



Ciencia de datos: R

Conferencista: Felipe Calvo Cepeda
fcalvoc@unal.edu.co – fe.calvo@uniandes.edu.co

educación continua



UNIVERSIDAD
NACIONAL
DE COLOMBIA



Información

Fechas y horario



Martes, jueves: 6pm a 9pm
Sábado: 9am a 12pm

Metodología

- Clases magistrales teóricas
- Participación de las y los estudiantes
- Actividades de práctica en R
- Break intermedio

Evaluación

- Quices: 40%
- Taller: 30%
- Evaluación individual: 30%

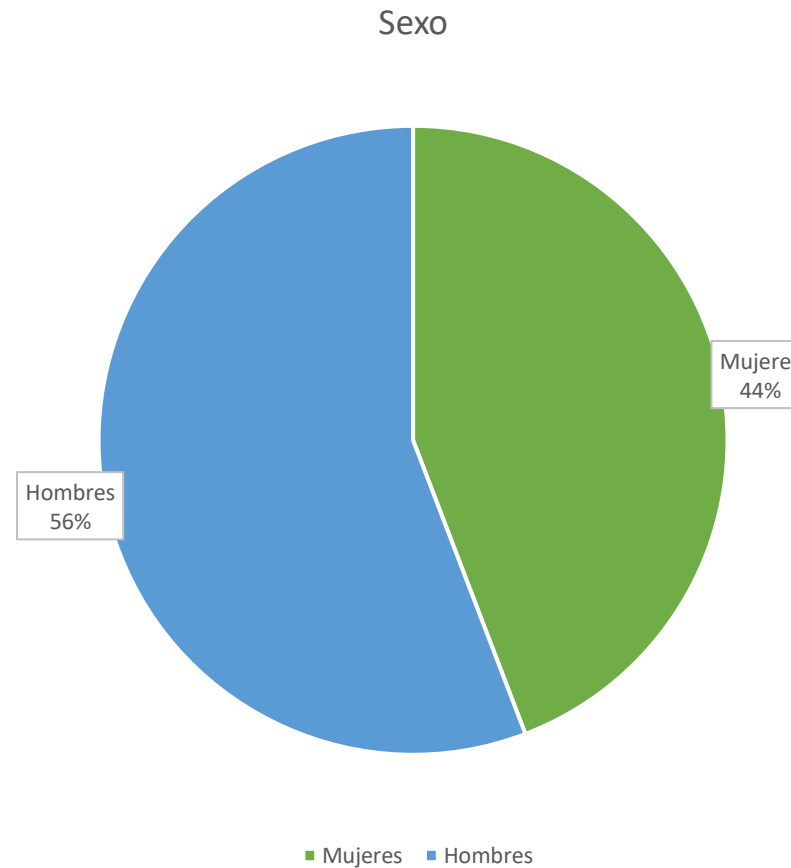
Certificados

- La Facultad de Ciencias Económicas de la Universidad Nacional de Colombia, otorgará un certificado de asistencia y/o aprobación del programa de Educación Continua, así:
- El certificado de asistencia se otorga a los estudiantes que cumplan con mínimo el ochenta por ciento (80%) de asistencia a los mismos.
- Los certificados de aprobación se entregan únicamente a quienes, además de cumplir con el mínimo de asistencia establecida obtengan un promedio de calificación final igual o superior a tres punto cero (3.0). Los certificados de aprobación son obligatorios para los Diplomados y para los cursos correspondientes a Formación a escala.

Temario

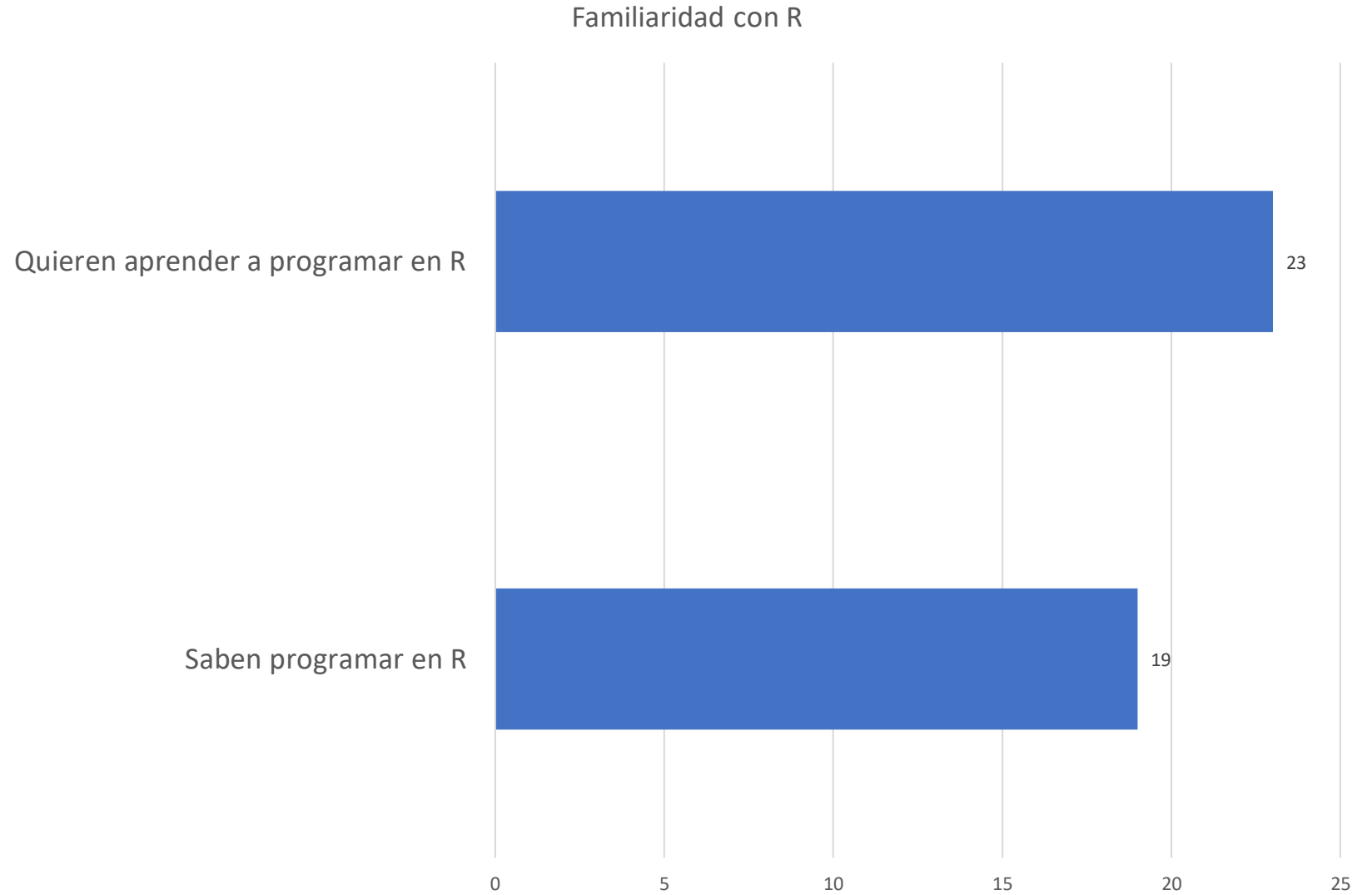
- Introducción, qué es R, instalación, paquetes, informes automáticos, proyectos y documentación. Carga de datos.
- Programación. Objetos y clases.
- Programación. Operaciones y funciones.
- Programación. Loops. Limpieza de datos, datos faltantes, datos atípicos, discretización de variables, trabajo con fechas y horas. Transformación de tablas de datos, crear nuevas columnas, generar resúmenes, desplegar y colapsar tablas. Operaciones entre tablas de datos. Inner join, left join., right join, full join.
- Datos univariados. Promedio, mediana, moda, varianza, cuartiles, rango intercuartílico.
- Datos multivariados. Covarianza, correlación, matriz de varianzas y covarianzas.
- Valor esperado y probabilidad condicionales.

