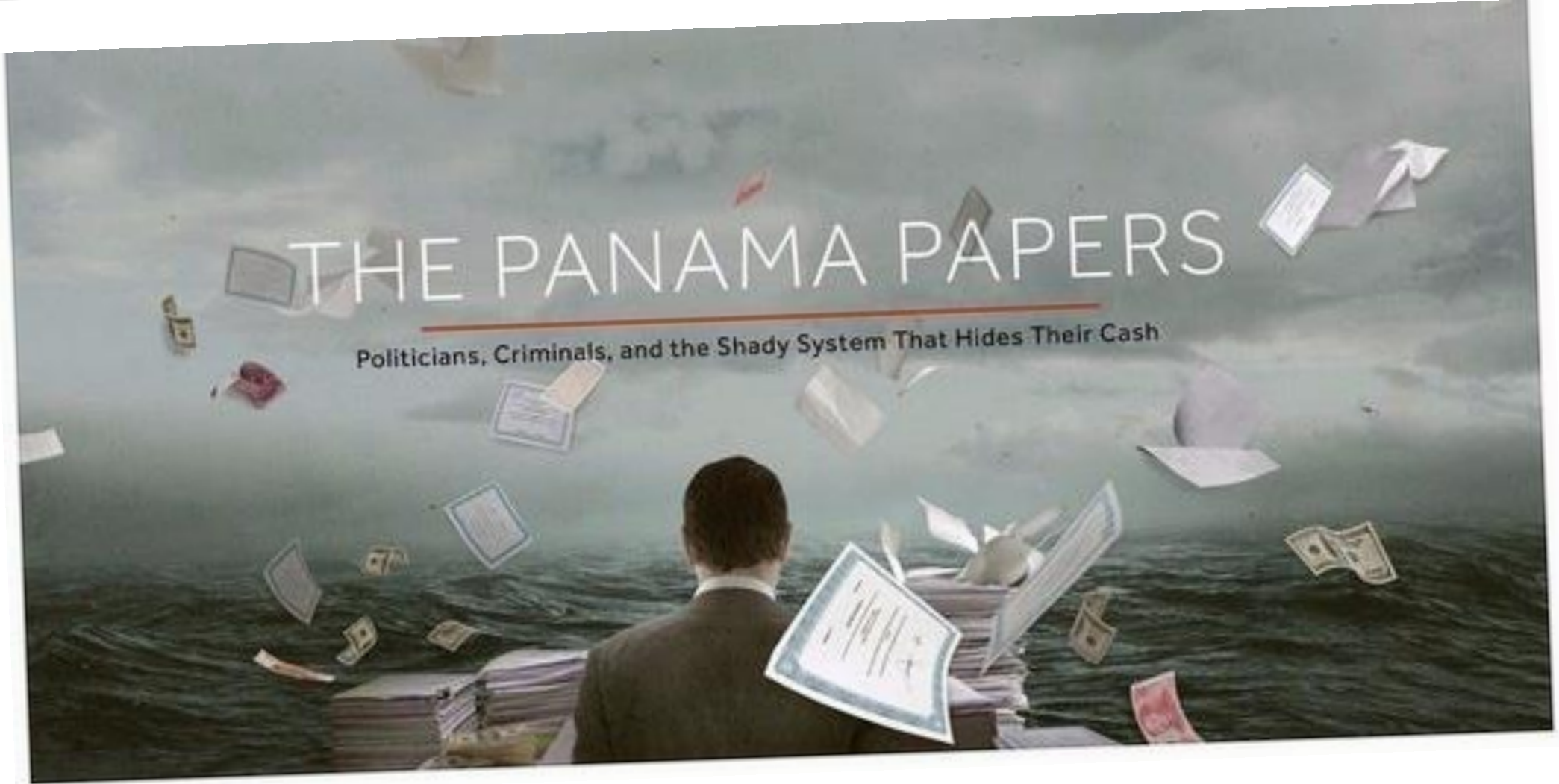


Clase 1: Periodismo y Ciencia de Datos.

