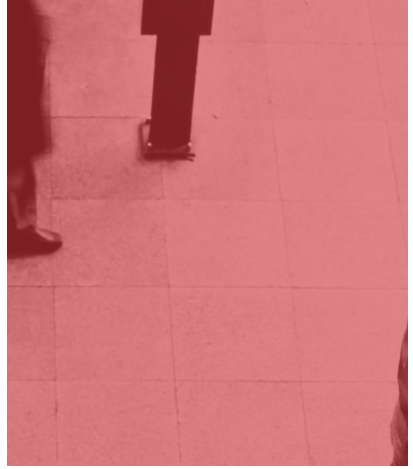
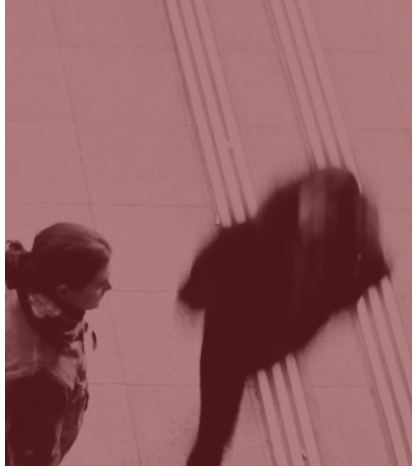
PROCESAMIENTO DEL LENGUAJE NATURAL Y OTRAS TECNICAS SOBRE DATOS POCO ESTRUCTURADOS

JAIME MARTEL ROMERO-VALDESPINO
CEO DE ITELLIGENT

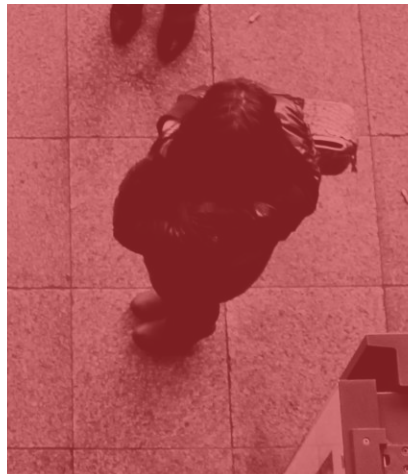




insights analytics i+a España

LA COMUNIDAD DE MARKET RESEARCH Y DATA SCIENCE








jmartel@itelligent.es

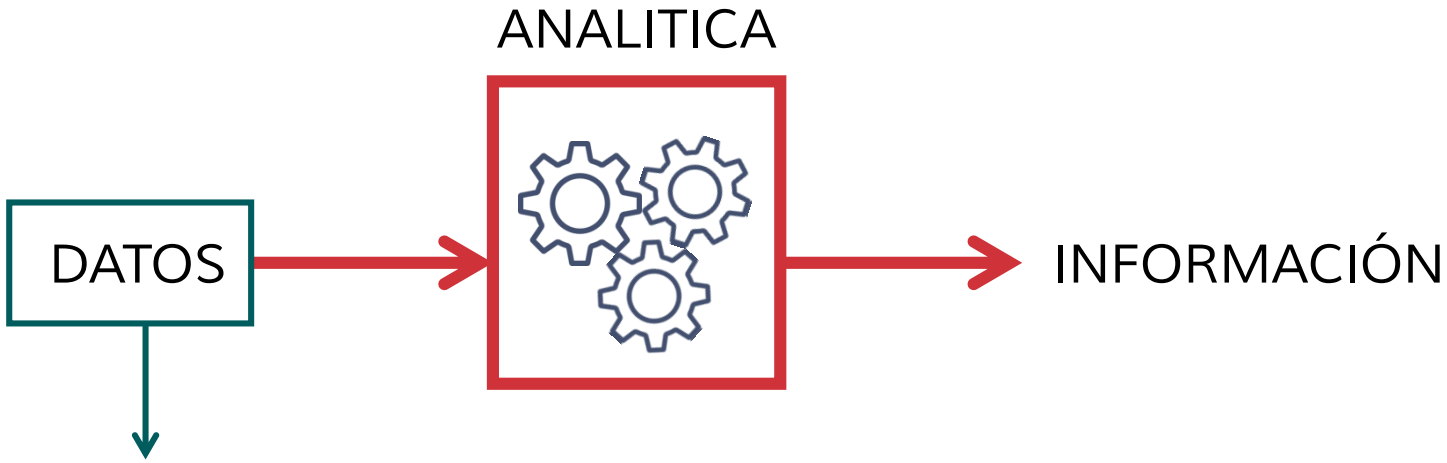
https://www.linkedin.com/in/jaime-martel/

Índice

DATOS POCO ESTRUCTURADOS Y SU PUESTA EN VALOR

- CASO 1: SISTEMAS DE RECOMENDACIÓN
- CASO 2: SISTEMAS DE PREGUNTAS Y RESPUESTAS
- CASO 3: IMÁGENES: DETECCIÓN DE OBJETOS Y CLASIFICACIÓN
- CASO 4: PLATAFORMA EXPLICACIÓN DE CONSUMOS ELÉCTRICOS

Datos estructurados vs poco estructurados



¿Cuándo decimos que un dato es estructurado?

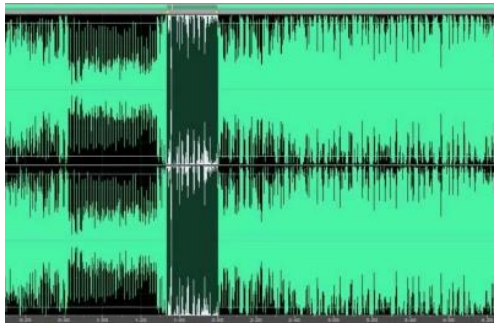
Decimos que un dato es estructurado si es fácil de automatizar, esto es si es fácil de generar información mediante medios digitales.

Datos estructurados vs poco estructurados

DATOS MUY ESTRUCTURADOS

DATOS POCO ESTRUCTURADOS

	A	B	C	D	E
1	MES	SEMANA	CLIENTE	ARTICULO	CANTIDAD
2	Enero	1	13	105	6
3	Enero	2	15	103	15
4	Enero	3	13	104	2
5	Enero	4	15	110	15
6	Febrero	1	12	108	8
7	Febrero	2	12	105	25
8	Febrero	3	13	110	1
9	Febrero	4	14	106	12
10	Marzo	1	15	105	16
11	Marzo	2	12	102	8
12	Marzo	3	13	103	5
13	Abril	1	15	105	11
14	Abril	2	13	110	10
15	Abril	3	12	103	5



