Ciencia de Datos para Políticas Públicas

Clase 01 - Introducción

Pablo Aguirre Hormann

05/08/2020

Logística del curso

Algunos acuerdos

- Cámaras abiertas cuando sea posible (y si el internet lo permite)
- Micrófonos cerrados
  - A menos que tengan preguntas
- PPT se subirá a CANVAS después de cada clase
- Aprender haciendo
  - Demostraciones, ejercicios, tareas
- **Dedicación de tiempo**

Información general (i)

- 10 clases
  - última clase: presentaciones
- Sin ayudante
- Consultas por CANVAS fuera del horario de clases
- Hora de consulta (Zoom): Todos los lunes 18:00-19:00
  - A menos que se diga lo contrario

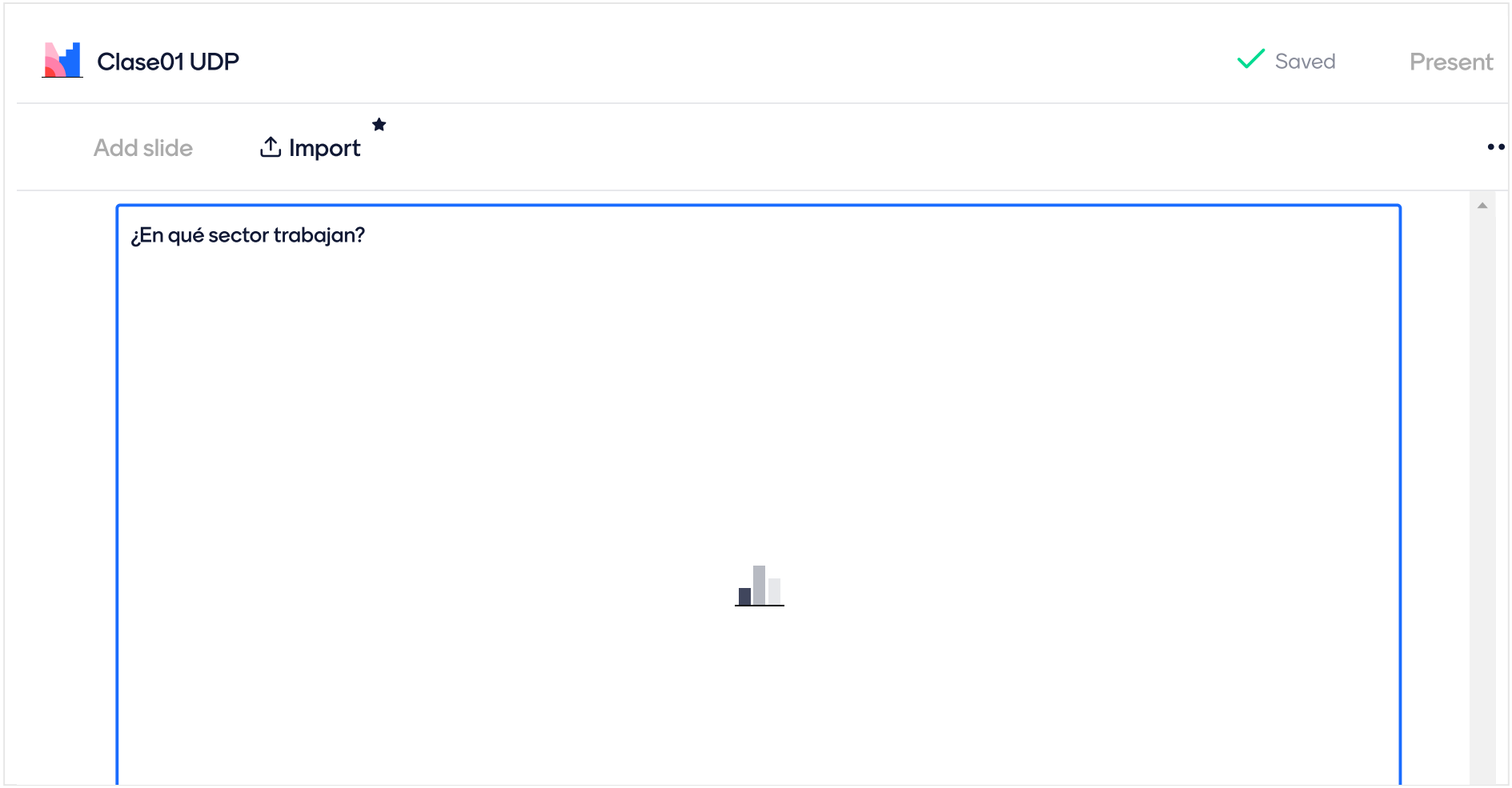
Información general (ii)

- 4 tareas: 40% (10% c/u)
- Trabajo: 50%
  - Informe preliminar (10%)
  - Informe final (20%)
  - Presentación (20%)
- Participación

Me presento

- Pablo Aguirre Hörmann - pjaguirreh@gmail.com
  - Ing. Agrónomo (UC) y MPP (U. of Chicago)
  - Análisis e Inteligencia de Negocios - Superintendencia del Medio Ambiente
  - https://github.com/pjaguirreh
  - @PAguirreH

Ahora ustedes



¿De qué se trata este curso?

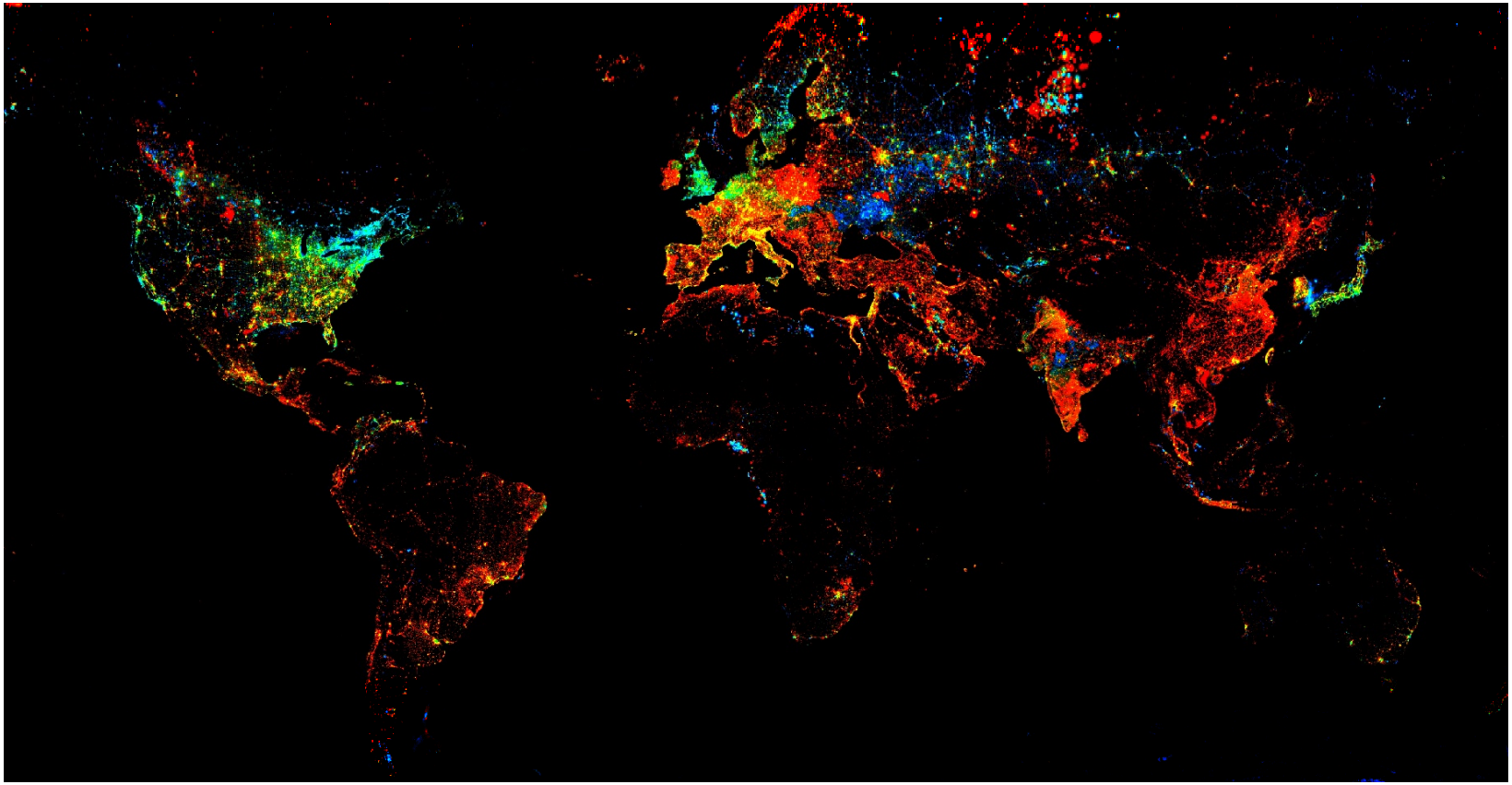
¿De qué se trata este curso?

