Clase 1

Políticas públicas y ciencia de datos

Aproximación a las políticas públicas desde los datos | UC | 01 de junio, 2023

José D. Conejeros | jdconejeros@uc.cl

Guía

1. Presentación y Motivación
2. Big Data para las Ciencias sociales
3. Taller: introducción a R

1. Presentación y Motivación

Sobre el profesor

José Daniel Conejeros

Sociología y Estadística, Pontificia Universidad Católica

Vulnerablidad social, salud y métodos estadísticos aplicados

jdconejeros@uc.cl

JDConejeros

jd-conejeros.com

¿De qué se trata este curso?

El concepto de políticas basadas en evidencia adquiere cada vez más relevancia en distintos sectores dedicados a el análisis y diseño de políticas públicas, por lo que se hace necesario contar con profesionales preparados para estos efectos.

La era digital ha producido una explosión de datos conductuales y relacionales de distinta naturaleza comúnmente denominados big data. La ciencia social computacional (CSC) utiliza herramientas computacionales para procesar, analizar, visualizar y modelar estos datos. Este curso introduce a 2 áreas de interés para la CSC: **1) la extracción y manipulación de bases de datos** y **2) técnicas aplicadas al análisis de datos geográficos (GD), redes sociales (SNA) y el procesamiento del lenguaje natural (NLP)**.

¿Qué tipo de curso es este?

Este curso tiene un **carácter práctico donde se utilizará de base el lenguaje de programación R**. Los estudiantes formarán grupos y realizarán proyectos de investigación a lo largo del curso, revisaran artículos aplicados en el área y realizaran tareas individuales. Al terminar el curso, los estudiantes sabrán realizar extracciones de datos vía web scraping o el uso de API's y manipular bases de datos en R. Además, podrán aplicar técnicas para datos de distinta naturaleza y preguntas de investigación.

Contenidos

- Introducción a la programación para la política pública
- Manipulación de datos
- Web scrapping y APIs
- Técnicas de análisis descriptivo e inferencia estadística
- Análisis de redes sociales
- Análisis de datos geográficos
- Análisis automatizado de texto

¿Qué se espera al final de este curso?

Al terminar el curso, los estudiantes sabrán:

1. Identificar tipos de datos para distintas preguntas de investigación.
2. Identificar ventajas y desventajas de distintos métodos empíricos.
3. Aplicar métodos de ciencia de datos utilizando análisis estadístico

Los puntos anteriores se realizarán utilizando R, por lo que **los/as estudiantes profundizarán su conocimiento de este lenguaje de programación**.

¿Cuál es la metodología de este curso?

El curso es práctico por lo que tiene el propósito de enseñar a programar para proyectos, por lo que tendremos actividades:

- **Interactivas**: los estudiantes deberán programar análisis a partir de una pregunta/objetivo de investigación
- **De análisis**: los estudiantes deberán leer investigaciones en el área y exponer a sus compañeros los aspectos principales
- **Aplicación**: los estudiantes deberán aplicar alguna de las técnicas vistas a una pregunta **relevante para las políticas públicas**

¿Cómo se evalua este curso?

El curso tendrá 4 evaluaciones:

- **Tarea 1 (25%) y Tarea 2 (25%)**

A partir de sus preguntas de investigación los estudiantes deberán procesar, analizar y visualizar datos. Se penalizará con 0,5 décimas por día de atraso. Después de 3 días no se recibirá la evaluación y será calificada con nota 1,0.

¿Cómo se evalua este curso?

- **Presentación paper de investigación (20%)**

La presentación de paper es en grupo de 3 a 5 personas. Se designara a cada grupo un *paper* que deberán leer y estudiar de forma autodidacta. Luego deberán realizar una presentación a la clase explicando el objetivo del artículo, la metodología implementada y las limitaciones observadas. Los compañeros/as de la clase deberá participar con preguntas y comentarios. Se evaluará la claridad y precisión en cada una de las presentaciones.

¿Cómo se evalua este curso?

- **Proyecto final de investigación (30%)**

El proyecto final corresponde a la aplicación de una de las técnicas revisadas y equivale al examen del curso. En grupos de 5 personas, deberán definir una pregunta de investigación viable para ser desarrollada a lo largo del curso. Luego, se debe realizar el procesamiento, análisis, visualización e implementación de la técnica. Debe justificar de la mejor manera posible la elección de la técnica utilizada.

Deberá entrega un documento con los códigos de análisis y realizar una presentación. La inasistencia a clases ese día será calificada con nota 1,0.

Calendarización del curso

- Tarea 1: se publica el enunciado el 30/05/2023 (2 semanas para realizarlo)
- Tarea 2: se publica el enunciado el 30/06/2023 (2 semanas para realizarlo)
- Presentación paper empírico: Inscribirse aquí
- Entrega proyecto de investigación: presentación el 05/08/2023 (Última clase del curso)

Literatura del curso

El curso tendrá lecturas obligatorias y complementarias clase a clase.

Introducción al big data y a la ciencia social computacional

Salganik, Matthew J. 2018. Bit by Bit: Social Research in the Digital Age. Princeton, New Jersey: Princeton University Press. Click aquí

Cioffi-Revilla, C. (2014). Introduction to computational social science. London and Hei-delberg: Springer. 2ºEdición. Click aquí

Lazer, David et al 2009. 'Computational Social Science'. Science 323(5915):721–23. Click aquí

Wickham, H., & Grolemund, G. (2016). R for data science: import, tidy, transform, visualize, and model data. " O'Reilly Media, Inc.". Recurso en línea: Click aquí

Literatura del curso

El curso tendrá lecturas obligatorias y complementarias clase a clase.

Introducción al big data y a la ciencia social computacional

Urdinez, F., & Cruz, A. (2020). R for Political Data Science: A Practical Guide. CRC Press. Recurso en línea en español: Click aquí

Healy, K. (2018). Data visualization: a practical introduction. Princeton University Press. Recurso en línea: Click aquí

Breiman, L. (2001). Statistical modeling: The two cultures (with comments and a rejoinder by the author). Statistical science, 16(3), 199-214. **Presentación de estudiantes 1** Click aquí

Grimmer, J., Roberts, M. E., & Stewart, B. M. (2021). Machine learning for social science: An agnostic approach. Annual Review of Political Science, 24, 395-419. **Presentación de estudiantes 2** Click aquí

Literatura del curso

Recolección y manipulación de datos

Urdinez, F., & Cruz, A. (2020). R for Political Data Science: A Practical Guide. CRC Press. Recurso en línea en español: Click aquí

Wickham, H., & Grolemund, G. (2016). R for data science: import, tidy, transform, visualize, and model data. " O'Reilly Media, Inc.". Recurso en línea: Click aquí

Cioffi-Revilla, C. (2014). Introduction to computational social science. London and Hei-delberg: Springer. 2ºEdición. Click aquí

Wickham, H. (2019). Advanced r. CRC press. Recurso en línea: Click aquí

Literatura del curso

Aplicaciones CSC

van ‘t Wout, E., Pieringer, C., Torres Irribarra, D., Asahi, K., & Larroulet, P. (2021). Machine learning for policing: a case study on arrests in Chile. Policing and society, 31(9), 1036-1050. **Presentación de estudiantes 3** Click aquí

