

ECONOMÍA Y CIENCIA DE DATOS

Reproducibilidad y Repetitividad

Arturo Chian



Un presentación BEST <http://bestteamperu.org/>

¿Qué se nos viene a la mente con Ciencia de Datos?



Imagen extraída de Learning Tree

¿Qué se nos viene a la mente con Ciencia de Datos? 🧠

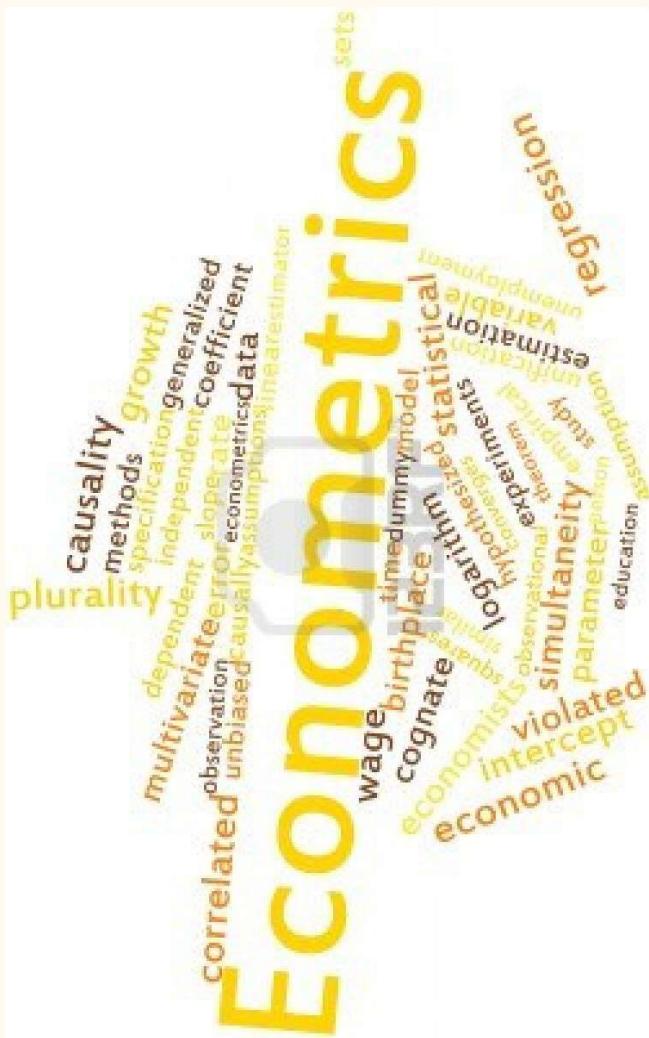


Imagen extraída de VSPS Education

Definiendo Data Science

“

For a long time I have thought I was a statistician, interested in inferences from the particular to the general. But as I have watched mathematical statistics evolve, I have had cause to wonder and to doubt. ... All in all I have come to feel that my central interest is in data analysis, which I take to include, among other things: procedures for analyzing data, techniques for interpreting the results of such procedures, ways of planning the gathering of data to make its analysis easier, more precise or more accurate, and all the machinery and results of (mathematical) statistics which apply to analyzing data.

Tukey (1962). *The future of Data Analysis, The Annals of Mathematical Statistics*

Definiendo Data Science



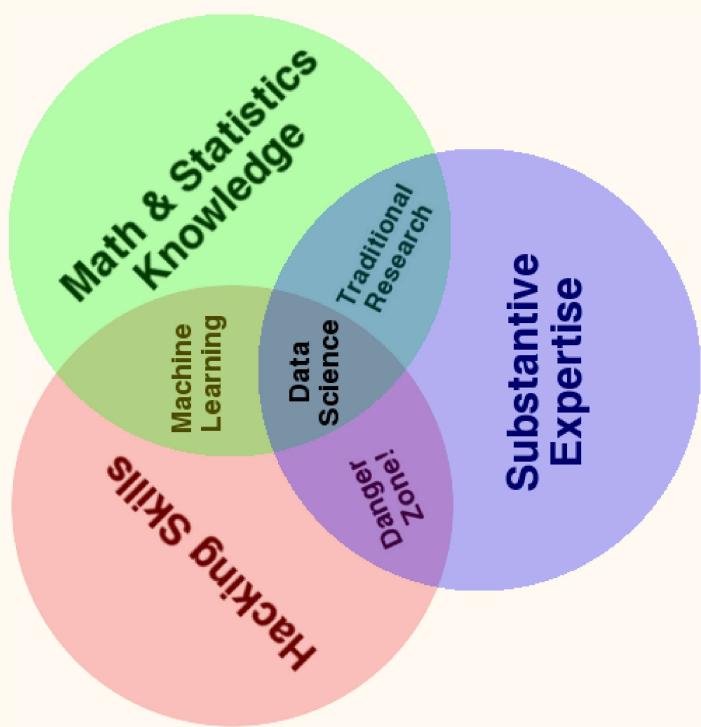
Hace 50 años, John Tukey llamó a una reforma académica en estadística, a través de uno de los más importantes papers de esa época, llamado “*The Future of Data Analysis*”, donde señalaba la necesidad futura de una ciencia cuyo interés sea aprender de la data o análisis de datos. Hace unos 20 a 10 años, John Chamber, Jeff Wu, Bill Cleveland y Leo Breiman, dieron una serie de argumentos, de forma independiente sobre expandir los límites de la estadística teórica: Chambers enfatizaba la importancia de la preparación de datos, más que el modelaje estadístico; Breiman, prefería enfatizar la predicción antes que la inferencia; y Cleveland y Wu sugerían llamar a este nuevo campo Data Science por su estrecha relación a la data.

