

Machine Learning en las Ciencias Sociales

Algunas nociones y posibilidades de uso

Germán Rosati (CONICET-UNSAM, PIMSA)

Machine learning proliferates in particle physics

Illustration by Sandbox Studio, Chicago with Corinne Mucha

06/01/18 | By Manuel Gnida

A new review in *Nature* chronicles the many ways machine learning is popping up in particle physics research.

Experiments at the Large Hadron Collider produce about a million gigabytes of data every second. Even after reduction and compression, the data amassed in just one hour at the LHC is

nature.com > nature > technology features > article

MENU ▾

nature
International Journal of science

News & Comment Research

News Opinion Research Analysis Careers Books & Culture

TECHNOLOGY FEATURE • 20 FEBRUARY 2018 • CORRECTION 07 MARCH 2018

Deep learning for biology

A popular artificial-intelligence method provides a powerful tool for surveying and classifying biological data. But for the uninitiated, the technology poses significant difficulties.

nature.com > nature > news > article

a nature

MENU ▾

nature
International Journal of science



Search



Email

News & Comment Research

News Opinion Research Analysis Careers Books & Culture

NEWS • 28 MARCH 2018

Need to make a molecule? Ask this AI for instructions

Artificial-intelligence tool that has digested nearly every reaction ever performed could transform chemistry.

Hoja de ruta

- ¿Qué es Machine Learning?
- Machine Learning y Estadística
- Machine Learning en las Ciencias Sociales
- Los “usos” de Machine Learning en Ciencias Sociales

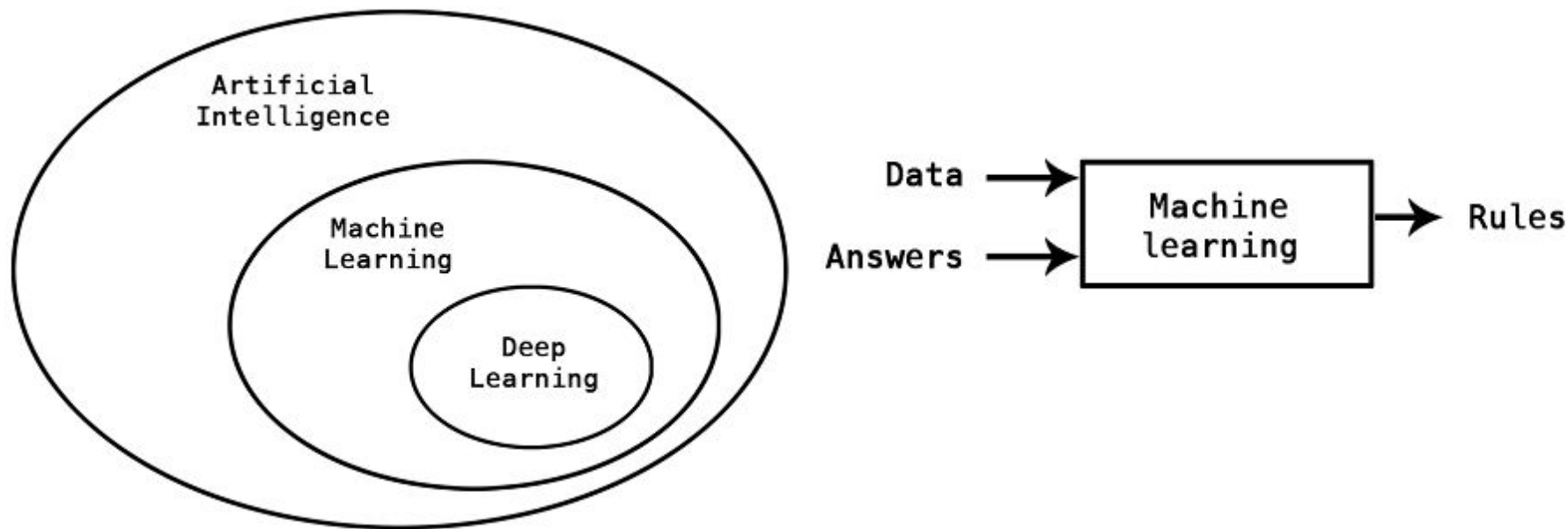
¿Qué es Machine Learning?

¿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 es Machine Learning?

[Chollet, 2018]



¿Qué es Machine Learning? - Elementos

- Elementos de un sistema de Machine Learning
 - **Conocimiento del problema**
 - Proceso de Generación de Datos
 - Algoritmos de “propósito general”

[Taddy, 2018]

¿Qué es Machine Learning? - Elementos

- Elementos de un sistema de Machine Learning
 - Conocimiento del problema
 - **Proceso de Generación de Datos**
 - Algoritmos de “propósito general”

[Taddy, 2018]

Procesos de Generación de Datos - Big Data



Dan Ariely ✓

6. Januar 2013 · 🌐

Big data is like teenage sex: everyone talks about it, nobody really knows how to do it, everyone thinks everyone else is doing it, so everyone claims they are doing it...

“Big data es como el sexo adolescente: todo el mundo habla de eso, nadie sabe realmente cómo hacerlo, todos piensan que el resto del mundo lo está haciendo, por lo tanto, todos afirman que lo están haciendo...”

Procesos de Generación de Datos - Big Data

- Dos sentidos (no necesariamente contrapuestos)
 - Nuevas fuentes de datos (redes sociales, logs, información mobile, etc.) disponibles en “tiempo real” y en gran volumen
 - Conjunto de tecnologías para recolectar, almacenar y procesar dichos grandes volúmenes (Spark, Hadoop, etc.)