Rossetti, T., Lobel, H., Rocco, V., & Hurtubia, R. (2019). Explaining subjective perceptions of public spaces as a function of the built environment: A massive data approach. Landscape and urban planning, 181, 169-178. **Presentación de estudiantes 4** Click aquí

Méndez, J. T., Lobel, H., Parra, D., & Herrera, J. C. (2019). Using Twitter to infer user satisfaction with public transport: the case of Santiago, Chile. IEEE Access, 7, 60255-60263. **Presentación de estudiantes 5** Click aquí

Literatura del curso

Aplicaciones CSC

Bro, N., & Mendoza, M. (2021). Surname affinity in Santiago, Chile: A network-based approach that uncovers urban segregation. PloS one, 16(1), e0244372. **Presentación de estudiantes 6** Click aquí

Bronfman, N. C., Repetto, P. B., Guerrero, N., Castañeda, J. V., & Cisternas, P. C. (2021). Temporal evolution in social vulnerability to natural hazards in Chile. Natural hazards, 107(2), 1757-1784. **Presentación de estudiantes 7** Click aquí

Beytía, P., & Müller, H. P. (2022). Towards a Digital Reflexive Sociology: Using Wikipedia's Biographical Repository as a Reflexive Tool. Poetics, 95, 101732. **Presentación de estudiantes 8** Click aquí

Literatura del curso

Aplicaciones CSC

Cioffi-Revilla, C. (2014). Introduction to computational social science. London and Hei-delberg: Springer. 2ºEdition. Cap. 4. Click aquí

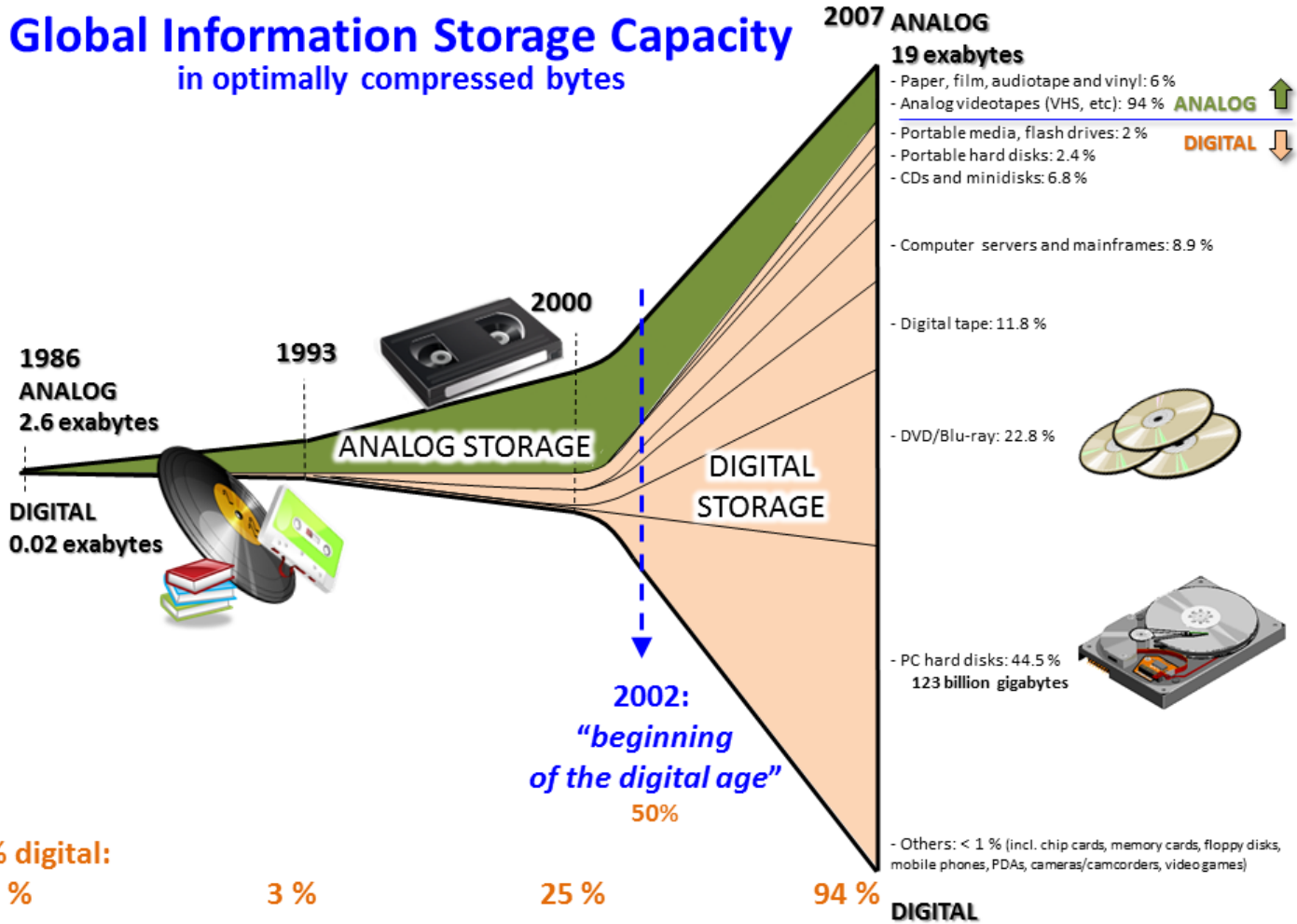
Lovelace, R., Nowosad, J., & Muenchow, J., (2019), Geocomputation with R, Chapman and Hall/CRC. Click aquí

Silge, J., & Robinson, D. (2017). Text mining with R: A tidy approach. " O'Reilly Media, Inc.". Recurso en línea: Click aquí