Usar datos…

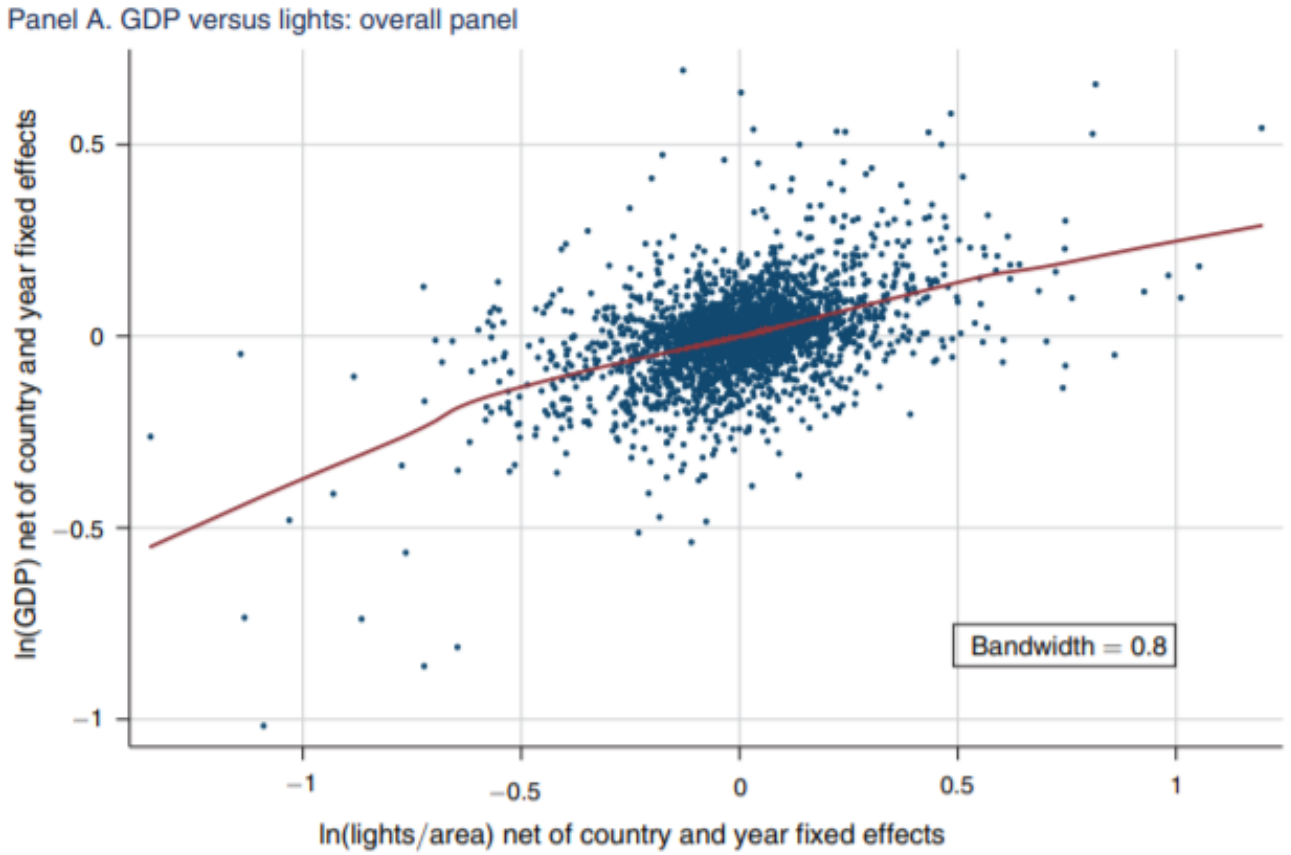
- para facilitar tareas
- para aprender algo
- para informar decisiones
- para el bien común (¿?)

