

Extracción y Análisis de Datos No Estructurados: Aplicaciones usando texto, audio, imágenes y video

Ramiro H. Gálvez
Departamento de Computación, FCEyN, UBA

Workshop: "Big Data, Economía y Políticas Públicas: Aplicaciones,
Herramientas y Aprendizaje Automático"

Objetivo

Mostrar que existe toda una serie de datos "no tradicionales" (imágenes, audio, texto, etc.) que pueden ser analizados y de los que cuales se puede extraer valor.

Las aplicaciones que voy a mostrar no son necesariamente de economía, pero deberían servir para disparar ideas.

¿Son todos los datos iguales?

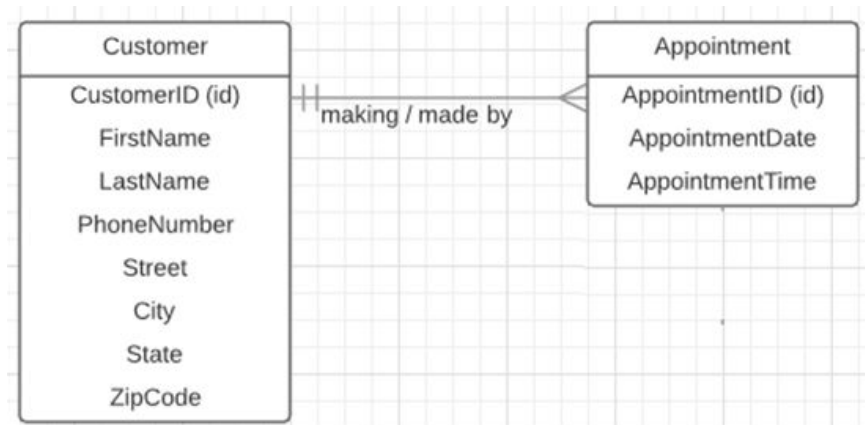
No, pensemos una dimensión en donde difieren: de qué momento vienen.

- **Pasado**: datos almacenados que se analizan en ex-post (ej: datos administrativos).
- **Presente**: datos en tiempo real que se analizan apenas se generan (ej: smart meters, subastas de publicidades online, cotización de activos financieros, etc.).

Datos estructurados

Otra dimensión en la que difieren es en sí detrás de los mismos hay un "modelo de datos", que indique, entre otras cosas:

- Qué tipos de datos hay en la base.
- Cómo se relacionan.
- Qué restricciones deben cumplir las datos.



Datos estructurados

A las colecciones de datos que siguen un modelo de datos o que siguen una organización clara, se los suele llamar **datos estructurados**.

En economía son los que tradicionalmente hemos usado, por ejemplo:

- Datos administrativos.
- Datos tributarios.
- Cuentas nacionales.
- Datos de encuestas de hogares.
- Bases de datos de ventas de una empresa.

Datos no estructurados

La definición es por negación... **no siguen un modelo de datos** o no siguen una organización clara.

Esto hace difícil para una computadora responder de manera simple preguntas sobre los mismos.

Por ejemplo:

- ¿De qué tema trata un correo electrónico?
- ¿Qué objetos hay en una determinada foto?
- ¿Qué se dijo en una determinada grabación de audio?
- ¿Cuánto tiempo aparecen un personaje en una película?

¿Son muchos los datos no estructurados?

No hay una estimación rigurosa de cuánta de la información útil para una organización es no estructurada.

Pero **se da por sentado que es mucha**, estimaciones:

- Merrill Lynch 80%.
- IBM 80%.
- 70-80% Computer World Magazine.

¿Pueden aprovecharse los datos no estructurados?

Extraer información útil de datos no estructurados presenta un mayor desafío que hacerlo de datos estructurados, sin embargo **es posible explotar esta información**.

El análisis de cada tipo de dato no estructurado es en sí **una disciplina distinta**:

- Texto: natural language processing (NLP) / text mining.
- Habla: speech technologies / NLP.
- Imágenes / video: image processing.

Aplicaciones: texto

CHAPTER 5

ARYA

They traveled dawn to dusk, past woods and orchards and neatly tended fields, through small villages, crowded market towns, and stout holdfasts. Come dark, they would make camp and eat by the light of the Red Sword. The men took turns standing watch. Arya would glimpse firelight flickering through the trees from the camps of other travelers. There seemed to be more camps every night, and more traffic on the kingsroad by day.

Morn, noon, and night they came, old folks and little children, big men and small ones, barefoot girls and women with babes at their breasts. Some drove farm wagons or bumped along in the back of ox carts. More rode: draft horses, ponies, mules, donkeys, anything that would walk or run or roll. One woman led a milk cow with a little girl on its back. Arya saw a smith pushing a wheelbarrow with his tools inside, ...

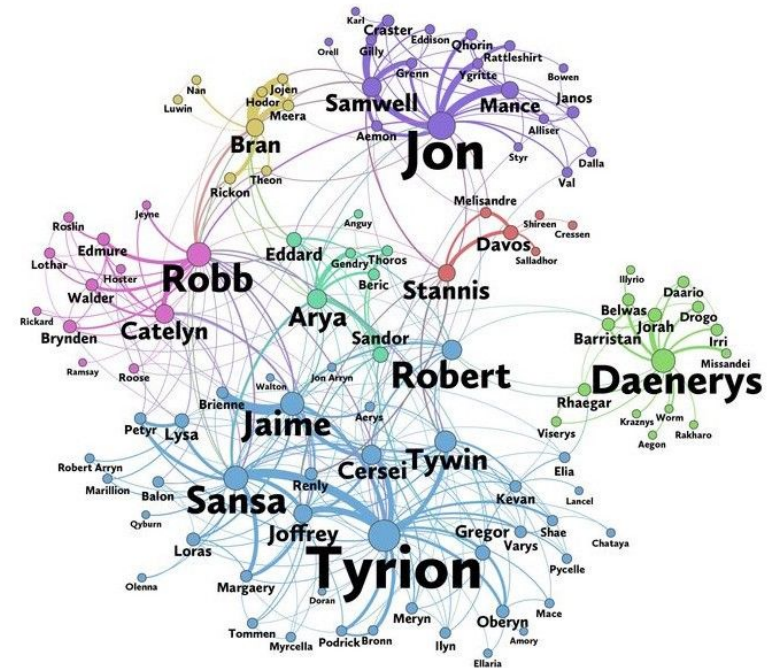
Aplicaciones: texto

CHAPTER 5

ARYA

They traveled dawn to dusk, past woods and orchards and neatly tended fields, through small villages, crowded market towns, and stout holdfasts. Come dark, they would make camp and eat by the light of the Red Sword. The men took turns standing watch. **Arya** would glimpse firelight flickering through the trees from the camps of other travelers. There seemed to be more camps every night, and more traffic on the kingsroad by day.

