

Procesamiento de datos en R y estadística para Ciencias Sociales

Clase 1. Presentación e introducción a R

¿Qué son las Ciencias Sociales Computacionales?

- No parece haber una definición clara y consensuada
- Muchas definiciones por extensión
- Tratemos de definirlo mediante un caso de aplicación...

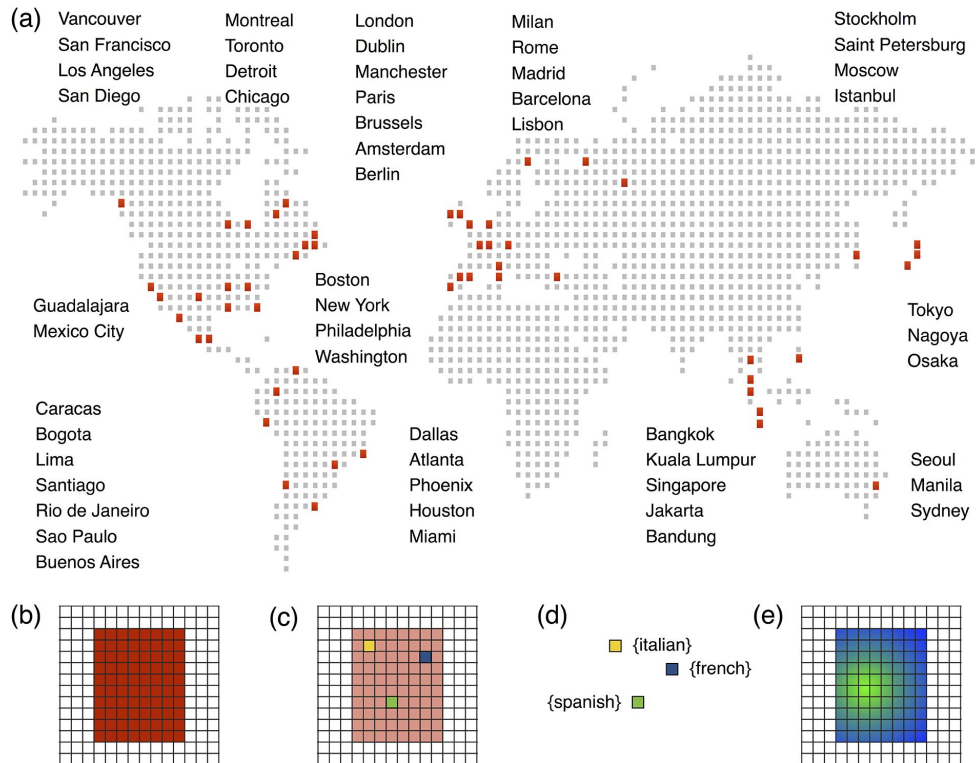
Antes de empezar...



Immigrant community integration in world cities

- Medir la segregación/ integración de comunidades migrantes
- 52 ciudades
- Datos de Twitter
- Búsqueda de lugares de residencia habituales
- Detección de lenguaje
- Cálculo de índices de segregación

[\[Lamanna, Lenormand, et al 2016\]](#)



A global map of travel time to cities to assess inequalities in accessibility in 2015

- Acceso a grandes centros urbanos como indicador de desigualdad
- Uso de fuentes abiertas (OSRM, Google, imágenes satelitales)
- Mapa de alta resolución 1kmx1km

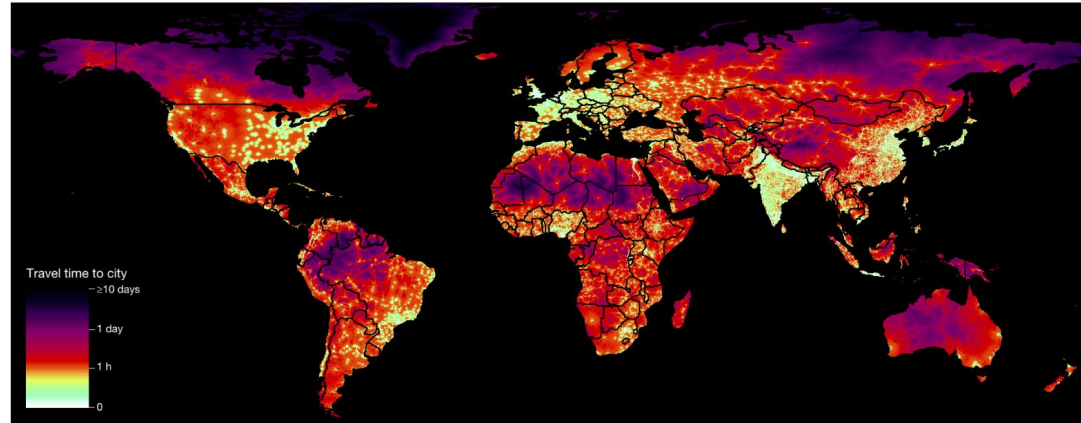


Figure 1 | Global map of travel time to cities for 2015. The accessibility map has a spatial resolution of approximately 1×1 km, spans 60° south to 85° north latitude, and enumerates travel time to the city with the shortest associated journey.