Una ‘foto’ de ustedes

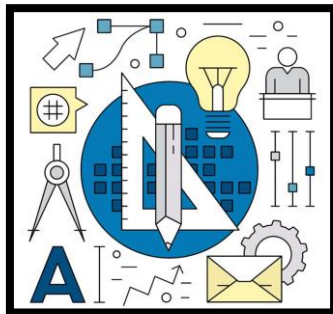
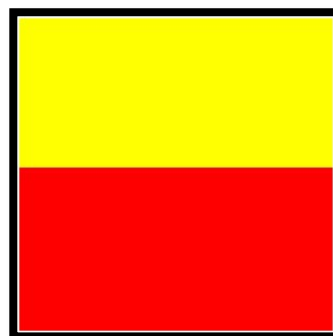


Profesión	Conteo
Economía	18
Ingeniería	6
Administración	5
Contaduría	2
Finanzas	2
No reportan	2
Bibliotecología	1
Biología	1
Comercio Internacional	1
Estadística	1
Periodismo	1
Ciencias Políticas	1
Química	1

Una “foto” de ustedes



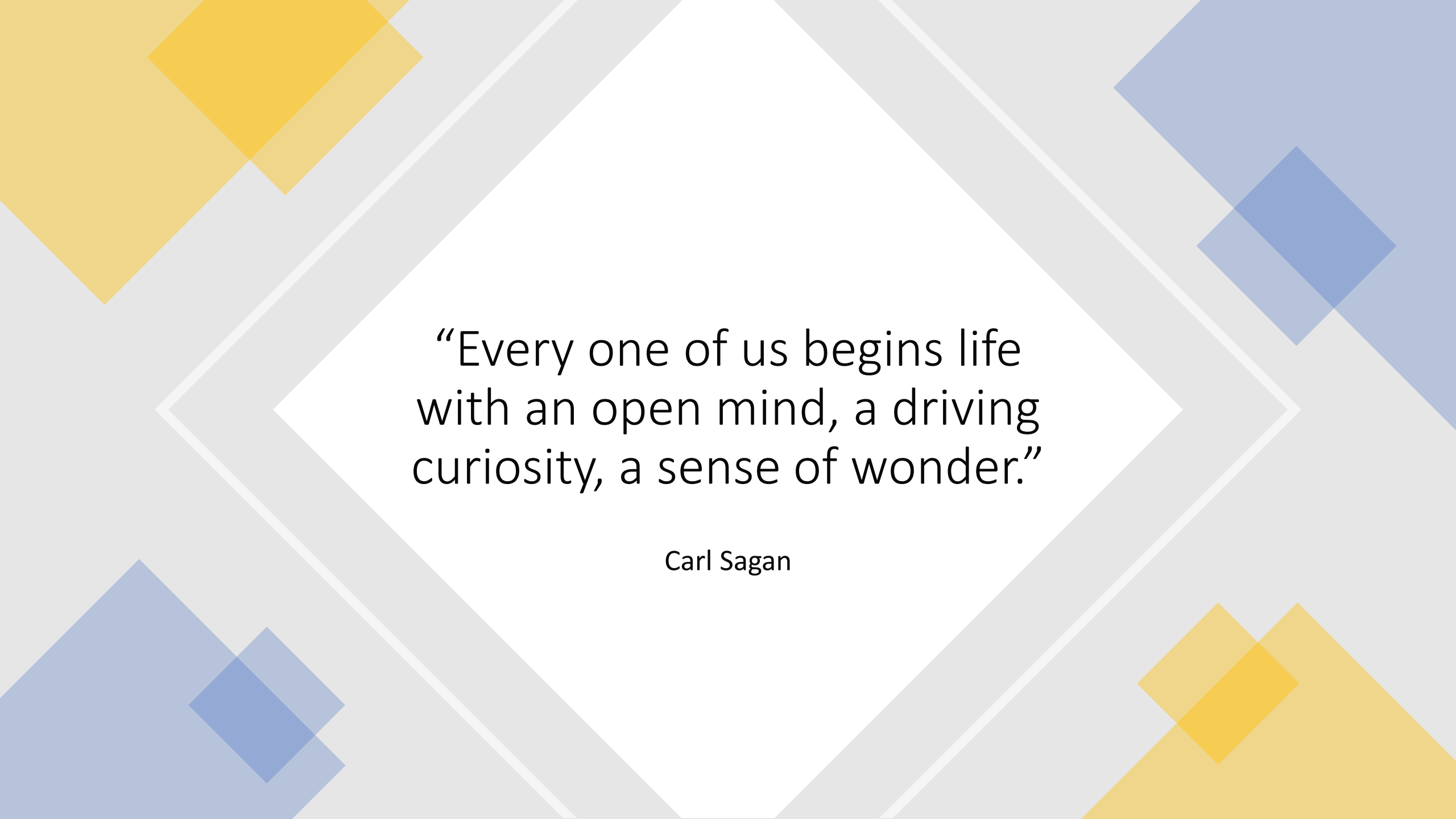
Una “foto”
mía



Algunos retos que ustedes tienen

- Aprender a programar
- Habilidades de análisis de datos
- Investigación
- Modelado
- Visualización de datos
- Estadística
- Machine learning
- Automatización de procesos
- Ir más allá de Excel

Sus retos son
importantes



“Every one of us begins life
with an open mind, a driving
curiosity, a sense of wonder.”