Morn, noon, and night they came, old folks and little children, big men and small ones, barefoot girls and women with babes at their breasts. Some drove farm wagons or bumped along in the back of ox carts. More rode: draft horses, ponies, mules, donkeys, anything that would walk or run or roll. One woman led a milk cow with a little girl on its back. **Arya** saw a smith pushing a wheelbarrow with his tools inside, ...



Aplicaciones: texto

NBER WORKING PAPER SERIES

TEXT AS DATA

Matthew Gentzkow
Bryan T. Kelly
Matt Taddy

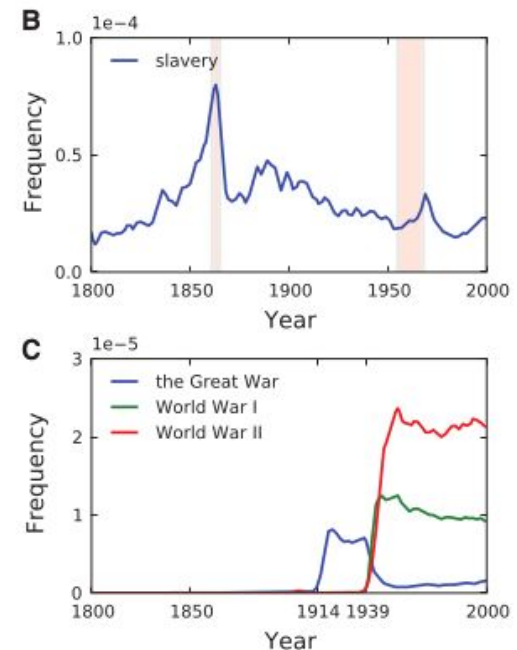
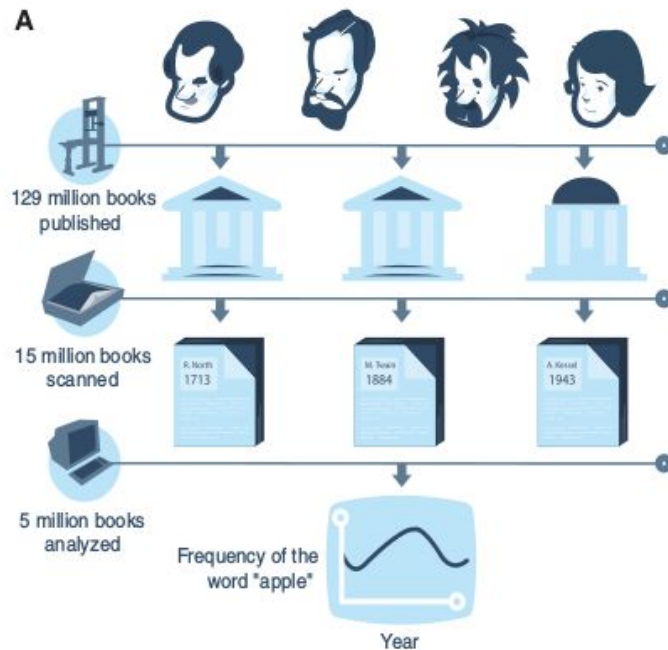
Working Paper 23276
<http://www.nber.org/papers/w23276>

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
March 2017

Aplicaciones: texto

Culturomics

Fig. 1. Culturomic analyses study millions of books at once. **(A)** Top row: Authors have been writing for millennia; ~129 million book editions have been published since the advent of the printing press (upper left). Second row: Libraries and publishing houses provide books to Google for scanning (middle left). Over 15 million books have been digitized. Third row: Each book is associated with metadata. Five million books are chosen for computational analysis (bottom left). Bottom row: A culturomic time line shows the frequency of “apple” in English books over time (1800–2000). **(B)** Usage frequency of “slavery”. The Civil War (1861–1865) and the civil rights movement (1955–1968) are highlighted in red. The number in the upper left (10^{-4}) is the unit of frequency. **(C)** Usage frequency over time for “the Great War” (blue), “World War I” (green), and “World War II” (red).



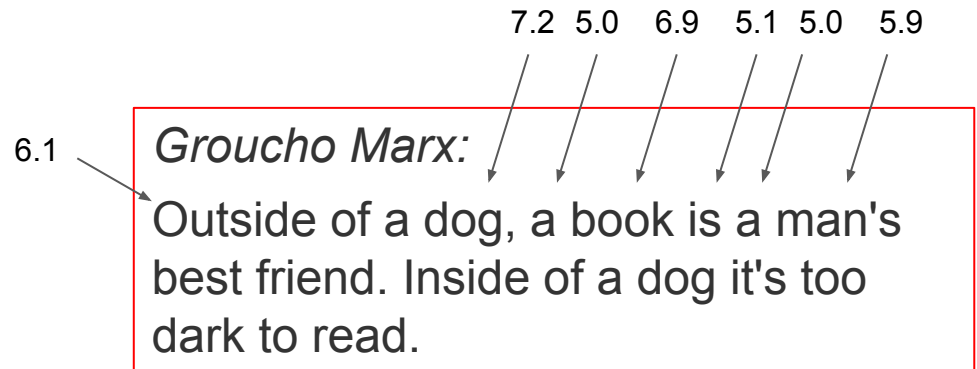
Aplicaciones: texto

Análisis de sentimiento

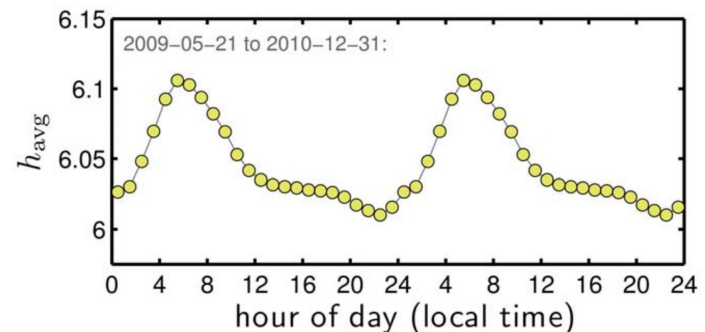
Ejemplo: Hedonometer

word	Happiness
Laughter	8.50
Happiness	8.44
love	8.42
...	
rape	1.44
suicide	1.30
terrorist	1.30