Procesos de Generación de Datos - Big Data

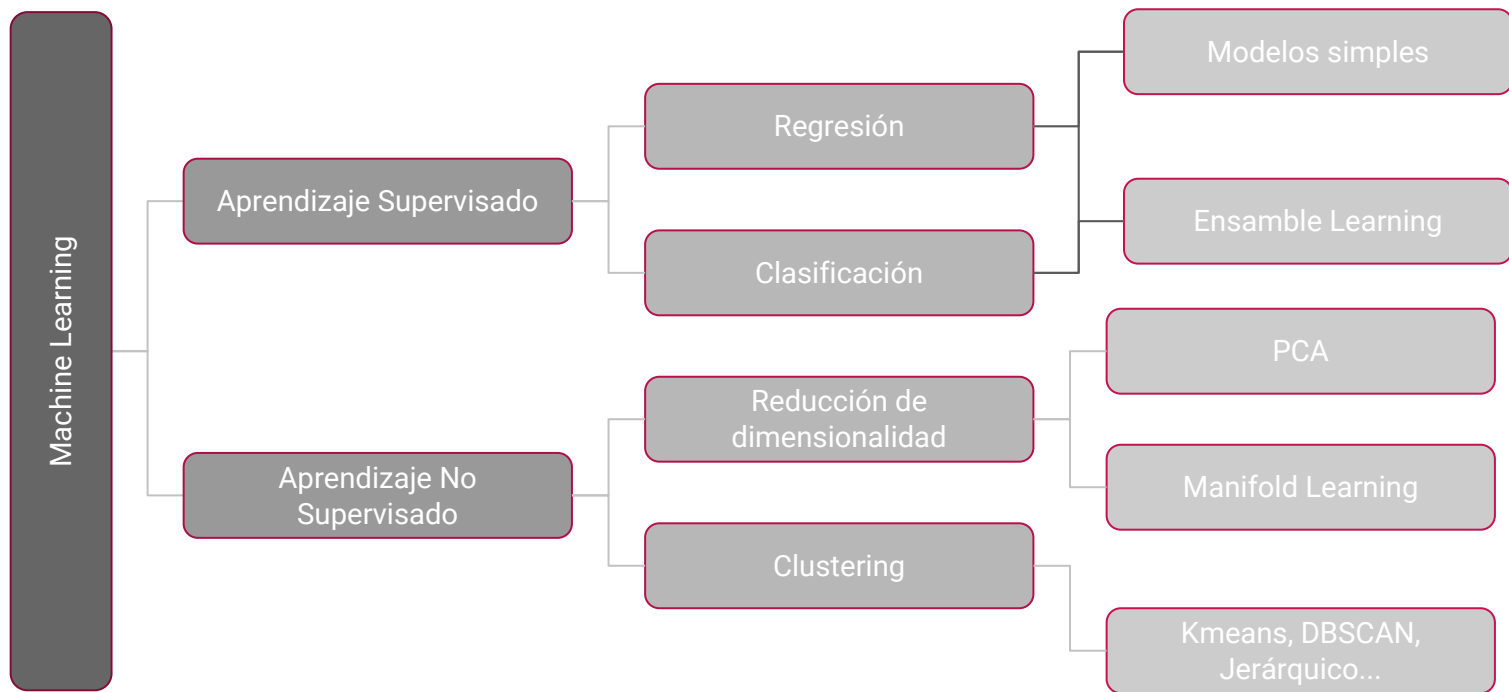
Small Data	Big Data
Estructurados	No estructurados
Volumen “mediano”	Volúmen grande
Muestras	Universo (?)
Pocas fuentes homogéneas	Muchas fuentes diversas
Proceso de producción largo	“Real Time”
Fuentes tradicionales	Datos relacionales, redes sociales, móviles
Costoso	“Marginalmente” barato

¿Qué es Machine Learning? - Elementos

- Elementos de un sistema de Machine Learning
 - Conocimiento del problema
 - Proceso de Generación de Datos
 - **Algoritmos de “propósito general”**

[Taddy, 2018]

¿Qué es Machine Learning? - Problemas



Machine Learning y Estadística

Statistical Science
2001, Vol. 16, No. 3, 199-231

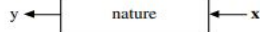
Statistical Modeling: The Two Cultures

Leo Breiman

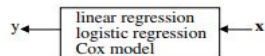
Abstract. There are two cultures in the use of statistical modeling to reach conclusions from data. One assumes that the data are generated by a given stochastic data model. The other uses algorithmic models and treats the data mechanism as unknown. The statistical community has been committed to the almost exclusive use of data models. This commitment has led to irrelevant theory, questionable conclusions, and has kept statisticians from working on a large range of interesting current problems. Algorithmic modeling, both in theory and practice, has developed rapidly in fields outside statistics. It can be used both on large complex data sets and as a more accurate and informative alternative to data modeling on smaller data sets. If our goal as a field is to use data to solve problems, then we need to move away from exclusive dependence on data models and adopt a more diverse set of tools.

1. INTRODUCTION

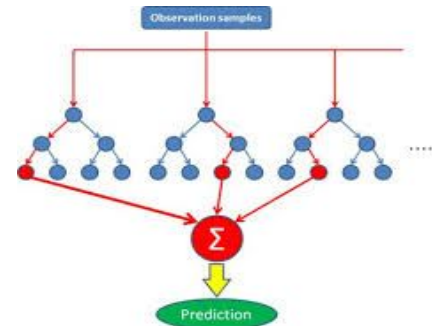
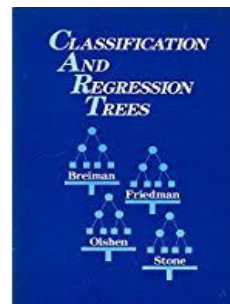
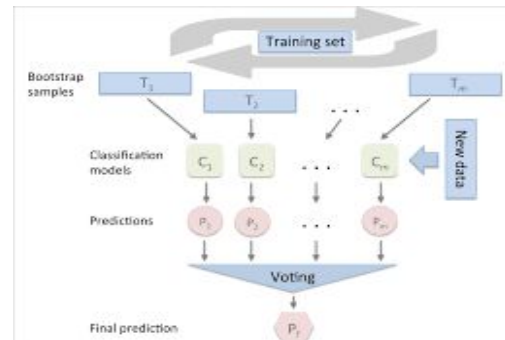
Statistics starts with data. Think of the data as being generated by a black box in which a vector of input variables \mathbf{x} (independent variables) go in one side, and on the other side the response variables \mathbf{y} come out. Inside the black box, nature functions to associate the predictor variables with the response variables, so the picture is like this:



The values of the parameters are estimated from the data and the model then used for information and/or prediction. Thus the black box is filled in like this:



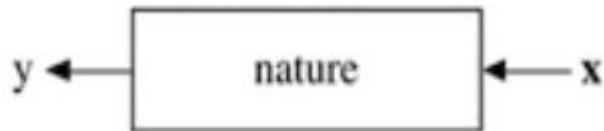
Model validation. Yes—no using goodness-of-fit tests and residual examination.
Estimated culture population. 98% of all statisticians.