Arturo Chian (2018). A propósito de los 25 años de R y 50 años de Data Science (Parte 1), Blog de Behavioral Economics & Data Science Team

Definiendo Data Science

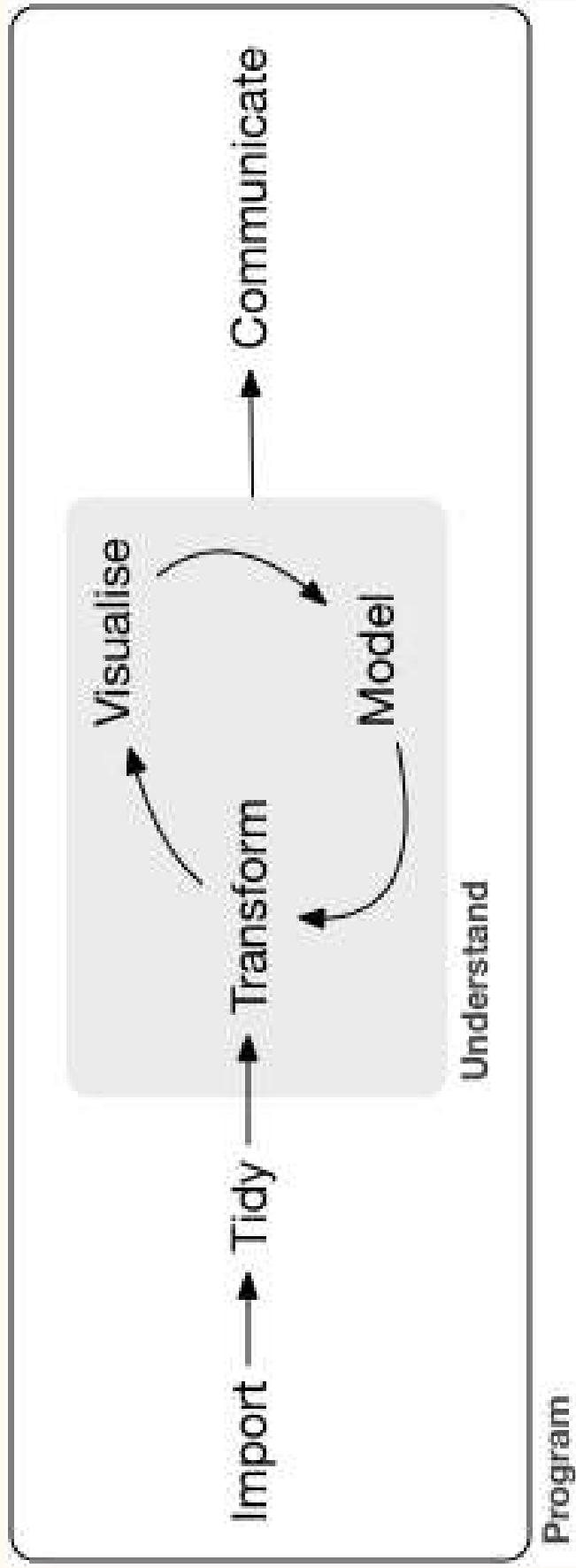
Diagrama de Venn - Drew Conway

1. **Hacking Skills:** Capacidad de resolver problemas programando.
2. **Math & Statistics knowledge:** Aplicar de forma correcta estadística.
3. **Conocimiento de experto:** Comprender la data en su campo de investigación (economía, biología, psicología, derecho, etc).