Cristóbal Alcázar
alcazar.cristobal90@gmail.com

THE PANAMA PAPERS

Politicians, Criminals, and the Shady System That Hides Their Cash

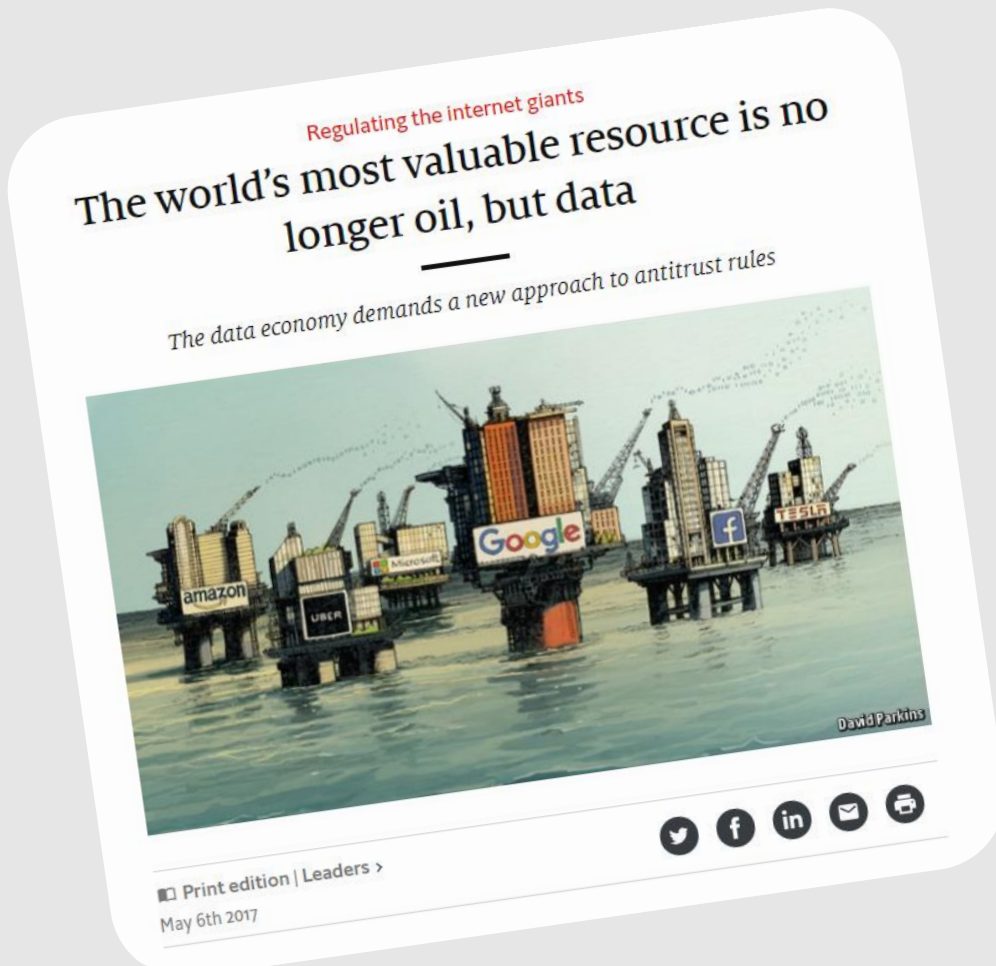


Panama Papers.



- Investigación de carácter internacional, en que trabajaron periodistas en 25 idiomas diferentes y cerca de 80 países distintos.
- Contribuyeron más de 100 medios de comunicación de todo el mundo.
- Basada en la información extraída de 11,5 millones de archivos filtrados que constituyeron una base de datos de 2,6 terabytes de datos.

~ El **90%** de todos los datos del mundo fue generado en los últimos 2 años.



¿Tiene sentido para un periodista desarrollar habilidades para analizar datos?

¿Porqué no externalizar este trabajo?



COMPUTER SCIENCE



COLUMBIA | ENGINEERING
The Fu Foundation School of Engineering and Applied Science



COLUMBIA UNIVERSITY
IN THE CITY OF NEW YORK

CONTACT US | DIRECTIONS

Search



General		Academics		People		Press	Career & Jobs Center		Affiliated Programs	
Theory	Graphics & User Interfaces	NLP & Speech	Security & Privacy	Computational Biology	Software Systems	Computer Engineering	Networking	Vision & Robotics	Machine Learning	Artificial Intelligence

Dual M.S. in Journalism and Computer Science

This dual degree program is designed to provide students with skills in Computer Science and Journalism to prepare them for new digital-media oriented careers in journalism. Students will earn Master's degrees in Computer Science and in Journalism.

Students will enroll for a total of four semesters. In addition to taking classes already offered at the Journalism and Engineering schools, students will attend a seminar and workshop designed specifically for the dual degree program. The seminar will teach students about the impact of digital techniques on journalism; the emerging role of citizens in the news process; the influence of social media; and the changing business models that will support newsgathering. In the workshop, students will use a hands-on approach to delve deeply into

Columbia Engineering is committed to an open and welcoming community for all students, faculty, researchers, and staff. Click for Dean Mary Boyce's [full statement](#).

Upcoming Events

SEP
20

High-dimensional Sampling and Integration

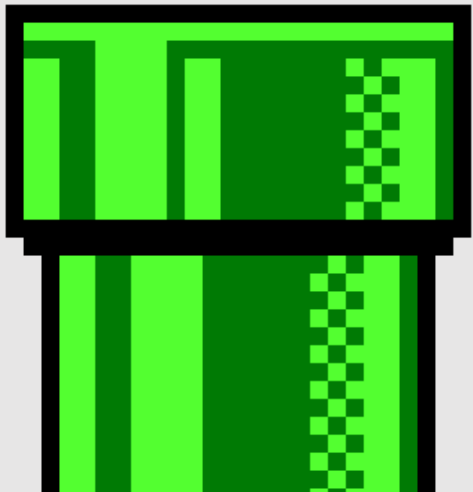
Thursday 1:00 pm

CS Conference Room

Ciencia de Datos.

Ciencia de datos es el **proceso** en que los **datos** se **convierten** en **conocimiento** y comprensión.

~ tubería de datos!