Muchos datos en la actualidad

Los datos pueden ser bonitos



Los datos pueden ser informativos



Los datos pueden ser poderosos

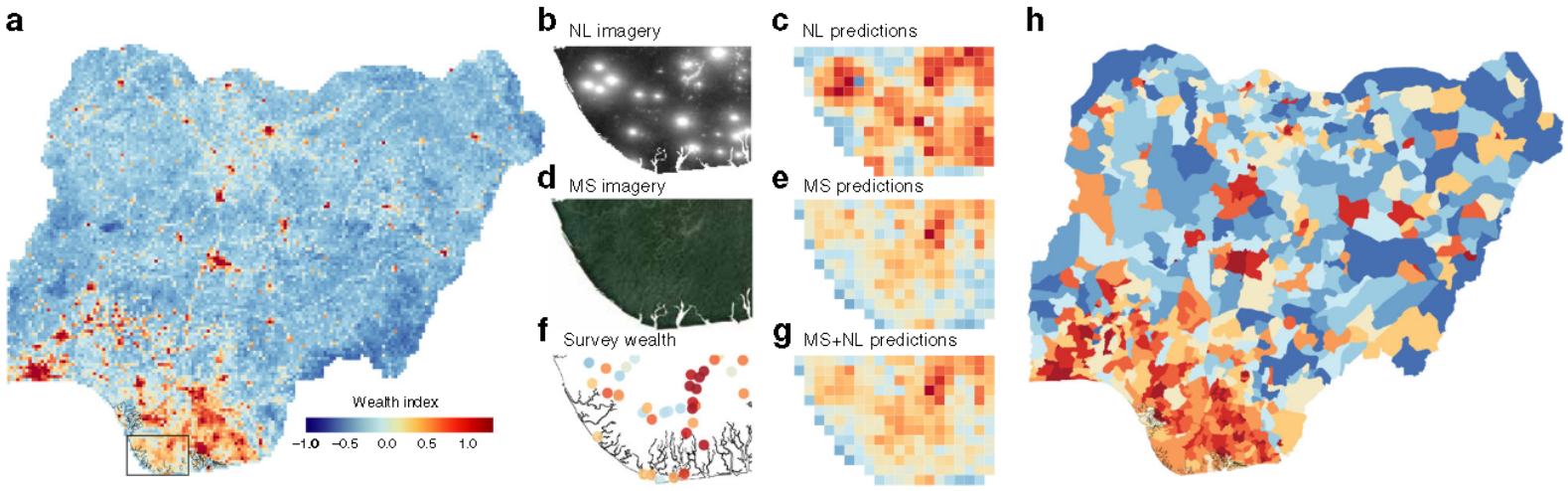
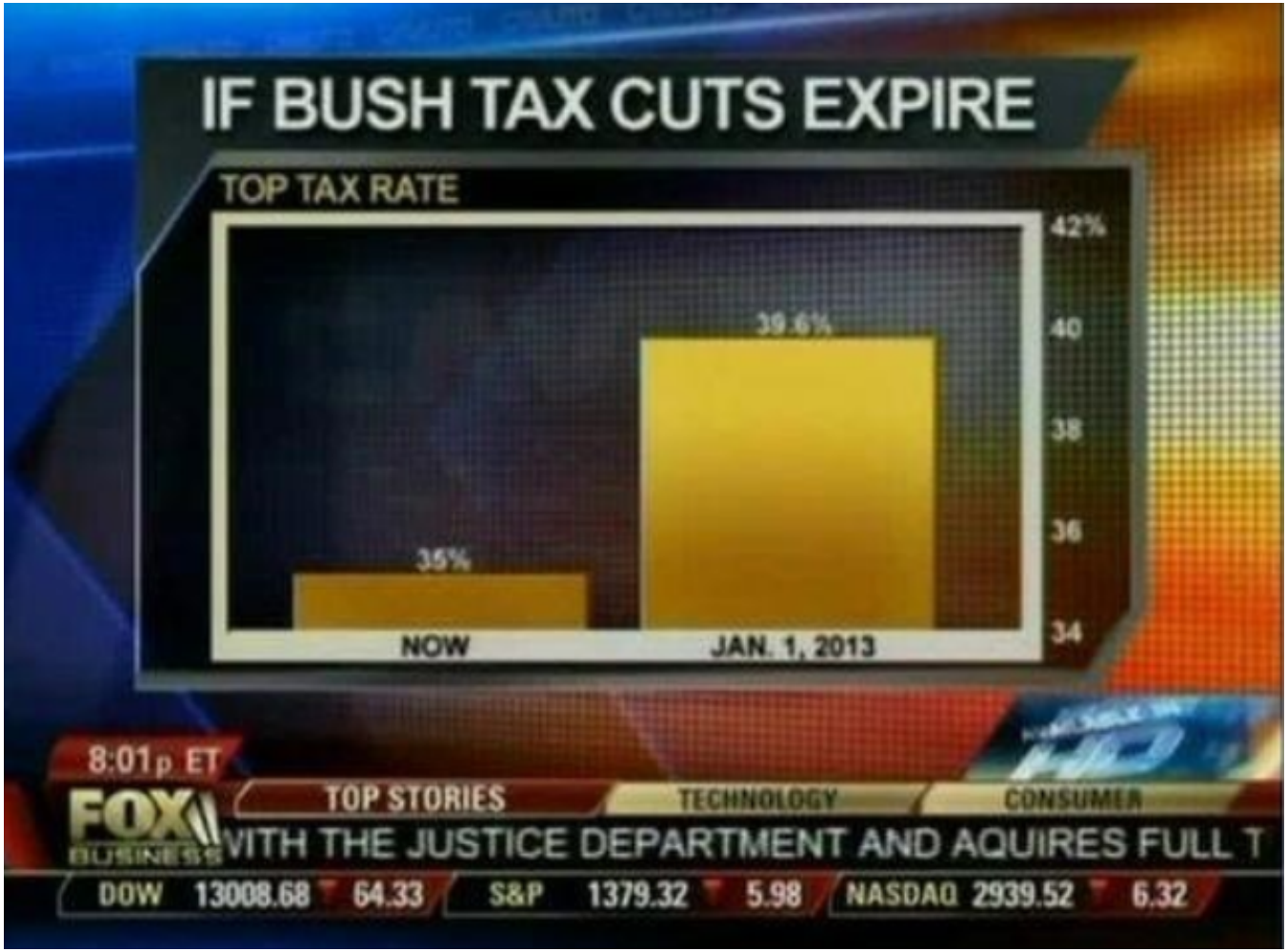


Fig. 6 Spatial extent of imagery allows wealth predictions at scale. a Satellite-based wealth estimates across Nigeria at pixel level. **b**, **d** Imagery inputs to model over region in Southern Nigeria depicted in box in **a**. **f** Ground truth input to model over the same region. **c**, **e**, **g** Model predictions with just nightlights (NL) as input, just multispectral (MS) imagery as input, and the concatenated NL and MS features as input. In this region, the model appears to rely more heavily on MS than NL inputs, ignoring light blooms from gas flares visible in **b**. **h** Deciles of satellite-based wealth index across Nigeria, population weighted using Global Human Settlement Layer population raster, and aggregated to Local Government Area level from the Database of Global Administrative Areas.

Yeh et al., 2020. Nature

Los datos pueden ser engañosos



Los datos pueden ser peligrosos (i)

› Lancet. 1998 Feb 28;351(9103):637-41. doi: 10.1016/s0140-6736(97)11096-0.

Ileal-lymphoid-nodular hyperplasia, non-specific colitis, and pervasive developmental disorder in children

A J Wakefield [1], S H Murch, A Anthony, J Linnell, D M Casson, M Malik, M Berelowitz, A P Dhillon, M A Thomson, P Harvey, A Valentine, S E Davies, J A Walker-Smith

Affiliations + expand

PMID: 9500320 DOI: 10.1016/s0140-6736(97)11096-0

Los datos pueden ser peligrosos (ii)

EARLY REPORT

Early report

Ileal-lymphoid-nodular hyperplasia, non-specific colitis, and pervasive developmental disorder in children

A J Wakefield, S H Murch, A Anthony, J Linnell, D M Casson, M Malik, M Berelowitz, A P Dhillon, M A Thomson, P Harvey, A Valentine, S E Davies, J A Walker-Smith

Summary

Background We investigated a consecutive series of children with chronic enterocolitis and regressive developmental disorder.

Methods 12 children (mean age 6 years [range 3–10], 11 boys) were referred to a paediatric gastroenterology unit with a history of normal development followed by loss of acquired skills, including language, together with diarrhoea and abdominal pain. Children underwent gastroenterological, neurological, and developmental assessment and review of developmental records. Ileocolonoscopy and biopsy sampling, magnetic-resonance imaging (MRI), electroencephalography (EEG), and lumbar puncture were done under sedation. Barium follow-through radiography was done where possible. Biochemical, haematological, and immunological profiles were examined.

Findings Onset of behavioural symptoms was associ... by the parents, with measles, mumps, and rub... vaccination in eight of the 12 children, with measl... infection in one child, and otitis media in as... All 1... children had intestinal abnormalities ... from lymphoid nodular hyperplasia to a... ulceration. Histology showed patchy chronic inf... tion in 11 children and reactive ileal lymph... perplasia in seven, but no granulomas. Be... oural disorders included autism (nine), disintegrative ... sis (one), and possible postviral or vaccinal enceph... (one). There were no focal neurological abn... alities and ... and EEG tests were normal. Abno... al laboratory results were significantly raised urinary ... thylmal... acid compared with age-matched control... (p=0·03), low ... haemoglobin in four children, ... low s... m IgA in ... children.

Interpretation We identi... associated gastrointestinal d... e and developmental regression in a group of previ... mal children, which was generally associated in time ... possible environmental triggers.

Lancet 1998; **351**: 637–41
See Commentary page

Inflammatory Bowel Disease Study Group, University Departments of Medicine and Histopathology (A J Wakefield FRCS, A Anthony MB, J Linnell PhD, A P Dhillon MRCPath, S E Davies MRCPath) **and the University Departments of Paediatric Gastroenterology** (S H Murch MB, D M Casson MRCP, M Malik MRCP, M A Thomson FRCP, J A Walker-Smith FRCP,), **Child and Adolescent Psychiatry** (M Berelowitz FRCPsych), **Neurology** (P Harvey FRCP), **and Radiology** (A Valentine FRCR), **Royal Free Hospital and School of Medicine, London NW3 2QG, UK**

Correspondence to: Dr A J Wakefield

Introduction

We saw several children who, after a p... of apparent normality, lost acquired skills, includ... com... nication. They all had gastrointestinal ... mptoms, ... luding abdominal pain, diarrhoea, and ... ating and, ... some cases, food intolerance. We d... scribe ... clinical f... dings, and gastrointestinal feature... of these chi... en.