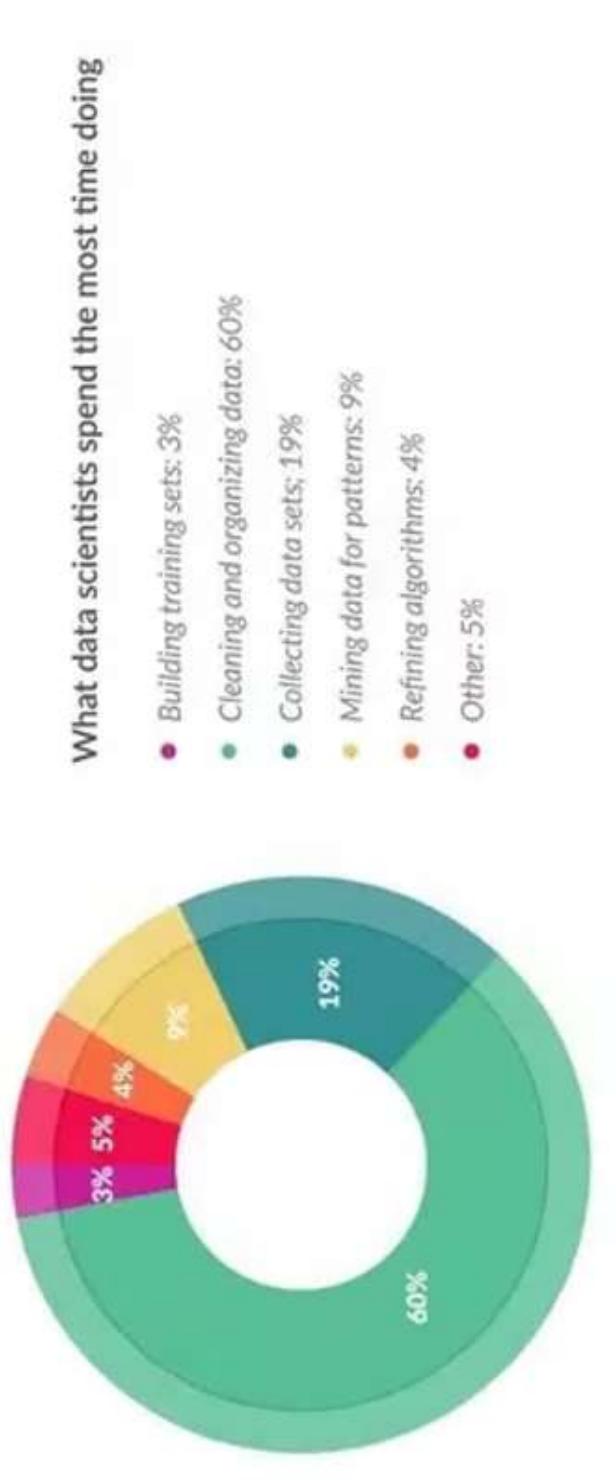
Fuente: Drew Conway

¿Qué hace un Data Scientist en el día a día? 🎈



Fuente: *I/O-World*

¿Qué hace un Data Scientist en el día a día? 🧠



Fuente: I/O-World

El Ciudadano (¿Economista?) Data Scientist



BIG DATA | 21 mar 2019

Imagen extraída de VSPS Education

¿Cómo muchos economistas hemos aprendido Data Science en la práctica?



¿Y qué tan útil puede ser Data Science para un economista?



Creación de libros abiertos 😊

The screenshot shows the front cover and the table of contents for the book 'Forecasting: Principles and Practice' by Rob J Hyndman and George Athanasopoulos.

Front Cover:

- Title: Forecasting: Principles and Practice
- Authors: Rob J Hyndman and George Athanasopoulos
- Monash University, Australia
- Image: A dark blue background with the title and authors' names in white.

Table of Contents:

- Preface
- 1 Getting started
 - 1.1 What can be forecast?
 - 1.2 Forecasting, planning and goals
 - 1.3 Determining what to forecast
 - 1.4 Forecasting data and methods
 - 1.5 Some case studies
 - 1.6 The basic steps in a forecasting process
 - 1.7 The statistical forecasting perspective
 - 1.8 Exercises
 - 1.9 Further reading
- 2 Time series graphics
 - 2.1 ts objects
 - 2.2 Time plots
 - 2.3 Time series patterns
 - 2.4 Seasonal plots
 - 2.5 Seasonal subseries plots
 - 2.6 Scatterplots

Fuente: *Forecasting: Principles and Practice*

Nuevas formas de hacer tesis 😊

thesisdown



README.md

This project was inspired by the bookdown package and is an updated version of my Senior Thesis template in the reedtemplates package [here](#).

Currently, the PDF and gitbook versions are fully-functional. The word and epub versions are developmental, have no templates behind them, and are essentially calls to the appropriate functions in bookdown.

