¿Qué es Machine Learning?

Fundamentos conceptuales

German Rosati
german.rosati@gmail.com

UNSAM - CONICET

22 de octubre de 2020

Preguntas

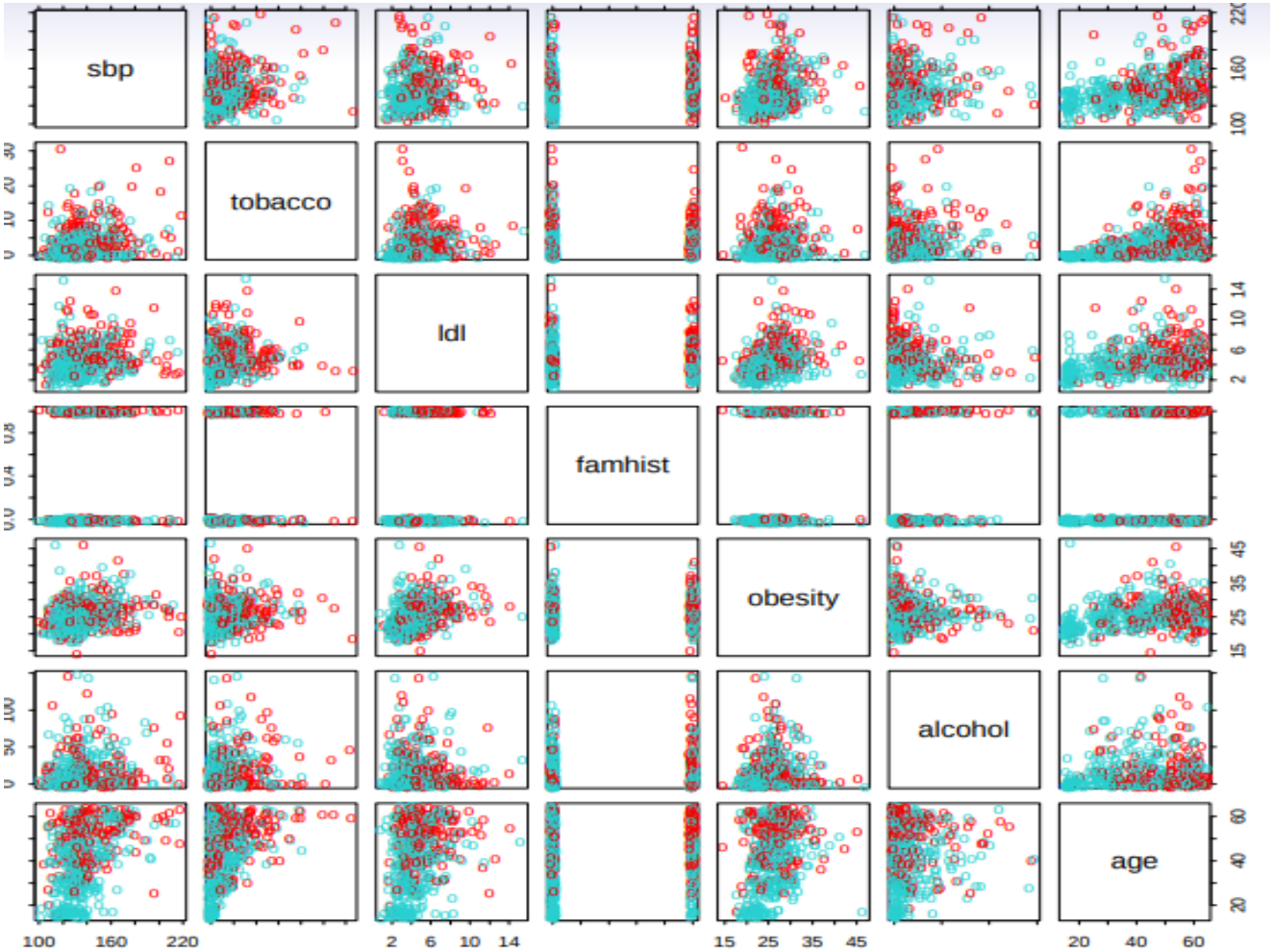
- ¿Podría una computadora ir más allá de "lo que sea que sepamos decirle que haga" y realmente "aprender" por su cuenta como realizar una determinada tarea?
- ¿Podría ser posible el aprendizaje automático de estas reglas a partir de los datos?

¿Qué problemas se pueden abordar?

- Riesgo de cáncer prostático
- Predicción de ataques cardíacos en función de variables demográficas, orgánicas, médicas, etc.
- Generar un detector de spam automático para tu cuenta de correo
- Reconocer dígitos en números manuscritos
- Clasificar muestras de tejido en diferentes tipos de cáncer basado en ciertos perfiles genéticos
- Establecer relaciones entre el ingreso y variables demográficas
- Clasificar imágenes satelitales
- Identificar tópicos en corpus textuales
- Clasficar textos según su sentimiento -positivos o negativos-
- Detectar objetos en imágenes

¿Qué es Machine Learning?

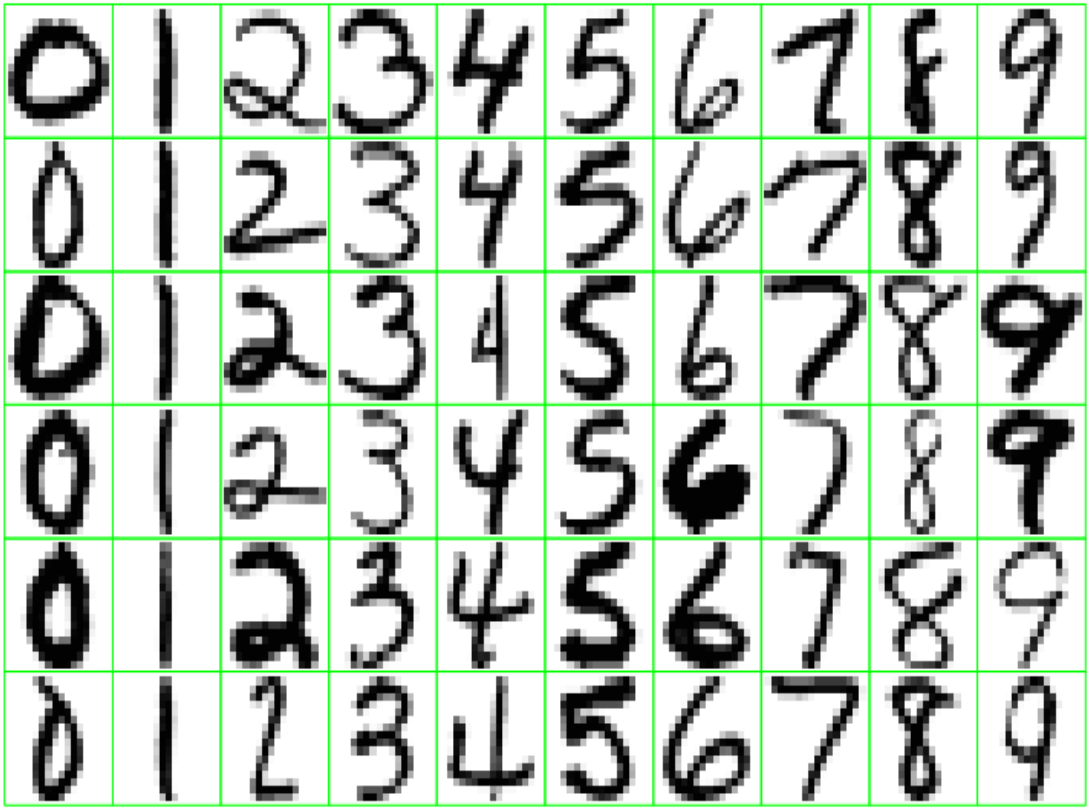
Ataques cardíacos



- Hastie, et al. 2013 [8]

¿Qué es Machine Learning?