MÁS FÁCIL SACAR INFORMACIÓN DE FORMA AUTOMÁTICA

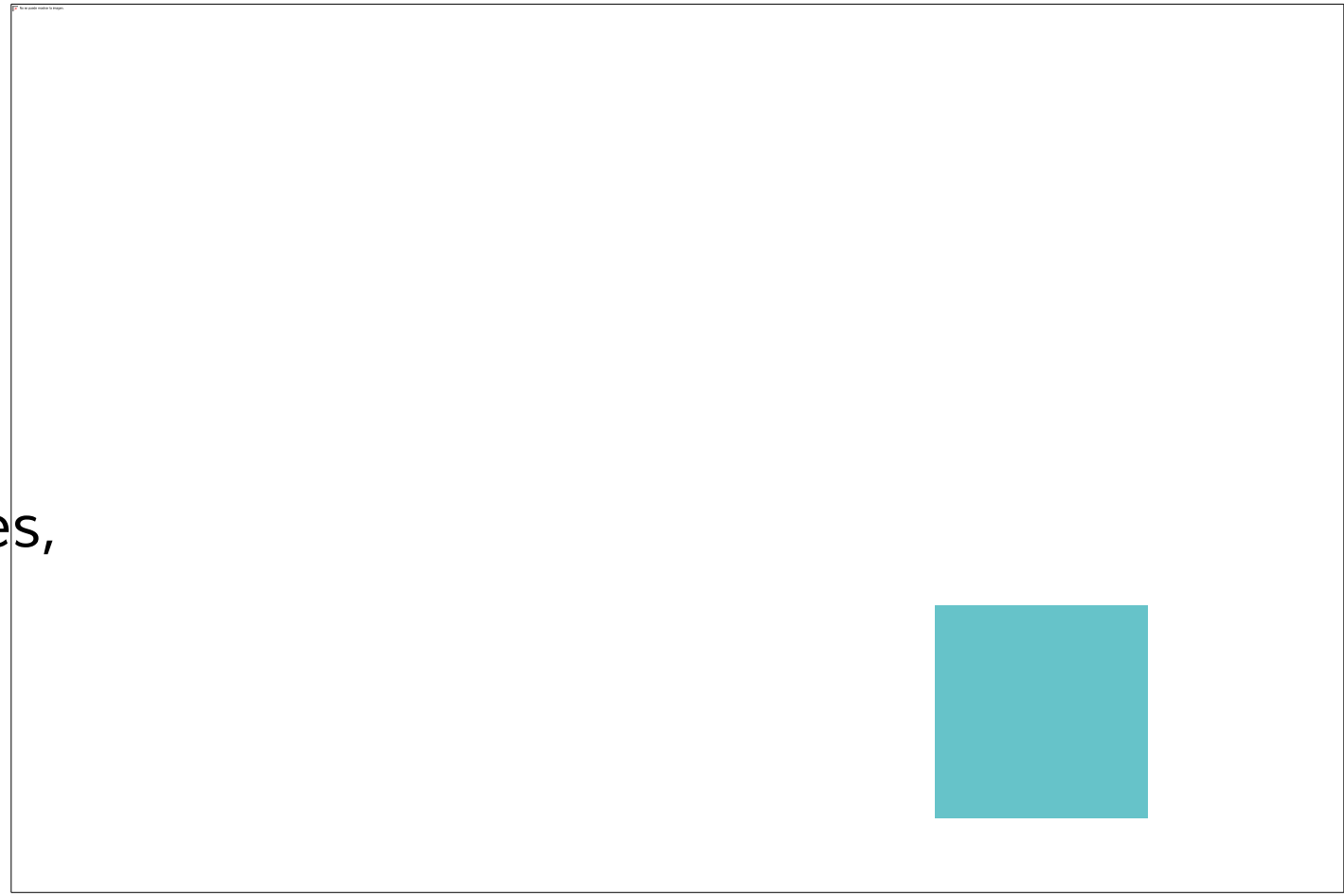
MÁS DIFICIL SACAR INFORMACIÓN DE FORMA AUTOMÁTICA

Datos poco estructurados

¿Cómo podemos utilizar los datos poco estructurado?

Inteligencia artificial:

- Procesamiento lenguaje natural: Textos, audios, ...
- Procesamiento de imágenes: Imágenes, videos,..



¿Qué es el Procesamiento del Lenguaje Natural?

El objetivo último del Procesamiento del Lenguaje Natural o PLN es conseguir que las maquinas **entiendan en profundidad** el lenguaje (significado del lenguaje), no simplemente que sean capaces de "procesar" el lenguaje (ej. contar palabras, buscar palabras,...)

¿Qué es el Procesamiento de imágenes?

El objetivo último del Procesamiento de imágenes es conseguir que las maquinas **entiendan** las imágenes (ej. qué contienen o qué significa), y no simplemente que sean capaces de “procesar” las imágenes (ej. colores, pixeles, formas, ...)

Una historia cargada de fracasos

Procesamiento del Lenguaje Natural

‘the spirit is willing, but the flesh is weak’
(Mateos, 26:41)

El espíritu es voluntarioso, pero la carne es débil

‘the vodka is agreeable, but the meat is spoiled’

El vodka es agradable pero la carne esta podrida.

Sistema de traducción Ruso/Inglés en IBM (Watson presidente de IBM el tercero desde la izquierda)

Una historia cargada de fracasos

Procesamiento de imágenes

El ejercito americano entrenó una red neuronal para distinguir entre tanques rusos y tanques americanos. Para ello, se entrenaban con fotos de tanques rusos y tanques americanos.

Aunque el resultado fue muy bueno en laboratorio, cuando se testeó en un test real el resultado fue muy pobre.

ruso

americano

Comienzan los éxitos

Procesamiento del Lenguaje Natural

Durante la década del 2000s IBM desarrolla Watson, que en el 2011 derrota al juego del Jeopardy a dos expertos en dicho juego. Watson es un sistema diseñado para una tarea de PLN denominada Q&A (preguntas y respuestas), adaptado a las particularidades de Jeopardy y con capacidad de interactuar por voz con el presentador.

Una historia llena de esperanza…Deep Learning

Ejemplo Procesamiento del Lenguaje Natural: Word2Vec, 2013

turismo → (1.234, 3.034, 5.201, …) Vector de coordenadas
EMBEDDINGS

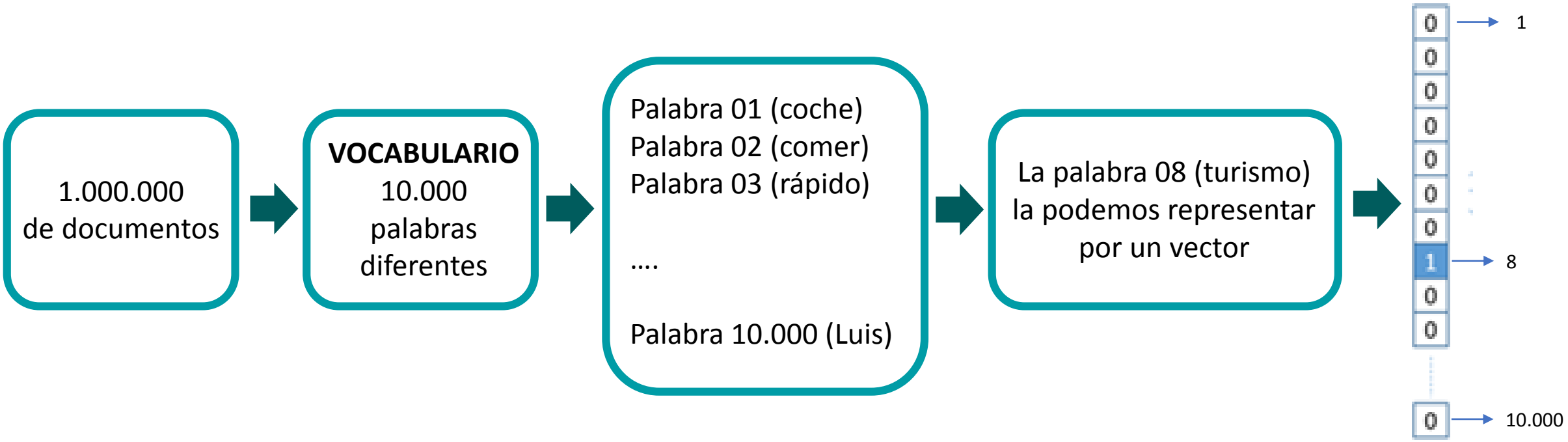
… salió el hombre alto riéndose …

… salió la mujer alta riéndose …

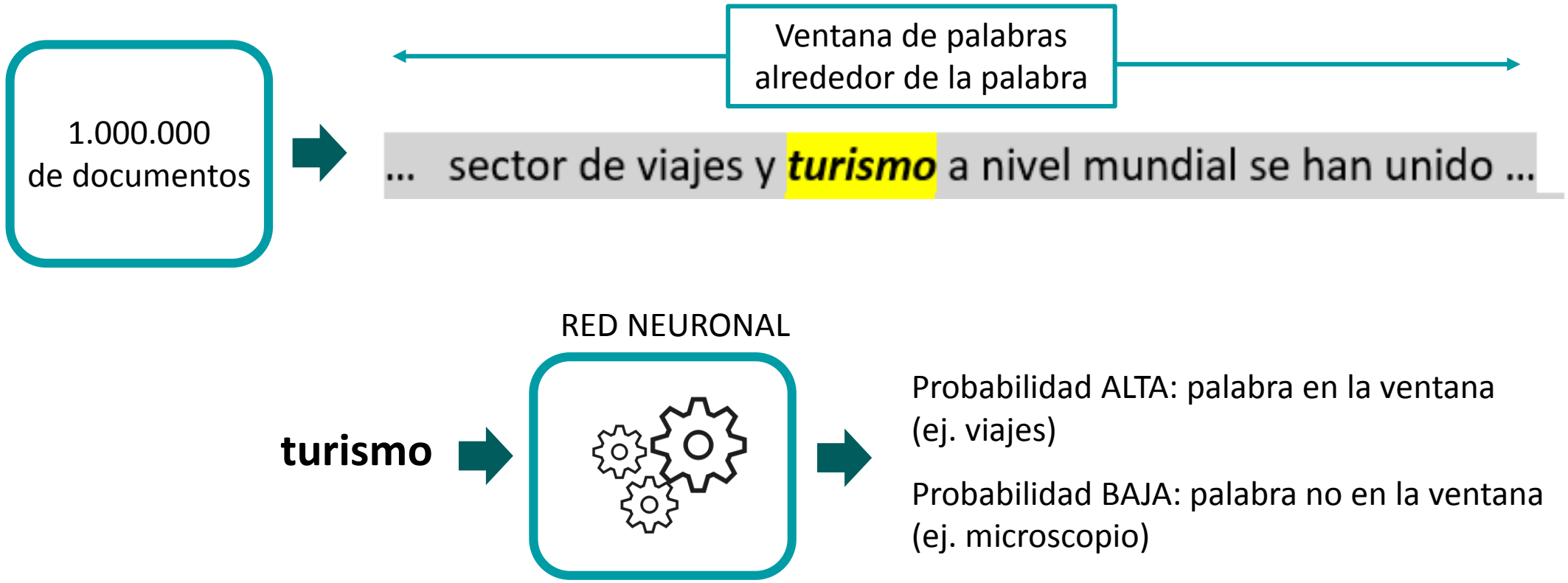
"You shall know a word by the company it keeps"

J.R. Firth, 1957

Explicación intuitiva al Deep Learning
(Word2Vec)