Léxico de felicidad para twitter

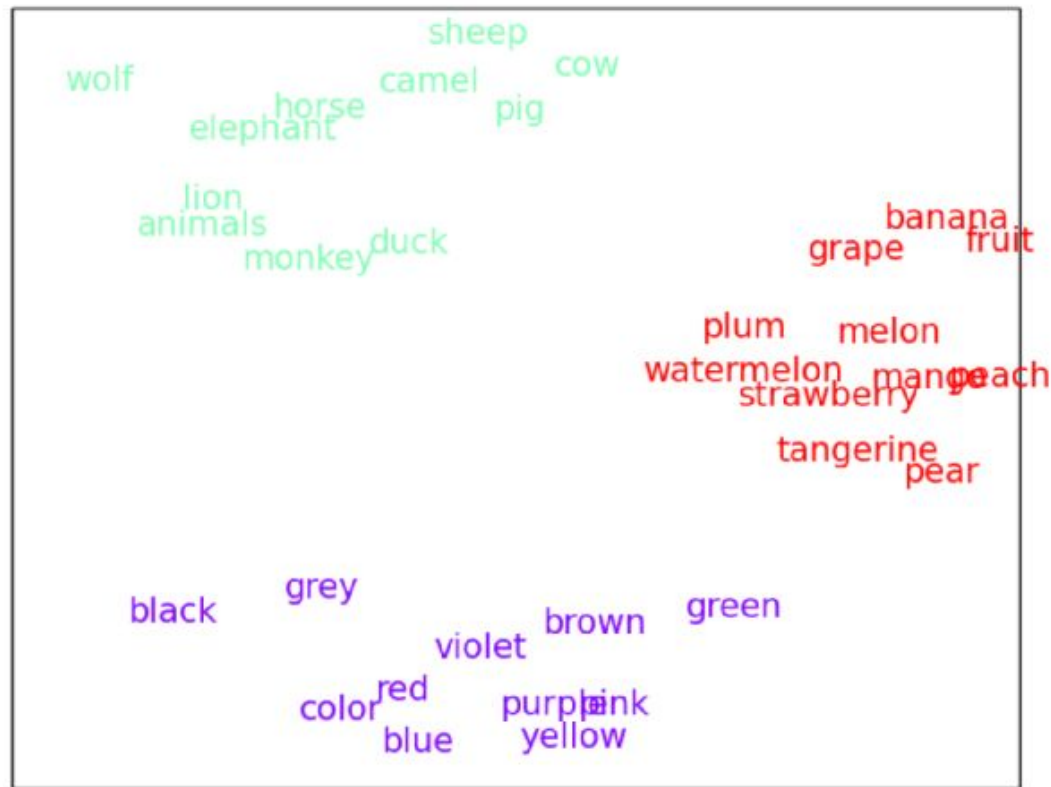


Felicidad promedio = 5.09

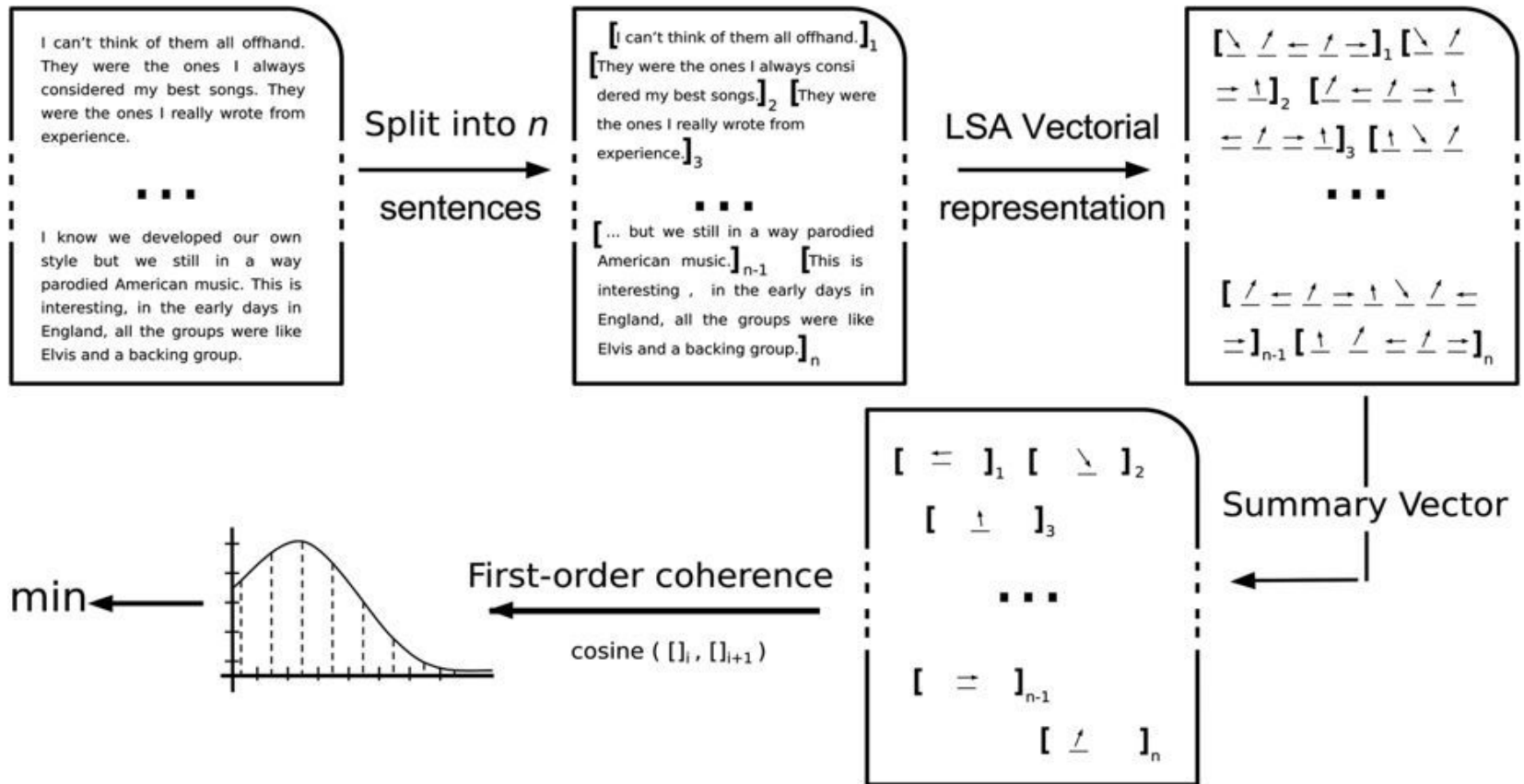


Aplicaciones: texto

Word embeddings (LSA, W2V, etc.)



