

# **M5. Minería de Texto + webscraping**

## **Clase 6. Embeddings y clasificación de texto**



# Hipótesis distribucional

- “El significado deriva del uso de las palabras en el lenguaje” (Wittgenstein)
- “Conocerás a una palabra por su compañía” (Firth)
- Palabras cercanas tienen sentidos “cercanos”
- Ítems lingüísticos con distribuciones similares tienen significados similares”
- Idea de co-ocurrencia => términos que ocurren juntos



# Aplicaciones en Ciencias Sociales

npj Schizophrenia

www.nature.com/npjSch  
All rights reserved 2334-265X/15

ARTICLE OPEN

## Automated analysis of free speech predicts psychosis onset in high-risk youths

Gillinder Bedi<sup>1,2,9</sup>, Facundo Carrillo<sup>3,9</sup>, Guillermo A Cecchi<sup>4</sup>, Diego Fernández Slezak<sup>3</sup>, Mariano Sigman<sup>5</sup>, Natália B Mota<sup>6</sup>, Sidarta Ribeiro<sup>6</sup>, Daniel C Javitt<sup>1,7</sup>, Mauro Copelli<sup>8</sup> and Cheryl M Corcoran<sup>1,7</sup>

**BACKGROUND/OBJECTIVES:** Psychiatry lacks the objective clinical tests routinely used in other specializations. Novel computerized methods to characterize complex behaviors such as speech could be used to identify and predict psychiatric illness in individuals.

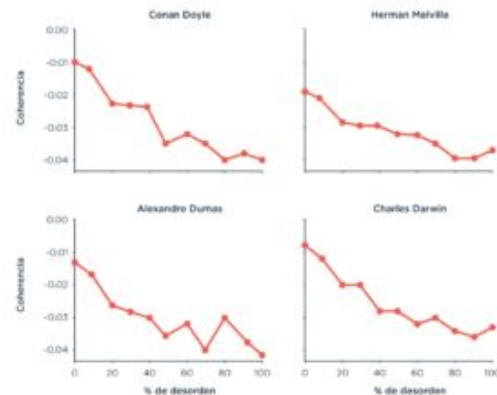
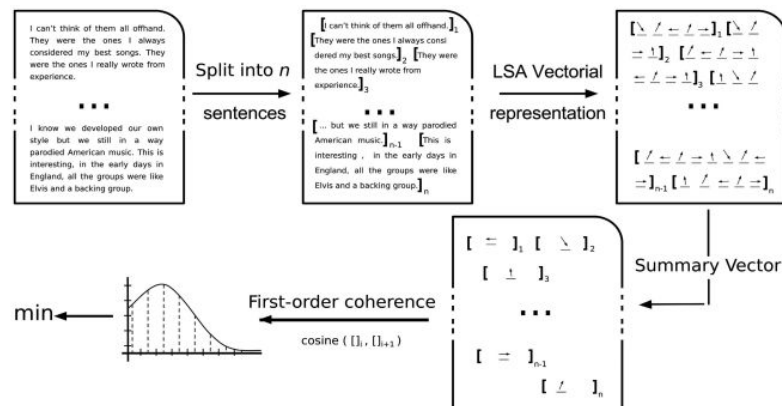
**AIMS:** In this proof-of-principle study, our aim was to test automated speech analyses combined with Machine Learning to predict later psychosis onset in youths at clinical high-risk (CHR) for psychosis.

**METHODS:** Thirty-four CHR youths (11 females) had baseline interviews and were assessed quarterly for up to 2.5 years; five transitioned to psychosis. Using automated analysis, transcripts of interviews were evaluated for semantic and syntactic features predicting later psychosis onset. Speech features were fed into a convex hull classification algorithm with leave-one-subject-out cross-validation to assess their predictive value for psychosis outcome. The canonical correlation between the speech features and prodromal symptom ratings was computed.

**RESULTS:** Derived speech features included a Latent Semantic Analysis measure of semantic coherence and two syntactic markers of speech complexity: maximum phrase length and use of determiners (e.g., *which*). These speech features predicted later psychosis development with 100% accuracy, outperforming classification from clinical interviews. Speech features were significantly correlated with prodromal symptoms.