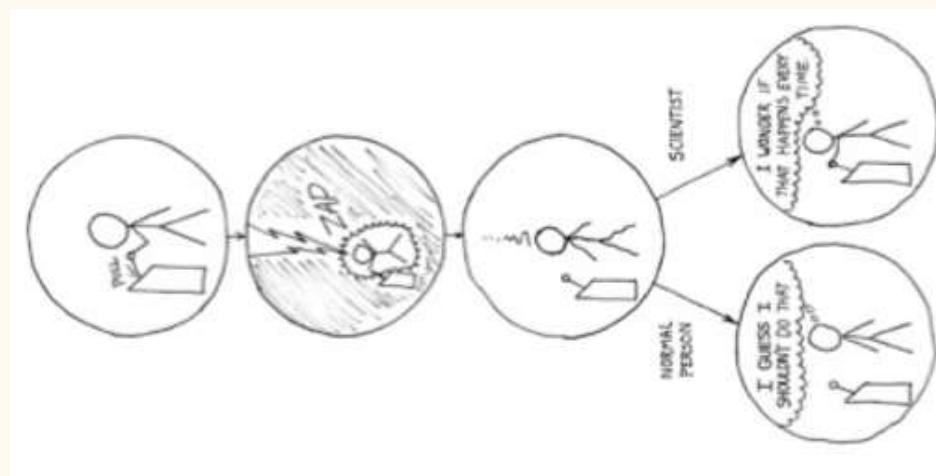
If you are new to working with bookdown / rmarkdown , please read over the documentation available in the gitbook template at <https://thesisdown.netlify.com/>. This is also available below at http://ismayc.github.io/thesisdown_book.

The current output for the four versions is [here](#):

- PDF (Generating LaTeX file is available [here](#) with other files at in the book directory.)
 - Word
 - ePUB
 - gitbook

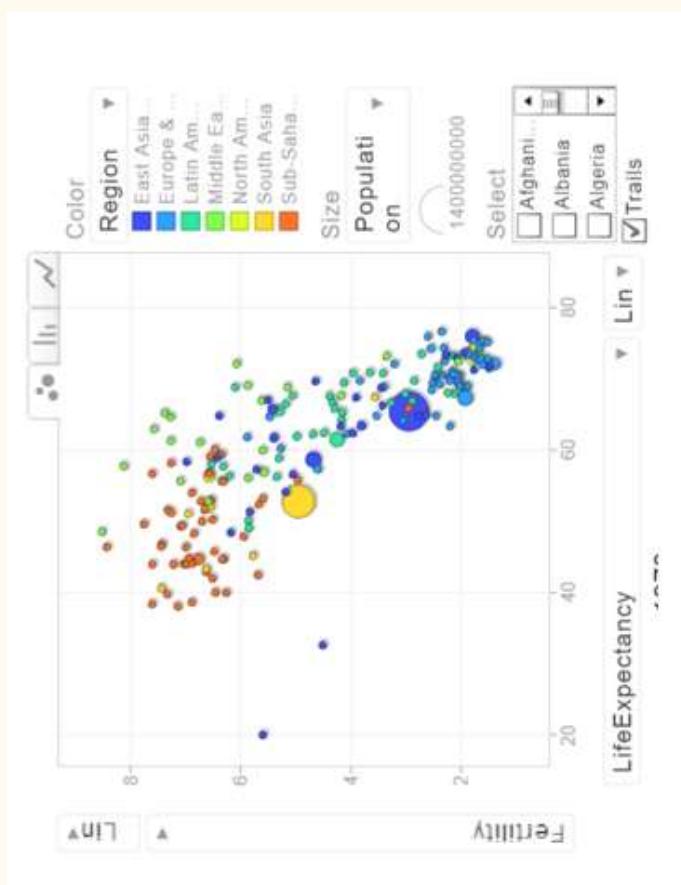
Fuente: *ThesisDown*

Promueve la investigación reproducible



Fuente: ThesisDown

Gráficos Dinámicos ☺



Fuente: Mages Blog

Desarrollo de aplicaciones fácil de realizar ☺

Pasa-Segura-Medellín

Aplicación web sobre accidentalidad vial en Medellín usando Ciencia de Datos

Vers. 3 (2018-1). Años de observación: 2014 - 2017_1 (ene-jul 2017). Grupo de investigación IDINNOV investigacion@idinnov.com

1. DESCRIPCIÓN DE VARIABLES DE ACCIDENTALIDAD PARA PERIODOS 2014 - 2017_1

Periodo de observación:

Elegir: Dia Mes Año ▾

2016

SEGMENTACIÓN DE ACCIDENTES VIALES

Elegir: Horas ▾

Segmentos: Horas ▾

Gravedad ▾

Todos ▾

Frecuencia de todos

Hour	Percentage
0	11.9%
1	13%
2	14.6%
3	11.9%
4	12.5%
5	13.1%
6	13.1%
7	13.1%
8	13.1%
9	14.6%
10	13.1%
11	11.9%
12	11.9%

SEGMENTACIÓN DE ACCIDENTES VIALES

Elegir: Año ▾

Accidentes según gravedad por día (promedio)

Severity Level	Average Number of Accidents
Leve	56 %
Moderado	43.5 %
grave	50.9 %

Fuente: Pasa Segura Medellín

Y otras más potentes... El límite es tu imaginación 😊

Fuente: Analytic Health

La Llave para desbloquear el poder del ultra instinto del economista del siglo XXI