¿Qué son las Ciencias Sociales Computacionales?

- **Problemas/preguntas** de investigación más o menos clásicas

¿Qué son las Ciencias Sociales Computacionales?

- Problemas/preguntas de investigación más o menos clásicas
- Uso intensivo de algoritmos, cálculos y métodos de predicción
 - Métodos cuantitativos/estadísticos clásicos
 - Machine Learning

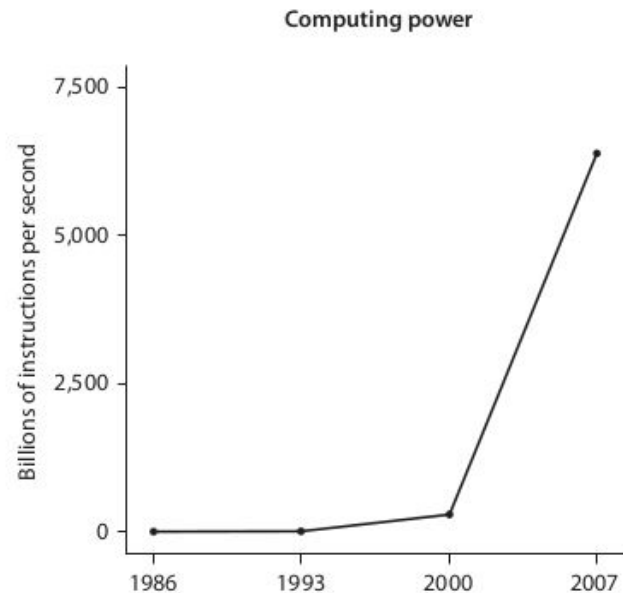
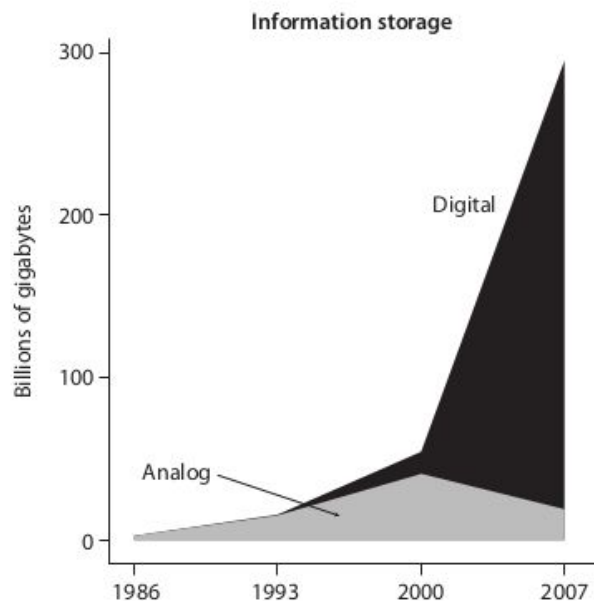
¿Qué son las Ciencias Sociales Computacionales?

- Problemas/preguntas de investigación más o menos clásicas
- Uso intensivo de algoritmos, cálculos y métodos de predicción
 - Métodos cuantitativos/estadísticos clásicos
 - Machine Learning
- Combinación de datos
 - diferentes orígenes,
 - diferentes procesos de producción
 - diferentes grados de estructuración

¿Qué son las Ciencias Sociales Computacionales?

- Problemas/preguntas de investigación más o menos clásicas
 - Uso intensivo de algoritmos, cálculos y métodos de predicción
 - Métodos cuantitativos/estadísticos clásicos
 - Machine Learning
 - Combinación de datos
 - diferentes orígenes,
 - diferentes procesos de producción
 - diferentes grados de estructuración
- } ● Rudimentos de programación / escritura de código (R, Python, JS, lo que sea necesario)

Los datos, los algoritmos y la ciencia social



Los datos, los algoritmos y la ciencia social

CHRIS ANDERSON SCIENCE 06.23.08 12:00 PM

THE END OF THEORY: THE DATA DELUGE MAKES THE SCIENTIFIC METHOD OBSOLETE



“Olvídense de la teoría del comportamiento humano, desde la lingüística hasta la sociología. Olvídense de las taxonomías, de la ontología y de la psicología. ¿Quién sabe por qué la gente hace lo que hace?. El punto es que lo hacen y que podemos trackear y medir eso que hacen con una precisión sin precedentes. Con suficientes datos, los números hablan por sí mismos.”