**CONCLUSIONS:** Findings support the utility of automated speech analysis to measure subtle, clinically relevant mental state changes in emergent psychosis. Recent developments in computer science, including natural language processing, could provide the foundation for future development of objective clinical tests for psychiatry.



npj Schizophrenia (2015) 1, Article number: 15030; doi:10.1038/npjSch.2015.30; published online 26 August 2015



factor-data  
EIDAEs\_UNSAM

# Aplicaciones en Ciencias Sociales - Estereotipos

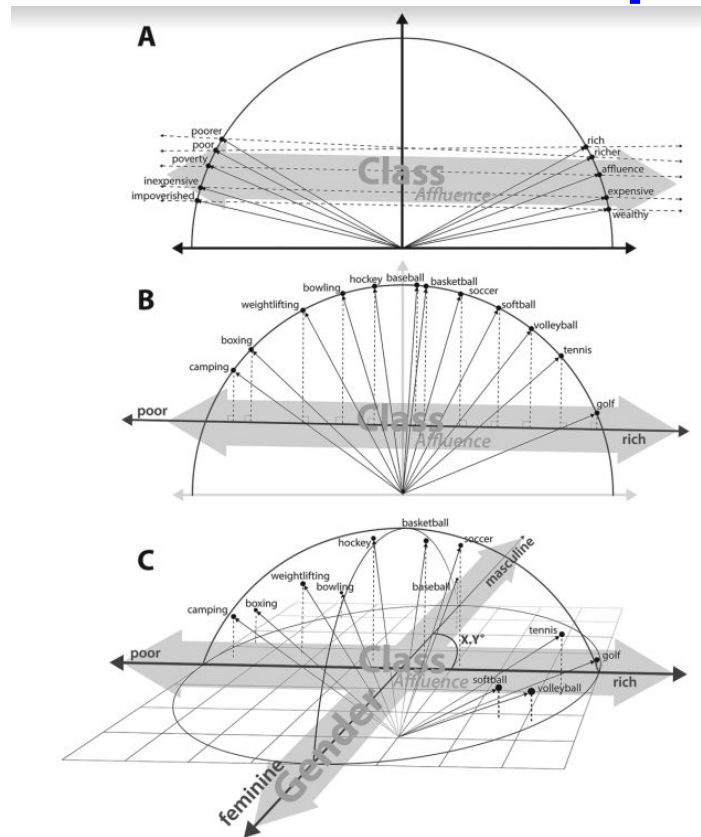
## The Geometry of Culture: Analyzing the Meanings of Class through Word Embeddings

Austin C. Kozlowski,<sup>a</sup>  Matt Taddy,<sup>b</sup>  
and James A. Evans<sup>a,c</sup> 

American Sociological Review  
2019, Vol. 84(5) 905–949  
© American Sociological  
Association 2019  
DOI: 10.1177/0003122419877135  
[journals.sagepub.com/home/asr](https://journals.sagepub.com/home/asr)

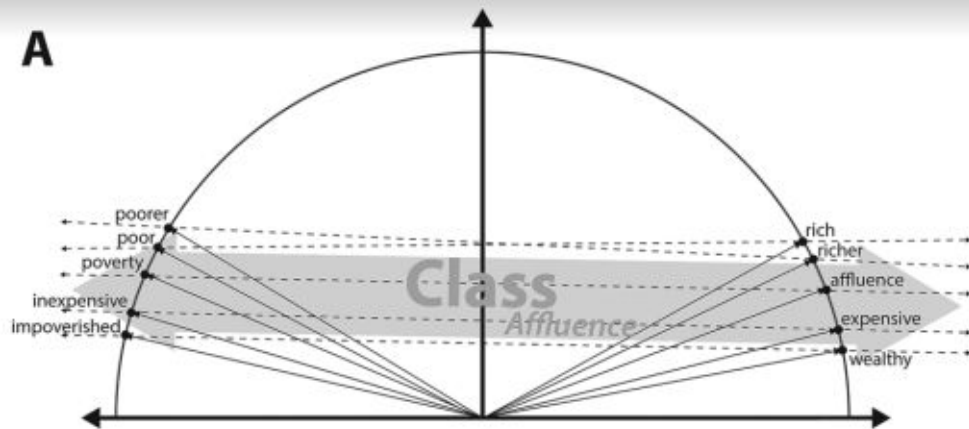


factor-data  
EIDAES\_UNSAM



**Figure 2.** Conceptual Diagram of (A) the Construction of a Cultural Dimension; (B) the Projection of Words onto That Dimension; and (C) the Simultaneous Projection of Words onto Multiple Dimensions

