Procesamiento de datos en R y estadística para Ciencias Sociales

Clase 1. Presentación e introducción a R

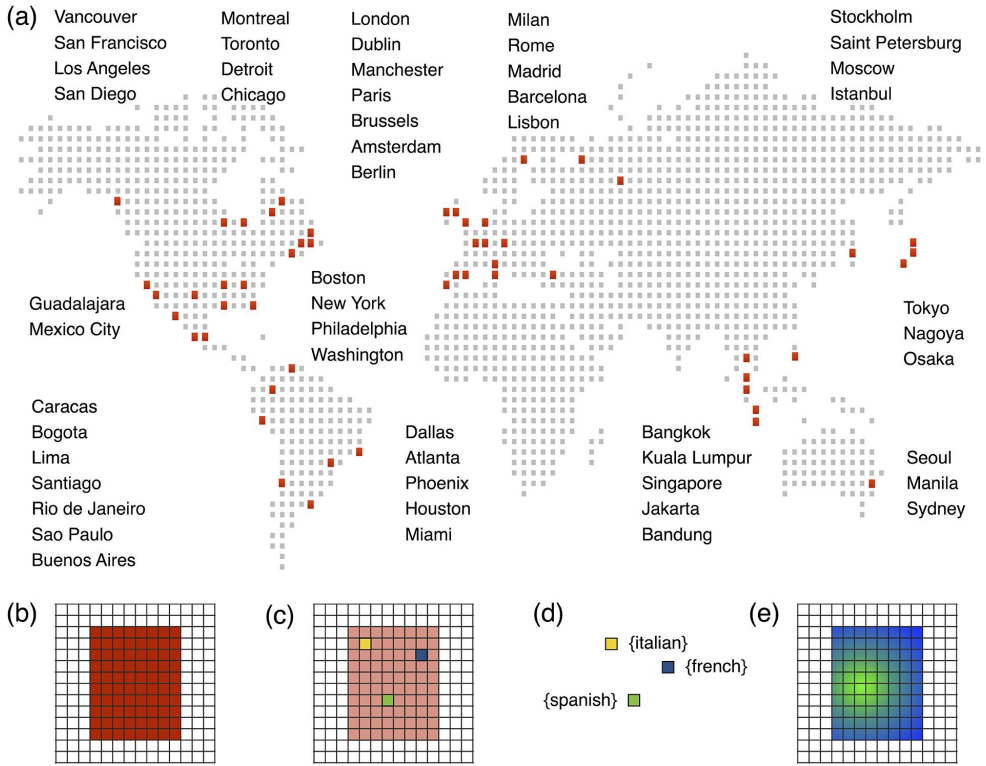
factor~data
IDAES_UNSAM

¿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...

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]

¿Qué son las Ciencias Sociales Computacionales?

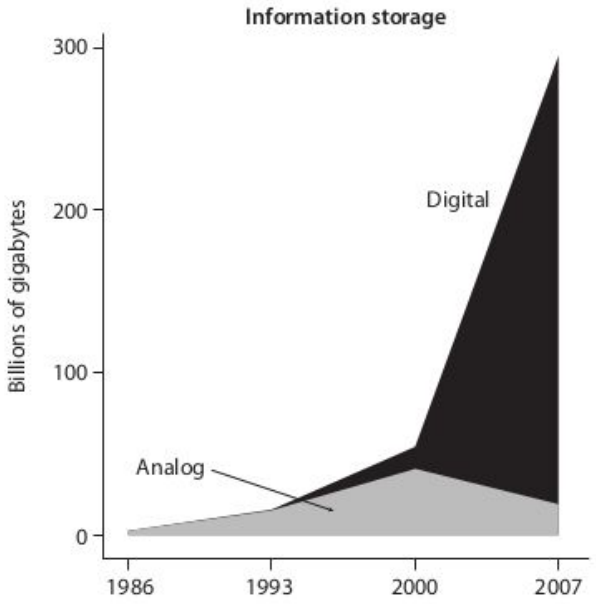
- **Problemas/preguntas** de investigación más o menos clásicas

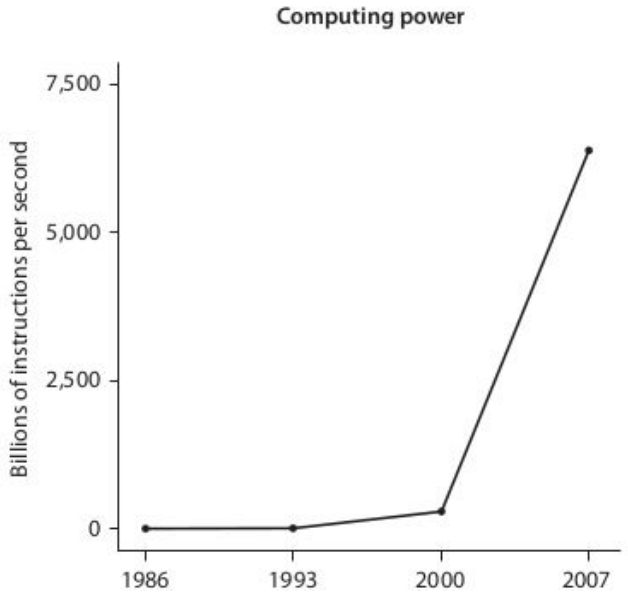
- Uso intensivo de algoritmos, cálculos y métodos de predicción
  - Métodos cuantitativos/estadísticos clásicos
  - Machine Learning

