

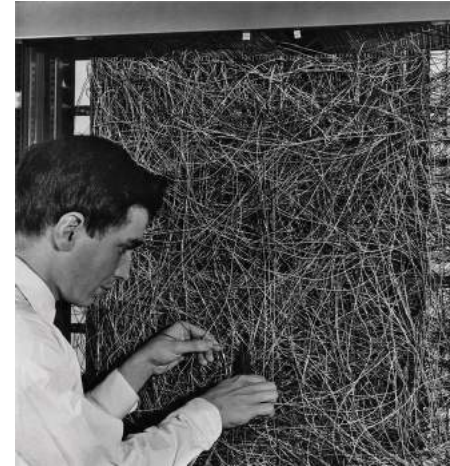
Introducción a deep learning

antecedentes

Máquinas electrónicas

Mark I, 1944

1 operación cada 3 segundos



Máquinas mecánicas

Mecanismo de Anticitera 200 a. C



La teoría

Frank Rosenblatt 1957

Alan Turing 1950

antecedentes

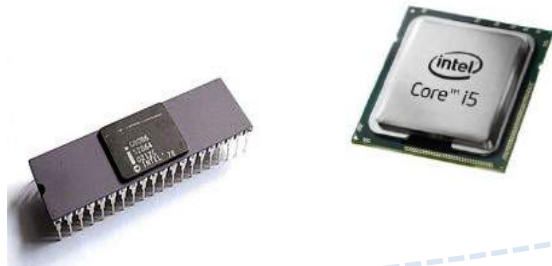
Los microprocesadores

Intel 8086, 1978

50 mil operaciones por segundo

Intel i5, 2018

25 mil millones de operaciones por segundo



2010s La era de las GPUs

Playstation 4s, 2016

1.8 TFlops (~90 x intel i5)

Playstation 5s, 2020

10.2 TFlops (~411 x intel i5)

Nvidia RTX Titan, 2018

16 TFlops (~640 x intel i5)

Nvidia RTX 3090, 2020

35 TFlops (~1400 x intel i5)



antecedentes

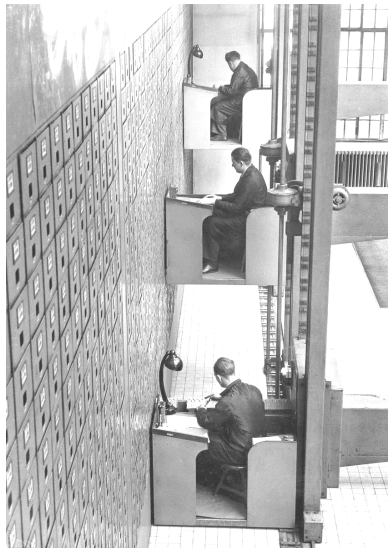
Velocidad de almacenamiento

Disco duro 2000 18GB (48MB/s)
HD estado sólido 2021 1TB (7000 MB/s)



Almacenamiento

Sistema mecánico 1937 (República Checa)



Capacidad de almacenamiento

Cinta perforada 1970 <1 KB
Disco 3 1/2 1987 1.4 MB
DVD 1995 4.7 GB



antecedentes



*Proceedings paper 1995 /
Revistas paper*



Buscadores internet 1998



TensorFlow

PYTORCH

*Software gratuito
y
Toolkits 2010*

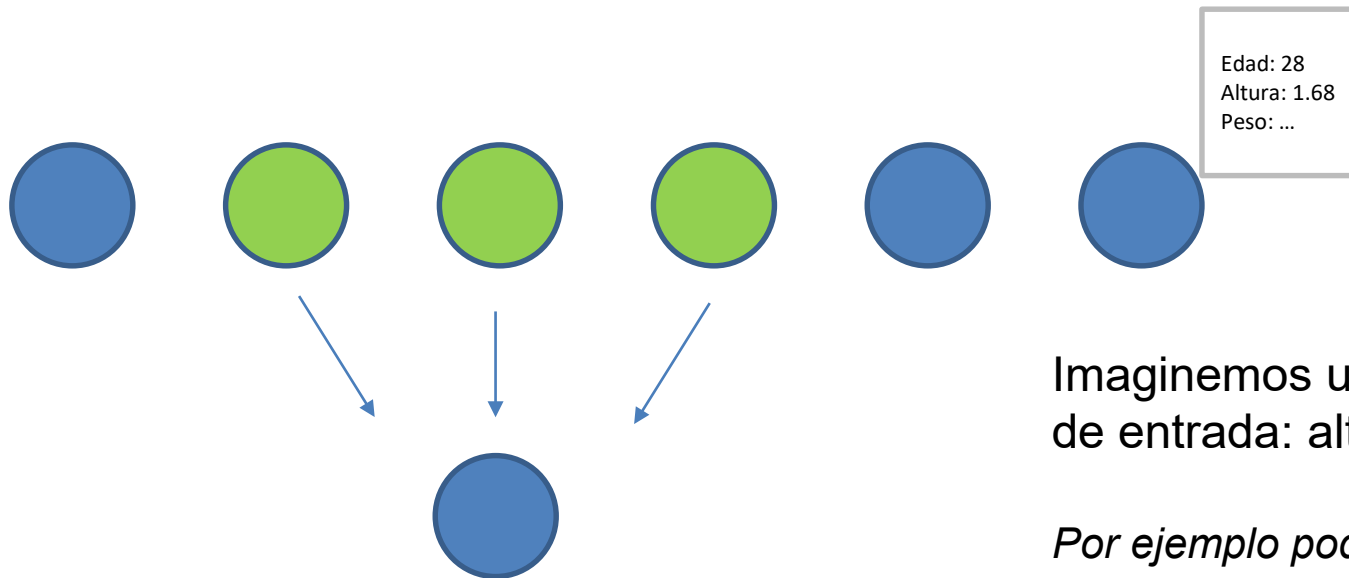


GitHub

*2008 Redes sociales /
plataformas de desarrollo colaborativo*

fundamentos

- Procesado en fases, capas

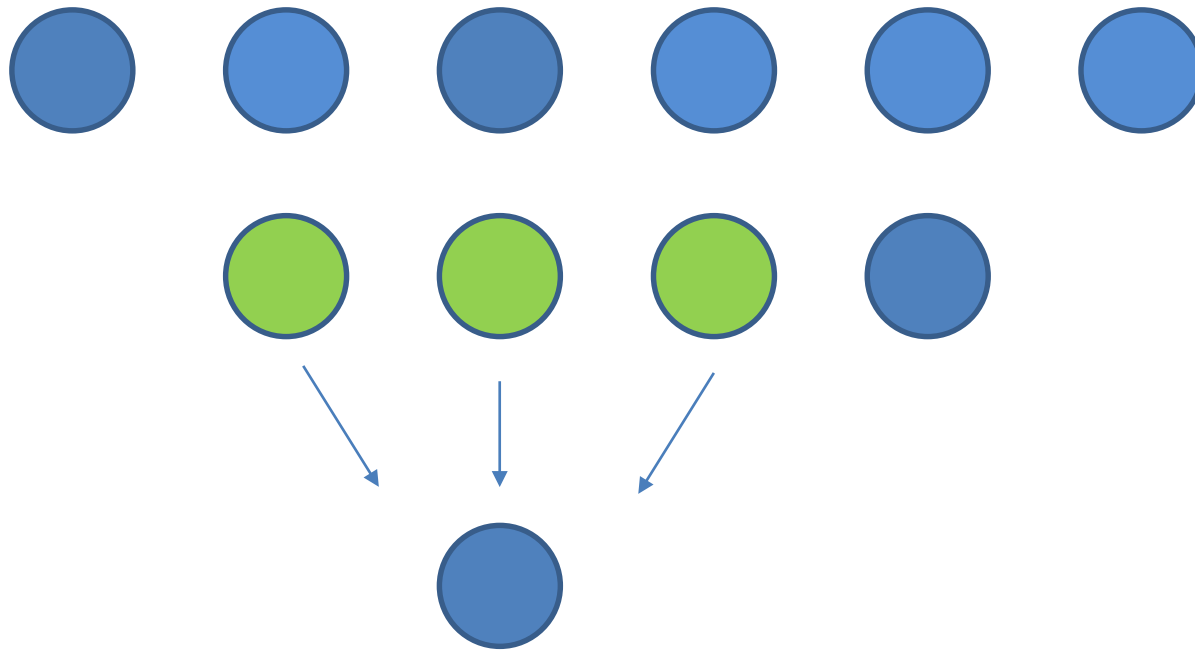


Imaginemos unos datos de entrada: altura, edad,..

Por ejemplo podríamos decir: Pregunta a los tres que tengas en la fila de delante y quédate con el máximo, mínimo, etc

fundamentos

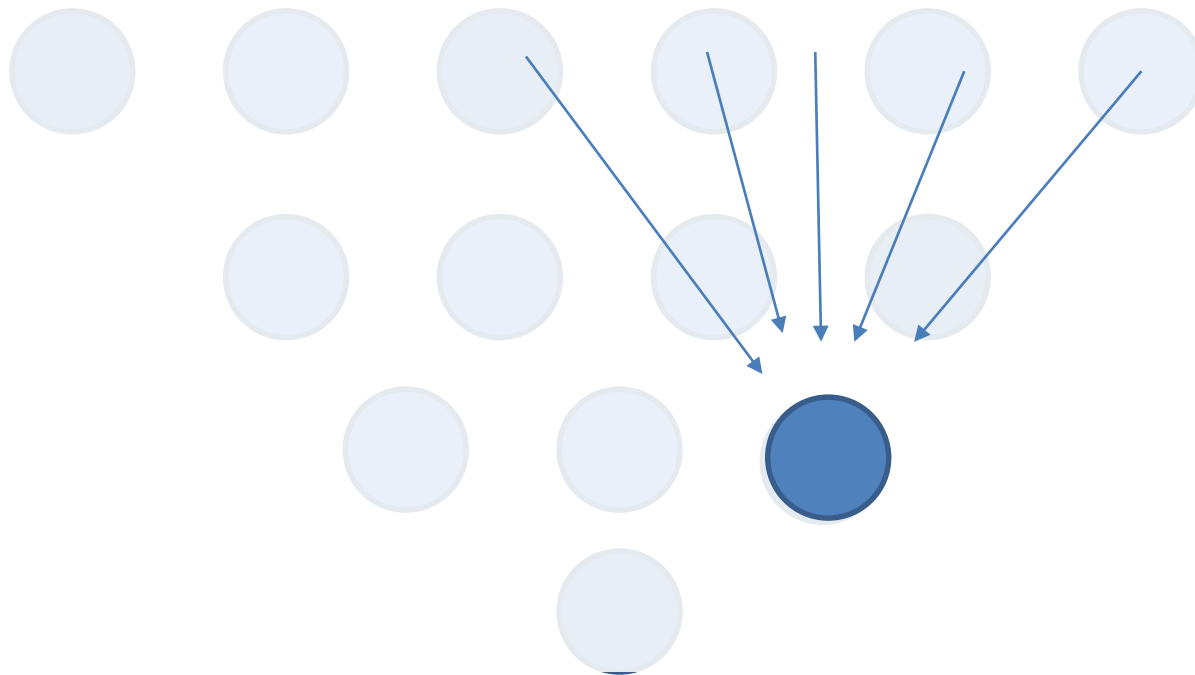
- Procesado en fases, capas



Y repetimos,
en todas las
filas...

fundamentos

- Procesado en fases, capas

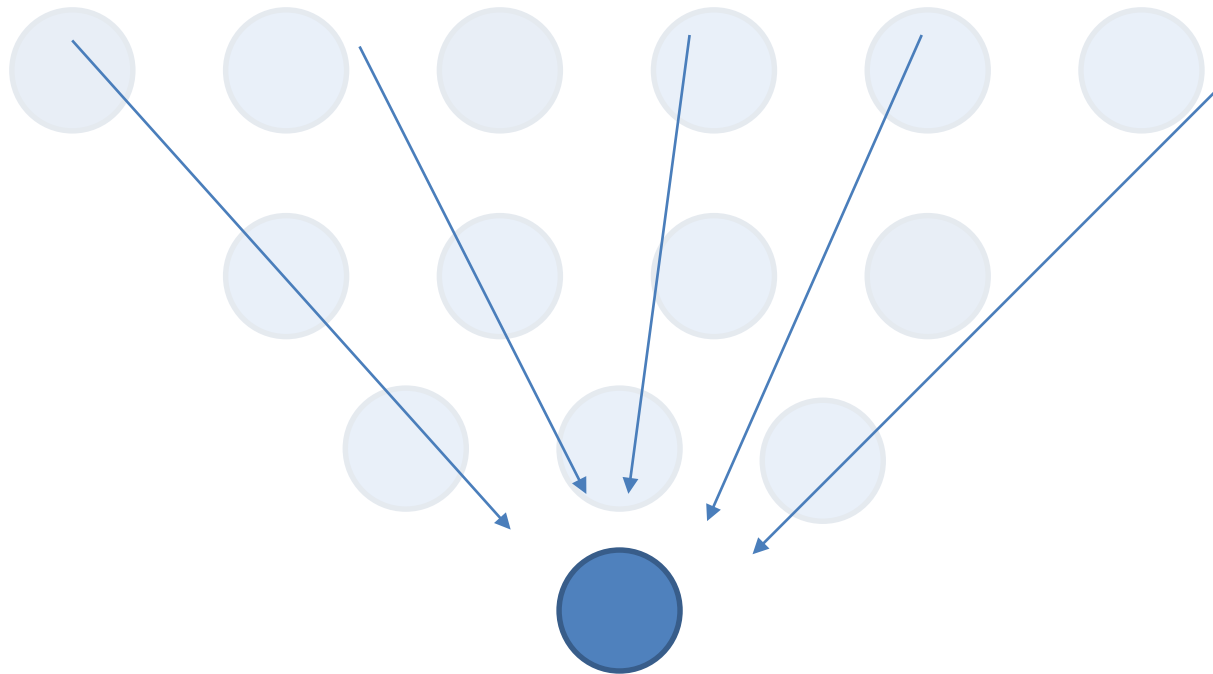


Algunos nodos

reciben
parte de la
información

fundamentos

- Procesado en fases, capas

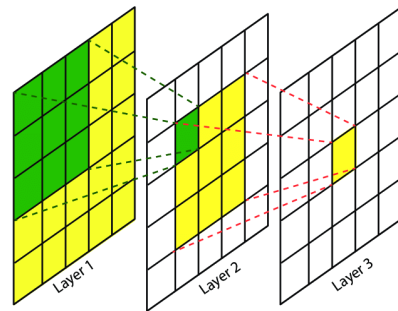


Ahora al último le podríamos preguntar, **¿quién es el más joven ?** Ha recibido **toda** la información

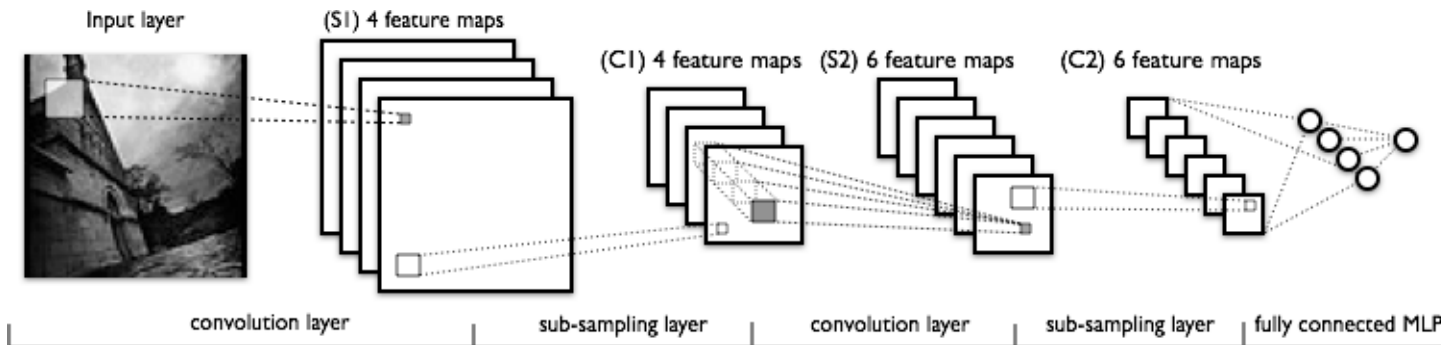
fundamentos

- Redes convolucionales

- Cada capa suma varios valores de entrada con distinto peso, normalmente 9 entradas: 3×3



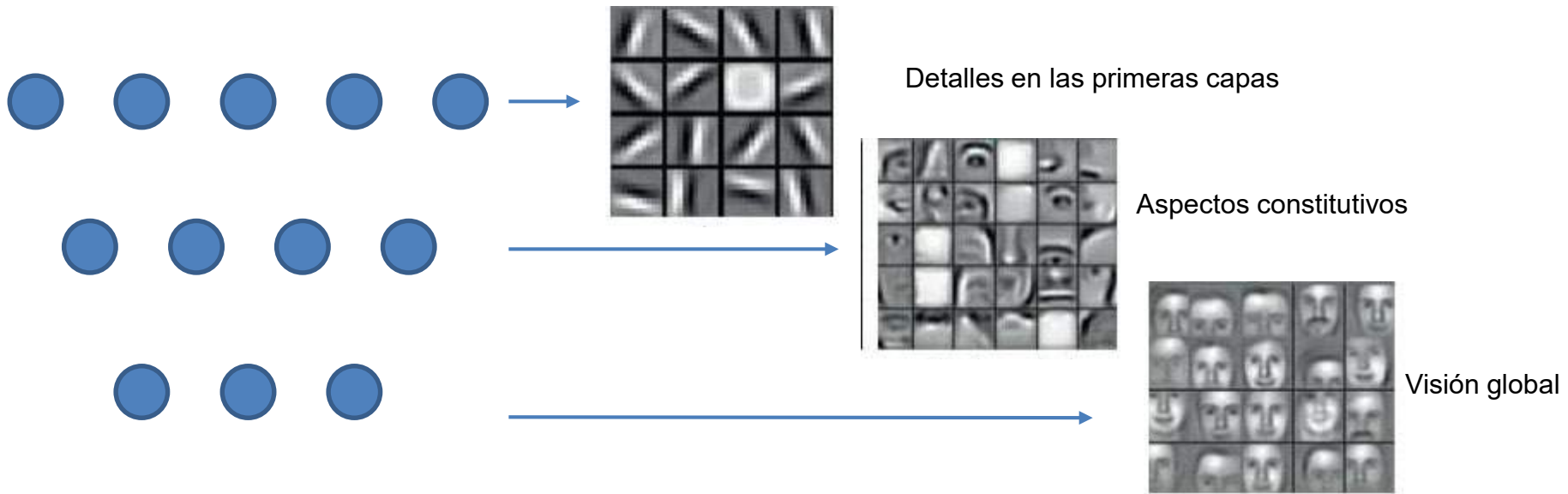
Yann LeCun



LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11), 2278-2324.

fundamentos

- Redes convolucionales
- Con mayor **profundidad** (*depth*) se logra mayor **abstracción**
 - Las primeras redes profundas tenían 7 capas
 - Hoy en día en cuestión de minutos se tiene acceso a redes de más de 100 capas ya entrenadas



fundamentos

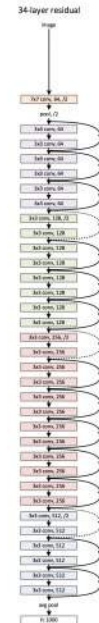
- **Redes convolucionales**

- *Las primeras redes profundas tenían 7 capas*
- *Hoy en día en cuestión de minutos se tiene acceso a redes de más de 100 capas ya entrenadas*

Conv	96 (11x11) dw: 4
Maxpool	(3 x 3) dw: 2
Conv	256 (5x5)
Maxpool	(3 x 3) dw: 2
Conv	384 (3x3)
Conv	384 (3x3)
Conv	256 (3x3)
Linear	4096
Linear	4096
Linear	1000
Softmax	



Geoffrey Hinton



2012: 7 capas
84.6 % aciertos

2014: Inception 25 capas
93.3% aciertos

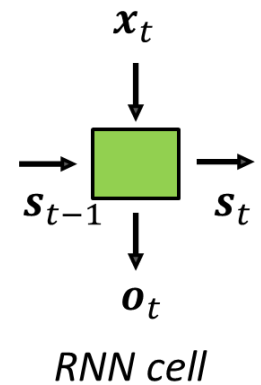
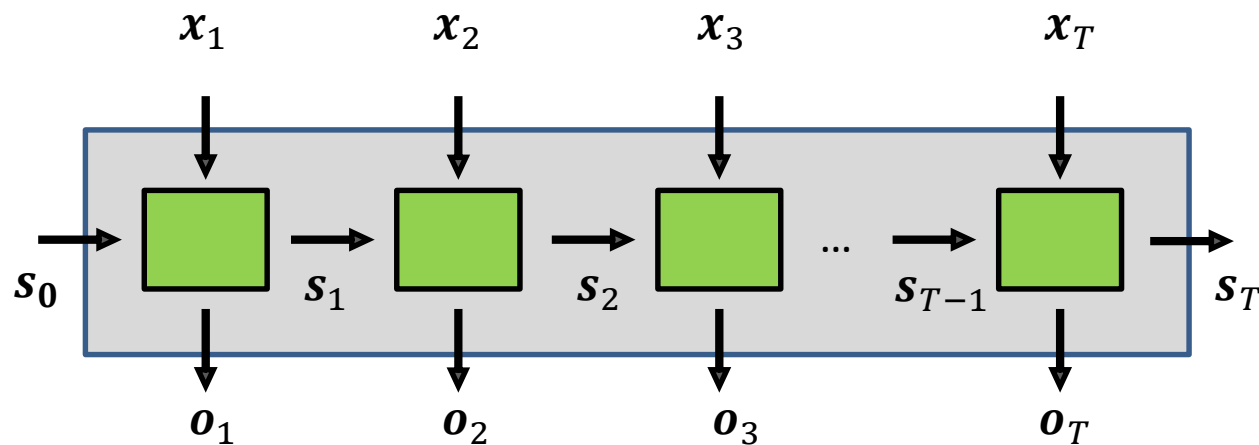
2015: Resnet >100 capas
96.43% aciertos

fundamentos

- **Redes recurrentes**

- Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8), 1735-1780

- *Analizan la entrada en orden (sentido temporal, orden del texto)*
- *Cada celda tiene una memoria finita para recibir información de los instantes previos y escribir nueva información para el futuro*



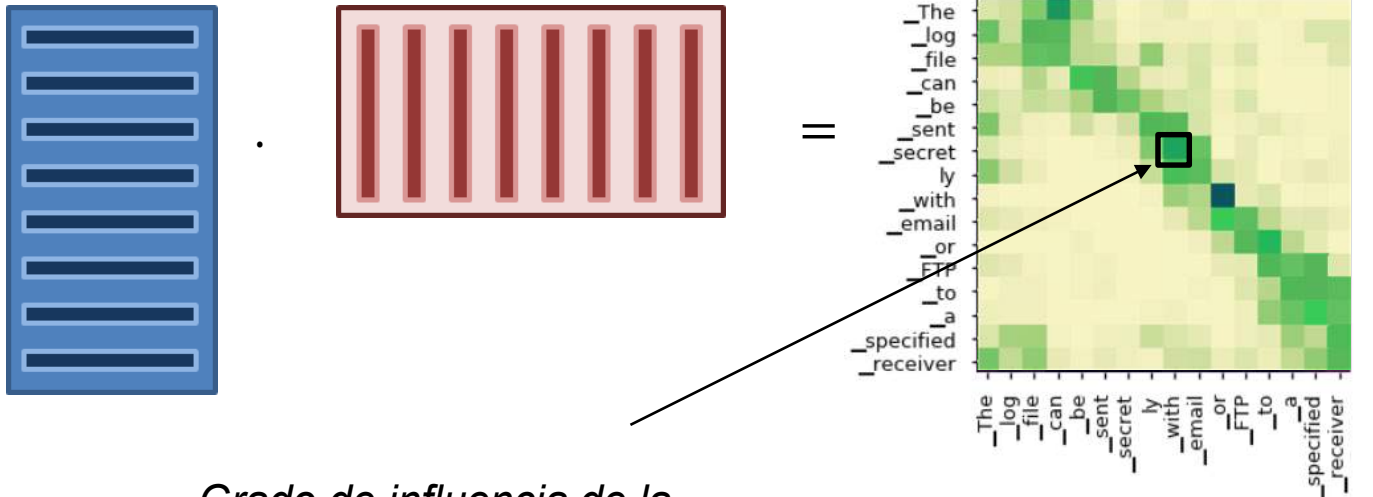
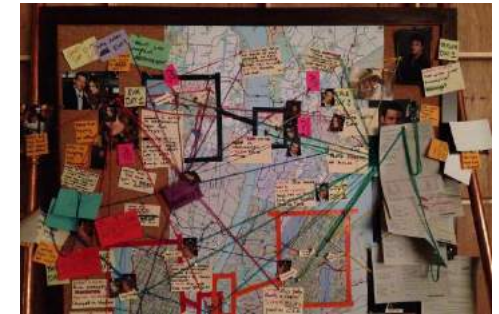
RNN

fundamentos

- Transformers

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N. Kaiser L, Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30, 5998-6008

- Son capaces de analizar la relación de todas las entradas



Grado de influencia de la entrada i sobre la j

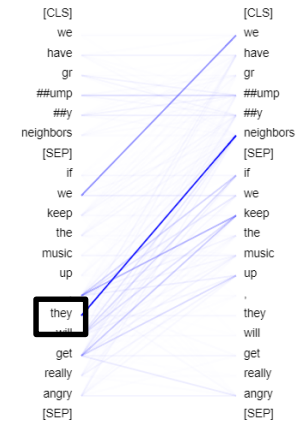
<https://nlp.seas.harvard.edu/2018/04/03/attention.html>

RNN

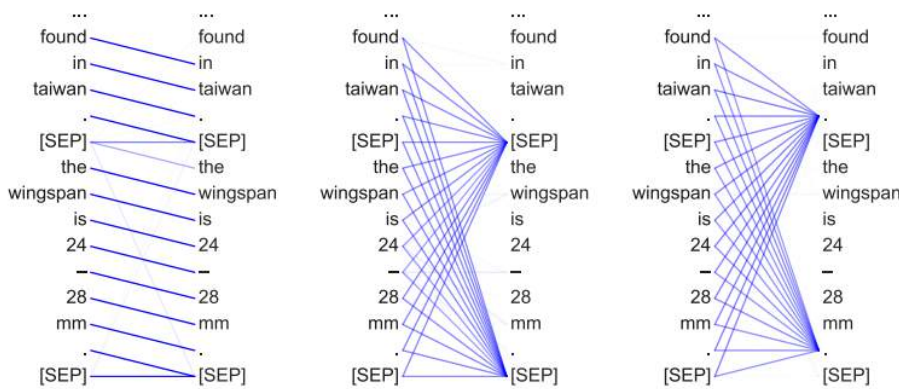
fundamentos

- Transformers

- Son capaces de analizar la relación de todas las entradas*

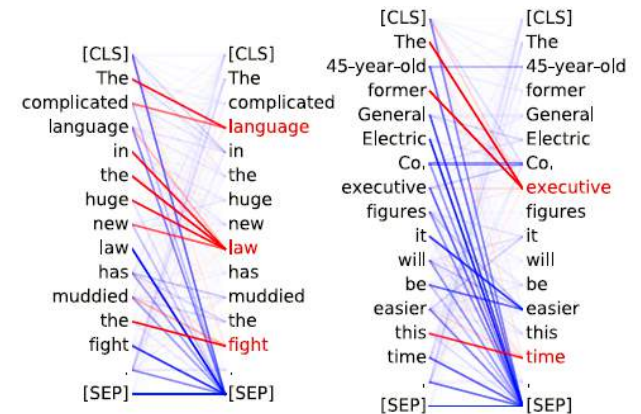


Desambiguación ellos -> vecinos



palabra anterior

Final de frase



Determinantes y modificadores de un nombre

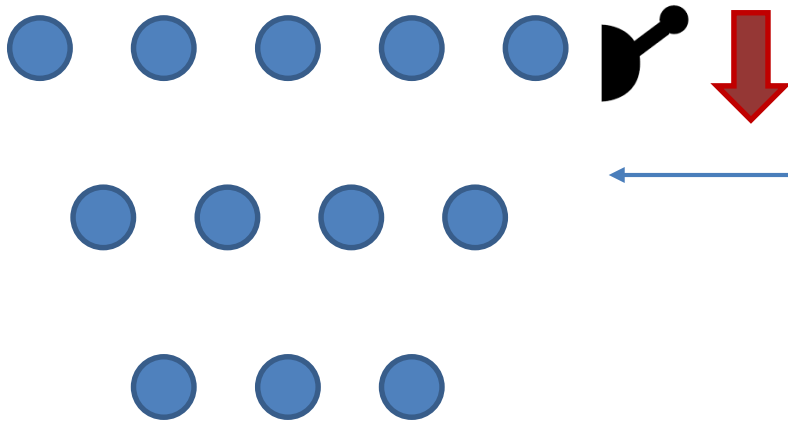
RNN

fundamentos

- **Para aprender a realizar la tarea**

Un modelo de Deep learning

Actual puede tener desde unos pocos millones de parámetros a **miles de millones!!**



Se podría probar prueba y error hasta que se encontrara alguna buena combinación de todas las palancas ... pero tardaríamos demasiado

fundamentos



- **Repetir el proceso de corrección**

- **miles de veces**



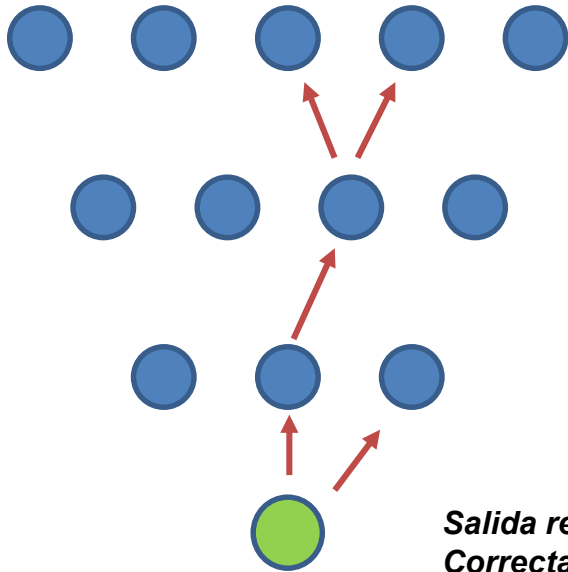
- Depende de lo complicada que sea la tarea pueden ser **millones** de correcciones

- **Hay que disponer de datos y respuestas, coste**

- Corpus, bases de datos
 - Miles o millones de ejemplos con su etiqueta

- **Problema sesgos en los datos**

- Si mostramos más veces un ejemplo y la respuesta que otros ejemplos aparecerá un sesgo en el sistema



Salida red: oso 99.9%
Correcta: gato

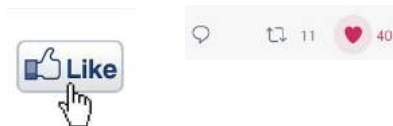
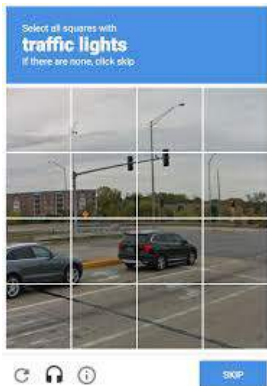
fundamentos

- **Etiquetado de datos**

- ¿Quién etiqueta?
- Freelance,
 - Mechanical turk amazon
- Empresas de etiquetado de datos
- **Todos nosotros**

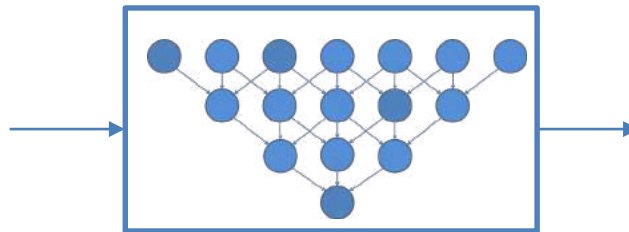
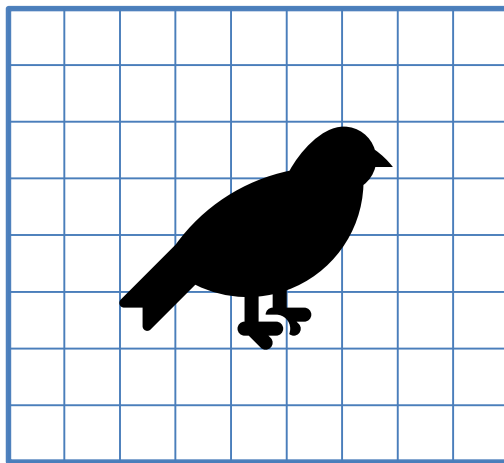


<https://time.com/5518339/china-ai-farm-artificial-intelligence-cybersecurity/>



Tipos de problemas (1 / 4)