# Aplicaciones en Ciencias Sociales - Estereotipos



## *Measuring Cultural Dimensions*

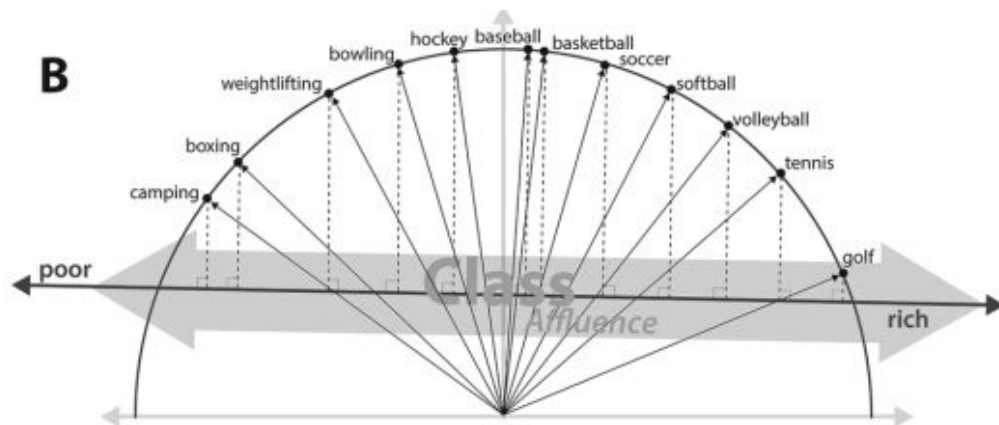
To identify cultural dimensions in word embedding models, we average numerous pairs of antonym words. Cultural dimensions are calculated by simply taking the mean of all word pair differences that approximate a

given dimension,  $\frac{\sum_p |\vec{p}_1 - \vec{p}_2|}{|P|}$ , where  $p$  are

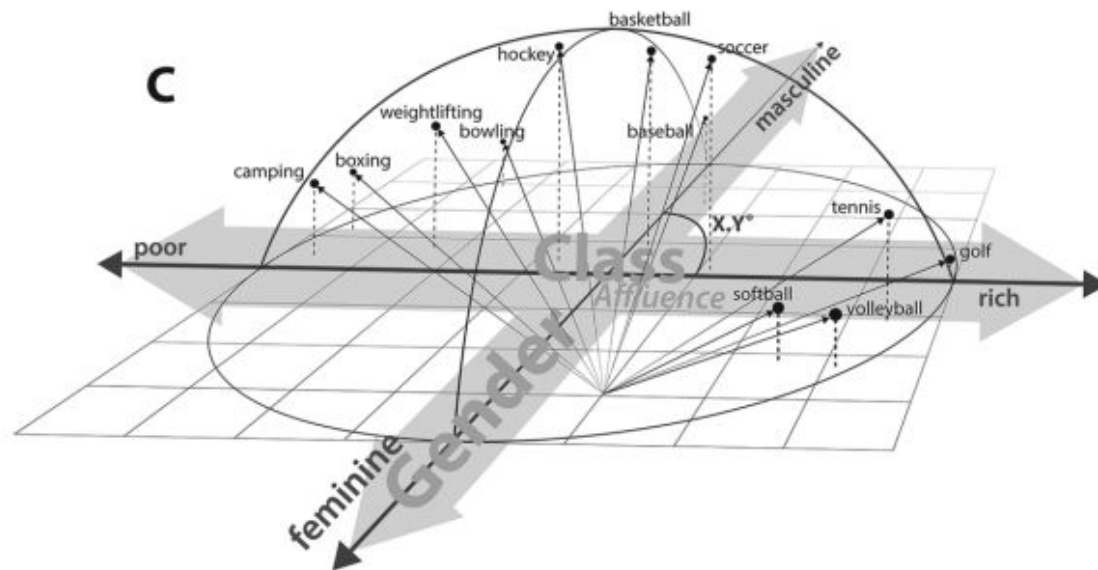
all antonym word pairs in relevant set  $P$ , and  $\vec{p}_1$  and  $\vec{p}_2$  are the first and second word vectors of each pair.<sup>17</sup> The projection of a normalized word vector onto a cultural dimension is calculated with cosine similarity, as is the angle between cultural dimensions.