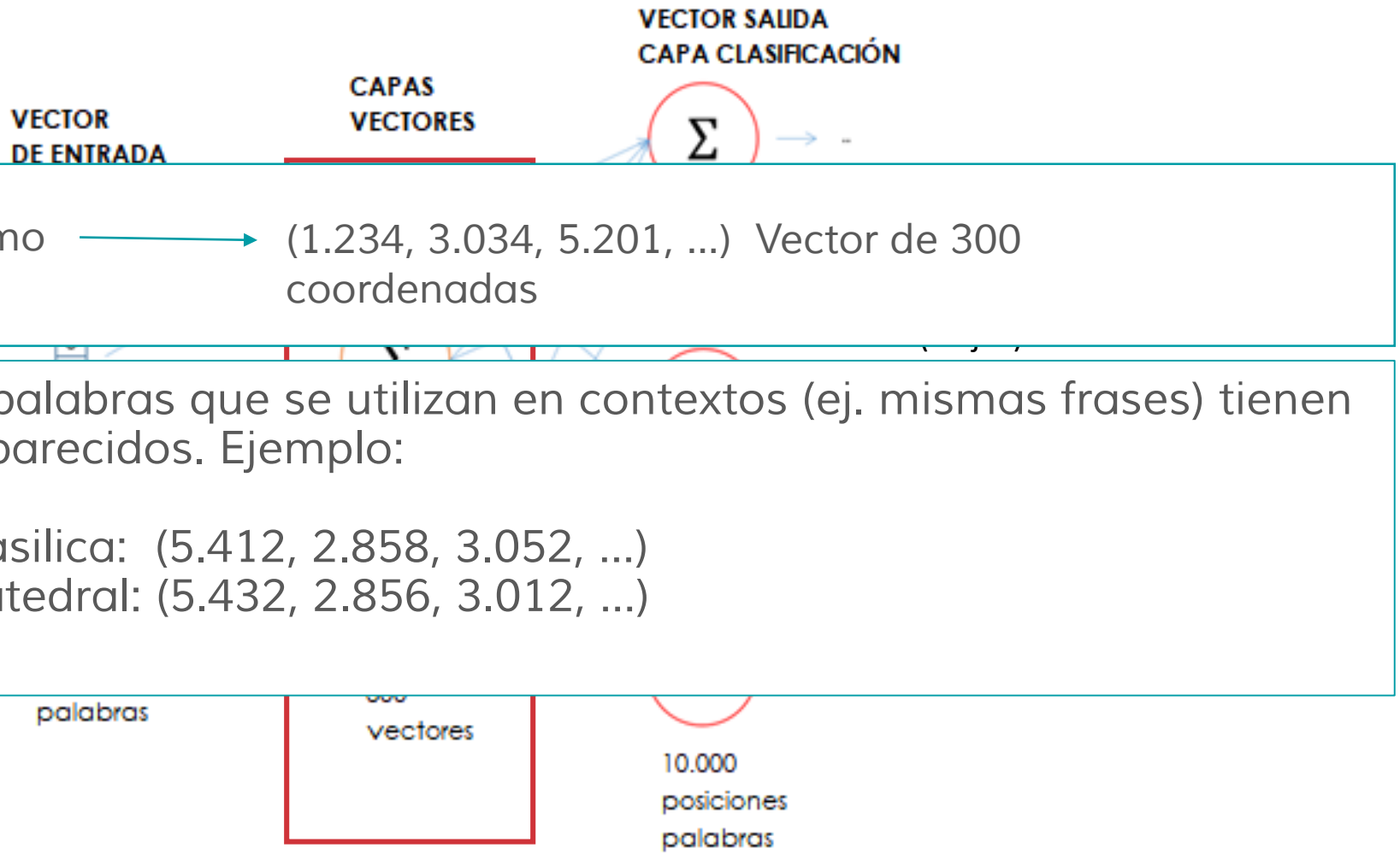
Explicación intuitiva al Deep Learning
(Word2Vec)



Explicación intuitiva al Deep Learning
(Word2Vec)

... sector de viajes y ***turismo*** a nivel mundial se han unido ...

Red Neuronal



turismo → (1.234, 3.034, 5.201, ...) Vector de 300 coordenadas

Aquellas palabras que se utilizan en contextos (ej. mismas frases) tienen vectores parecidos. Ejemplo:

basilica: (5.412, 2.858, 3.052, ...)
catedral: (5.432, 2.856, 3.012, ...)

Explicación intuitiva al Deep Learning
(Word2Vec)

... la magnífica basílica renacentista sobresale ...

... la magnifica catedral renacentista sobresale ...

Nearest words to frog:

1. frogs
2. toad
3. litoria
4. leptodactylidae
5. rana
6. lizard
7. eleutherodactylus

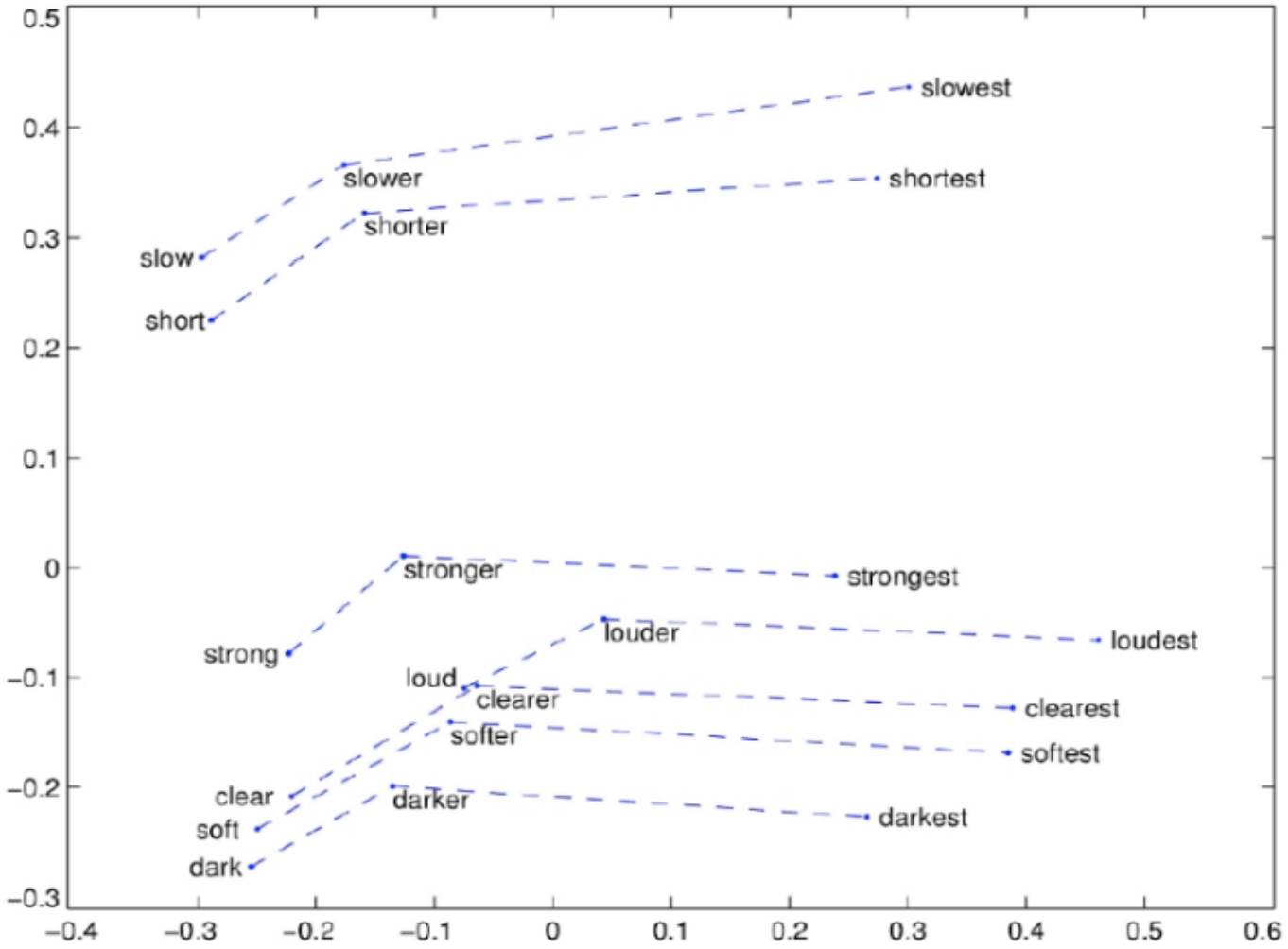
litoria

leptodactylidae

rana

eleutherodactylus

Explicación intuitiva al Deep Learning
(Word2Vec)



... salió el hombre alto riéndose ...

... salió la mujer alta riéndose ...