Aplicaciones: texto



Aplicaciones: texto

Estereotipos de género en películas

SubRip File

279
00:24:09,973 --> 00:24:11,665
I want you to rest well, and a
month from now...

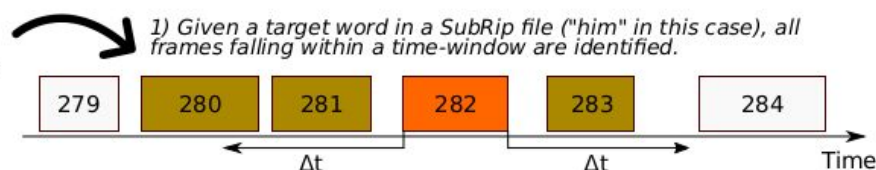
280
00:24:11,839 --> 00:24:15,203
This Hollywood big shot's
gonna give you what you want.

281
00:24:15,373 --> 00:24:18,032
Too late. They start shooting in
a week.

282
00:24:18,573 --> 00:24:21,561
I'm gonna make **him** an offer he
can't refuse.

283
00:24:24,606 --> 00:24:26,334
Now just go outside, enjoy
yourself...

284
00:24:26,572 --> 00:24:30,505
and forget about all this
nonsense.



2) Text contained in all context frames is cleaned, tokenized and lemmatized.

Tokens in the current context

this, hollywood, big, shot, 's, gon, na, give, you, what, you, want, too, late, they, start, shoot, in, a, week, i, 'm, gon, na, make, an, offer, he, ca, n't, refuse, now, just, go, outside, enjoy, yourself

Co-occurrence matrix

		Context Tokens				
		her	him	offer	smart	
Target Tokens	her	7	5	2	2	
	him	5	12	...	9	11
			:	:		
	offer	3	9	...	0	3
	smart	5	11	3	3	

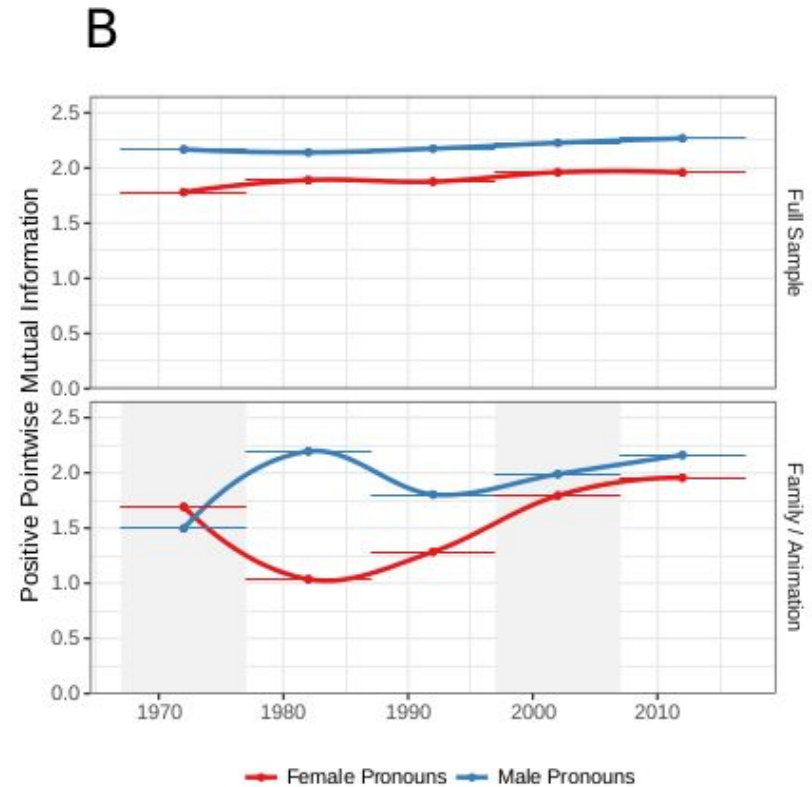
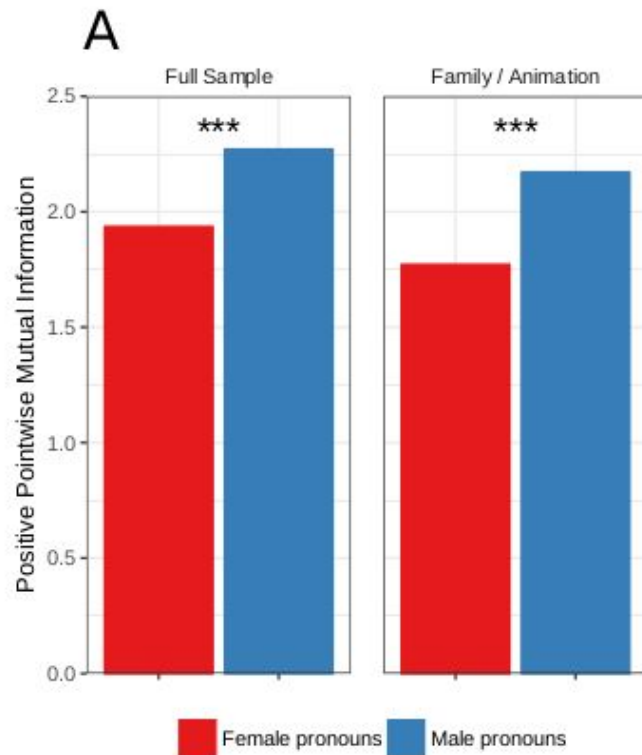
3) For each token, its number of appearances in the current context is added to the row corresponding to the target word in the co-occurrence matrix.