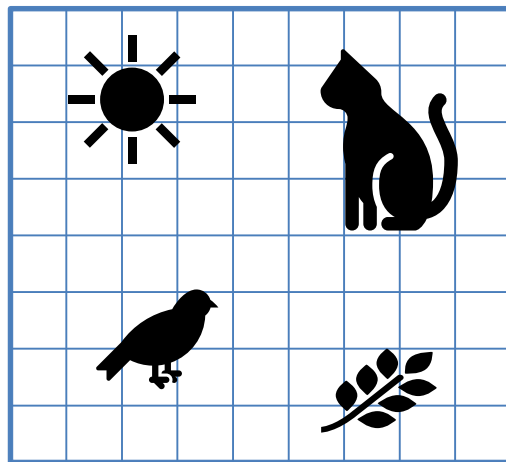
- Aunque cada día hay más variantes las dos formas principales de usar DNNs hoy en día:
 - **Clasificación:**
 - Decir **qué concepto** hay en una imagen/texto/audio



Una sola salida/concepto para toda la entrada

Tipos de problemas (2 / 4)

- Aunque cada día hay más variantes las dos formas principales de usar DNNs hoy en día:
 - **Clasificación múltiple:**
 - Decir **qué concepto** hay en cada zona/pixel: imagen/texto/audio
 - Decir **varias propiedades/conceptos** de una imagen/texto/audio

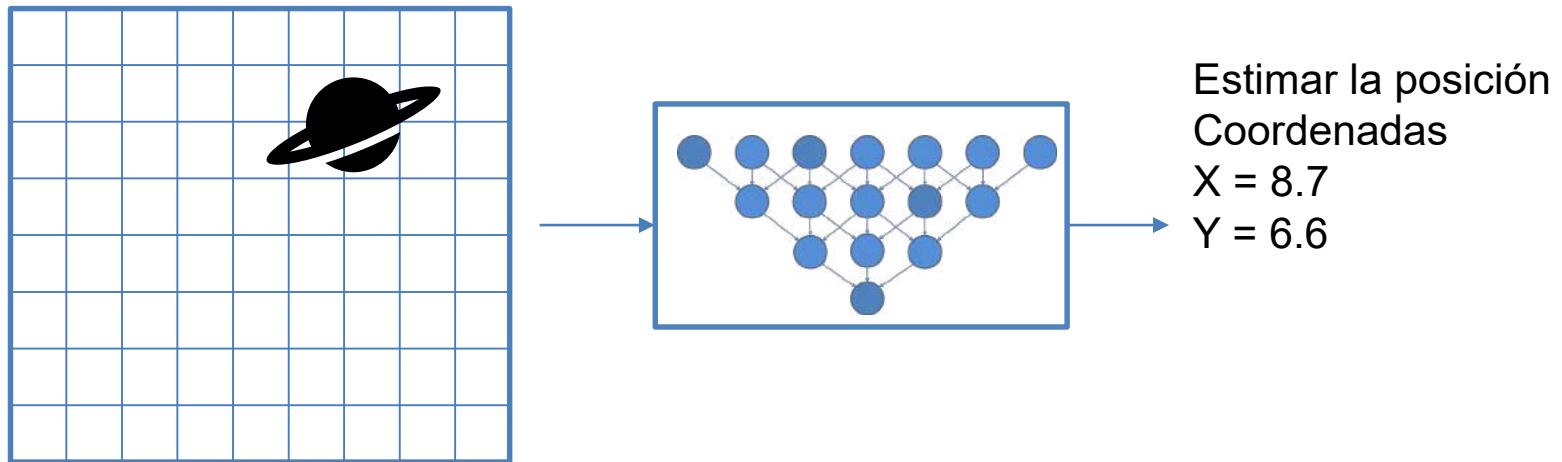


Una salida/concepto para cada pixel de entrada

sol	sol	gato	gato
sol	sol	gato	gato
pájaro			
pájaro	hoja	hoja	

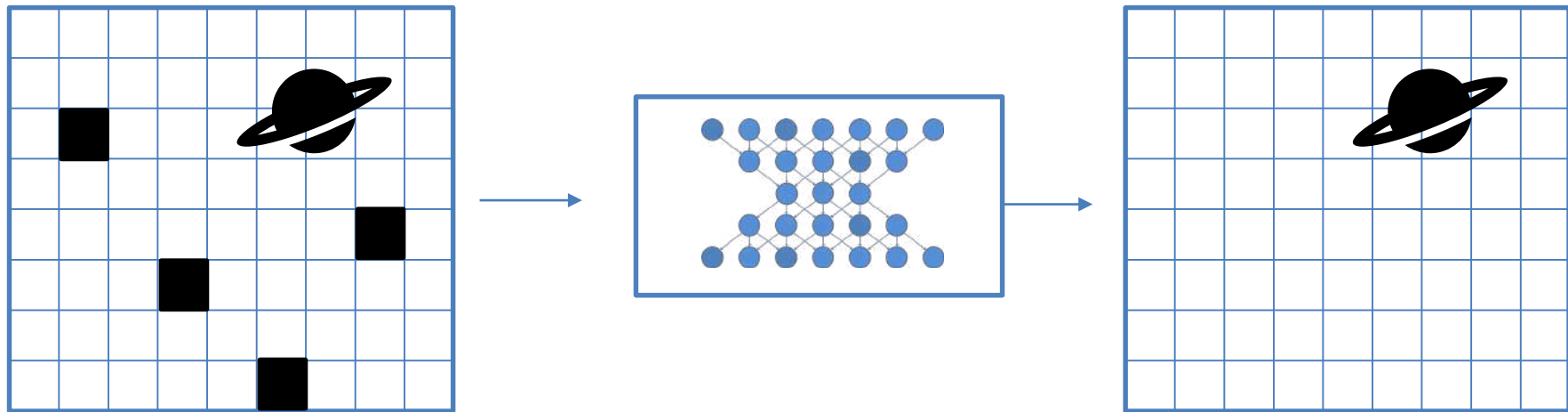
Tipos de problemas (3 / 4)

- Aunque cada día hay más variantes las dos formas principales de usar DNNs hoy en día:
 - **Regresión:**
 - Utilizar los datos para obtener algún tipo de **predicción numérica**

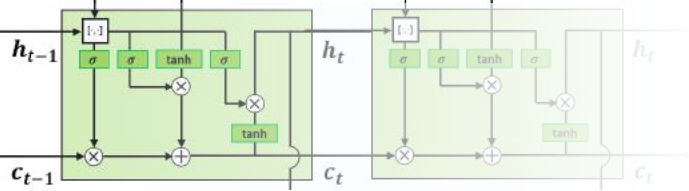


Tipos de problemas (4/4)

- Aunque cada día hay más variantes las dos formas principales de usar DNNs hoy en día:
 - Regresión múltiple:
 - Predecimos varios valores numéricos: por cada zona, pixel...

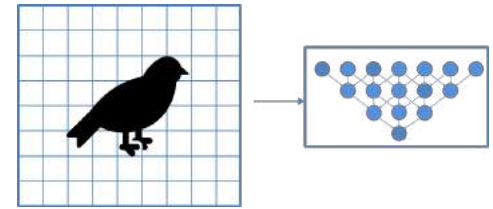


aplicaciones: análisis



– Clasificación:

- Decir **qué concepto** hay en una imagen/texto/audio



En este ejemplo la red neuronal se prepara para resolver el problema de clasificación:

¿Qué hay en esta imagen? -> 1 respuesta

Entre las posibles respuestas hay 120 razas de perro

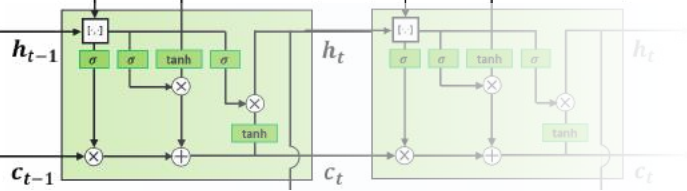
En 2012 el error top5 era del 25%,

Hinton y Krizhevsky red de 7 capas 15%

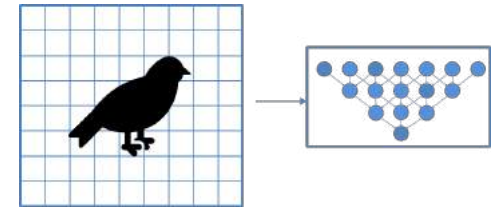
Hoy en día decenas, cientos de capas, alrededor del 2%,



aplicaciones: análisis

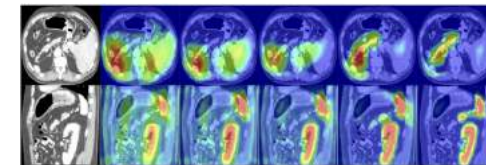


– **Clasificación:** ¿ nos podemos fiar ?

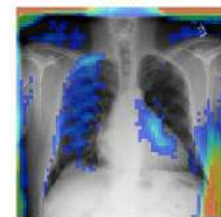


¿ cómo es ese pequeño porcentaje de fallos... ?

Hay modelos que pueden mostrar **qué zonas** han considerado más

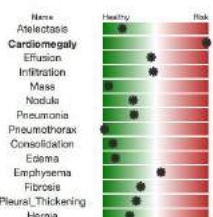


Out Of Distribution reconstruction error
Heatmap shows the image values from the training distribution.

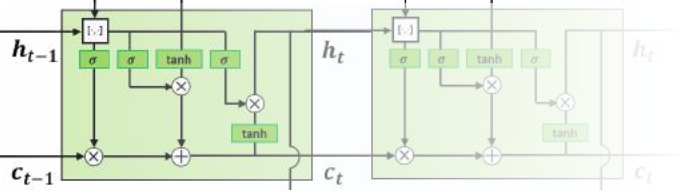


recScore: 0.27, ssim: 0.39

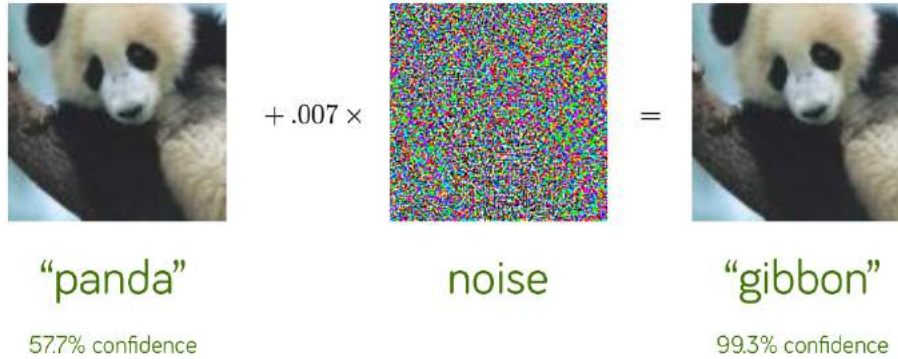
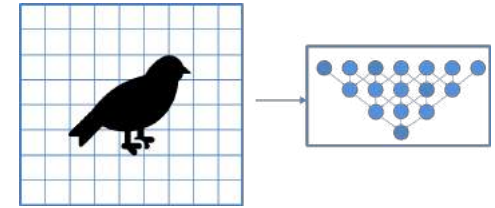
Disease Predictions
Probability of a disease.



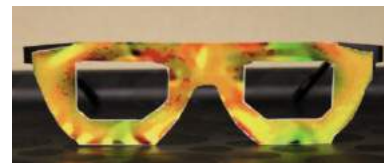
aplicaciones: análisis



– **Clasificación:** ¿ nos podemos fiar ?



Ataques adversarios



Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition

Mahmood Sharif, Sruti Bhagavatula, Lujio Bauer, Michael K. Reiter
ACM Conference on Computer and Communications Security (CCS 2016)

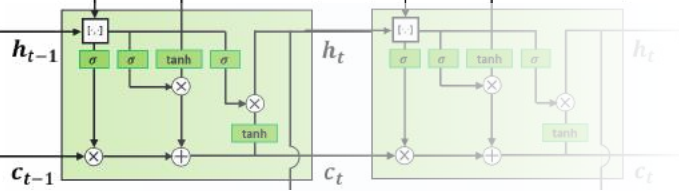


Speed Limit 80
(88% confidence)

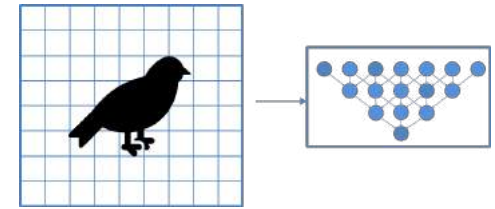
Robust physical-world attacks on deep learning visual classification.

Eykholt, K., Evtimov, I., Fernandes, E., Li, B., Rahmati, A., Xiao, C., ... & Song, D. (2018). In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1625-1634).

aplicaciones: análisis



– **Clasificación:** ¿ nos podemos fiar ?



paperswithcode.com

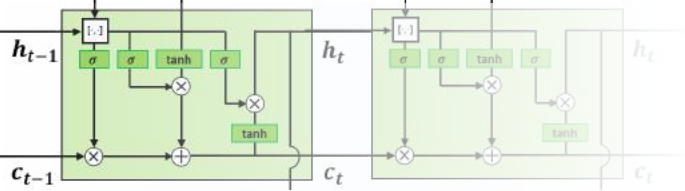


Entrenamiento

Hoy en día se entrenan facilitando múltiples versiones de las imágenes/sonidos

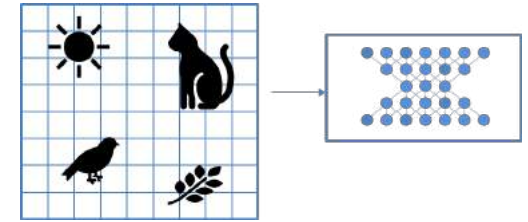
Se conoce como:
Aumento de datos

aplicaciones: análisis



– Clasificación múltiple:

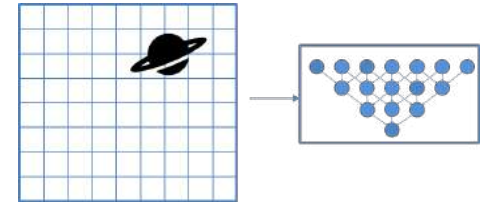
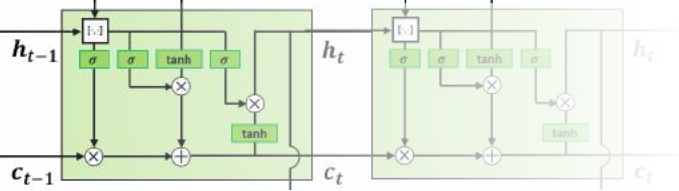
- **varias propiedades/conceptos** de una imagen/texto/audio



En este ejemplo la red neuronal se prepara para resolver muchas respuestas sí o no:

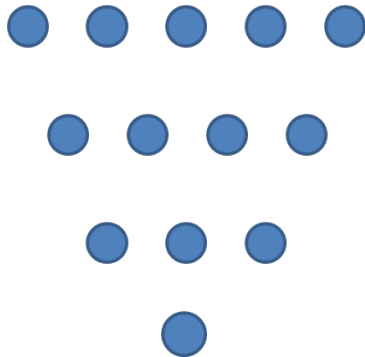
- | | |
|-----------------|----|
| ¿Hay un perro? | No |
| ¿Hay un gato? | Sí |
| ¿Hay árboles? | No |
| ¿Hay un pájaro? | Sí |
| ¿Hay cielo? | No |
| ¿Hay hierba? | Sí |

aplicaciones: análisis



– Regresión:

- Utilizar los datos para obtener algún tipo de **predicción** numérica



Edad?

En este ejemplo la red neuronal se prepara para resolver el problema:

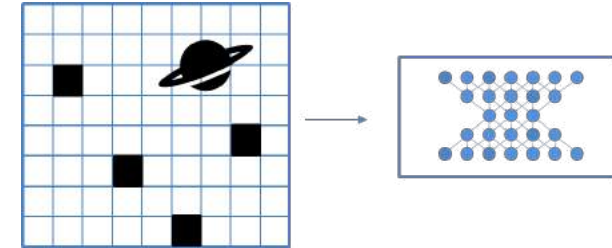
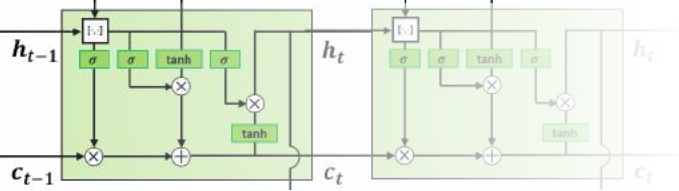
¿Qué edad tienen estas persona?

La respuesta sería un número con la edad en años

Entrenaríamos el sistema con muchas imágenes

Aplicaríamos las correcciones necesarias cuando la red se equivoca

aplicaciones: síntesis



– Regresión múltiple:

- **Transformar los datos** con alguna finalidad, que se parezcan a algo, que mejoren de calidad...



En este ejemplo la red neuronal se prepara para resolver el problema:

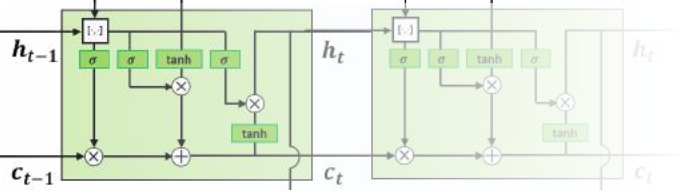
Convertir una imagen de BN en color



En este ejemplo la red neuronal se prepara para resolver el problema:

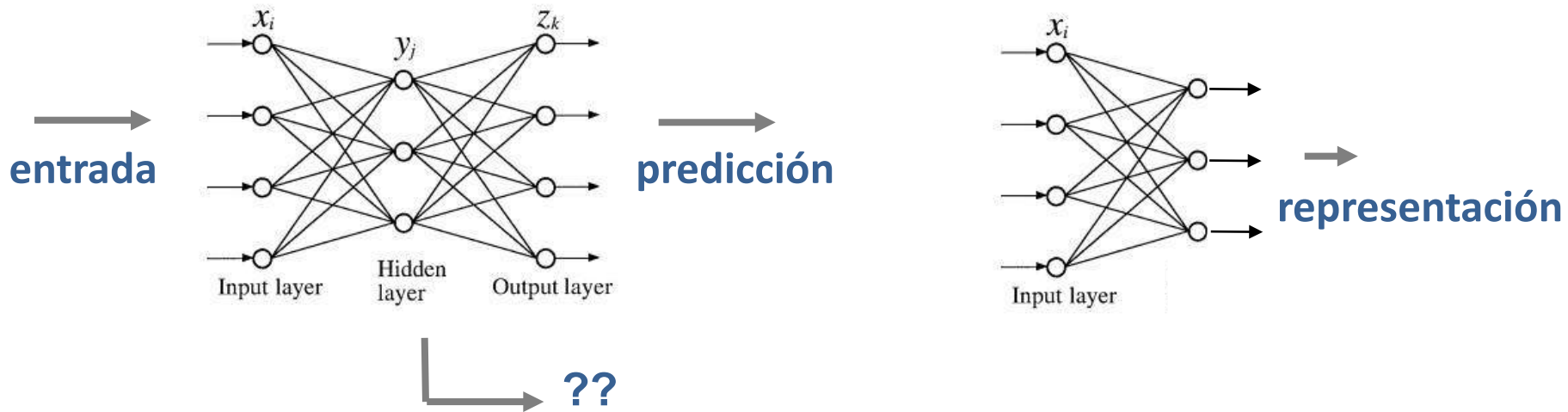
Mejorar la calidad de la imagen

aplicaciones: análisis

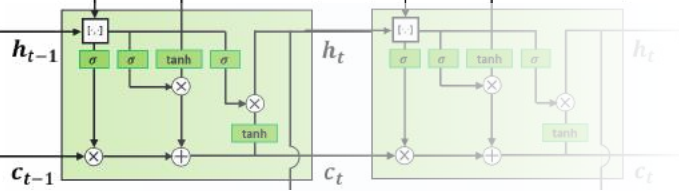


■ Representation learning

- Podemos utilizar representaciones internas de la red
 - Objetivo comparar imágenes/sonidos/textos



aplicaciones: análisis

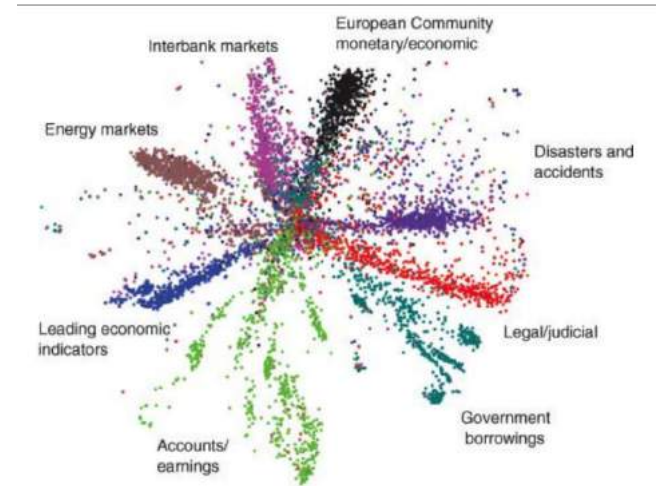


■ Representation learning

- Podemos utilizar representaciones internas de la red
 - imágenes/sonidos/textos similares están más próximos en ese espacio

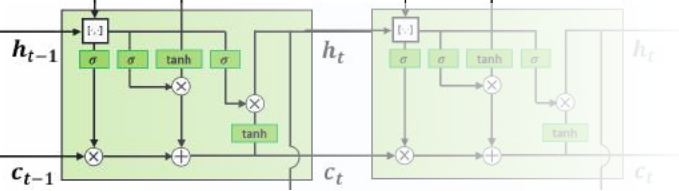


*Handwritten digits,
A. Karpathy, Stanford University*



*Text topic classification
G. Hinton, Toronto University*

aplicaciones: análisis



Imagenet classification with deep convolutional neural networks

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012).. *Advances in neural information processing systems*, 25.

■ Representation learning

- Podemos utilizar representaciones internas de la red

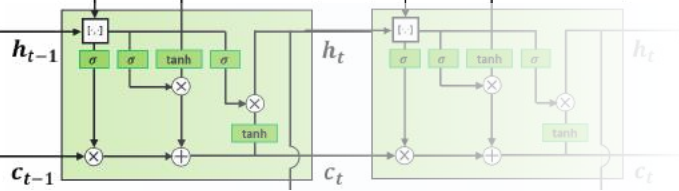


Buscar imágenes similares ...



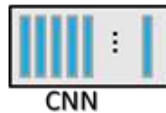
Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25.

aplicaciones: análisis

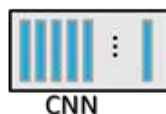


■ Representation learning

- Podemos utilizar representaciones internas de la red

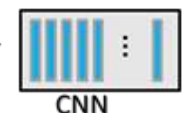


CNN



CNN

Comparar si dos imágenes corresponden a la misma identidad



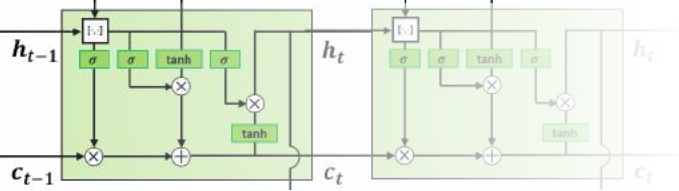
CNN



CNN

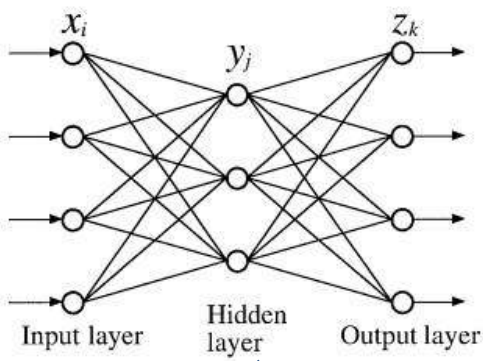


aplicaciones: análisis



■ Generación

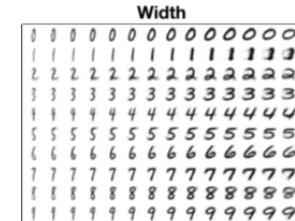
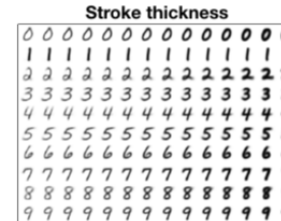
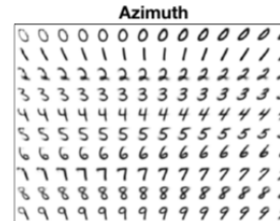
- Podemos aprender a manipular las imágenes



- ¿Qué pasa si cambio la representación para conseguir otra imagen distinta?

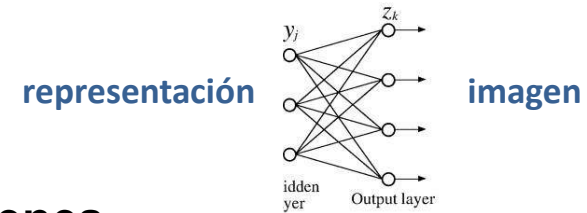
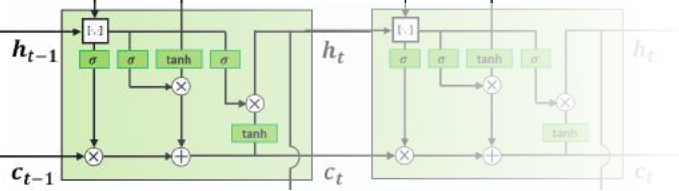


representación



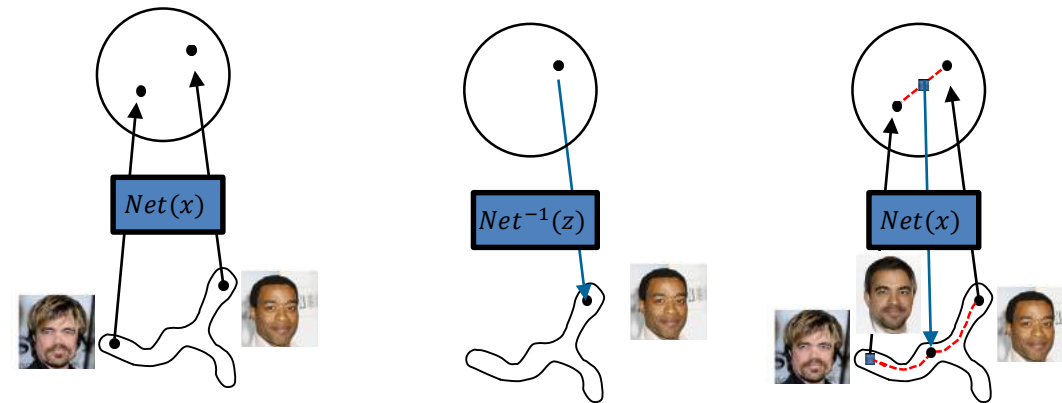
Antoran, J., & Miguel, A. (2019, December). Disentangling and Learning Robust Representations with Natural Clustering. In 2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA) (pp. 694-699). IEEE.

aplicaciones: síntesis



■ Generación

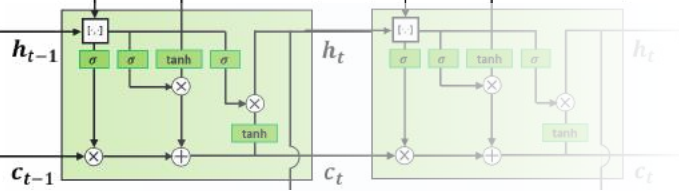
- Podemos aprender a manipular las imágenes
 - ¿Qué pasa si cambio la representación para conseguir otra imagen distinta?



Generación de nuevas imágenes que nunca han existido

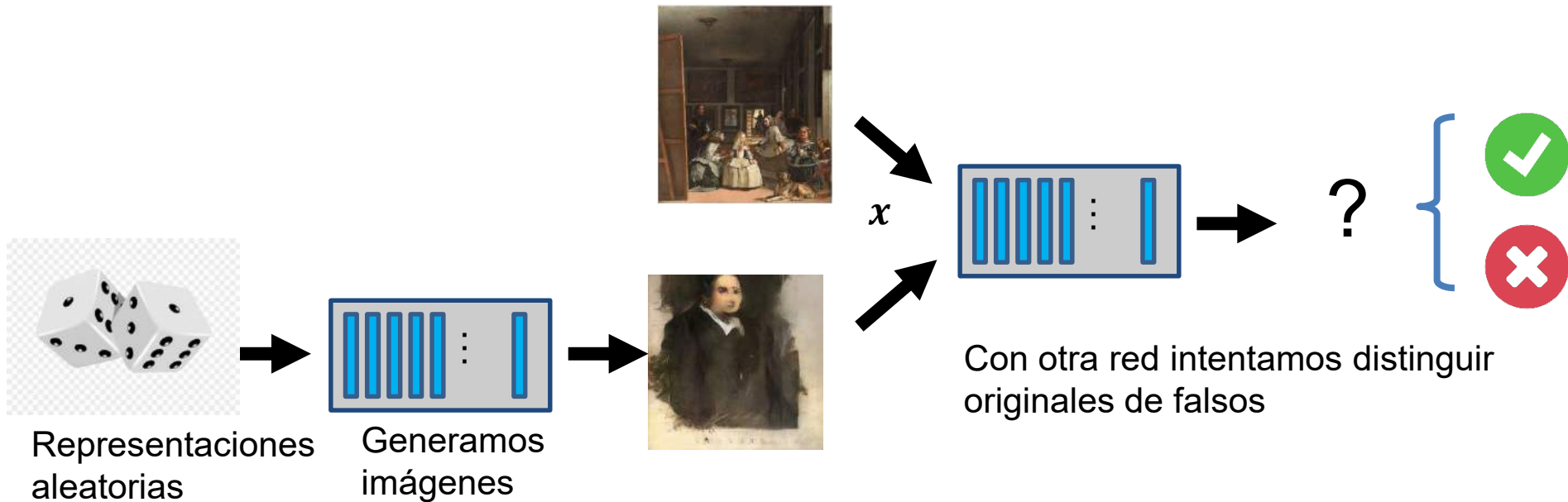
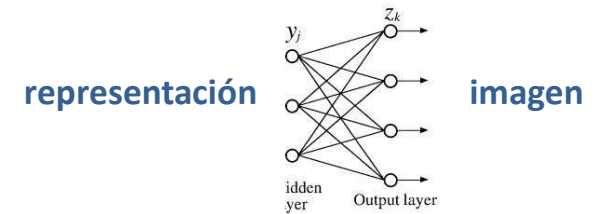
Kingma, D. P., & Dhariwal, P. (2018). Glow: Generative flow with invertible 1x1 convolutions. In Advances in neural information processing systems (pp. 10215-10224). Kingma, D. P., & Dhariwal, P. (2018). Glow: Generative flow with invertible 1x1 convolutions. In Advances in neural information processing systems (pp. 10215-10224).

aplicaciones: síntesis

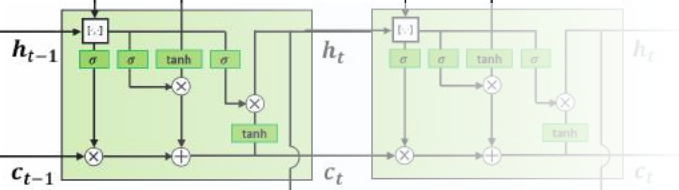


■ Generación

- Hay modelos en los que directamente se aprende a generar imágenes
- **Generative adversarial network**

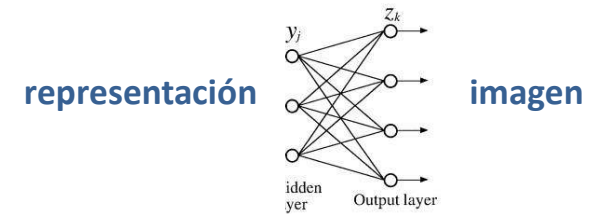


aplicaciones: síntesis



■ Generación

- Hay modelos en los que directamente se aprende a generar imágenes
- **Generative adversarial network**



¿Qué imagen es artificial?