# Aplicaciones en Ciencias Sociales - Estereotipos

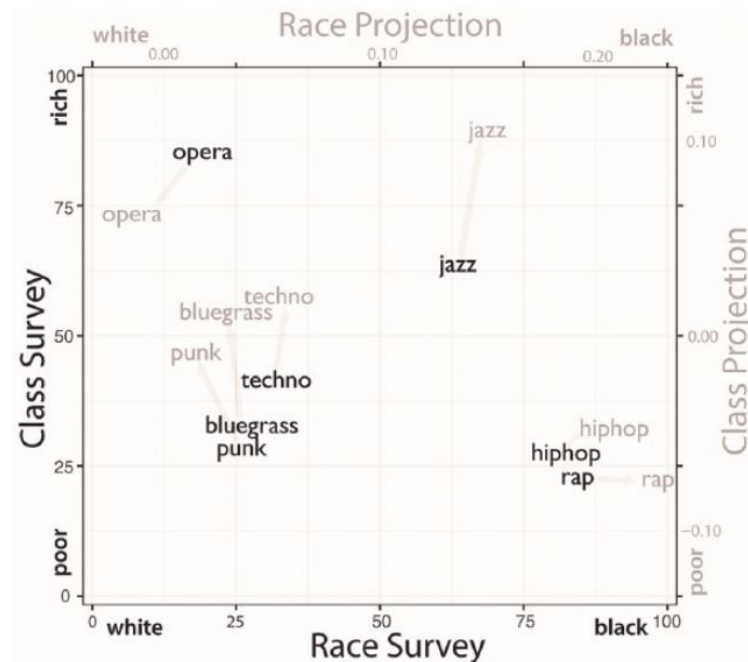


# Aplicaciones en Ciencias Sociales - Estereotipos



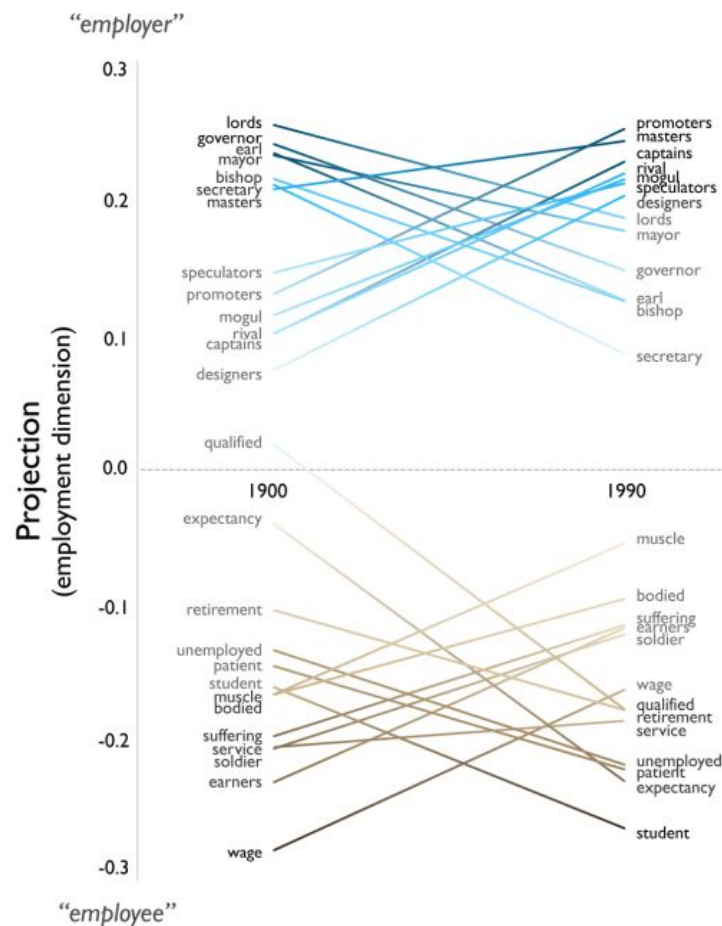
**Figure 2.** Conceptual Diagram of (A) the Construction of a Cultural Dimension; (B) the Projection of Words onto That Dimension; and (C) the Simultaneous Projection of Words onto Multiple Dimensions

# Aplicaciones en Ciencias Sociales - Estereotipos



**Figure 3.** Projection of Music Genres onto Race and Class Dimensions of the Google News Word Embedding (Gray) and Average Survey Ratings for Race and Class Associations (Black)