2. Big Data para las Ciencias sociales

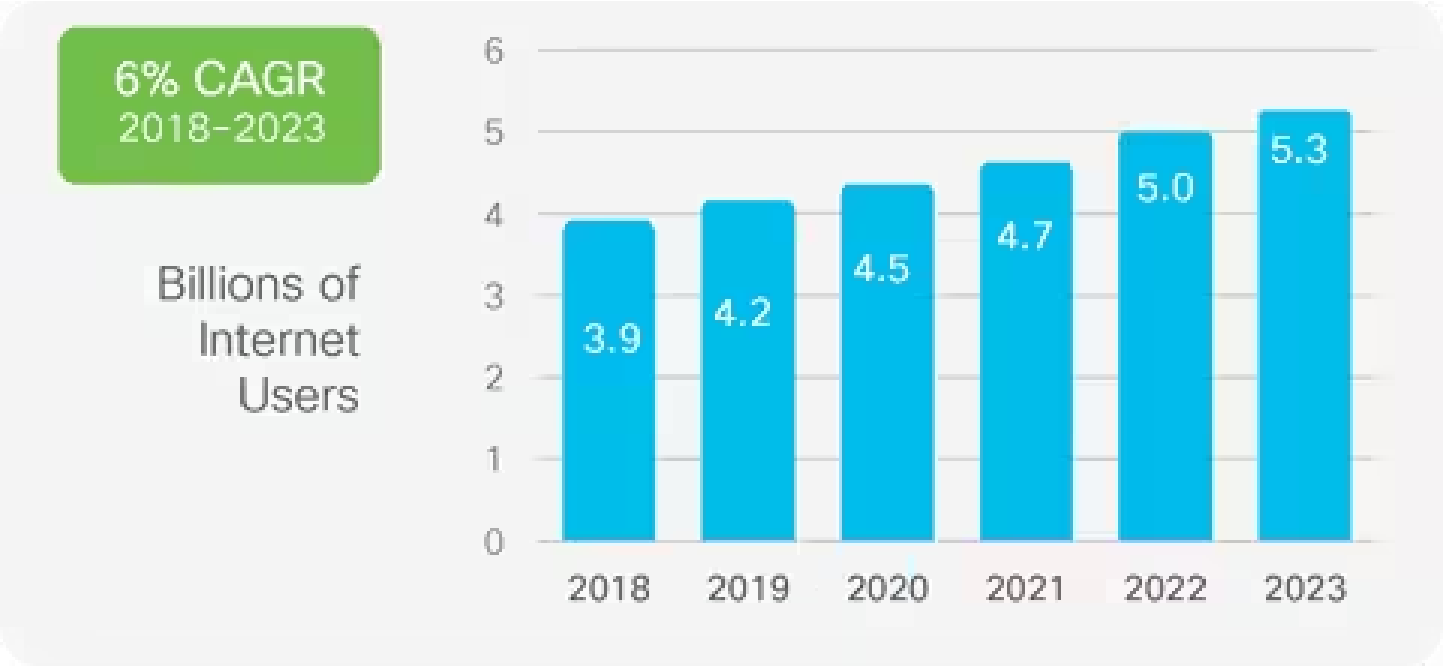
"The digital age is everywhere, it’s growing, and it changes what is possible for researchers (Salganik, 2018)."

Big Data



Source: Hilbert, M., & López, P. (2011). The World's Technological Capacity to Store, Communicate, and Compute Information. *Science*, 332(6025), 60 –65. http://www.martinhilbert.net/WorldInfoCapacity.html

Big Data



Fuente: Cisco Annual Internet Report (2018–2023) White Paper

Big Data

El paso de lo digital a lo análogo implica grandes desafíos.

- Cambios rápidos en la informática: computadores personales, portátiles, celulares, etc.
- Internet de las cosas $\rightarrow$ sensores digitales
- Nuevas formas de comunicación y **socialización**
- Experimentación rápida con la información
- Grandes volúmenes de información
- Nuevas arquitecturas y métodos
- Ética

La investigación social debe tomar ideas de las ciencias sociales y la ciencia de datos para aprovechar las oportunidades de la era digital. Para esto, el diseño computacional será clave

Big Data

> Big data are created and collected by companies and governments for purposes other than research. Using this data for research therefore requires repurposing.

Salganik, M. (2018)

Big Data

¿3V, 5V o 7V? Podemos caracterizar el big data como: Volumen, variedad y velocidad

- Veracidad, valor (P)
- Vago y Vacuo (N)

Sin embargo, desde la investigación es mejor preguntarse por **5W**: Who, What, Where, When, and **Why**.

Los datos son más allá de la investigación $\rightarrow$ Empresas y Gobiernos

Oportunidades y desafíos

Big Data

¿Cuál sería la mejor opción para investigar opinión pública?

Big Data

Características del Big Data:

Ventajas para la investigación

- Big: datos como un medio para alcanzar un fin.
- Always on: permiten estudiar fenómenos en tiempo real.
- Nonreactive: menos probable de cambiar el comportamiento de las personas (sesgo de deseabilidad).

Big Data

Características del Big Data:

Desafíos para la investigación

- Incomplete: puede que no tenga la información que requiera nuestra pregunta de investigación.
- Inaccessible: es difícil acceder a los datos que manejan empresas y gobiernos.
- No representative: los datos no representativos son malos para las generalizaciones fuera de la muestra, pero pueden ser muy útiles para las comparaciones dentro de la muestra.
- Drifting: las herramientas de medida no son estables en el tiempo (población, comportamiento o el sistema).
- Algorithmically confounded: el comportamiento en las plataformas no es natural sino que viene condicionado por los objetivos del propio sistema que los soporta. En otras palabras, inducen el comportamiento. Es fundamental entender el **proceso de generación de los datos**,
- Dirty: es necesario trabajar en la limpieza de datos.
- Sensitive: trabajar con información sensible.

Ciencias Sociales Computacionales

Las ciencias sociales investigan la dinámica y la organización humana y social en todos los niveles de análisis (consiliencia), incluidos la cognición, la toma de decisiones, el comportamiento, los grupos, las organizaciones, las sociedades y el sistema mundial.

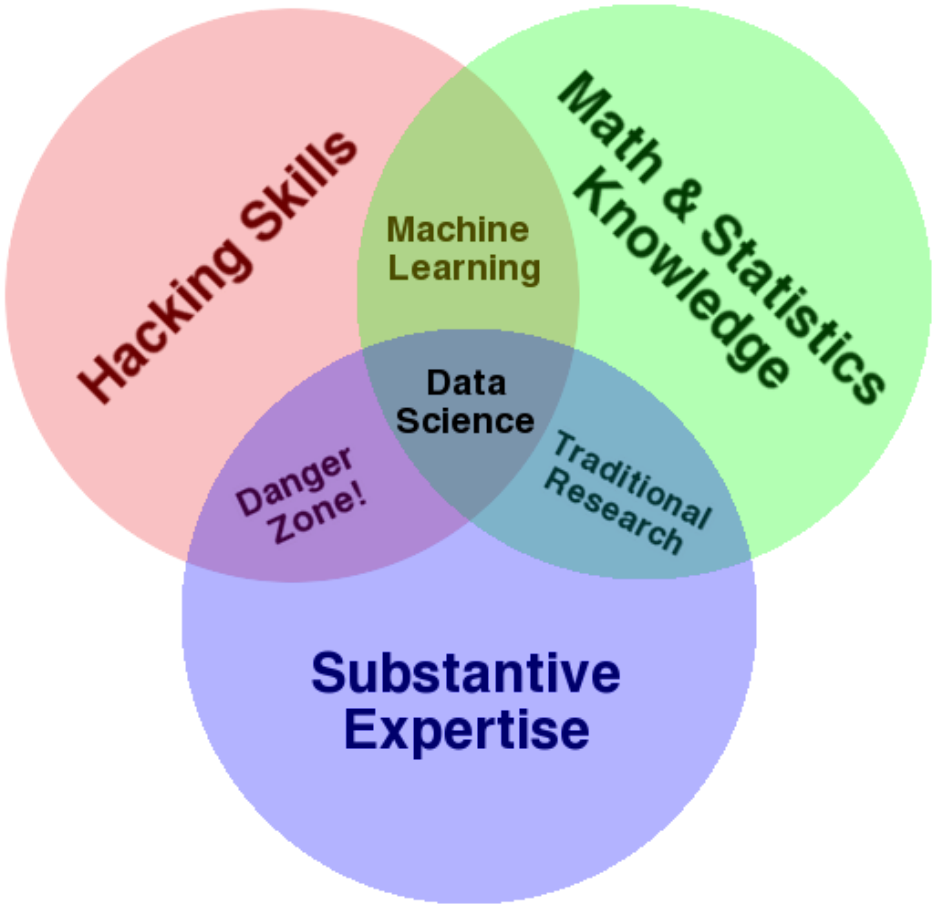
- Sociología
- Psicología social
- Antropología
- Economía
- Ciencias Políticas
- Políticas Públicas