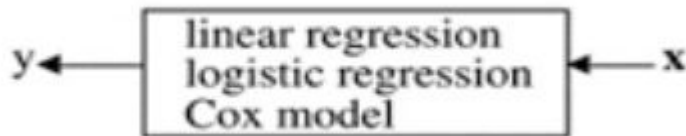
Machine Learning y Estadística

[Breiman, 2001]

- “El mundo como un productor de outputs en base a features”
- Caja negra
- ¿Cuál es la manera en que el mundo produce resultados?



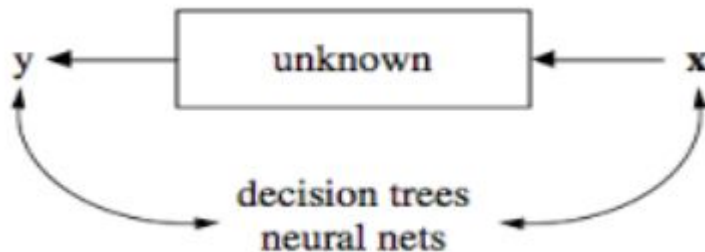
Machine Learning y Estadística



[Breiman, 2001]

- Estadística clásica:
 - Énfasis en $f(x)$. Efectos causales. Interpretación de los parámetros
 - El modelo se “postula” en base a supuestos sobre $f(x)$ en base a conocimiento acumulado sobre el problema (teoría) o en base al diseño de un experimento
 - Probabilidad. Tests estadísticos, error estándar, etc.
 - Modelo bueno: estimadores insesgados, robustos, varianza mínima

Machine Learning y Estadística

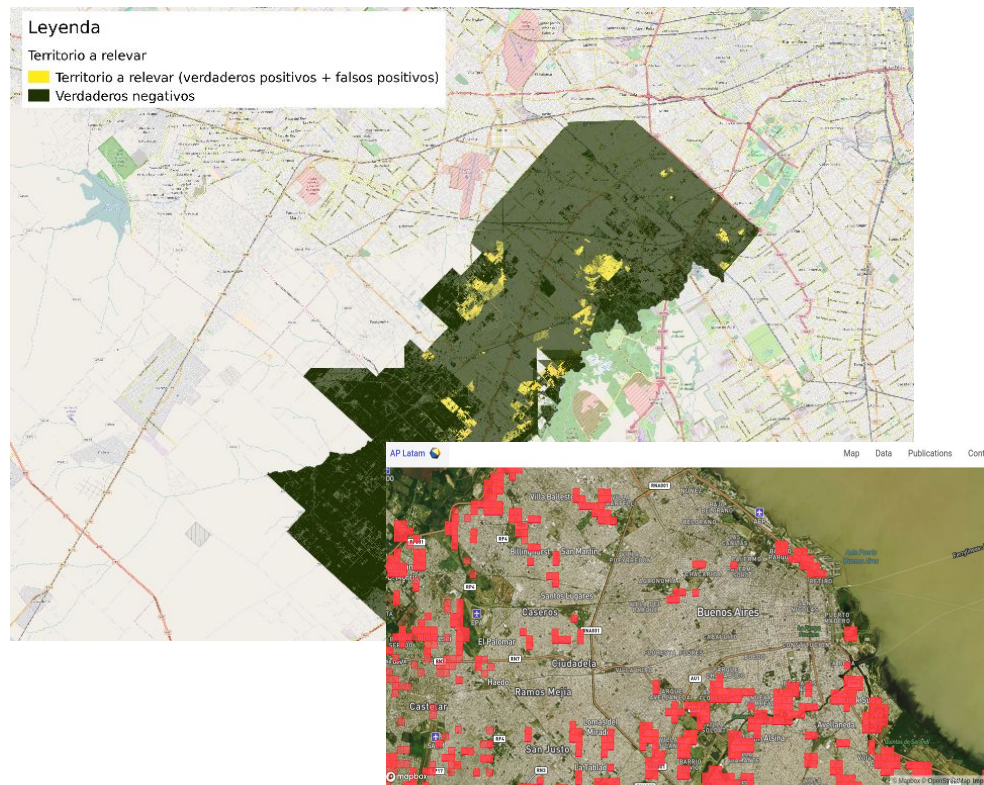


[Breiman, 2001]

- Machine Learning
 - Énfasis en Y: predicción, clasificación, medición
 - El modelo se “aprende” de los datos
 - No hay inferencia sino predicciones puntuales
 - Modelo bueno: buena performance predictiva

Aplicaciones de ML en Ciencias Sociales

Detección automática de
asentamientos informales
[Baylé, 2016]



Aplicaciones de ML en Ciencias Sociales



Ensemble Learning para imputación de datos perdidos [Rosati, 2017]

Aplicaciones de ML en Ciencias Sociales

Integración de comunidades inmigrantes en grandes ciudades

[Lamanna, Lenormand, et al 2016]