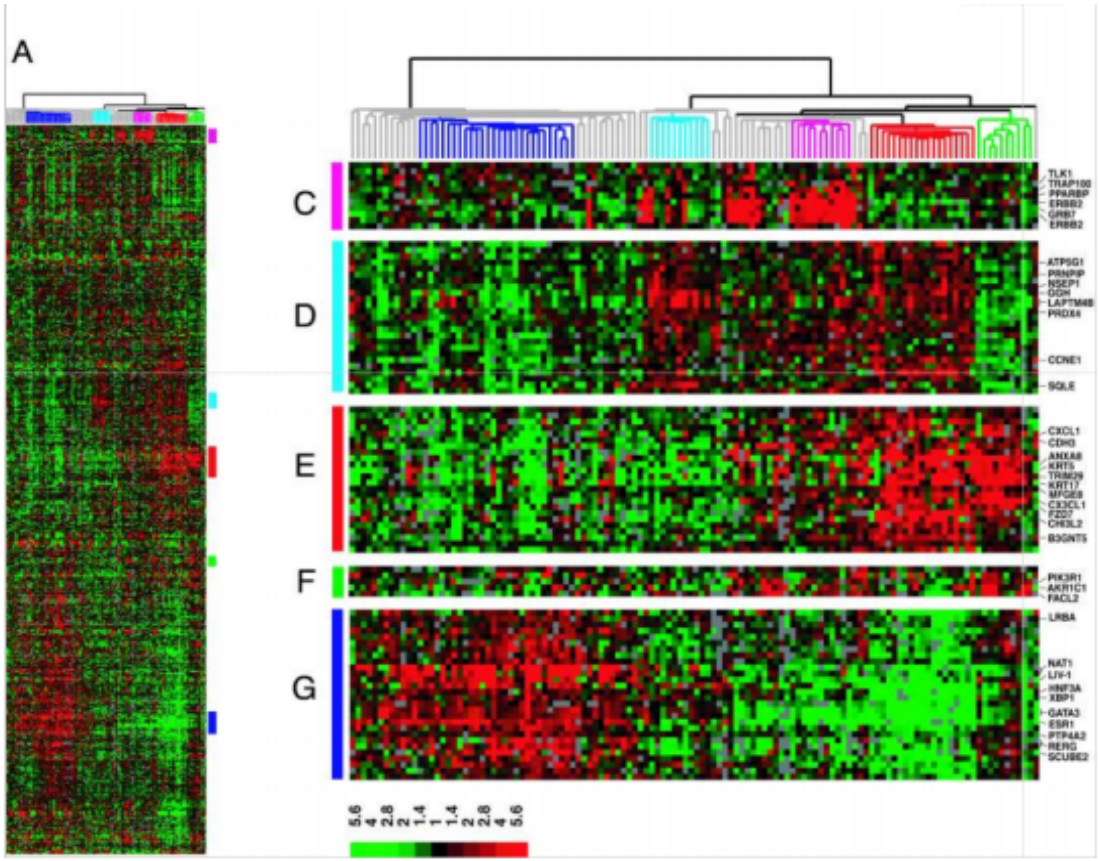
Reconocimiento de dígitos



- Hastie, et al. 2013 [8]

¿Qué es Machine Learning?

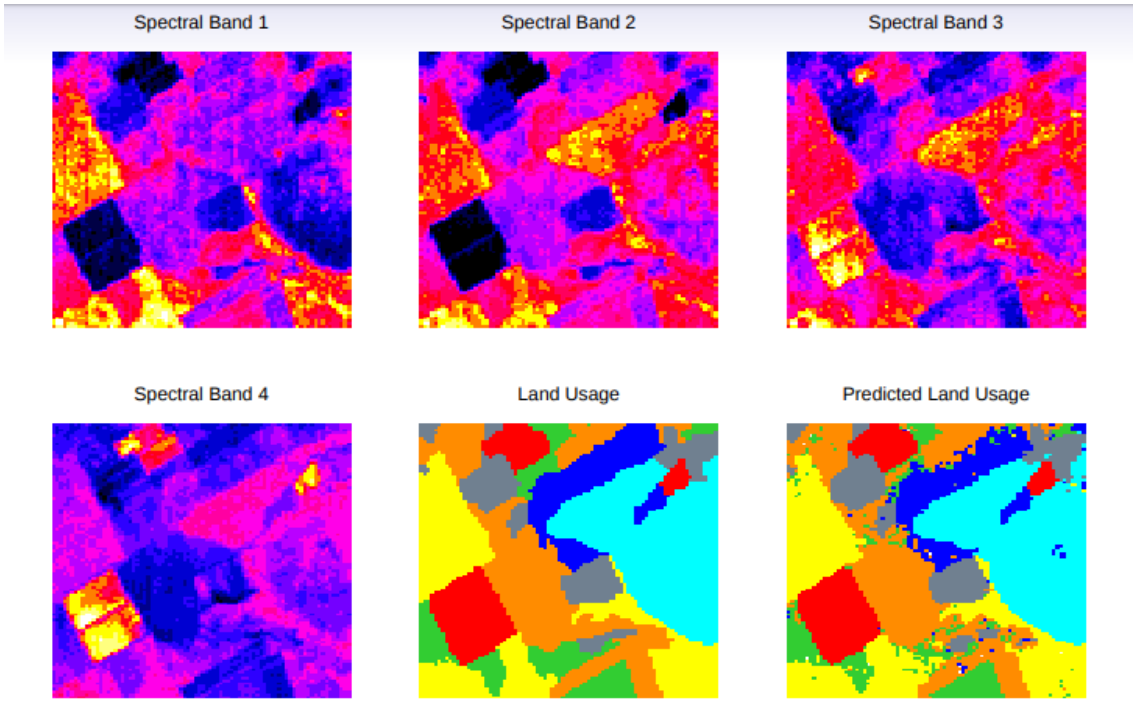
Cáncer y genes



- Hastie, et al. 2013 [8]

¿Qué es Machine Learning?