Desde la segunda mitad del siglo XX y la invención de los ordenadores comenzó a emerger una disciplina que utilizaba la estadística para el análisis de datos sociales. Algunos referentes:

- Herbert A. Simon (1916-2001)
- Karl W. Deutsch (1912-1992)
- Harold Guetzkow (1915-2008)
- Thomas C. Schelling (1921-2016)

La estadística como método científico es importante para todas estas áreas

Ciencias Sociales Computacionales



Fuente: http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram

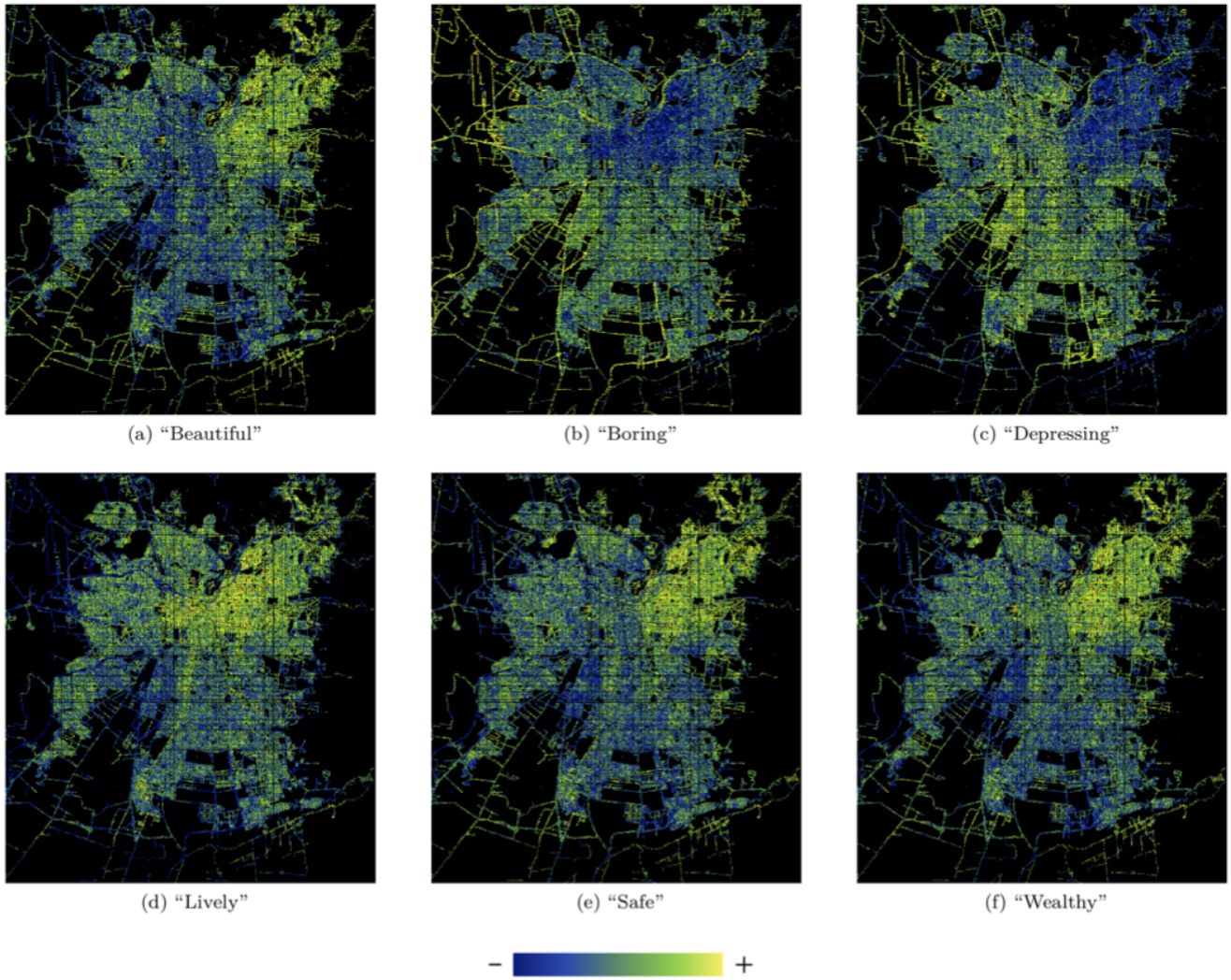
Ciencias Sociales Computacionales

La ciencia social computacional es una disciplina científica instrumentalizada, científicamente similar a la microbiología, la radioastronomía o la nanociencia -nuevos campos científicos de investigación facilitados por el microscopio, el radar y el microscopio electrónico, respectivamente-. En este caso el instrumento de investigación es el que impulsa el desarrollo de la teoría y comprensión:

- **Extracción automatizada de información**
- **Análisis de redes sociales (SNA)**
- **Análisis geoespacial (sistemas de información geográfica o GIS social)**.
- Modelización de la complejidad
- Modelos de simulación social

Esto nos permite observar más allá de lo disponible en la visiones tradicionales y aproximarnos a problemas y preguntas de políticas públicas relevantes.

Aplicaciones: Rossetti, T., et. al. (2019)



Aplicaciones: Bro, N., & Mendoza, M. (2021).

