (2022).

El gráfico utilizado no comunica las características de los datos (ni el diseño experimental).

Tercer ejemplo - Datos correlacionados

¿Cómo deberíamos proceder entonces?

Simulemos primero un conjunto de datos que estén correlacionados. Lo podemos hacer con funciones del paquete `faux`.



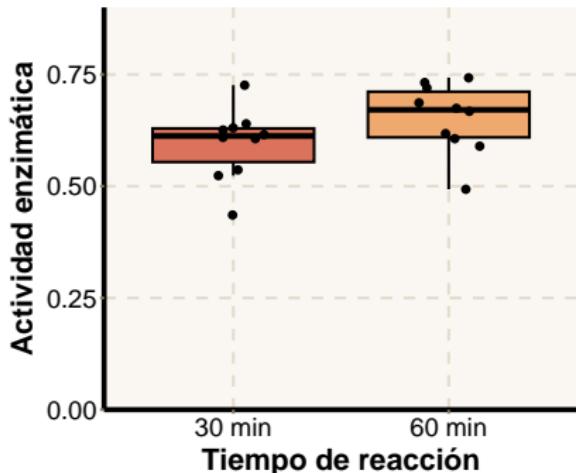
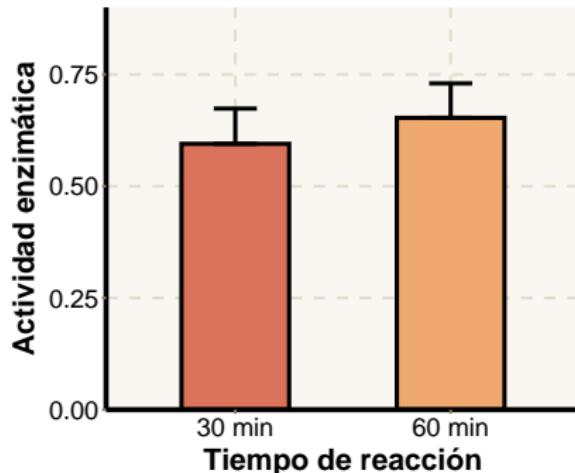
Tercer ejemplo - Datos correlacionados

Los datos generados están basados en un experimento en el que se mide la actividad de una enzima en 10 muestras de extractos celulares luego de 30 y 60 min de reacción de formación de un aducto fluorescente:

Muestra	Actividad enzimática	
	30 min	60 min
1	0.630	0.674
2	0.609	0.686
3	0.524	0.606
4	0.537	0.590
5	0.435	0.493
6	0.626	0.720
7	0.639	0.732
8	0.615	0.618
9	0.726	0.743
10	0.606	0.668

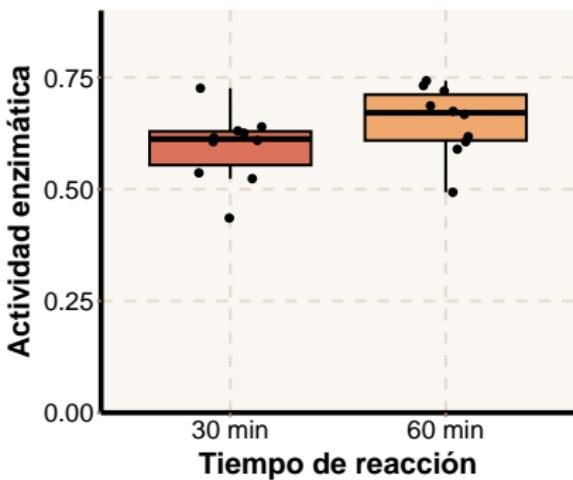
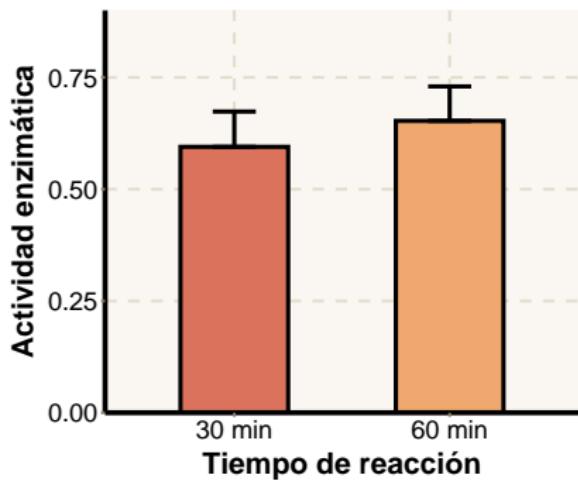
Tercer ejemplo - Datos correlacionados

¡Grafiquemos! El gráfico dinamita ya sabemos que presenta muchos problemas, pero ¿qué pasa con el boxplot?



