



# Nociones fundamentales

Proyecto VIP: Aplicaciones de Machine Learning en políticas públicas y economía

---

**Bastián González-Bustamante**

University of Oxford

Universidad de Santiago de Chile

✉ [bastian.gonzalezbustamante@politics.ox.ac.uk](mailto:bastian.gonzalezbustamante@politics.ox.ac.uk)

Presentación preparada para taller del proyecto

Vertically Integrated Project (VIP), 26 de agosto de 2021

# Tabla de contenidos

1. Inteligencia artificial y ML
2. Training Data Lab
3. Proyectos y ejemplos
4. Datos proyecto VIP



# Inteligencia artificial y ML

---

## Turing test

During the Turing test, the human questioner asks a series of questions to both respondents.

After the specified time, the questioner tries to decide which terminal is operated by the human respondent and which terminal is operated by the computer.

■ QUESTION TO RESPONDENTS ■ ANSWERS TO QUESTIONER

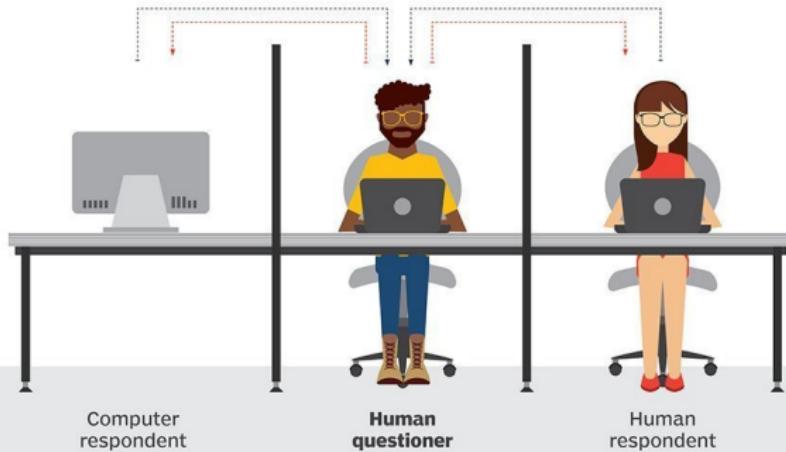


ILLUSTRATION: ESTUDIO GROUPOHOSE STOCK

©2021 TECHTARGET. ALL RIGHTS RESERVED TechTarget

Se asocia la IA con la capacidad de crear programas que puedan realizar operaciones como los humanos, por ejemplo, razonamiento lógico y aprendizaje.

Para pasar el test de Turing sería necesario:

- Procesamiento de lenguaje natural
- Almacenar información
- Razonamiento (econometría clásica) y aprendizaje automático (ML)
- Percepción del entorno y robótica (interactuar)

# Aprendizaje de máquinas

ML se puede considerar un subcampo de computer science que construye **algoritmos** para resolver un problema práctico básicamente usando reco-pilación de **big data** y la construcción de modelos a partir de esos datos.

- Requiere de cierta capacidad de procesamiento
- No necesariamente hay teoría para entender relaciones
- La inteligencia artificial puede aprender de relaciones subyacentes en big data

Aprendizaje supervisado para encontrar  $Y = f(X)$ :  $(X_i, Y_i)_{i=1}^n$

Aprendizaje no supervisado (patrones):  $(X_i)_{i=1}^n$

# **Training Data Lab**

---

# Training Data Lab

© 2020 Training Data Lab es un grupo de investigación que se enfoca en aplicaciones de ciencia de datos en ciencias sociales en tres áreas interconectadas: **minería de datos, modelamiento econométrico y aprendizaje automático**. Por una parte, buscamos recoger datos con técnicas de minería para elaborar modelos econométricos con técnicas observacionales o de emparejamiento.

Por otro lado, nos enfocamos en entrenar modelos con técnicas de aprendizaje automático y profundo etiquetando conjuntos de datos para diferentes proyectos. Lo anterior, nos permite clasificar datos no codificados usando nuestros modelos entrenados incorporando validación humana en el flujo de trabajo, lo que mejora la inteligencia artificial en los procesos de aprendizaje.