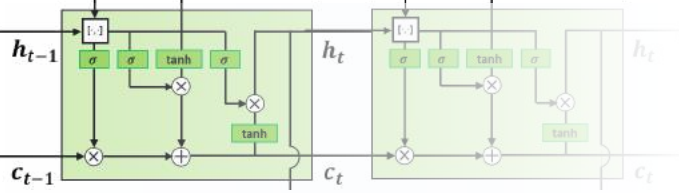
One hour of imaginary celebrities

<https://www.youtube.com/watch?v=36lE9tV9vm0>

Dos redes compiten:

- **Red 1 genera imágenes realistas de manera que la Red 2 falle**
- **Red 2 intenta distinguir las imágenes reales de las falsas**
- Es una combinación de generación y clasificación
- Genera datos de forma que sean indistinguibles de los originales
- Generador ilimitado de datos: imágenes, audio, texto...

aplicaciones: síntesis



■ Generación

- Este proceso se ha sofisticado mucho en menos de 10 años

Goodfellow et al., 2014; Radford et al., 2016; Liu & Tuzel, 2016; Karras et al., 2018; Karras et al., 2019; Goodfellow, 2019; Karras et al., 2020, Karras 2021



2014



2015



2016



2017



2018



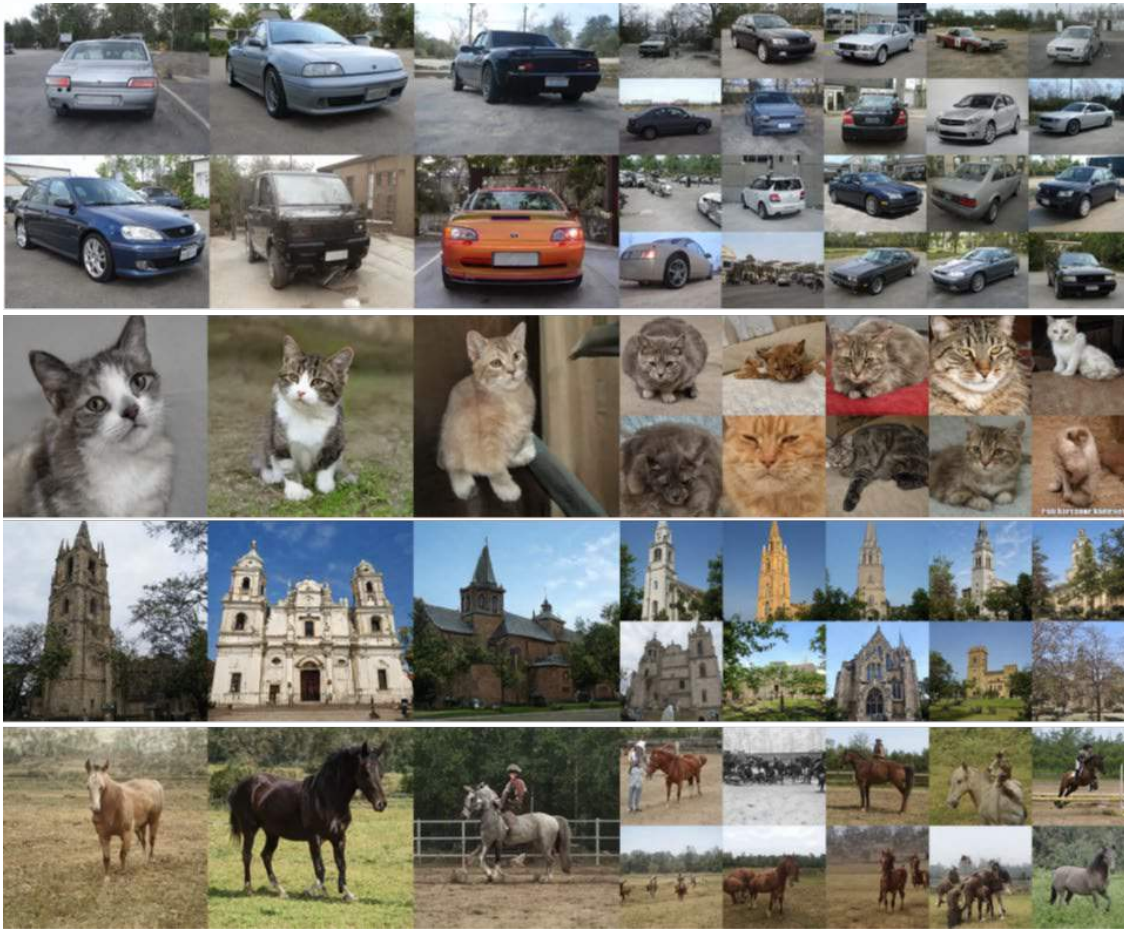
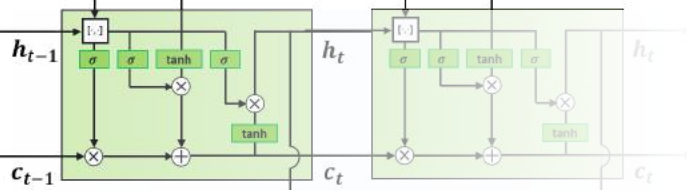
2020



2021

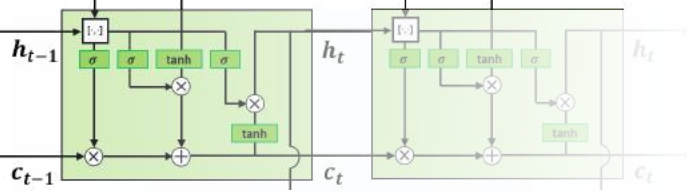
StyleGAN3 (Karras 2021)

aplicaciones: síntesis



StyleGAN2 (Karras 2020)

aplicaciones: síntesis



Monet ↔ Photos

Zebras ↔ Horses

Summer ↔ Winter



Monet → photo



zebra → horse



summer → winter



photo → Monet



horse → zebra



winter → summer



Photograph



Monet



Van Gogh



Cezanne

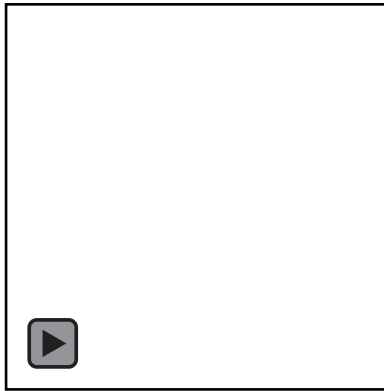


Ukiyo-e

<https://junyanz.github.io/CycleGAN/>

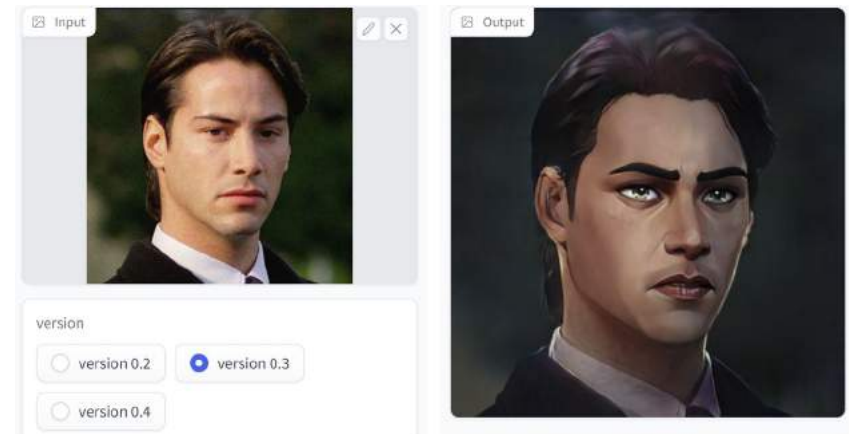
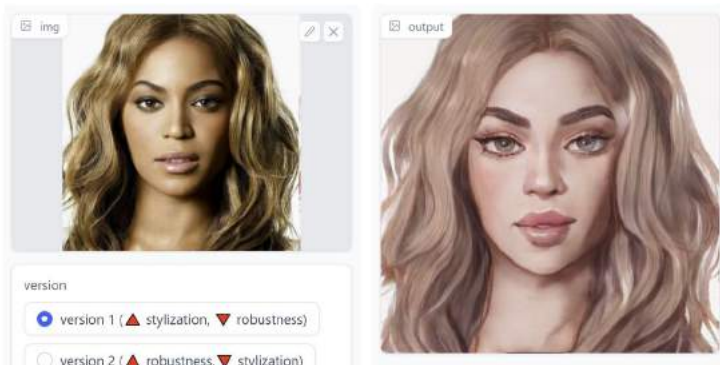
aplicaciones: síntesis

- Generación de vídeos realistas: Deep fakes



aplicaciones: síntesis

- Modificación de estilos



<https://huggingface.co/spaces/akhaliq/AnimeGANv2>

<https://huggingface.co/spaces/akhaliq/ArcaneGAN>

aplicaciones: síntesis

Ho, J., Jain, A., & Abbeel, P. (2020). Denoising diffusion probabilistic models. arXiv preprint arXiv:2006.11239

■ Generación:

Podemos añadir ruido hasta que no se reconozca la imagen



Con una red aprendemos a “limpiar” ese ruido



Son capaces de generar imágenes tan realistas como los GANs



unreal engine gaudi house in a field of poppy



unreal engine building by gaudi

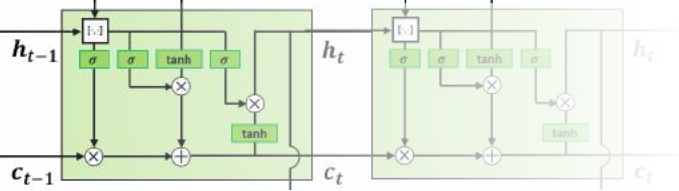


the angel of air. unreal engine @arankomatsuzaki



treehouse in the style of studio ghibli animation @danielrusruss

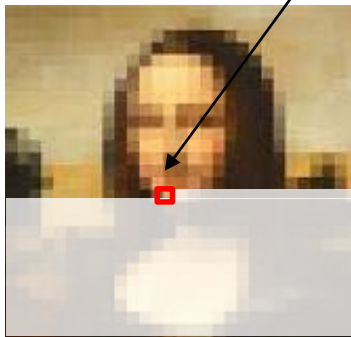
aplicaciones: síntesis



▪ Aprendiendo sin etiquetas: no supervisado

- Podemos conseguir que los sistemas automáticos comprendan los datos **forzando a que hagan predicciones** sobre lo que no han visto

Viendo los pixels anteriores: ¿cómo es el siguiente ?

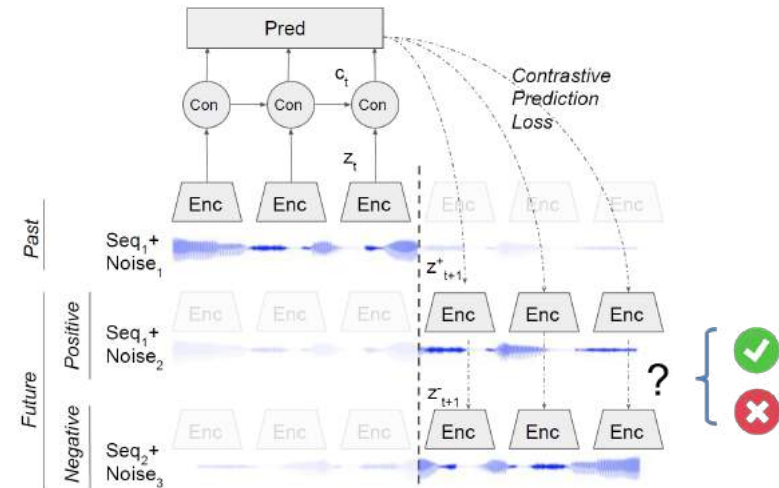
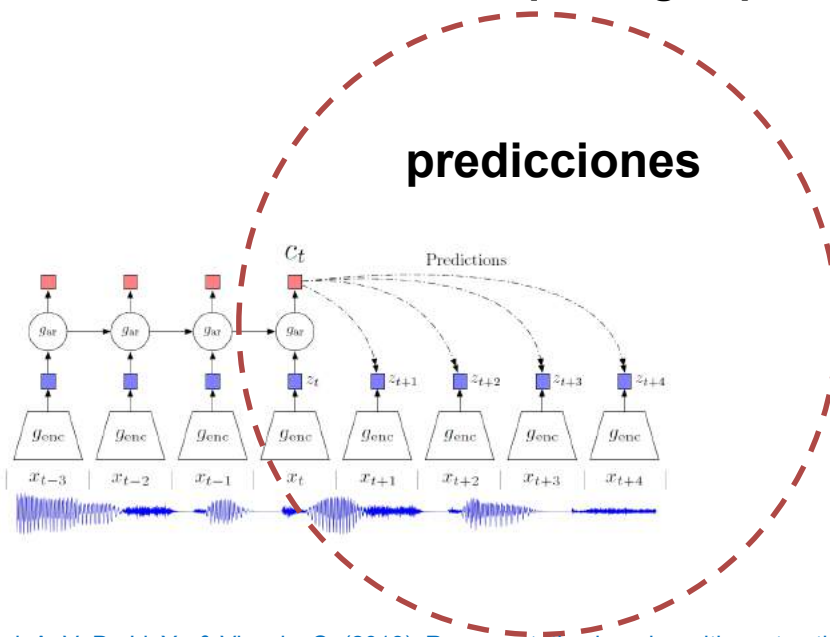


*A Oord, N Kalchbrenner, O Vinyals, L Espeholt, A Graves, K Kavukcuoglu
Conditional Image Generation with PixelCNN Decoders 2016*

aplicaciones: análisis

■ Aprendiendo sin etiquetas: no supervisado

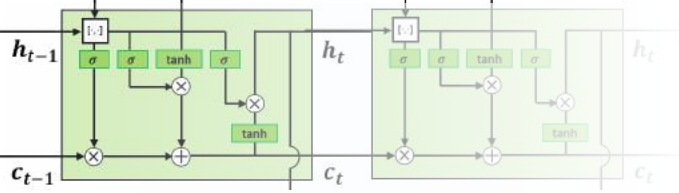
- Podemos conseguir que los sistemas automáticos comprendan los datos **forzando a que hagan predicciones** sobre lo que no han visto



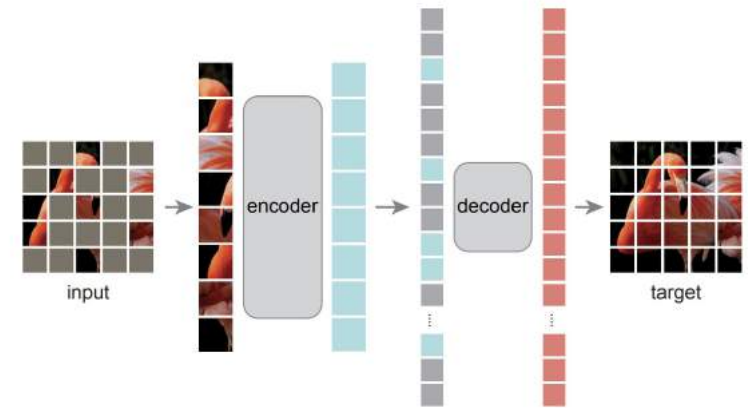
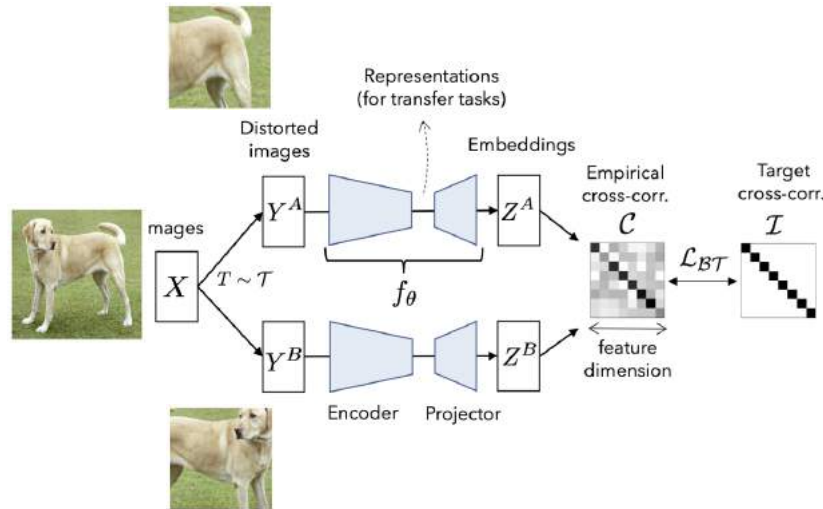
Oord, A. V. D., Li, Y., & Vinyals, O. (2018). Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748
Oord, A. V. D., Li, Y., & Vinyals, O. (2018). Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748

Una estrategia es dar varias opciones como si fuera un examen

aplicaciones: análisis



■ Aprendiendo sin etiquetas: no supervisado



Resolver la pregunta son partes de la misma imagen

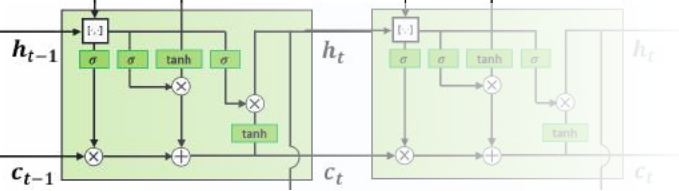
Reconstruir la imagen a partir de una con oclusiones

Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020, November). A simple framework for contrastive learning of visual representations. In *International conference on machine learning* (pp. 1597-1607). PMLR

Zbontar, J., Jing, L., Misra, I., LeCun, Y., & Deny, S. (2021, July). Barlow twins: Self-supervised learning via redundancy reduction. In *International Conference on Machine Learning* (pp. 12310-12320). PMLR.

He, K., et al (2022). Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*

aplicaciones: análisis



▪ Aprendiendo sin etiquetas: no supervisado

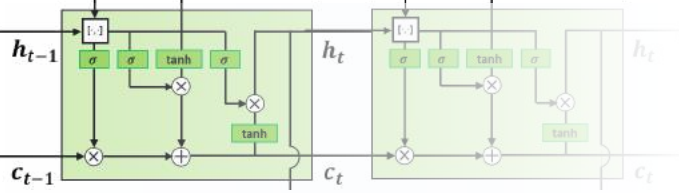
- Motivación: hay muchos datos no etiquetados
- Las representaciones obtenidas se pueden usar en otras tareas



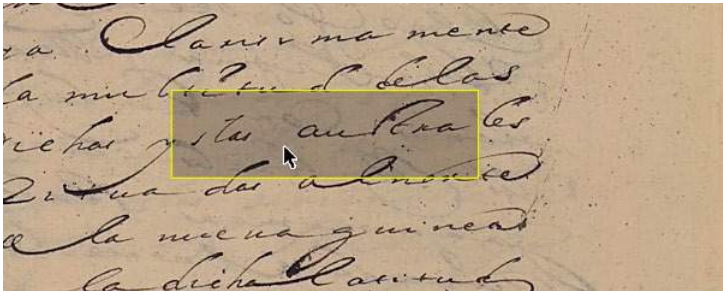
Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., & Joulin, A. (2021). Emerging properties in self-supervised vision transformers. arXiv preprint arXiv:2104.14294..



aplicaciones



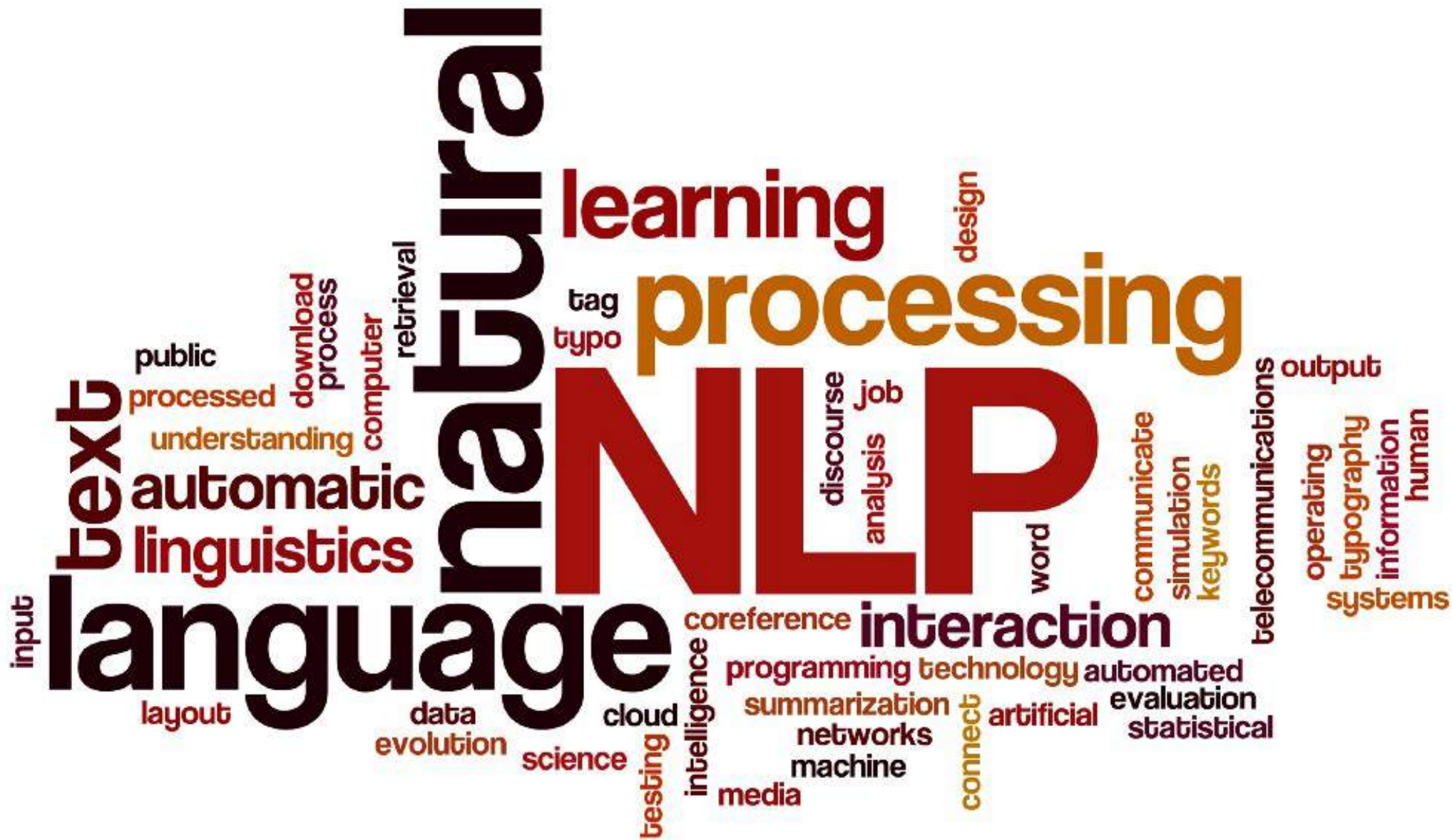
- La inteligencia artificial halla rastros del descubrimiento español de Australia



Un grupo de investigación en la UPV lleva años desarrollando sistemas de reconocimiento de texto manuscrito antiguo.

¿Se puede usar ya la tecnología?

- En un texto concreto un experto es más fiable
- La tecnología actual puede permitir **buscar**
- “escalar” un sistema básico permite hacer frente a documentos que no podrían ser tratados.
- **El Archivo General de Indias**, tiene 80 millones de páginas que no se han procesado en su totalidad.
- **Objetivo: asistir al profesional**



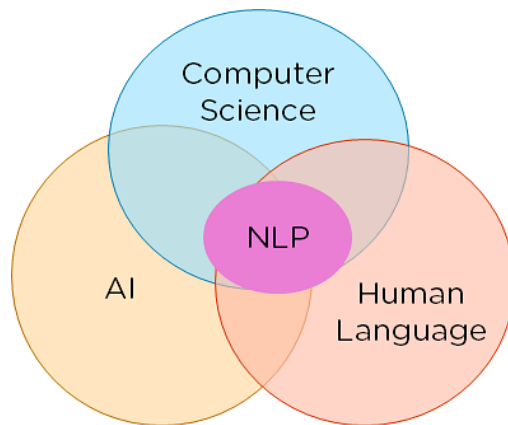
Procesamiento del Lenguaje Natural

¿Qué es el procesamiento del lenguaje natural?

- ✓ Término genérico que abarca todo aquello que permite a las máquinas *procesar* el *lenguaje humano* tanto en forma *escrita, verbal, o visual*.

¿Porqué es importante el procesamiento del lenguaje natural?

- ✓ Componente/Capacidad fundamental de los sistemas de IA.



Capacidades de un sistema de IA

- Percepción
- Aprendizaje
- Representación del conocimiento
- Razonamiento

Procesamiento del Lenguaje Natural

Procesado masivo de datos

Marc Márquez fue el más rápido en la última sesión de calentamiento de MotoGP de la temporada 2016 en Valencia, superando a Maverick Viñales por poco más de una décima de segundo.

Cantidades masivas de datos no estructurados (raw data) : texto, audio e imágenes

Datos estructurados
Representación numérica adecuada

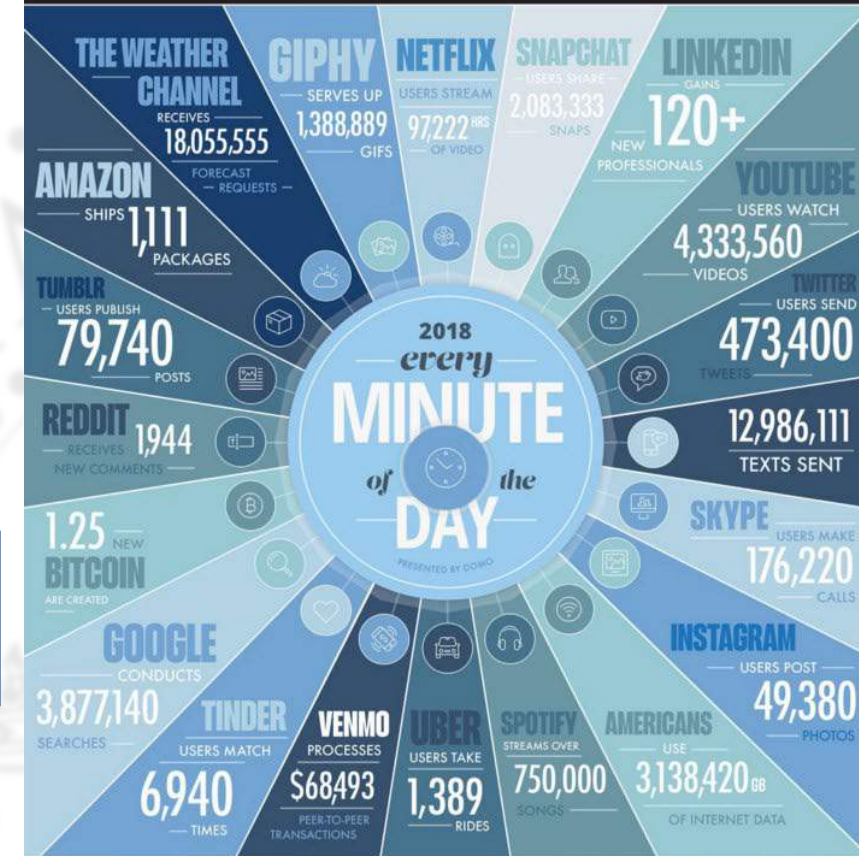
Persona: Marc Márquez

Evento: MotoGP

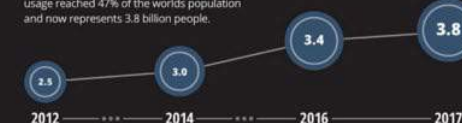
Ciudad: Valencia

Fecha: 2016

Personas relacionadas: Maverick Viñales



The world's internet population is growing significantly year-over-year. In 2017, internet usage reached 47% of the world's population and now represents 3.8 billion people.



GLOBAL INTERNET POPULATION GROWTH 2012-2017
(IN BILLIONS)

The ability to make data-driven decisions is crucial to any business. With each click, swipe, share, and like, a world of valuable information is created. Domo puts the power to make those decisions right into the palm of your hand by connecting your data and your people at any moment, on any device, so they can make the kind of decisions that make an impact.

Learn more at domo.com

SOURCES: STATISTA, LINKEDIN, INTERNET LIVE STATS, EXPANDED HAMBURGERS, SLASH FILM, RAA, BUSINESS OF APPS, INTERNATIONAL TELECOMMUNICATIONS UNION, INTERNATIONAL DATA CORPORATION



Procesamiento del Lenguaje Natural

Comprensión de la información

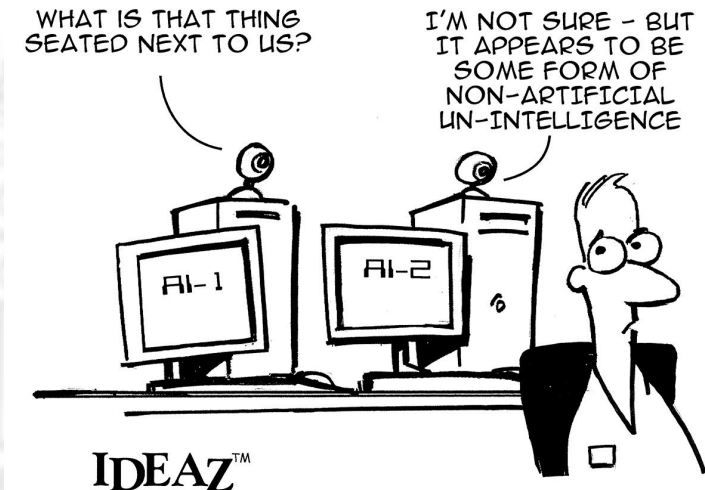
El objetivo final es **comprender** el mensaje codificado en el lenguaje.

Comprender

Percibir y tener una idea clara de lo que se dice, se hace o sucede o descubrir el sentido profundo de algo.

Implica entender conceptos y procesos para poder explicarlos y describirlos de forma adecuada.

➔ Nos proporciona herramientas para representar el **conocimiento**



Procesamiento del Lenguaje Natural

Representación del conocimiento

formalismos de representación del conocimiento:

Redes Semánticas (relaciones semánticas entre objetos en una red)

Frames(marcos), colección de datos estructurados
slots (propiedades) & fillers (valores) & Métodos

Reglas

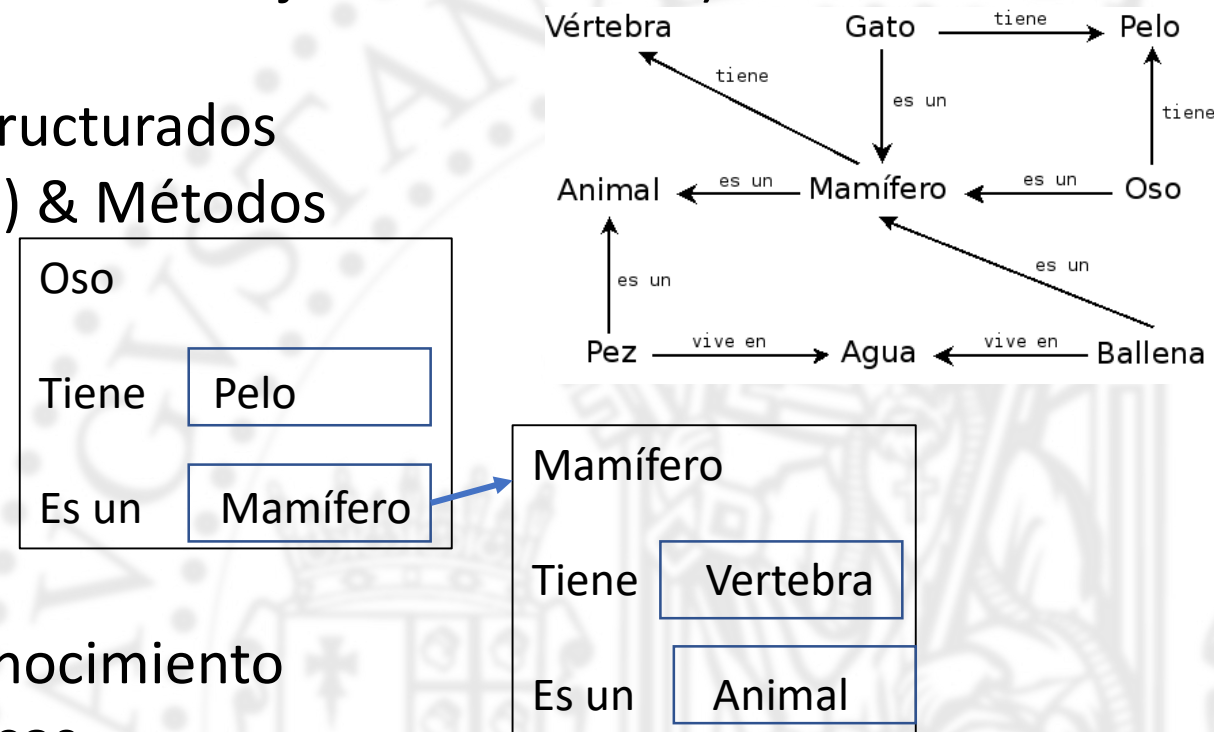
Si <condición>
entonces <conclusión>

Ontologías

esquemas de representación del conocimiento
basados en redes semánticas o marcos.