Imágenes satelitales

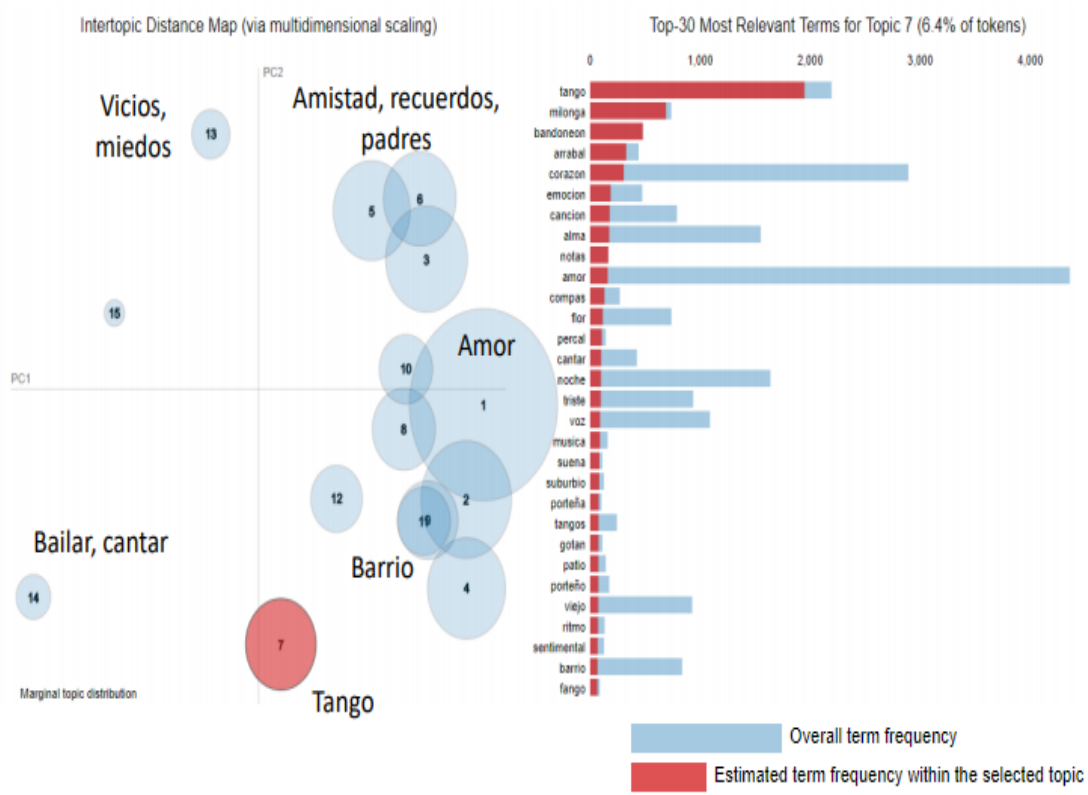


$Usage \in \{red\ soil,\ cotton,\ vegetation\ stubble,\ mixture,\ gray\ soil,\ damp\ gray\ soil\}$

- Hastie, et al. 2013 [8]

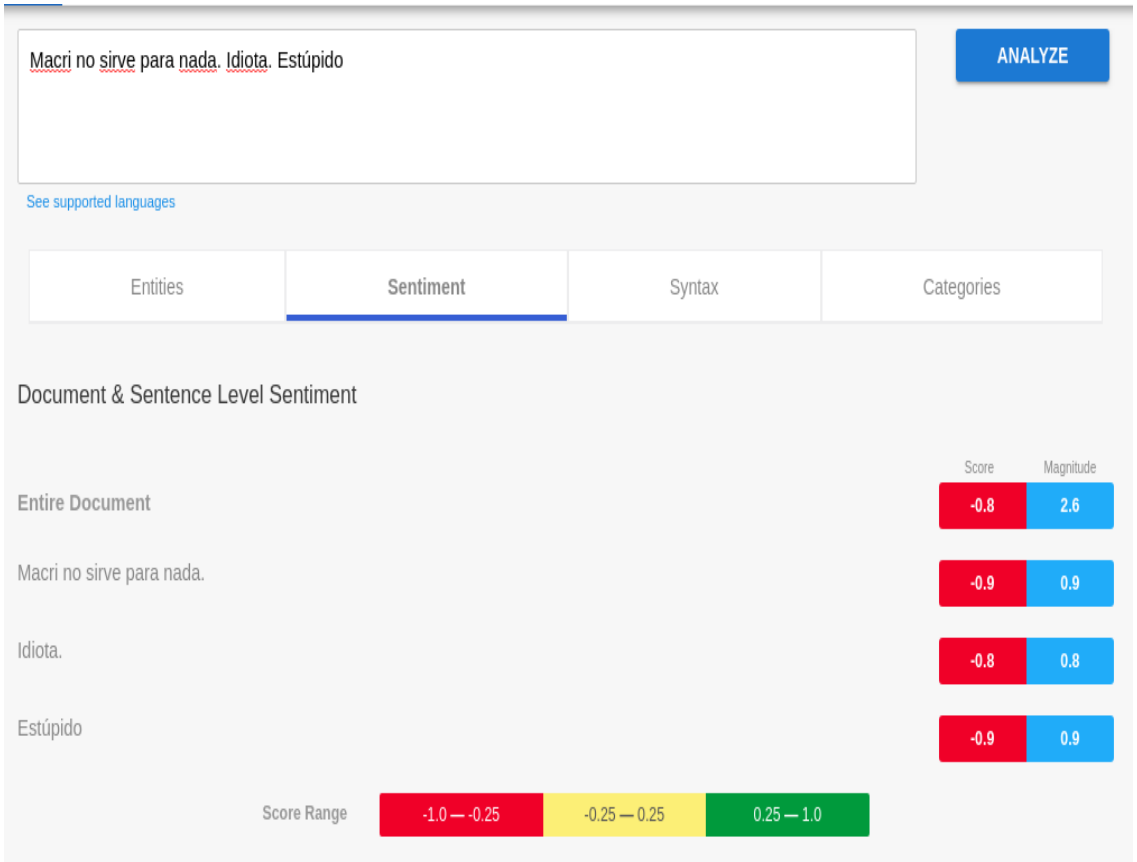
¿Qué es Machine Learning?