Importar

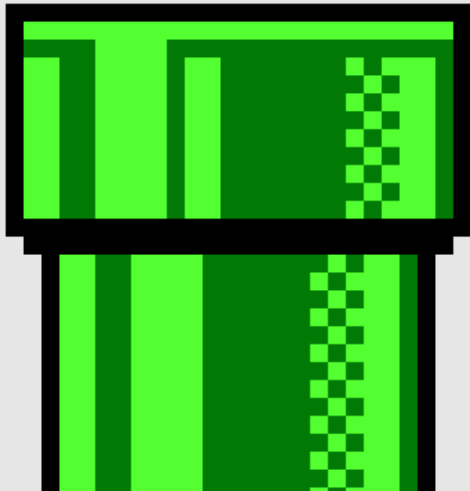


Tidy



Almacenar los
datos de manera
consistente

Descubrir y entender



Importar



Tidy



Almacenar los
datos de manera
consistente

Transformar

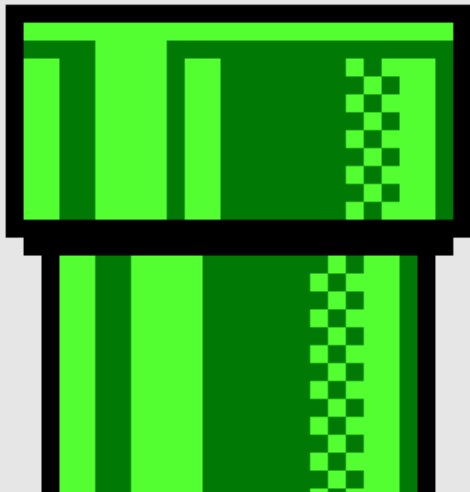
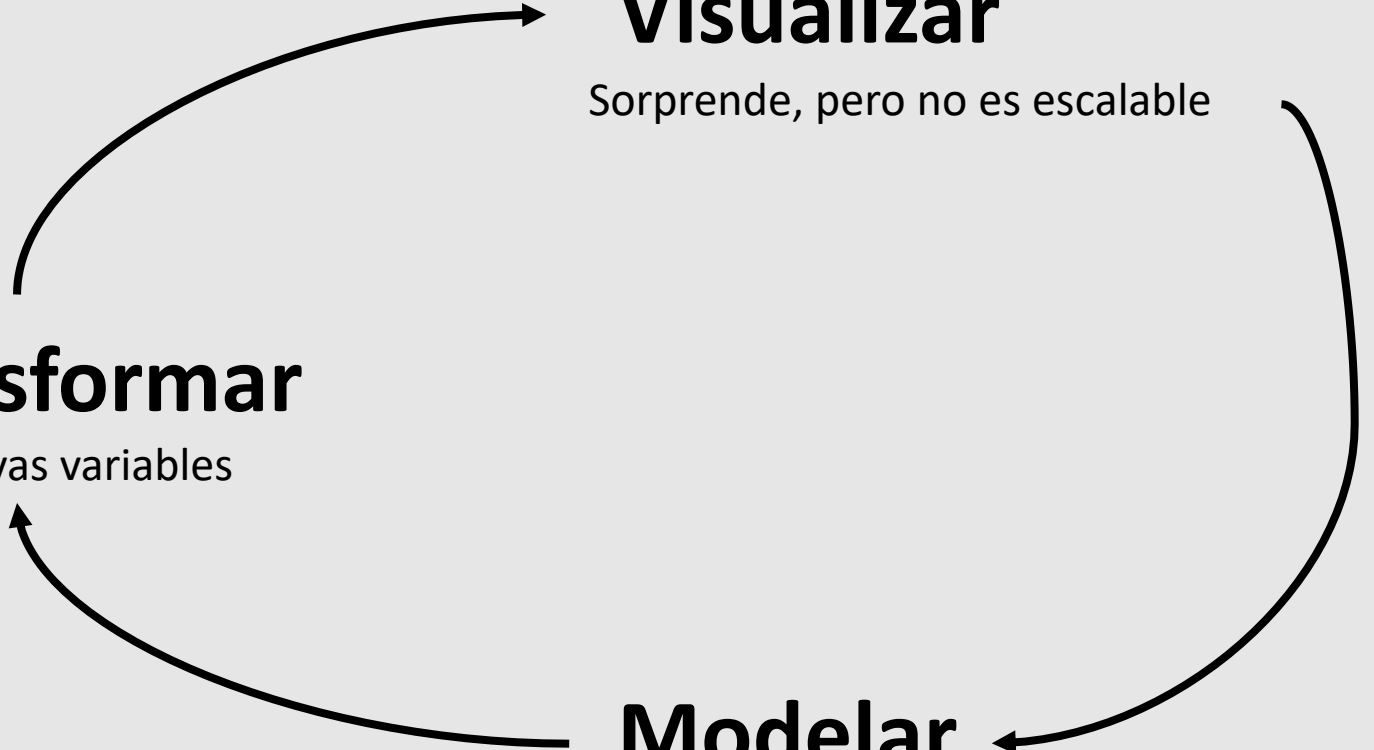
Crear nuevas variables