$$\text{PPMI}_{ij} = \max(\log_2 \frac{p_{ij}}{p_{i*} p_{*j}}, 0)$$

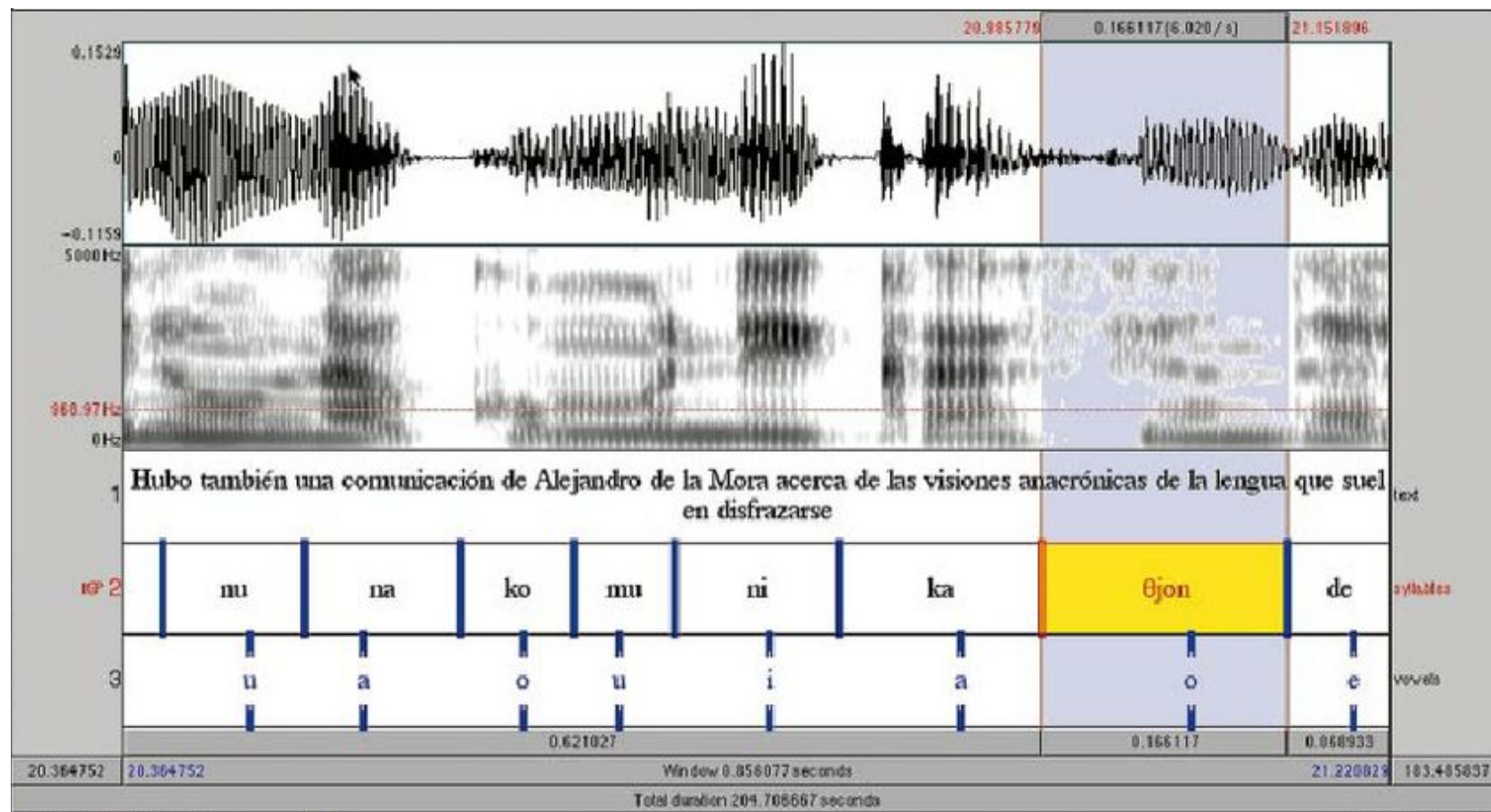
The process is repeated for every word in every subtitle under analysis.

Aplicaciones: texto

Estereotipos de género en películas



Aplicaciones: habla



Aplicaciones: habla

Modelando la **relación entre audio y palabras dichas** se pueden implementar sistemas que hagan las siguientes tareas:

- ***Automatic speech recognition*** (nos permite pasar a análisis de texto).
- ***Speech synthesis***.

Componentes claves detrás de los asistentes virtuales (*Google Assistant, Siri, Alexa, Cortana, etc.*)

Aplicaciones: habla

Pero ojo, el habla es **más que sólo la transcripción del texto**, posee también características prosódicas que hacen al mensaje. Por ejemplo:

- Tono / frecuencia.
- Velocidad del habla.
- Intensidad del habla.
- Calidad del habla.

Esto se analiza en tiempo real en los sistemas de diálogo (por ejemplo para cambios de turnos).