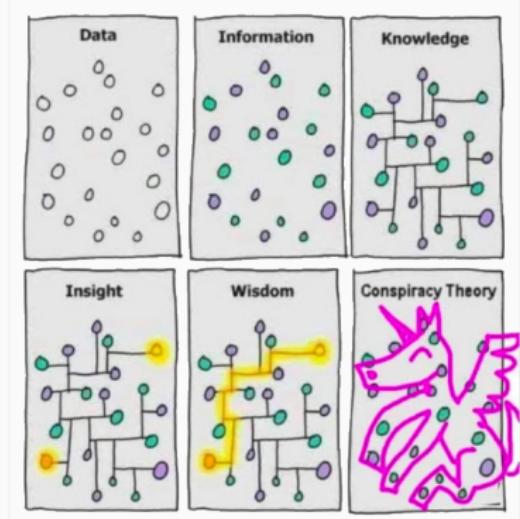
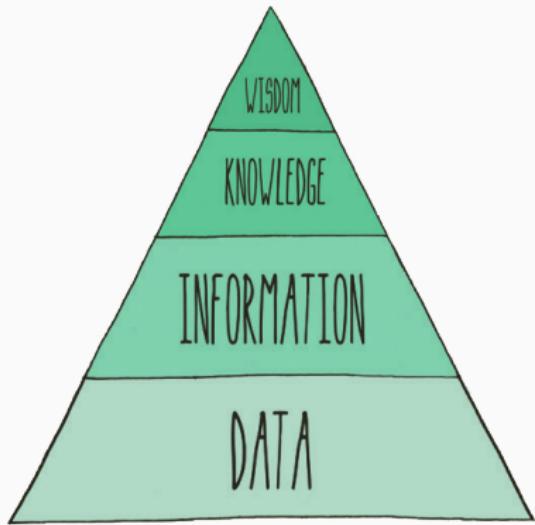


Universiteit  
Leiden



UNIVERSIDAD  
MAYOR

# Training Data Lab



## Proyectos y ejemplos

---

# Algoritmo OCR para servicio civil chileno

 Bastián González-Bustamante, Matías Astete y Berenice Orvenes

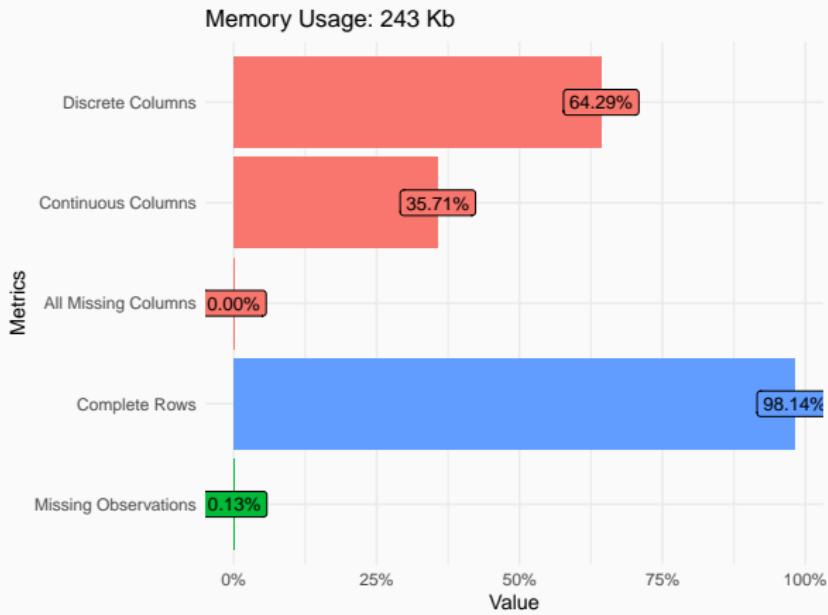
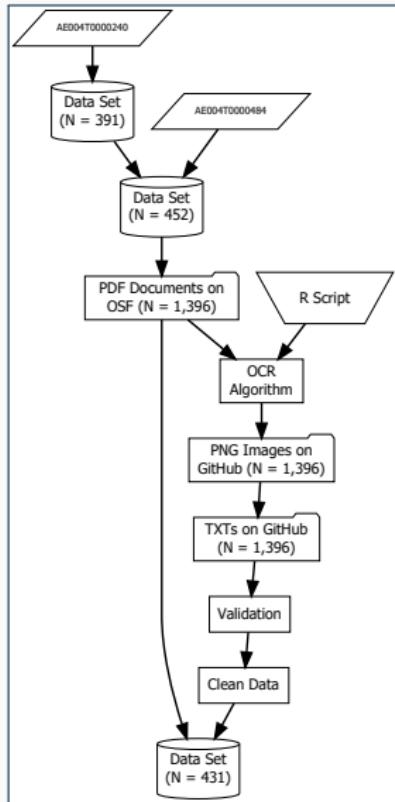
 DOI: [10.17605/OSF.IO/WBF6M](https://doi.org/10.17605/OSF.IO/WBF6M)