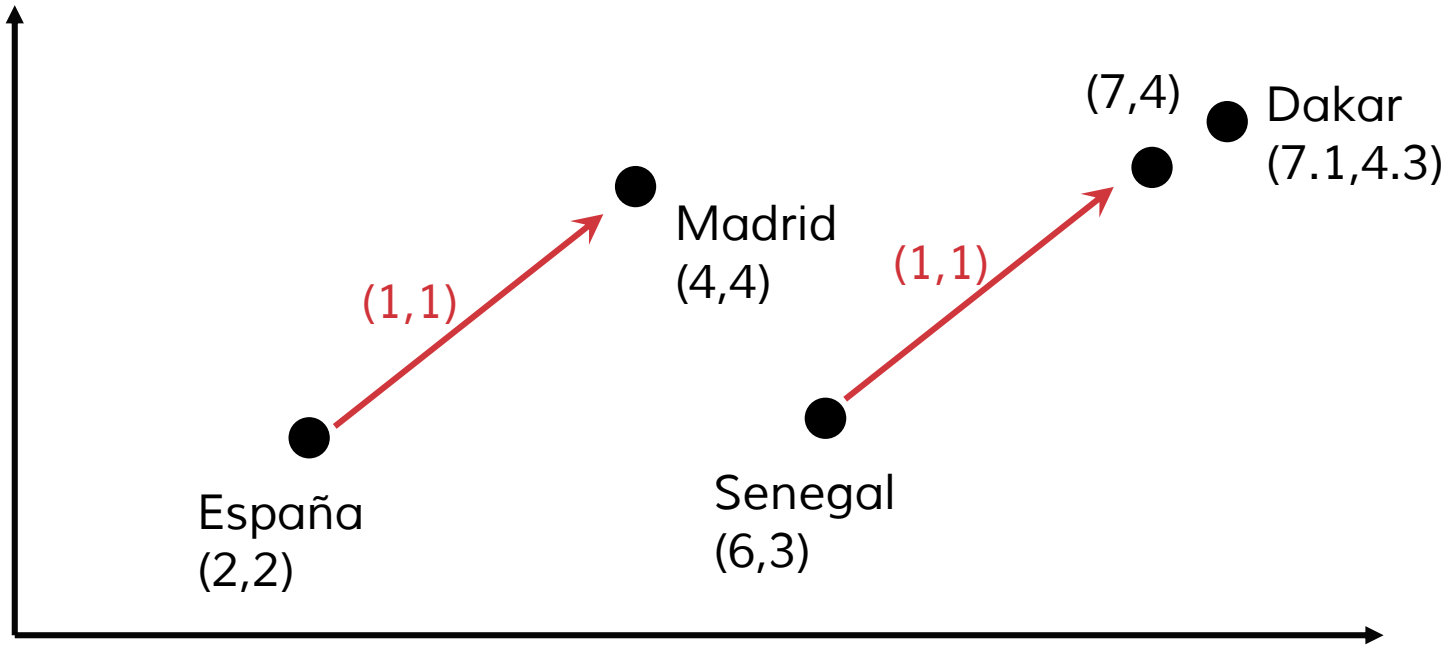
Figura Referencia: GLOVE STANDFORD

Explicación intuitiva al Deep Learning
(Word2Vec)

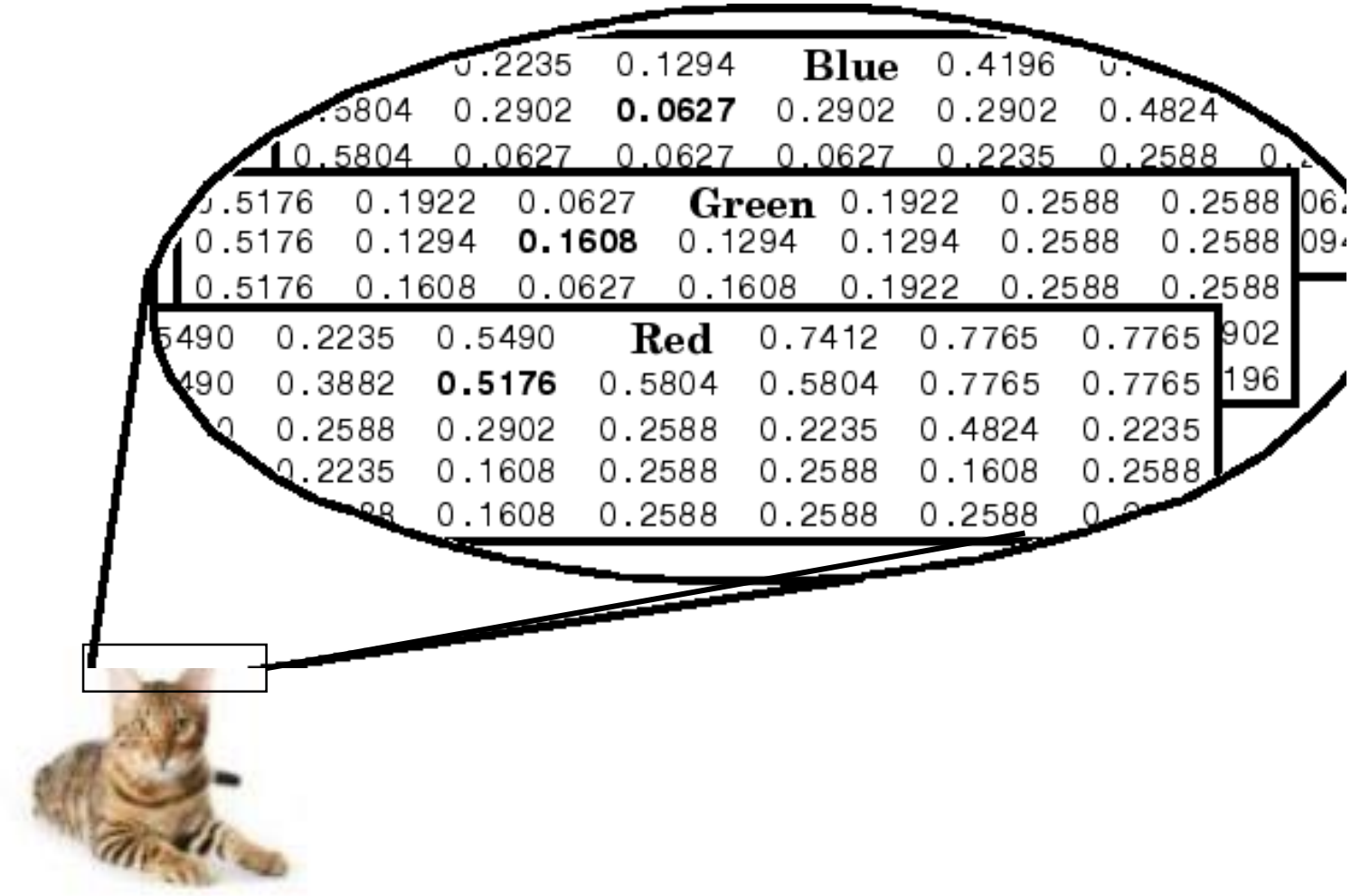
Madrid es a España como ___Dakar___ es a Senegal

Madrid (4,4) – España (2,2)= (1,1)

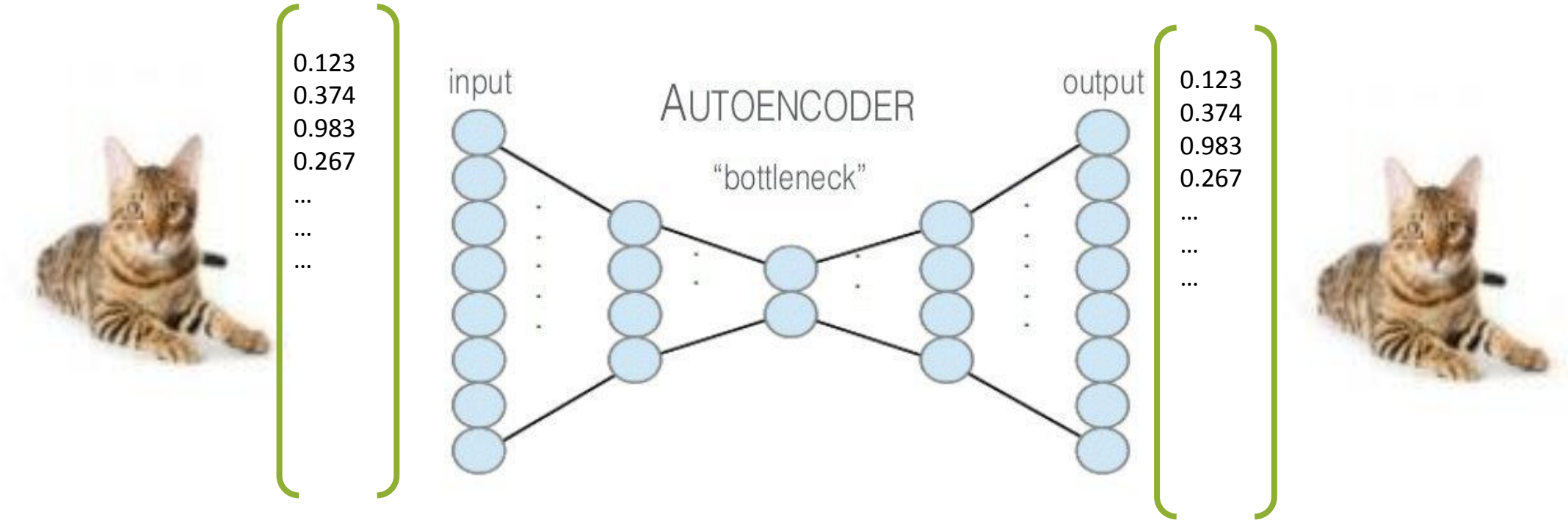
Senegal (6,3) + (1,1) = (7,4) → Dakar (7.1,4.3)