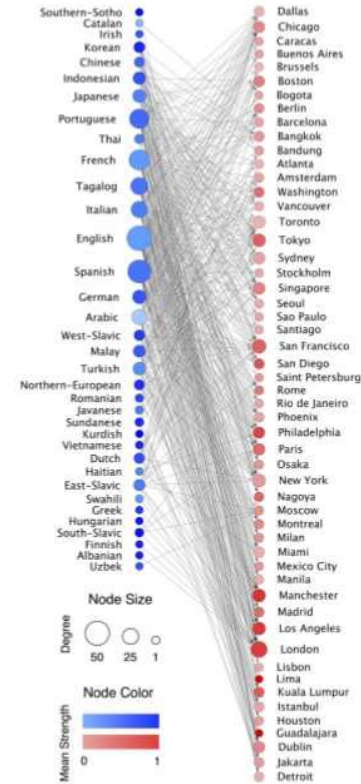


Fig 2. Bipartite spatial integration network. The network comprises of two sets: L of Languages and C of cities; the languages detected are connected to the cities set where the corresponding community of immigrants has been found. The weight of the edge corresponds to the values of h_{lc} . The size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

Machine Learning y Ciencias Sociales

- Avances relevantes para las Ciencias Sociales
 - Interpretable ML
 - Cajas Negras
 - Sesgo Algorítmico
 - Posibilidad de Interpretar resultados
 - **Masificación de APIs “amigables” para entrenar modelos y para usar modelos ya entrenados**
 - Aprendizaje No Supervisado
 - Text Mining, Natural Language Processing

Machine Learning y Ciencias Sociales

- Avances relevantes para las Ciencias Sociales
 - Interpretable ML
 - Cajas Negras
 - Sesgo Algorítmico
 - Posibilidad de Interpretar resultados
 - Masificación de APIs “amigables” para entrenar modelos y para usar modelos ya entrenados
 - **Aprendizaje No Supervisado**
 - **Text Mining, Natural Language Processing**

NLP - Text Mining - Topic Modelling

Seeking Life's Bare (Genetic) Necessities

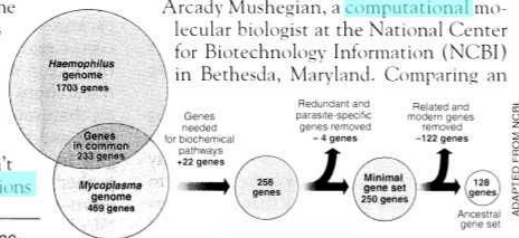
COLD SPRING HARBOR, NEW YORK—How many **genes** does an **organism** need to **survive**? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for **life**. One research team, using **computer** analyses to compare known **genomes**, concluded that today's **organisms** can be sustained with just 250 genes, and that the earliest life forms required a mere 128 **genes**. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those **predictions**

* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

SCIENCE • VOL. 272 • 24 MAY 1996

"are not all that far apart," especially in comparison to the 75,000 **genes** in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a **genetic numbers** game, particularly as more and more **genomes** are completely mapped and sequenced. "It may be a way of organizing any newly **sequenced genome**," explains Arcady Mushegian, a **computational** molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an

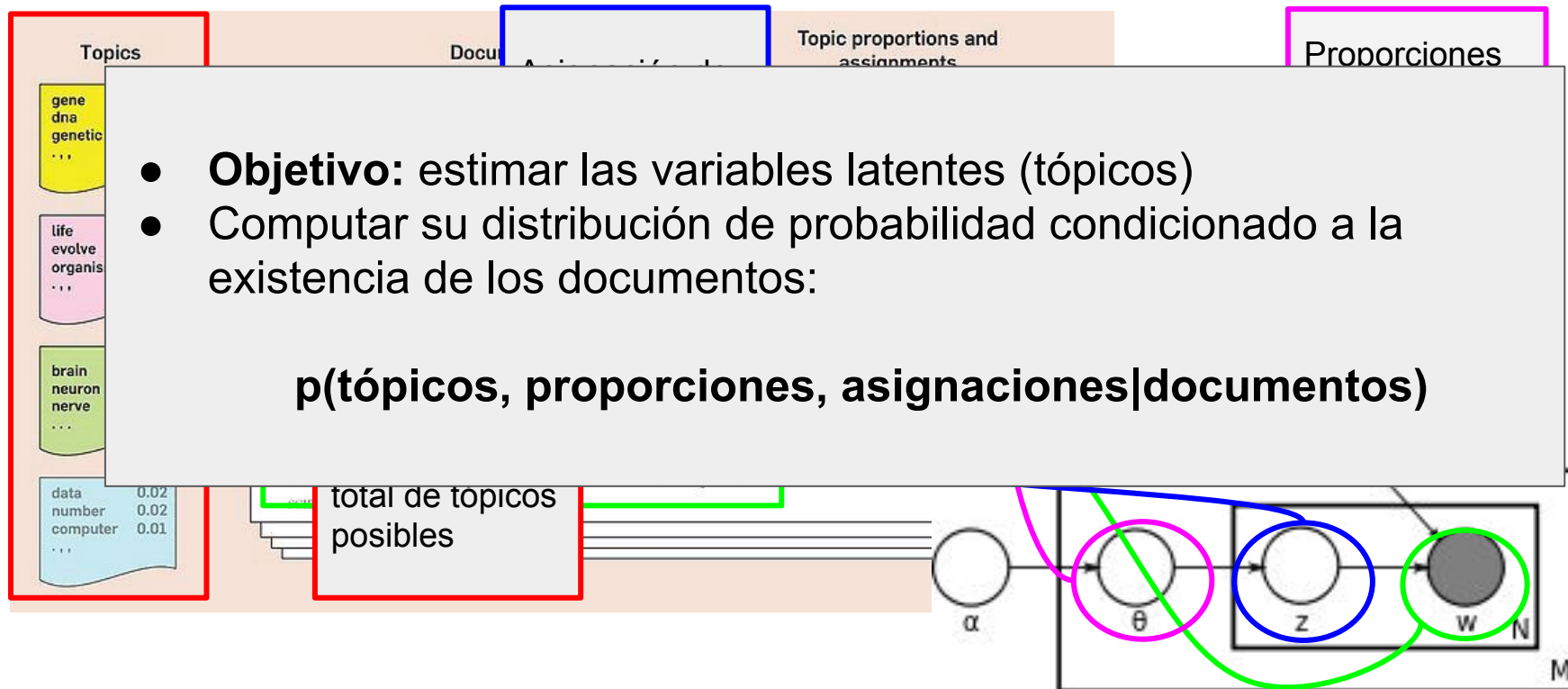


Stripping down. **Computer analysis** yields an estimate of the minimum modern and ancient genomes.