Detección de tópicos -tango-



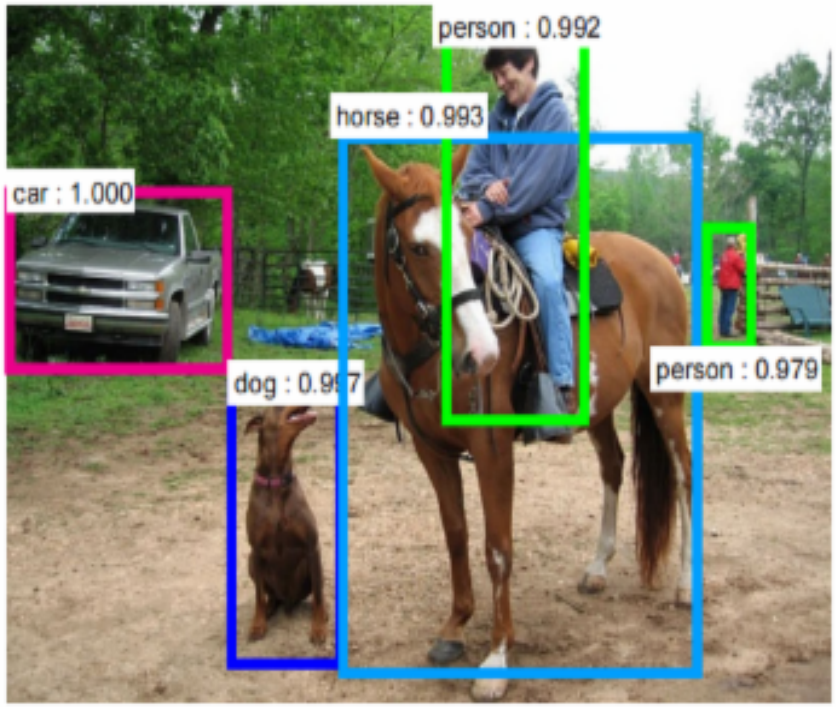
¿Qué es Machine Learning?

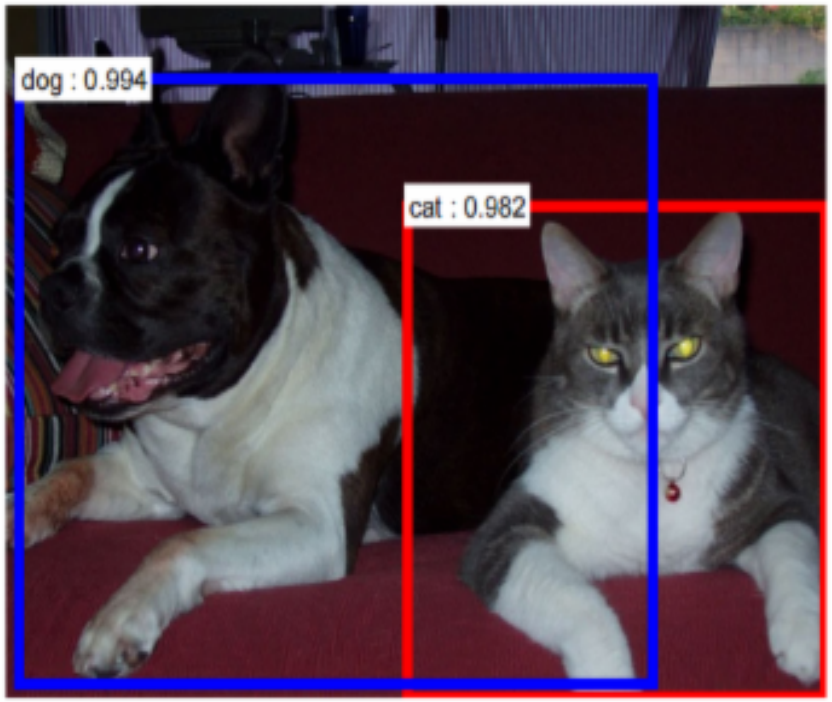
Sentiment Analysis



¿Qué es Machine Learning?

Detección de objetos en imágenes





Delimitación del campo

Chollet, 2017 [4]

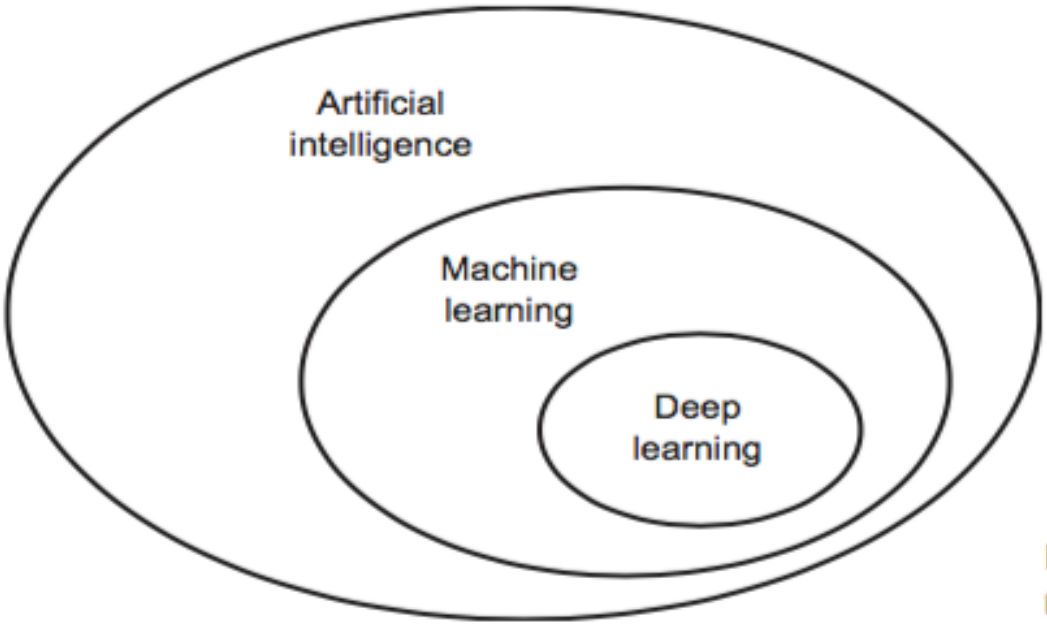
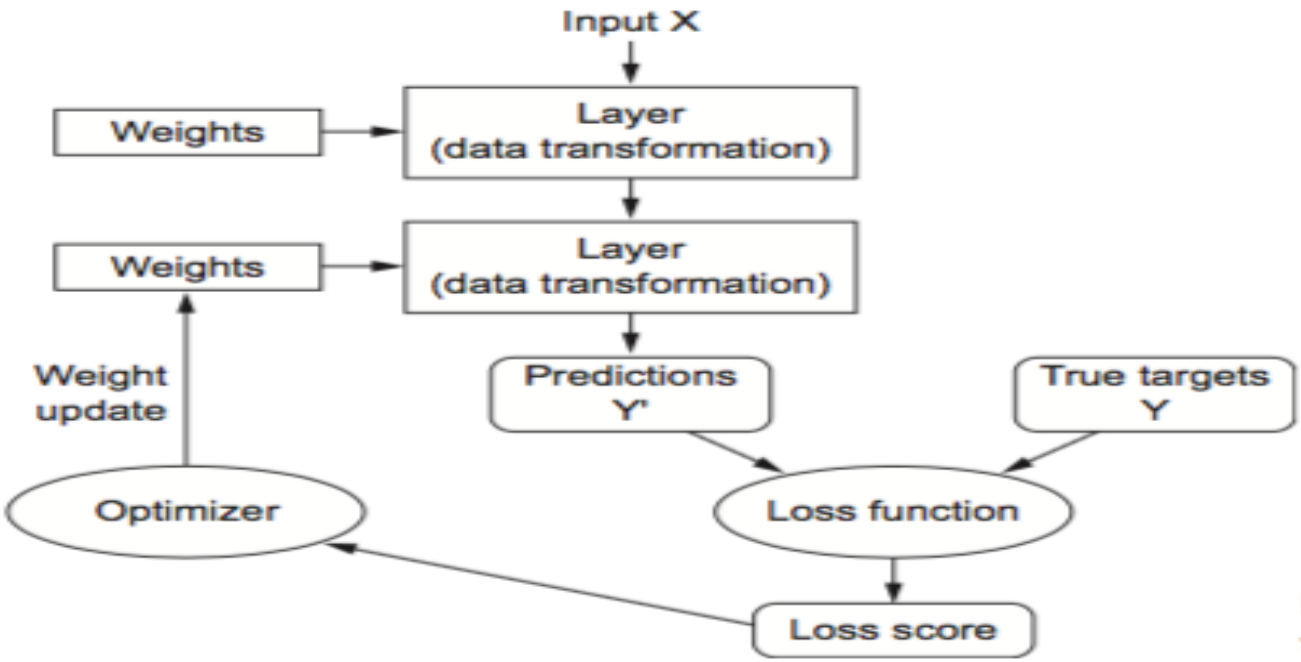