Aplicaciones: habla

Sistema para calificar la pronunciación de estudiantes de lenguas no nativas

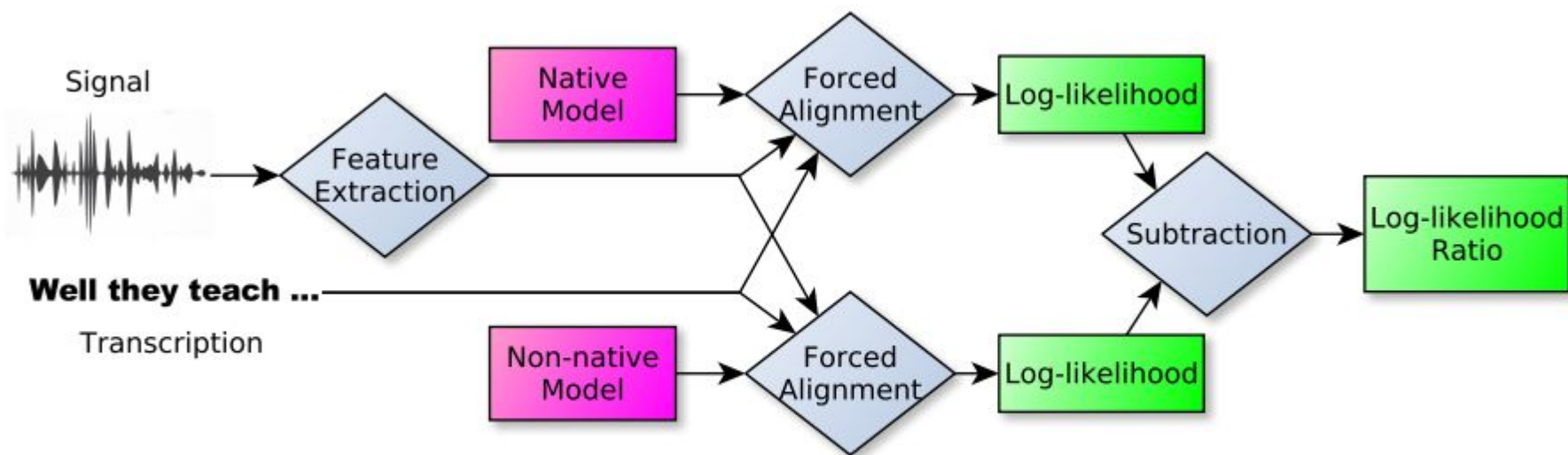
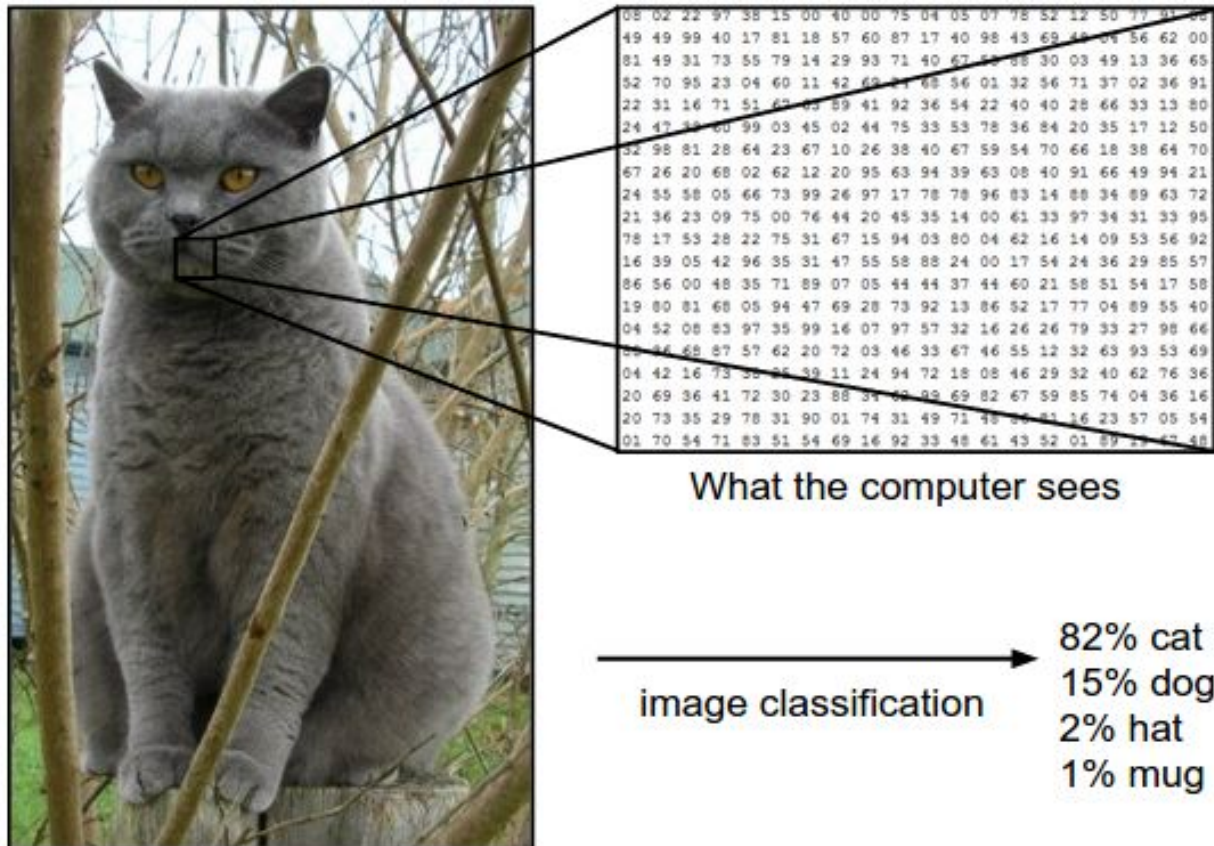