- Intuición: Un documento se compone de muchos tópicos
- Supuestos:
 - Cada tópico es una distribución de palabras con diferentes probabilidades
 - Cada documento es una mezcla de diferentes tópicos
 - Cada palabra se “extrae” de alguno de estos tópicos
- Objetivo: queremos estimar los tópicos en un corpus

[Blei, 2012]

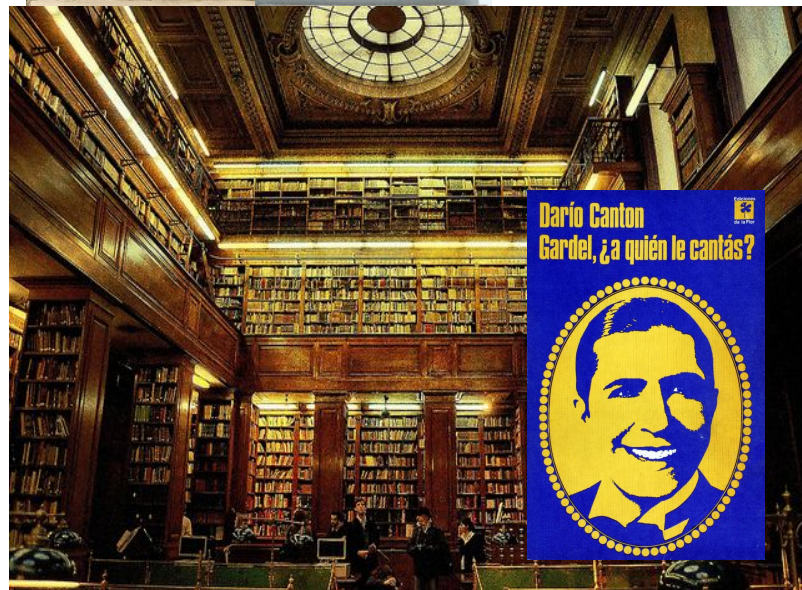
NLP - Text Mining - Topic Modelling



NLP - Text Mining - Topic Modelling

Enfoque “tradicional”

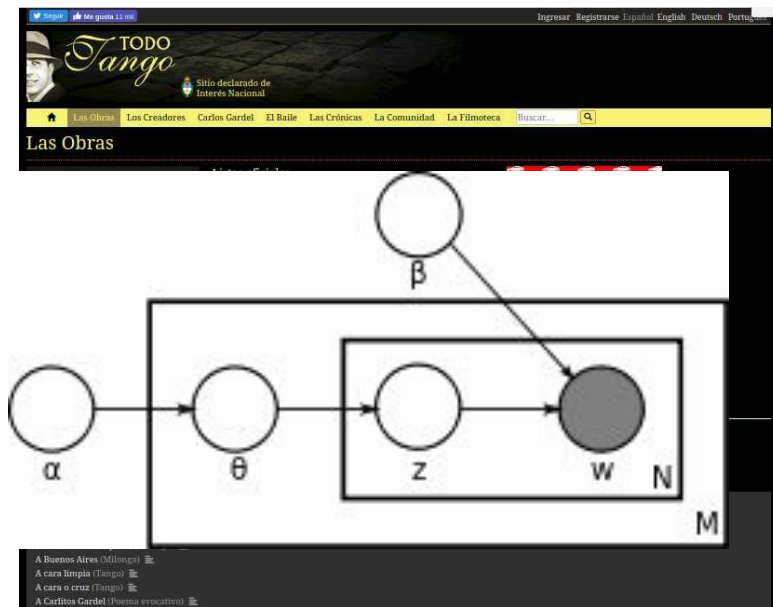
- Problema: queremos analizar un corpus de documentos grande
 - ~5.700 letras de tango
- Cantón (1972), analiza ciertos aspectos relevantes de las letras de los tangos cantados por Gardel



NLP - Text Mining - Topic Modelling

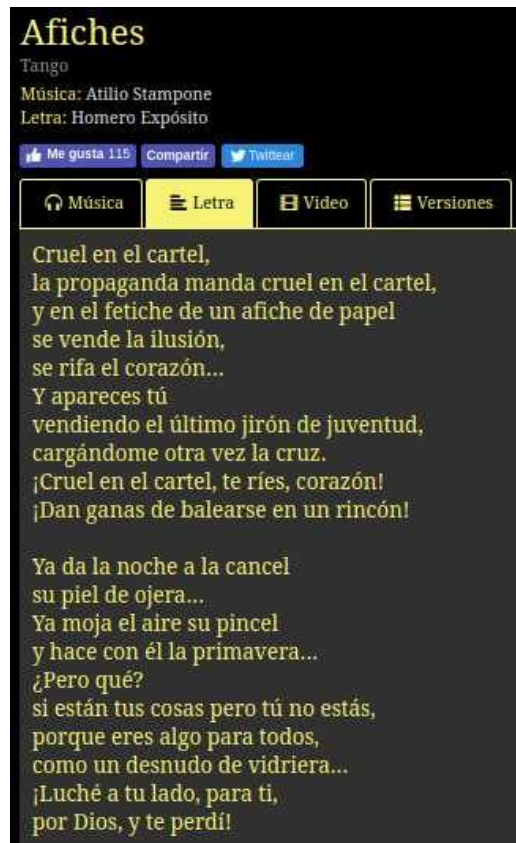
Enfoque Machine Learning

- Scrap de letras del sitio todotango.com
- Detección automática de tópicos
-Topic Modelling-: Latent Dirichlet Allocation



NLP - Text Mining - Topic Modelling

- Corpus: 5.700 letras
- Problema: analizar un corpus de ~5.700 letras de tango para detectar “tópicos”
- Pasar del “texto libre” a un formato de matriz



A screenshot of a digital interface for Tango lyrics. At the top, the title "Afiches" is displayed in a large, bold, yellow font. Below it, the artist "Tango" is listed, followed by "Música: Atilio Stampone" and "Letra: Homero Expósito". There are three interactive buttons: "Me gusta 115" (with a thumbs-up icon), "Compartir" (with a share icon), and "Twitter" (with a bird icon). Below these buttons is a navigation bar with four tabs: "Música" (with a headphones icon), "Letra" (with a document icon and highlighted in yellow), "Video" (with a video camera icon), and "Versiones" (with a list icon). The main content area displays the lyrics of the song "Afiches" in a yellow font on a dark background. The lyrics are arranged in two paragraphs, with line breaks corresponding to the original text.