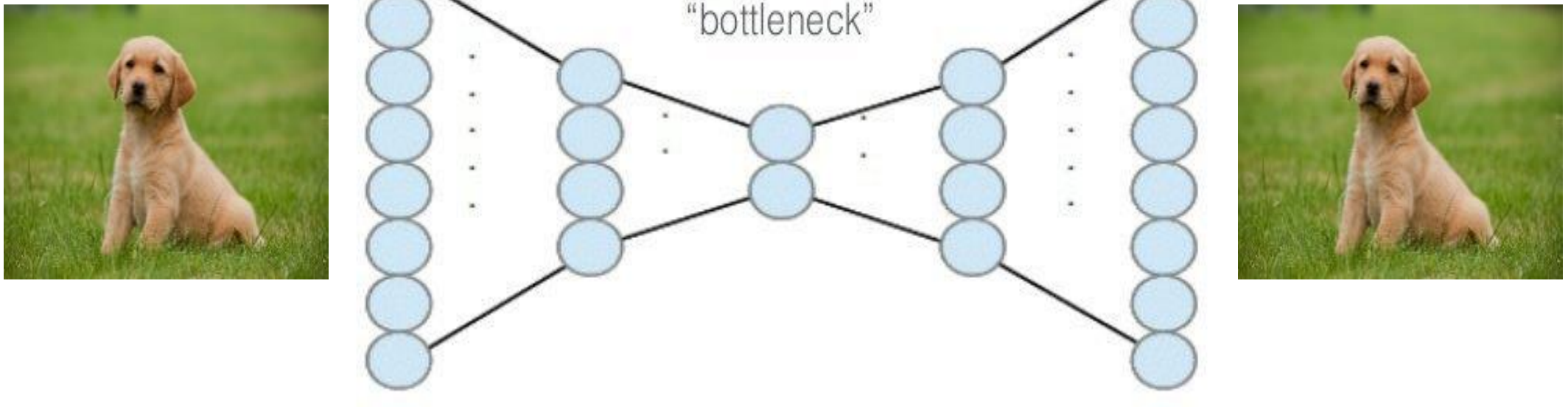
Deep Learning aplicado a las imágenes
(Autoencoders)



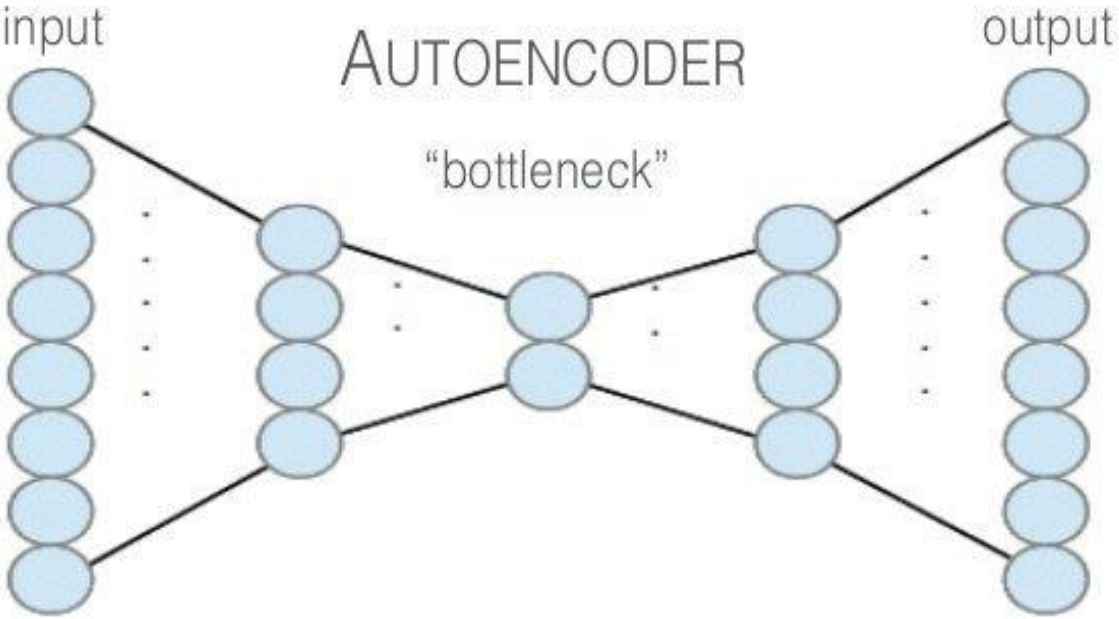
Deep Learning aplicado a las imágenes
(Autoencoders)



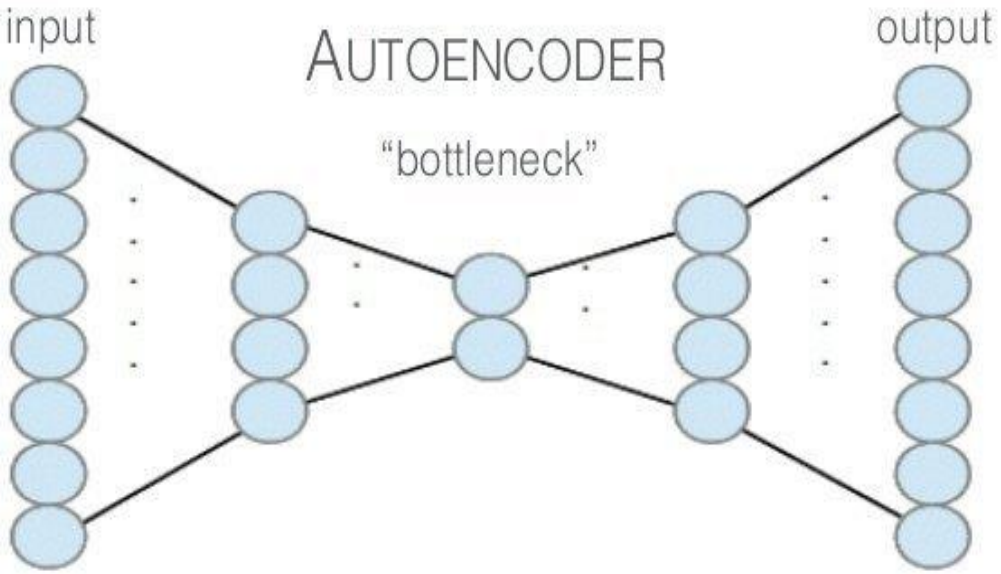
Deep Learning aplicado a las imágenes
(Autoencoders)



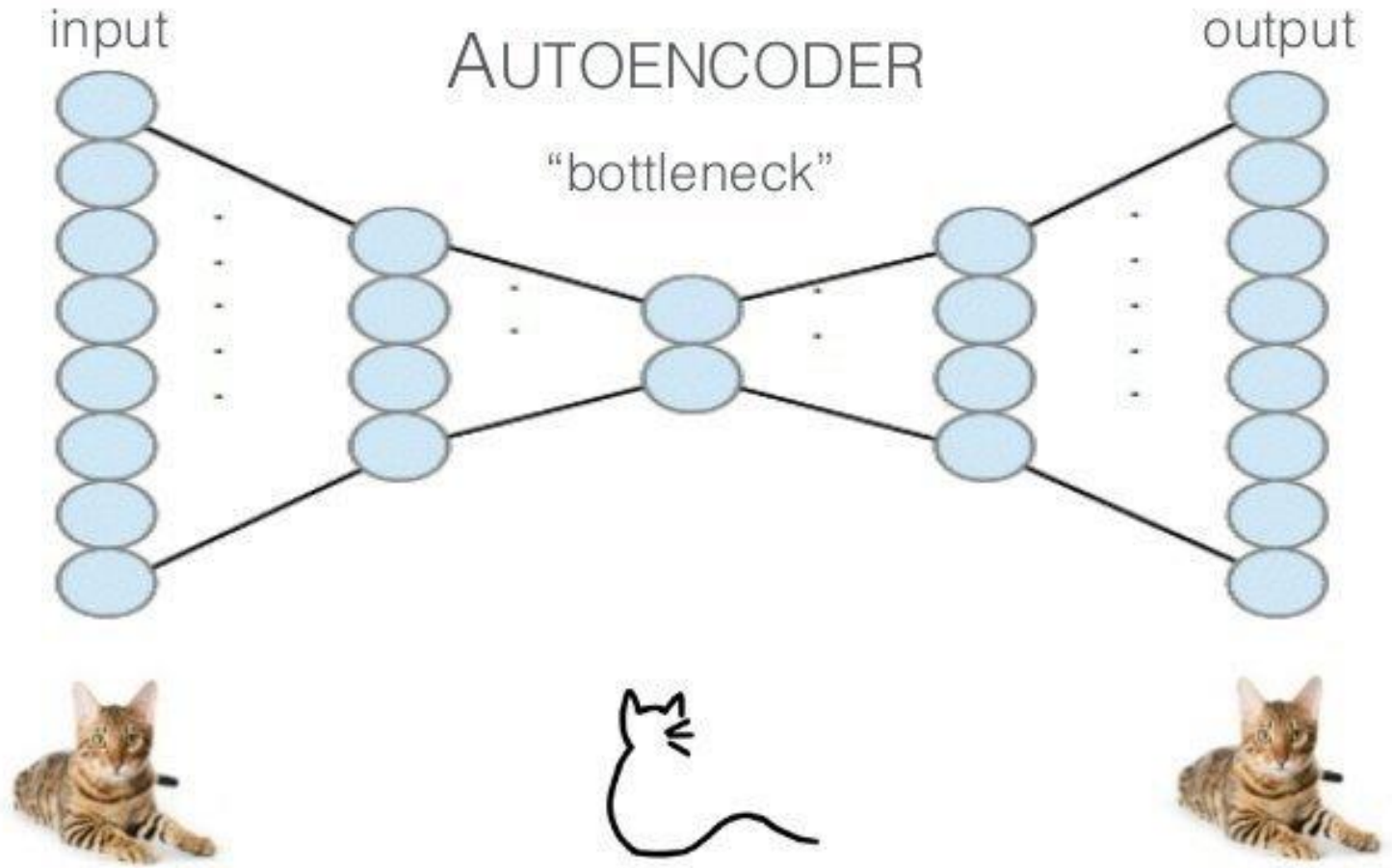
Deep Learning aplicado a las imágenes
(Autoencoders)