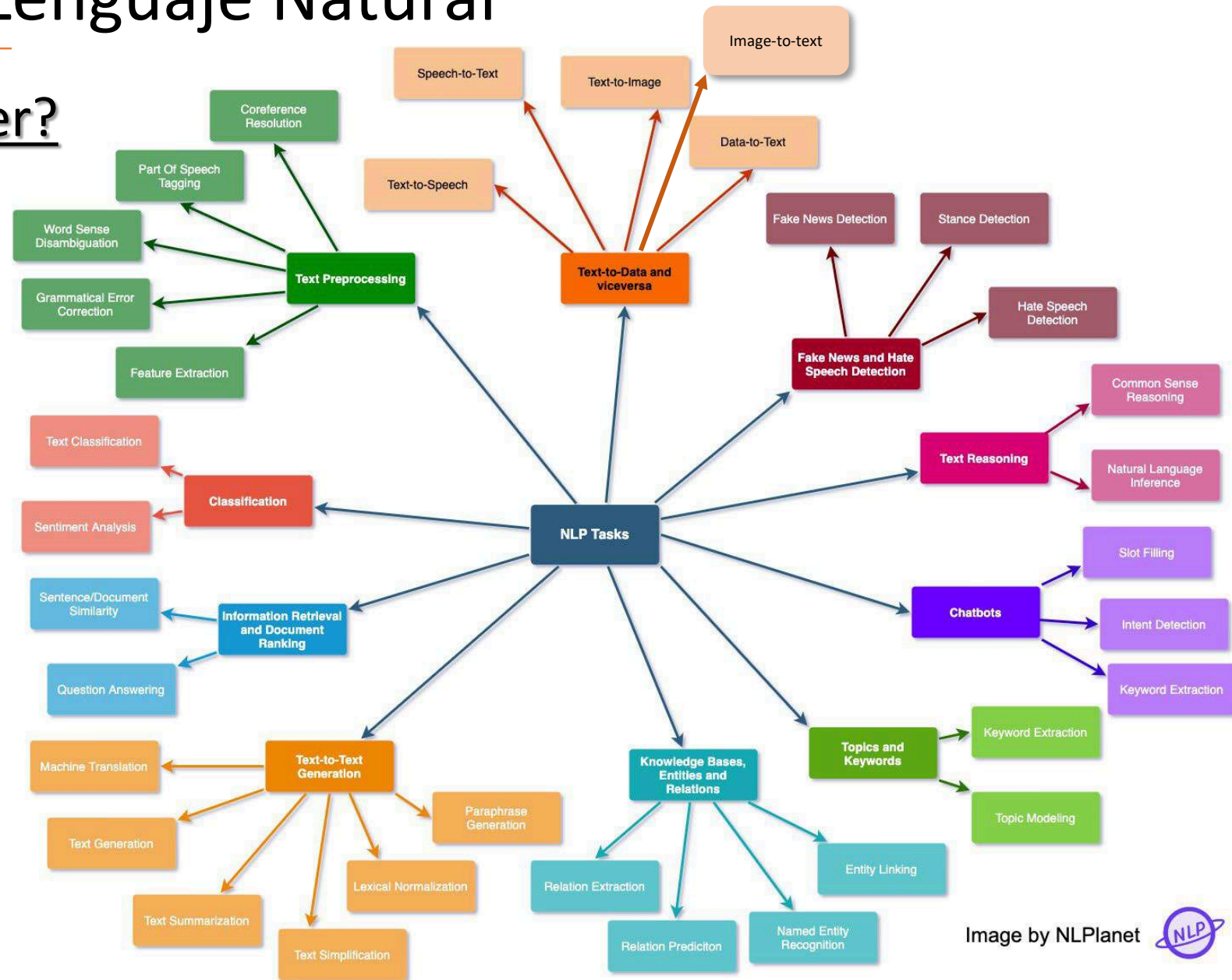
Espacios semánticos

Representación vectorial capaz de capturar el significado



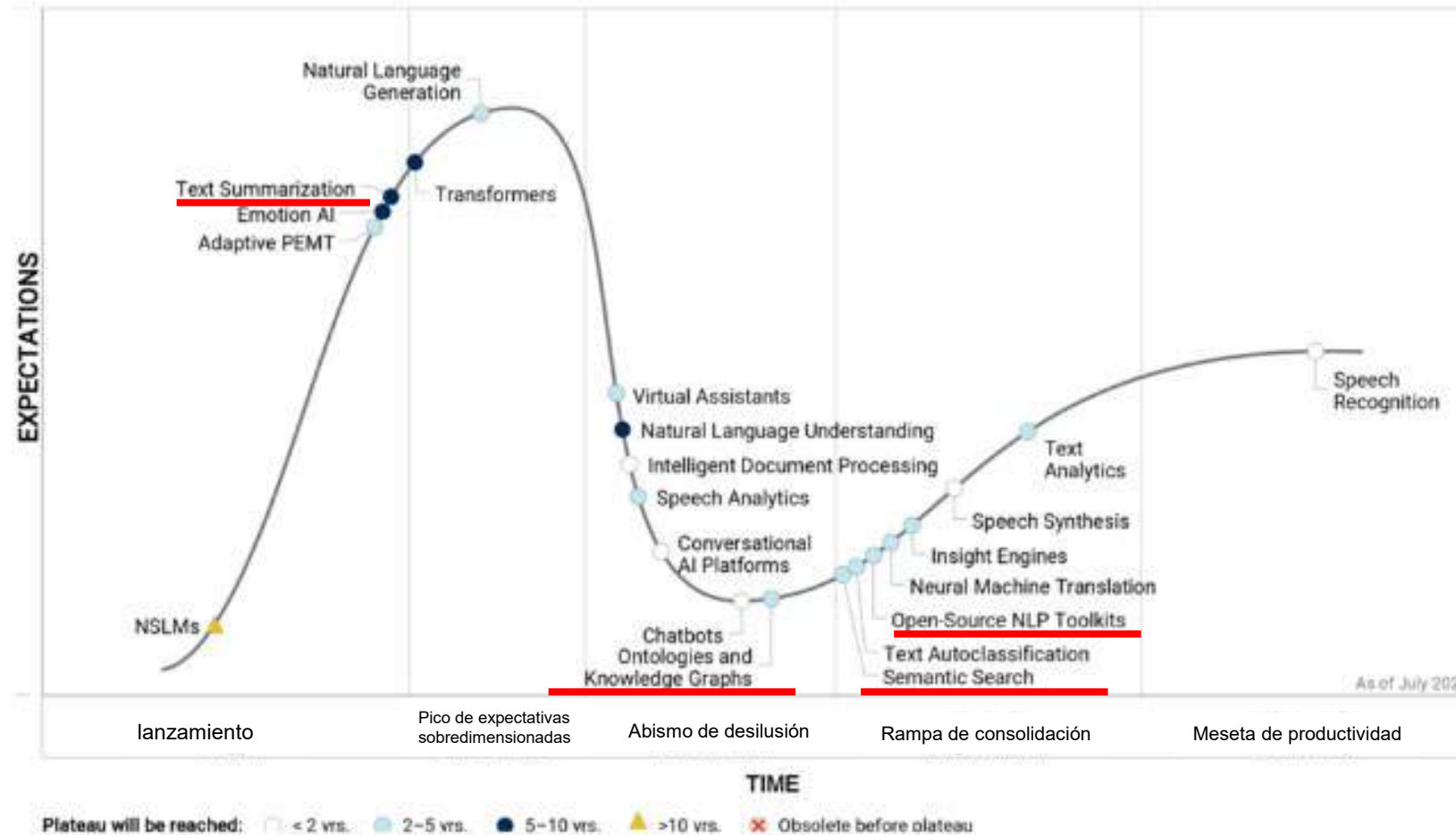
Procesamiento del Lenguaje Natural

¿Qué tareas podemos hacer?



Procesamiento del Lenguaje Natural

Hype Cycle for Natural Language Technologies, 2021



Source: Gartner (July 2021)

748656

Procesamiento del Lenguaje Natural

¿Cómo funciona el procesamiento de lenguaje natural?



OCR

voz a texto

Descripción de imagen

Pre-procesado texto:

- Segmentación en frases
- Segmentación en tokens
- Eliminar palabras comunes (stopwords)
- Lematización/Stemming
- Análisis morfológico
- Etiquetado gramatical (Part-Of-Speech tagging)
- Dependencias sintácticas

Sistemas/Algoritmos basados en:

- reglas
- datos (aprendizaje automático)
 - ✓ Espacios semánticos
 - ✓ Modelos de lenguaje

Speech and Language Processing (3rd ed. draft)

[Dan Jurafsky](#) and [James H. Martin](#)

<https://web.stanford.edu/~jurafsky/slp3/>

Procesamiento del Lenguaje Natural



voz a texto

Raw data

Segmentación en frases

Tokenización

Lematización

Netflix ha encontrado en el juego del calamar su nuevo fenómeno mundial ni siquiera en la propia plataforma contaban con ello como seguro que tampoco esperaban recibir multitud de quejas por una escena del cuarto episodio sin embargo han estado rápidos para responder a la indignación del público y ha introducido un cambio en el equipo

Netflix ha encontrado en el juego del calamar su nuevo fenómeno mundial. Ni siquiera en la propia plataforma contaban con ello como seguro que tampoco esperaban recibir multitud de quejas por una escena del cuarto episodio. Sin embargo han estado rápidos para responder a la indignación del público y ha introducido un cambio en el equipo.

| netflix | ha | encontrado | en | el | juego | de | el | calamar | su | nuevo | fenómeno | mundial | . |
| ni | siquiera | en | la | propia | plataforma | contaban | con | ello | ...

| netflix | haber | encontrar | en | el | juego | de | el | calamar | su | nuevo | fenómeno | mundial | . |
| ni | siquiera | en | el | propio | plataforma | contar | con | ello | ...

Procesamiento del Lenguaje Natural

POS tagging



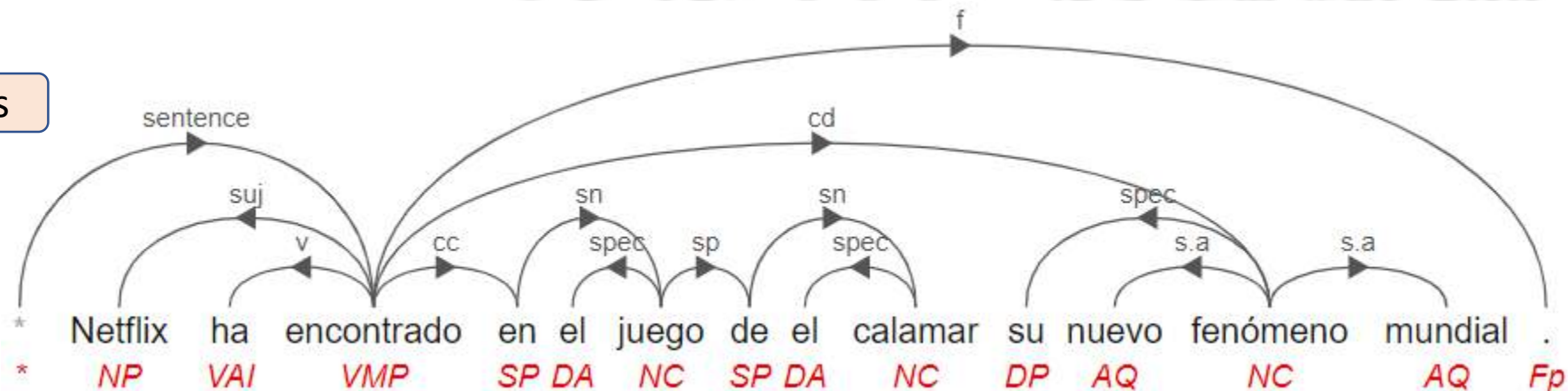
| Netflix [NP00SP0] | haber [VAIP3S0] | encontrar [VMP00SM] | en [SP] | el [DA0MS0] | juego [NCMS000] | de [SP] | el [DA0MS0] | calamar [NCMS000] | su [DP3CSN] | nuevo [AQ0MS00] | fenómeno [NCMS000] | mundial [AQ0CS00] | . [Fp]

| ni [CC] | siquiera [RG] | en [SP] | el [DA0FS0] | propio [AQ0FS00] | plataforma [NCFS000] | contar [VMII3P0] | con [SP] | ello [PD00S00] |

Quitar stopwords

| Netflix [NP00SP0] encontrar [VMP00SM] | juego [NCMS000] | calamar [NCMS000] | nuevo [AQ0MS00] | fenómeno [NCMS000] | mundial [AQ0CS00] | . [Fp] | ni [CC] | siquiera [RG] | propio [AQ0FS00] | plataforma [NCFS000] | contar [VMII3P0] |

Dependencias sintácticas



Procesamiento del Lenguaje Natural

Recursos:

Freeling (<https://nlp.lsi.upc.edu/freeling/index.php/>) permite: análisis morfológico, detección de entidades, POS-tagging, desambiguación del significado de palabras, análisis sintáctico, etiquetado de la función semántica...

Demo on-line:

<https://nlp.lsi.upc.edu/freeling/demo/demo.php>

SPACY (<https://spacy.io/>) toolkit en Python con el estado del arte en técnicas de procesamiento del lenguaje natural

Demos:

<https://spacy.io/universe>

GENSIM (<https://radimrehurek.com/gensim/>) toolkit en Python para el modelado de documentos y recuperación de información

Demos:

https://radimrehurek.com/gensim/auto_examples/index.html#documentation

NLTK (<https://www.nltk.org/>), Natural Language ToolKit *“leading platform for building Python programs to work with human language data”*

PATTERN (<https://github.com/clips/pattern/wiki>), *“web mining module for the Python programming language”*, NLP

TextBlob (<https://textblob.readthedocs.io>), *“API for diving into common natural language processing (NLP) tasks such as part-of-speech tagging, noun phrase extraction, sentiment analysis, classification, translation, and more.”*

Basada en NLTK y PATTERN

Procesamiento del Lenguaje Natural

Pero, ¿cómo representamos los tokens/palabras en una máquina?

La máquina trabaja con números,

.... luego *debemos transformar las palabras a números*

Opción 1.

Utilizamos un código numérico, p.e. las numeramos de forma correlativa

netflix	1
encontrar	2
juego	3
calamar	4
nuevo	5
fenómeno	6
mundial	7
ni	8
siquiera	9
propio	10
plataforma	11
contar	12

¿tiene algún significado el valor numérico?

¿Podemos calcular la proximidad semántica?, ¿tiene sentido?

Procesamiento del Lenguaje Natural

Opción 2.

Definimos un espacio matemático donde todas las palabras estén a la misma distancia.

El espacio matemático lo definimos con unos ejes de coordenadas que serán las dimensiones del espacio.

P.e. un vocabulario de 3 palabras \rightarrow 3 dimensiones.

Un punto en un espacio de tres dimensiones estará definido por el valor de las 3 coordenadas (x,y,z) , que llamaremos **vector**

Palabra A $(1,0,0)$

Palabra B $(0,1,0)$

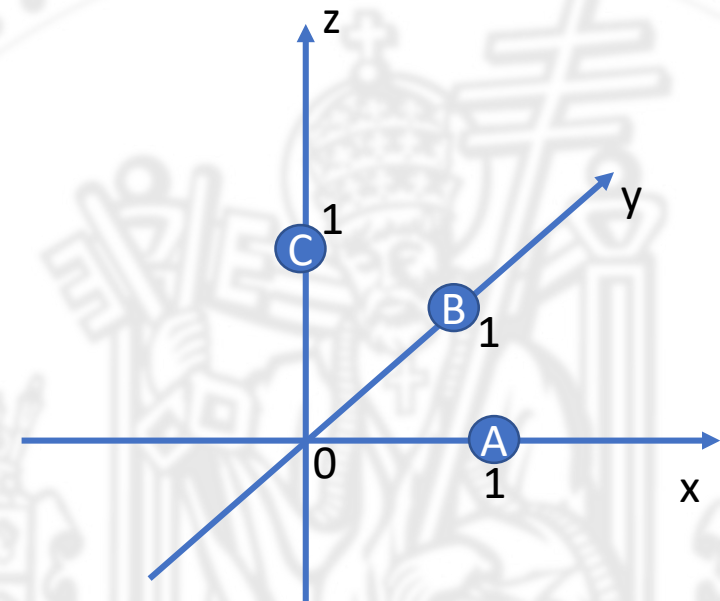
Palabra C $(0,0,1)$

En este espacio las tres palabras están a la misma distancia.

Para un vocabulario de N palabras, necesitaremos N dimensiones.

Representación *one-hot vector*, todos los valores a 0 excepto uno a 1

netflix	$(1,0,0,0,0,0,0,0,0,0,0,0)$
encontrar	$(0,1,0,0,0,0,0,0,0,0,0,0)$
juego	$(0,0,1,0,0,0,0,0,0,0,0,0)$
....	
plataforma	$(0,0,0,0,0,0,0,0,0,0,1,0)$
contar	$(0,0,0,0,0,0,0,0,0,0,0,1)$



¿tiene algún significado el valor numérico?

Representación dispersa (muchos 0)

Procesamiento del Lenguaje Natural

Opción 3.

Reflexionemos,
¿qué buscamos? (carta a los reyes magos)

- ✓ Queremos representar el significado de unidades lingüísticas (tokens/palabras)
- ✓ Queremos definir una medida de similitud semántica entre unidades
- ✓ Queremos que sea una representación numérica densa: “embeddings”

En definitiva:

Un espacio matemático de representación compacto donde la posición de los vectores que me identifican a las unidades contenga información semántica y que llamaremos *espacio semántico*

¿cómo lo construimos?

Semántica distribucional

Procesamiento del Lenguaje Natural

Semántica distribucional

¿Cómo conocemos el significado de una palabra?

John Rupert Firth, “You shall know a word by the Company it keeps”

“Similar words occur in similar contexts”

Ludwig Wittgenstein, “The meaning of a word is its use in language”

Hay una botella de *Belikin* sobre la mesa

A todo el mundo le gusta la *Belikin*

No bebas *Belikin* si tienes que conducir

La *Belikin* se fabrica con granos de cebada germinada

¿qué podemos deducir sobre la palabra *Belikin*?

Miramos las palabras que acompañan

Buscamos la similitud semántica con otras palabras ya conocidas

... y deducimos que la *Belikin* debe ser una bebida similar a...

Procesamiento del Lenguaje Natural

Semántica distribucional

Hipótesis

Si dos palabras w_1 y w_2 tienen distribuciones similares, podemos asumir que tienen significados similares

Vector semántico

Vector que representa la distribución de una palabra

Matriz de coocurrencias o contextuales

cuando a menudo los términos (palabras/lexemas/...) coocurren en un contexto

Matriz de términos-documentos

Matriz de términos-términos

	As You Like It	Twelfth Night	Julius Caesar	Henry V
battle	1	1	8	15
soldier	2	2	12	36
fool	37	58	1	5
clown	5	117	0	0

Figure 15.1 The term-document matrix for four words in four Shakespeare plays. Each cell contains the number of times the (row) word occurs in the (column) document.

	aardvark	...	computer	data	pinch	result	sugar	...
apricot	0	...	0	0	1	0	1	
pineapple	0	...	0	0	1	0	1	
digital	0	...	2	1	0	1	0	
information	0	...	1	6	0	4	0	

Figure 15.4 Co-occurrence vectors for four words, computed from the Brown corpus.

Daniel Jurafsky, James H. Martin

Procesamiento del Lenguaje Natural

Semántica distribucional

Matriz términos-documentos: *buscar documentos similares*

Cada fila representa una palabra del vocabulario

Cada columna representa un documento

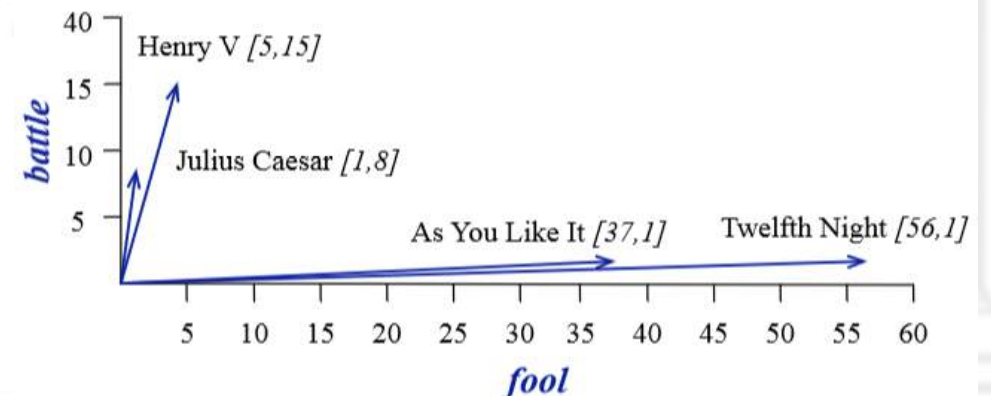
Cada celda: Cuenta del término t en el documento d : $tf_{t,d}$

Cada documento: un vector de cuentas en \mathbb{N}^v

Dos documentos son similares si sus vectores son similares

	As You Like It	Twelfth Night	Julius Caesar	Henry V
battle	1	1	8	15
soldier	2	2	12	36
fool	37	58	1	5
clown	5	117	0	0

Figure 15.2 The term-document matrix for four words in four Shakespeare plays. The red boxes show that each document is represented as a column vector of length four.



Daniel Jurafsky, James H. Martin

Procesamiento del Lenguaje Natural

Semántica distribucional

Matriz términos-contexto: *buscar palabras similares*

Cada fila representa una palabra del vocabulario

Cada columna representa un contexto/palabra

Cada celda: # veces la palabra t y c coocurren en el algún contexto

	c_1	c_2	c_3	c_4	...	c_v
t_1	12	0	4	0		2
t_2	1	6	0	8		2
t_3	6	1	14	2		0
t_4	0	0	4	16		10
...						
t_v	5	5	0	2		0

Contexto:

- ✓ Mismo documento
- ✓ Párrafo
- ✓ Ventana de $\pm D$ palabras alrededor la palabra t_v

Procesamiento del Lenguaje Natural

Semántica distribucional

Matriz de palabras-palabras con un contexto de ± 7 palabras

sugar, a sliced lemon, a tablespoonful of their enjoyment. Cautiously she sampled her first well suited to programming on the digital for the purpose of gathering data and	apricot pineapple computer. information	preserve or jam, a pinch each of, and another fruit whose taste she likened In finding the optimal R-stage policy from necessary for the study authorized in the
---	--	--

Matrices de dimensión $|V| \times |V|$, donde $|V| < 50.000$

Muchos valores son 0. Matrices dispersas (sparse)

El tamaño del contexto depende de la finalidad

Representaciones más sintácticas, ventanas más cortas $\pm 1-3$

Representaciones más semánticas, ventanas más largas $\pm 4-10$

Daniel Jurafsky, James H. Martin

Procesamiento del Lenguaje Natural

Representaciones densas

La mayoría de los elementos del vector son cero

Reducir el tamaño del vector: representaciones densas

- Descomposición en valores singulares
 - LSA: Latent Semantic Analysis
- Word embeddings:
 - Word2vec: cbow y skip-gran (<https://code.google.com/archive/p/word2vec>)
 - Glove: global vectors (<https://nlp.stanford.edu/projects/glove>)
 - fastText: Text representation (<https://fasttext.cc/>)

Representación gráfica de vectores densos

- t-SNE: t-Distributed Stochastic Neighbor Embedding (<https://lvdmaaten.github.io/tsne/>)
- UMAP: Uniform Manifold Approximation and Projection (<https://umap-learn.readthedocs.io>)

Procesamiento del Lenguaje Natural

Representaciones densas: Descomposición en Valores Singulares

Idea:

Aproximar un espacio N-dimensional con otro M-dimensional donde $M \ll N$

¿Cómo?

Rotar los ejes a un nuevo espacio donde los ejes están orientados a las dimensiones de mayor varianza.

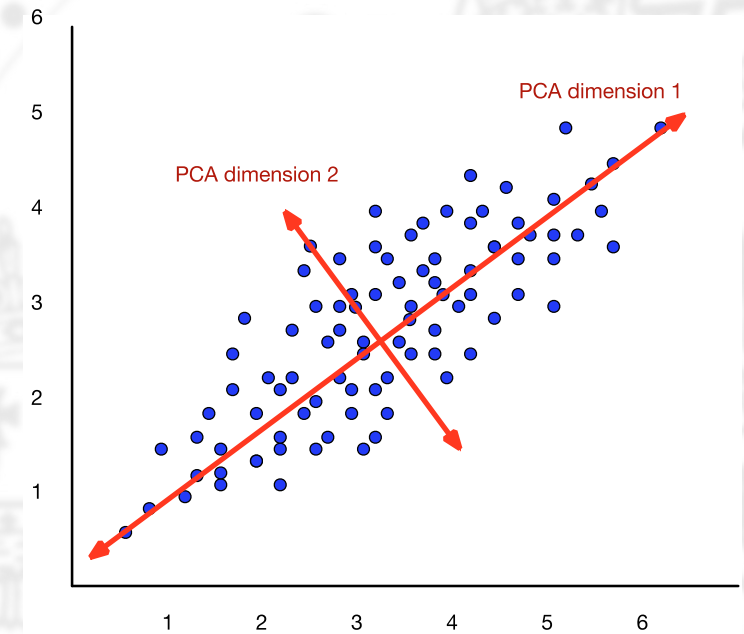
Utilizar las dimensiones de mayor varianza

Algunos métodos:

PCA – principal components analysis

FA – Factor Analysis

SVD – Singular Value Decomposition



Daniel Jurafsky, James H. Martin

Procesamiento del Lenguaje Natural

Representaciones densas: Descomposición en Valores Singulares

Una matriz rectangular X de dimensiones $|V| \times |C|$ se puede factorizar en el producto de tres matrices

$$X = W_k \Sigma_k C_k^T$$

donde

W es una matriz de dimensiones $|V| \times k$, las filas se corresponden con el espacio original y cada columna es una nueva dimensión en lo que se denomina el “espacio latente” (espacio no observado directamente)

$$k < \min(|V|, |C|)$$

Σ es una matriz diagonal de dimensiones $k \times k$. Cada elemento o valor singular informa sobre la importancia de la dimensión que representa

C es una matriz de dimensiones $|C| \times k$, las filas se corresponden con el espacio original y cada columna es un vector en el “espacio latente”

Procesamiento del Lenguaje Natural

Representaciones densas: Descomposición en Valores Singulares

LSA - Latent Semantic Analysis

$$\begin{bmatrix} X \\ |V| \times c \end{bmatrix} = \begin{bmatrix} W_k \\ |V| \times k \end{bmatrix} \begin{bmatrix} \sigma_1 & 0 & 0 & \dots & 0 \\ 0 & \sigma_2 & 0 & \dots & 0 \\ 0 & 0 & \sigma_3 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & \sigma_k \\ k \times k \end{bmatrix} \begin{bmatrix} C^T \\ k \times c \end{bmatrix}$$

La matriz W_k nos da la representación de las palabras en el espacio latente

Word embedding

La matriz C_k nos da la representación de los documentos en el espacio latente

Document embedding

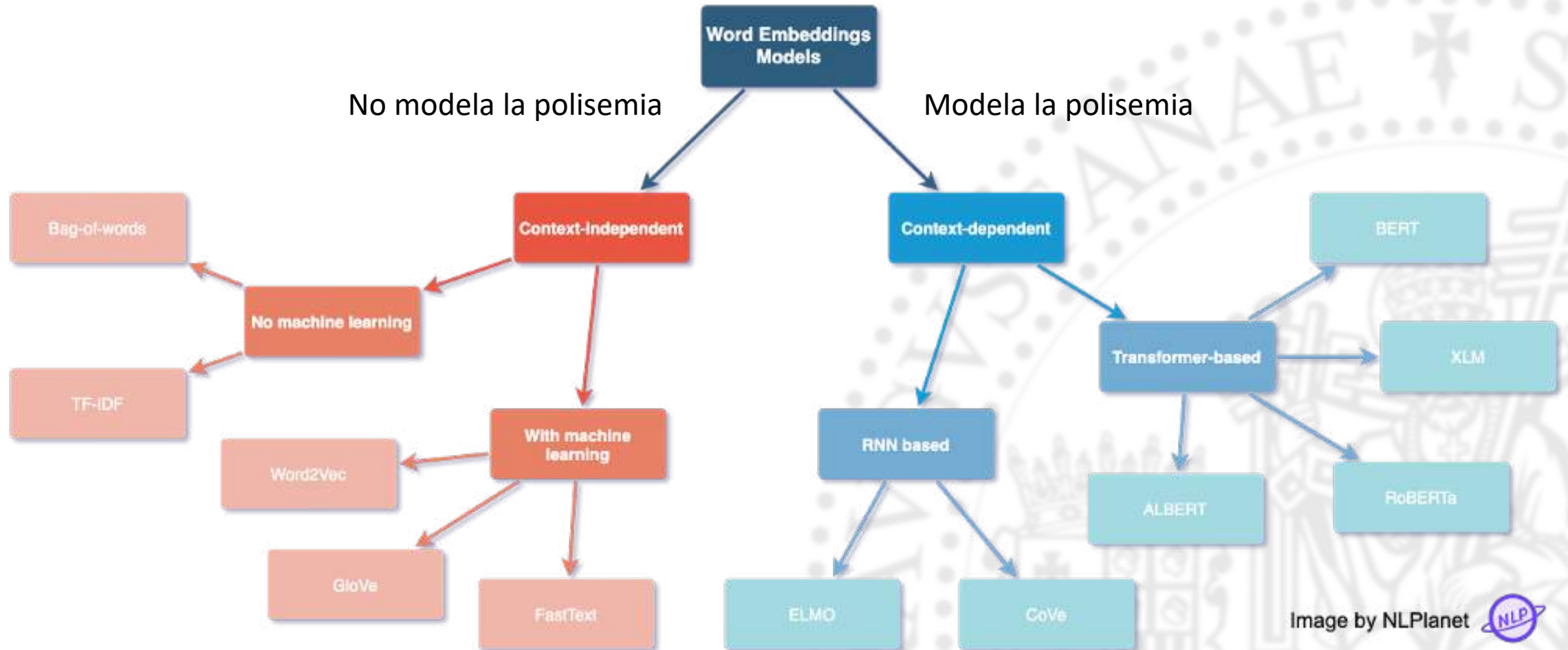
Ejemplo indexado:

Pregunta o query q ($|V| \times 1$) se convierte al espacio latente q^* ($k \times 1$)

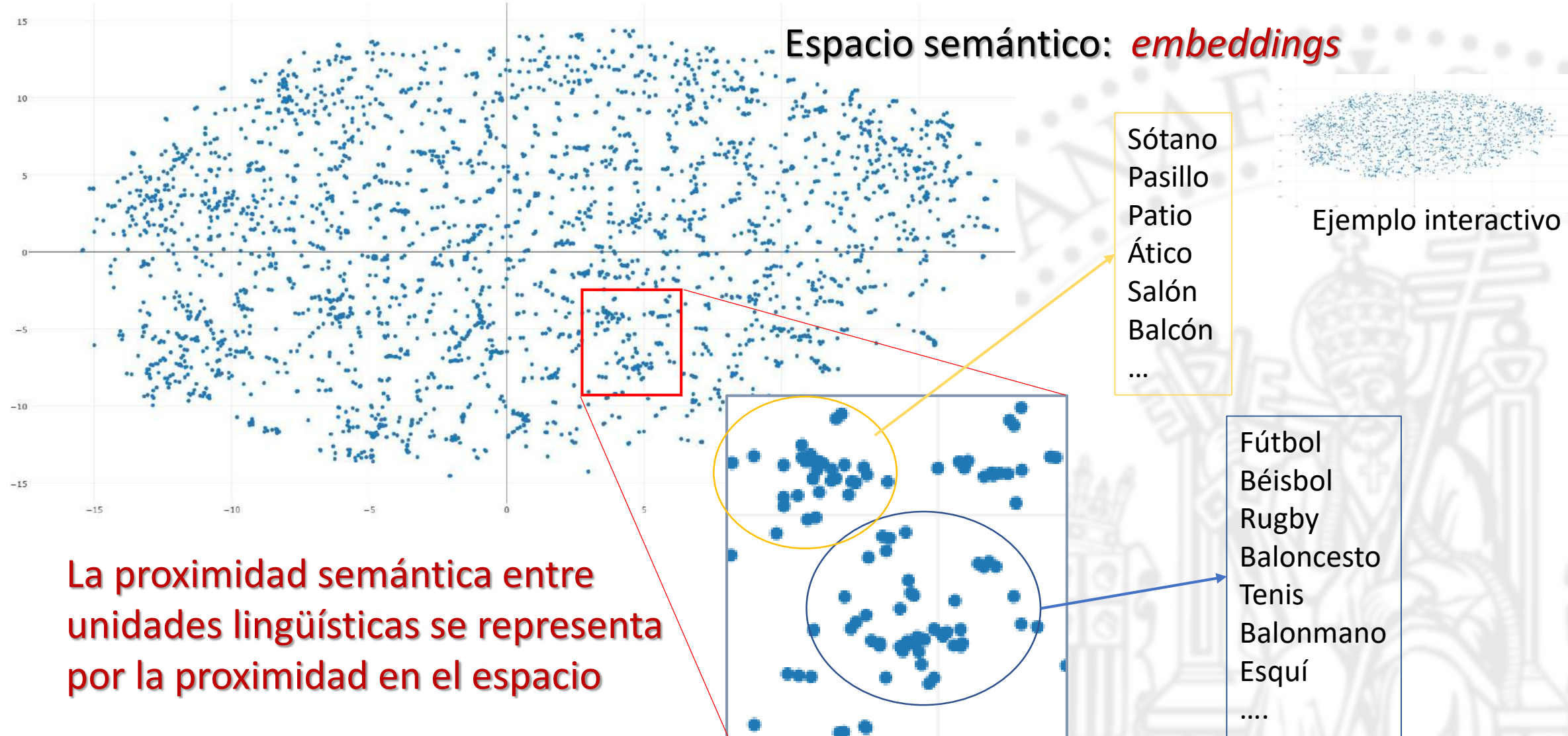
$$q^* = \Sigma_k^{-1} W_k^T q$$

Se compara con todos los embeddings de los documentos (veremos un ejemplo)

Procesamiento del Lenguaje Natural



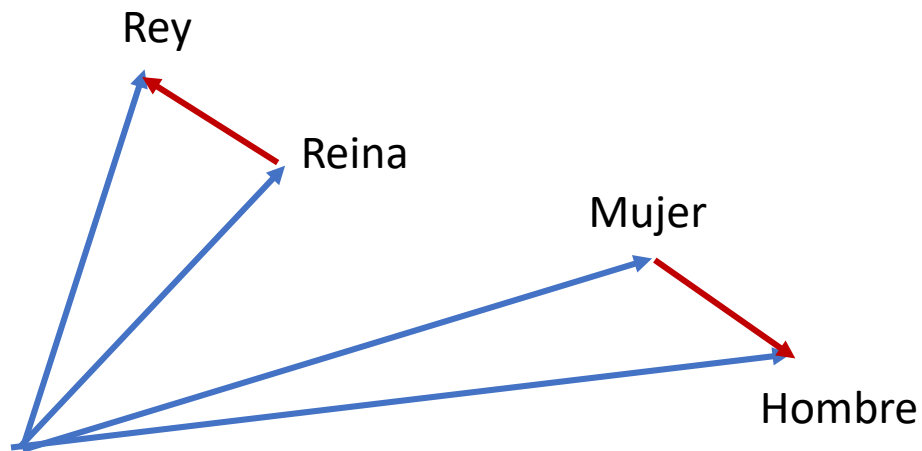
Procesamiento del Lenguaje Natural



Procesamiento del Lenguaje Natural

Relaciones semánticas

$$\text{vector}[\text{Reina}] = \text{vector}[\text{Rey}] - \text{vector}[\text{Hombre}] + \text{vector}[\text{Mujer}]$$



- día + noche =

-volar+navegar =

- taza + caja =

- caja + taza =

Imágenes próximas



(Kiros, Salakhutdinov, Zemel, TACL 2015)

Procesamiento del Lenguaje Natural

Generalización de los espacios semánticos:

Embeddings de Palabras, Frases, Documentos, Audio, Imagen, Vídeos, ...