Visualizar

Sorprende, pero no es escalable

Modelar

Escalable, pero no sorprende



Importar



Tidy



Almacenar los
datos de manera
consistente

Transformar

Crear nuevas variables

Visualizar

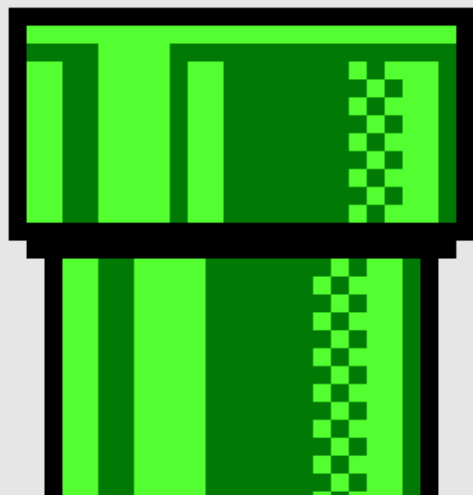
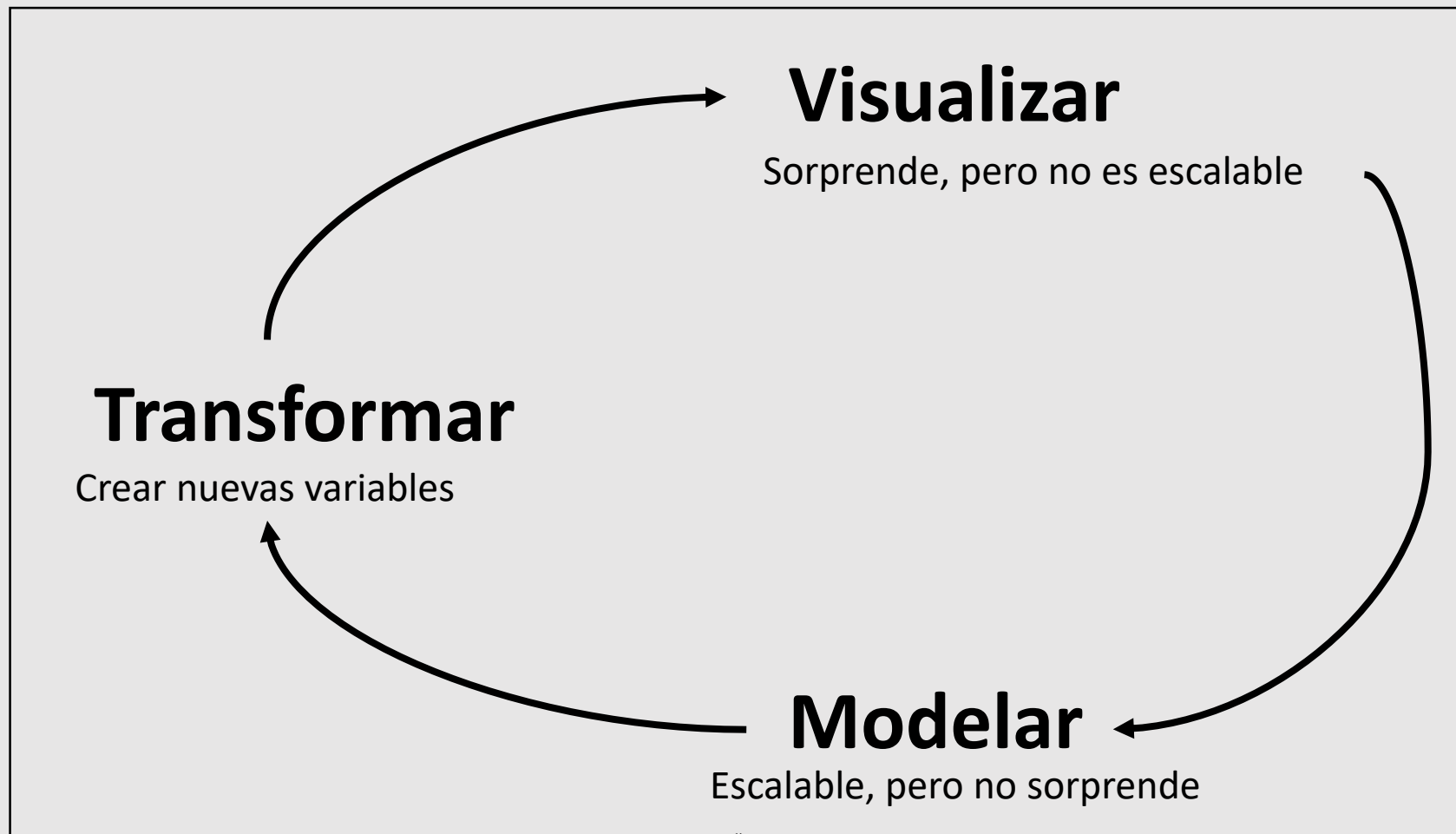
Sorprende, pero no es escalable