- 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





[Salganik, 2018]

El problema de los datos

MAS_500 Aglomerados segun tamanio	AGLOMERADO Codigo de Aglomerado	PONDERA Ponderacion	CH03 Relacion de parentesco	CH04 Sexo	CH05 Fecha de nacimiento (dia, mes y anio)
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

14/06/1976

Presentación Track

Trayecto de Métodos Cuantitativos y Ciencias Sociales Computacionales

- 4 materias optativas
- Computan 100 horas de investigación
- Opcionalmente, se puede computar un taller de tesis
- Correlatividades:
  - Metodología de la Investigación
  - Metodologías Cuantitativas

Equipo

- Germán Rosati
- Adriana Chazarreta
- Laia Domenech
- Tomás Maguire

factor~data
IDAES_UNSAM

Trayecto de Métodos Cuantitativos y Ciencias Sociales Computacionales

Procesamiento de datos con R para ciencias sociales

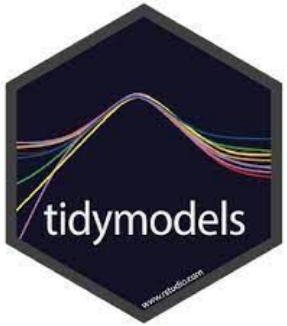
- Programación estadística en R.
- Limpieza y procesamiento de datos
- Estadística descriptiva e inferencial
- Fundamentos de visualización de datos

factor~data
IDAES_UNSAM

Trayecto de Métodos Cuantitativos y Ciencias Sociales Computacionales

Métodos de análisis cuantitativos multivariados

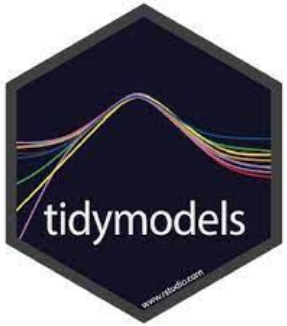
- Regresión lineal y logística
- Introducción a las técnicas de clustering
- Metodología del aprendizaje automático (machine learning).
- Introducción a tidymodels.



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

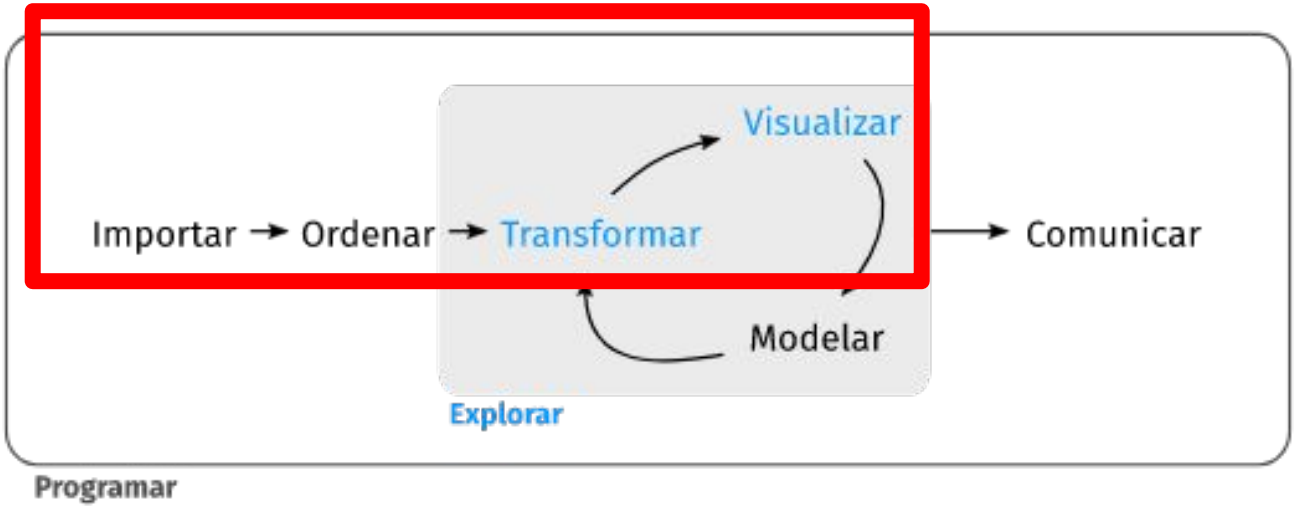
Medios de comunicación

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

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.

Las etapas de análisis de datos



Fuente: https://es.r4ds.hadley.nz/explorar-introduccion.html

¿Qué es R?

- Lenguaje de programación => análisis y visualización de datos
- Básicamente, 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 Ithaka y Robert Gentelman
- 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 >= hay un 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 adapatarlo 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