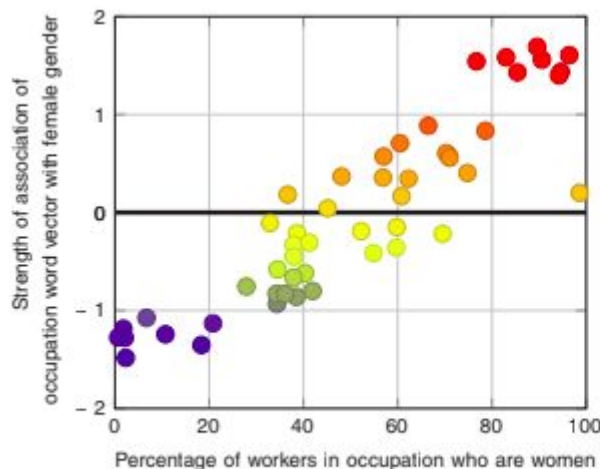


**Figure 10.** Words That Project High and Low on the Employment Dimension of Word Embedding Models Trained on Texts Published at the Beginning and End of the Twentieth Century; 1900–1919 and 1980–1999 Google Ngrams Corpus

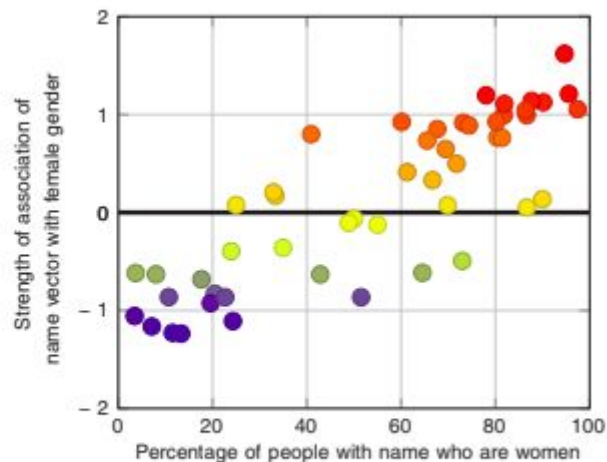
# Aplicaciones en Ciencias Sociales - Estereotipos

Semantics derived automatically  
from language corpora contain  
human-like biases

Aylin Caliskan,<sup>1\*</sup> Joanna J. Bryson,<sup>1,2\*</sup> Arvind Narayanan<sup>1\*</sup>



**Fig. 1. Occupation-gender association.** Pearson's correlation coefficient  $\rho = 0.90$  with  $P < 10^{-18}$ .



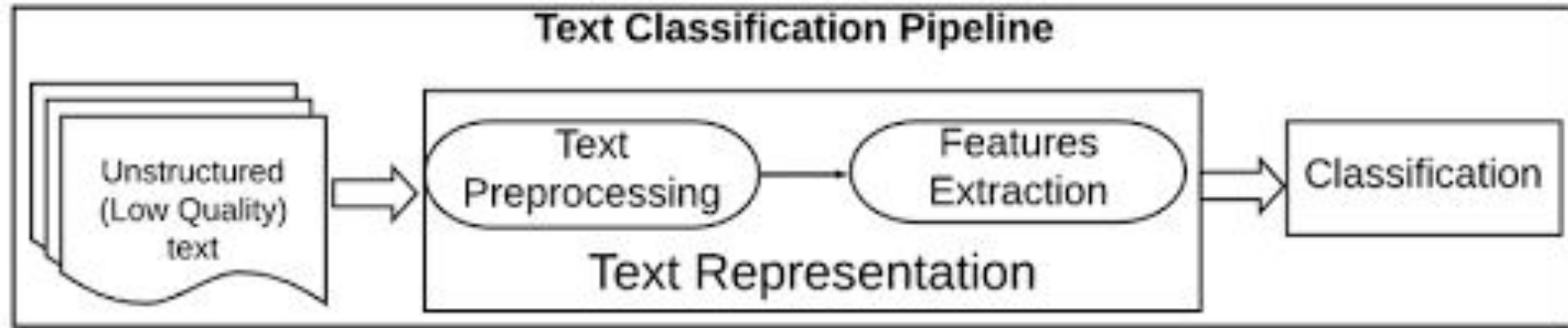
**Fig. 2. Name-gender association.** Pearson's correlation coefficient  $\rho = 0.84$  with  $P < 10^{-13}$ .