Figure 1.1 Artificial intelligence, machine learning, and deep learning

Delimitación del campo

Chollet, 2017, [4]



Tipos de problemas

Aprendizaje Supervisado:

- Variable dependiente, resultado, target Y
- Matriz de p predictores X, featuras, variables independientes, etc...
- En *problemas de regresión* Y es cuantitativa
- En *problemas de clasficación* Y es cualitativa
- Tenemos datos de entremiento $(X_1, y_1, ..., (X_N, y_1)$. Son observaciones (ejemplos, instancias) de las mediciones

Tipos de problemas

Aprendizaje No Supervisado:

- No hay una variable target... no existe Y
- Sí existe la matriz de p predictores X, featuras, variables independientes, etc...
- Es más complejo: es más difícil medir qué tan bien funciona el modelo

¿Qué es un modelo?

- Básicamente: una manera de proponer hipótesis sobre la forma en que se combinan variables
- En general, vamos a estar tratando de generar modelos de esta forma

$$Y = f(X) + \epsilon \tag{1}$$

- Todo el problema es estimar $f(X)$, es decir, de qué forma(s) se combinan las X para generar un output
- Una posibilidad es suponer que Y es una combinación lineal de las X

¿Qué es un modelo?

Las dos culturas (Breiman, 2001) [3]

"Todos los modelos son equivocados. Algunos son útiles."
George Box

- El mundo como productor de outputs -y- en base a features -X-
- Problemas: ¿cuál es la manera en que el mundo produce resultados?
- Una forma común es asumir que los datos son generados por extracciones independientes de
  $output = f(predictores, ruido, parametros)$

¿Qué es un modelo?

Las dos culturas (Breiman, 2001)[3]

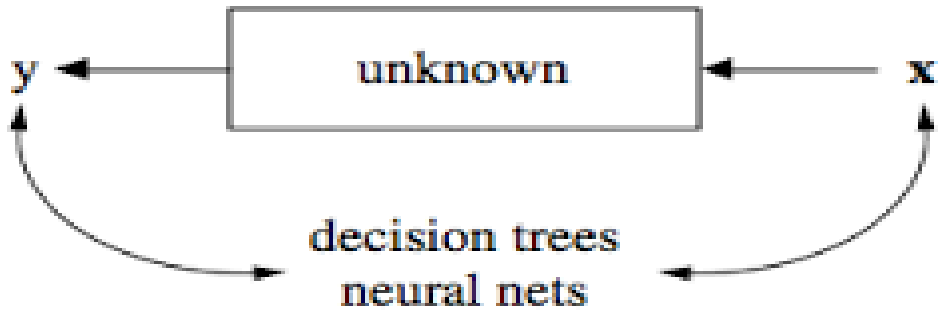
Modelado estadístico

- Énfasis en $f(x)$. El modelo se postula en base a supuestos sobre $f(x)$
- Conocimiento acumulado, teoría, diseño de experimentos
- Los parámetros son estimados con los datos y luego se realizan las predicciones.
- Evaluación del modelo: estimadores insesgados, robustos, mínima varianza

¿Qué es un modelo?

Las dos culturas (Breiman, 2001)[3]

Modelado algorítmico (o *Machine Learning*, *Data Mining*, etc.)



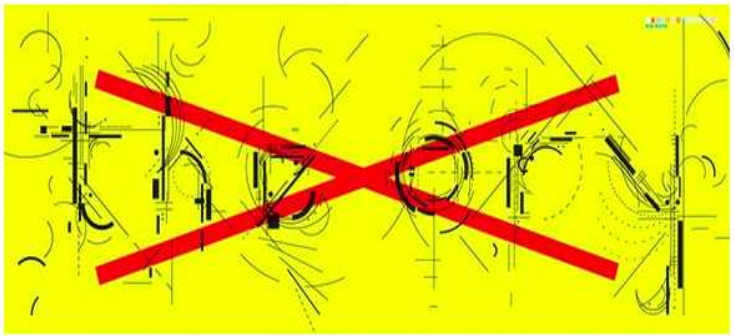
- Énfasis en $\hat{y}$
- El enfoque es encontrar una función $f(x)$ -un algoritmo- que opera sobre las x para predecir las y.
- El modelo se "aprende"de los datos
- Evaluación del modelo: performance predictiva

¿Qué es un modelo?

El fin de la teoría (Anderson, 2008)[1]