Afiches
Tango
Música: Atilio Stampone
Letra: Homero Expósito

Me gusta 115 Compartir Twitter

Música **Letra** Video Versiones

Cruel en el cartel,
la propaganda manda cruel en el cartel,
y en el fetiche de un afiche de papel
se vende la ilusión,
se rifa el corazón...
Y apareces tú
vendiendo el último jirón de juventud,
cargándome otra vez la cruz.
¡Cruel en el cartel, te ríes, corazón!
¡Dan ganas de balearse en un rincón!

Ya da la noche a la cancel
su piel de ojera...
Ya moja el aire su pincel
y hace con él la primavera...
¿Pero qué?
si están tus cosas pero tú no estás,
porque eres algo para todos,
como un desnudo de vidriera...
¡Luché a tu lado, para ti,
por Dios, y te perdí!

NLP - Text Mining - Topic Modelling

Vectorizando texto: del texto libre a una matriz

TANGO	agua	blanda	cartel	cruel	el	en	era	la	manda	más	propaganda	que
Cruel en el cartel, la propaganda manda cruel en el cartel,	0	0	2	2	2	2	0	1	1	0	1	0
Era más blanda que el agua que el agua blanda	2	2	0	0	2	0	1	0	0	1	0	2

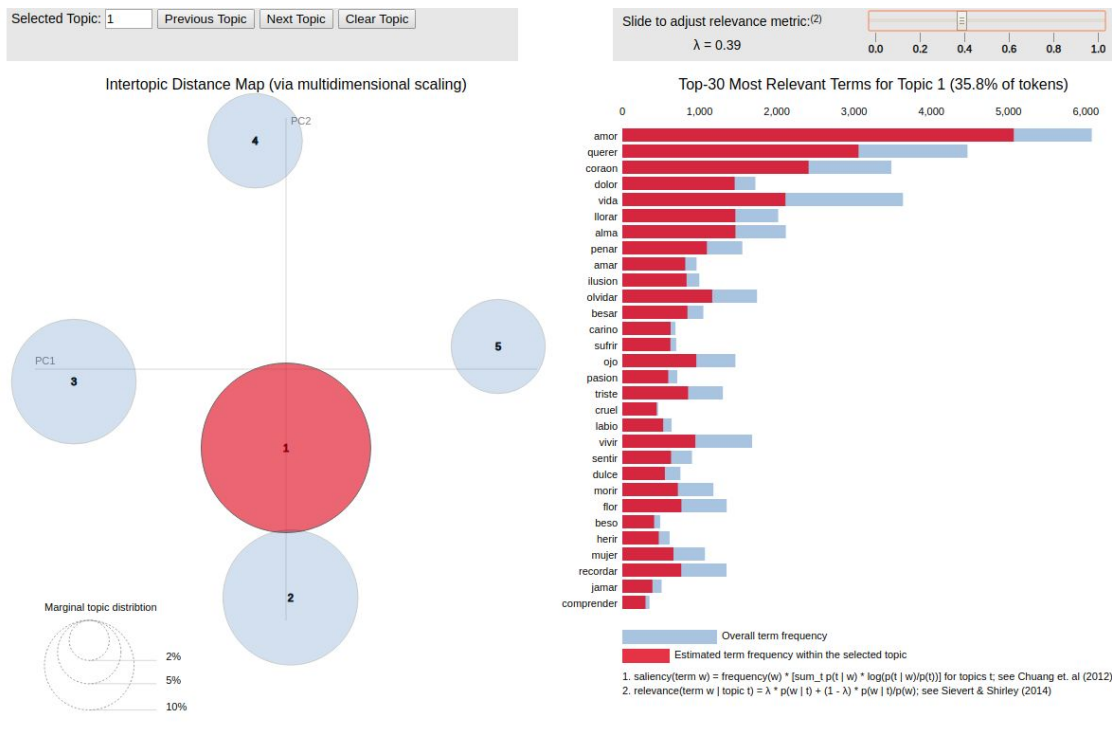
NLP - Text Mining - Topic Modelling

Vectorizando texto: del texto libre a una matriz

TANGO	agua	blanda	cartel	cruel	era	manda	propaganda
Cruel en el cartel, la propaganda manda cruel en el cartel,	0	0	2	2	0	1	1
Era más blanda que el agua que el agua blanda	2	2	0	0	1	0	0