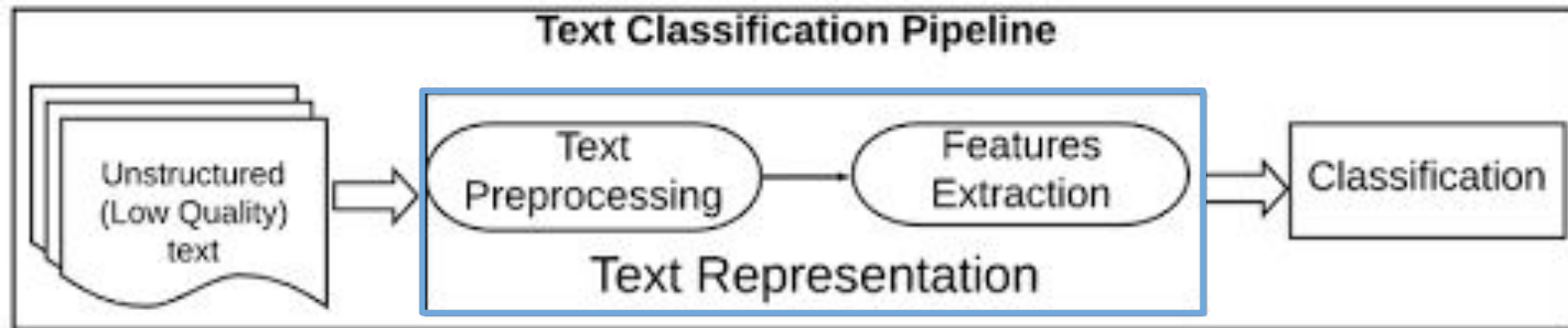
# Clasificación de texto



# Clasificación de texto

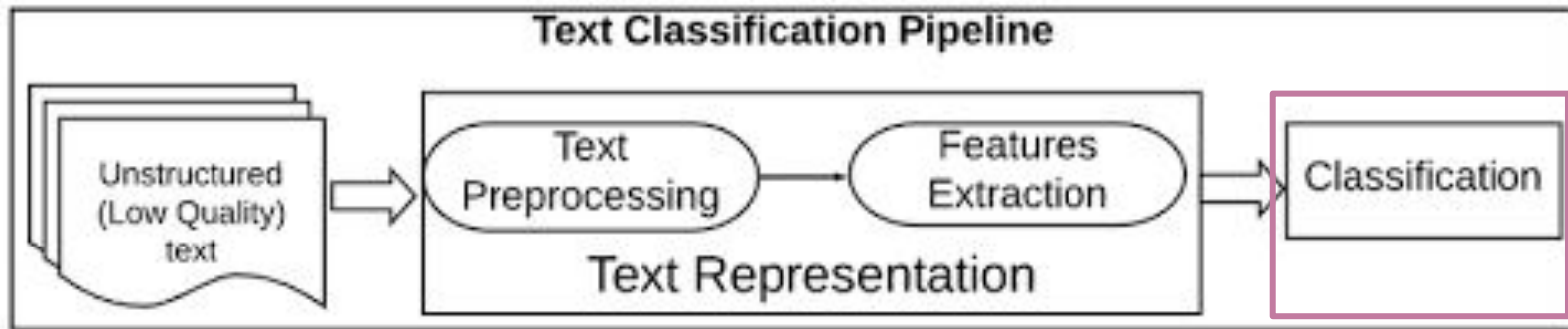


# Clasificación de texto



- Dos formas (que vimos):
  - Bag of Words => Term Frequency Matrix
  - Word Embeddings => Dense word Matrix

# Clasificación de texto



- Cualquiera de los métodos supervisados que vimos:
  - Regresión (¿lineal o logística?)
  - Random Forest
  - Gradient Boosting ...
- O que no vimos... RNN, LSTM, Transformers, etc...

# Vamos al notebook...



# Algunas cuestiones para cerrar...

- NLP “Del giro lingüístico al giro (lingüístico) computacional”.
- Posibilidades metodológicas para las ciencias sociales
- Discusiones “no metodológicas” que suscitan

Innatismo o no del lenguaje (discusión con Chomsky)

“Los límites de mi mundo son los de mi lenguaje”.