Modelar

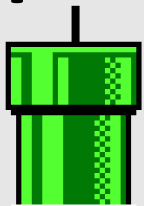
Escalable, pero no sorprende

Comunicar

Automatizar

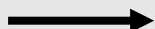


Importar



Tidy

Almacenar los
datos de manera
consistente



Transformar

Crear nuevas variables

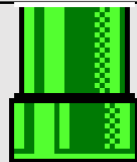
Visualizar

Sorprende, pero no es escalable

Modelar

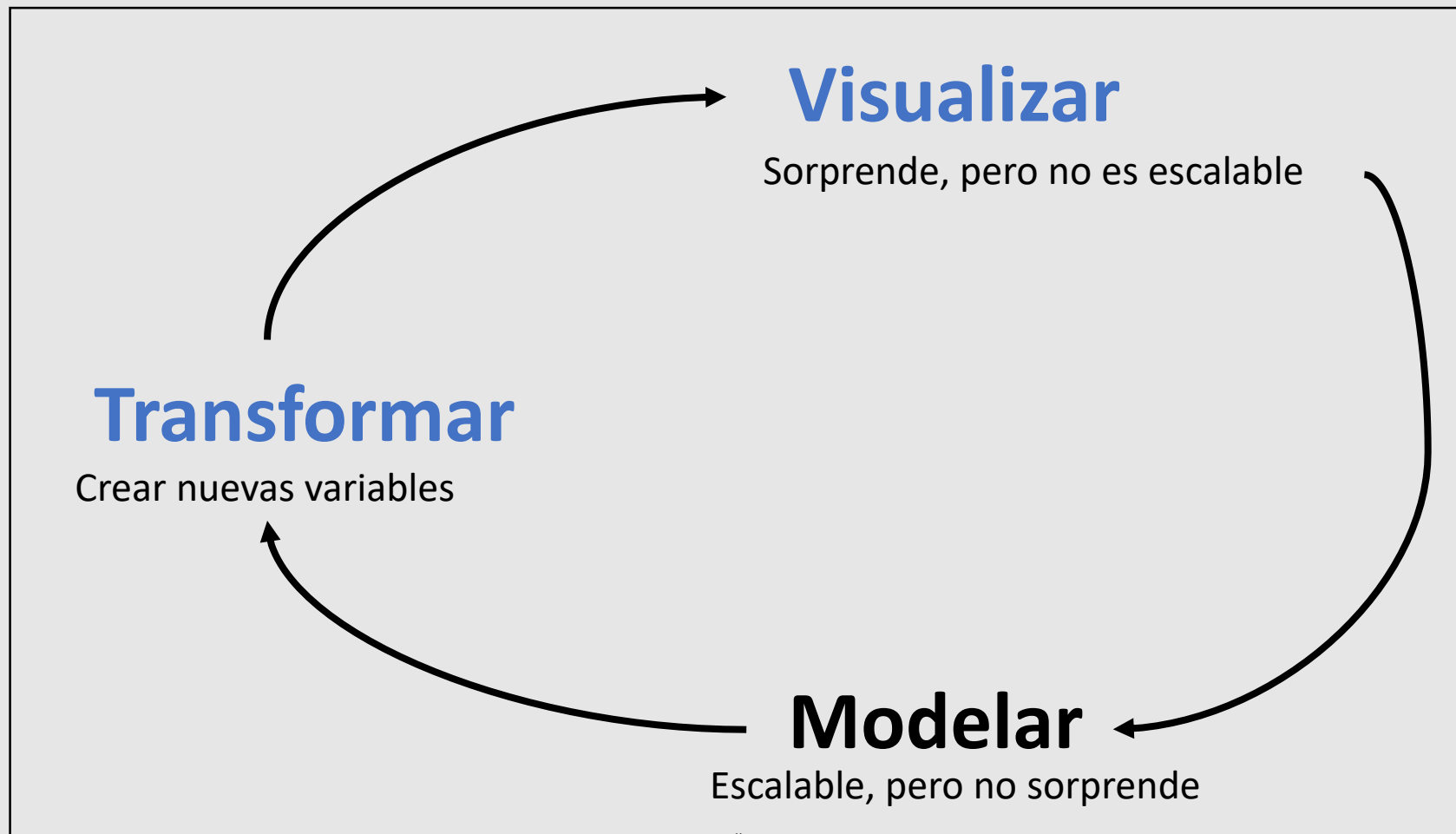
Escalable, pero no sorprende

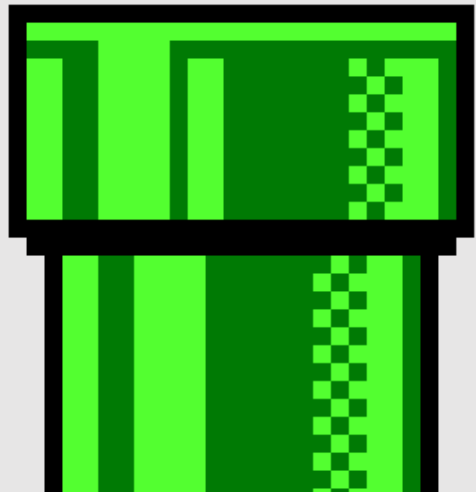
~ tubería de datos!



Comunicar

Automatizar





Implementaremos
esta **tubería** de
datos utilizando las
herramientas R y
Rstudio.

¿Qué es R?

Partamos con la definición de la página oficial.

“R is a free software environment for statistical computing and graphics”



Aterrizando la respuesta...

- R es **un lenguaje de programación**.
- Es **un software libre**, esto quiere decir que su licencia es gratis para su uso y distribución.
- Esta inspirado en el lenguaje de programación llamado S, creado en Bell Labs en los años 70 **para los proyectos de análisis de datos** del departamento de investigación estadística.

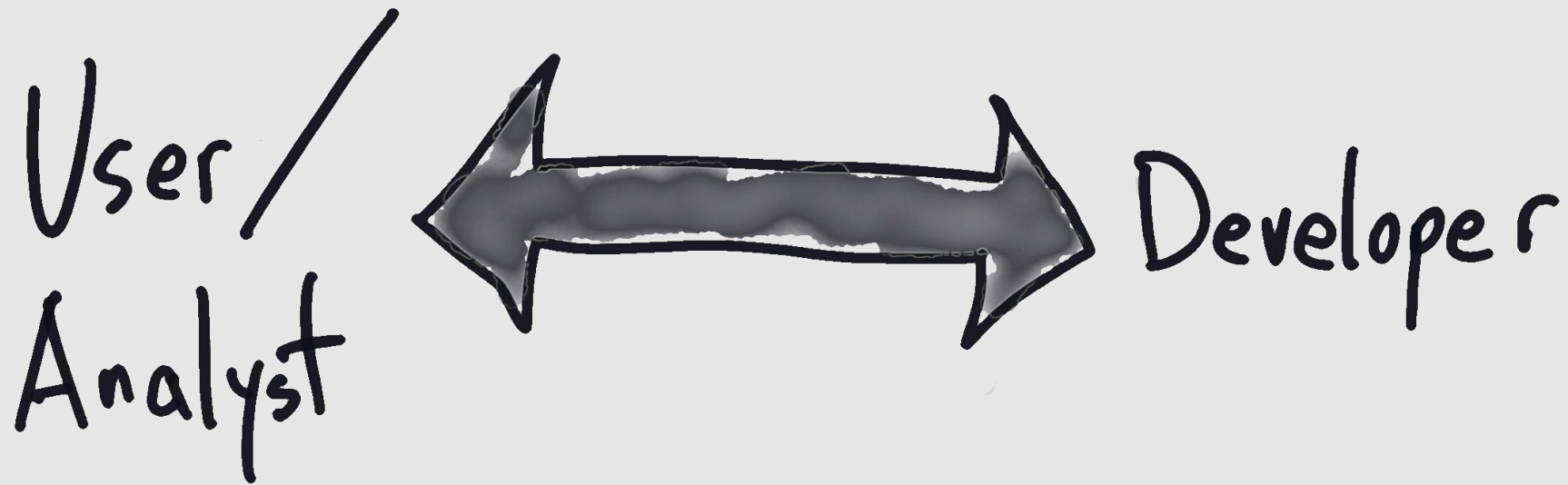
Aterrizando la respuesta...