El problema de los datos

MAS_500 Aglomerados segun tamaño	AGLOMERADO Codigo de Aglomerado	PONDERA Ponderacion	CH03 Relacion de parentesco	CH04 Sexo	CH05 Fecha de nacimiento (dia, mes y año)
N	8	108	2	2	03/06/1990
N	8	108	3	2	29/12/2005
N	8	108	3	1	26/01/2018
N	8	108	1	2	30/03/1978
N	8	108	3	2	20/09/2009
N	8	141	1	1	26/04/1967
N	8	221	1	1	15/03/1955
N	8	221	2	2	25/04/1956
N	8	221	3	2	10/06/1994
N	8	221	1	1	22/07/1944
N	8	221	3	1	23/08/1985
N	8	309	1	1	14/06/1976
N	8	309	2	2	17/06/1978
N	8	309	3	2	20/07/1997
N	8	309	3	1	19/10/2001
N	8	309	1	2	02/01/1967
N	8	309	3	2	29/06/1982
N	8	88	1	1	15/08/1974

El problema de los datos

ENVÍA TU COMENTARIO Ver legales ▾

Para poder comentar tenés que ingresar con tu usuario de LA NACION.

860 comentarios

INGRESAR

10 personas siguiendo



Esta nota se encuentra cerrada a comentarios.

Más nuevos Más viejos



crisel11
Larreta, imrepresentable.

12:55 12/03/2020

Reportar Compartir

Me gusta



aleman1943
"en terapia intensiva "por desaturación"; Desaturación" de que !!!! habra sido dehidratación o descubrieron un nuevo "ente" fisiológico en la RA ??

06:42 12/03/2020

Reportar Compartir

Me gusta



dani20010
[@aleman1943](#) se mide la saturación de oxígeno en sangre. Desaturación es la falta de oxigenación, posiblemente por problemas respiratorios.

18:41 12/03/2020

Reportar Compartir



Josef_Radetzky
Tranquilos. Ahora el gobierno manda el Proyecto de Ley de Aborto al Congreso, y se arregla todo.

Reportar Compartir



transilium
Cuando todo ésto pase en unas semanas, se tendrán que hacer responsables los que propagaron el pánico inútilmente causando estragos económicos y sanitarios en todo el planeta.hoy en día estamos mucho mas preparados para el coronavirus que en otras enfermedades del pasado pero peor informados pese a tener datos en directo minuto a minuto.

05

Reportar Compartir

MÁS COMENTADAS

- 1 El Gobierno sacará por DNU el congelamiento de alquileres y créditos hipotecarios
- 2 Coronavirus: Alfredo Casero estalló contra Marcelo Tinelli y Diego Brancatelli
- 3 "No nos abandonen, no somos unos chetos", la frustración de los argentinos varados
- 4 Coronavirus: con 4492 nuevos casos, volvió a dispararse el contagio en Italia
- 5 Es hora de pensar si hay una alternativa mejor que cerrar todo

```
<<SimpleCorpus>>
Metadata: corpus specific: 1, document level (indexed): 0
Content: documents: 3
```

[1] a bailar a bailar | que la orquesta se va | sobre el fino garabato | de un tango nervioso y lerdo | se ira borrando el recuerdo | a bailar a bailar | que la orquesta se va | el ultimo tango perfuma la noche | un tango dulce que dice adios | la frase callada se asoma a los labios | y canta el tango la despedida! | vamos! a bailar! | tal vez no vuelvas a verla nunca | y el ultimo tango perfuma la noche | y te es el tango que dice el adios | a bailar a bailar | que la orquesta se va! | quedara el salon vacio | con un monton de esperanzas | iran camino al olvido | a bailar a bailar | que la orquesta se va!