Carl Sagan

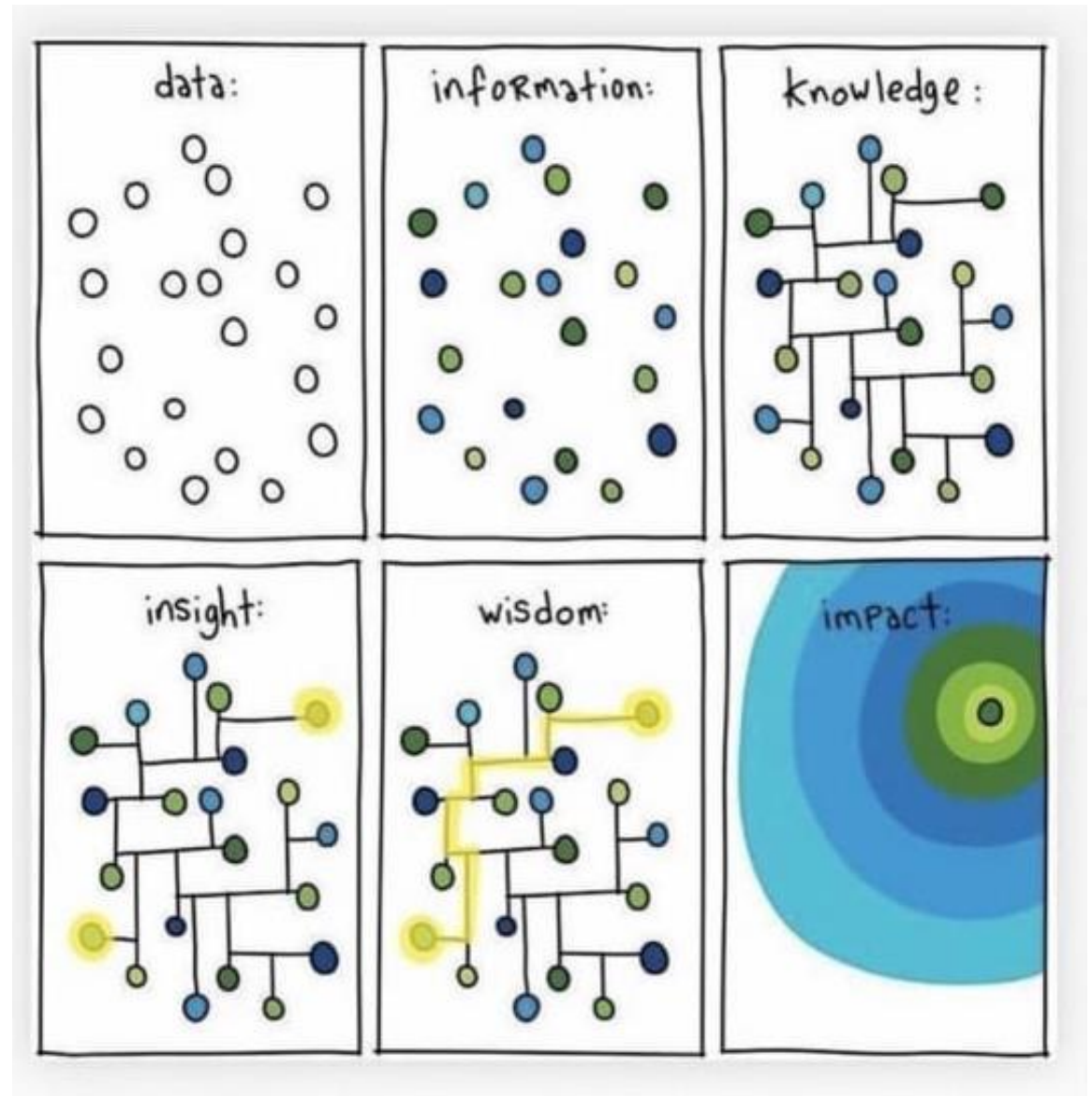
En la ciencia de
datos, los datos
son la segunda
cosa más
importante

- Lo más importante es una pregunta
- Lo segundo más importante son los datos
- Generalmente los datos limitan o permiten las preguntas
- Sin embargo, tener datos no habilita nada si detrás no hay una pregunta
- Los métodos estadísticos no sustituyen un buen diseño de investigación

Preguntas del mundo hoy

- ¿Donald Trump será reelegido?
- ¿Qué podemos hacer con la epidemia de noticias falsas?
- ¿Está la democracia liberal en crisis? Si sí, ¿por qué?
- ¿Se aproxima una nueva guerra mundial?
- ¿Qué civilización domina el mundo?
- ¿Tendría Europa que abrir sus puertas a los inmigrantes?
- ¿Puede el nacionalismo resolver los problemas de desigualdad y de cambio climático?