CHRIS ANDERSON SCIENCE 06.23.08 12:00 PM

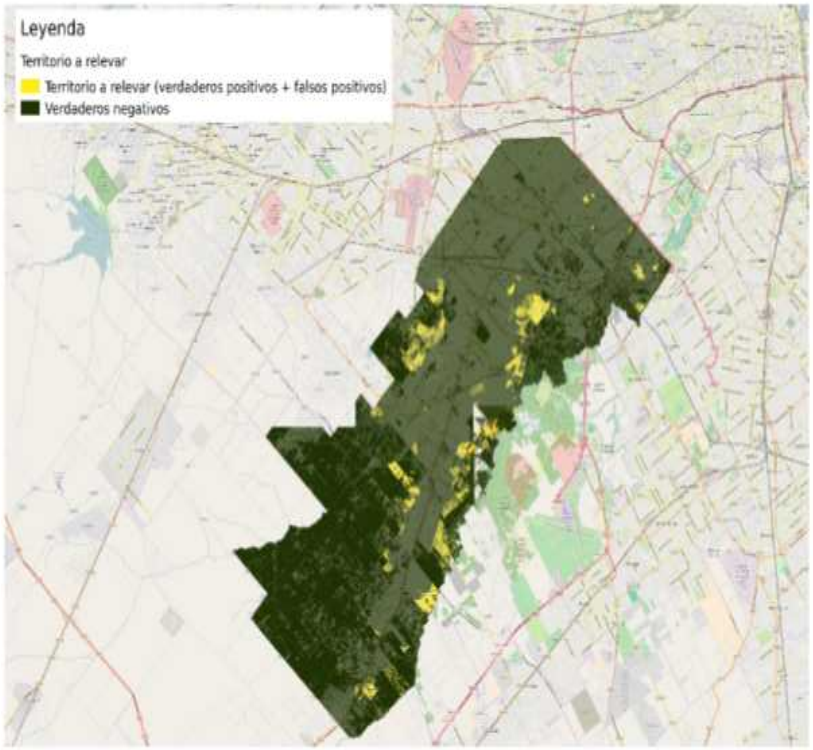
THE END OF THEORY: THE DATA DELUGE MAKES THE SCIENTIFIC METHOD OBSOLETE



Out with every theory of human behavior, from linguistics to sociology. Forget taxonomy, ontology, and psychology. Who knows why people do what they do? The point is they do it, and we can track and measure it with unprecedented fidelity. With enough data, the numbers speak for themselves.

Aplicaciones de ML en Ciencias Sociales

Detección de asentamientos informales en base a imágenes satelitales

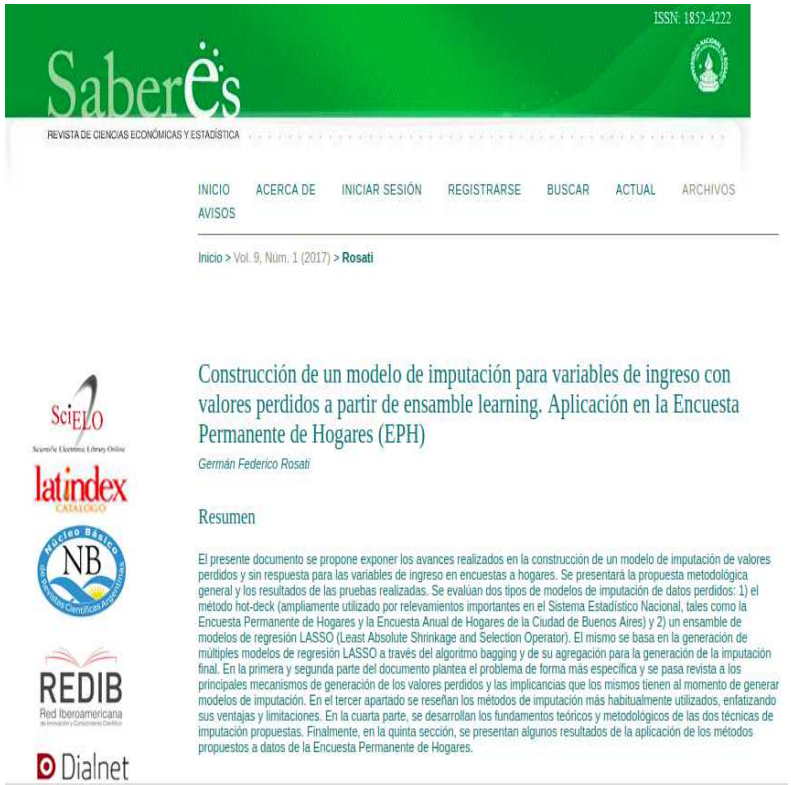


- Baylé, 2016 [2]

Aplicaciones de ML en Ciencias Sociales

Imputación de datos perdidos en encuestas sociomodemográficas

- Rosati, 2017 [10]



Aplicaciones de ML en Ciencias Sociales

Integración de comunidades migrantes

- Lamanna,Lenormand, et al 2016 [9]

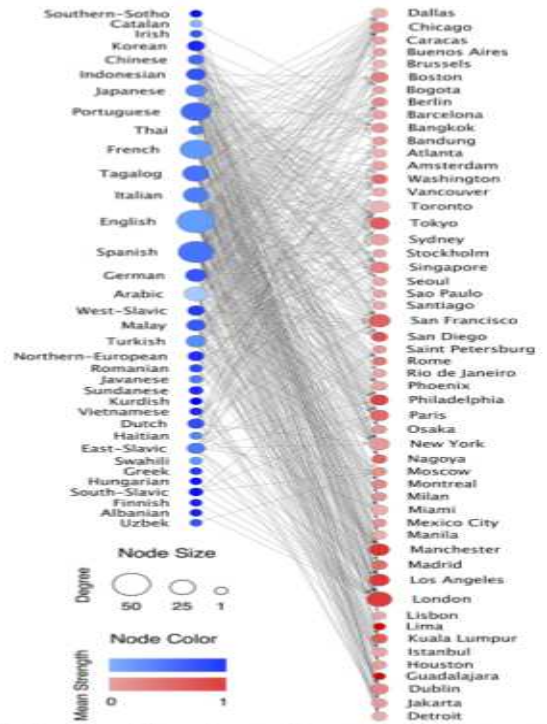


Fig 2. Bipartite spatial integration network. The network comprises of two sets: **L** of Languages and **C** of cities; the languages detected are connected to the cities set where the corresponding community of immigrants has been found. The weight of the edge corresponds to the values of k_{lc}. The size of the nodes is proportional to its degree and the color to its mean strength.

Aplicaciones de ML en Ciencias Sociales

Predicción de enfermedades mentales mediante text mining

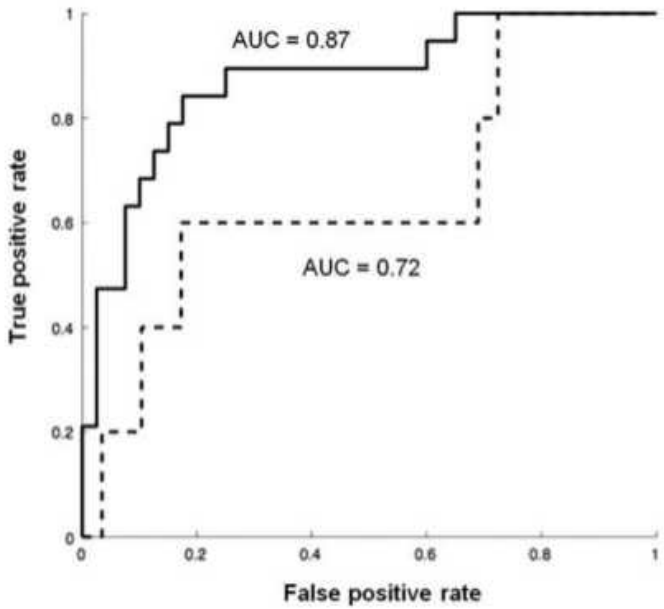


Figure 2 Receiver operating characteristics (ROC) for the University of California Los Angeles (UCLA) clinical high-risk (CHR) classifier of psychosis outcome as applied to the UCLA dataset (solid line) and to the realigned New York City (NYC) dataset (dotted line). AUC – area under the curve.