Patients and meth... ds

12 children, cons... cutive ... ed to ... department of paediatric gastr... terology ... a hi... y of a pervasive developmental ... der with los... ... ed skills and intestinal symptoms ... arrh... abdominal ... ain, bloating and food intolerance), were inv... gated. All children were admitted to the ward for ... week, accomp... ed by their parents.

Clinical investigations

... took histor... including details of immunisations and ... osure to infec... diseases, and assessed the children. In 11 ca... the histor... was obtained by the senior clinician (JW-S). Neur... ... d psychiatric assessments were done by consultant staff (PH, MB) with HMS-4 criteria.[5] Developmental ... included a review of prospective developmental records from parents, health visitors, and general practitioners. Four children did not undergo psychiatric assessment in hospital; all had been assessed professionally elsewhere, so these assessments were used as the basis for their behavioural diagnosis.

After bowel preparation, ileocolonoscopy was performed by SHM or MAT under sedation with midazolam and pethidine. Paired frozen and formalin-fixed mucosal biopsy samples were taken from the terminal ileum; ascending, transverse, descending, and sigmoid colons, and from the rectum. The procedure was recorded by video or still images, and were compared with images of the previous seven consecutive paediatric colonoscopies (four normal colonoscopies and three on children with ulcerative colitis), in which the physician reported normal appearances in the terminal ileum. Barium follow-through radiography was possible in some cases.

Also under sedation, cerebral magnetic-resonance imaging (MRI), electroencephalography (EEG) including visual, brain stem auditory, and sensory evoked potentials (where compliance made these possible), and lumbar puncture were done.

Laboratory investigations

Thyroid function, serum long-chain fatty acids, and cerebrospinal-fluid lactate were measured to exclude known causes of childhood neurodegenerative disease. Urinary methylmalonic acid was measured in random urine samples from eight of the 12 children and 14 age-matched and sex-matched normal controls, by a modification of a technique described previously.[6] Chromatograms were scanned digitally on computer, to analyse the methylmalonic-acid zones from cases and controls. Urinary methylmalonic-acid concentrations in patients and controls were compared by a two-sample *t* test. Urinary creatinine was estimated by routine spectrophotometric assay.

Children were screened for antiendomysial antibodies and boys were screened for fragile-X if this had not been done

RETRACTED

Los datos pueden ser peligrosos (iii)

Category:Anti-vaccination organizations

From Wikipedia, the free encyclopedia

Pages in category "Anti-vaccination organizations"

The following 26 pages are in this category, out of 26 total. This list may not reflect recent changes (learn more).

A

- *Anti-Vaccination League of America*
- Anti-Vaccination Society of America
- Association of American Physicians and Surgeons
- Australian Vaccination-risks Network

C

- Children's Health Defense
- Children's Medical Safety Research Institute
- Church of Conscious Living

F

- Freedom Angels Foundation

H

- Homeopathy Plus!
- Humanitarian League

I

- Informed Consent Action Network
- Informed Medical Options Party

L

- Learn The Risk

N

- National Anti-Vaccination League
- National League for Liberty in Vaccination
- National Vaccine Information Center
- New Jersey Coalition for Vaccination Choice
- New Zealand Public Party

P

- Palmetto Family Council
- Pioneer Club (women's club)