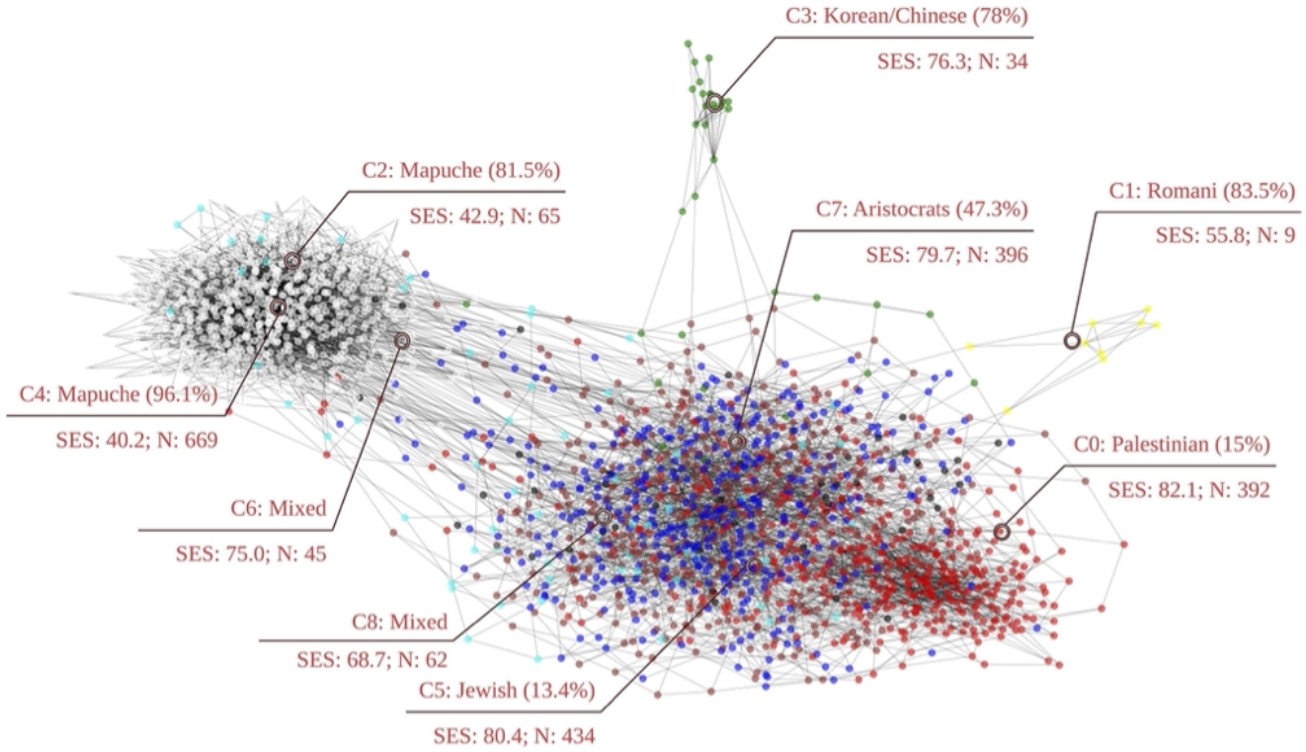
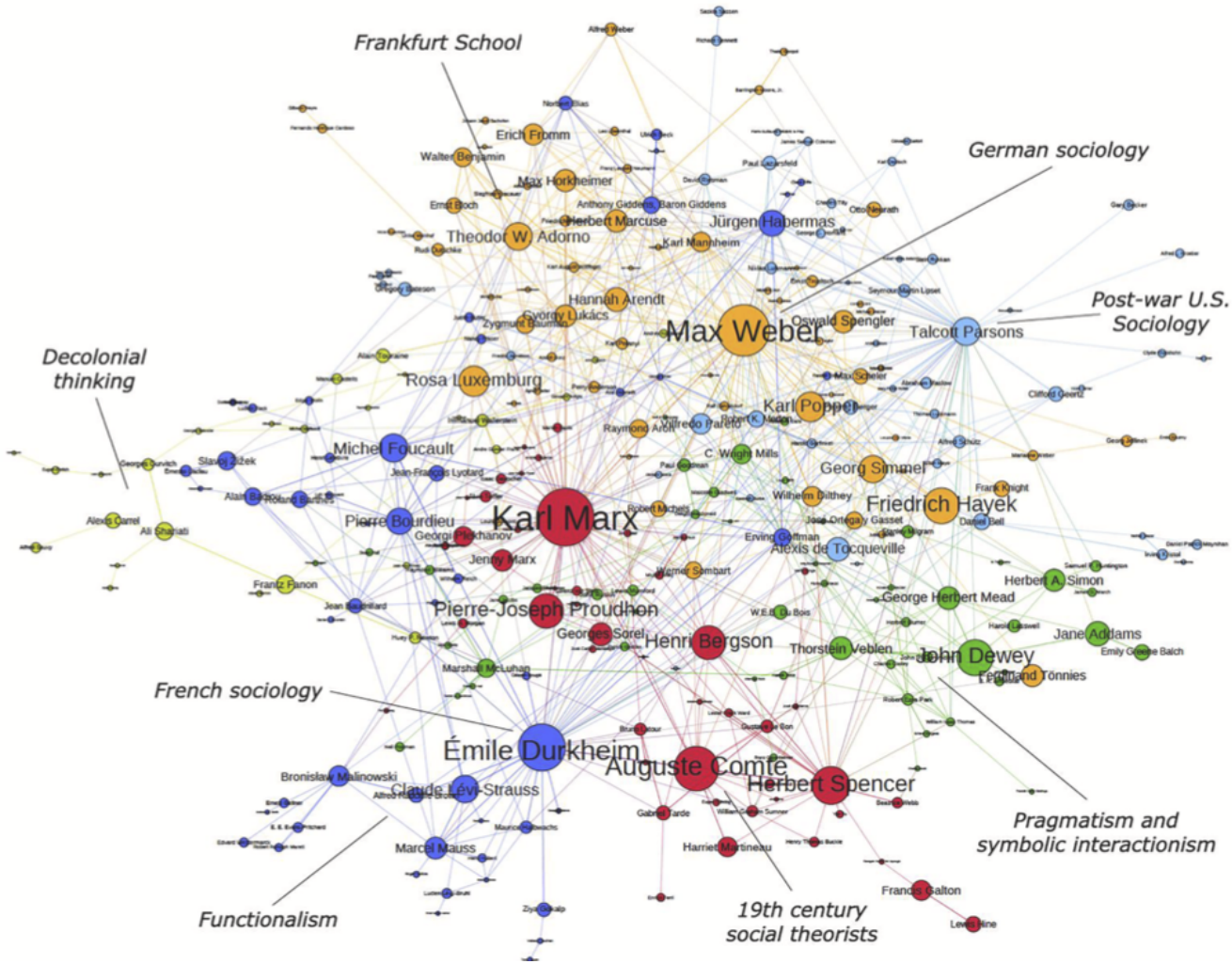


Fig 1. Communities detected on the paternal-maternal surname network. Some communities show a strong presence of surnames related to an ethnic group, while others are mixed. The two communities with the most significant presence of Mapuche are also those with the lowest SES (socioeconomic status) index. The communities with higher SES exhibit a strong presence of aristocratic, Jewish, and Palestinian surnames. Finally, some smaller communities show groups with little connection with the rest of society but a robust internal connectivity, as is the Romani and the Korean communities.

Aplicaciones: Beytía, P., & Müller, H. P. (2022).



Aplicaciones: Bronfman, N. C., et. al. (2021).