Una metáfora



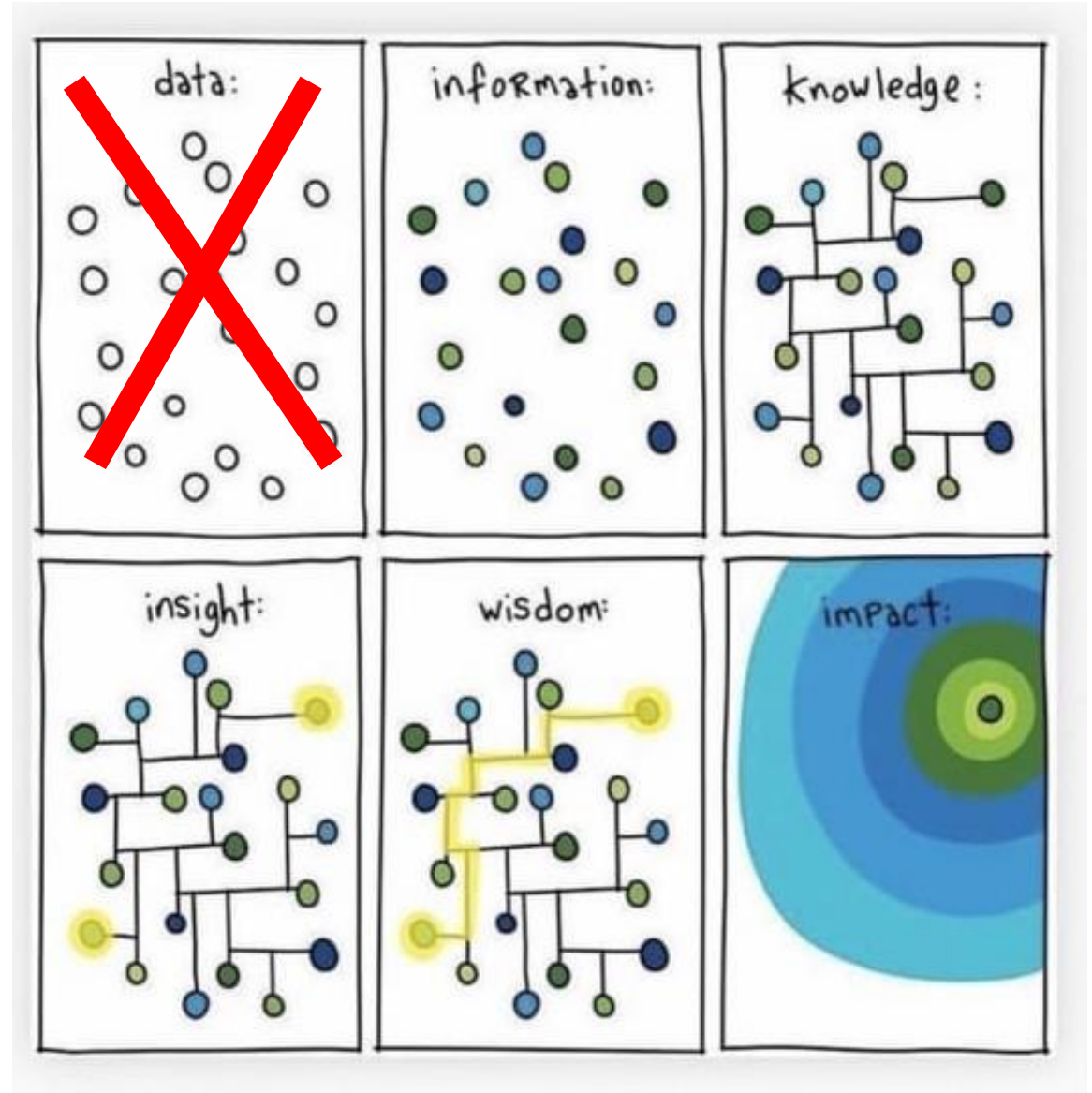
Hipótesis para algunas preguntas del mundo hoy

- ¿Donald Trump será reelegido? **Sí**
- ¿Qué podemos hacer con la epidemia de noticias falsas? **No censurarlas**
- ¿Está la democracia liberal en crisis? Si sí, ¿por qué? **Sí**
- ¿Se aproxima una nueva guerra mundial? **No**
- ¿Qué civilización domina el mundo? **NS/NR**
- ¿Tendría Europa que abrir sus puertas a los inmigrantes? **Sí**
- ¿Puede el nacionalismo resolver los problemas de desigualdad y de cambio climático? **No**

¿Tenemos datos para resolver esas preguntas?

- ¿Donald Trump será reelegido?
- ¿Qué podemos hacer con la epidemia de noticias falsas?
- ¿Está la democracia liberal en crisis? Si sí, ¿por qué?
- ¿Se aproxima una nueva guerra mundial?
- ¿Qué civilización domina el mundo?
- ¿Tendría Europa que abrir sus puertas a los inmigrantes?
- ¿Puede el nacionalismo resolver los problemas de desigualdad y de cambio climático?

Otra metáfora



En la ciencia de datos

- Podemos
 - aprender
 - tomar decisiones
 - presentar conclusionessi tenemos datos.
- Pero a veces (bastantes veces) no los tenemos.

El camino largo

- ¡Hay que recolectarlos!
- Diseñar una estrategia de recolección

A, B y C

- A. El 80% de la ciencia de datos tiene que ver con tener datos y tenerlos bien estructurados y limpios*.
- B. Para el curso vamos a trabajar con algunos datos bien estructurados.
- C. Les invito a poner en práctica los temas vistos con datos de su propio interés profesional o académico o recreativo.

Cómo se ven
los datos

```
@HWI-EAS121:4:100:1783:550#0/1
CGTTACGAGATCGGAAGAGCGGTTTCAGCAGGAATGCCGAGACGGATCTCGTATGCCGTCTGCTGCGTGACAAGACAGGGG
+HWI-EAS121:4:100:1783:550#0/1
aaaaa`b_aa`aa`YaX]aZ`aZM^Z]YRa]YSG[ [ZREQLHESDHNDHNMEEDDMPENITKFLFEEDDDHEJQMEDDD
@HWI-EAS121:4:100:1783:1611#0/1
GGGTGGGCATTTCCACTCGCAGTATGGGTTGCCGCACGACAGGCAGCGGTCAGCCTGCGCTTGGCCTGGCCTTCGGAAA
+HWI-EAS121:4:100:1783:1611#0/1
a``^`_`_`_`^a`a`a`a`_`a`_]a`_]`a`_`_`_`^`^`]X`_]XTV`_]NX_XVX`]T`TTTG[VTHPN]VFDZ
@HWI-EAS121:4:100:1783:322#0/1
CGTTTATGTTTTTGAATATGTCTTATCTTAACGGTTATATTTTAGATGTTGGTCTTATTCTAACGGTCATATATTTTCTA
+HWI-EAS121:4:100:1783:322#0/1
abaa`^aaaaabbaababbbbbb`bbbb_bbbbbbbb`bbbaV^_a``a``]``aT]a__V`]]`^a`]a_abbaV__
@HWI-EAS121:4:100:1783:1394#0/1
GGGTCTTTATTGGTCTGGTGATCCCCCATATTCTCCGGTTGTGTGGTTTAACCGATCATCGCGCATTACTTCCCGGCTGC
+HWI-EAS121:4:100:1783:1394#0/1
````[aa\b^^[ ]aabb]`a`abbb`a``bbbbbbababaaaab_VZa`^__bab_X`[a\HV`[_]`[_X`T_VQQ
@HWI-EAS121:4:100:1783:207#0/1
CCCTGGGAGATCGGAAGAGCGGTTTCAGCAGGAATGCCGAGACCGATCTCGTATGCCGTCTTCTGCTTGAAAAAAAAAACA
+HWI-EAS121:4:100:1783:207#0/1
abba`Xa\`^\`aa]ba__bba[a_O_a`aa`aa`a]^V]X_a^YS\R_\H_[]\ZTDUZZUSOPX]]POP\GS\WSHHD
@HWI-EAS121:4:100:1783:455#0/1
GGGTAATTCAGGGACAATGTAATGGCTGCACAAAAAATACATCTTTCATGTTCCATTGCACCATTGACAAATACATATT
+HWI-EAS121:4:100:1783:455#0/1
abb_babbababbbbbbbbbbbbbbbba\`b`\abbbabbbbabbbbbbaabbbbb`bb`ab_O_bab_Q_bbabaa_a
```



Cómo se ven  
los datos

## Example Response