Ejemplo: Búsqueda semántica en periódicos

Buscador de noticias similares <http://signal4.cps.unizar.es:5000/>

- ✓ Cada noticia es un embedding
- ✓ Calcular embedding del texto a buscar
- ✓ Buscar los embeddings de noticias más próximos al del texto

Pero además permite

- ✓ Clasificar las noticias por categorías/temas
- ✓ Reconocer entidades
- ✓ Descubrir estereotipos y sesgos
- ✓ Evolución temporal/espacial de la semántica de palabras
- ✓ Componente principal de los modelos de lenguaje con redes neuronales
- ✓

Procesamiento del Lenguaje Natural

Bases de datos basadas en vectores

MILVUS (<https://milvus.io/>)

Milvus se creó en 2019 con el único objetivo de almacenar, indexar y administrar vectores de embeddings masivos.

Usos:

- Búsqueda masiva de texto, imágenes, audio y vídeo por similitud sin necesidad de estructurar los datos
- Sistemas de recomendación
- Sistemas de pregunta-respuesta

Ejemplos

<https://github.com/milvus-io/bootcamp/tree/master/solutions>

<https://milvusio.medium.com/>

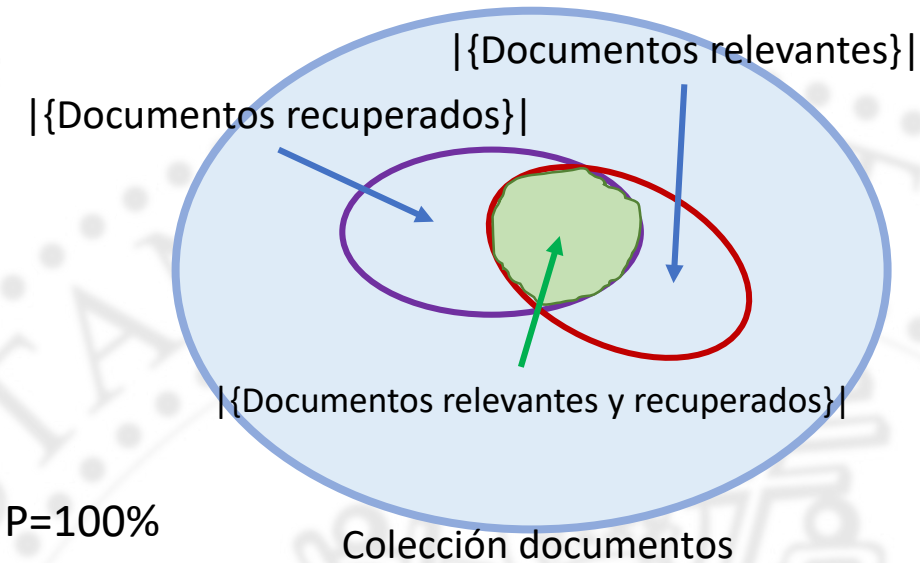
Procesamiento del Lenguaje Natural

Medida de prestaciones en búsquedas y recuperación de información:

Precisión y Exhaustividad/Sensibilidad (Precision/Recall), Valor-F (F-score)

Precisión: Probabilidad de que un documento recuperado sea relevante.

$$P = \frac{|\{\text{documentos relevantes y recuperados}\}|}{|\{\text{documentos recuperados}\}|}$$



Exhaustividad: Probabilidad de que un documento relevante sea recuperado en una búsqueda.

$$R = \frac{|\{\text{documentos relevantes y recuperados}\}|}{|\{\text{documentos relevantes}\}|}$$



Valor-F: Valor único ponderado de la Precisión y la Exhaustividad.

$$F = 2 \frac{\text{Precisión} \times \text{Exhaustividad}}{\text{Precisión} + \text{Exhaustividad}}$$

$$F_{\beta} = (1 + \beta^2) \frac{\text{Precisión} \times \text{Exhaustividad}}{\beta^2 \times \text{Precisión} + \text{Exhaustividad}}$$

Procesamiento del Lenguaje Natural

Modelos de Lenguaje

Asignar una probabilidad a una frase o secuencia de palabras

Calcular la probabilidad de una frase

W="la alta comisionada de las naciones unidas para los derechos humanos ha advertido que atacar viviendas supone una violación de las convenciones de ginebra que regulan los conflictos"

$$P(W) = P(w_1, w_2, w_3, w_4, \dots, w_M)$$

Predecir la siguiente palabra en una secuencia

W="la alta comisionada de las naciones..."

$$P(w_n | w_1, w_2 \dots w_{n-1})$$

Modelos estadísticos: N-gramas

Bigramas, Trigramas, ...

Modelos neuronales:

Large Language Models (GPT-3 (175 Gparámetros), PaLM (540 Gparámetros),...)

Fine-tuned Language Modes (OpenAI Codex, orden de Gparámetros)

Edge Language Models (adaptados al hardware local, muchos menos parámetros que LLMs)

Procesamiento del Lenguaje Natural

¿Dónde estamos?

<https://huggingface.co/>

<https://openai.com/>



Completion
Generate or manipulate text and code



Semantic search
Score text based on relevance



Fine-tuning Beta
Train a model for your use case



Classification Beta
Classify text into different categories



Question answering Beta
Generate high-accuracy answers



The AI community building the future.

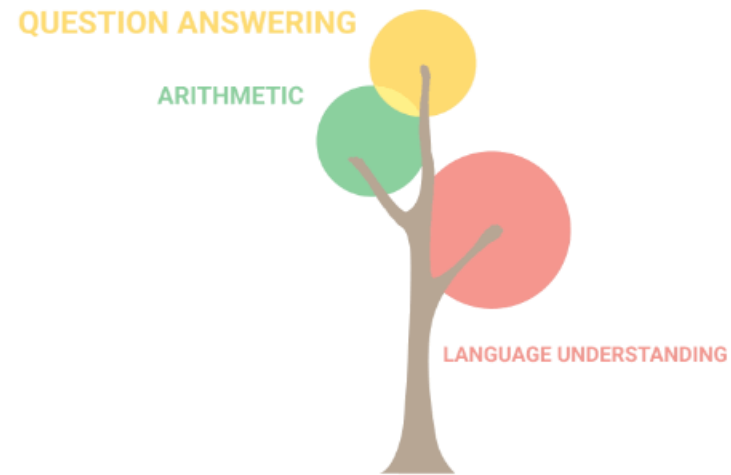
Build, train and deploy state of the art models powered by
the reference open source in machine learning.

GPT-3 Access Without the Wait

We've made improvements to our API and safety
features so developers can get started right away.

<https://beta.openai.com/examples>

Procesamiento del Lenguaje Natural



8 billion parameters

Pathways Language Models (PaLM) Google AI

<https://ai.googleblog.com/2022/04/pathways-language-model-palm-scaling-to.html>

Procesamiento del Lenguaje Natural

GPT-3: engine="text-davinci-001"

Hazme una lista en formato json con la persona, oficio, nacionalidad y año de nacimiento.

Bryan Adams, el fotógrafo encargado de realizar el calendario, es un cantante, guitarrista, compositor, fotógrafo y filántropo canadiense.

Anne Erin Annie Clark, conocida artísticamente como St. Vincent, es una cantautora y multiinstrumentista estadounidense. Es ganadora de tres Premios Grammy por Mejor Canción de Rock.

Kali Uchis, es una cantante, compositora, actriz, directora y diseñadora colomboestadounidense, saltó a la fama internacional en dos mil veintiuno con el gran éxito de su canción Telepatía.

Diamonte Quiava Valentin Harper, es una conocida rapera estadounidense.

Cher es una de las figuras más destacadas del este calendario.

```
[{
  "person": "Bryan Adams",
  "oficio": "cantante, guitarrista, compositor, fotógrafo y filántropo",
  "nacionalidad": "canadiense",
  "año de nacimiento": "1958"
},
{
  "person": "Anne Erin Annie Clark",
  "oficio": "cantautora y multiinstrumentista",
  "nacionalidad": "estadounidense",
  "año de nacimiento": "1982"
},
{
  "person": "Kali Uchis",
  "oficio": "cantante, compositora, actriz, directora y diseñadora",
  "nacionalidad": "colomboestadounidense",
  "año de nacimiento": "1994"
},
{
  "person": "Diamonte Quiava Valentin Harper",
  "oficio": "rapera",
  "nacionalidad": "estadounidense",
  "año de nacimiento": "1996"
},
{
  "person": "Cher",
  "oficio": "cantante",
  "nacionalidad": "estadounidense",
  "año de nacimiento": "1946"
}]
```

Procesamiento del Lenguaje Natural

Razonamiento: PaLM

Standard Prompting

Example Input

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

Example Output

A: The answer is 11.

Prompt

The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

Model Response

The answer is 50.



Chain of thought prompting

Example Input

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

Example Output

Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. $5 + 6 = 11$. The answer is 11.

Prompt

The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

Model Response

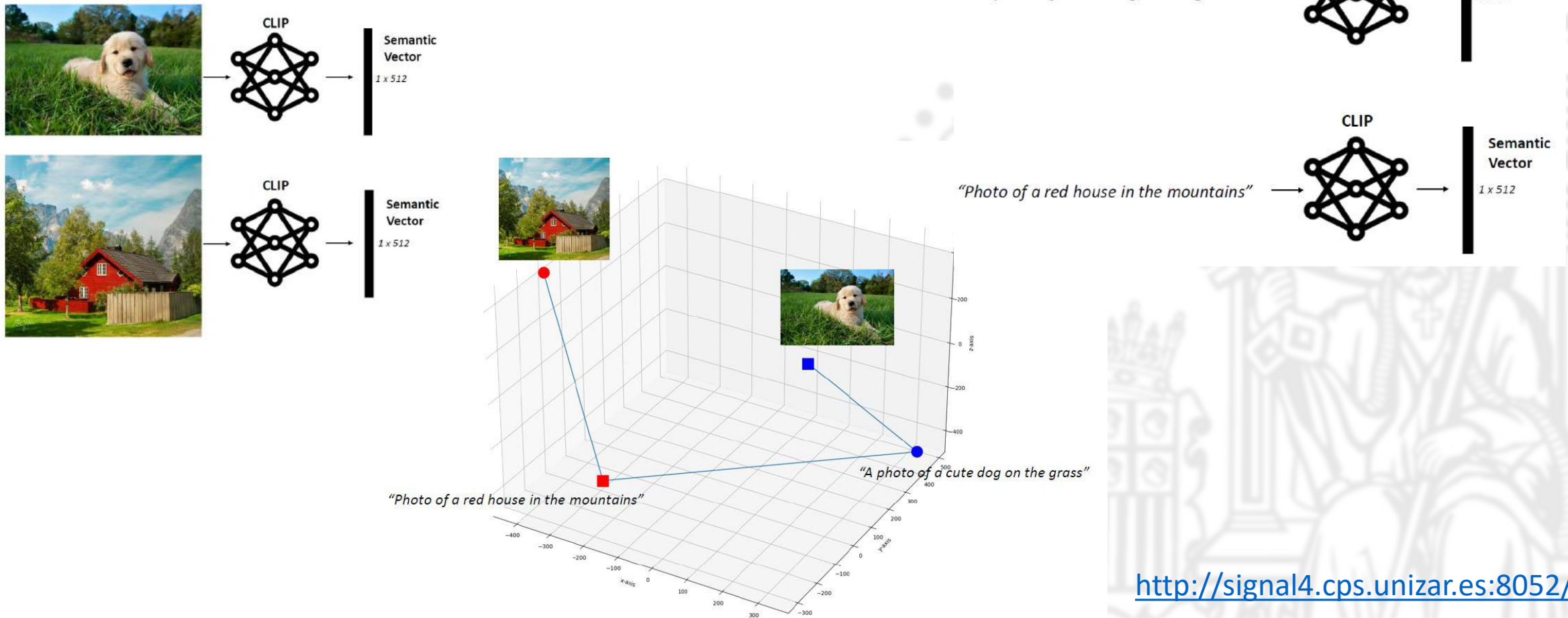
The cafeteria had 23 apples originally. They used 20 to make lunch. So they had $23 - 20 = 3$. They bought 6 more apples, so they have $3 + 6 = 9$. The answer is 9.



Procesamiento del Lenguaje Natural

CLIP: modelo multimodal de OpenAI

Combina un modelo de lenguaje de inglés con un modelo semántico de conocimiento de imágenes
Entrenado con mas de 400M de pares imagen+texto



<http://signal4.cps.unizar.es:8052/>

Procesamiento del Lenguaje Natural

Algunas aplicaciones

Comparación textos - imagen



```
"matches": [ {"text": "the blue car is on the left, the red car is on the right"}, {"text": "the blue car is on the right, the red car is on the left"}, {"text": "the blue car is on top of the red car"}, {"text": "the blue car is below the red car"} ]]
```

```
"the blue car is on the left, the red car is on the right" 0.5232442617416382  
"the blue car is on the right, the red car is on the left" 0.32878655195236206  
"the blue car is below the red car" 0.11064132302999496  
"the blue car is on top of the red car" 0.03732786327600479
```


Procesamiento del Lenguaje Natural

Algunas aplicaciones

CLIP + GPT2: Descripción de imágenes

Búsqueda en vídeos con lenguaje natural



A couple of people standing next to an elephant.



A wooden table sitting in front of a window.



A bunch of bananas sitting on top of a table.



A woman holding a plate with a piece of cake in front of her face.

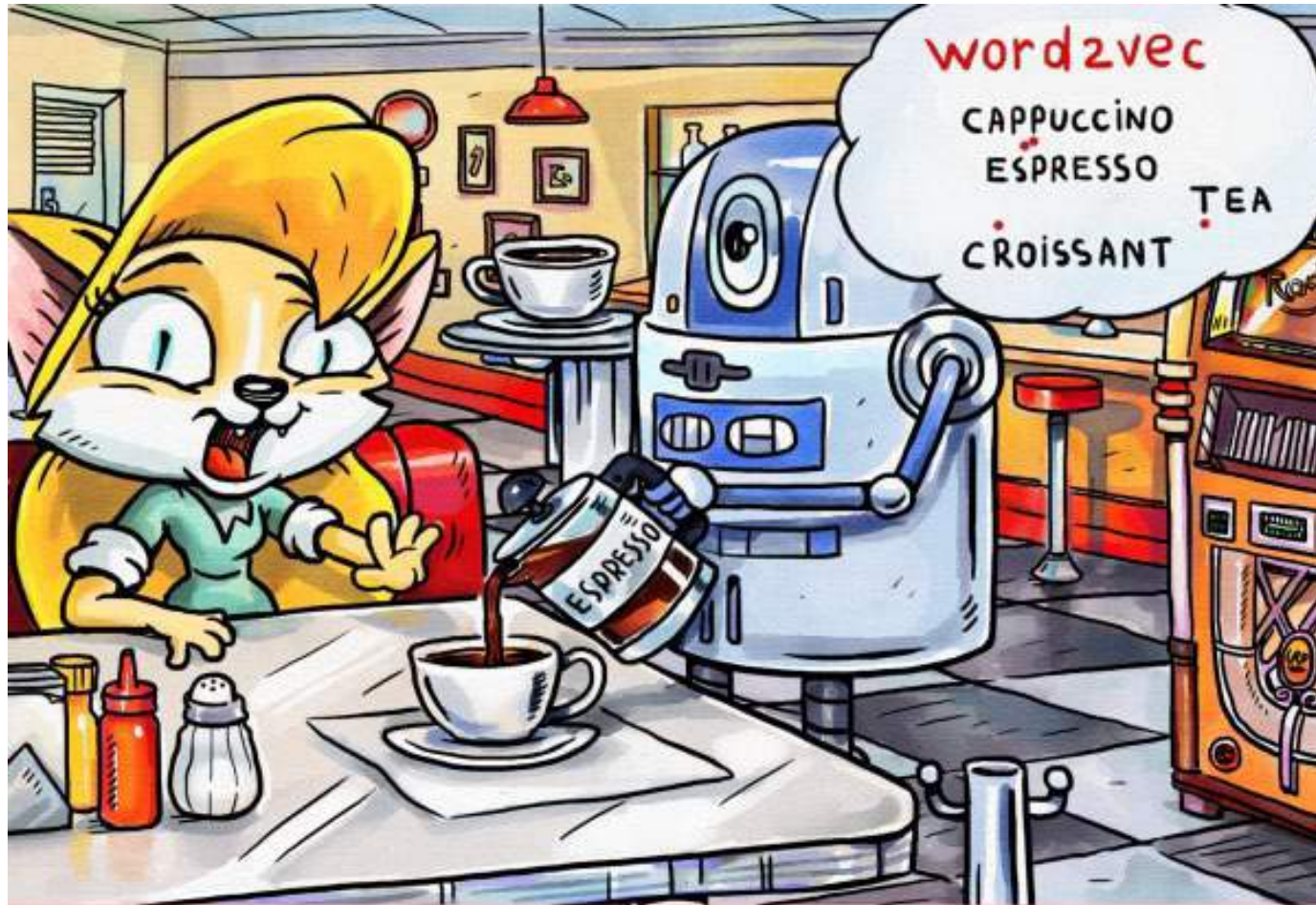


A wooden table topped with lots of wooden utensils.



A red motorcycle parked on top of a dirt field.

Procesamiento del Lenguaje Natural



- Espresso? But I ordered a cappuccino!
- Don't worry, the cosine distance between them is so small that they are almost the same thing.



Universidad
Zaragoza

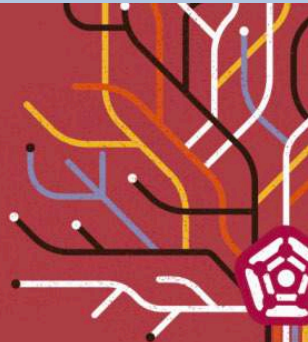
Introducción a las Tecnologías de Inteligencia Artificial: Tecnologías del Habla



Cursos Extraordinarios
Universidad Zaragoza

Vicerrectorado de Cultura y Proyección Social

CURSOS
EXTRAORDINARIOS
UNIVERSIDAD DE ZARAGOZA



VIVOLAB

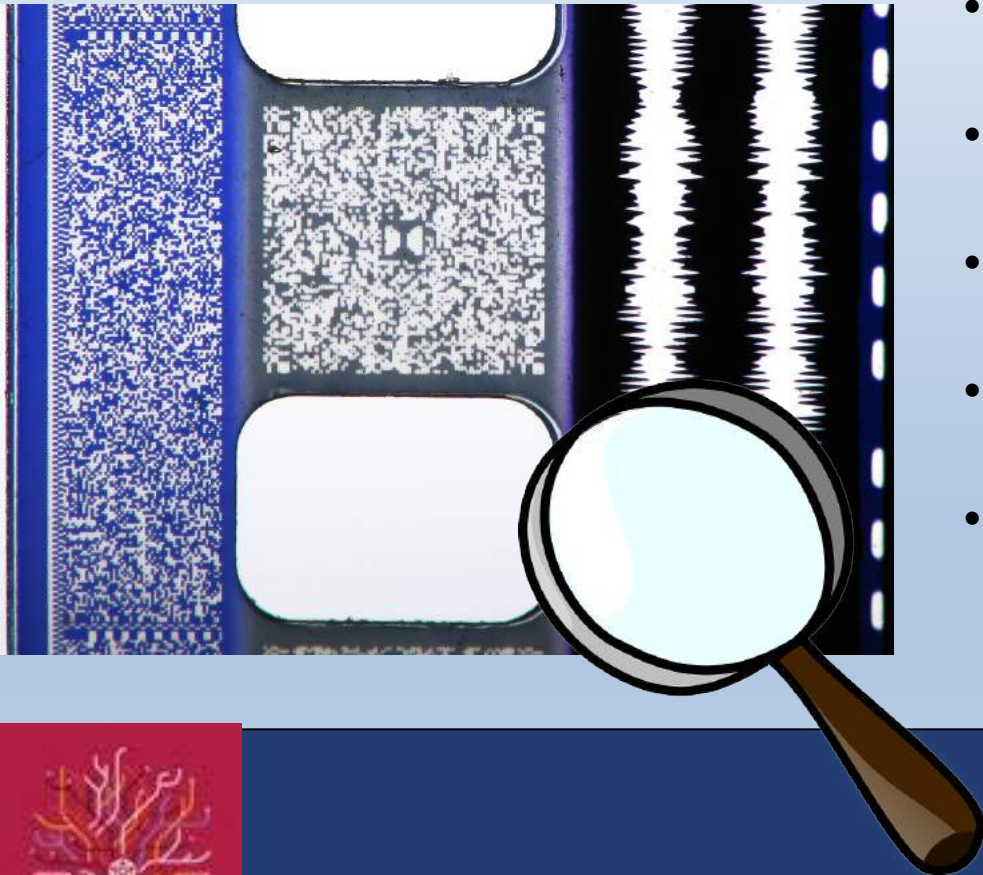


Universidad
Zaragoza

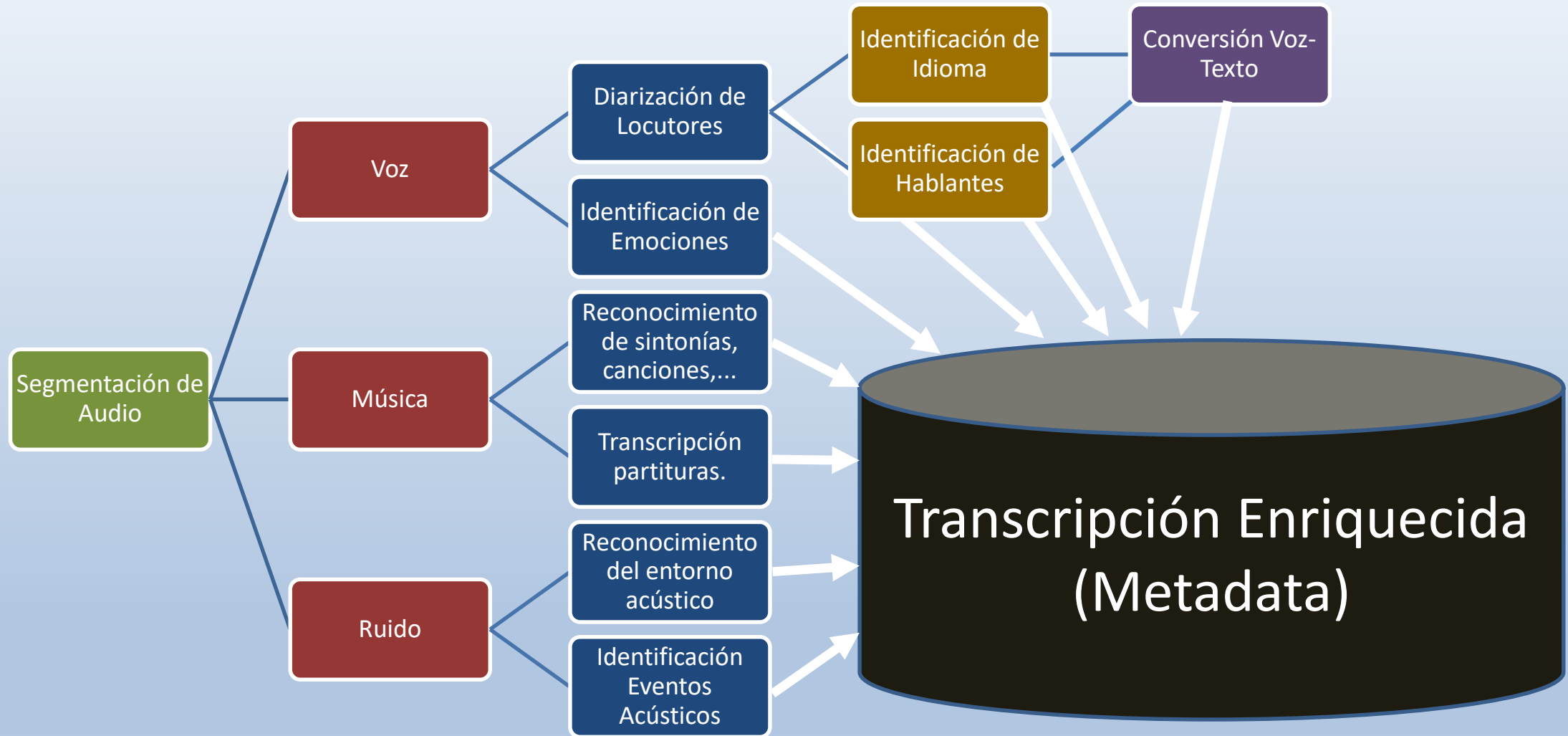
Extracción de Información: Audio

¿Qué información podemos encontrar en un audio?

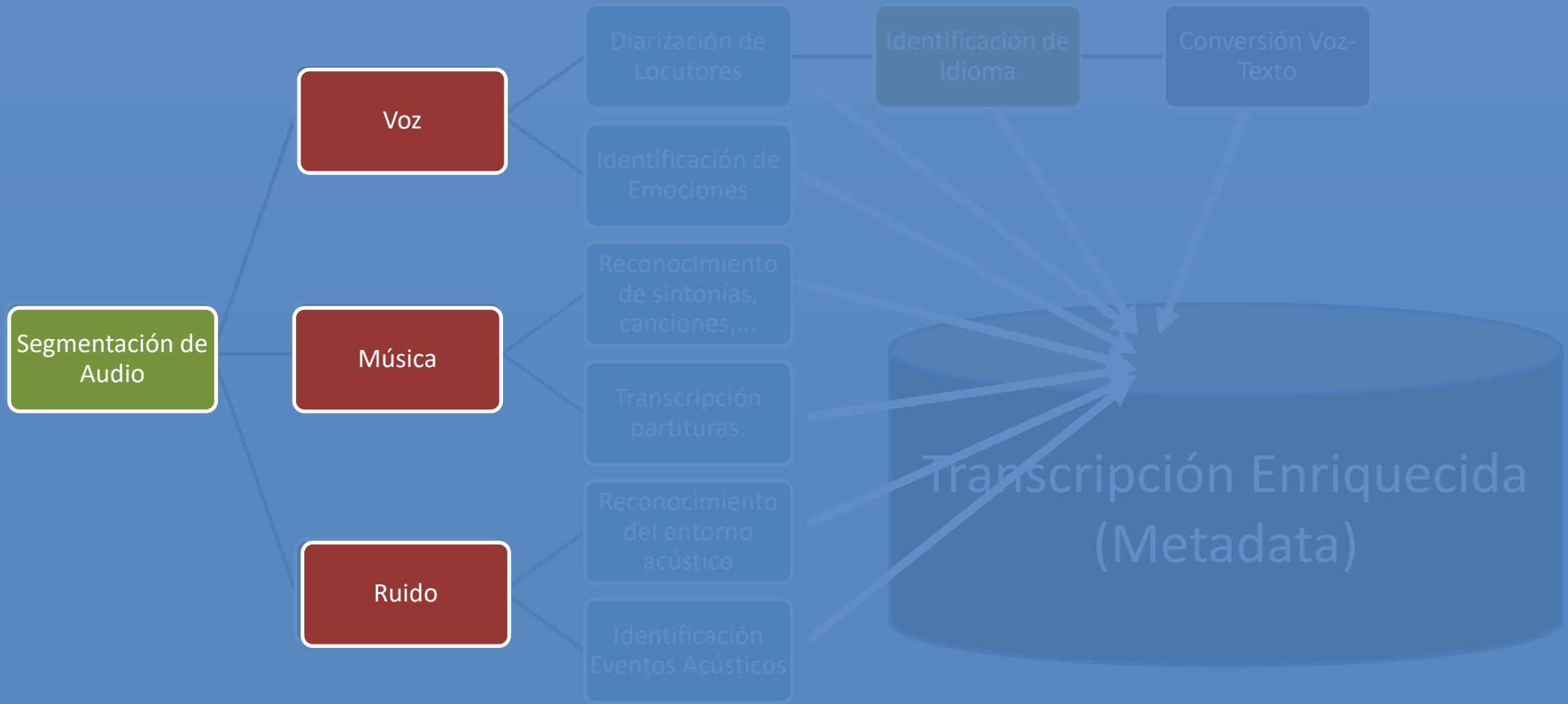
- Hay ruido, música, habla, ...
- Cuántas personas hablan y cuándo habla cada una de ellas
- Cuáles son las identidades de las personas que hablan
- En qué idioma están hablando
- Qué dice cada una de ellas
- Cuál es el estado emocional de cada una de ellas



Tecnologías



Segmentación de Audio



Segmentación de Audio

- ¿Qué es?:
 - Dividir el audio de entrada en fragmentos atendiendo al tipo de contenido acústico: Voz / Música / Ruido y combinaciones de estos.

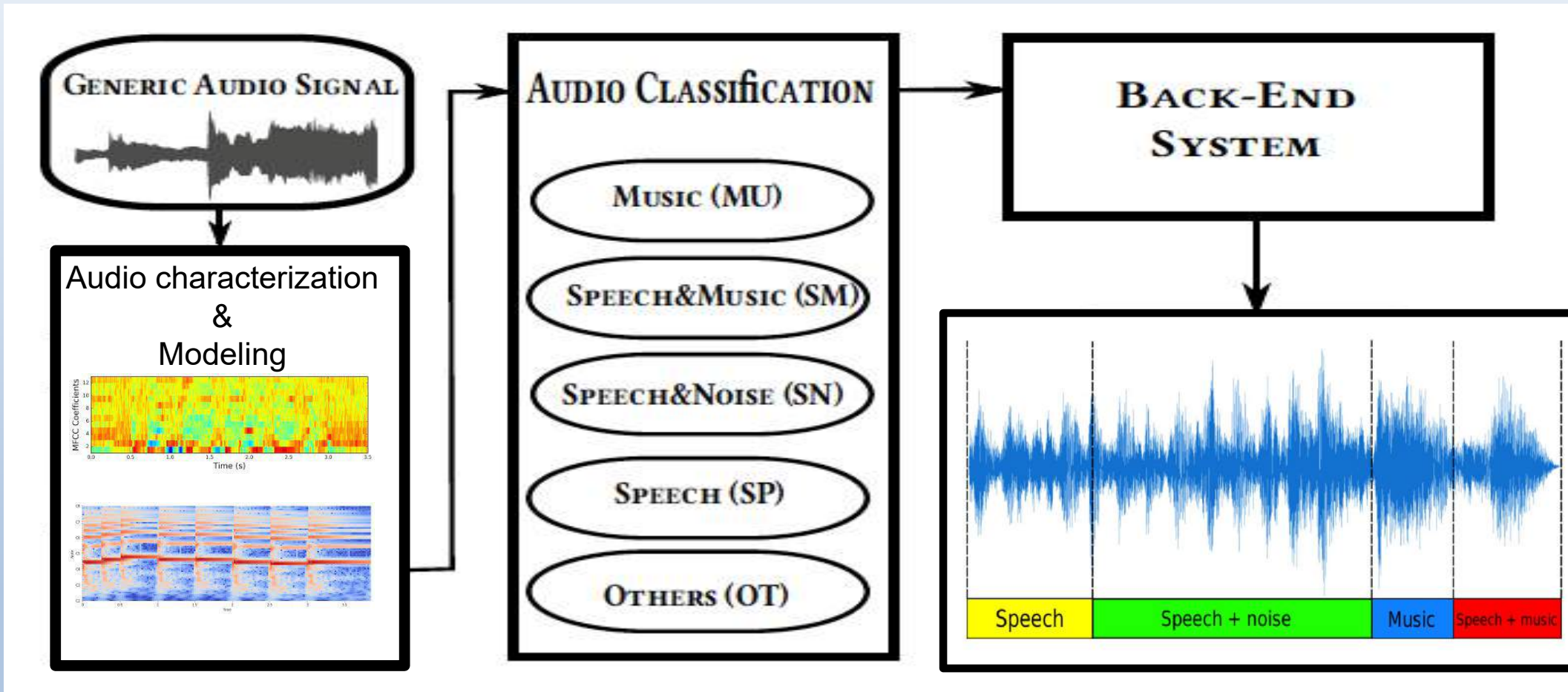


Segmentación de Audio

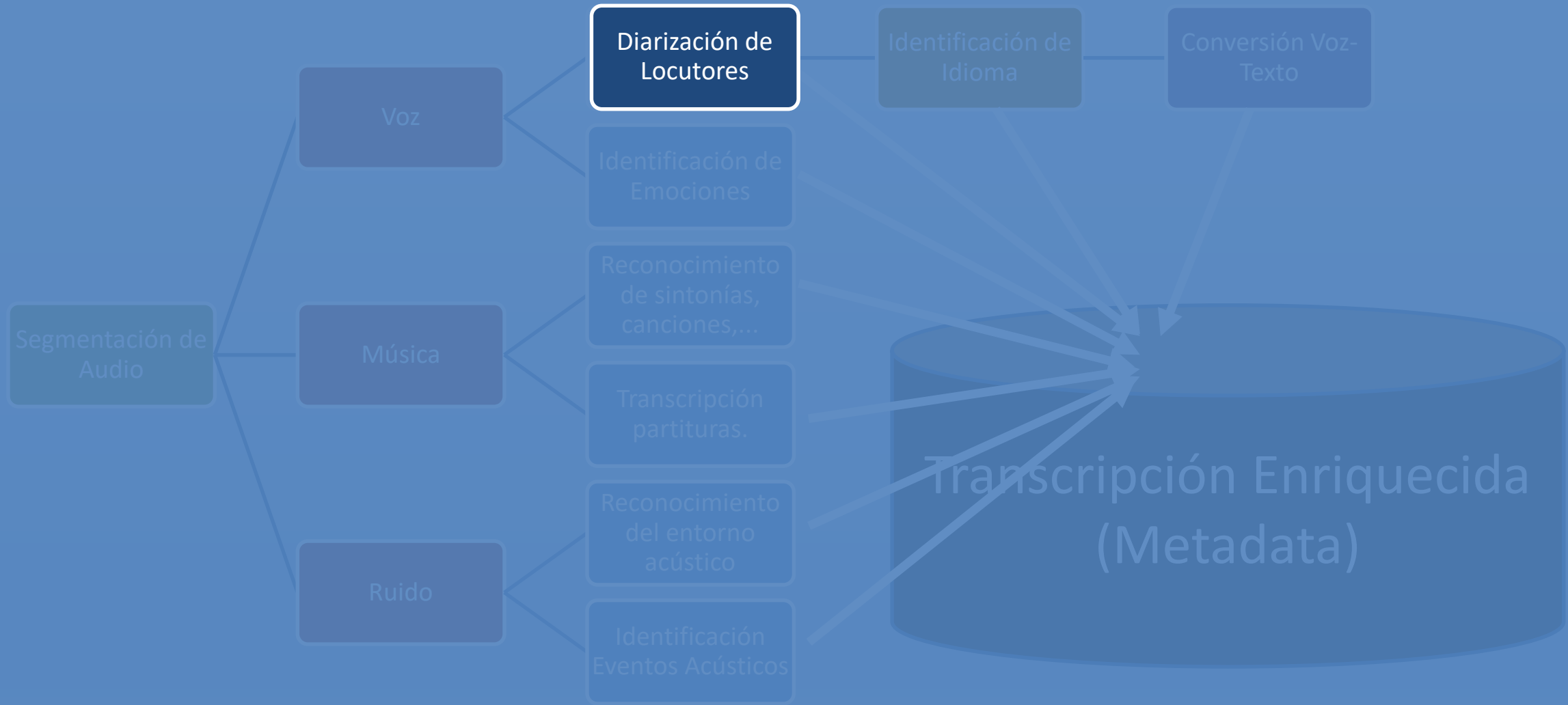
- ¿Para qué sirve?:
 - Da soporte a otras tareas de extracción de información como:
 - Diarización
 - Identificación del hablante
 - Conversión Voz-Texto
 - ...