Fig. 1. LLR-based pronunciation scoring system

Aplicaciones: imágenes

Clasificación en imágenes



Aplicaciones: imágenes

Classification



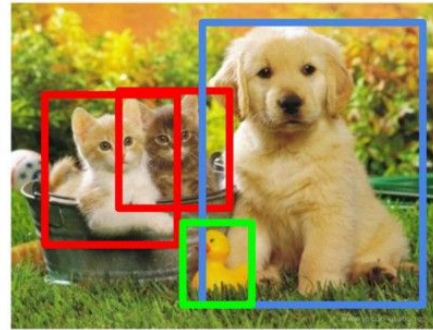
CAT

**Classification
+ Localization**



CAT

Object Detection



CAT, DOG, DUCK

**Instance
Segmentation**



CAT, DOG, DUCK

Single object

Multiple objects

Aplicaciones: imágenes

Clasificación en imágenes



Aplicaciones: imágenes

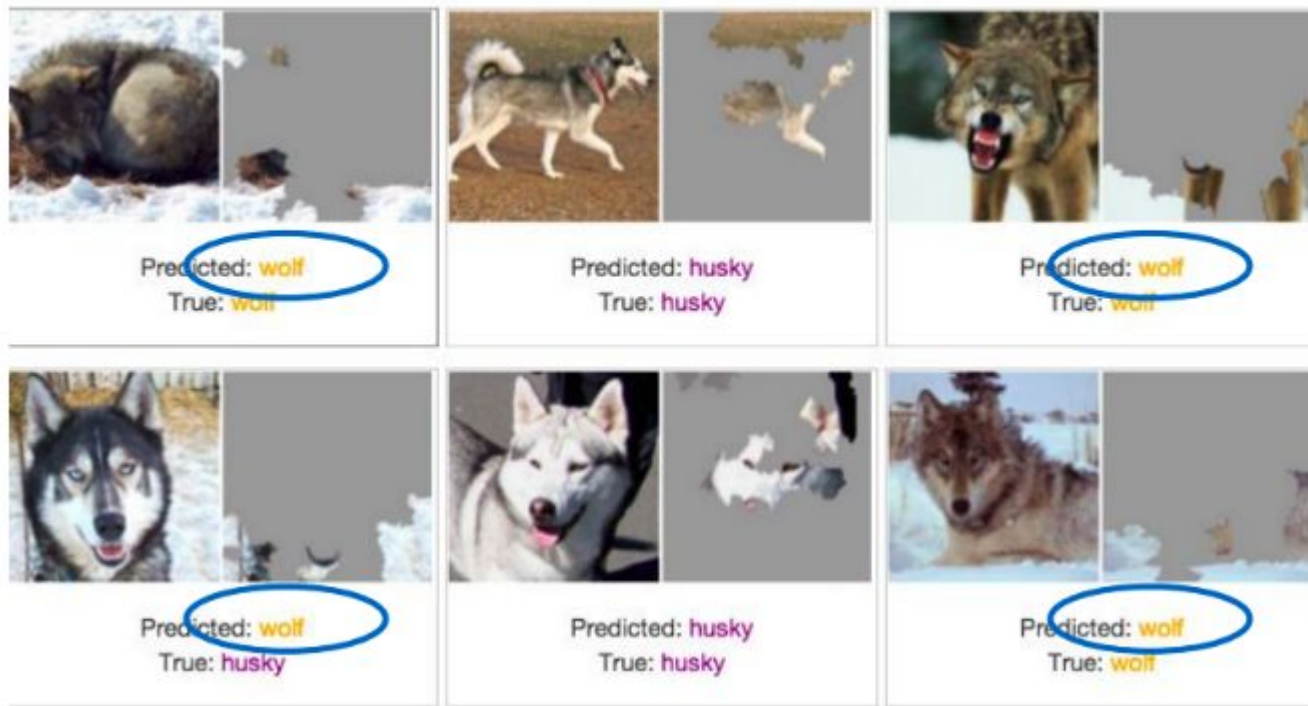
Clasificación en imágenes

Only 1 mistake!



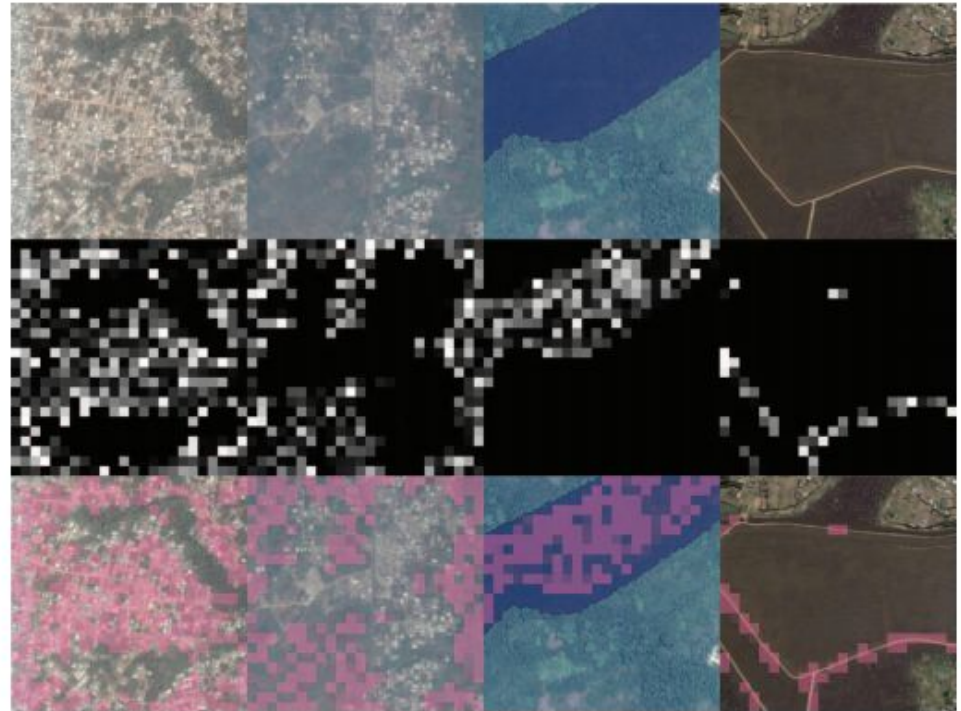
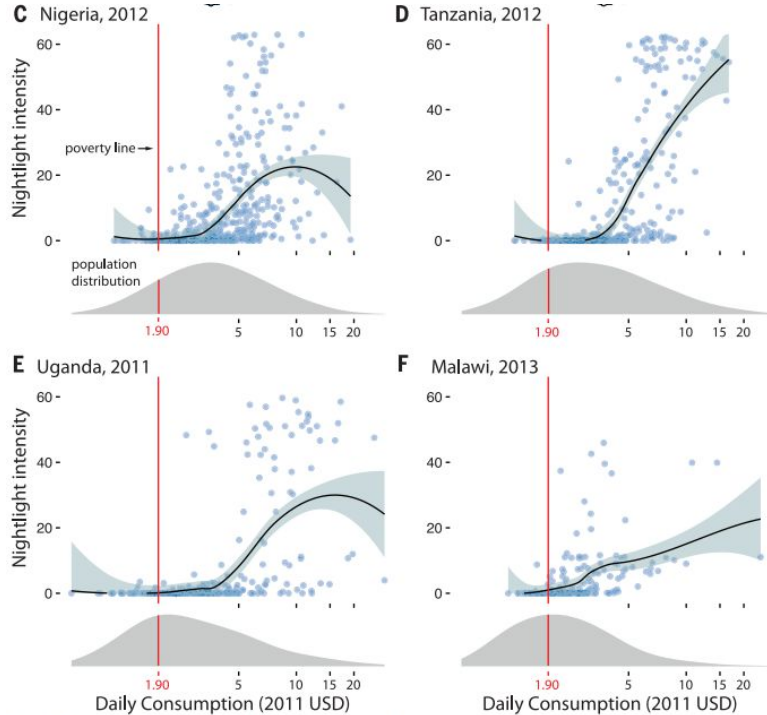
Aplicaciones: imágenes