 [training-datalab.com/projects/chilean-civil-service](https://training-datalab.com/projects/chilean-civil-service)

**A Novel Dataset on Members of the Chilean Civil Service.** Este conjunto de datos contiene información detallada de 431 altos directivos públicos del primer nivel jerárquico del servicio civil chileno durante el período 2009-2017. Fue creado con dos solicitudes de acceso a información pública realizadas a la DNSC y una revisión de 1.396 documentos públicos, principalmente decretos y noticias institucionales. Estos documentos fueron digitalizados con algoritmos de minería de datos y revisados de forma semi-automatizada exhaustivamente.

 Revisar el [preprint en SocArXiv](#) y el [artículo publicado en español](#).

# Algoritmo OCR para servicio civil chileno



# Algoritmo OCR para servicio civil chileno

**Tesseract.** Motor para reconocimiento óptico que se comenzó a desarrollar en 1995 en Bristol y desde 2005 está disponible como un código abierto y actualmente es usado por Google. Repositorio en [GitHub](#).

La aplicación de OCR sigue varias etapas. Primero, se identifican componentes conectados y se anidan los contornos y las líneas de texto. Luego, las líneas se dividen en palabras considerando espacios. Cada palabra se intenta reconocer con diccionarios de datos entrenados (*baseline*).

Este proceso implica aprendizaje automático (*machine learning*), por tanto, se realiza una iteración con el fin de reconocer palabras que en una primera instancia no fueron identificadas. Finalmente, se revisan los espaciados difusos.

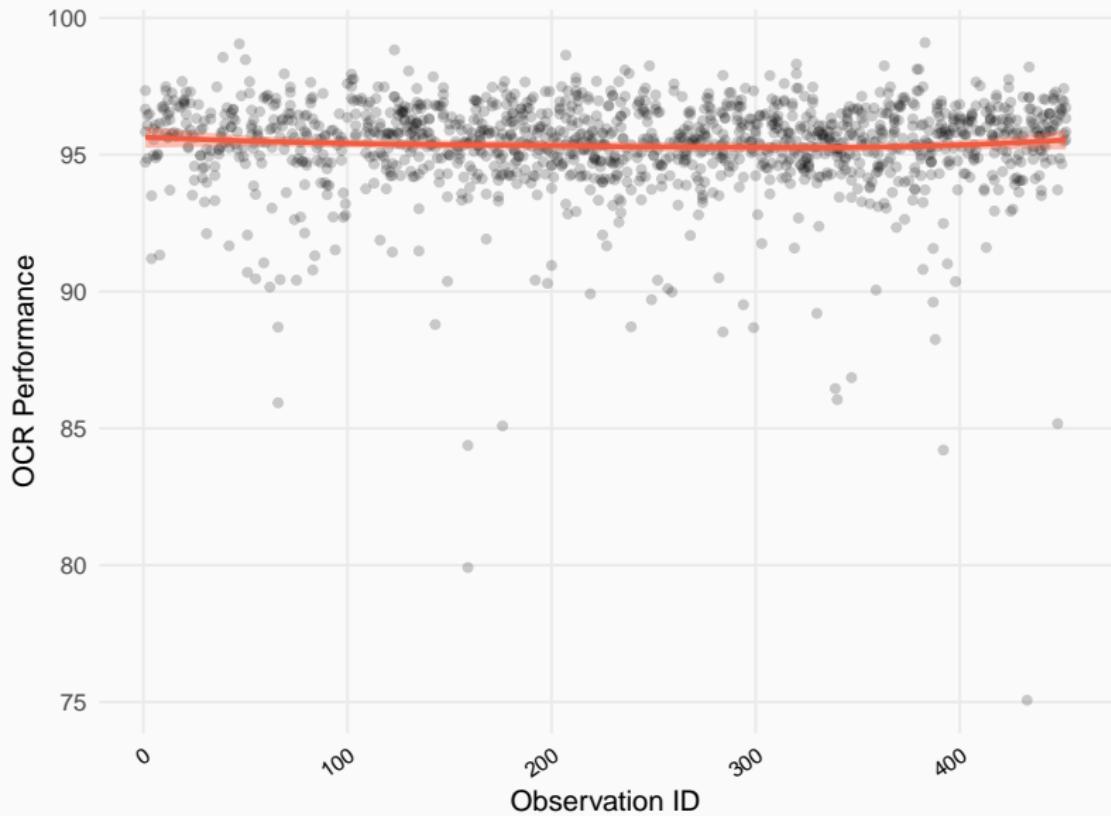
# Algoritmo OCR para servicio civil chileno