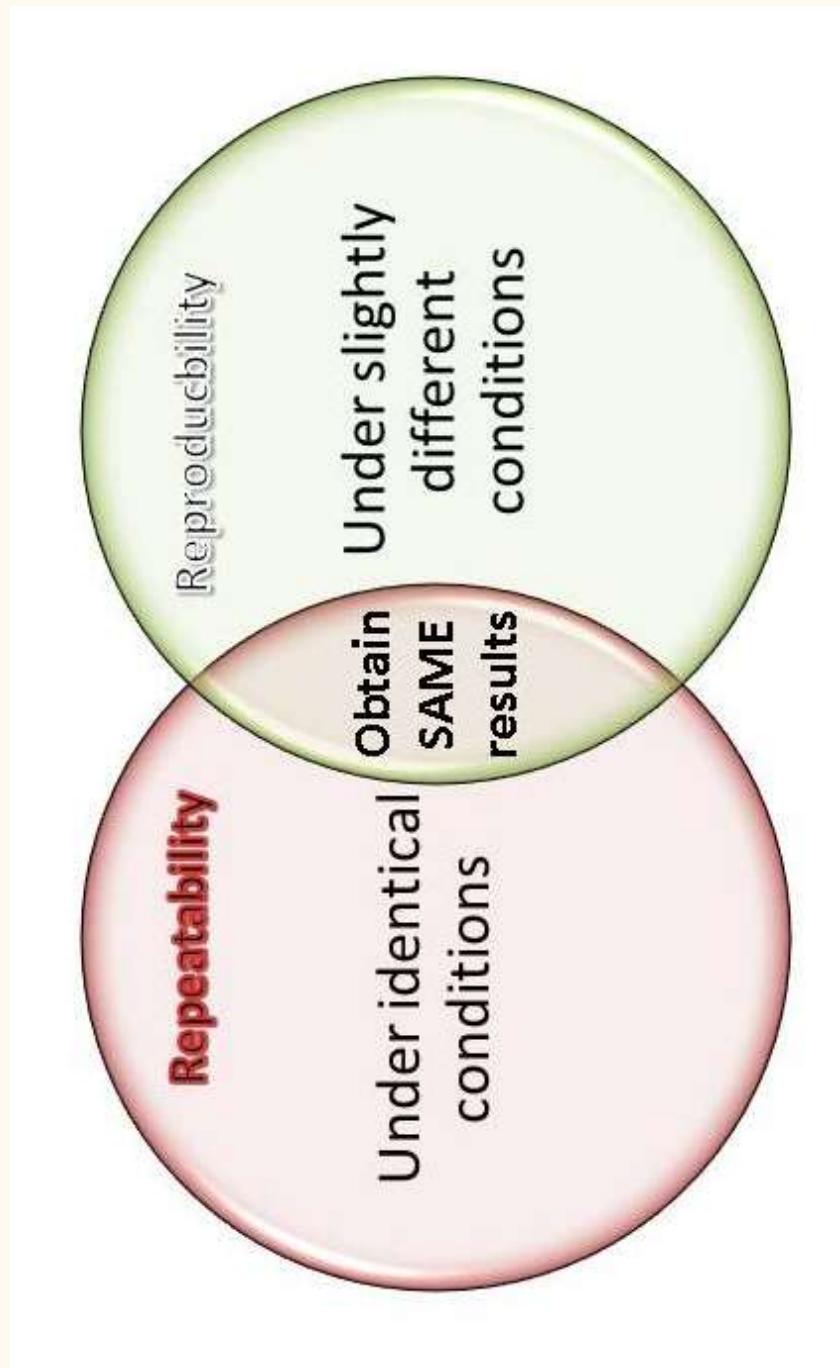
El poder del Guantalete del Data Science

En manos de un economista para dominar al universo: Consultorías, academia, trabajos, etc. Las posibilidades son infinitas. 😊

Reproducibilidad y Repetibilidad



Conociendo los conceptos



¿Por qué sería importante para economía?



¿Crisis de replicabilidad/reproducibilidad en economía?

Economics

Why Economics Is Having a Replication Crisis

Recreating research by gathering data from the real world and analyzing it statistically often fails to produce the same result.

By Noah Smith

September 17, 2018, 8:00 AM GMT-5

Fuente: *Bloomberg opinion*

¿Crisis de replicabilidad/reproducibilidad en economía?



Fuente:Vox

24 / 48

Tipos de reproducibilidad en Economía

1. Reproducirlo con otra población/tiempo/espacio.
2. Usar la misma data, pero modificar la metodología de datos cualitativos.
3. Replicar el paper sin modificar la metodología ni la data.