Aplicaciones: Ginsberg, et.al. (2009)

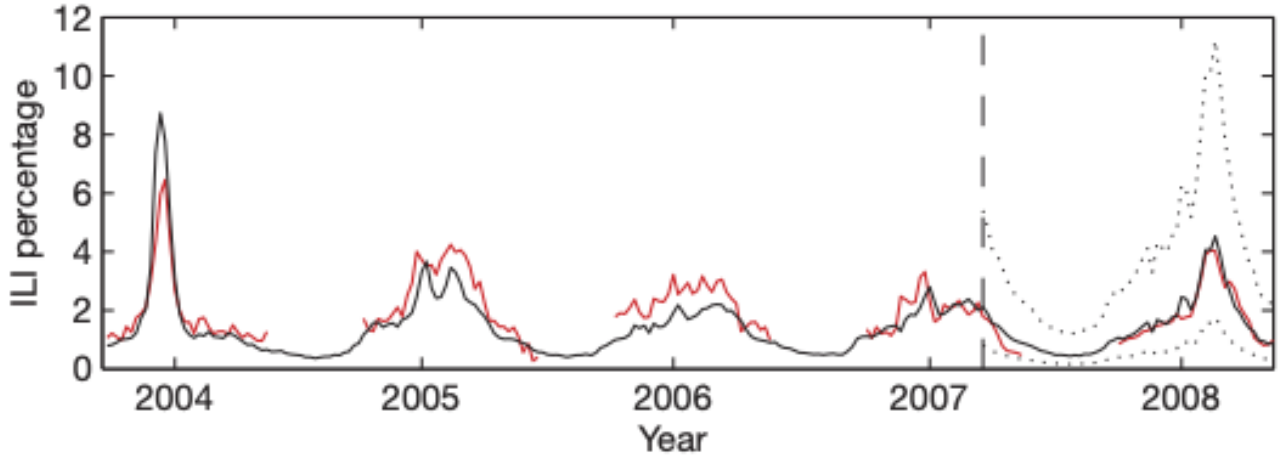


Figure 2 | A comparison of model estimates for the mid-Atlantic region (black) against CDC-reported ILI percentages (red), including points over which the model was fit and validated. A correlation of 0.85 was obtained over 128 points from this region to which the model was fit, whereas a correlation of 0.96 was obtained over 42 validation points. Dotted lines indicate 95% prediction intervals. The region comprises New York, New Jersey and Pennsylvania.

Aplicaciones: Blumenstock, et.al. (2009)

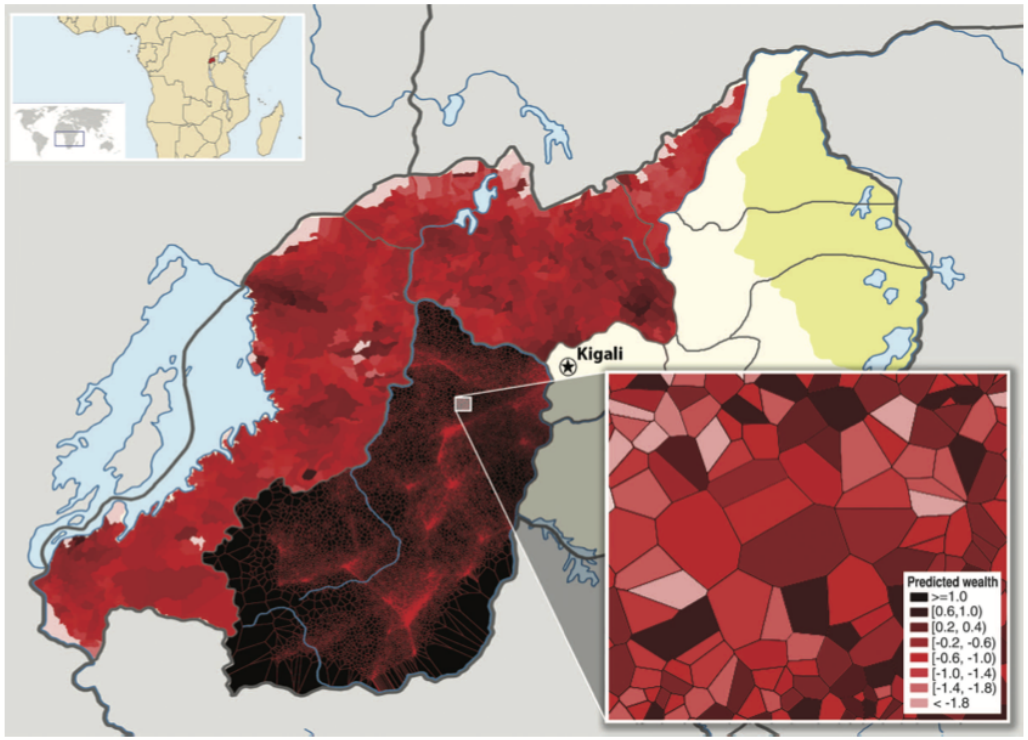


Fig. 2. Construction of high-resolution maps of poverty and wealth from call records. Information derived from the call records of 1.5 million subscribers is overlaid on a map of Rwanda. The northern and western provinces are divided into cells (the smallest administrative unit of the country), and the cell is shaded according to the average (predicted) wealth of all mobile subscribers in that cell. The southern province is overlaid with a Voronoi division that uses geographic identifiers in the call data to segment the region into several hundred thousand small partitions. (**Bottom right inset**) Enlargement of a 1-km^2 region near Kiyonza, with Voronoi cells shaded by the predicted wealth of small groups (5 to 15 subscribers) who live in each region.

Uso de los lenguajes de programación

Tweet

Sergio Toro Maureira
@yuriflame

Vengo de una reunión con varios Ministerios por temas de información y Política Pública basada en evidencia. Resumen, el nuevo profesional que no sepa R y Python va perdiendo 0-40.

Translate Tweet
3:24 PM · Jan 14, 2020 from Pudahuel, Chile · Twitter for Android