S

- Stop Mandatory Vaccination

T

- Talk About Curing Autism
- Texans for Vaccine Choice

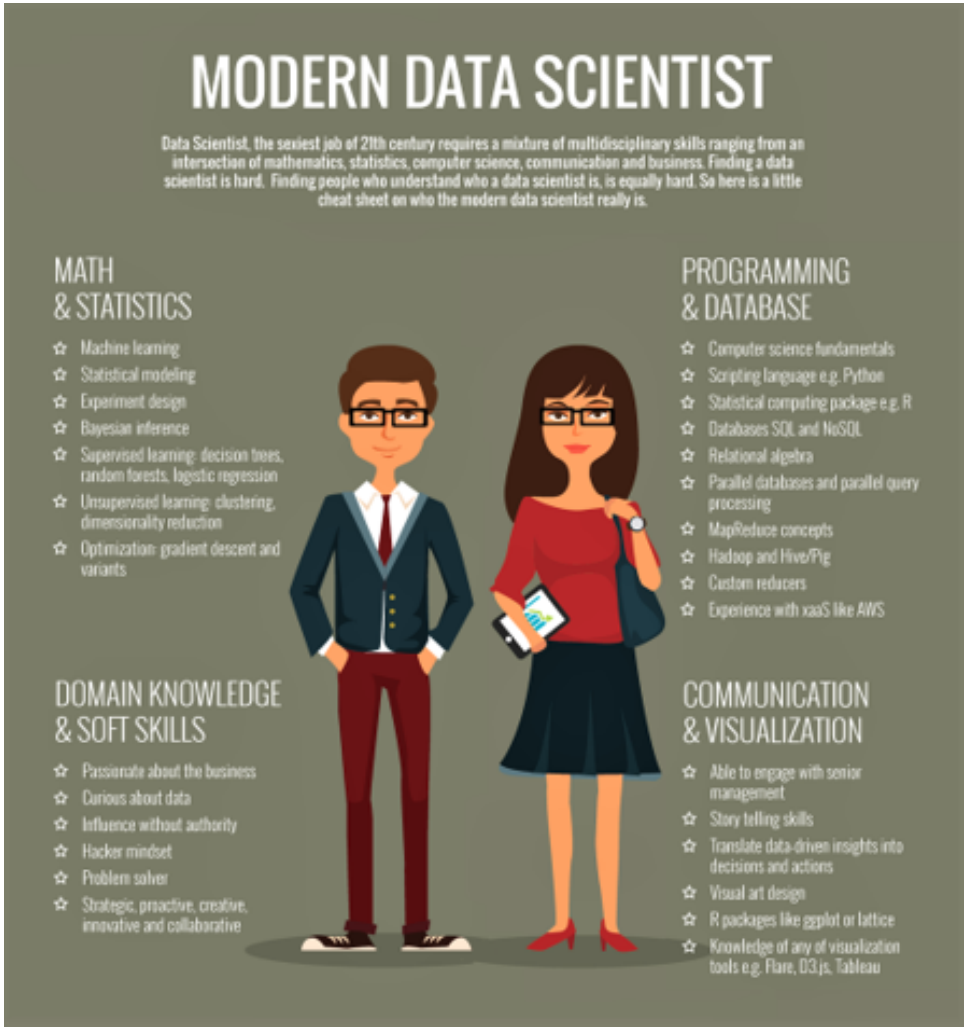
V

- Vaccine Choice Canada

W

- Warnings About Vaccination Expectations NZ
- World Chiropractic Alliance

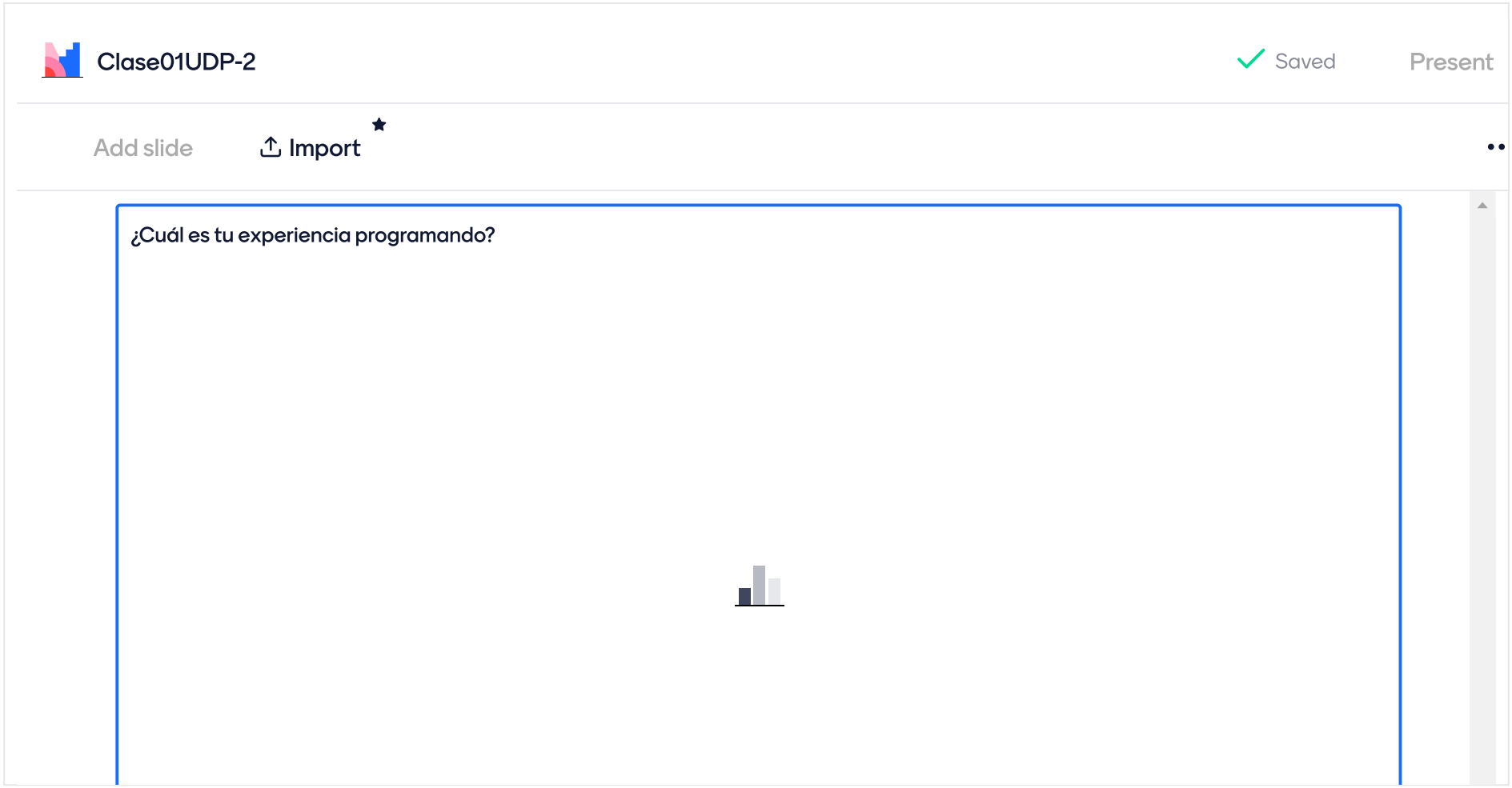
¿Qué hacer entonces?



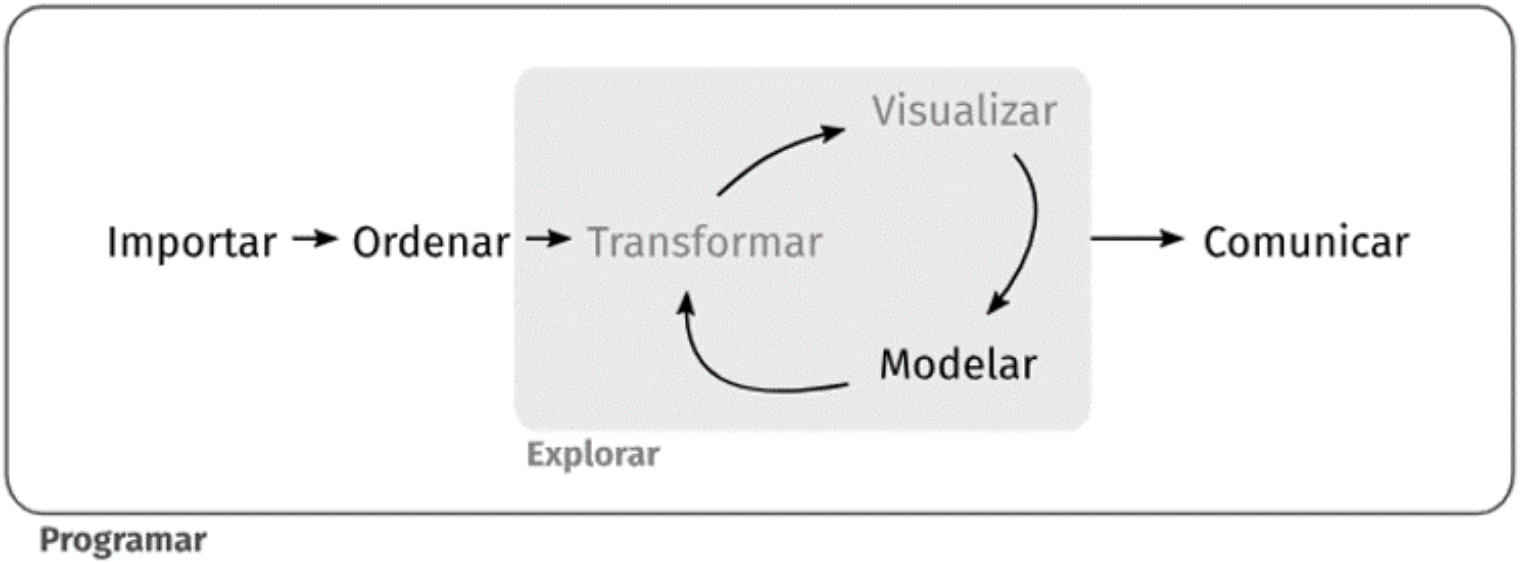
Ustedes ya saben de estas cosas

- Métodos econométricos
- Gestión, representación, y análisis de datos
- Evaluación de impacto
- Otros…