```
{
 "contributors_enabled": true,
 "created_at": "Sat May 09 17:58:22 +0000 2009",
 "default_profile": false,
 "default_profile_image": false,
 "description": "I taught your phone that thing you like. The Mobile Partner Engineer @Twitter. ",
 "entities": {
 "description": {
 "urls": []
 }
 },
 "favourites_count": 586,
 "follow_request_sent": false,
 "followers_count": 10622,
 "following": false,
 "friends_count": 1181,
 "geo_enabled": true,
 "id": 38895958,
 "id_str": "38895958",
 "is_translator": false,
 "lang": "en",
 "listed_count": 190,
 "location": "San Francisco",
 "name": "Sean Cook",
 "notifications": false,
 "profile_background_color": "1A1B1F",
 "profile_background_image_url": "http://a0.twimg.com/profile_background_images/495742332/purty_wood.png",
 "profile_background_image_url_https": "https://si0.twimg.com/profile_background_images/495742332/purty_"
```

Cómo se ven  
los datos

ALLERGIES	MEDICATION HISTORY
Last Updated: 01 Dec 2011 @ 0851	Last Updated: 11 Apr 2011 @ 1737
Allergy Name: TRIMETHOPRIM	Medication: AMLODIPINE BESYLATE 10MG TAB
Location: DAYT29	Instructions: TAKE ONE TABLET BY MOUTH TAKE ONE-HALF TABLET FOR GRAPEFRUIT JUICE--
Date Entered: 09 Mar 2011	Status: Active
Reaction:	Refills Remaining: 3
Allergy Type: DRUG	Last Filled On: 20 Aug 2010
Drug Class: ANTI-INFECTIVES,OTHER	Initially Ordered On: 13 Aug 2010
Observed/Historical: HISTORICAL	Quantity: 45
Comments: The reaction to this allergy was MILD (NO SQUELAE)	Days Supply: 90
Allergy Name: TRAMADOL	Pharmacy: DAYTON
Location: DAYT29	Prescription Number: 2718953
Date Entered: 09 Mar 2011	Medication: IBUPROFEN 600MG TAB
Reaction: URINARY RETENTION	Instructions: TAKE ONE TABLET BY MOUTH FOUR TIMES A DAY WITH FOOD
Allergy Type: DRUG	Status: Active
Drug Class: NON-OPIOID ANALGESICS	Refills Remaining: 3
Observed/Historical: HISTORICAL	Last Filled On: 20 Aug 2010
Comments: gradually worsening difficulty emptying bladder	Initially Ordered On: 01 Jul 2010
Tramadol was not taken because it was not needed for relief	Quantity: 300

Cómo se ven  
los datos



Cómo se ven  
los datos



Cómo se ven  
los datos  
*(rara vez)*

iris					