- R aparte de ser una implementación gratuita de S, siguió con la filosofía original de este, **traducir las ideas en programas de manera rápida y confiables**.
- Esto se traduce en:
 - Una serie de rutinas ya implementadas para realizar análisis de datos, como computar un promedio, simular datos o visualizar una distribución de valores.
 - Un sistema flexible, con capacidades para programar y poder resolver situaciones de carácter no estándar.

En palabras del arquitecto de este sistema,
John Chambers:

*“Buscamos que los usuarios puedan iniciar en un **entorno interactivo**, en el que no se vean, conscientemente, a ellos mismos como programadores. Conforme sus necesidades sean más claras y su complejidad se incremente, **deberían gradualmente poder profundizar en la programación**, es cuando los aspectos del lenguaje y el sistema se vuelven más importantes”*

La filosofía en la génesis de R: *el espectro usuario – desarrollador.*



Los perfiles de los nuevos usuarios de la comunidad de R.

- No necesariamente hay una familiaridad con *software* de análisis de datos (Eviews, SAS, Stata, SPSS, etcétera).
- No hay experiencia previa en algún lenguaje de programación.
- Muchas veces es el primer encuentro con análisis de datos o estadística aplicada.

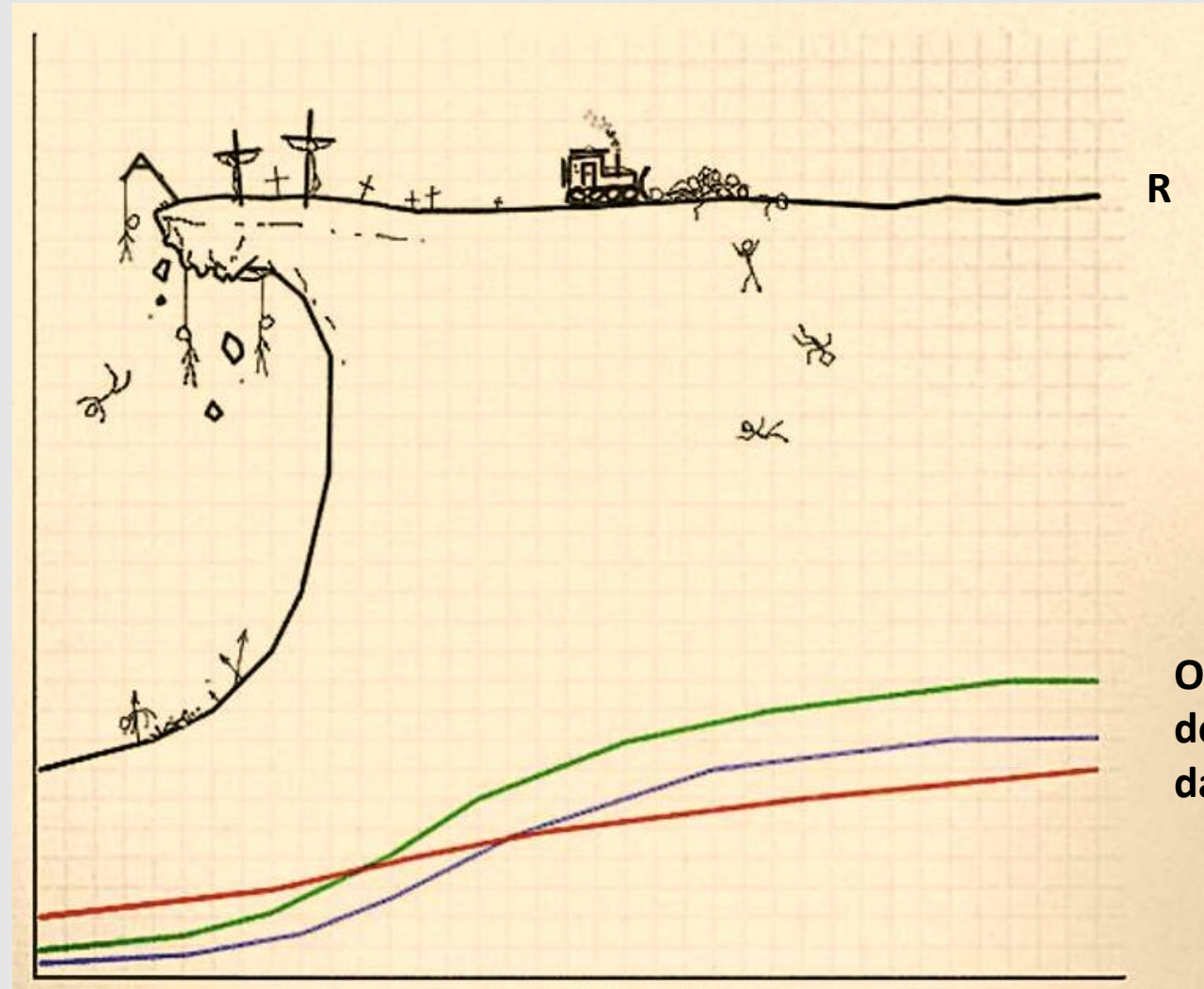
¿Qué ofrece de distintivo R a estos nuevos usuarios?

- La **propuesta de valor** es **flexibilidad y habilidad para programar** en el contexto de análisis de datos.



- Pero esto implica **mayor complejidad y requisitos en programación**.
- Nos encontramos más cerca del extremo “desarrollador” en el espectro.
- Esto también porque el **usuario objetivo inicial era un perfil muy distinto** a la amplia diversidad de nuevos usuarios en la actualidad.

Esta propuesta de valor tiene una **curva de aprendizaje bastante desmotivadora...**



Otros software
de análisis de
datos

No parece una buena idea.