- Corcoran, Carrillo, Fernández Slezak et al, 2018 [5]

Aplicaciones de ML en Ciencias Sociales

Automatización de procesos para la construcción de bases de datos de protestas

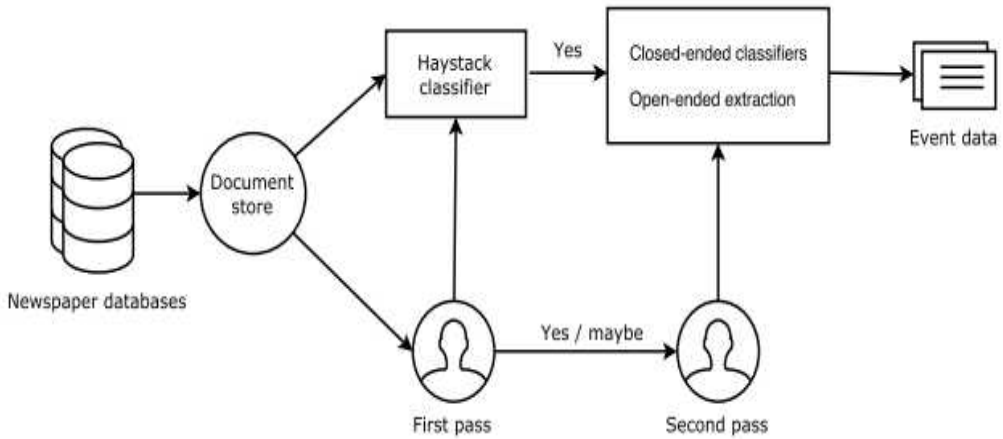


Figure 1: MPEDS pipeline with training.

- Hannah, 2017 [7]

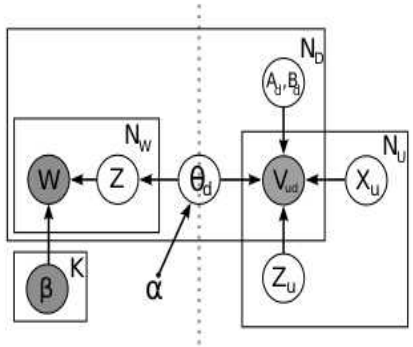
Aplicaciones de ML en Ciencias Sociales

Posiciones ideológicas en proyectos de ley

- Gerrish y Blei, 2015 [6]

Terrorism	Commemorations	Transportation
terrorist	nation	transportation
september	people	minor
attack	life	print
nation	world	tax
york	serve	land
terrorist attack	percent	guard
hezbolah	community	coast guard
national guard	family	substitute

Labeled topics

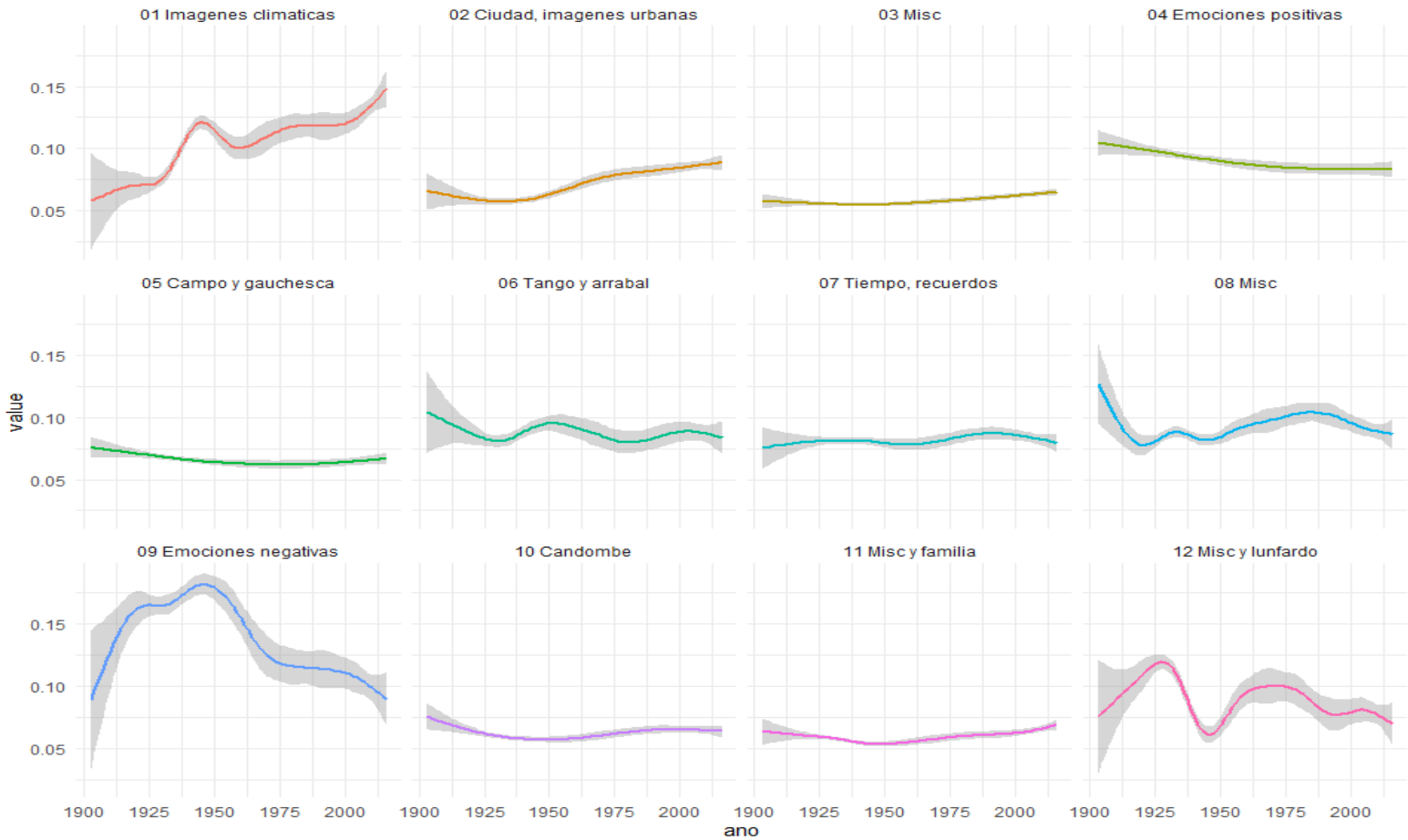


The issue-adjusted ideal point model

Figure 3: Left: Top words from topics fit using labeled LDA [6]. Right: the issue-adjusted ideal point model, which models votes v_{ud} from lawmakers and legislative items. Classic item response theory models votes v using x_u and a_d, b_d. For our work, documents' issue vectors $\boldsymbol{\theta}$ were estimated fit with a topic model (left of dashed line) using bills' words w and labeled topics β. Expected issue vectors $\mathbb{E}_q[\boldsymbol{\theta}|\boldsymbol{w}]$ are then treated as constants in the issue model (right of dashed line).