	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
9	4.4	2.9	1.4	0.2	setosa
10	4.9	3.1	1.5	0.1	setosa
11	5.4	3.7	1.5	0.2	setosa
12	4.8	3.4	1.6	0.2	setosa
13	4.8	3.0	1.4	0.1	setosa
14	4.3	3.0	1.1	0.1	setosa
15	5.8	4.0	1.2	0.2	setosa
16	5.7	4.4	1.5	0.4	setosa
17	5.4	3.9	1.3	0.4	setosa
18	5.1	3.5	1.4	0.3	setosa
19	5.7	3.8	1.7	0.3	setosa
20	5.1	3.8	1.5	0.3	setosa
21	5.4	3.4	1.7	0.2	setosa
22	5.1	3.7	1.5	0.4	setosa
23	4.6	3.6	1.0	0.2	setosa
24	5.1	3.3	1.7	0.5	setosa
25	4.8	3.4	1.9	0.2	setosa
26	5.0	3.0	1.6	0.2	setosa
27	5.0	3.4	1.6	0.4	setosa
28	5.2	3.5	1.5	0.2	setosa
29	5.2	3.4	1.4	0.2	setosa
30	4.7	3.2	1.6	0.2	setosa



## Roles en proyectos con datos



- Ingeniería de datos
  - Obtener los datos
  - Limpiarlos y estructurarlos para posterior análisis
  - Crear pipelines de análisis automatizado
  - Utilización de herramientas en la nube
  - Análisis descriptivo de los datos
- Ciencia de datos
  - Análisis matemático de los datos
  - Identificación de variables relevantes / features
  - Generación de modelos predictivos y prescriptivos
- Profesionales de modelado ML
  - Creación de sistemas predictivos y prescriptivos de gran escala
  - Mantenimiento y ajuste del modelo



# Practica



- Configuración del ambiente de trabajo
- Importar datos a R

Consulte los repositorios del curso, los cuales quedarán alojados en una carpeta de Github que las y los estudiantes podrán clonar a sus cuentas personales o descargar a sus equipos personales para referencia y consulta posterior.

# Practica



- Programación básica
  - Algoritmos
  - Tipos de variables
  - Conversiones
  - Operaciones
  - Vectores, factores, matrices, arreglos y tablas
  - Extracción
  - Control flow

Consulte los repositorios del curso, los cuales quedarán alojados en una carpeta de Github que las y los estudiantes podrán clonar a sus cuentas personales o descargar a sus equipos personales para referencia y consulta posterior.

# Practica



- Limpieza de datos
  - Datos faltantes
    - Imputación de datos
  - Datos atípicos
    - Abordaje univariado
    - Abordaje multivariado
- Naming
- Valores duplicados
- Discretización de variables