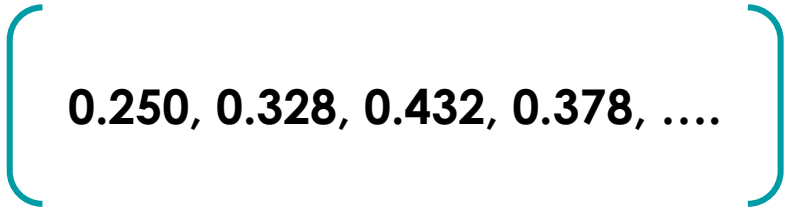
Deep Learning aplicado a las imágenes
(Autoencoders)



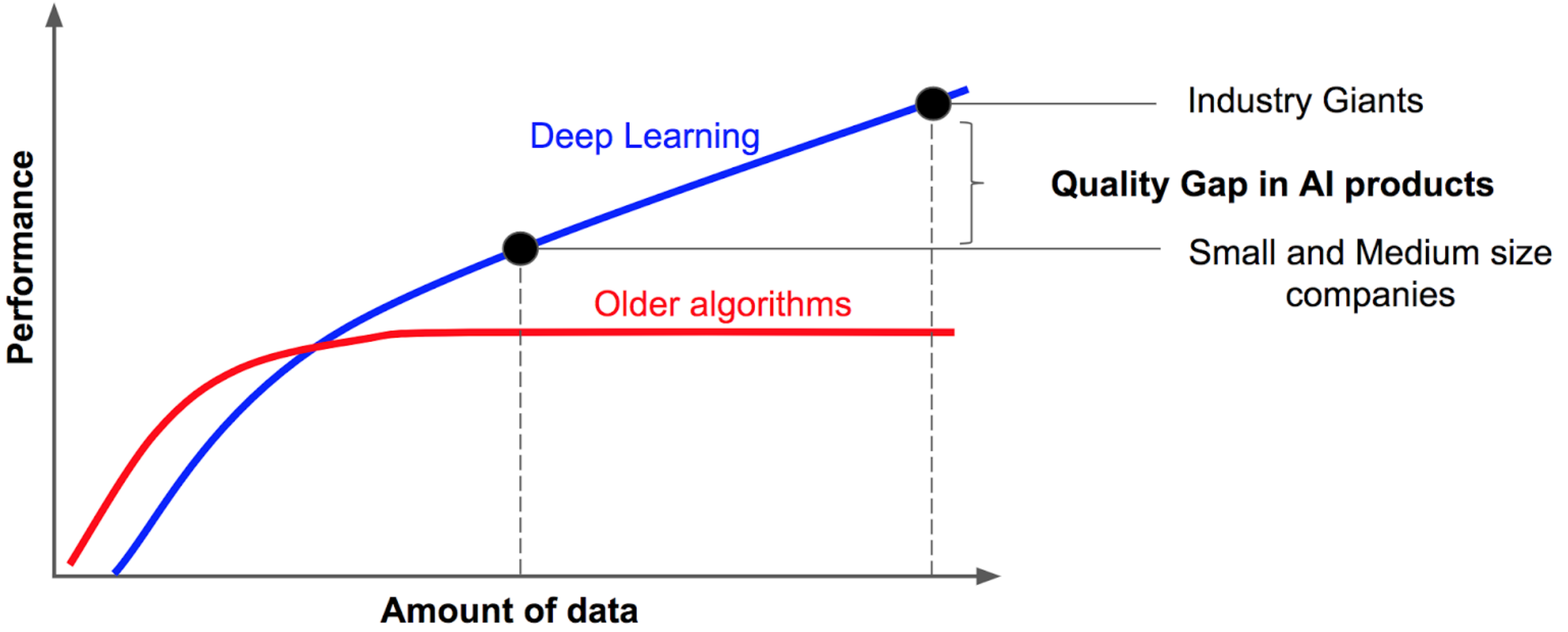
Deep Learning aplicado a las imágenes
(Autoencoders)



Deep Learning aplicado a las imágenes
(Autoencoders)



¿Por qué Deep Learning?

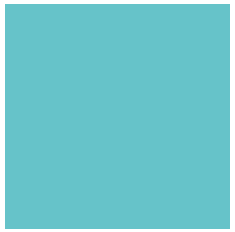


CASO 1
SISTEMAS DE COMENDACIÓN

SISTEMAS DE RECOMENDACIÓN

TIPOS:

1. RECOMENDADOR POR INTERACCIONES
2. RECOMENDADOR POR CONTENIDOS



Sistemas de recomendación

1. RECOMENDADOR POR INTERACCIONES

"Dime con quién andas, y te diré quién eres"

Sistemas de recomendación

1. RECOMENDADOR POR INTERACCIONES

Pero… ¿qué pueden ser estos Id? Películas, comidas, bebidas, libros, coches, …

LUIS

Id 01 | Id 20 | Id 65 | Id 80

IDEA FUNDAMENTAL

Las recomendaciones por interacciones no necesitan información sobre los productos que recomiendan SOLO sobre las INTERACCIONES.

ROSA

Id 20 | Id 35 | Id 65 | Id 80 | a Luis: Id 90 | Id 98

2. RECOMENDADOR POR CONTENIDOS

"Si te gustó el libro
¿Sueñan los androides con ovejas eléctricas? probablemente te guste la película
Blade Runner "

Sistemas de recomendación

2. RECOMENDADOR POR CONTENIDOS

***INTUICIÓN**: Si a Luis le gusta Interestellar, puede que le guste una película del mismo director, género, etc.*

CONTENIDOS de la película

RECOMENDACIONES de películas

→ Director → Origen

IDEA FUNDAMENTAL

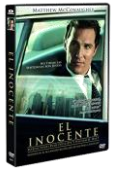
Utilizamos los contenidos y características de los productos para encontrar similitudes con otros productos