Una pregunta más

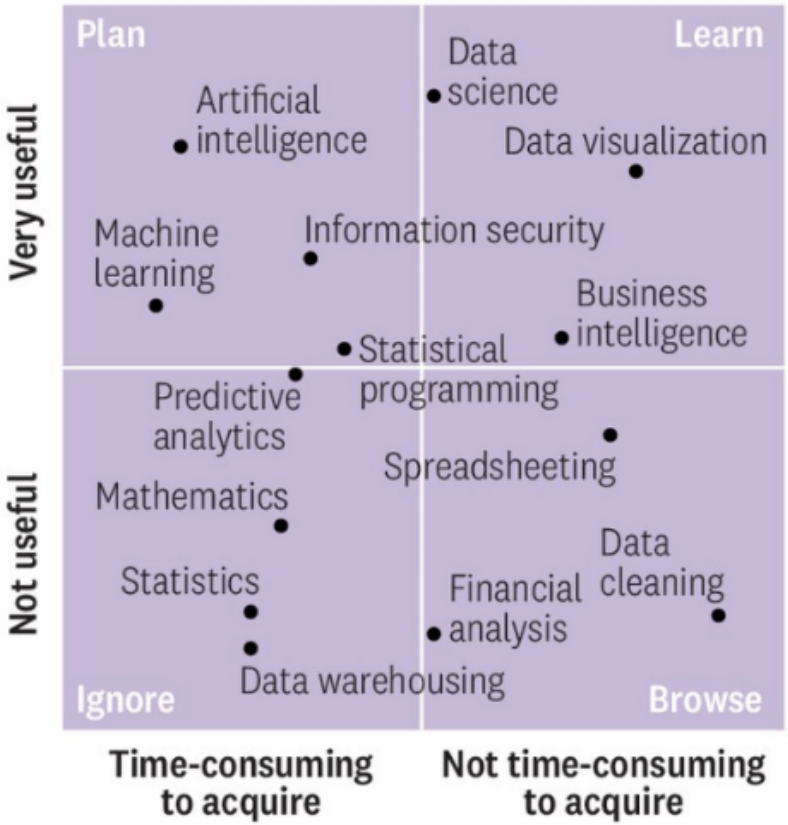


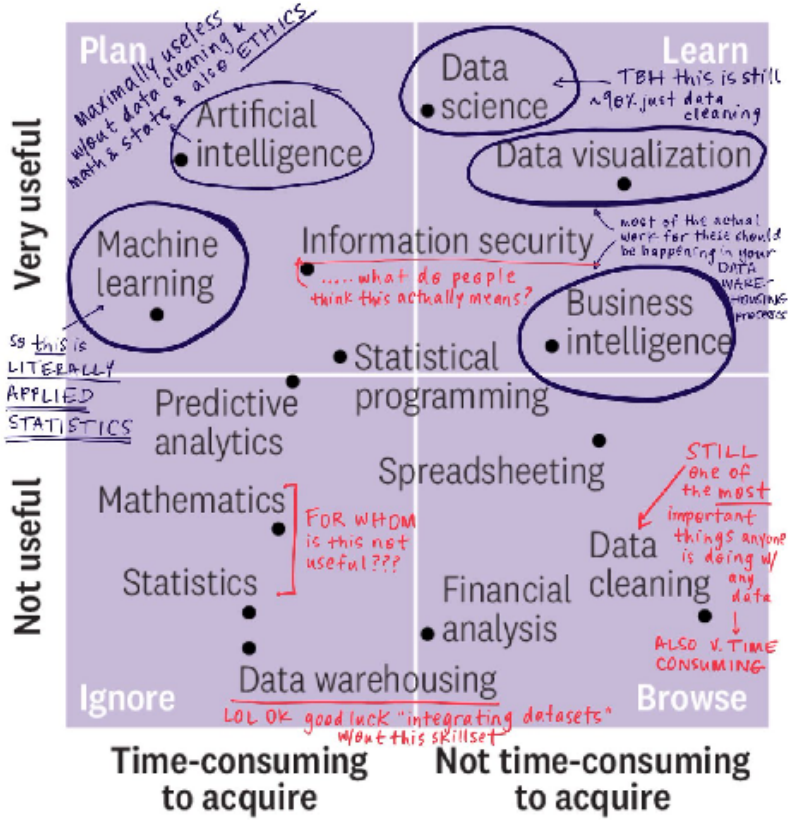
Foco de este curso



Productores de análisis

Pero también mejores consumidores…





https://hbr.org/2018/10/prioritize-which-data-skills-your-company-needs-with-this-2x2-matrix

En específico

- Visualización de datos
- Manejo de datos
- Regresión/Clasificación
- Predicción vs Inferencia
- Aprendizaje no supervisado
- Automatización (¿?)
- **PROGRAMACIÓN**

¿Por qué programar?

Reproducibilidad (i)

Objetivo: realizar un reporte basado en datos de Datos Públicos

Reproducibilidad (i)

Objetivo: realizar un reporte basado en datos de Datos Públicos

SIN PROGRAMACIÓN

1. Ingresar a la web y descargar datos
2. Limpiar datos en *MS Excel*
3. Analizar datos en *Stata*
4. Escribir documento en *MS Word*

Reproducibilidad (i)

Objetivo: realizar un reporte basado en datos de Datos Públicos

SIN PROGRAMACIÓN

1. Ingresar a la web y descargar datos
2. Limpiar datos en *MS Excel*
3. Analizar datos en *Stata*
4. Escribir documento en *MS Word*

CON PROGRAMACIÓN

1. Crear una carpeta específica para el proyecto/tarea
   - `datos`
   - `gráficos`
   - `resultados`
2. Descargar datos desde `R`
3. Limpiar datos en `R`
4. Analizar datos en `R`
5. Escribir documento en `R Markdown`

Reproducibilidad (ii)

Seis meses después quieres repetir la tarea (o un/a coleg@)

Reproducibilidad (ii)

Seis meses después quieres repetir la tarea (o un/a coleg@)

SIN PROGRAMACIÓN

1. Recordar que se hizo
2. Ingresar a la web y descargar datos
3. Limpiar datos en MS Excel y esperar no haberse olvidado de nada
4. Analizar datos en Stata
5. Escribir documento en MS Word

Reproducibilidad (ii)

Seis meses después quieres repetir la tarea (o un/a coleg@)

SIN PROGRAMACIÓN

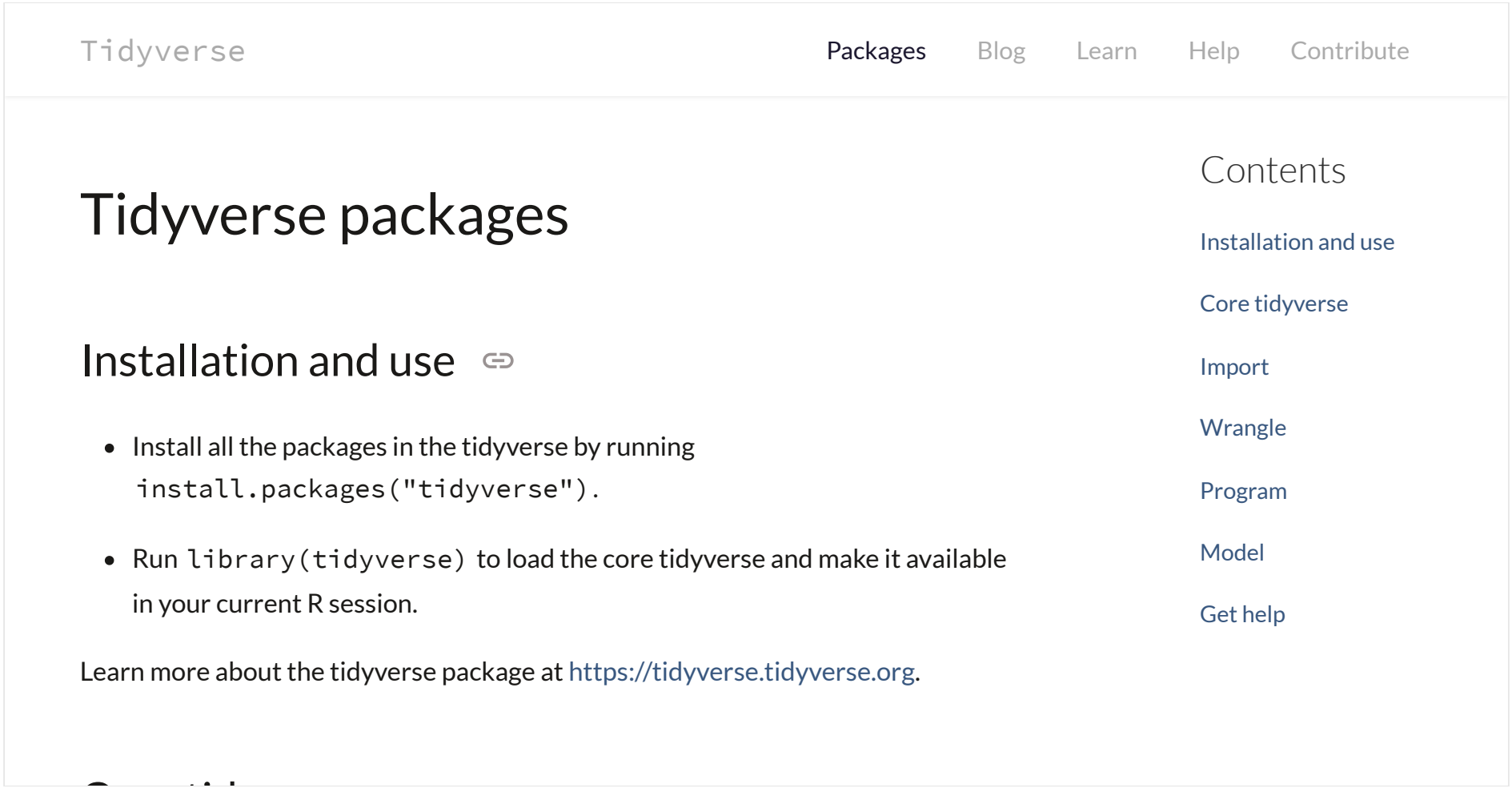
1. Recordar que se hizo
2. Ingresar a la web y descargar datos
3. Limpiar datos en MS Excel y esperar no haberse olvidado de nada
4. Analizar datos en Stata
5. Escribir documento en MS Word