Tipo 1: Reproducibilidad en otra población

Perú



Alemania



Tipo 2: Variación metodológica en definiciones cualitativas

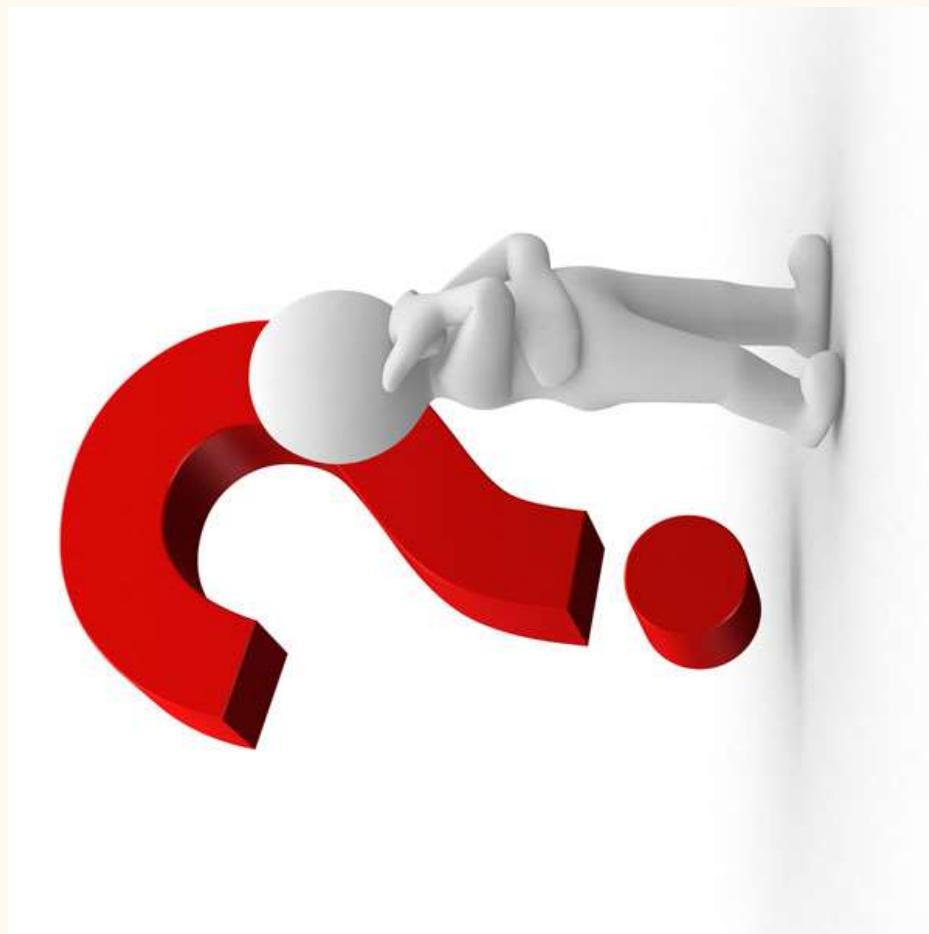
- ↳ ¿Cómo definir qué es pobre?
- ↳ ¿Cómo definir qué es feliz?
- ↳ ¿Cómo definir qué es bienestar?
- ↳ ¿Cómo definir qué es un buen estudiante?

Tipo 3: Replicación al 100%

- ↳ Se debe realizar de preferencia con el mismo software.
- ↳ Se debe realizar con la base de datos en bruta.
- ↳ De preferencia sí, con el mismo código, en caso aplique.

Tipo 3: Replicación al 100%

¿Esto sería relevante en Economía?



Tipo 3: Replicación al 100%

- ↳ En el 2016, los economistas Andrew Chang y Phillip Li trataron de reproducir los resultados de 65 papers publicados en importantes Journals. Ellos usaron la data original e incluso contactaron a los autores para que les señalen los pasos que usaron. Sólo lograron replicar el 49%.
- ↳ Uno de los casos más conocidos es la replicación de un paper del 2013 de Reinhart y Rogoff, los cuales alegaban una correlación alta entre alta deuda de gobierno y crecimiento; pero encontraron errores/manipulaciones en la base de datos que usaron vs la original.

Otros problemas en las investigaciones: El p-hacking

- ↳ Es un problema que afecta a la ciencia.
- ↳ Se trata de buscar p value significativo probando diversas técnicas sin rigor científico.
- ↳ Hay muchas publicaciones científicas que se desarrollan y se publican usando el p-hacking.

Rmarkdown: ¿Una solución?



El inicio: knitr

El paquete que dio inicio a todo se llama knitr.



El creador de knitr: Yihui Xie

Creador de diversos paquetes de R y uno de los más relevantes Data Scientist del mundo de R. Actualmente cuenta con un PhD y trabaja en RStudio.



Markdown vs Latex

¿Qué ventajas tiene markdown sobre Latex?

- ↳ **Rápido de aprender:** En minutos lo aprendes.
- ↳ **Variedad de outputs:** No sólo latex, word, ppt, html, markdown, etc.
- ↳ **flexible a tu medida:** Si necesitas más detalles específicos, puedes usar CSS o incluso Latex.
- ↳ **Combinar lenguajes de programación:** No sólo R, sino Python, Julia, C, etc.