LUIS

→ Espacial →

Marte

→ Actores →



El inocente

Sistemas de recomendación

PROBLEMA HABITUAL DE LOS RECOMENDADORES

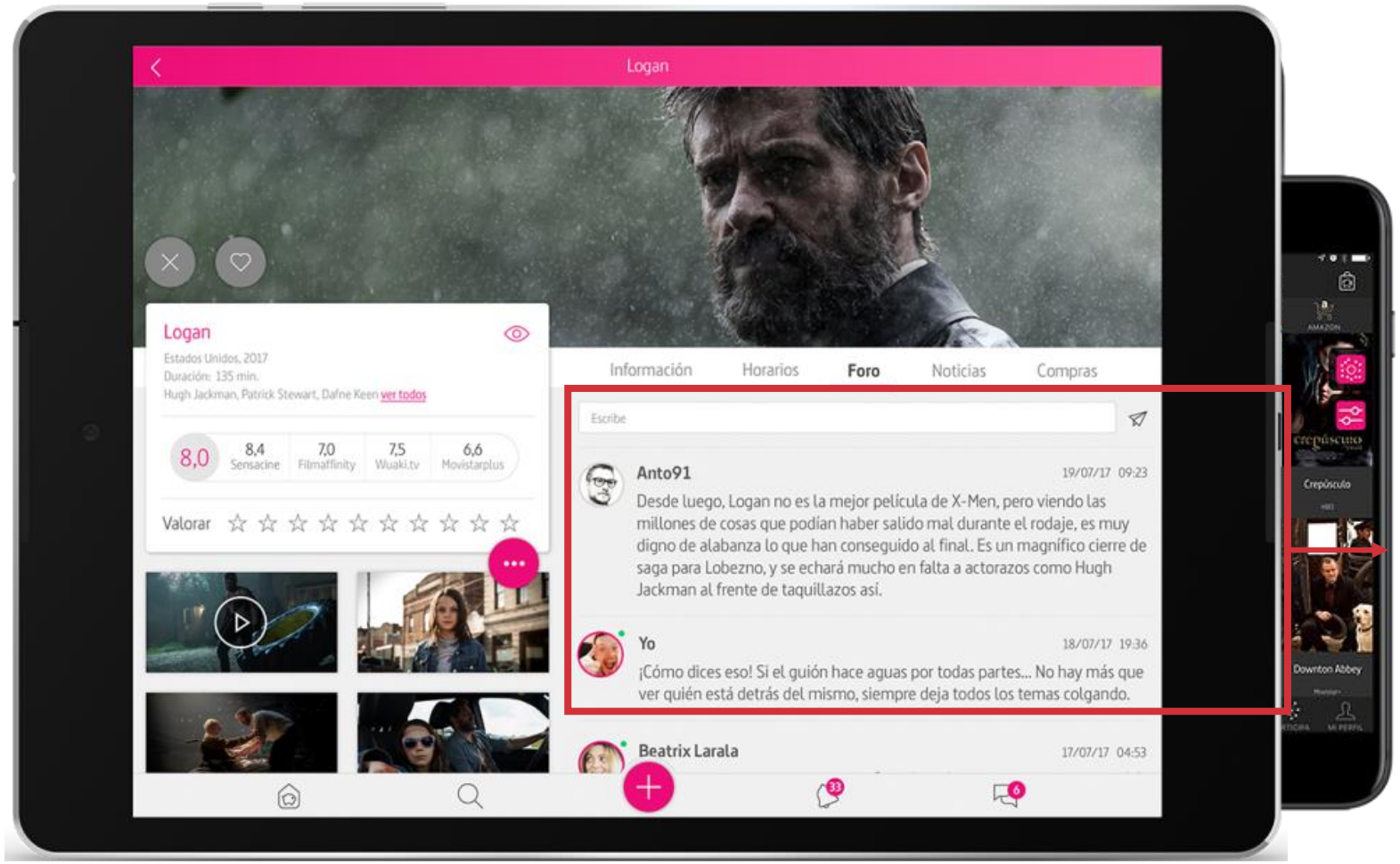
Un problema habitual de los sistemas de recomendación es: EL ARRANQUE EN FRIO

Una forma habitual de resolver este problema de “arranque en frio” es utilizar inicialmente un recomendador por contenidos y posteriormente, utilizar un recomendador por interacciones.

Sistemas de recomendación

EJEMPLO DE APLICACIÓN

RECOMENDADOR 01. Contenidos audiovisuales



Doc2Vec
[3.5, 2.4, 7.8, ...]

Sistemas de recomendación

EJEMPLO DE APLICACIÓN

RECOMENDADOR 02. Artículo de moda

La información que poseemos es:

Zapatos

- Zapatillas
- Zapatos de salón
- Zapatos altos
- Bailarinas
- Sandalias
- Botines
- Zapatos con cordones
- Zapatos bajos
- Zuecos
- Botas
- Zapatillas deportivas
- Zapatillas outdoor
- Chanclas
- Pantuflas

Talla | Marca | Color | Precio | Material exterior | Estampado | Forma tacón | Puntera | Cierre | Ancho del zapato

Material exterior: Piel de imitación/tela
Material interior: Tela
Plantilla: Tela
Suela: Fibra sintética
Grosor del relleno: Relleno contra el frío
Material/composición: Cuero sintético

CARACTERÍSTICAS DEL PRODUCTO

Puntera: Redonda
Forma del tacón: Plano
Cierre: Con cordones
Número de artículo: P0G11A04I-A11

Sistemas de recomendación

EJEMPLO DE APLICACIÓN

RECOMENDADOR 02. Artículo de moda

Ana le gusta este tipo de zapato:

Sistemas de recomendación

EJEMPLO DE APLICACIÓN

RECOMENDADOR 02. Artículo de moda

Otro zapato con la misma descripción...

Material exterior: Piel de imitación/tela
Material interior: Tela
Plantilla: Tela
Suela: Fibra sintética
Grosor del relleno: Relleno contra el frío
Material/composición: Cuero sintético

CARACTERÍSTICAS DEL PRODUCTO

Puntera: Redonda
Forma del tacón: Plano
Cierre: Con cordones
Número de artículo: P0G11A04I-A11

Material exterior: Piel de imitación/tela
Material interior: Tela
Plantilla: Tela
Suela: Fibra sintética
Grosor del relleno: Relleno contra el frío
Material/composición: Cuero sintético

CARACTERÍSTICAS DEL PRODUCTO

Puntera: Redonda
Forma del tacón: Plano
Cierre: Con cordones
Número de artículo: NI111A0GI-A16