CON PROGRAMACIÓN

1. Re-correr el código

R y Tidyverse

- Hay muchos lenguajes de programación
- Hay muchas formas de escribir código en R



Tidyverse Packages Blog Learn Help Contribute

Tidyverse packages

Contents

Installation and use

Core tidyverse

Import

Wrangle

Program

Model

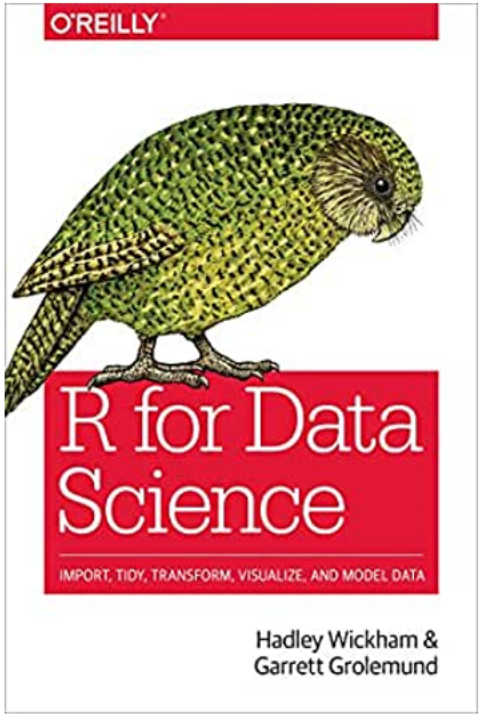
Get help

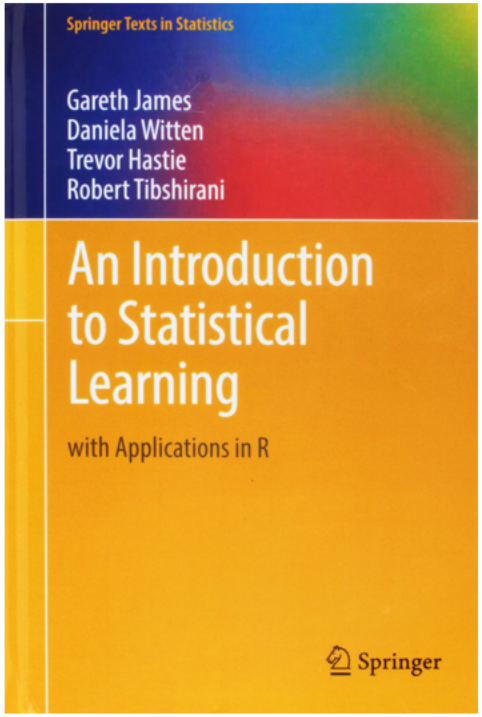
Installation and use

- Install all the packages in the tidyverse by running `install.packages("tidyverse")`.
- Run `library(tidyverse)` to load the core tidyverse and make it available in your current R session.

Learn more about the tidyverse package at https://tidyverse.tidyverse.org.

Fuentes valiosas de información (i)





- https://r4ds.had.co.nz/ | https://es.r4ds.hadley.nz/
- http://faculty.marshall.usc.edu/gareth-james/ISL/

Fuentes valiosas de información (ii)

Google

stackoverflow

R-bloggers

R news and tutorials contributed by hundreds of R bloggers

#rstats

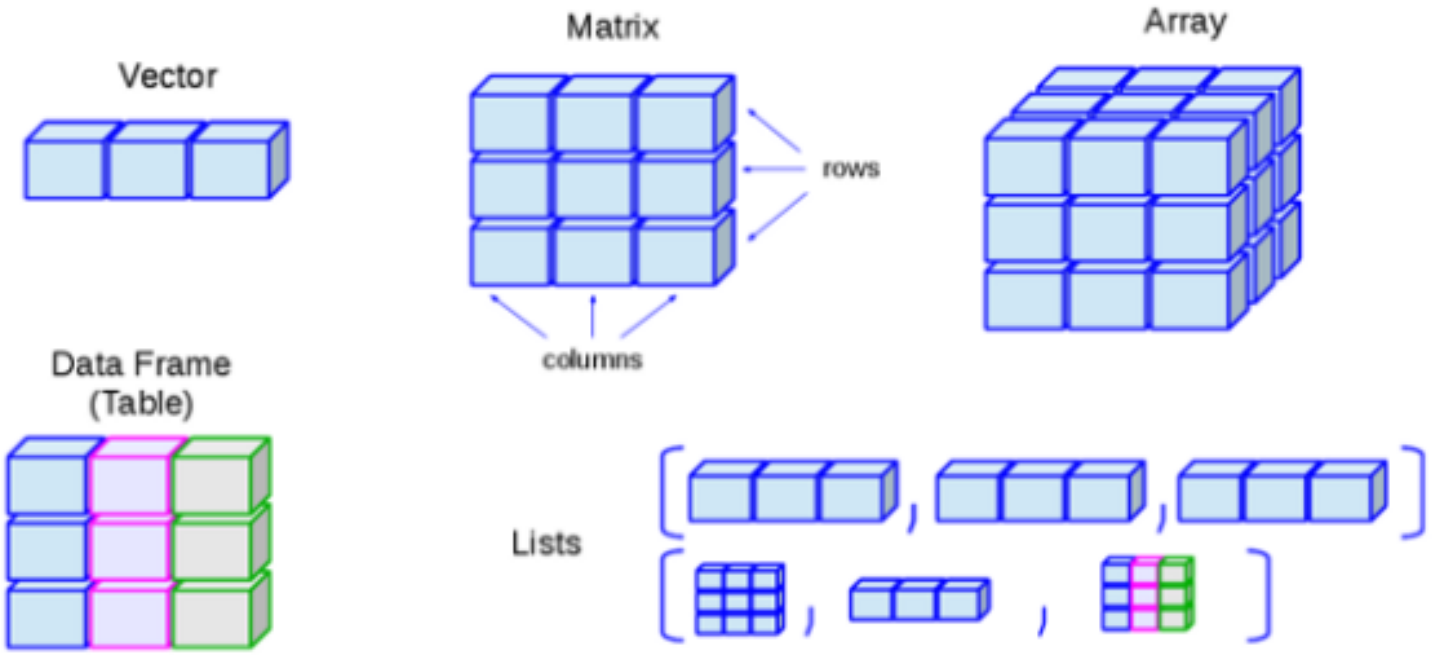
Suficiente bla bla…

Demo - Ejercicio

- script: `EjercicioRepasoR.R`

Cosas a tener en cuenta

Tipos de datos



Pipe

`%>%` nos permite definir nuestras acciones como una secuencia

- Código “anidado”

```
estacionar(manejar(buscar(llaves), hacia = "trabajo"))
```