[2] este tango nacio para bailarse | y asi hamacarse muy suavemente | oigan ustedes este compas... | es muy sencillo bailar el tango | un ble paso despues descanso | la media vuelta la vuelta entera | y siempre junto a la compañera | este tango nacio para bailarse | no hay e quedarse mirandolo

[3] nacio en la calle quito | entre boedo y colombres | barrio de tauras de hombres | de timbas y de garitos | mi recuerdo es muy estri... | de prosenio un corralon | modesto fue su blason | y la dulce purretita | se lavaba la carita | en el viejo pileton | amante del var... e | soñaba con ser artista | comenzo como corista | hasta llegar a vedette | piernas tipo mistinguette | cintura bien contorneada | ana... ia envidiada | y un rostro angelical | para que plumas y percal | lucieran como hermanadas | siempre causo sensacion | en cine radio y t... tro; | se volco al dos por cuatro | con sentida emocion | triunfo en television | y nadie podra dudar | fue figura consular | en todos... escenarios | recogio aplausos a diario | se llamaba beba bidart

Tweet fijado

Conflicts+Violence @JohanGaltung · 13 sept. 2018

A note to students on [#Militarism](#) & [#MilitarySecrets](#):
You have a SERIOUS problem. You need truthful data on military capacity & deployment i.e. in your [#conflictanalysis](#). Just remember this detail: Historically, they MOSTLY lie & distort facts for so called "strategic reasons".

Military?
#SolveTheUnderlyingConflict

"With the exception of defensive defense strategies, evolving military measures to solve socio-economic contradictions is the equivalent of healing anemia with bloodletting."
-Johan Galtung

3 37 51

Mostrar este hilo

Conflicts+Violence @JohanGaltung · 6h

Dear friends, there will come a point where I will no longer teach nor be available to consult. Here's a list of my closest analysts:
Antonio Rosa (Brasil)
Fernando Montiel (Mexico)
Rais Boneza (DRC)
Erika Degortes (Italy)
Karoline Weber (Germany)
Naakow Grant-Hayford (Ghana)
JG

15 24 97

Conflicts+Violence @JohanGaltung · 6h

Interesting. Ingredients for a nuclear winter. Readily available on Twitter.
[twitter.com/nhmck/status/1...](#)

No puedes ver este Tweet debido a que el titular de esta cuenta limita



factor-data
IDAES_UNSAM

Presentación Track

Trayecto de Métodos Cuantitativos y Ciencias Sociales Computacionales

Procesamiento de datos con R para ciencias sociales

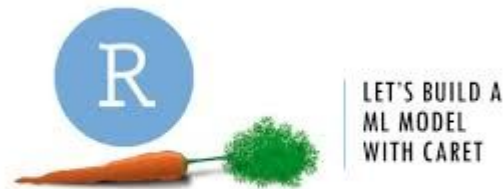
- Programación estadística en R.
- Limpieza y procesamiento de datos
- Estadística descriptiva e inferencial
- Fundamentos de visualización de datos



Trayecto de Métodos Cuantitativos y Ciencias Sociales Computacionales

Métodos de análisis cuantitativos multivariados

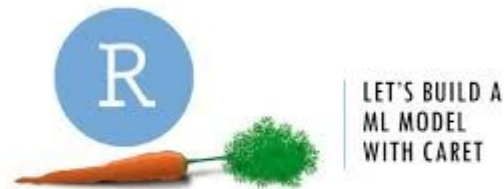
- Regresión lineal y logística
- Introducción a las técnicas de clustering
- Metodología del aprendizaje automático (machine learning).
- Introducción a tidymodels.



Trayecto de Métodos Cuantitativos y Ciencias Sociales Computacionales

Machine Learning aplicado a las Ciencias Sociales

- Clasificadores basados en árboles: CART.
- Algoritmos de Ensamble: bagging, random forest, boosting, Gradient Boosting.
- Introducción a las redes neuronales
- Machine Learning Interpretable: Herramientas para la interpretación de modelos de caja negra



Trayecto de Métodos Cuantitativos y Ciencias Sociales Computacionales

Laboratorio de datos: web scraping y procesamiento de lenguaje natural

- Webscraping y APIs
- Preprocesamiento de texto: tokenización, normalización (lemas y stemming), stopwords.
- Vectorización de texto:
- Modelado de tópicos
- Word embeddings



Programa M1, cuestiones administrativas, medios de comunicación

Dinámica de clases

- Bloques de 50-55 minutos
- Cortes de 15 minutos
- Actividades independientes

Herramientas



Medios de comunicación

- Clases presenciales o virtuales (según la situación epidemiológica imperante en cada momento)