Segmentación de Audio



Tecnologías



Segmentación y Agrupación de Hablantes

- ¿Qué es?:

- Dividir en fragmentos atendiendo al interviniente y agrupar dichos fragmentos en función de la identidad del locutor.

- Término usado por la comunidad: Diarización

- *Diarise: (Diarize) to make use of a diary to record past events or those planned for the future.*



Segmentación y Agrupación de Hablantes



Segmentación y Agrupación de Hablantes

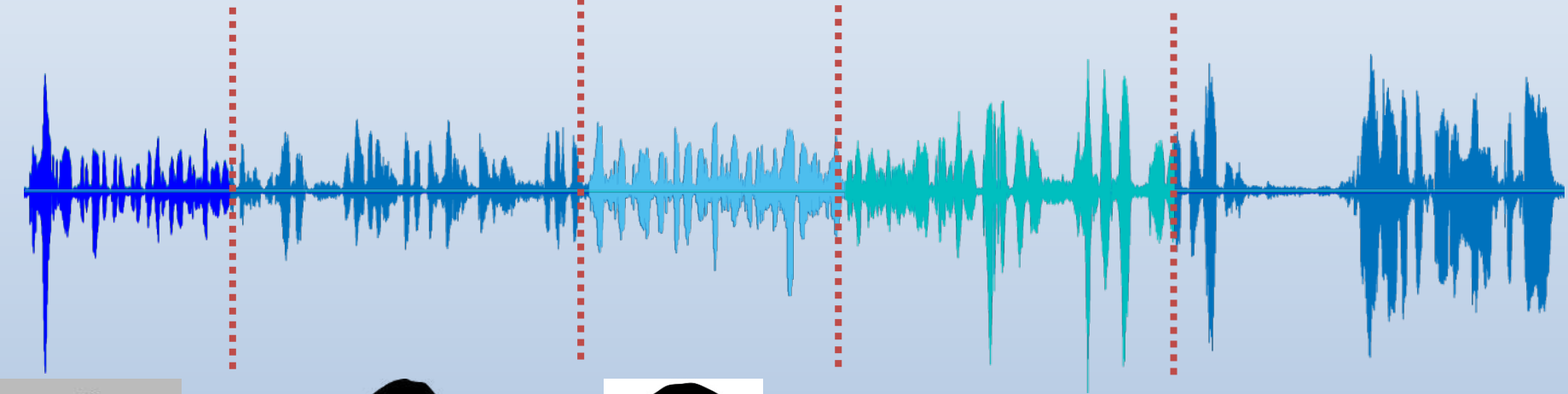
Segm 1

Segm 2

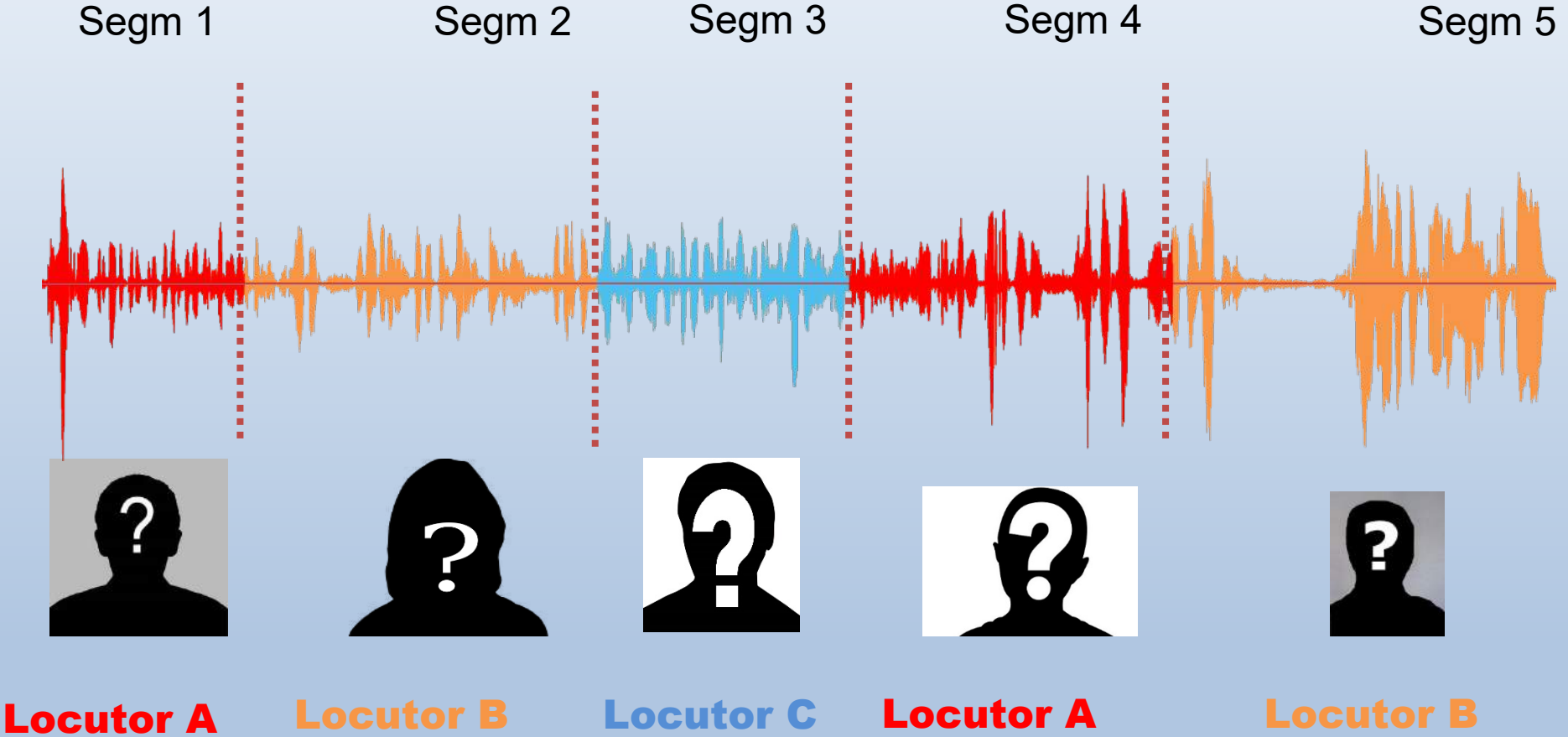
Segm 3

Segm 4

Segm 5

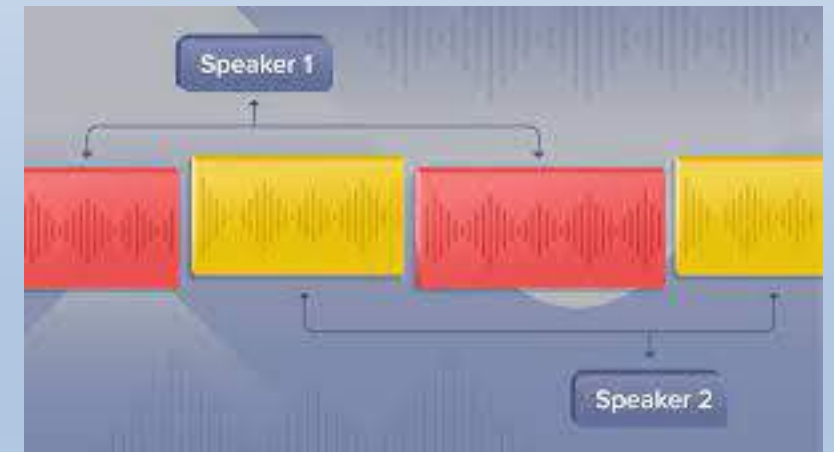


Segmentación y Agrupación de Hablantes



Segmentación y Agrupación de Hablantes

- ¿Para qué sirve?:
 - Tecnología soporte para mejorar prestaciones de:
 - Reconocimiento automático del habla
 - Reconocimiento del hablante
 - ...



Aplicaciones

- Documentación y Edición:



RADIODIFUSIÓN



PARLAMENTOS



REUNIONES



JUZGADOS

- Monitorización, Vigilancia y Seguimiento :



Aplicaciones

- Entornos Educativos, Científicos o Divulgativos:



CHARLAS



CONFERENCIAS



CLASES

- y más ...



SUBTITULADO



**ASISTENTES
PERSONALES**



....



Aplicaciones

- Calidad de Servicio:



CALL - CENTERS



EMERGENCIAS



CENTROS DE GESTIÓN

- Entornos Médicos:



CIRUGÍA



**DIAGNÓSTICO
PSICOLÓGICO**

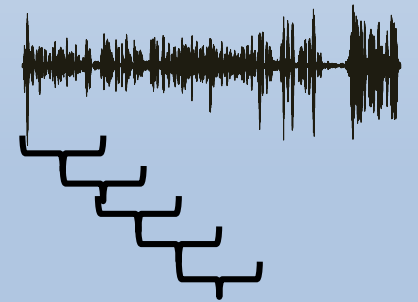
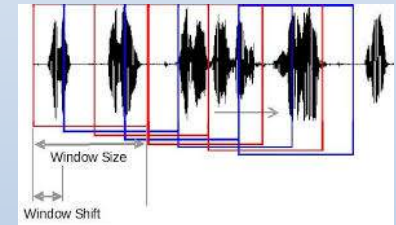
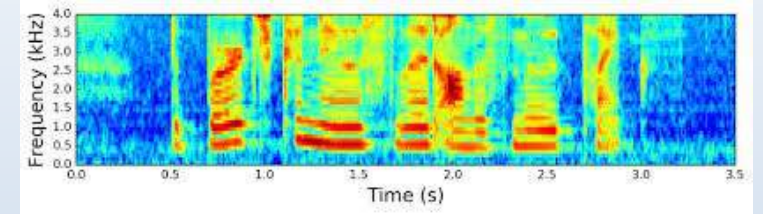


DIAGNÓSTICO PTSD



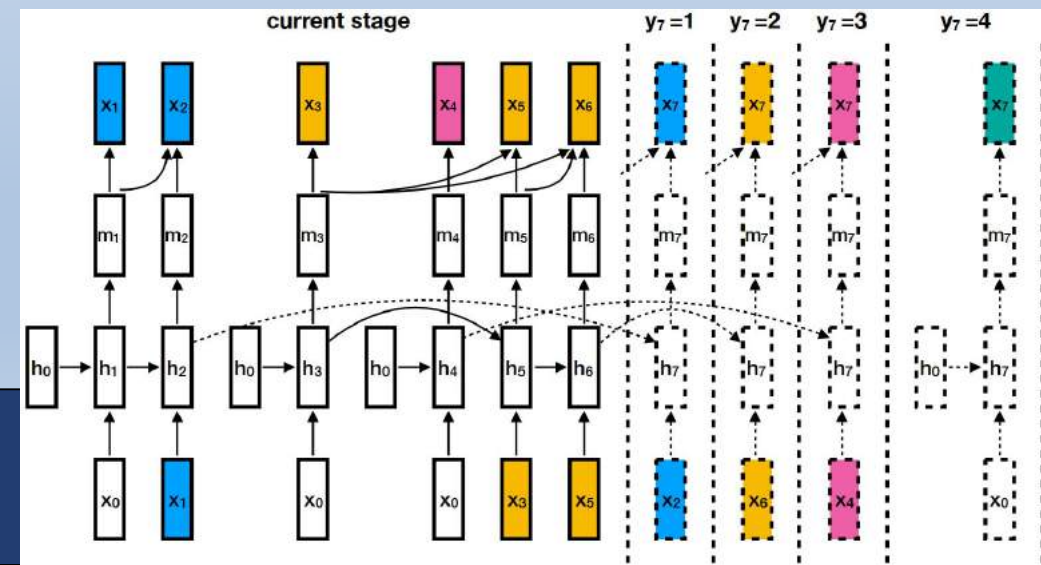
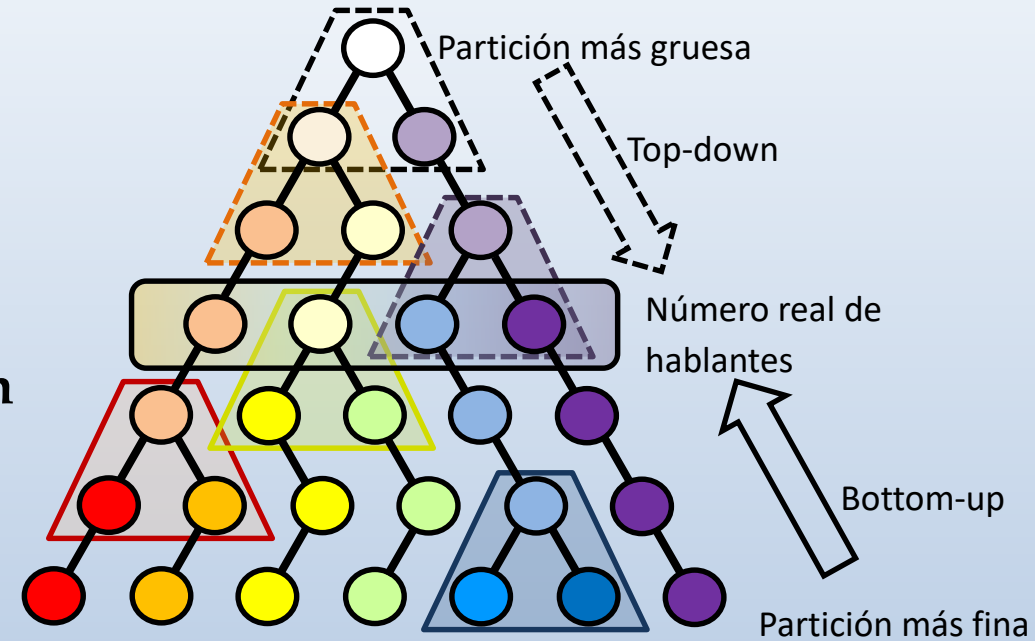
Etapas de la Diarización

- Caracterización de la señal de voz:
 - Descripción espectral a corto plazo
- Segmentación: Detección de Cambio de locutor
 - Estrategias
 - Segmentación y Clasificación
 - Búsqueda de puntos de cambio en las estadísticas de la voz
 - Agrupación de los segmentos mediante clasificación
 - Segmentación por Clasificación
 - División en segmentos de duración constante
 - Búsqueda de fronteras entre hablantes tras la clasificación



Etapas de la Diarización

- Agrupación (Clustering):
 - Estrategias
 - Bottom-up
 - Se parte de muchos fragmentos y se agrupan
 - Top-Down
 - Se parte de un grupo y se subdivide
 - Representación robusta de cada segmento:
 - Embeddings, ...



Medida del Error

- Diarization Error Rate:

$$DER = \frac{T_{incorrecto}}{T_{voz}}$$

- Componentes del error:

- Pérdida:

- Hay voz pero se ha confundido con silencio.

- Falsa Alarma:

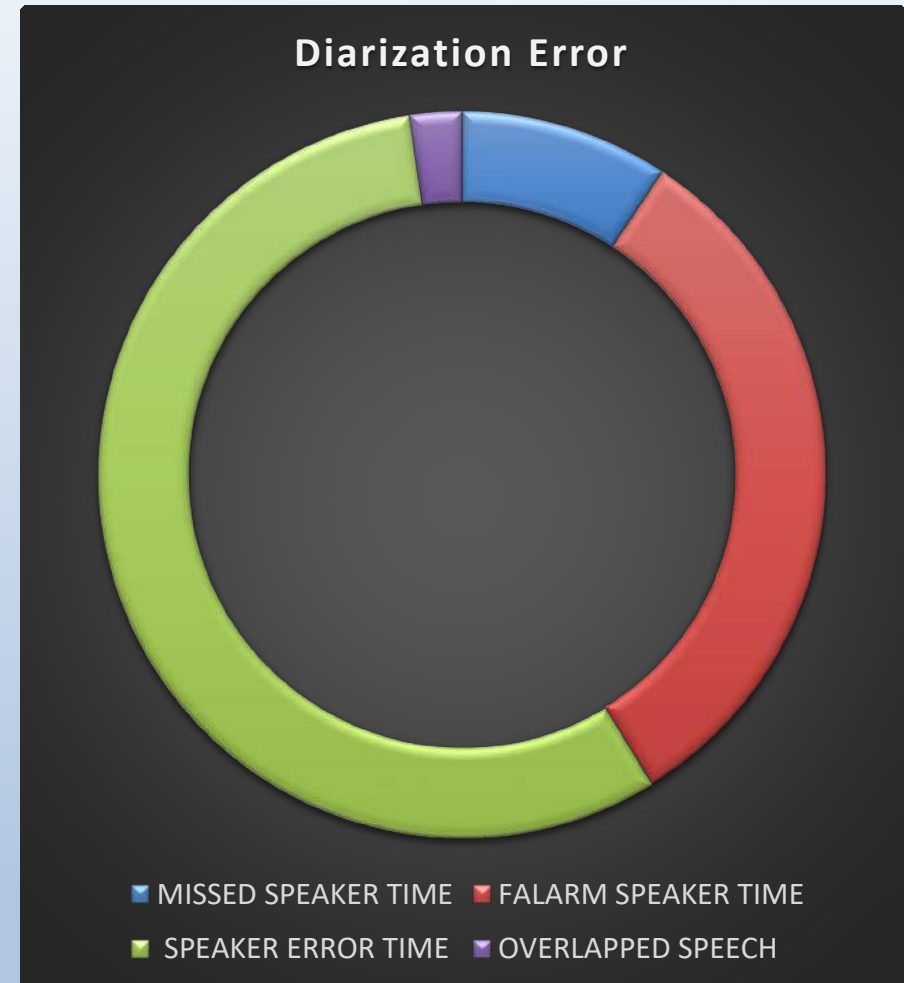
- No hay voz pero se ha detectado erróneamente.

- Error de Locutor:

- Se ha confundido un locutor con otro.

- Error por Solape:

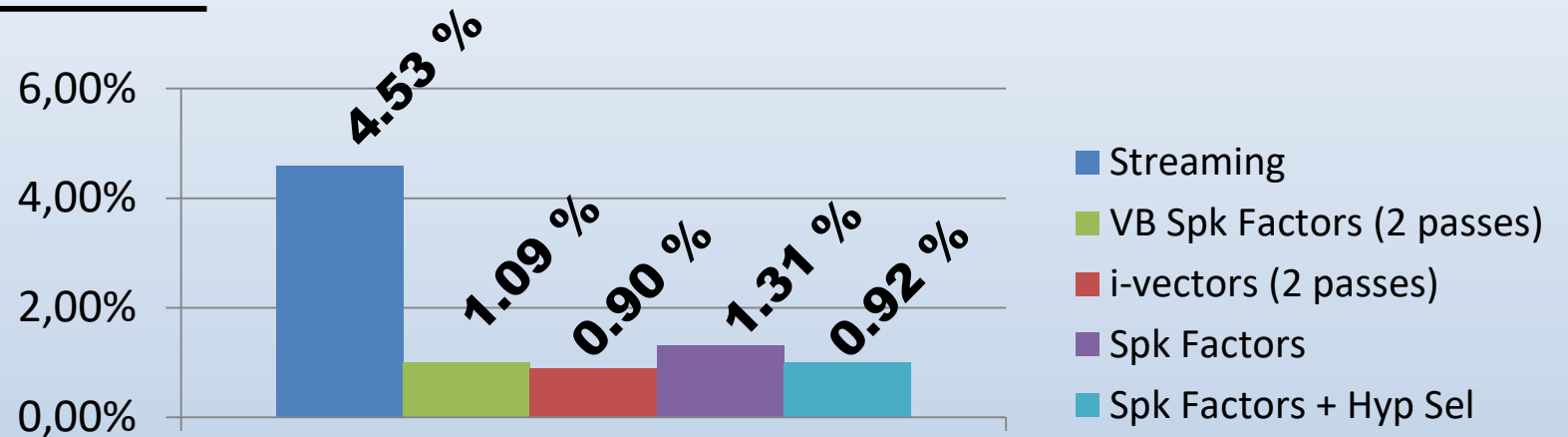
- Dos locutores hablan a la vez, pero solo se ha identificado a uno.



Prestaciones

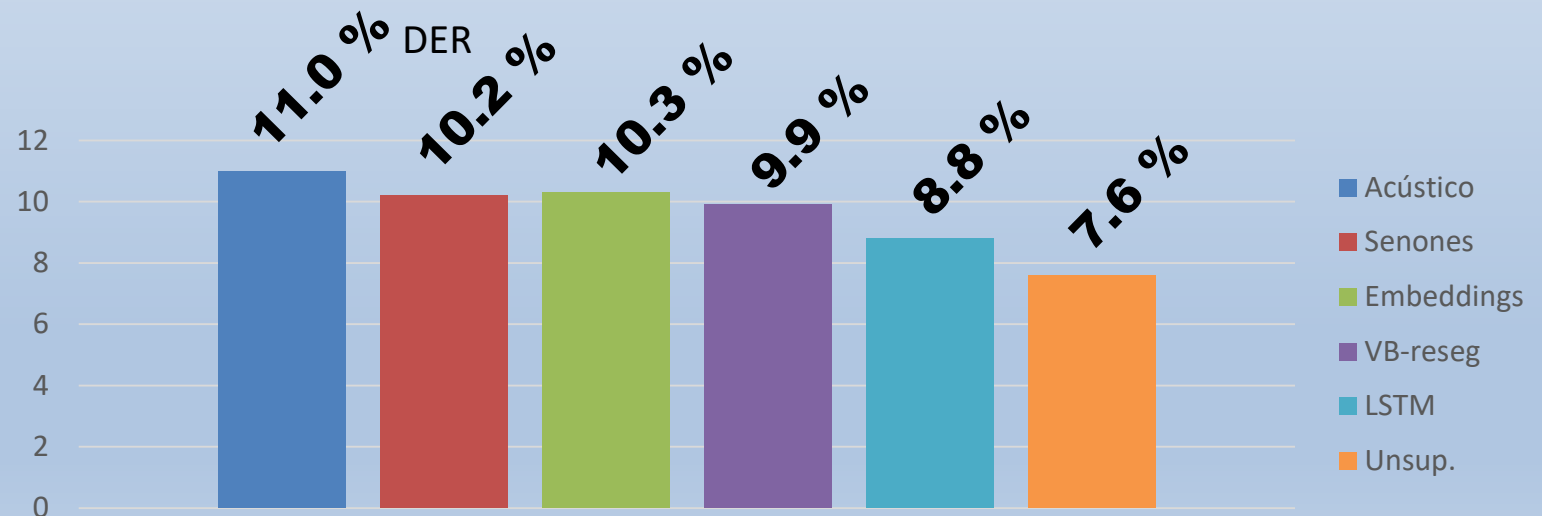
• Entorno Telefónico:

– Dos locutores:



– Dos o más Locutores:

• De 2 a 7

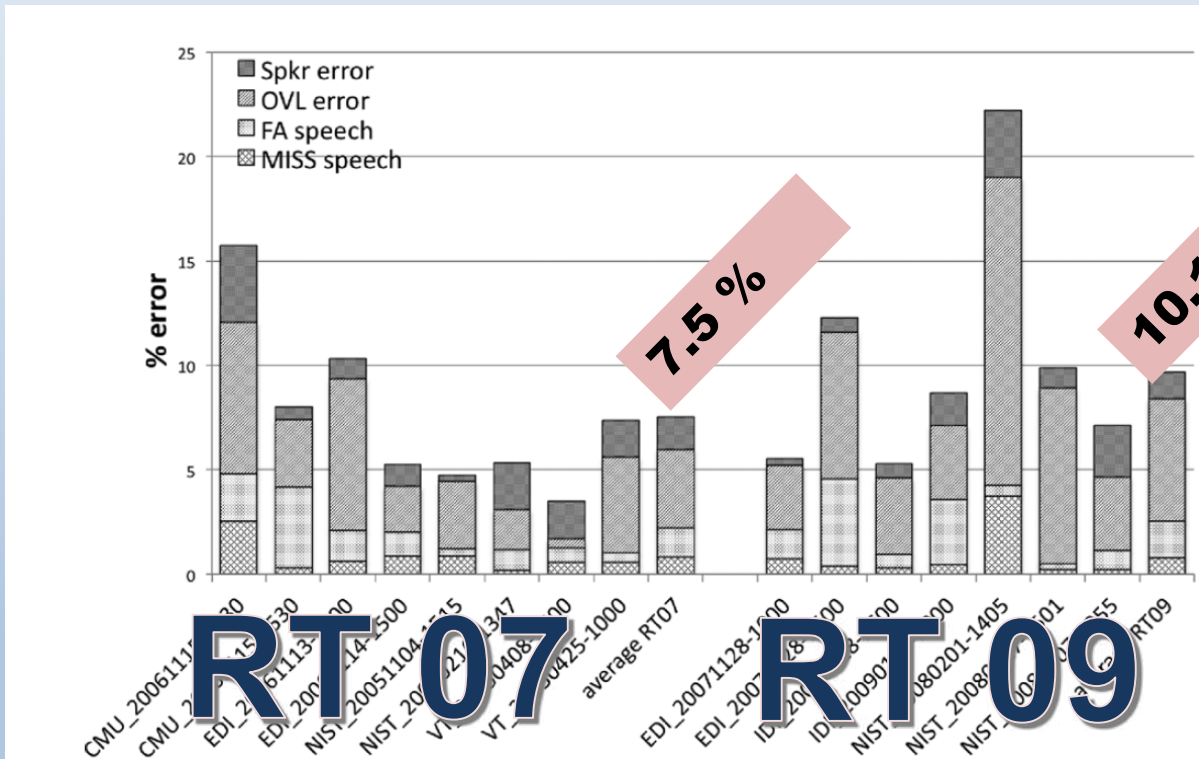


Prestaciones

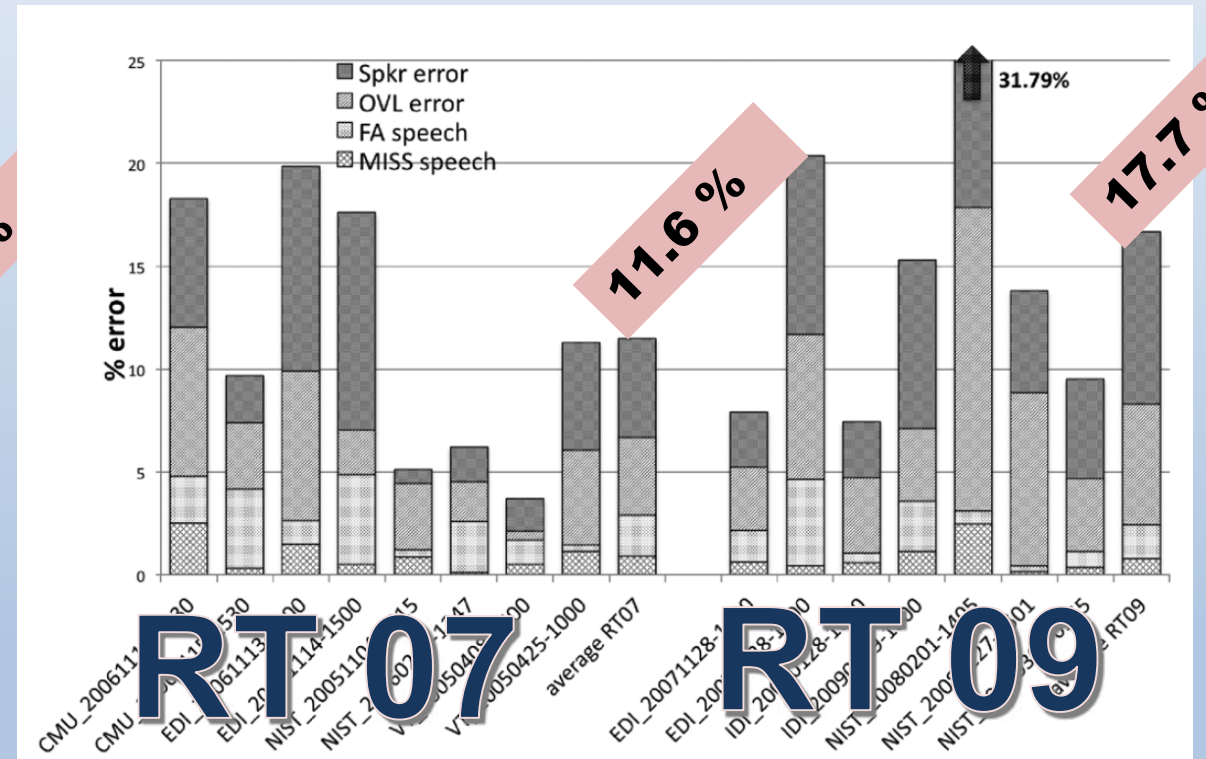


- Reuniones:

– EVALUACIONES NIST RICH TRANSCRIPTION 2007 Y 2009



MÚLTIPLES MICRÓFONOS



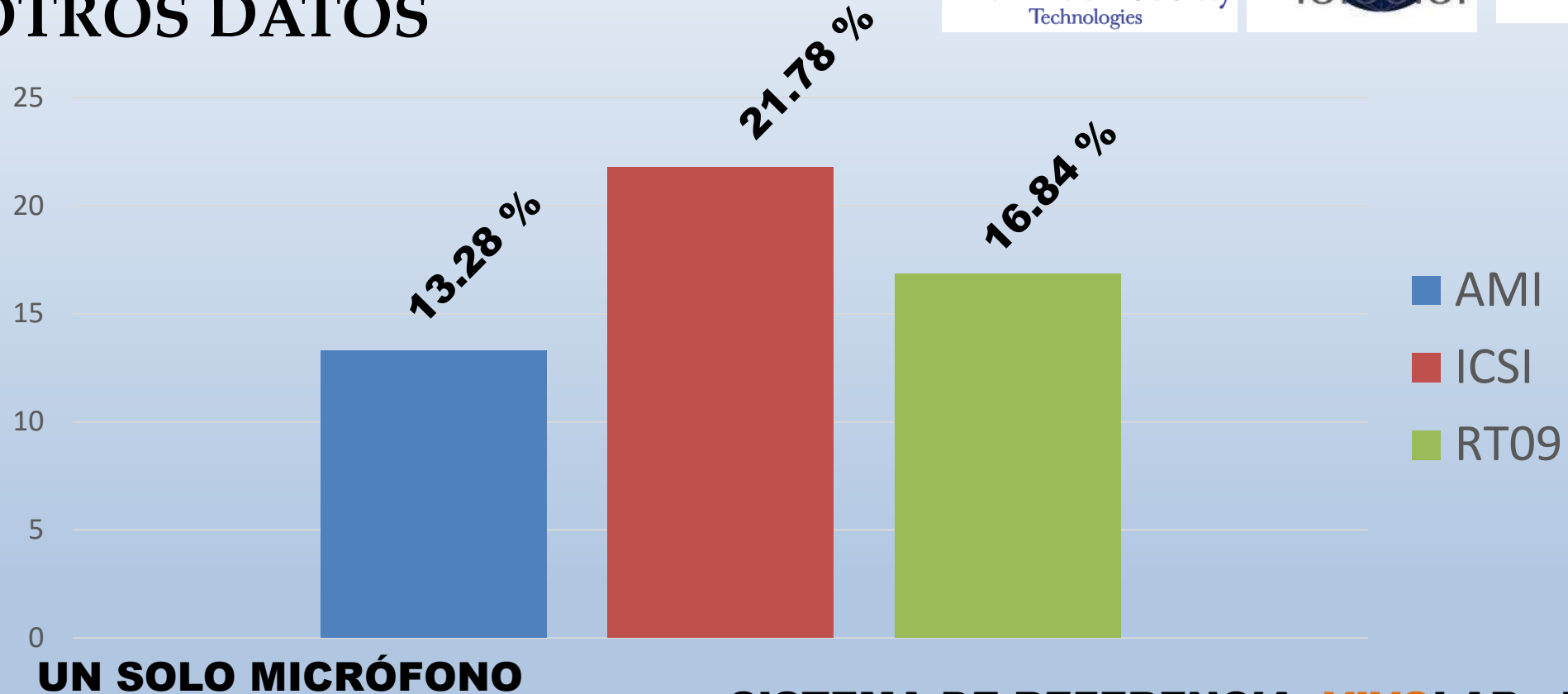
UN SOLO MICRÓFONO



Prestaciones

- Reuniones:

- OTROS DATOS



SISTEMA DE REFERENCIA: **VIVOLAB** - UNIZAR 2018

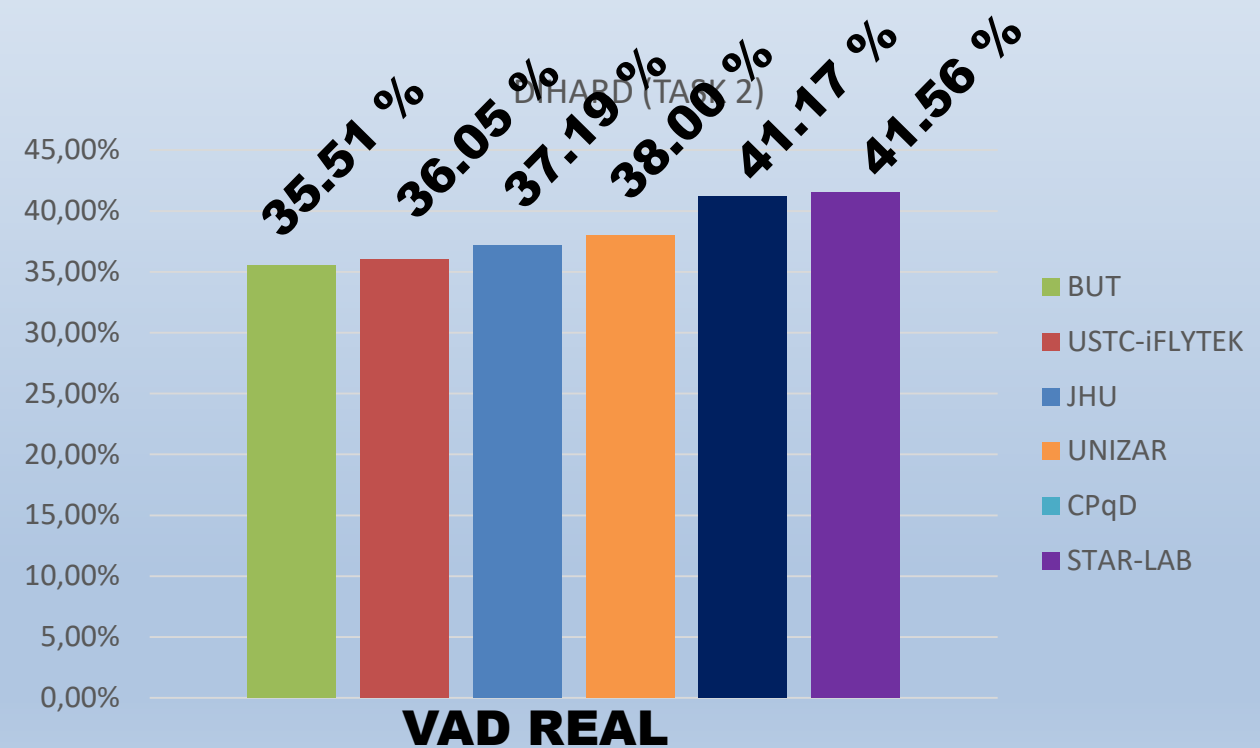
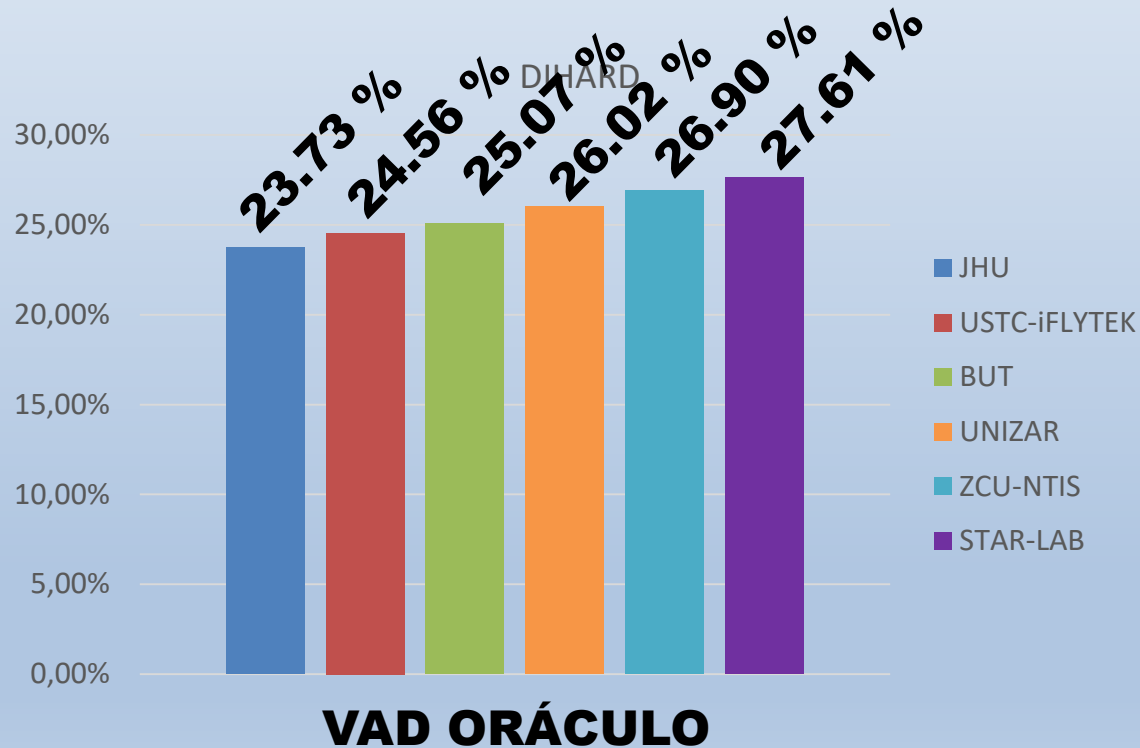


Prestaciones

- Entornos Variados:

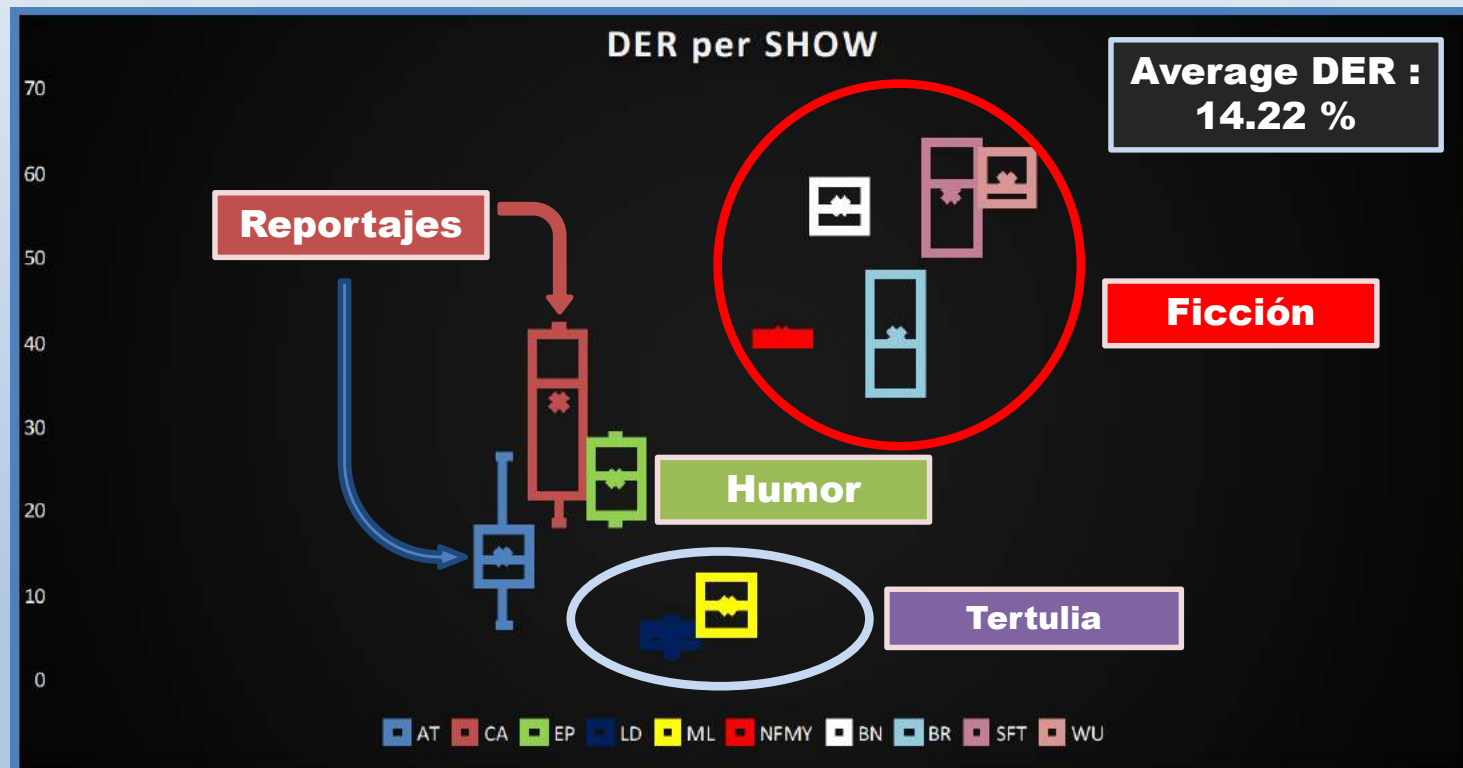
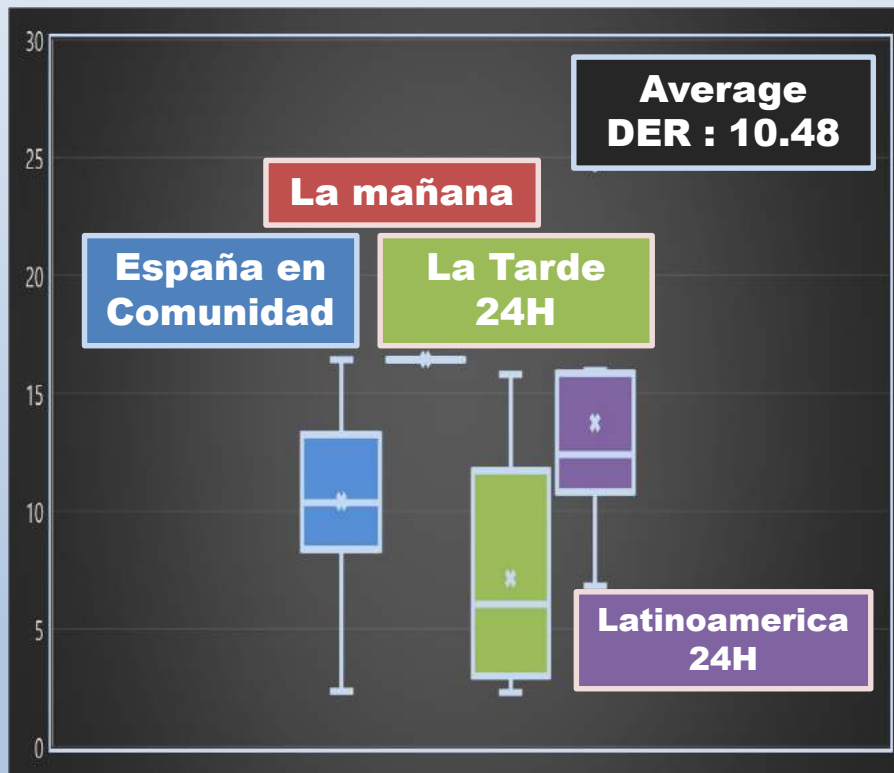
- DIHARD (2018)

- Reuniones, teléfono, habla leída, madre-bebé, restaurante, ...

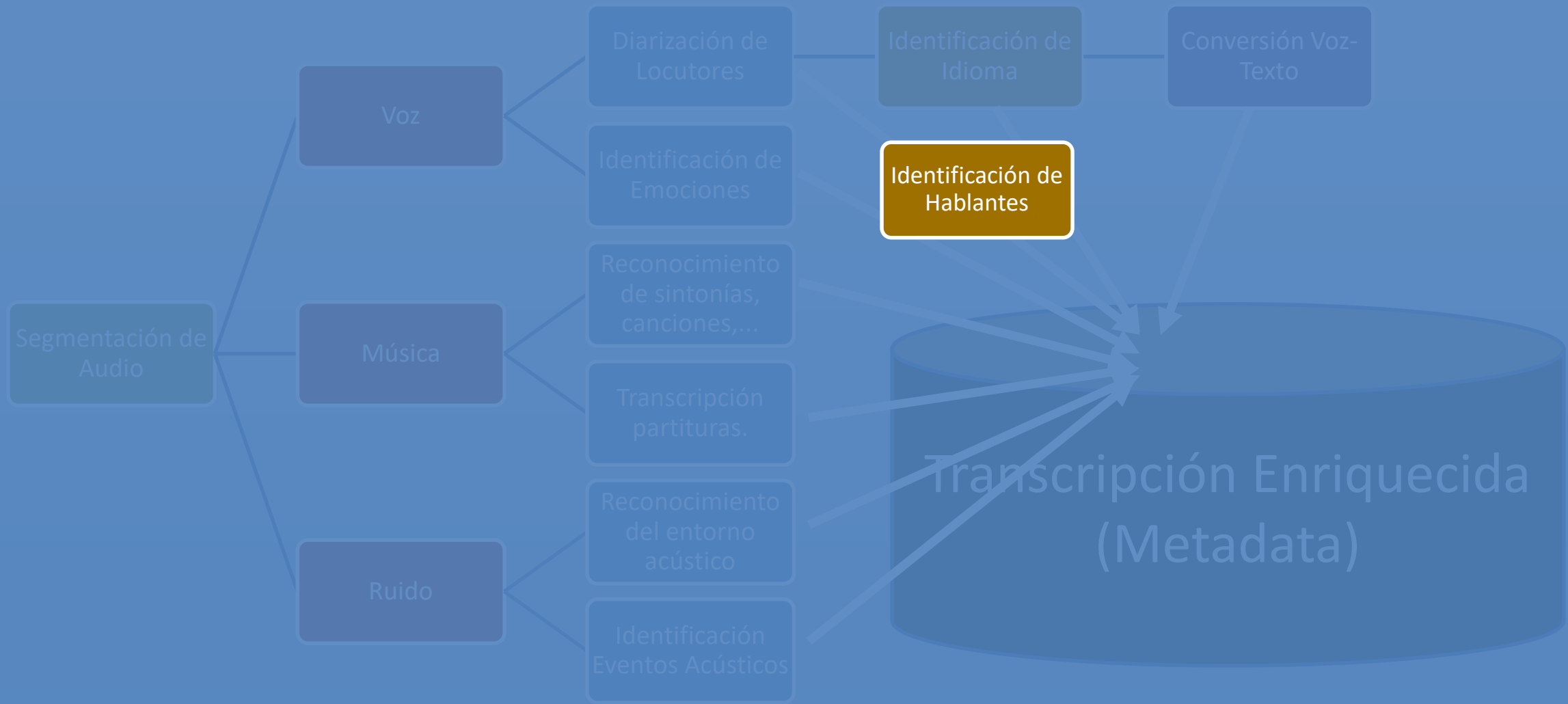


Prestaciones: Albayzin 2018 y 2020

- Broadcast:



Tecnologías



Identificación de Hablantes

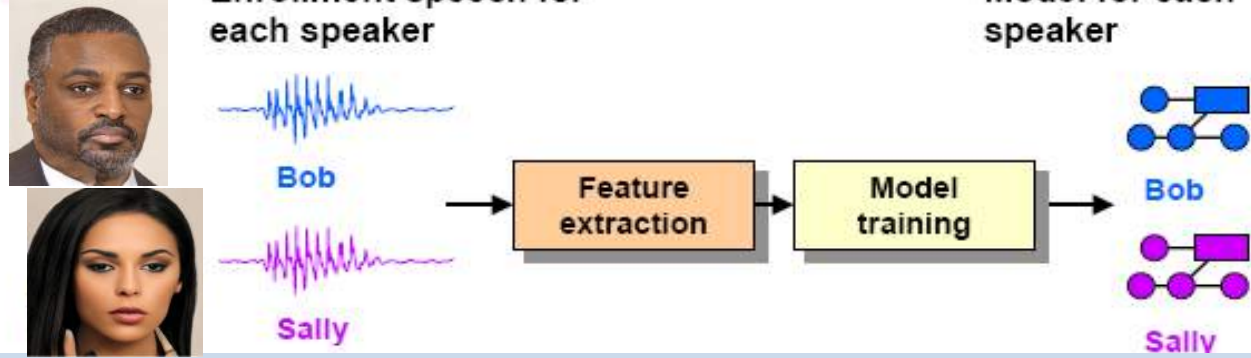
- ¿Para qué sirve?:

- Permite asignar identidades concretas a fragmentos de audio de un contenido analizado



Identificación de Hablantes

Enrollment Phase

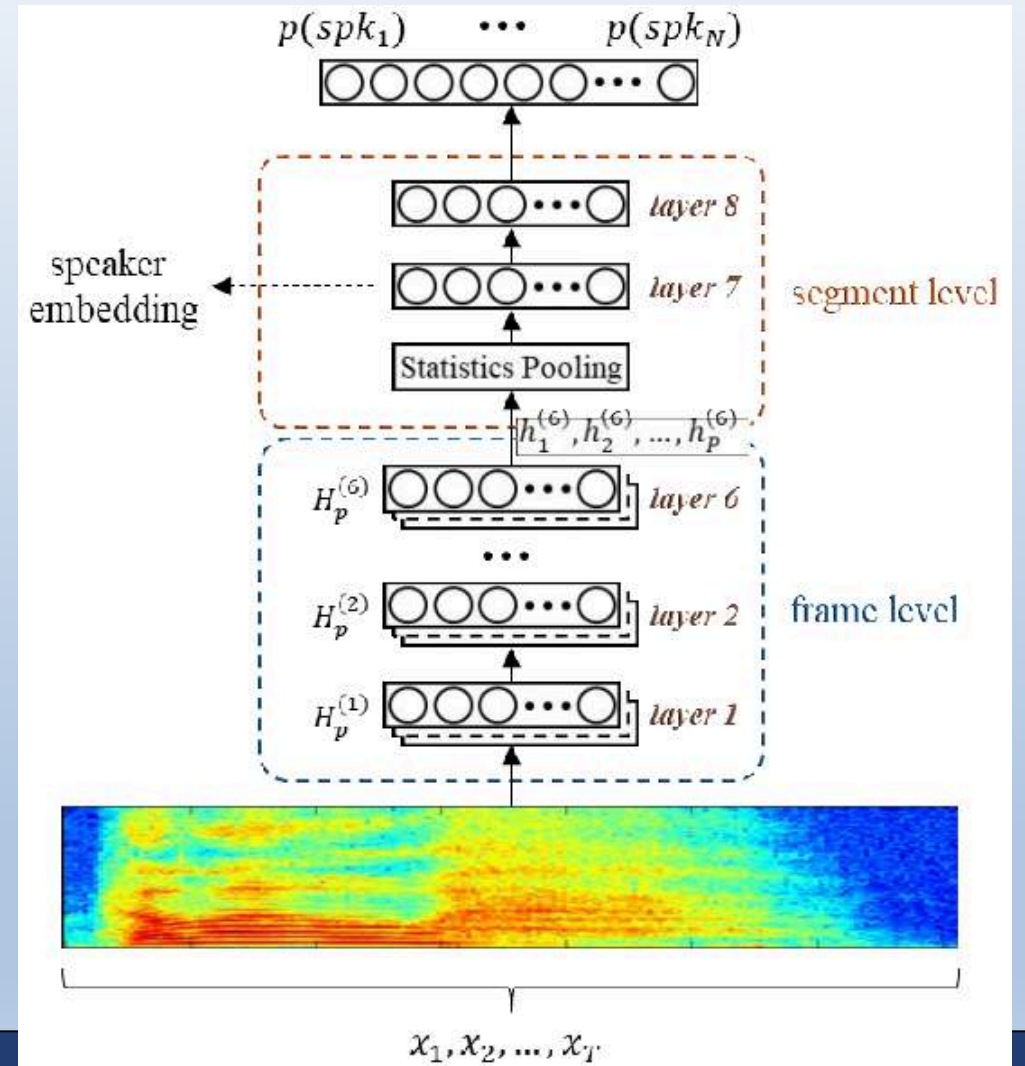


Audio



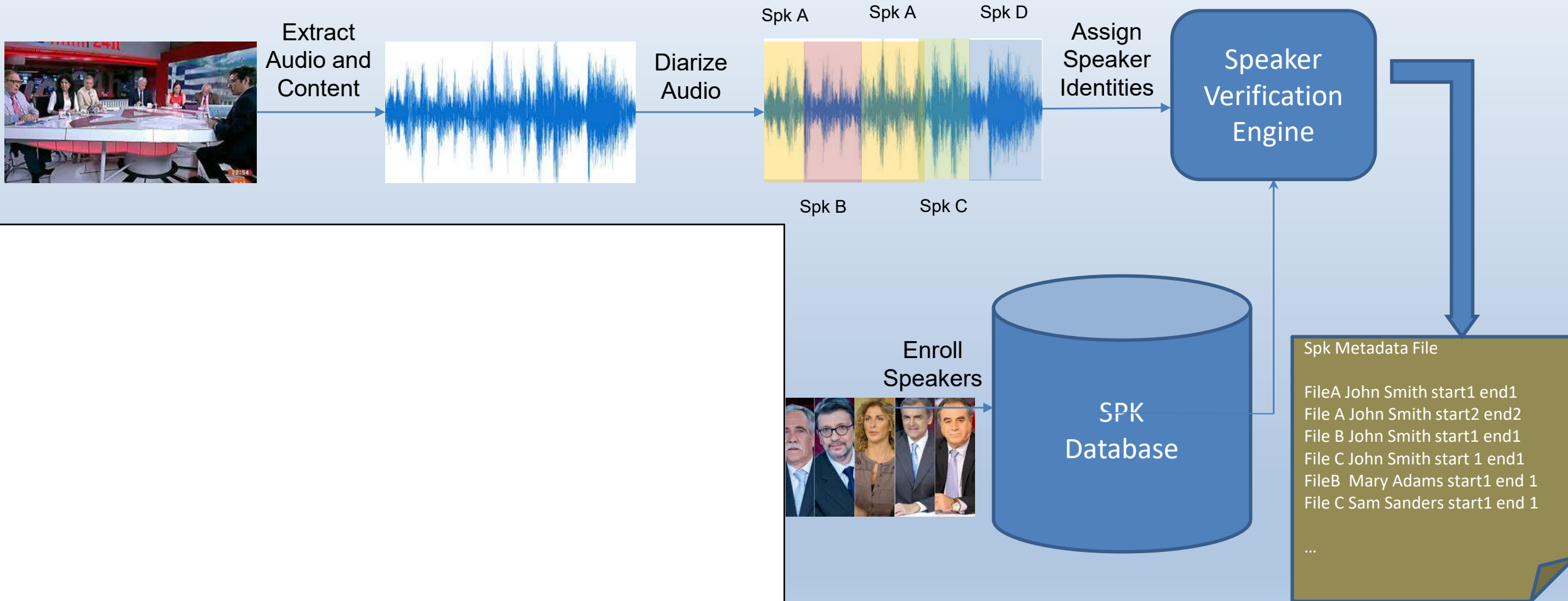
Attribution

Identities



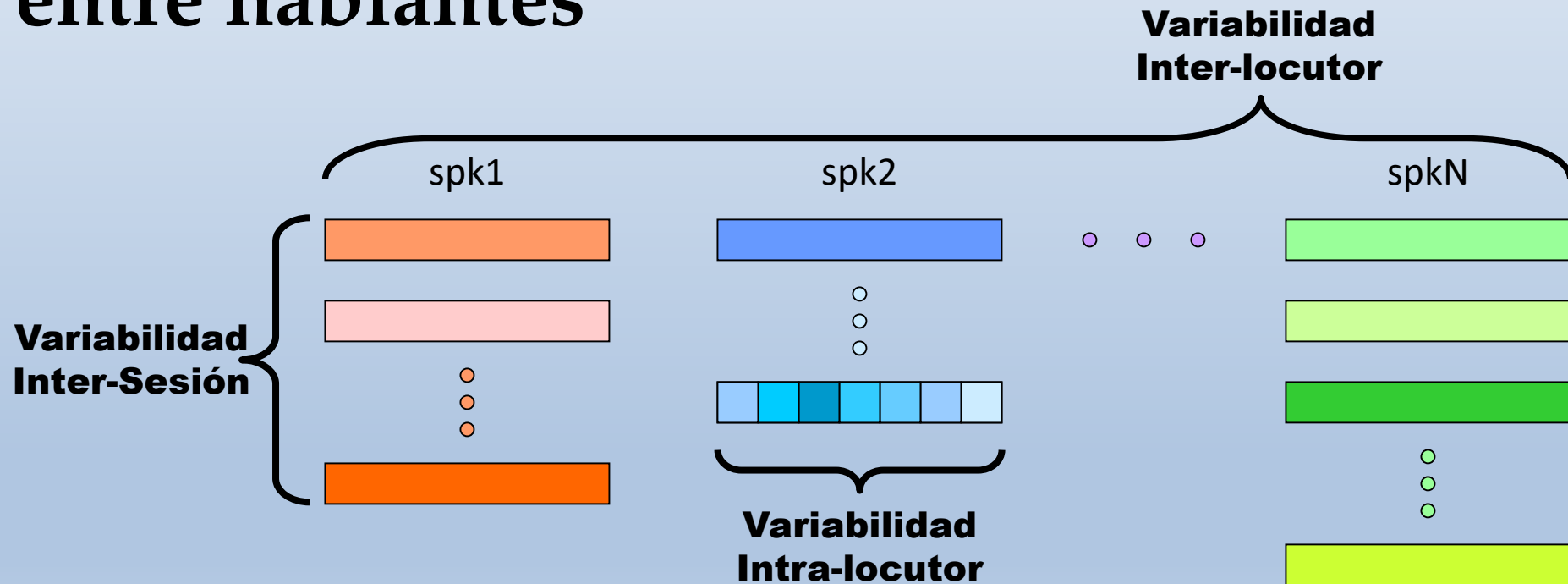
Identificación de Hablantes

- Reconocimiento de personajes en programas de TV:

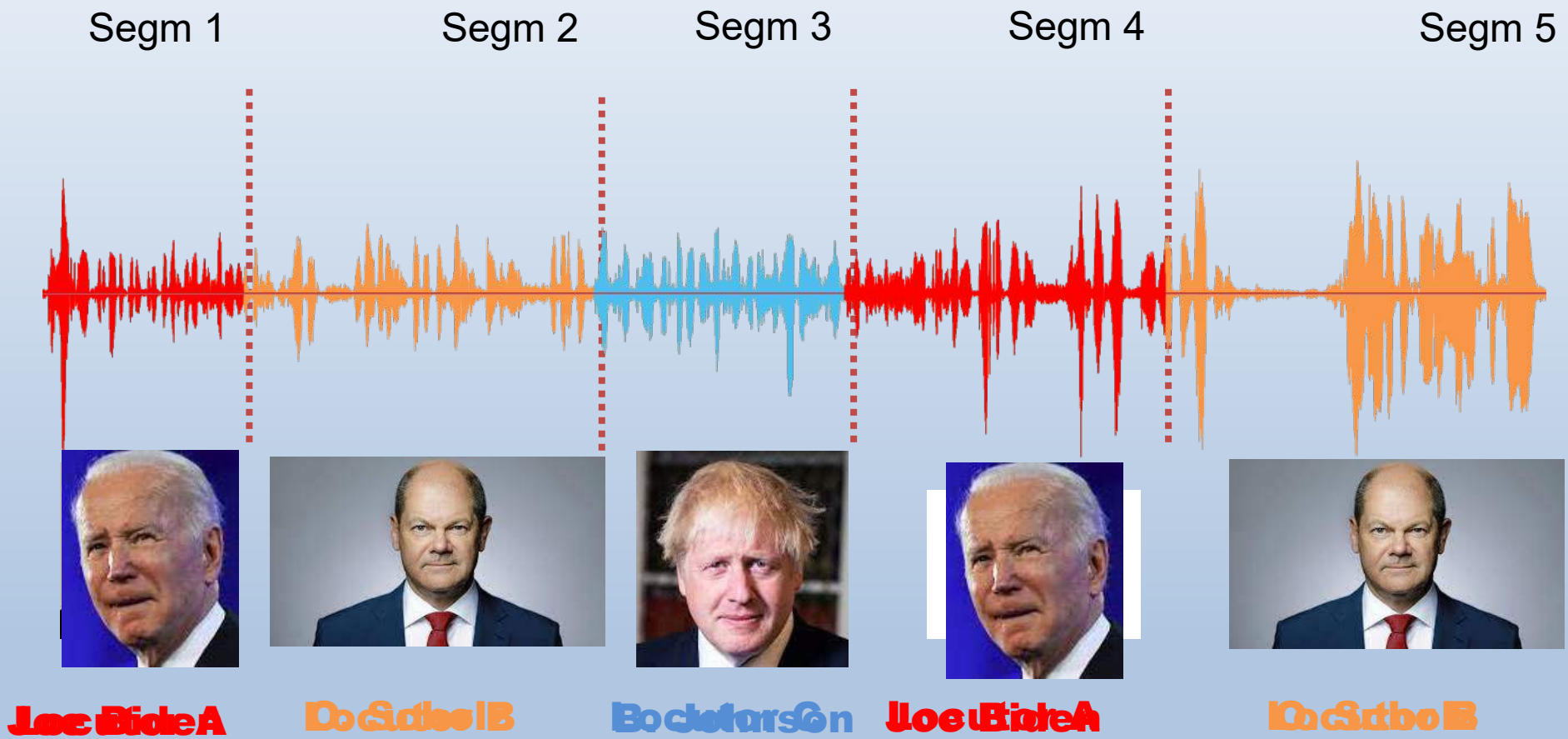


Retos en la Identificación de hablantes

- Alta variabilidad en la voz de los hablantes
- Diversidad de dominios acústicos
- Solape entre hablantes



Diarización Junto con Identificación de Hablante:



Prestaciones: Albayzin 2020

- Iberspeech-RTVE-Challenge :*



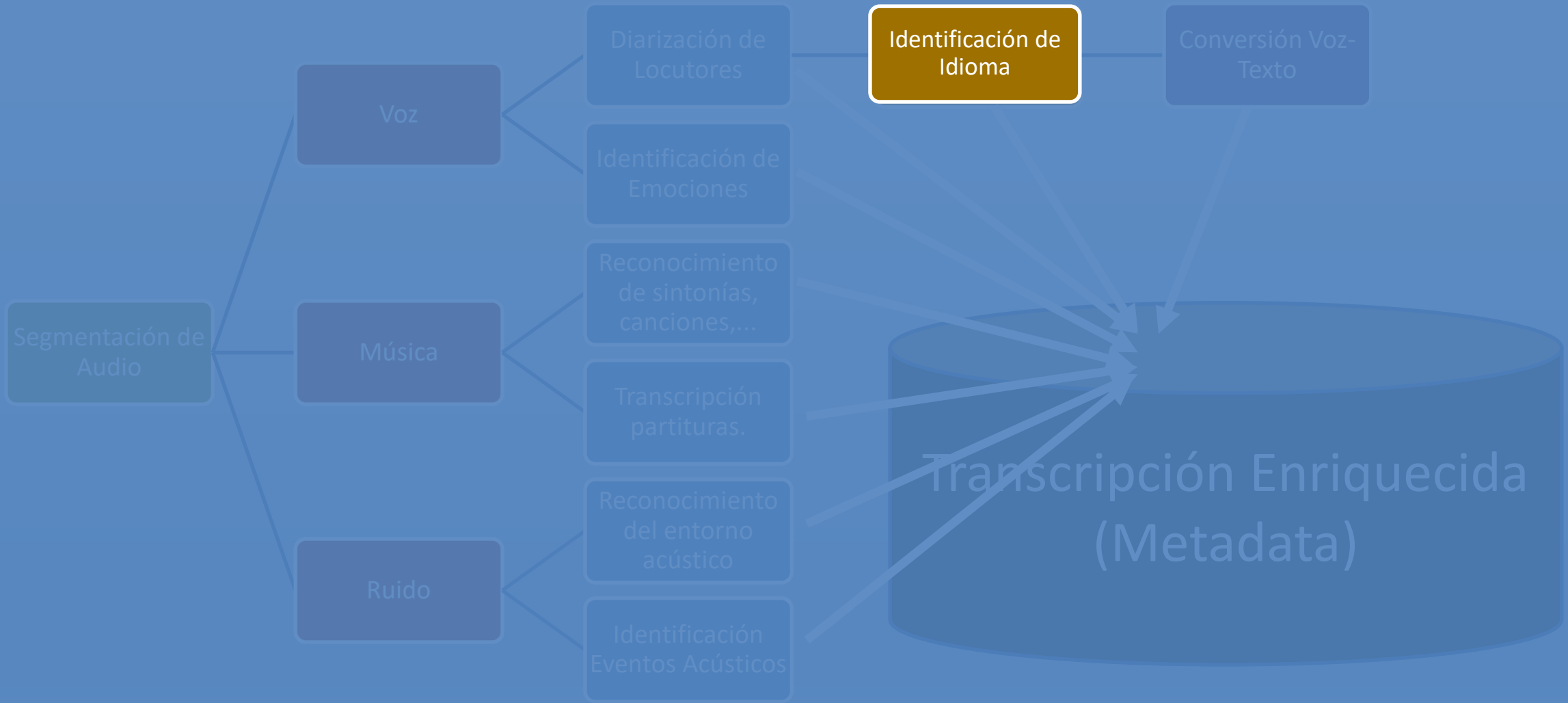
	AER
BIOMETRIC VOX	65.09 %
VIVOLAB	72.63 %

	MISS	FA	SPK ERR
BIOMETRIC VOX	47.0 %	9.2 %	8.9 %
VIVOLAB	5.1 %	53.3 %	14.2 %

Subset	Closed Condition			Open Condition		
	Direct	Indirect	Hybrid	Direct	Indirect	Hybrid
Dev. subset	13.73	15.27	15.89	41.91	37.45	37.68
Eval. subset	25.11	17.20	16.49	65.31	60.34	31.95



Tecnologías

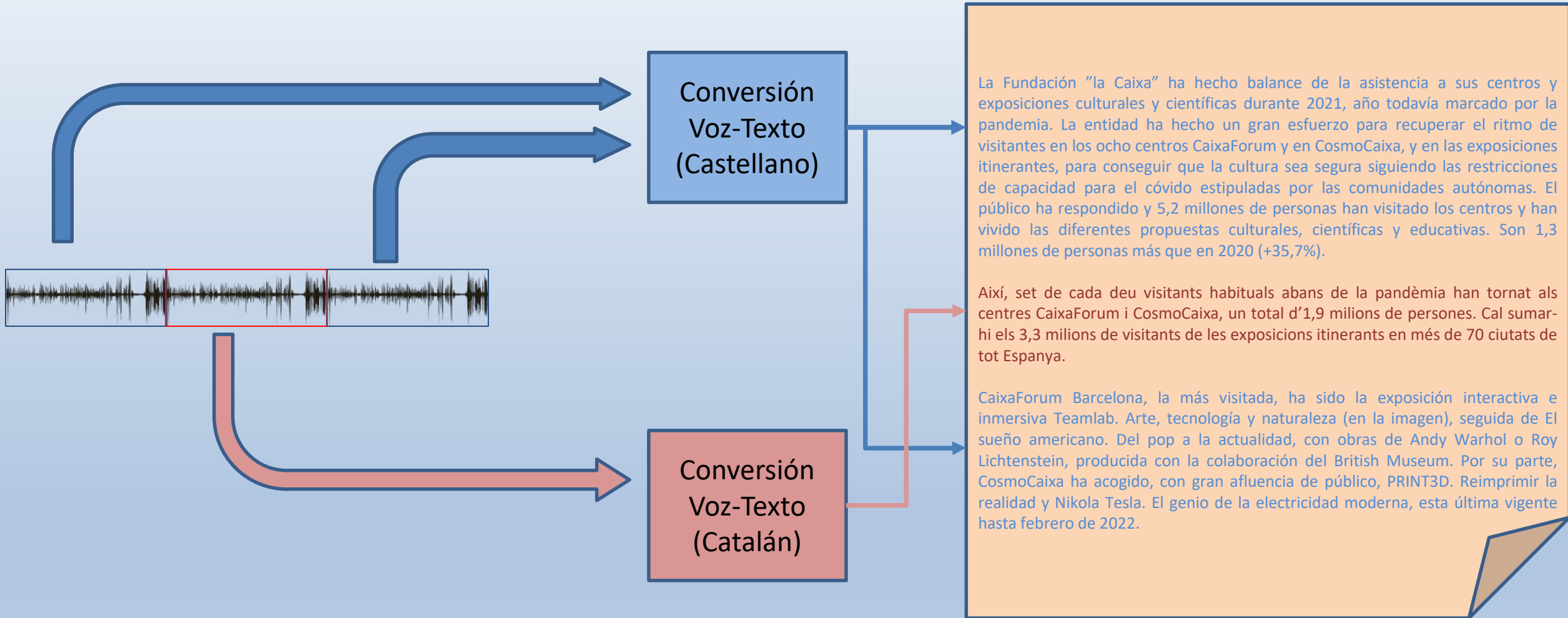


Identificación de Idioma

- ¿Para qué sirve?:
 - En entornos multilingüe, permite el indexado y la recuperación de documentos:
 - Esencial en esos entornos como soporte a:
 - Reconocimiento automático del habla



Identificación de Idioma



Identificación de Idioma

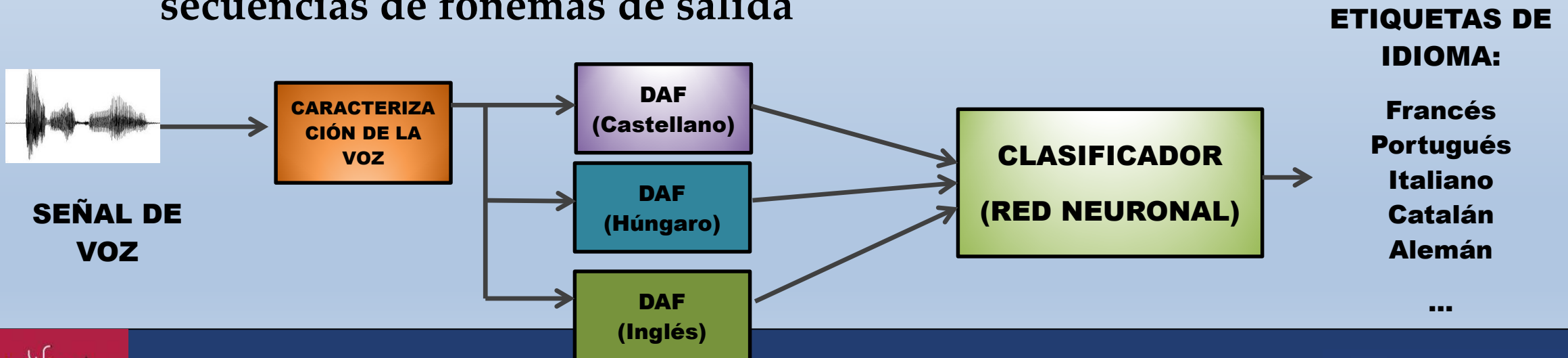
- **Tecnologías:**

- Acústicos

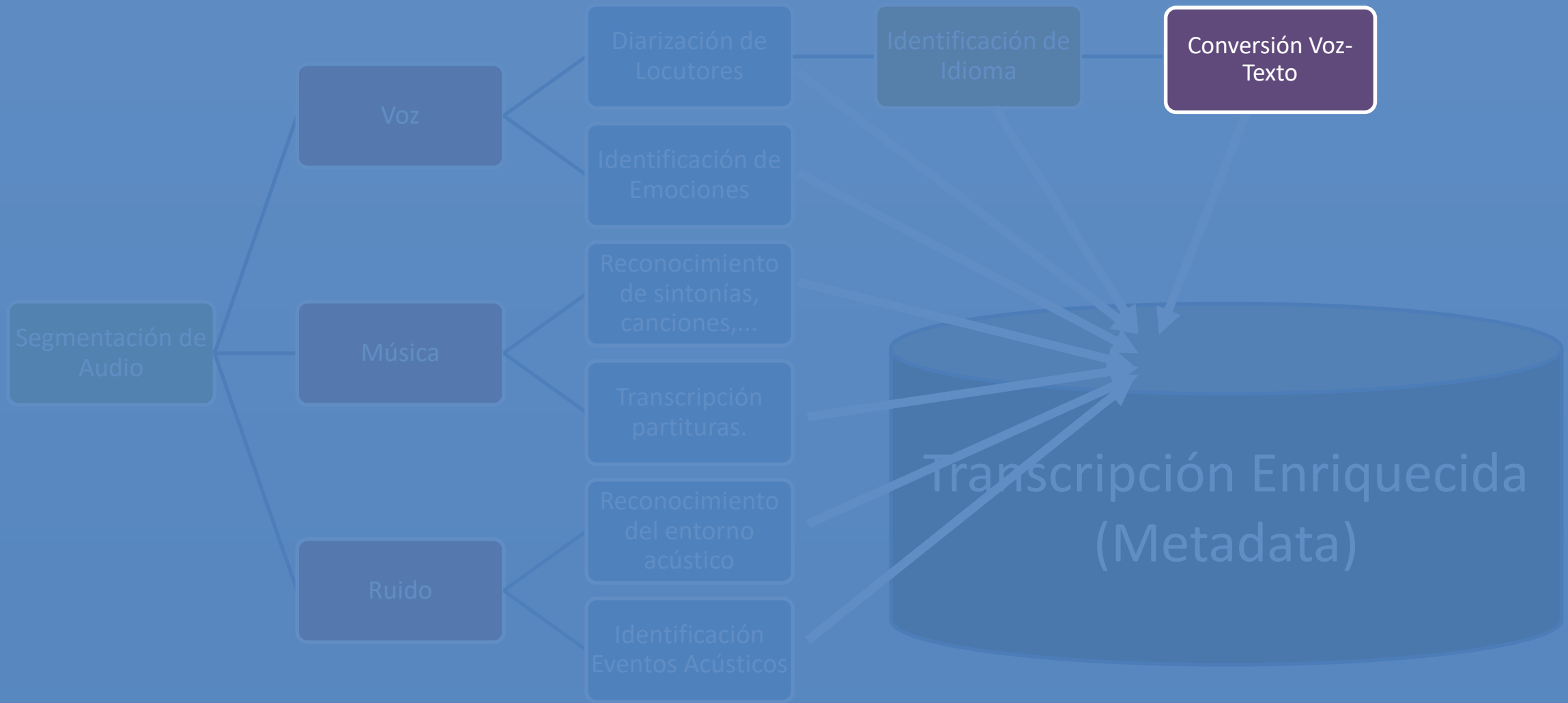
- Tratan de buscar patrones discriminativos directamente sobre la señal de voz

- Fonotácticos (Lingüísticos)

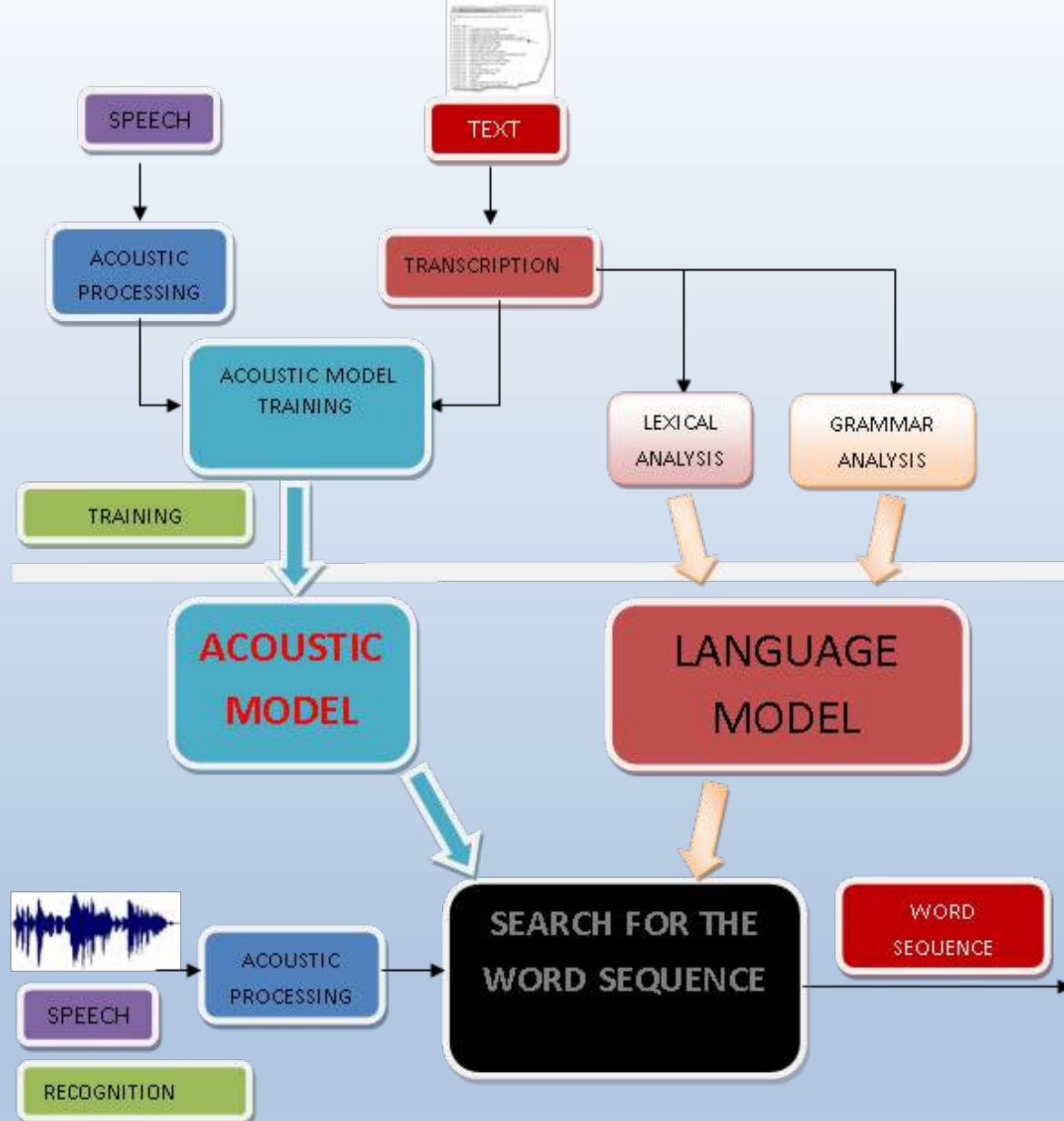
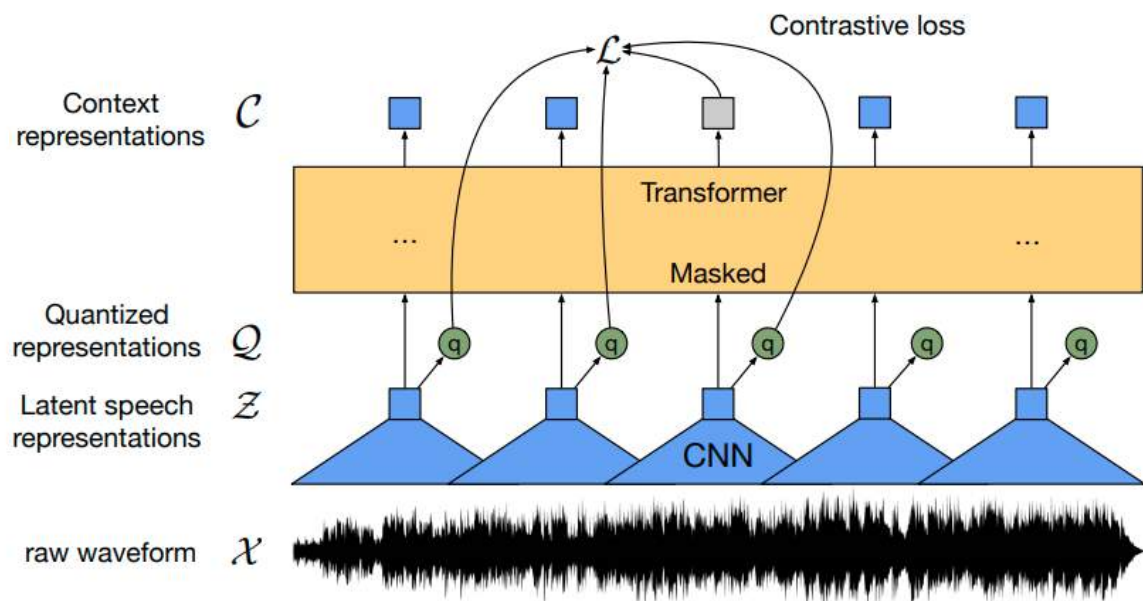
- Primero procesan la señal de entrada con un (o varios) reconocedor fonético (en varios idiomas) después buscan patrones discriminativos en las secuencias de fonemas de salida



Tecnologías



Etapas y Procesos Reconocimiento Automático del Habla:



Componentes de un Sistema de RAH

■ **MODELO ACÚSTICO**

- Describe las características de cada unidad desde el punto de vista de la señal de voz (espectralmente)

■ **MODELO DE LENGUAJE**

- Describe las relaciones entre palabras del vocabulario
- Cuantifica la probabilidad de las secuencias de palabras

■ **MODELO LÉXICO**

- Describe cómo se forma cada palabra del vocabulario a partir de las diferentes unidades del modelo acústico.



ERRORES EN UN SISTEMA RAH

- **Borrados**
 - El locutor dice algo pero el sistema no devuelve nada
- **Substituciones**
 - El sistema devuelve a su salida una palabra diferente de la pronunciada por el locutor.
- **Inserciones**
 - El locutor no dice nada, pero el sistema devuelve alguna palabra (generalmente debido a artefactos acústicos)



ERRORES EN UN SISTEMA RAH

- Métricas de Precisión y Error:

REF: a las tres **y siete** minutos de mañana
HYP: a las tres **diecisiete** minutos de **la** mañana

CORRECTO (C)

ERRORES:

Substituciones (S), Borrados (B), Inserciones (I)

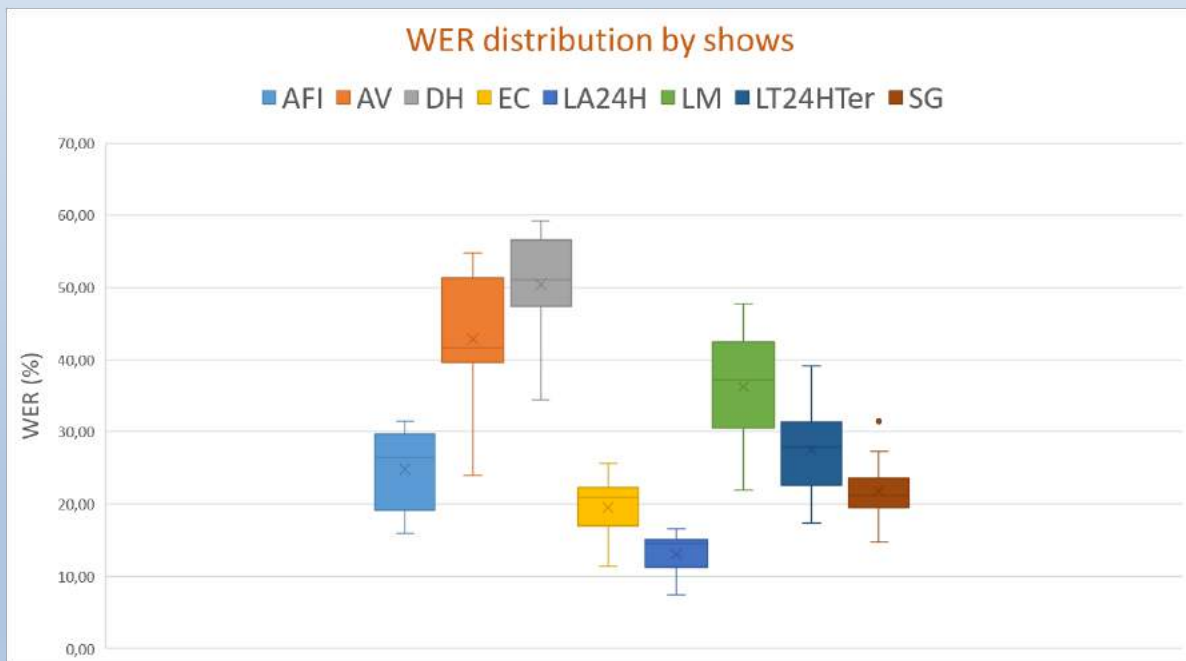
$$\% \text{ ACC} = \frac{C}{C+S+B+I} \times 100$$

$$\% \text{ WER} = \frac{S+B+I}{C+S+B} \times 100$$

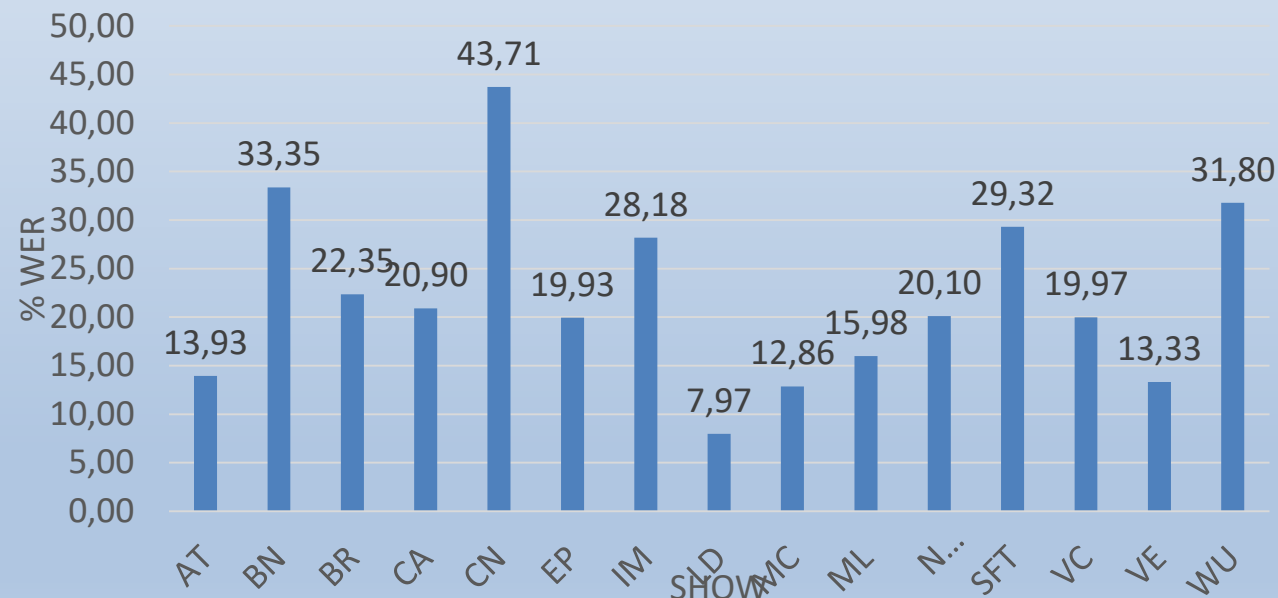


Prestaciones: Albayzin 2018 y 2020

- *Iberspeech-RTVE-Challenge* :



Minimum WER by show



Ejemplos de audios de evaluación:

ESPAÑA EN COMUNIDAD



del río hay muchísima la margarita africana de las entre las dunas y bueno así muy bonita y la uña de gato la red natura dos mil a la que pertenece a la playa de fresh urfé es una red europea de espacios protegidos diseñada para asegurar la supervivencia a largo plazo de las especies y los hábitats naturales en el territorio europeo además esta playa toda la zona costera de navia y el río están incluidas en la zona especial de conservación y de protección para las aves











LA TARDE EN 24H TERTULIA



al otro lado está esperanza aguirre fuera de la política también todos sabemos porque ignacio gonzález no está no estaba en la foto en fin para cuando la responsabilidad políticas hasta donde tiene que llegar el nivel del lodazal de lo que estamos viendo para que los responsables políticos últimos den la cara y asuman la responsabilidad que les toque es verdad que estamos todavía en una fase en este caso concreto de la gürtel valenciana sólo de juicio

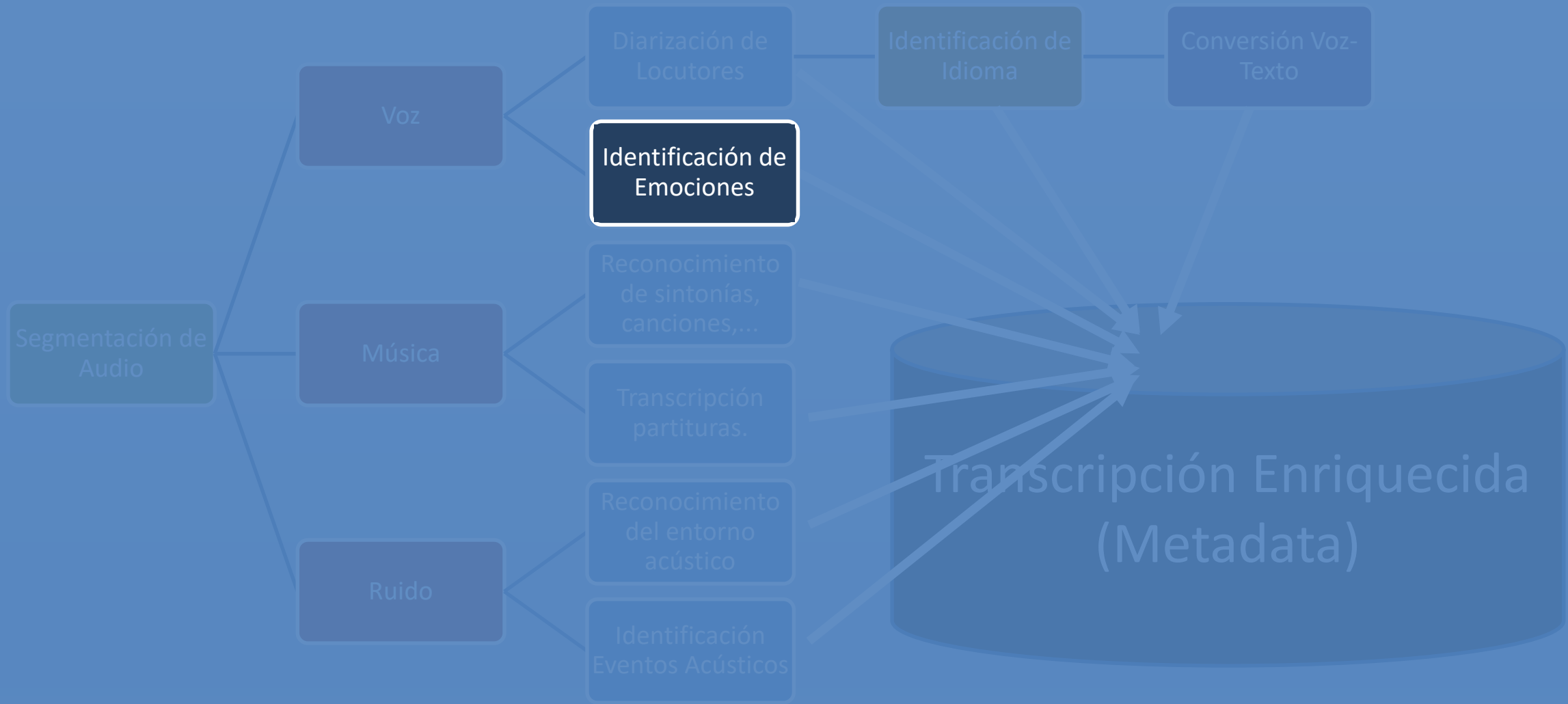


Diarización e Identificación de hablantes, Añadiendo Reconocimiento del Habla:

-   **J. Biden:** Contenido de la intervención 1 ... <Tcomienzo1> <Tfin1>
-   **O. Scholz :** Contenido de la intervención 2 ... <Tcomienzo2> <Tfin2>
-   **B. Johnson:** Contenido de la intervención 3 ... <Tcomienzo3> <Tfin3>
-   **J. Biden:** Contenido de la intervención 4 <Tcomienzo4> <Tfin4>
-   **A. Scholz :** Contenido de la intervención 5 ... <Tcomienzo5> <Tfin5>



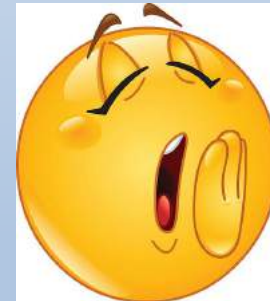
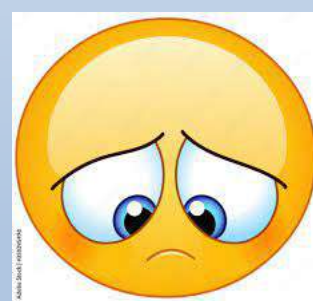
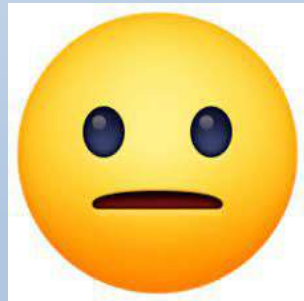
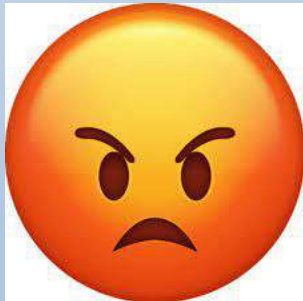
Tecnologías



Identificación de Emociones

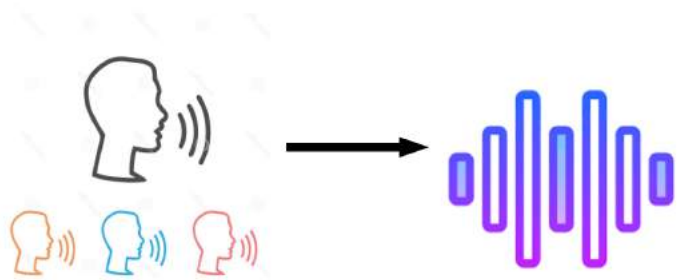
- ¿Para qué sirve?:

- Puede añadir información extra que enriquece el discurso de los protagonistas de un contenido

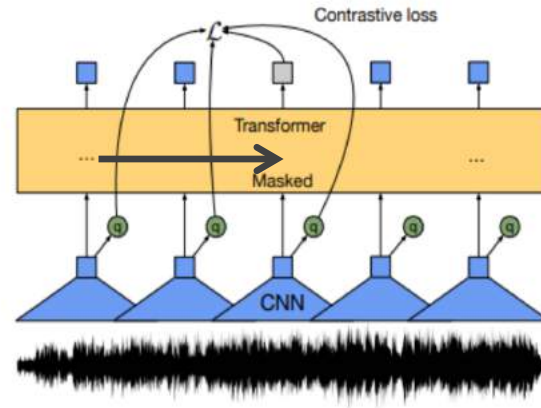


Identificación de Emociones

SEÑAL DE VOZ



Feature Extraction Using Wav2Vec2



Feature Vectors

Classify Layer



Emotion Recognition

enfado

alegría

tristeza

neutro

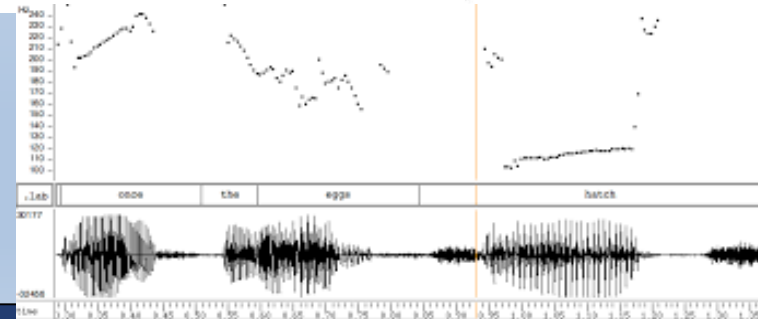
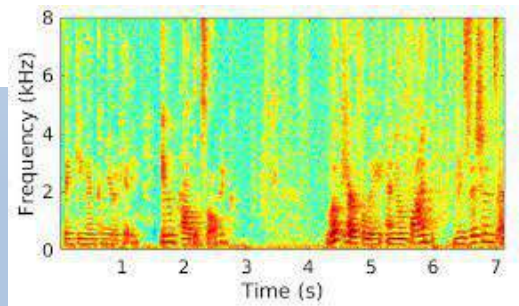
aburrimiento

ansiedad

Características:

**Espectrales
Prosódicas
Paralingüísticas**

...





Universidad
Zaragoza