**Precisión del algoritmo OCR.** Se evaluó la proporción de texto que logró identificar correctamente. Se contrastan las palabras identificadas con diccionarios del idioma usados para entrenar los modelos Long Short Term Memory (LSTM) usados por Tesseract.

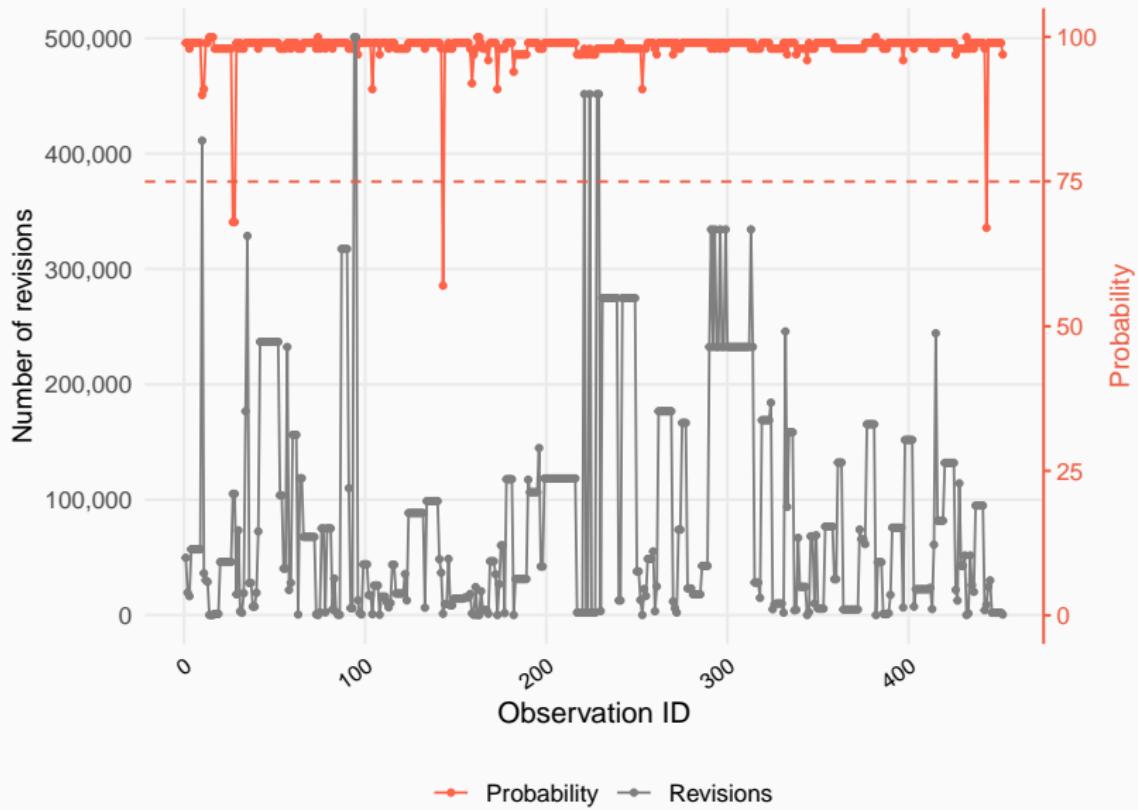
**Validación automática de sexo.** Usando el primer nombre de cada caso y una base de datos de nombres de diversos países del mundo, sexo y su predicción estimada creada en 2013. Esta base crece diariamente con datos extraídos de perfiles de redes sociales y en el momento de la revisión contaba con 111.541.298 observaciones, de las cuales 210.959 (0,19 %) correspondían a casos de Chile.