- Código como secuencia

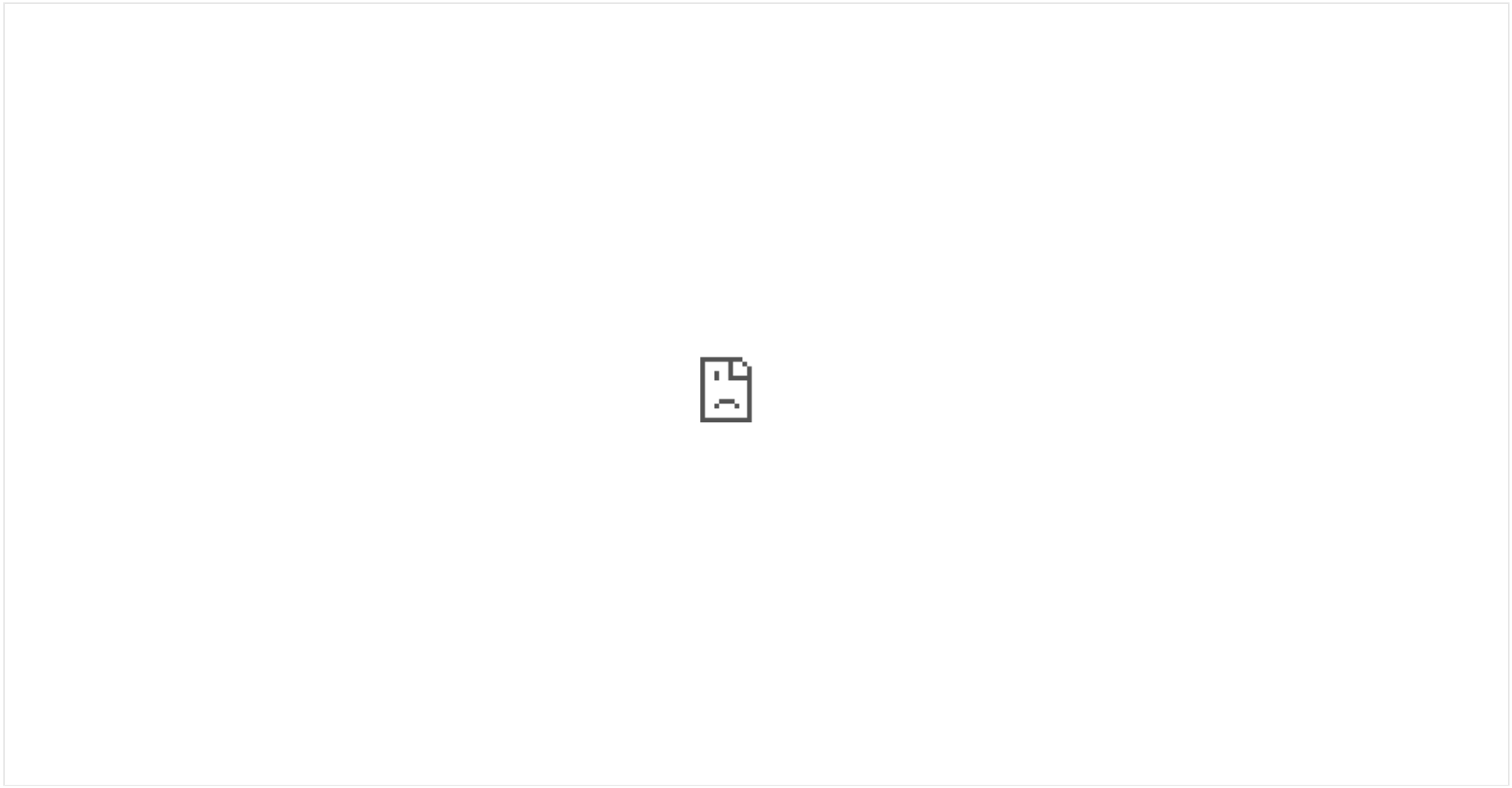
```
llaves %>%
  buscar() %>%
  manejar(hacia = "trabajo") %>%
  estacionar()
```

Simpre hay más de una forma de hacer lo mismo

```
datos_mundo[datos_mundo$anio == 2007,]
subset(datos_mundo, anio == 2007)
filter(datos_mundo, anio == 2007)
datos_mundo %>% filter(anio == 2007)
```

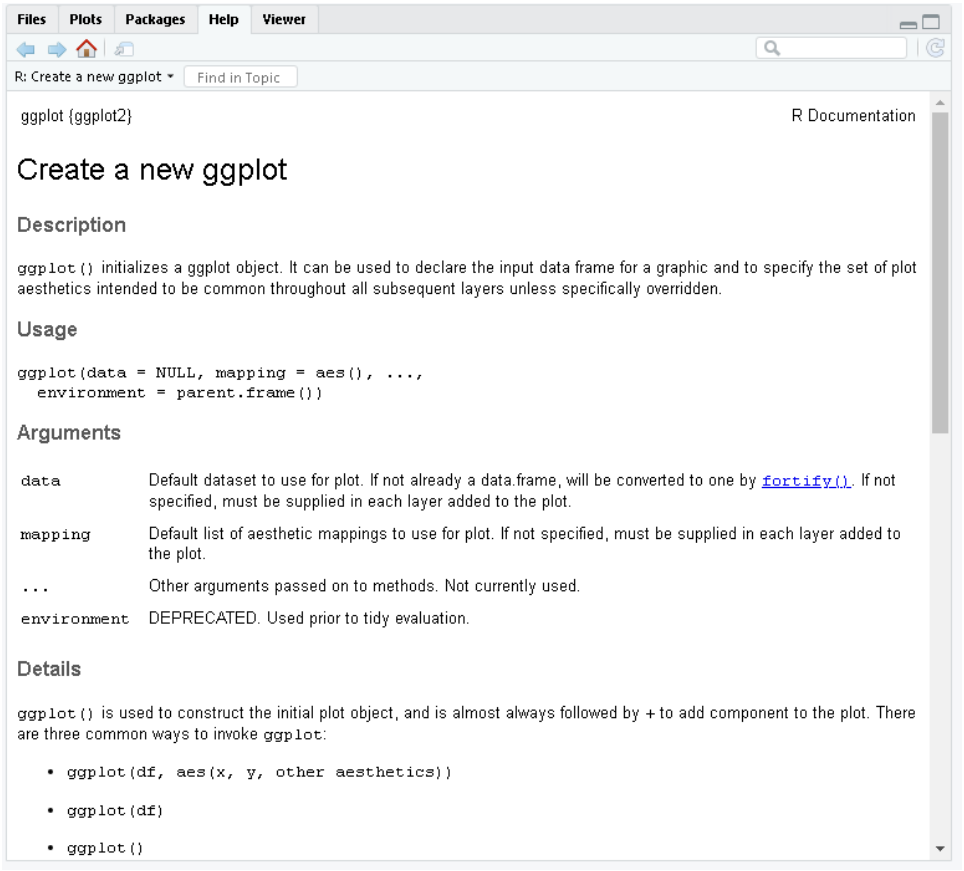
```
## # A tibble: 142 x 6
##    pais        continente  anio ExpVida       pob gdpPercap
##    <chr>       <chr>      <dbl>   <dbl>     <dbl>     <dbl>
##  1 Afghanistan Asia        2007    43.8  31889923      975.
##  2 Albania     Europe      2007    76.4   3600523     5937.
##  3 Algeria     Africa      2007    72.3  33333216     6223.
##  4 Angola      Africa      2007    42.7  12420476     4797.
##  5 Argentina   Americas    2007    75.3  40301927    12779.
##  6 Australia   Oceania     2007    81.2  20434176    34435.
##  7 Austria     Europe      2007    79.8   8199783    36126.
##  8 Bahrain     Asia        2007    75.6    708573    29796.
##  9 Bangladesh  Asia        2007    64.1 150448339     1391.
## 10 Belgium     Europe      2007    79.4  10392226    33693.
## # ... with 132 more rows
```

… incluyendo los gráficos



Además de libros y google…

Siempre consulten `?nombrefunción`. Ej: `?ggplot`



¿Qué se viene?

- Semana 2: Visualización de datos
- Semana 3: Manejo de datos
    - Entregables: Idea de trabajo y Tarea 1