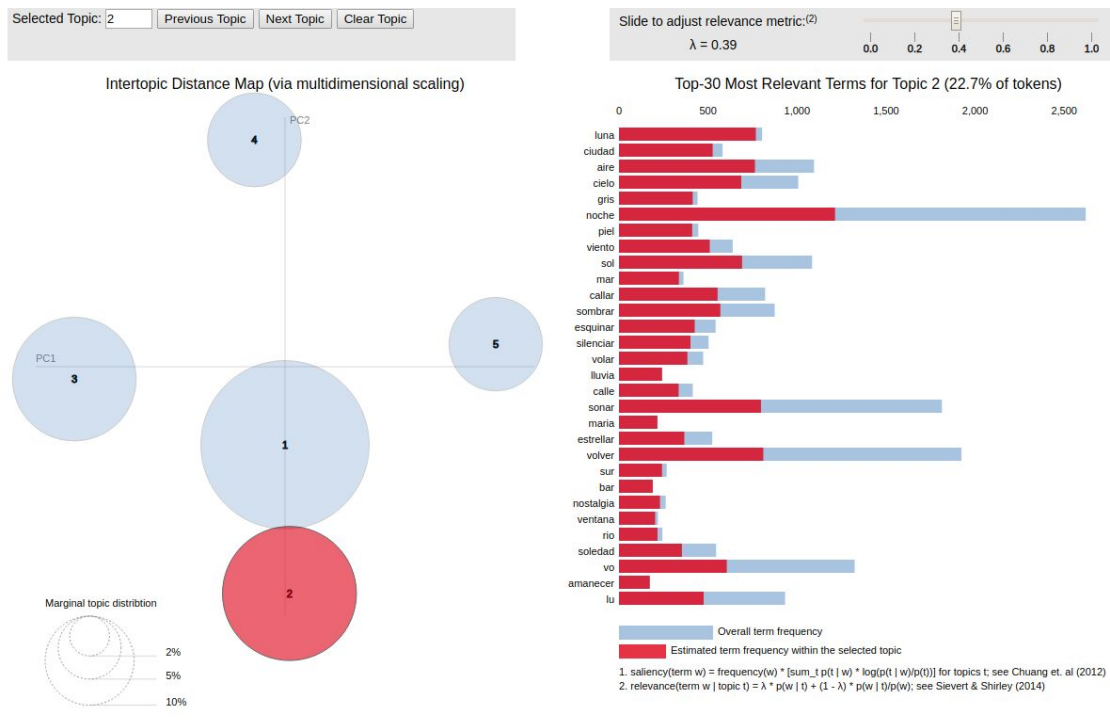
NLP - Text Mining - Topic Modelling

- Tópico 1. vinculado a las emociones (amor, odio, corazón, querer)



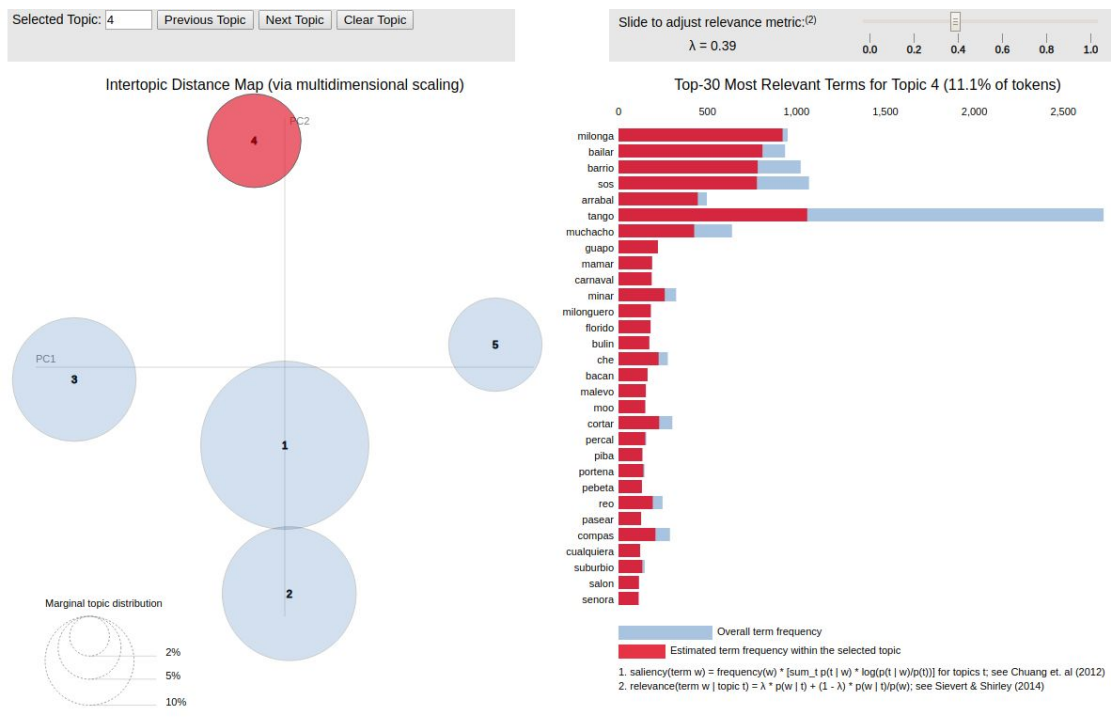
NLP - Text Mining - Topic Modelling

- Tópico 2: vinculado a una visión objetivista (luna, aire, ciudad, ...)