**Algoritmo criptográfico.** Las variables que contienen información personal fueron anonimizadas en la versión final del conjunto de datos con Secure Hash Algorithm de 256 caracteres (SHA256) basado en una función *hash* que bloquea la ingeniería-reversa.

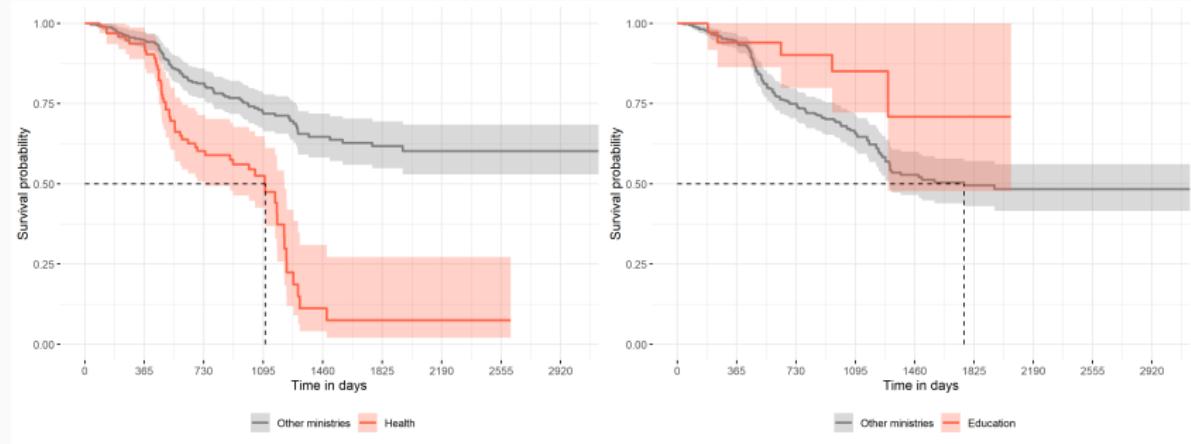
# Algoritmo OCR para servicio civil chileno



# Algoritmo OCR para servicio civil chileno



# Algoritmo OCR para servicio civil chileno



# Algoritmo clasificador para mociones legislativas

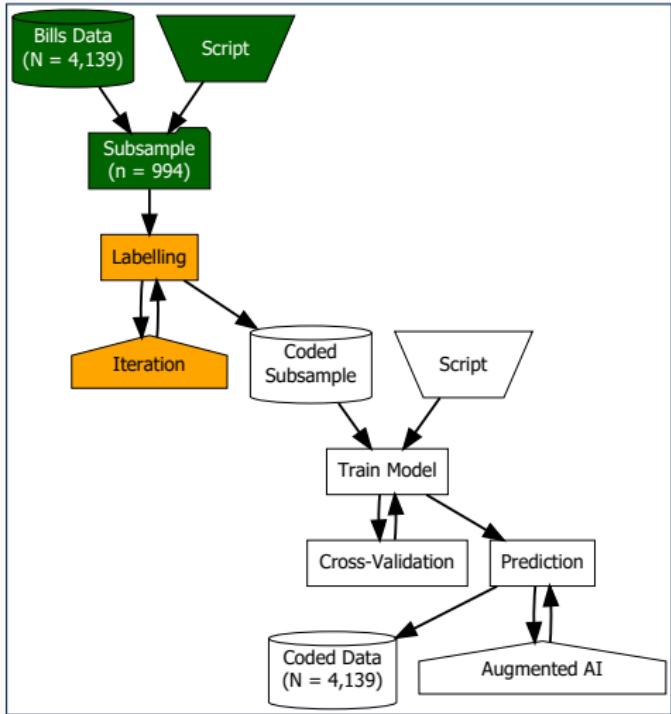
 Carla Cisternas y Bastián González-Bustamante