Fuente: [Blog Yihui Xie en inglés](#)

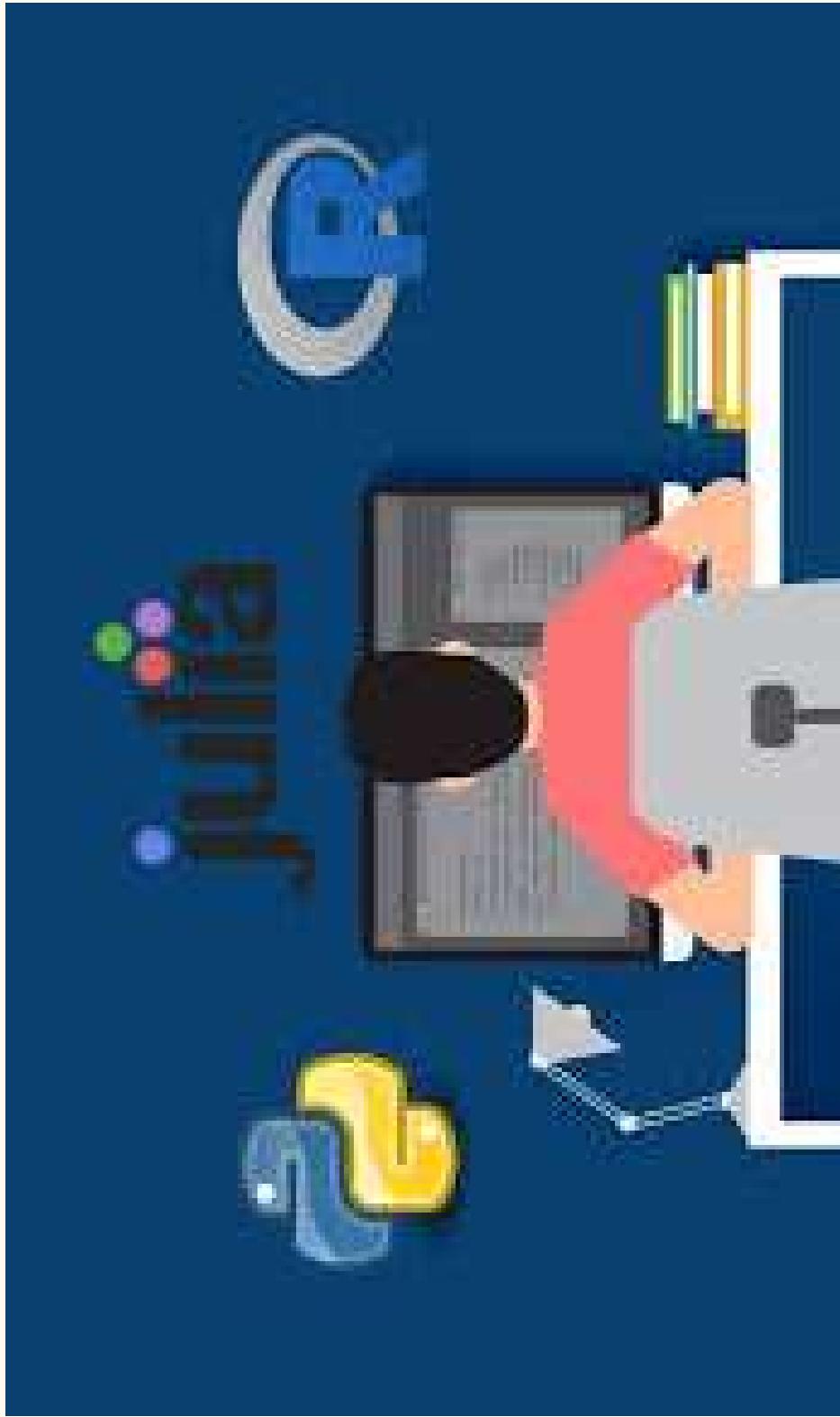
Estructura de un RMD

- ☞ **YAML:** Lenguaje de serialización. Sirve para meter datos como autor, fecha, opciones avanzadas, etc.
- ☞ **Títulos:** Usando michi para título 1, 2 michis para título 2, así sucesivamente.
- ☞ **Textos:** Tan fácil como tippear normal. En caso de negritas, poner 1, 2 o 3 subrayado o negritas (lo veremos en la práctica!).
- ☞ **Chunks:** Puedes correr código de R y otros lenguajes en este espacio, y puedes configurar de tal forma que sólo sea necesario.
- ☞ **Código en texto:** Aplica a tu paper o reporte un código de tal forma que te salga, por ejemplo, el coeficiente de regresión y no tengas que tippearlo.

¿Y qué tan difícil es programar?



¿Y qué tan difícil es programar?