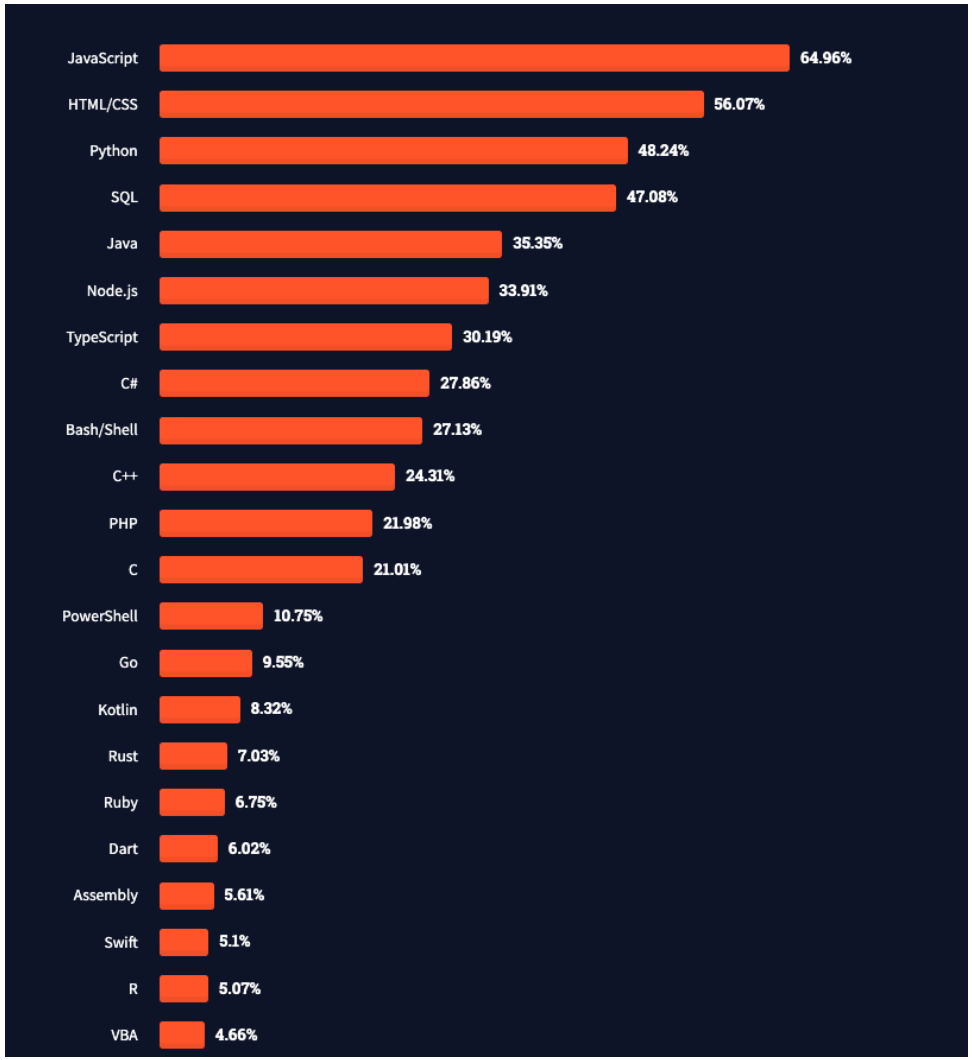
33 Retweets **139** Likes

Mr. Vine @miparra · Jan 14
Replying to @yuriflame
Igual, a veces, más que saber R y Python, es mas importante saber interpretar los resultados. Está complicado tener gente que justifica estadísticas y modelos sin tener ni la menor idea de lo que está hablando. Algo que veo con demasiada frecuencia.

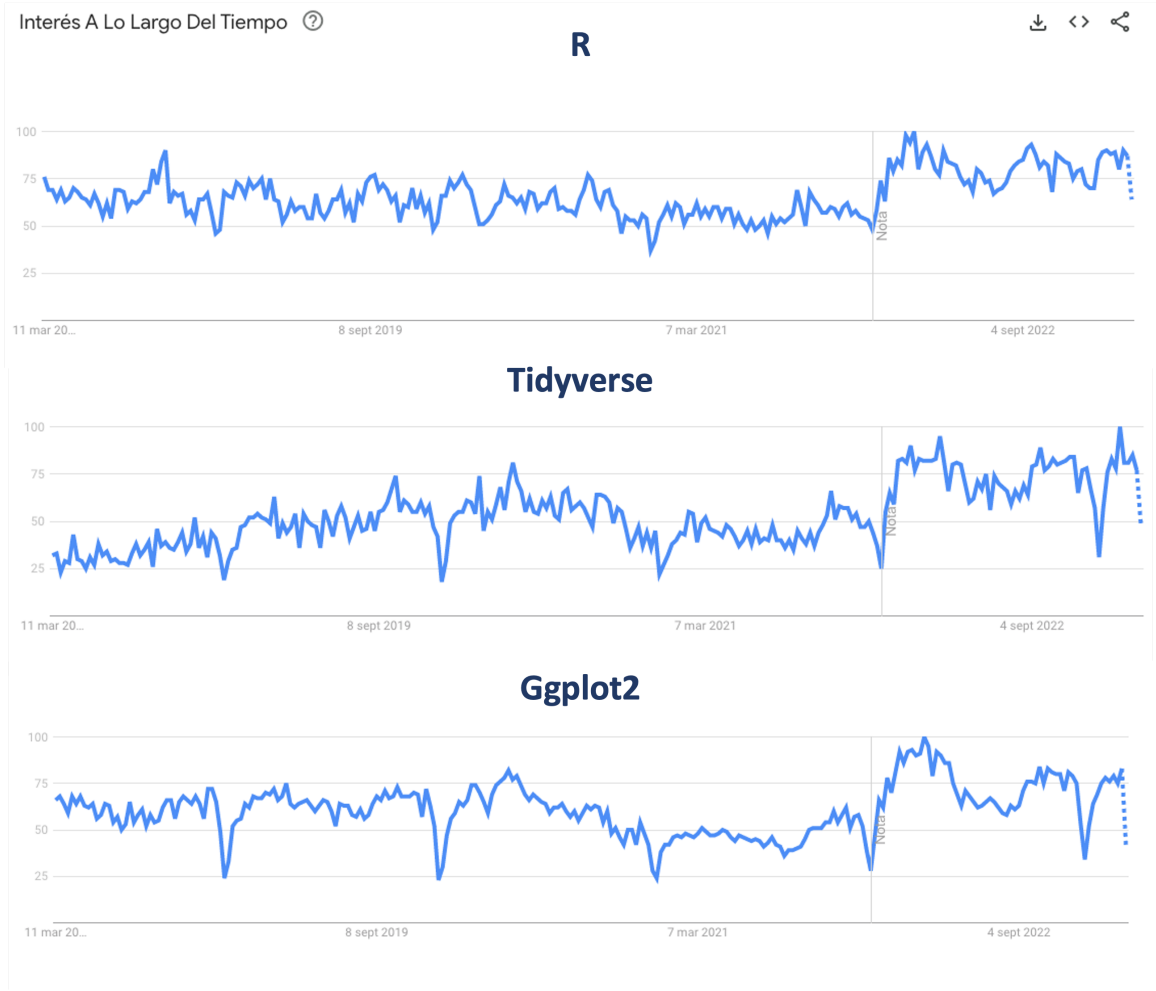
1 23

Uso de los lenguajes de programación

Uso de los lenguajes de programación



Uso de los lenguajes de programación



Introducción a R

¿Qué es R?

R es un Ambiente y lenguaje de programación libre con un enfoque estadístico para el desarrollo de herramientas, métodos, cálculos y gráficos. Utiliza un lenguaje específico, anidado principalmente en las palabras comunes de técnicas estadísticas en ingles.

En R, un análisis estadístico se realiza en una serie de pasos, con resultados intermedios que se van almacenando para ser observados o analizados posteriormente.

¿Qué es R?