 **Estamos reclutando colaboradores y ayudantes**

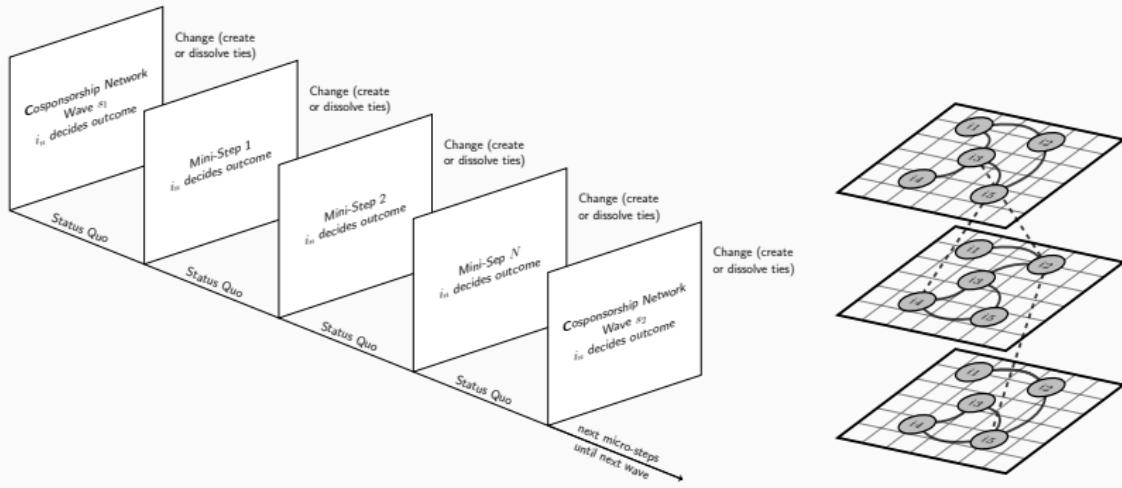
 [training-datalab.com/projects/chilean-congress-bills](http://training-datalab.com/projects/chilean-congress-bills)

**Training Data on Chilean Congress Bills.** A partir de un conjunto de datos de proyectos de ley de la Cámara de Diputados de Chile entre 2006 y 2018 ( $N = 4.139$ ), período que corresponde a tres administraciones, extraemos una submuestra aleatoria considerando algunos proyectos de ley por mes. En esta submuestra realizamos dos procedimientos de codificación de datos para identificar tanto el tema del proyecto de ley como su alcance territorial.

# Algoritmo clasificador para mociones legislativas



# Choice Modelling Networks Scheme



Source: Cisternas, C., & González-Bustamante, B. (2021). *Political Careers and Cosponsorship in the Chilean Lower House 2006-2018*. Presentation delivered at the XXVI World Congress of the Political Science, Lisbon.

## Datos proyecto VIP

---

# Algoritmo proyecto VIP

**Twitter Online Tracker of the Chilean Referendum for a New Constitution.** Rastreador online durante el plebiscito para una nueva Constitución en octubre de 2020. Contiene datos diarios de #Apruebo y #Rechazo entre el 26 de septiembre y 01 de noviembre ( $N = 2.529.134$ ).

Algunas variables disponibles son fecha, hora, usuario, texto, recuento de RTs y favs, ubicación, etc. Es necesario fusionar los conjuntos diarios y limpiar los datos.

El objetivo es entrenar un algoritmo clasificador y realizar un benchmarking de mediciones de emotividad y confiabilidad.