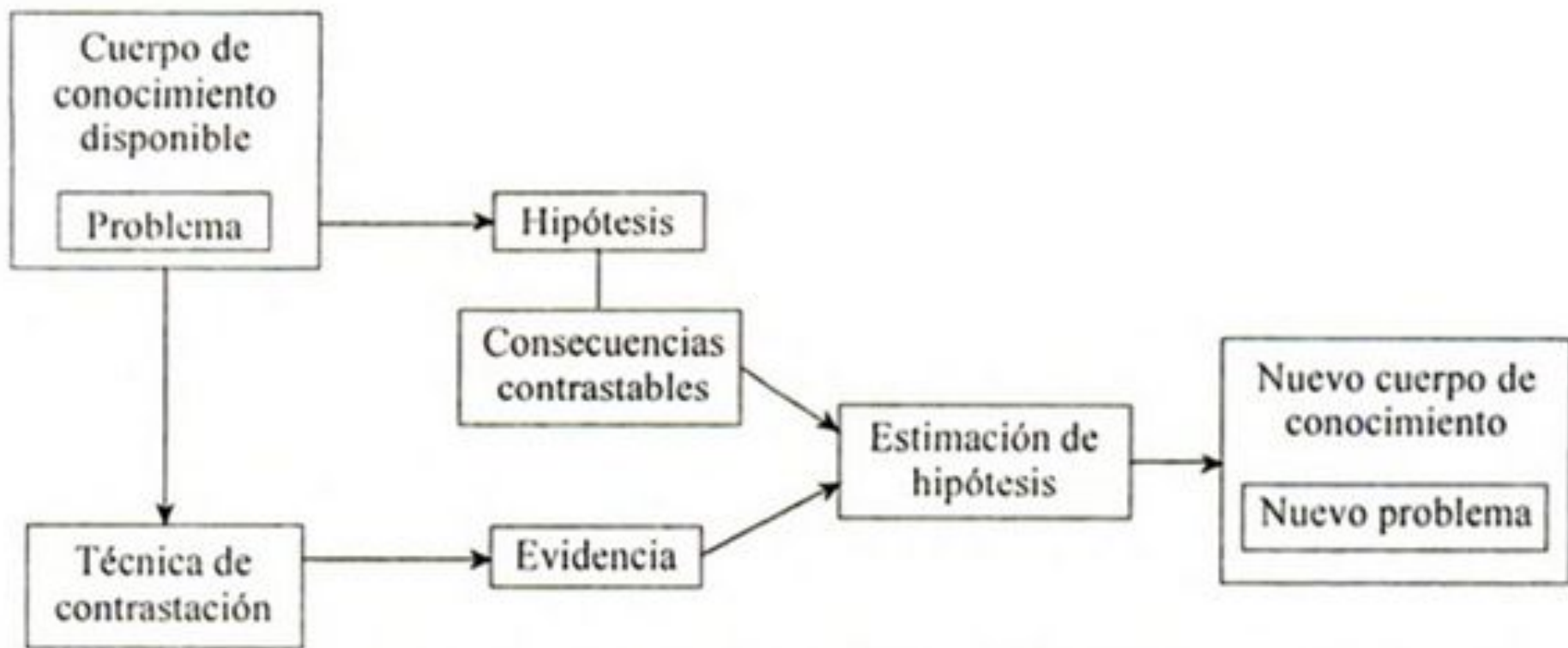


Google Classroom

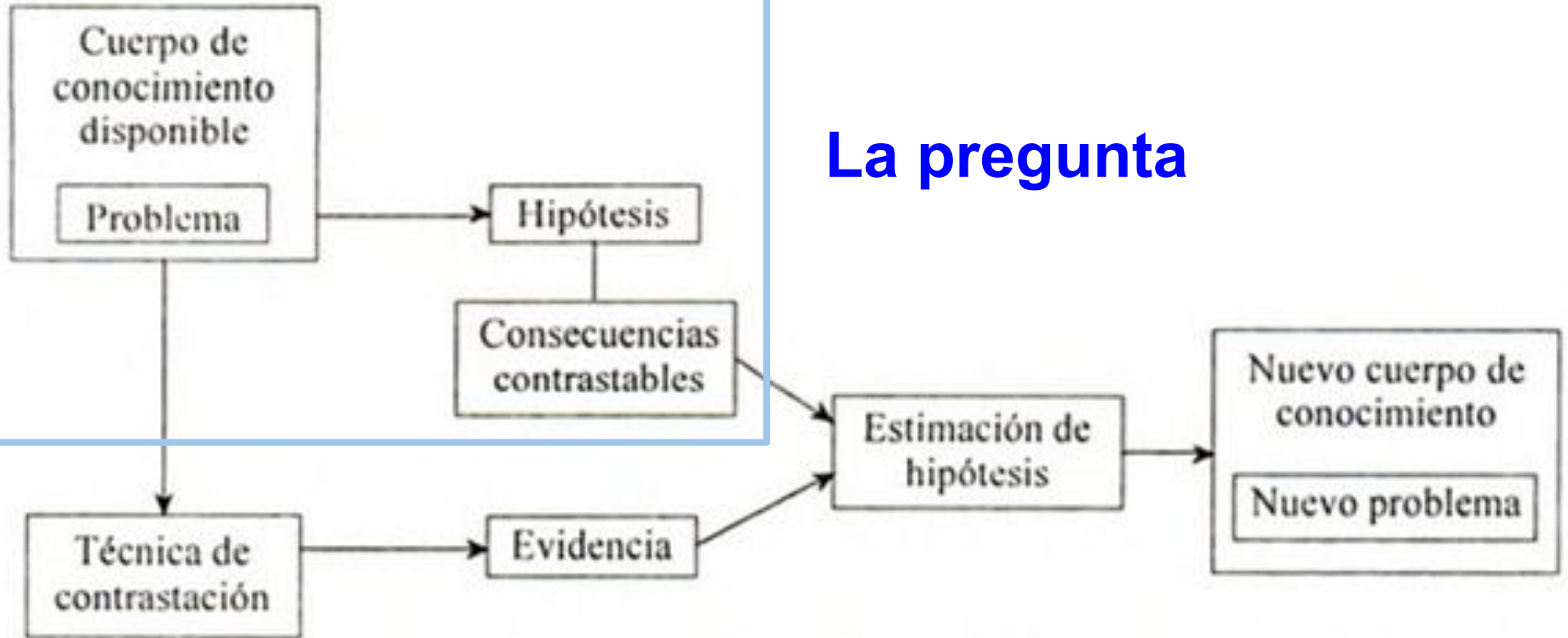
Herramientas

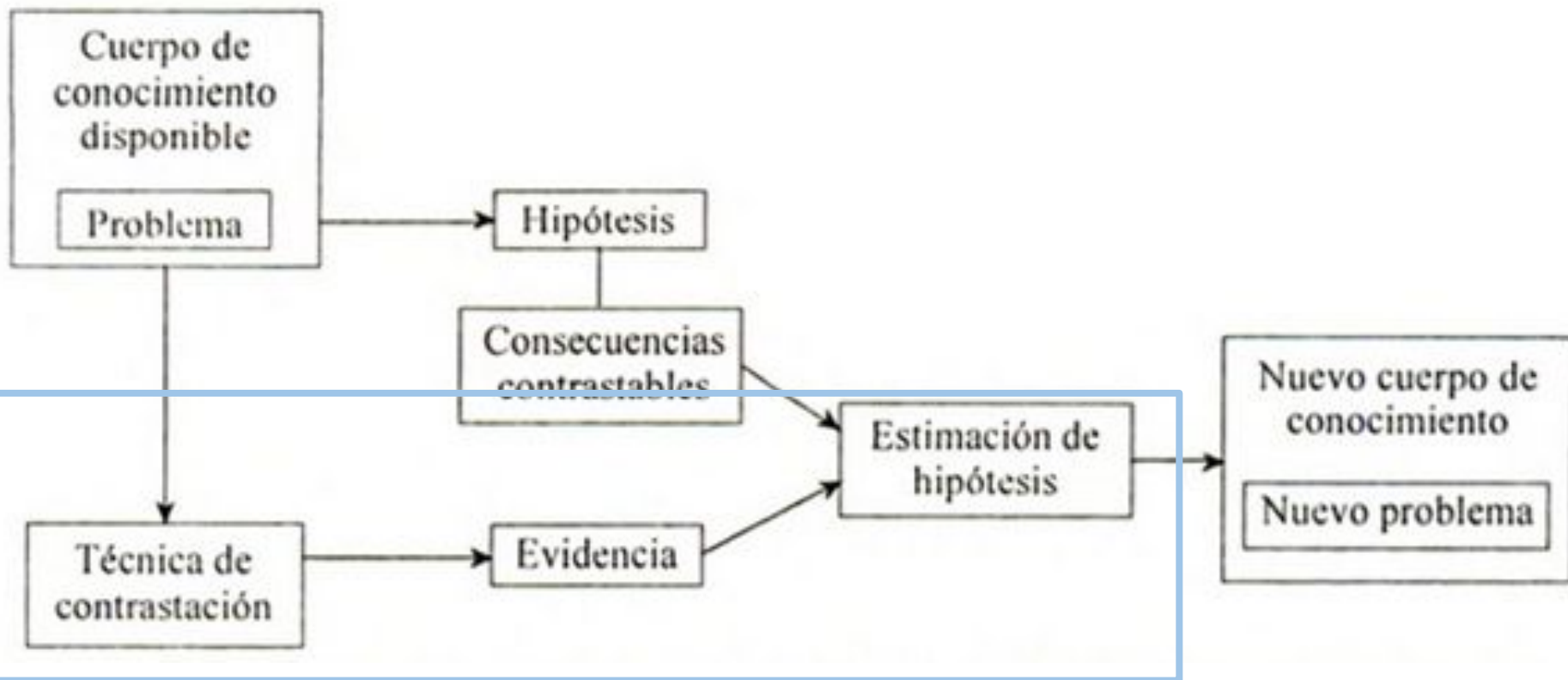
- Unidad 1. Intro R.
- Unidad 2. Análisis exploratorio.
- Unidad 3. Procesamiento de datos en R
- Unidad 4. Fundamentos de estadística inferencial
- Unidad 5. Pruebas de hipótesis.