- Sobre todo cuando estamos recién aprendiendo a analizar datos y queremos lograr operaciones bastantes puntuales.
- Una **promesa por el uso futuro** de una herramienta para hacer “análisis sofisticados”, **no es suficiente para sortear un montón de frustraciones** que incluso tendremos para tareas que se pueden hacer con excel.

Más usuarios de herramientas que desarrolladores de estas.

- Nuestro objetivo principal es **analizar datos** no programar. Por lo menos en un comienzo.

User/
Analyst ←

- Esto se justifica en que las operaciones más rutinarias en análisis de datos no requieren de mayores sofisticaciones y que **R base**, para los nuevos usuarios descritos, los hace bastante complejos.

Suavizando la curva de aprendizaje: **R base** y *tidyverse*.

- R es un *software* en constante desarrollo, sus **capacidades esenciales**, que llamaremos **R base**, pueden ser expandidas a través de paquetes. Un paquete es una forma estándar de compartir herramientas, esto a través de código y documentación.
- Por *tidyverse*, nos referimos a una **colección de paquetes** con herramientas para transformar y visualizar datos **que comparten una serie de convenciones**.



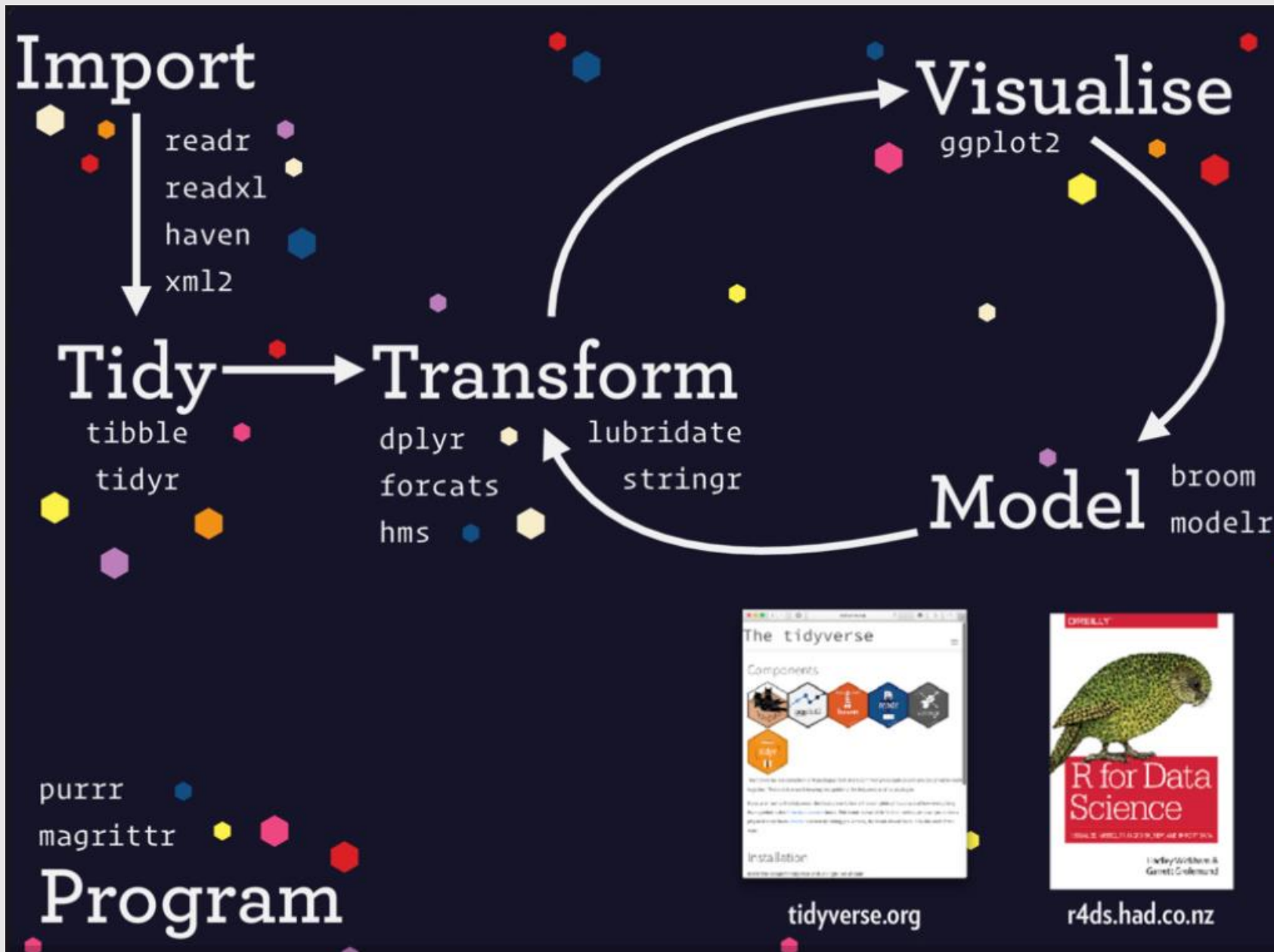
+



¿Paquetes de R?

- R es un celular nuevo.
- Podemos extender las funcionalidades nuevas de R descargando apps (paquetes). Esto en R sería: `install.packages("mi_app")`
- Una vez descargado, solo tenemos que abrir la app cada vez que queremos utilizarla. En R antes de ocupar las funcionalidades de un paquete, lo cargamos con: `library(mi_app)`





Resolver problemas
complejos al
**combinar simples
piezas** que tienen
una estructura
consistente.



Resolver problemas
complejos al
combinar **simples
piezas** que tienen
una estructura
consistente.