Consulte los repositorios del curso, los cuales quedarán alojados en una carpeta de Github que las y los estudiantes podrán clonar a sus cuentas personales o descargar a sus equipos personales para referencia y consulta posterior.

# Análisis exploratorio de datos

- Habilidad
  - Desarrollar una idea aproximada de cómo se ve un conjunto de datos y qué tipo de preguntas pueden responder.
- Proceso
  - Requiere el conocimiento y aplicación de diversas técnicas matemáticas, estadísticas y gráficas.
- Requisito
  - Antes de modelar o formular hipótesis.

# Análisis exploratorio de datos



# Análisis exploratorio de datos

Resúmenes numéricos  
de un conjunto de datos

Resúmenes gráficos de  
un conjunto de datos

# Análisis exploratorio de datos

## Resúmenes numéricos de un conjunto de datos

- Medidas de tendencia central
- Medidas de variabilidad

## Resúmenes gráficos de un conjunto de datos

- *Una imagen vale más que mil palabras*





## COVID-19 Dashboard by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University (JHU)



## Global Cases

22.252.446

Cases by  
Country/Region/Sovereignty

5.523.826 US

3.407.354 Brazil

2.767.273 India

935.066 Russia

596.060 South Africa

549.321 Peru

531.239 Mexico

489.122 Colombia

390.037 Chile

370.867 Spain

Admin0

Last Updated at (M/D/YYYY)

8/19/2020 5:27:47 p. m.



Cumulative Cases

Active Cases

Incidence Rate

Case-Fatality Ratio

Testing Rate

Hospitalization Rate

188

countries/regions

Lancet Inf Dis Article: [Here](#). Mobile Version: [Here](#). Data sources: [Full list](#). Downloadable database: [GitHub](#), [Feature Layer](#).Lead by JHU CSSE. Technical Support: [Esri Living Atlas team](#) and [JHU APL](#). Financial Support:

## Global Deaths

783.825

172.945 deaths  
US109.888 deaths  
Brazil57.774 deaths  
Mexico52.889 deaths  
India41.483 deaths  
United Kingdom

Global Deaths

## US State Level

## Deaths, Recovered

32.865 deaths, 74.258  
recovered  
New York US15.926 deaths, 33.403  
recovered  
New Jersey US11.606 deaths, recovered  
California US10.895 deaths, 415.903  
recovered

US Deaths, Reco...



Daily Cases