Clasificación en imágenes



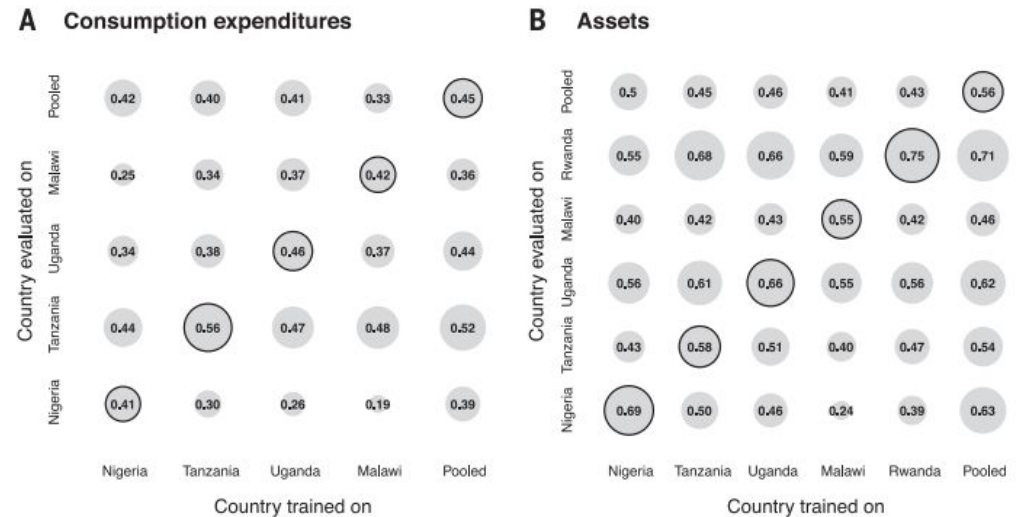
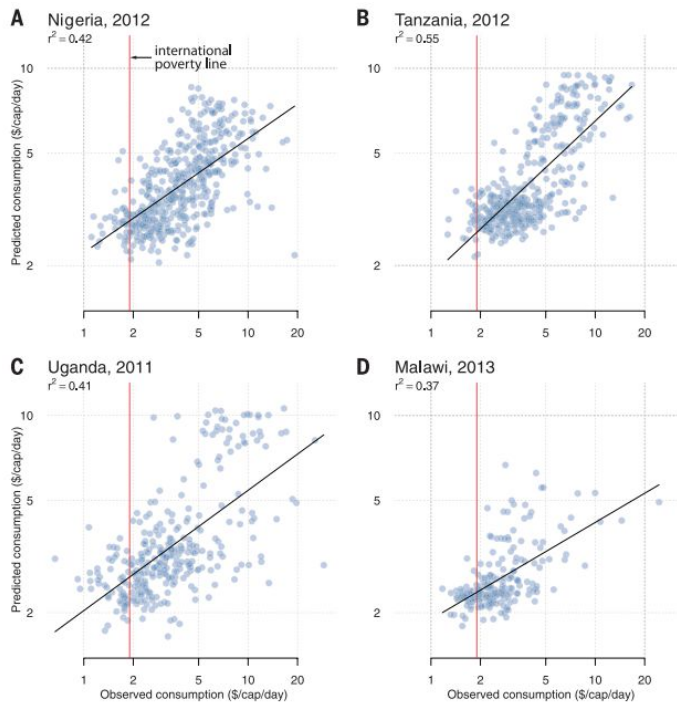
Aplicaciones: imágenes

Predicción de pobreza usando imágenes satelitales



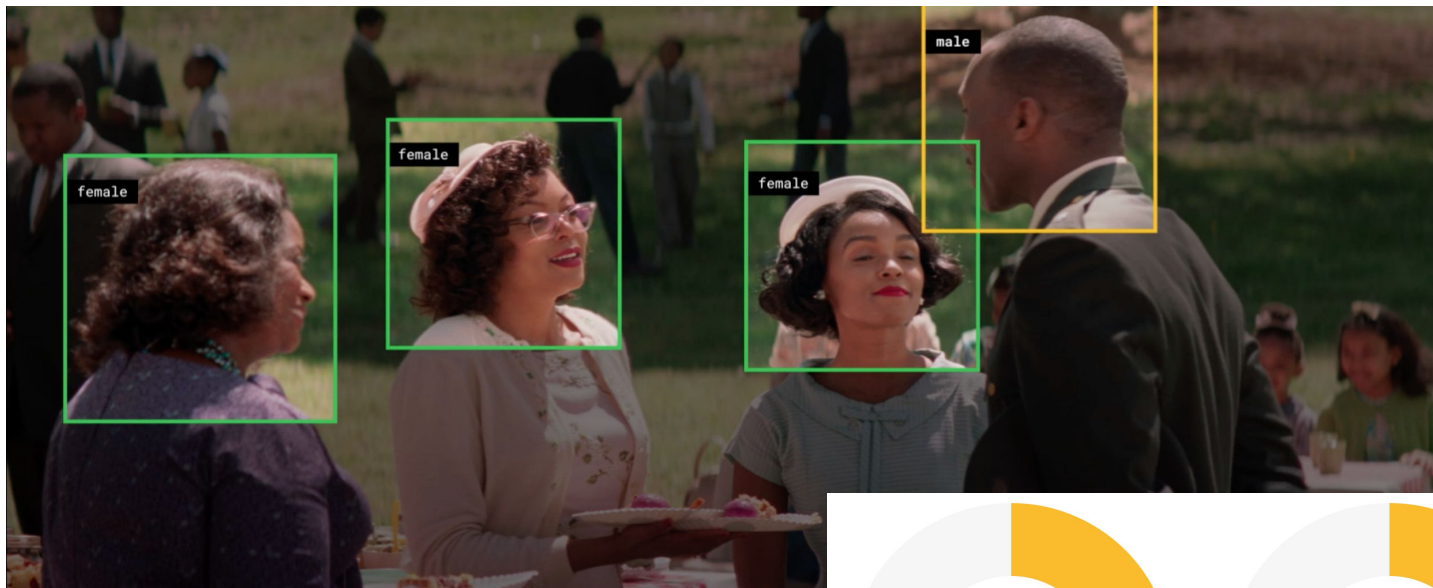
Aplicaciones: imágenes

Predicción de pobreza usando imágenes satelitales



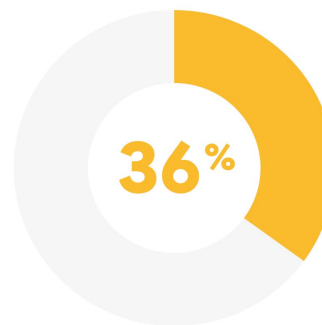
Aplicaciones: videos

Medición de tiempo en pantalla de mujeres en películas

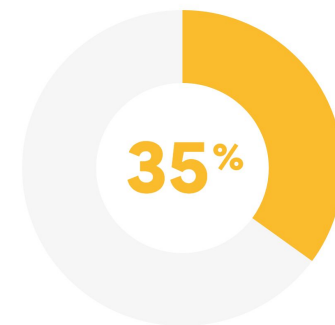


Tiempo de mujeres
en pantalla

Tiempo de
diálogo de
mujeres



Female on-screen time



Female speaking time

Conclusión

- No sólo los datos estructurados pueden ser analizado.
- Hay una gran disponibilidad de datos no estructurados.
- Existen múltiples técnicas para analizar distintos tipos de datos no estructurados.
- Hoy en día se están haciendo grandes avances en esta línea.

Consejo para economistas en formación: aprendan y practiquen mucho programar y manejar estructuras de datos eficientes.

¡Gracias!

rgalvez@dc.uba.ar