Tercer ejemplo - Datos correlacionados

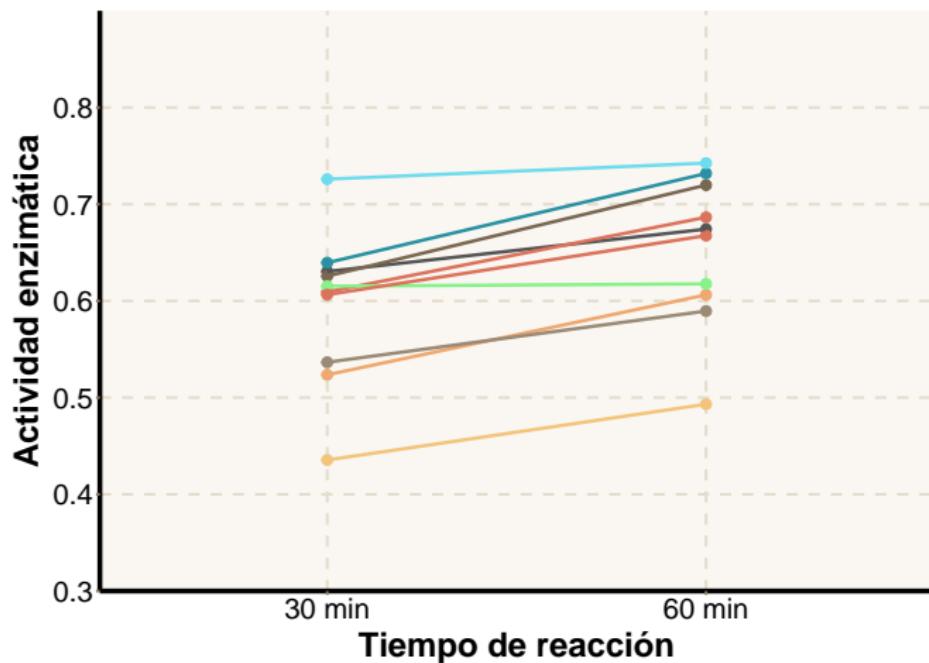
¡Grafiquemos! El gráfico dinamita ya sabemos que presenta muchos problemas, pero ¿qué pasa con el boxplot?



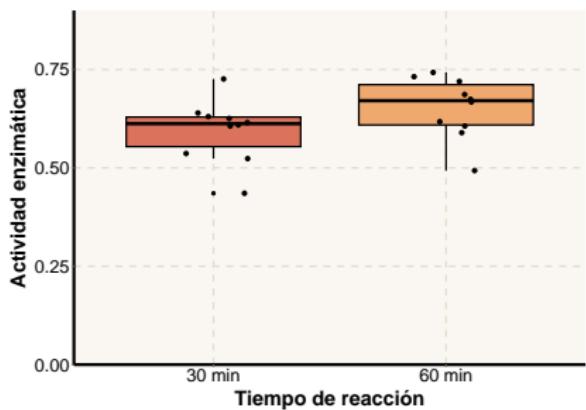
El boxplot tampoco refleja el hecho de que se trata de **muestras pareadas**. ¿Entonces?

Tercer ejemplo - Datos correlacionados

Un gráfico más adecuado para representar datos provenientes de muestras pareadas o dependientes es el **gráfico de perfiles**.



Tercer ejemplo - Datos correlacionados

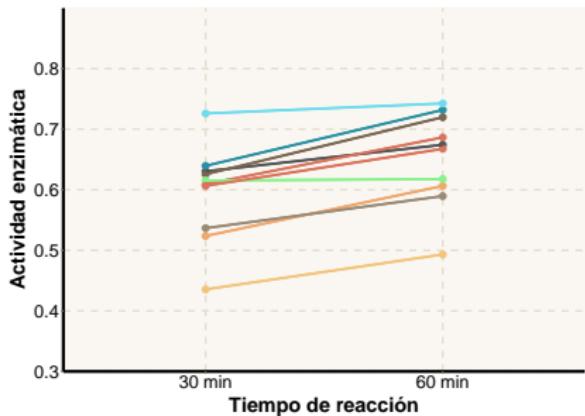


```
## Two sample t-test  
## p-value = 0.1131
```

En línea con la propuesta de representar los datos con un boxplot, si quisiéramos evaluar si existen diferencias estadísticamente significativas en la actividad enzimática promedio entre los dos tiempos de reacción de interés, utilizaríamos un **test-t para la comparación de dos promedios en base a muestras independientes.**

¿Cuál sería nuestra conclusión en este caso?

Tercer ejemplo - Datos correlacionados



```
## Two sample t-test
```

```
## p-value = 0.0002
```

Habiendo elegido un gráfico que tiene en cuenta que trabajamos con **muestras pareadas**, el test adecuado para este caso es **test-t para la comparación de dos promedios en base a muestras dependientes**.