Programación funcional (Ejm: Excel)

```
arrange(  
  summarize(  
    group_by(  
      filter(mtcars, carb > 1),  
      cyl  
    ),  
    Avg_mpg = mean(mpg)  
  ),  
  desc(Avg_mpg)  
)
```

Programación funcional (Ejm: Excel)

```
arrange(summarize(group_by(filter(mtcars, carb > 1), cyl), Avg_mpg = mean(mpg)) )
```

Programación basada en objetos (Stata)

```
a <- filter(mtcars, carb > 1)
b <- group_by(a, cyl)
c <- summarise(b, Avg_mpg = mean(mpg))
d <- arrange(c, desc(Avg_mpg))
print(d)
```

Programación con pipas (R + Tidyverse)

```
library(magrittr)
library(dplyr)

mtcars %>%
  filter(carb > 1) %>%
  group_by(cyl) %>%
  summarise(Avg_mpg = mean(mpg)) %>%
  arrange(desc(Avg_mpg))
```

Possibles soluciones

- ↳ Promover tesis reproducibles.
- ↳ Promover más Journals que sean reproducibles.
- ↳ Promover el uso de software libre.

Posibles soluciones

BEST aportando a la comunidad del software libre y a la comunidad de ciencia abierta.

Paquetes en R – Información Socioeconómica



Posibles soluciones

BEST aportando a la comunidad del software libre y a la comunidad de ciencia abierta.

The screenshot shows a Jupyter Notebook interface with the following details:

- Title:** Analizando a los 4 Principales Bancos de Peru
- Author:** Arturo Chian
- Date:** 2018-07-10
- Source:** vignettes/analisis-4-principales-bancos.Rmd
- Contents:** Riesgo de crédito, Rentabilidad, Liquidez, Eficiencia
- R Code (Visible Part):**

```
packs=c("SBSR", "xts", "dplyr", "tidyverse", "dygraphs", "ggplot2", "reshape2", "lubridate")
invisible(lapply(packs, library, character=T))

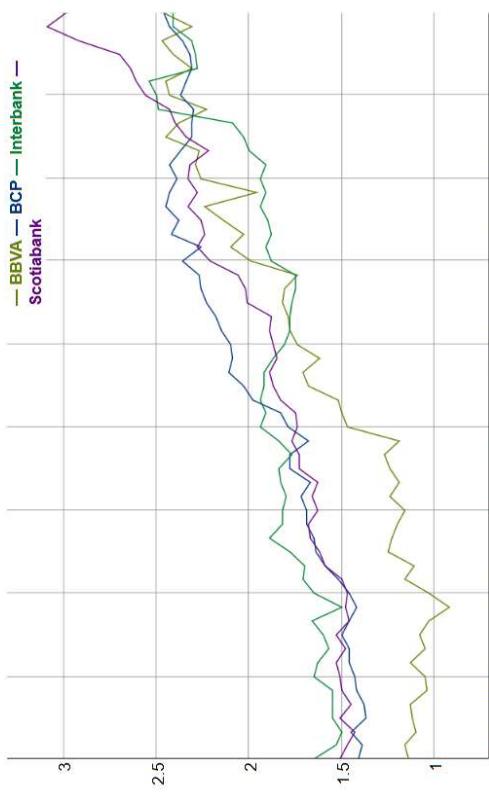
data("bancos")
bd<-bancos %>% filter (Entidad %in% c("BCP", "BBVA", "Scotiabank", "Interbank"))
```
- Notebook Footer:** Utilizemos el poder de R y BEST para analizar los 4 principales bancos. Primero activemos los paquetes!

Possibles soluciones

BEST aportando a la comunidad del software libre y a la comunidad de ciencia abierta.

Morosidad

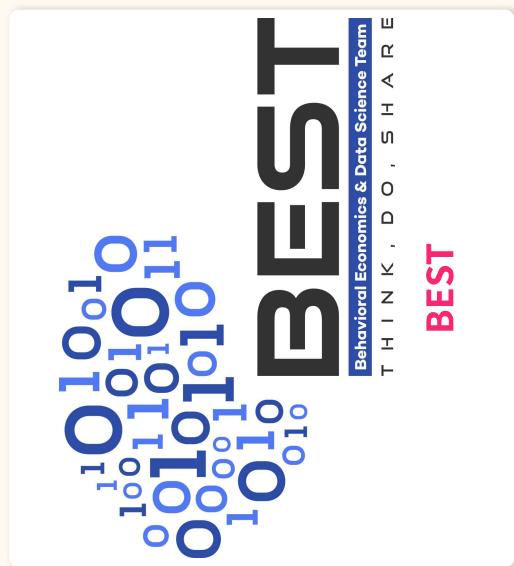
```
grafico1<-bd %% select(Entidad,morosidad,Fecha)
grafico1<-grafico1 %>%
  spread(Entidad, morosidad)
grafico1<-xts(grafico1[,2:5],order.by = as.Date(grafico1$Fecha))
dygraph(grafico1)
```



Contents

- Riesgo de crédito
- Riesgo de Solvencia
- Rentabilidad
- Liquidez
- Eficiencia

Economía y Ciencia de datos



IMUCHAS

AÑAYKI! • YUSPAGARA! • THANK YOU! • 謝謝!

GRACIAS!