Cada
herramienta
realiza una sola
tarea

Resolver problemas
complejos al
combinar simples
piezas que tienen
**una estructura
consistente.**



Estas herramientas
están diseñadas
para operar sobre
data rectangular
(*dataframe*)

Un simple ejemplo de data rectangular...

```
> gapminder
# A tibble: 1,704 x 6
  country    continent  year lifeExp      pop gdpPercap
  <fct>      <fct>    <int>   <dbl>    <int>    <dbl>
1 Afghanistan Asia      1952    28.8  8425333    779.
2 Afghanistan Asia      1957    30.3  9240934    821.
3 Afghanistan Asia      1962    32.0 10267083    853.
4 Afghanistan Asia      1967    34.0 11537966    836.
5 Afghanistan Asia      1972    36.1 13079460    740.
6 Afghanistan Asia      1977    38.4 14880372    786.
7 Afghanistan Asia      1982    39.9 12881816    978.
8 Afghanistan Asia      1987    40.8 13867957    852.
9 Afghanistan Asia      1992    41.7 16317921    649.
10 Afghanistan Asia      1997    41.8 22227415    635.
# ... with 1,694 more rows
```

Este *dataframe*
tiene 1.704
observaciones y
6 variables.

Una ilustración de la filosofía detrás de *tidyverse*

```
> gapminder
# A tibble: 1,704 x 6
  country    continent  year lifeExp      pop gdpPercap
  <fct>      <fct>    <int>  <dbl>    <int>    <dbl>
1 Afghanistan Asia      1952   28.8  8425333    779.
2 Afghanistan Asia      1957   30.3  9240934    821.
3 Afghanistan Asia      1962   32.0 10267083    853.
4 Afghanistan Asia      1967   34.0 11537966    836.
5 Afghanistan Asia      1972   36.1 13079460    740.
6 Afghanistan Asia      1977   38.4 14880372    786.
7 Afghanistan Asia      1982   39.9 12881816    978.
8 Afghanistan Asia      1987   40.8 13867957    852.
9 Afghanistan Asia      1992   41.7 16317921    649.
10 Afghanistan Asia      1997   41.8 22227415    635.
# ... with 1,694 more rows
```

- Estructura consistente: entra un *dataframe*, sale un *dataframe*.
- **Simples piezas**: cada herramienta (función) hace una sola cosa.
- **Componemos** con estas piezas para resolver tareas más complejas.

gapminder %>%

`filter(year == 2007)` %>%

`group_by(continent)` %>%

`summarize(gdpPercap_promedio = mean(gdpPercap))`

```
# A tibble: 5 x 2
  continent gdpPercap_promedio
  <fct>      <dbl>
1 Africa      3089.
2 Americas  11003.
3 Asia      12473.
4 Europe     25054.
5 Oceania    3329810.
```

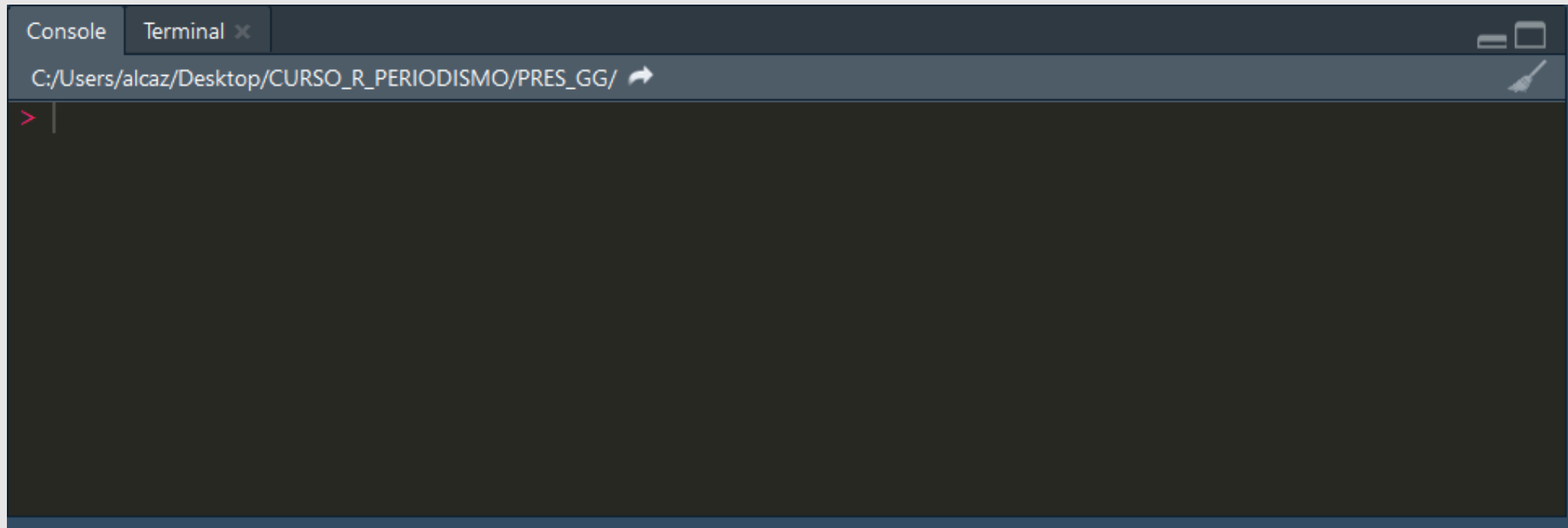
RStudio

- R es un lenguaje de programación.
- **Rstudio** es un **ambiente de desarrollo integrado**. En otras palabras, le agrega al lenguaje de programación R una interfaz con características y herramientas adicionales para tener una mayor productividad.
- En este curso utilizaremos R para el análisis de datos a través de Rstudio.



¿Porqué no usar un programa con interfaz para hacer análisis de datos?

¿Porqué usar esto?



Metodologías de enseñanza

- Mezcla de clases expositivas y de laboratorio.
- Sesiones de discusión.