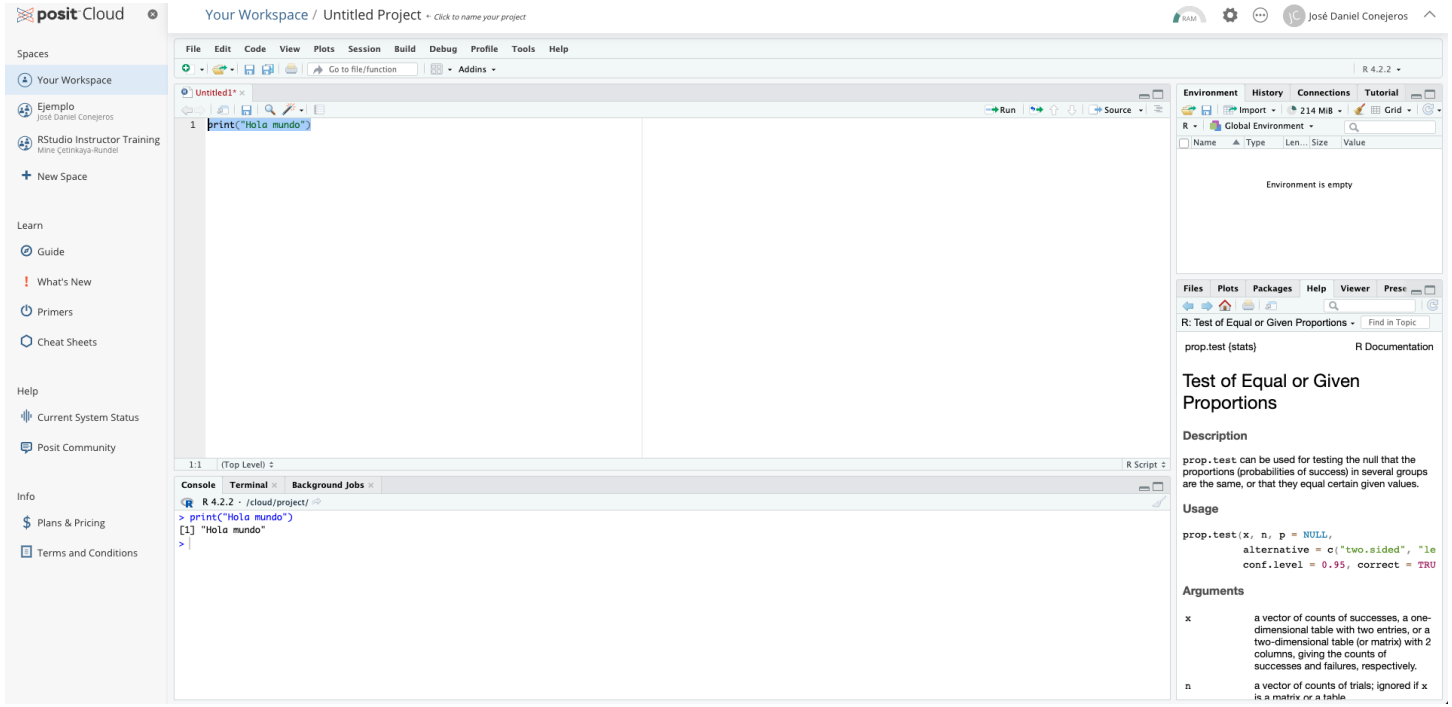
- Almacenamiento y manipulación efectiva de datos.
- Operadores para cálculo sobre variables indexadas, en particular, objetos.
- Una amplia, coherente e integrada colección de herramientas para análisis de datos.
- Ofrece un flujo completo e integrado para el análisis de datos.
- Posibilidades gráfcas para análisis de datos, que funcionan directamente sobre pantalla o para exportar.
- Lenguaje de programación de código abierto bien desarrollado, simple y efectivo.
- Amplia comunidad de desarrolladores.
- Interacción con otros lenguajes o softwares estadísticos.

¿Qué es RStudio?

Es un entorno de desarrollo integrado (IDE) para R.
Incluye varios elementos que facilitan las tareas durante el flujo de trabajo.

RStudio Cloud

Posit Cloud es un servicio web que ofrece una experiencia basada en navegador similar a RStudio, el IDE estándar para usuarios y desarrolladores de R.



Nota: Se debe crear una cuenta de usuario para acceder.

Pueden ver una guía con un click aquí

Objetos en R

```
saludo ← "Hola mundo"
print(saludo)
```

```
[1] "Hola mundo"
```

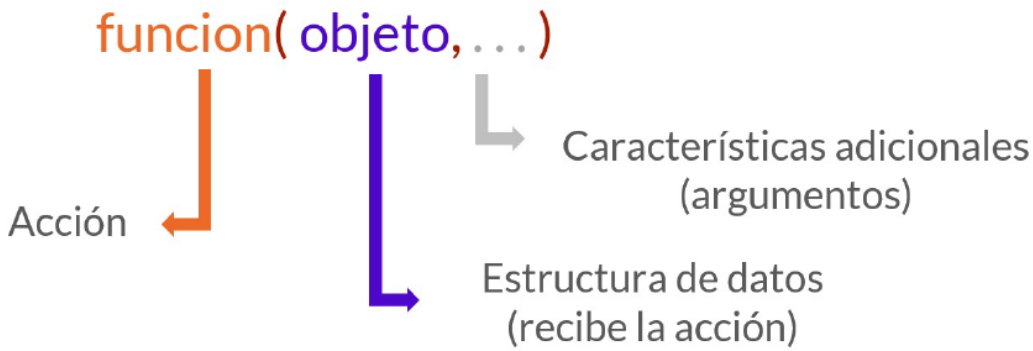
```
# Esto es un vector
nombre  ← c("José", "Constanza")

edad    ← c(29, 24)

altura  ← c(1.67, 1.65)

matriz  ← cbind(nombre, edad, altura)
matriz
```

```
     nombre      edad altura
[1,] "José"      "29" "1.67"
[2,] "Constanza" "24" "1.65"
```

```
lista   ← c(matriz, edad, nombre)
lista
```

Lógica de ejecución

Ejecutan una acción sobre nuestros datos. Algunas requieren `inputs` (argumentos) que van dentro del paréntesis.

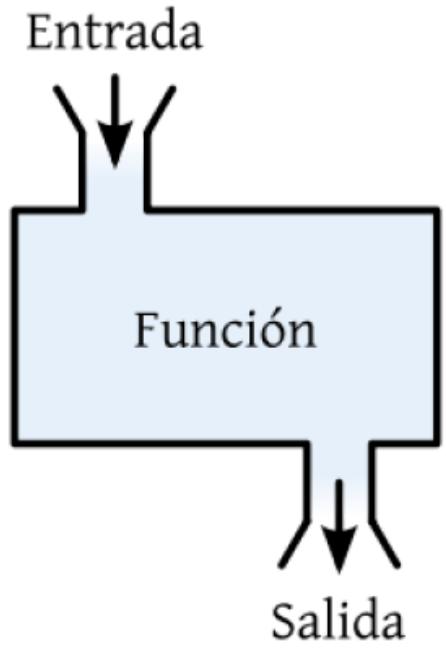


¿Qué hacer?
Función

¿A quién?
Objeto

Funciones y argumentos

Hacer que las cosas sucedan



Algunas requieren inputs y otras no, los inputs (argumentos) van dentro del paréntesis

Funciones y argumentos

```
?sum

sum( ... , na.rm = FALSE)

Arguments
 ...          numeric or complex or logical vectors.

na.rm         logical. Should missing values (including NaN) be removed?
```

```
edad
```

```
[1] 29 24
```

```
sum(edad)
```

```
[1] 53
```

Objetos $\longrightarrow$ **Valores** $\longrightarrow$ **Funciones**

Resolver dudas de código

Clase 1

Políticas Públicas y datos

01 de junio, 2023

José D. Conejeros | jdconejeros@uc.cl | JDConejeros