Aplicaciones de ML en Ciencias Sociales

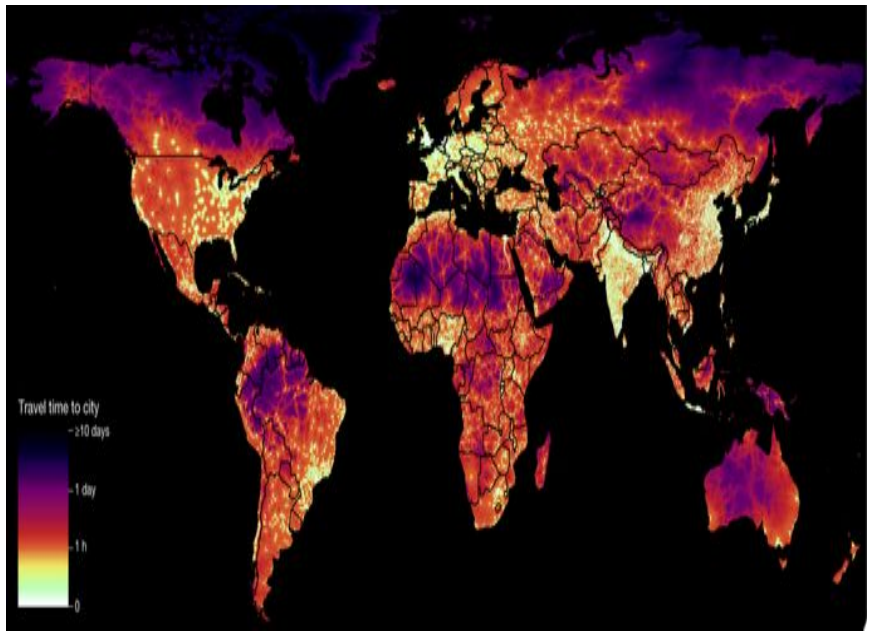
Tópicos en letras de tango, Rosati, 2020, inédito



Pero antes..

Algunos ejemplos de textos con problemas relevantes

Datos de accesibilidad como acercamiento a la desigualdad
Weiss et al 2018 [12]

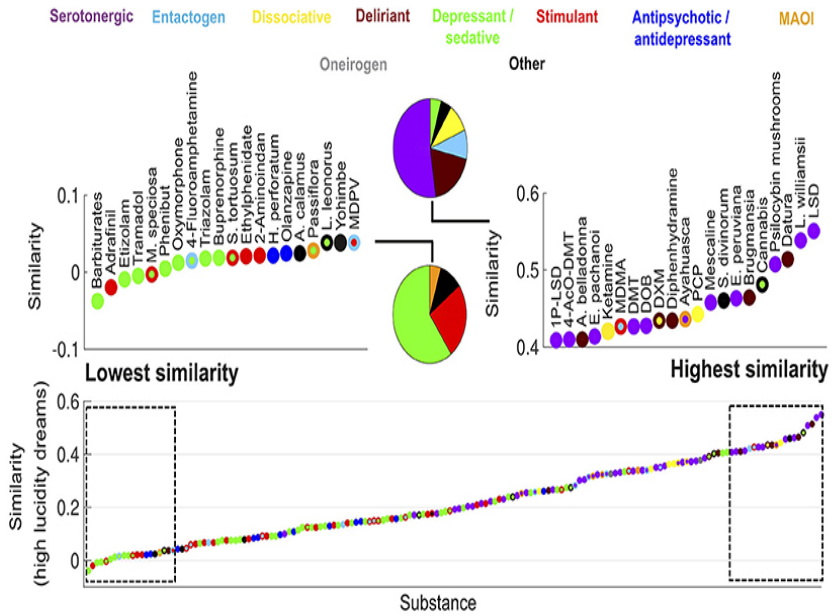


Pero antes..

Algunos ejemplos de textos con problemas relevantes

Relatos de experiencias de ”viajes”para estimar efectos de diferentes sustancias alucinógenas

Sanz et al 2018 [11]



Referencias bibliográficas I

Anderson, C.
The End of Theory: The Data Deluge Makes the Scientific Method Obsolete.
Wired 16, 07 (June 2008).

Baylé, F.
Detección de villas y asentamientos informales en el partido de la matanza mediante teledetección y sistemas de información geográfica.
Master's thesis, Tesis de Maestría, 2016.

Breiman, L.
Statistical modeling: The two cultures (with comments and a rejoinder by the author).
Statistical Science 16, 3 (08 2001), 199–231.

Chollet, F.
Deep Learning with Python.
Manning, Nov. 2017.

Corcoran, C. M., Carrillo, F., Fernández-Slezak, D., Bedi, G., Klim, C., Javitt, D., Bearden, C., and Cecchi, G.
Prediction of psychosis across protocols and risk cohorts using automated language analysis.
World Psychiatry 17, 01 (2 2018).

Referencias bibliográficas II


Gerrish, S., and Blei, D. M.
How they vote: Issue-adjusted models of legislative behavior.
In *NIPS* (2012), P. L. Bartlett, F. C. N. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds., pp. 2762–2770.


Hanna, A.
Mpeds: Automating the generation of protest event data, Jan 2017.


Hastie, T., Tibshirani, R., and Friedman, J.
The Elements of Statistical Learning.
Springer Series in Statistics. Springer New York Inc., New York, NY, USA, 2001.


Lamanna, F., Lenormand, M., Salas-Olmedo, M. H., Romanillos, G., Gonçalves, B., and Ramasco, J. J.
Immigrant community integration in world cities.
CoRR abs/1611.01056 (2016).


Rosati, G.
Construcción de un modelo de imputación para variables de ingreso con valores perdidos a partir de Ensamble Learning. Aplicación a la Encuesta Permanente de Hogares (EPH).
SaberES. Revista de Ciencias Económicas y Estadística 9, 01 (Febrero 2017).

Referencias bibliográficas III

Sanz, C., Zamberlan, F., Erowid, E., Erowid, F., and Tagliazucchi, E.
The experience elicited by hallucinogens presents the highest similarity to dreaming within a large database of psychoactive substance reports.
Frontiers in Neuroscience 12 (2018), 7.

Weiss, D. J., Nelson, A., Gibson, H. S., Temperley, W., Peedell, S., Lieber, A., Hancher, M., Poyart, E., Belchior, S., Fullman, N., Mappin, B., Dalrymple, U., Rozier, J., Lucas, T. C. D., Howes, R. E., Tusting, L. S.and Kang, S. Y., Cameron, Bisanzio, D., Battle, K. E., Bhatt, S., and Gething, P. W.
A global map of travel time to cities to assess inequalities in accessibility in 2015.
Nature 553, 333 (01 2018).