¿Cuál sería nuestra conclusión en este caso? ¿Es la misma a la que habíamos llegado anteriormente?

Conclusiones

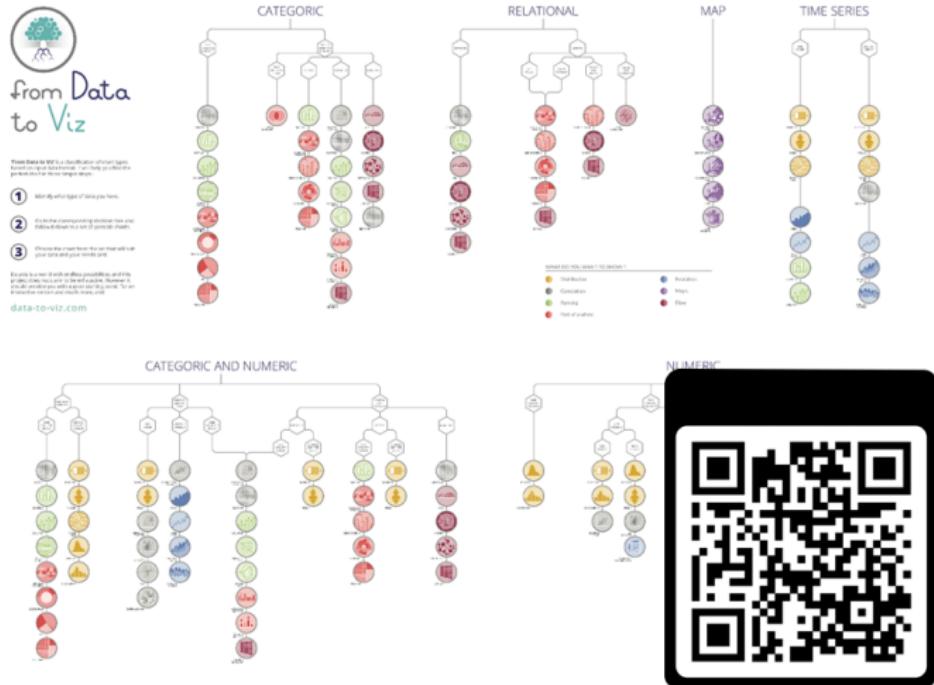
"The greatest value of a picture is when it forces us to notice what we never expected to see". (John W. Tukey)

- Un buen gráfico comunica ideas complejas con precisión, claridad y eficiencia.
- Todos los gráficos tienen sus potencialidades y limitaciones y no hay gráfico que cuente la historia completa sobre nuestros datos.
- Para construir un buen gráfico, se debe prestar atención a cuestiones conceptuales o estadísticas, perceptuales y estéticas.

*"Graphical excellence requires telling the truth about the data".
(Edward Tufte)*

Recursos interesantes

- **data-to-viz.com**: árboles de decisión que conducen desde un tipo particular de datos a un conjunto de gráficos posibles.



Recursos interesantes

- cedricscherer.com: blog de Cédric Scherer, *data visualization designer*.

The screenshot shows the homepage of cedricscherer.com. At the top, there is a navigation bar with links for Blog, Gallery, Portfolio, About Me, and Links. Below the navigation bar is a large, colorful graphic composed of many overlapping, wavy layers of different colors (purple, blue, green, orange). Overlaid on this graphic is the name "CÉDRIC SCHERER" in large, bold, white capital letters, followed by "Data Visualization & Information Design" in a smaller, white sans-serif font. At the bottom left, there is a section with the text "Hi, I am Cédric" followed by a yellow hand icon. Below this, it says "Data Visualization Designer, Consultant and Instructor for Engaging and Effective Graphical Storytelling." At the very bottom, there are three green links: "Gallery", "Portfolio", and "About Me". To the right of the main content area is a large QR code.

Recursos interesantes

- Healy, K. (2019). **Data Visualization. A practical introduction.** Princeton University Press.
- Tufte, E. R. (2001). **The visual display of quantitative information.** 2nd Ed. Graphics Press.
- Weissberger, T. L., Savic, M. et al. (2017) **Data visualization, bar naked: A free tool for creating interactive graphics.** *J. Biol. Chem.* 292(50) 20592-20598.
- Wong, B. (2012). **Visualizing biological data.** *Nat. Methods* 9, 1131.

Agradecimientos



RenRosario



ÁREA ESTADÍSTICA Y PROCESAMIENTO DE DATOS
Facultad de Ciencias Bioquímicas y Farmacéuticas
Universidad Nacional de Rosario



E-mails:

jferreyra@fbioyf.unr.edu.ar -
nlabadie@fbioyf.unr.edu.ar

Github:

github.com/natilab/datavis