# Algoritmo proyecto VIP

## twConstitution



Twitter Online Tracker of the Chilean Referendum for a New Constitution

[View the Project on GitHub](#)  
bgonzalezbustamante/twConstitution

### Twitter Online Tracker of the Chilean Referendum for a New Constitution

version v1.2.6 issues 1 open issues 4 closed DOI 10.17605/OSF.IO/73NDB

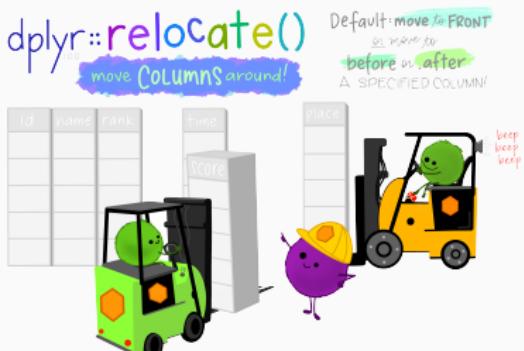
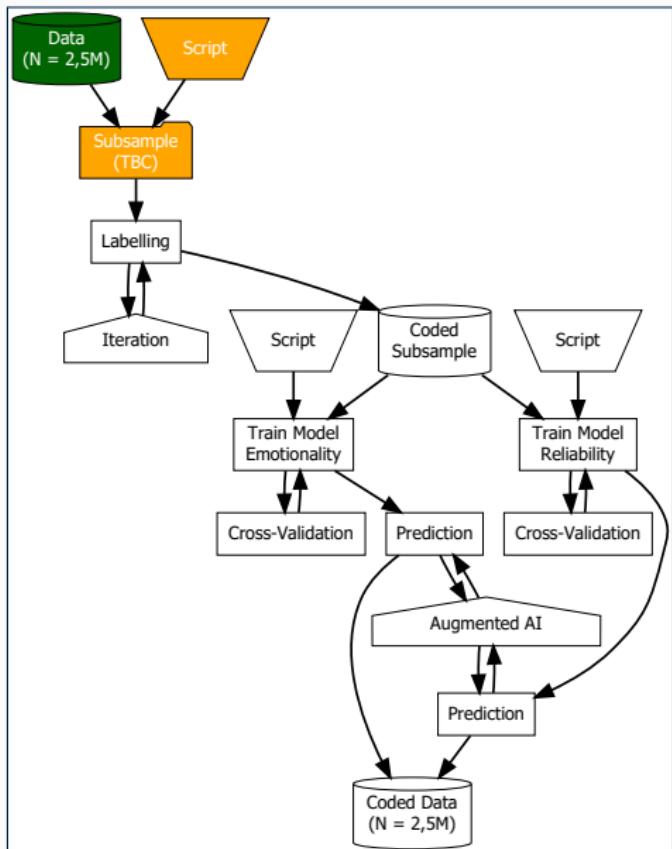
license CC-BY-4.0 made with R v4.0.2 made with Jekyll

This is a Twitter online tracker of the Chilean referendum for a new Constitution in October 2020, which contains daily datasets on **#Apruebo** (see words network, *forthcoming*) and **#Rechazo** (see words network, *forthcoming*) viewpoints on this social media.

Data sets are scraped and uploaded regularly. Some of the variables are date, hour, username, tweet text, RT count, fav count, location, among others. The data was collected during the afternoon each day. A couple of exceptions, such as October 6th and 26th, were collected early morning on the following day. This is not an issue because the data could be sliced, and in order to work with the whole period, it is necessary to merge the sets and retain unique cases.

Dataset	Date	Year	N	Size	Format
#Apruebo	Nov. 01	2020	45,195	46.5 MB	<a href="#">CSV</a>
#Rechazo	Nov. 01	2020	22,142	24.0 MB	<a href="#">CSV</a>
#Apruebo	Oct. 31	2020	50,244	52.1 MB	<a href="#">CSV</a>
#Rechazo	Oct. 31	2020	27,331	29.9 MB	<a href="#">CSV</a>
#Apruebo	Oct. 30	2020	53,889	56.1 MB	<a href="#">CSV</a>
#Rechazo	Oct. 30	2020	33,479	36.6 MB	<a href="#">CSV</a>

# Algoritmo proyecto VIP



**Solo un poco de información adicional...**

# Contacto

## Bastián González-Bustamante

DPhil (PhD) Researcher

Department of Politics and International Relations  
& St Hilda's College  
University of Oxford

📍 St Hilda's College, Cowley Place, Oxford OX4 1DY  
✉️ [bastian.gonzalezbustamante@politics.ox.ac.uk](mailto:bastian.gonzalezbustamante@politics.ox.ac.uk)  
🏡 <https://bgonzalezbustamante.com>

## Profesor Instructor

Departamento de Gestión y Políticas Públicas  
Facultad de Administración y Economía  
Universidad de Santiago de Chile  
📍 Av. Lib. B. O'Higgins 3363, Estación Central, Santiago  
✉️ [bastian.gonzalez.b@usach.cl](mailto:bastian.gonzalez.b@usach.cl)



Presentación compilada con **LATEX** y algunos ☕

⌚ Descargar la versión más reciente desde [GitHub](#)

♾ Artwork utilizado disponible en [GitHub](#)

Muchas gracias por su atención