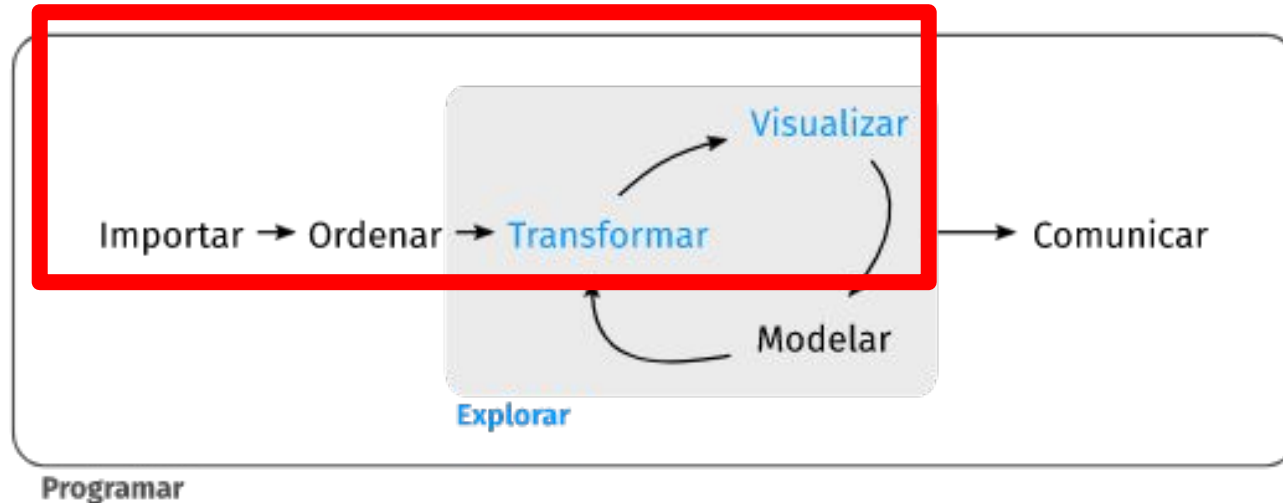
La pregunta



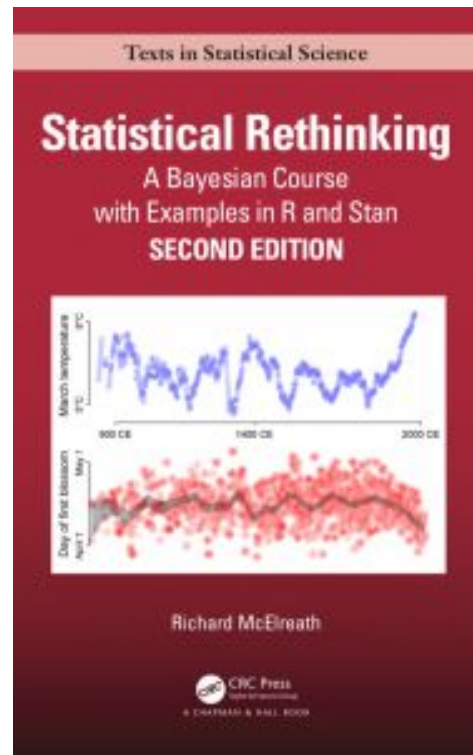


Los datos

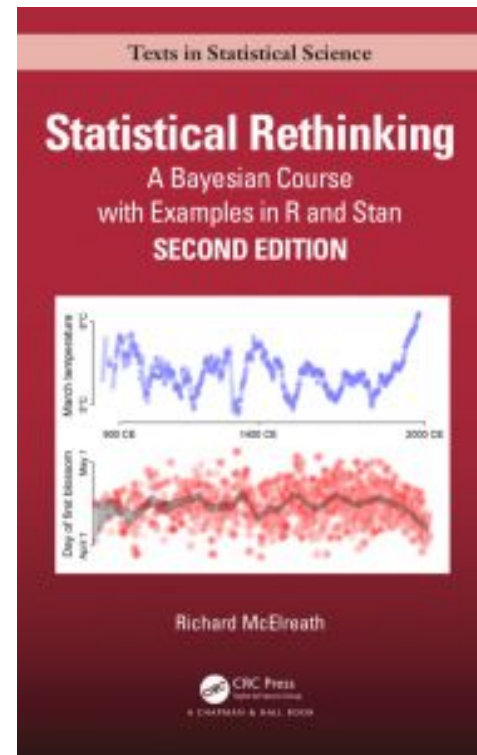
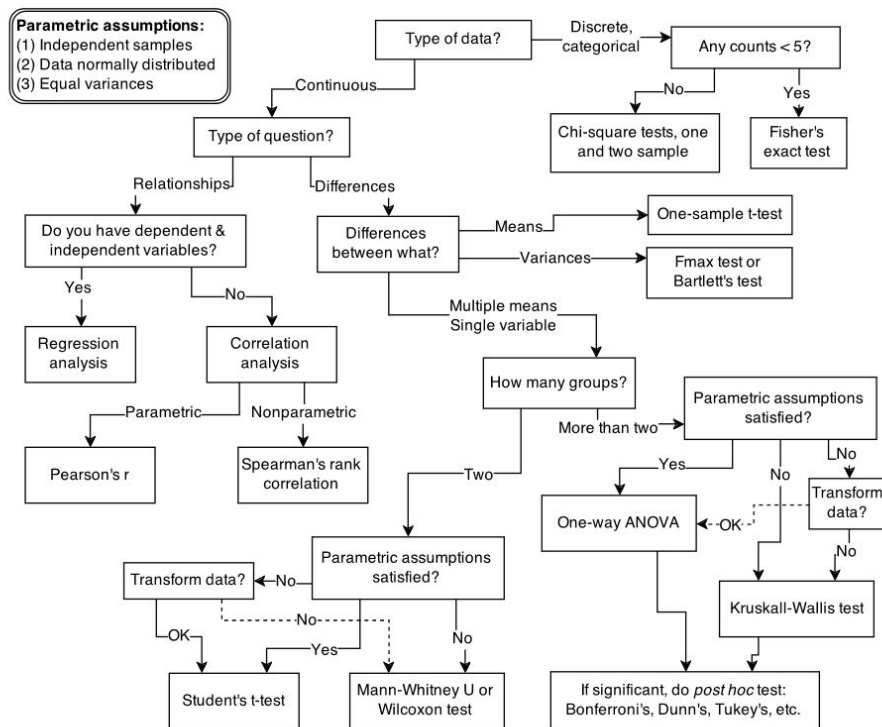
Las etapas de análisis de datos



De modelos y de golems



Golems estadísticos (y algorítmicos)



¿Qué es R?

- Lenguaje de programación => análisis y visualización de datos
- R es un “dialecto” de un lenguaje de los años '70: S-Language
- S fue creado en los Bell Labs
- R creado en 1991 por Ross Ihaka y Robert Gentleman
- 1993: R se anuncia por primera vez
- 2000: se lanza la primera versión R 1.0
- 2021: la versión más actual es la R 4.1.2

¿Por qué usar R?

- Modularidad: conjunto de funciones básicas al cual se le van agregando diferentes paquetes con funcionalidades específicas
- Paquetes instalables: siempre hay nuevas funcionalidades “customizables” para lo que queremos hacer.
- Corre en casi cualquier SO/plataforma (incluso en PS3)
- Muy buenas capacidades gráficas
- Lo mejor de todo: la comunidad. Cada estadístico que se le ocurre un algoritmo nuevo lo programa en R

¿Por qué usar R?

- Lo segundo mejor: GRATIS.
- Filosofía “free software”
- Libertad de correr el soft con cualquier propósito (grado 0)
- Libertad de estudiar cómo funciona el programa y adaptarlo a las necesidades (grado 1). Requisito: disponer del código fuente
- Libertad de redistribuir copias (grado 2)
- Libertad de mejorar el software y lanzar las mejoras al público (grado 3).
Mismo requisito que grado 1.

Herramientas

- Unidad 1. Intro R.
- Unidad 2. Análisis exploratorio.
- Unidad 3. Procesamiento de datos en R
- Unidad 4. Fundamentos de estadística inferencial
- Unidad 5. Pruebas de hipótesis.



Vamos al Notebook