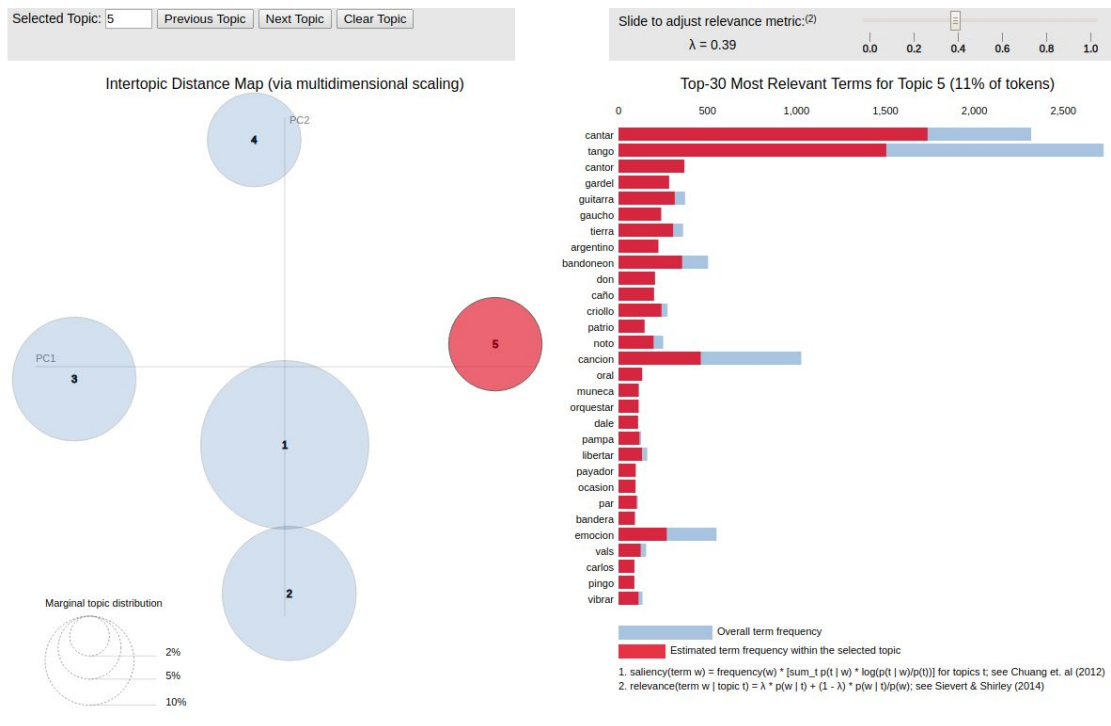
NLP - Text Mining - Topic Modelling

- Tópico 4: vinculado al tango, a la música y al barrio, arrabal



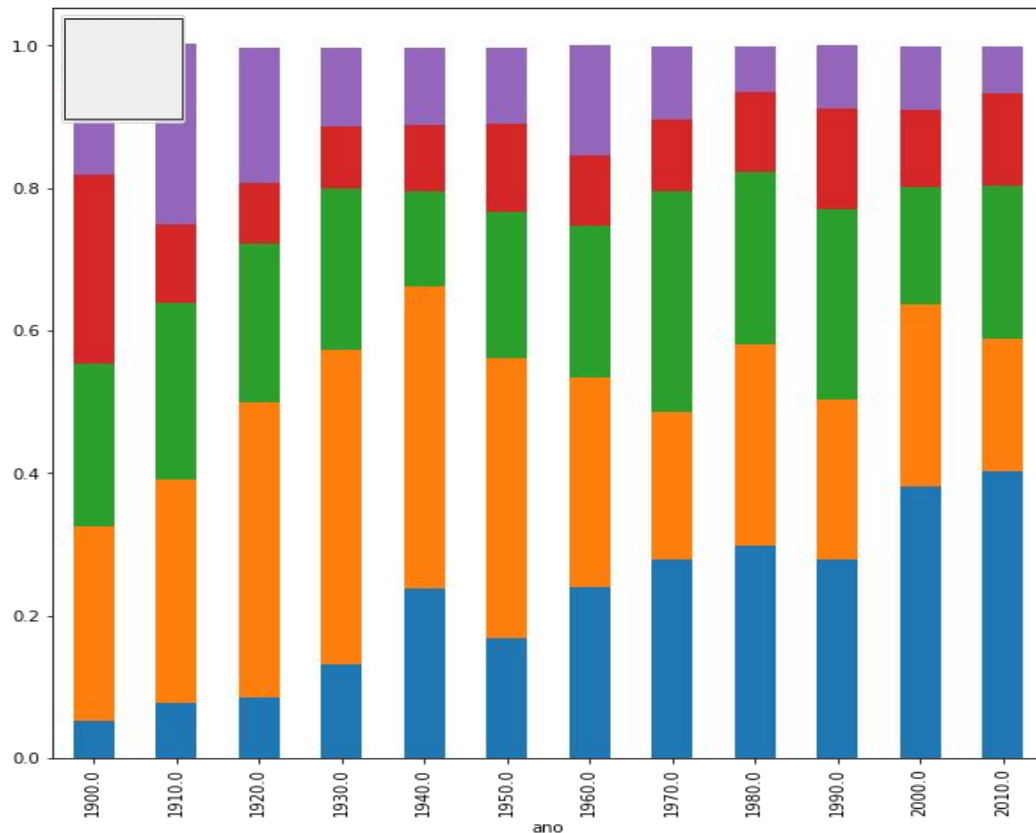
NLP - Text Mining - Topic Modelling

- Tópico 5: también vinculado a la música pero más “rural”



NLP - Text Mining - Topic Modelling

Evolución de los tópicos
(mediana de la
composición de los
documentos) 1880-2010



NLP - Text Mining - Topic Modelling

Composición de tópicos en algunos tangos

Título	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
Arrabal amargo	<u>0.56</u>	<u>0.18</u>	0.06	0	0.2
Barrio reo	<u>0.39</u>	0	0.08	0.18	<u>0.34</u>
Cafetin de Buenos Aires	<u>0.42</u>	<u>0.25</u>	0.21	0.01	0.11
Garua	<u>0.25</u>	<u>0.74</u>	0	0	0
Lejana tierra mia	<u>0.31</u>	<u>0.48</u>	0	0.18	0.03
Malena	0.14	<u>0.55</u>	0	<u>0.3</u>	0
Ventanita florida	<u>0.74</u>	0.12	0	0	0.13

¿Y entonces?

Dos formas de vinculación entre Cs. Sociales y Machine Learning

- **Como usuarios o consumidores**
 - \approx a la que tenemos con la estadística
 - Usuarios de métodos, APIS, etc.

¿Y entonces?

Dos formas de vinculación entre Cs. Sociales y Machine Learning

- **Como productores**
 - Planteo de nuevos problemas relevantes
 - Reformulación de nuevos métodos en base a problema

¿Preguntas?

german.rosati@gmail.com

Bibliografía

[\[Baylé, 2016\]](#)

[\[Blei, 2012\]](#)

[\[Breiman, 2001\]](#)

[\[Cantón, 1972\]](#)

[\[Corcoran, Carrillo, Fernández Slezak et al, 2018\]](#)

[\[Chollet, 2018\]](#)

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

[\[Rosati, 2017\]](#)

[\[Taddy, 2018\]](#)