## 1. Fin de la pobreza

### METAS

#### 1.1. Erradicar la extrema pobreza

Porcentaje de población que vive por debajo del umbral internacional de pobreza extrema

Incidencia de la Pobreza Monetaria Extrema

#### 1.2. Reducir la pobreza en, al menos, un 50%

Incidencia de la Pobreza Monetaria

Índice de Pobreza Multidimensional (IPM)

#### 1.3. Implementar sistemas de protección social

Porcentaje de población afiliada al sistema de seguridad social en salud

Porcentaje de población ocupada afiliada a

## Porcentaje de población que vive por debajo del umbral internacional de pobreza extrema

(%)



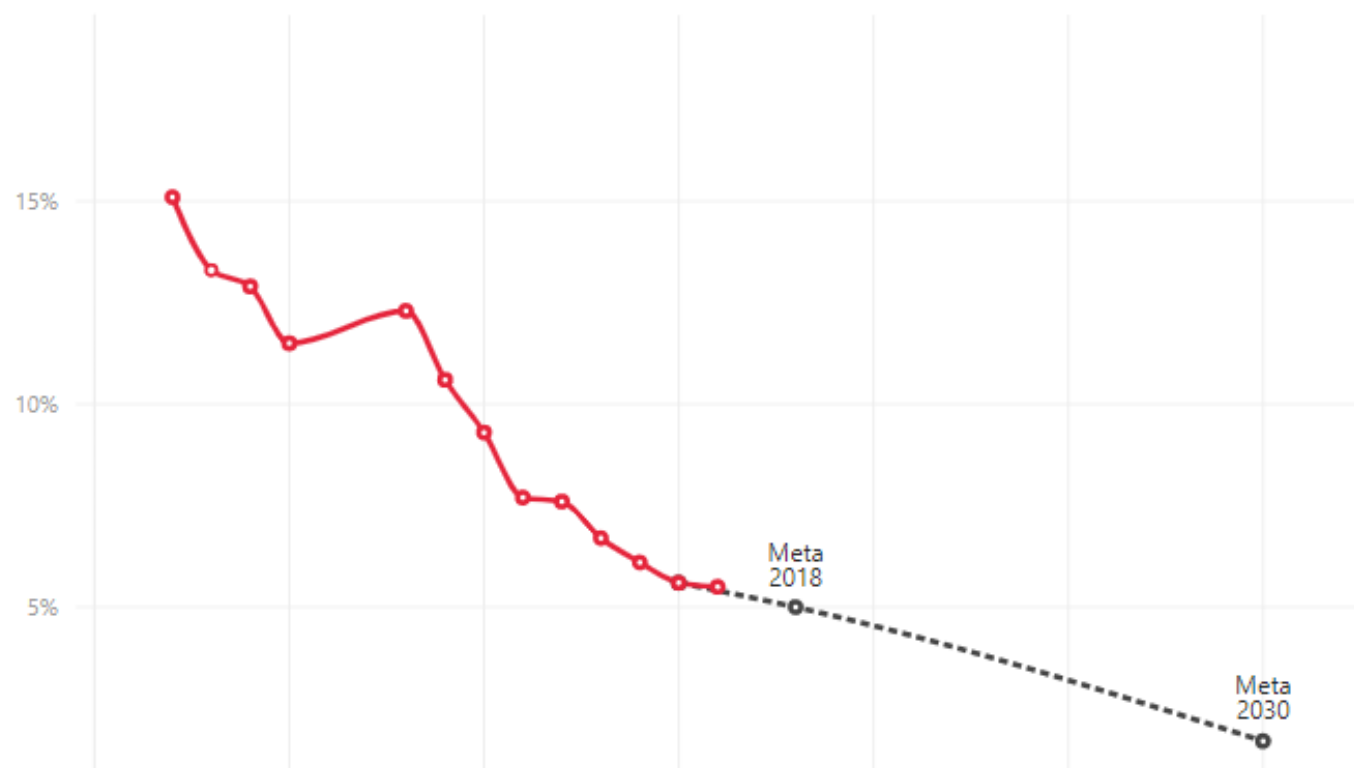
PAÍS

SEXO

ZONA

GRUPOS DE EDAD

MOSTRAR META A 2030



# Practica



- Estadística exploratoria
  - Descriptivos
  - Métodos de resumen
  - Gráficas
  - Visualizaciones avanzadas

Consulte los repositorios del curso, los cuales quedarán alojados en una carpeta de Github que las y los estudiantes podrán clonar a sus cuentas personales o descargar a sus equipos personales para referencia y consulta posterior.

# Dimensio- nalidad de los datos

- **Univariados**
  - Medición de una variable en un sujeto/unidad
- **Bivariados**
  - Medición de dos variables en un sujeto/unidad
- **Multivariado**
  - Medición de múltiples variables en un sujeto/unidad

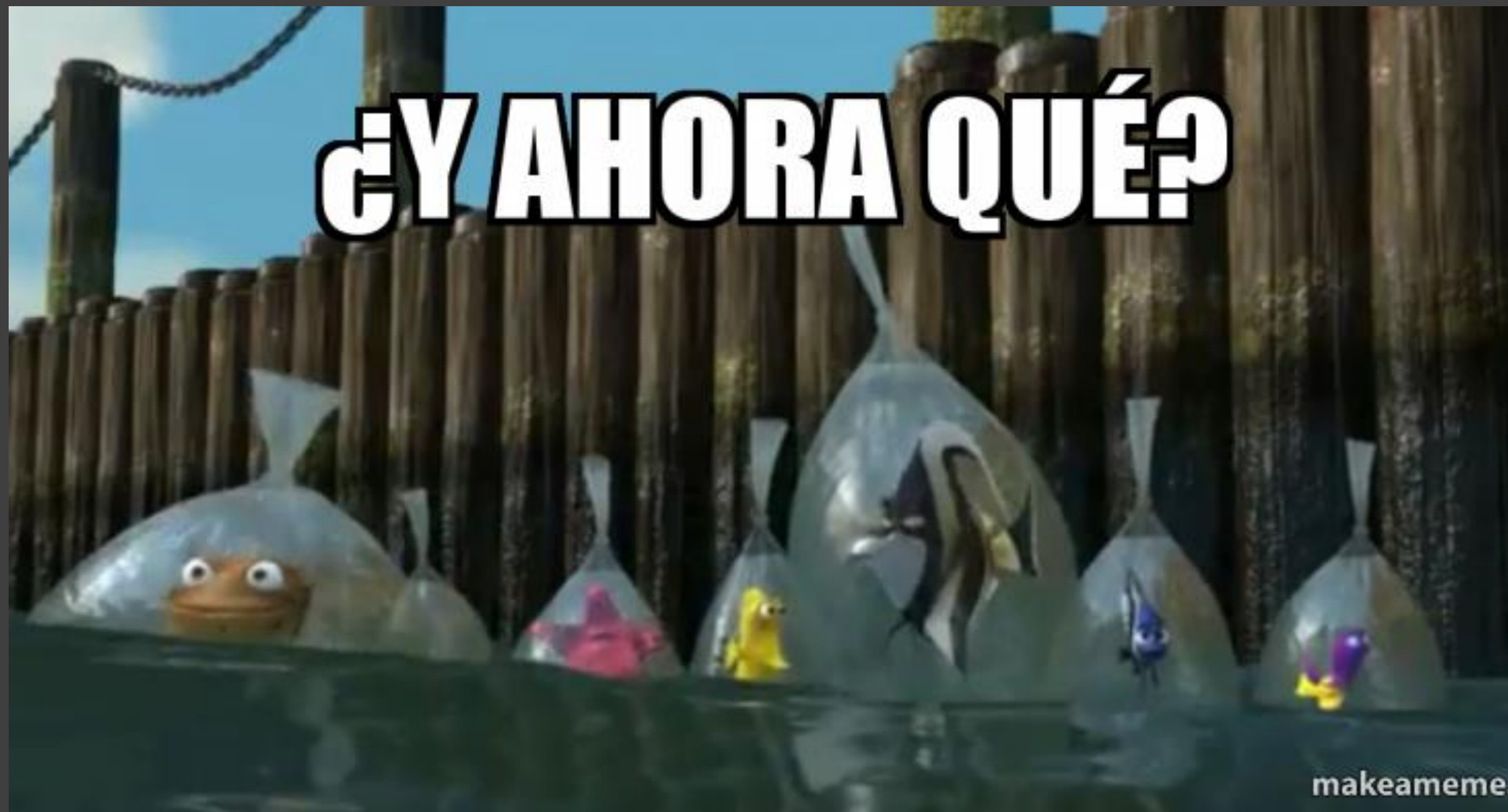
# Practica



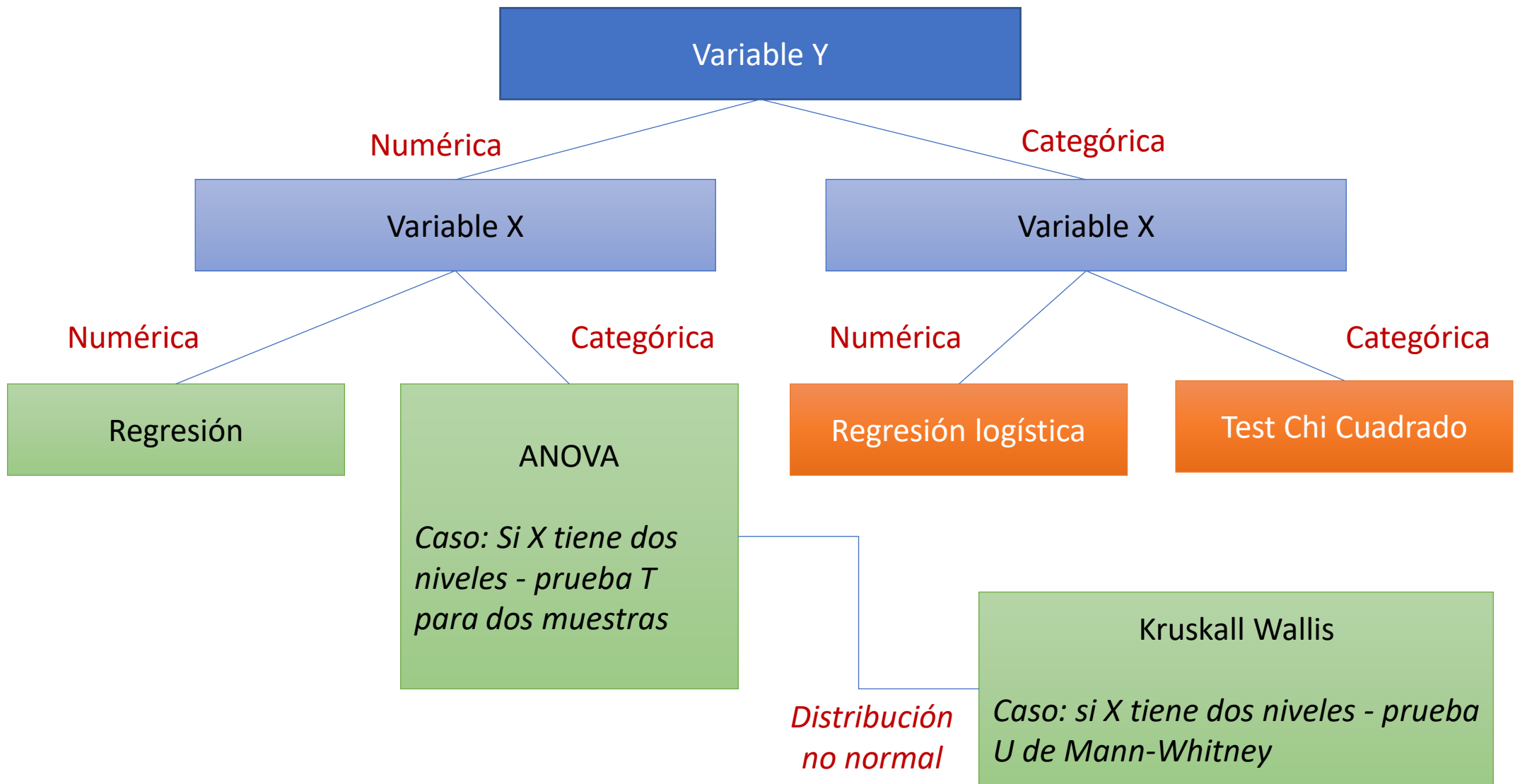
- Importación y descripción del conjunto de datos
- Limpieza de datos
- Análisis exploratorio
- Asociación e independencia
  - Métodos numéricos
  - Métodos gráficos
- Principios de análisis multivariado

Consulte los repositorios del curso, los cuales quedarán alojados en una carpeta de Github que las y los estudiantes podrán clonar a sus cuentas personales o descargar a sus equipos personales para referencia y consulta posterior.

**¿Y AHORA QUÉ?**

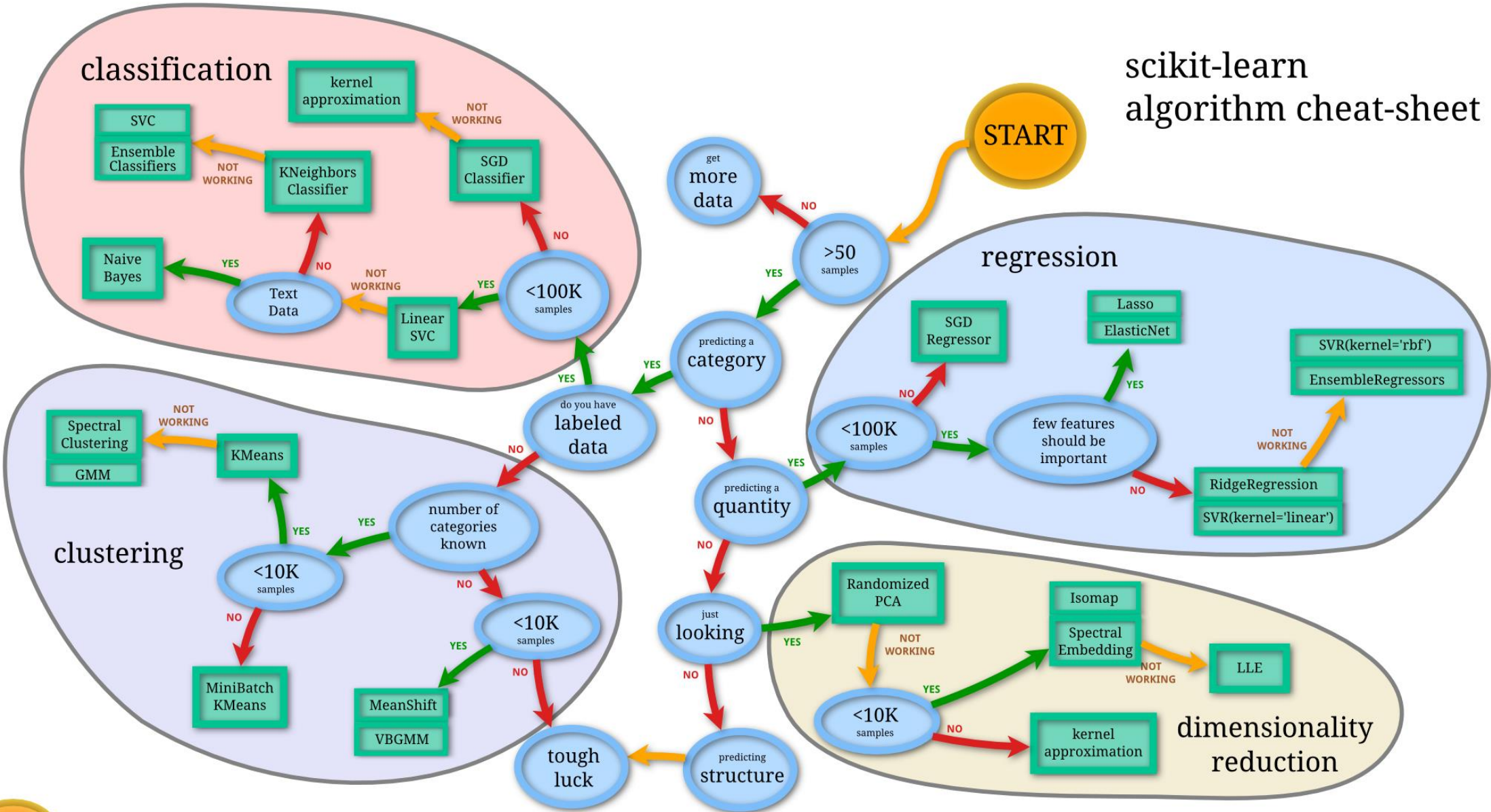


makeameme





scikit-learn  
algorithm cheat-sheet







# Ciencia de datos: R

---

Conferencista: Felipe Calvo Cepeda  
fcalvoc@unal.edu.co – fe.calvo@uniandes.edu.co

*educación continua*



UNIVERSIDAD  
**NACIONAL**  
DE COLOMBIA