Sistemas de recomendación

EJEMPLO DE APLICACIÓN

RECOMENDADOR 02. Artículo de moda

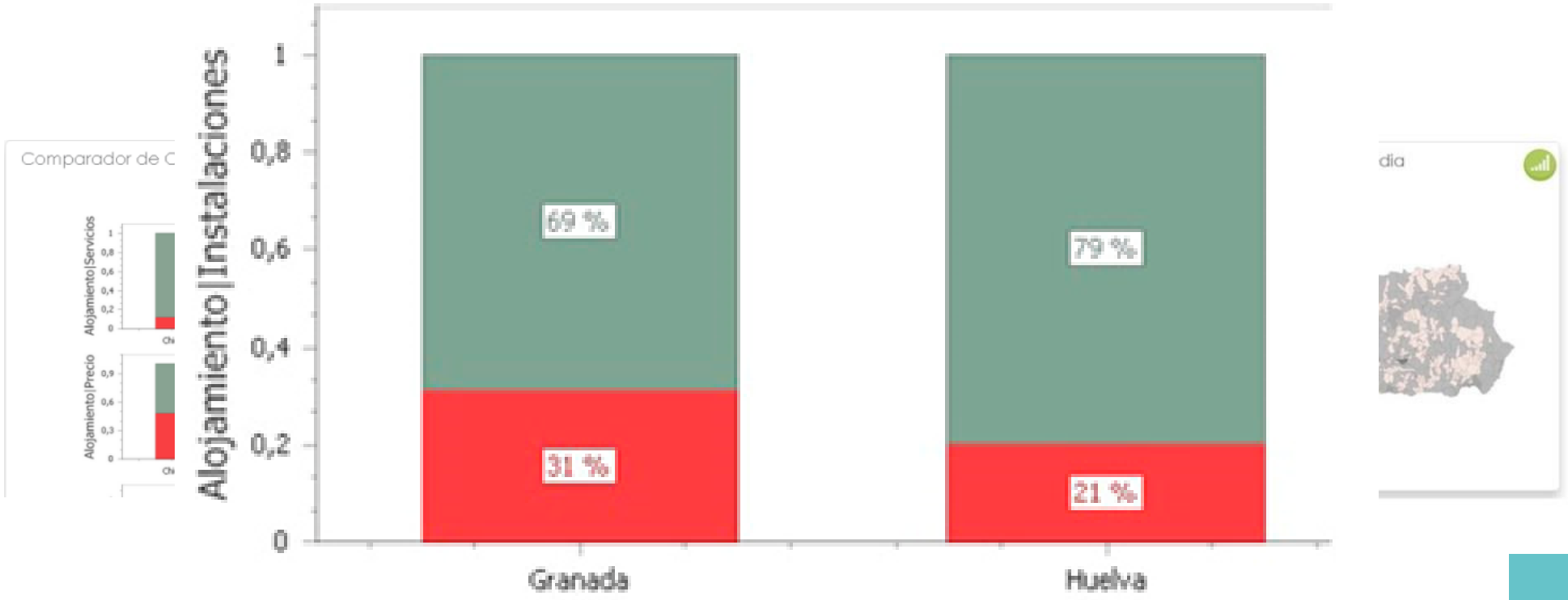
SOLUCIÓN: Deep Learning

CASO 2
SISTEMAS DE PREGUNTAS Y RESPUESTAS

Sistemas de preguntas y respuesta

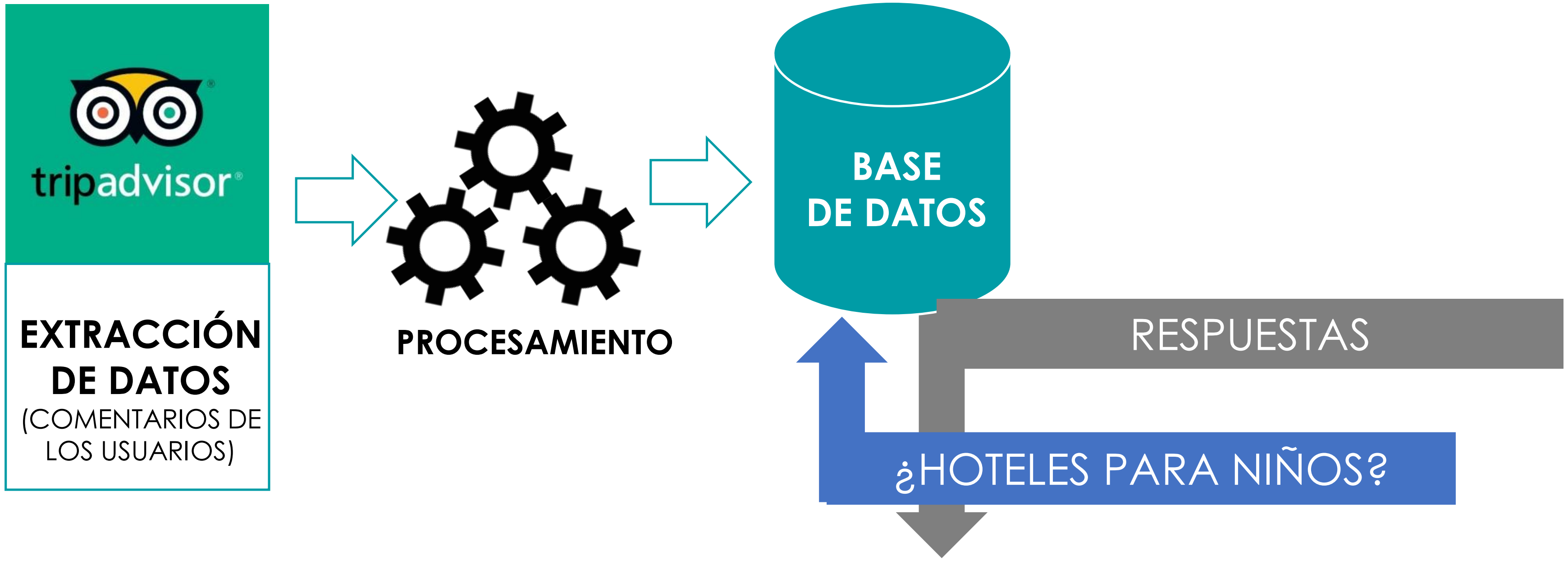
Sistemas de preguntas y respuesta

Sistemas de preguntas y respuesta



Sistemas de preguntas y respuesta

WORD EMBEDDING (texto)

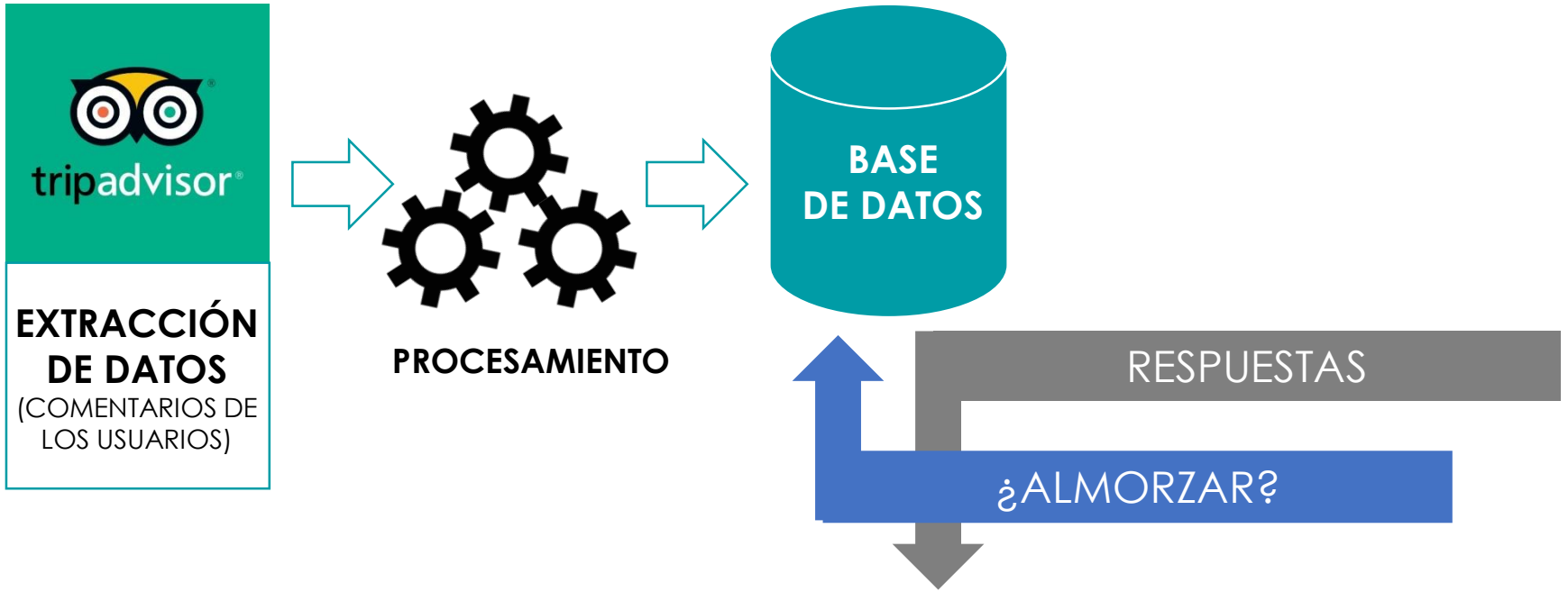


COMENTARIO 3: El apartamento genial. Los muebles bastantes nuevos. Piscina con

como niños que son...lloran, corretean..etc..etc.y el jacuzzi

Sistemas de preguntas y respuesta

WORD EMBEDDING (texto)

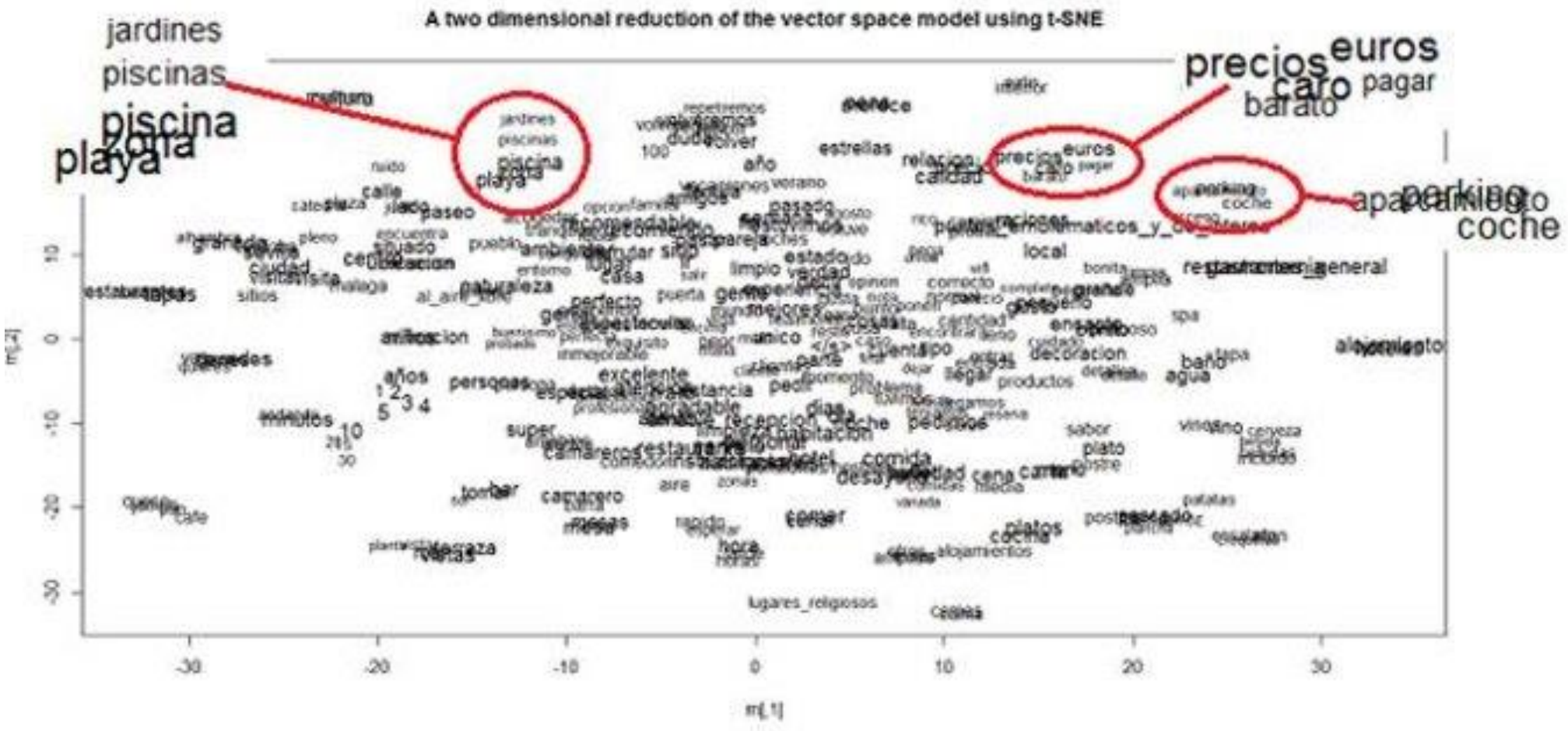


COMENTARIO 1: Una plaza llena de tabernas y terrazas, donde poder comer, cenar, beber algo. Muchísimo ambiente a todas las horas del día.

COMENTARIO 2: Un paseo con un ambiente impresionante. Lugar para poder comer, beber y disfrutar. Los precios de lo más ajustados. Se respira un ambiente diferente.

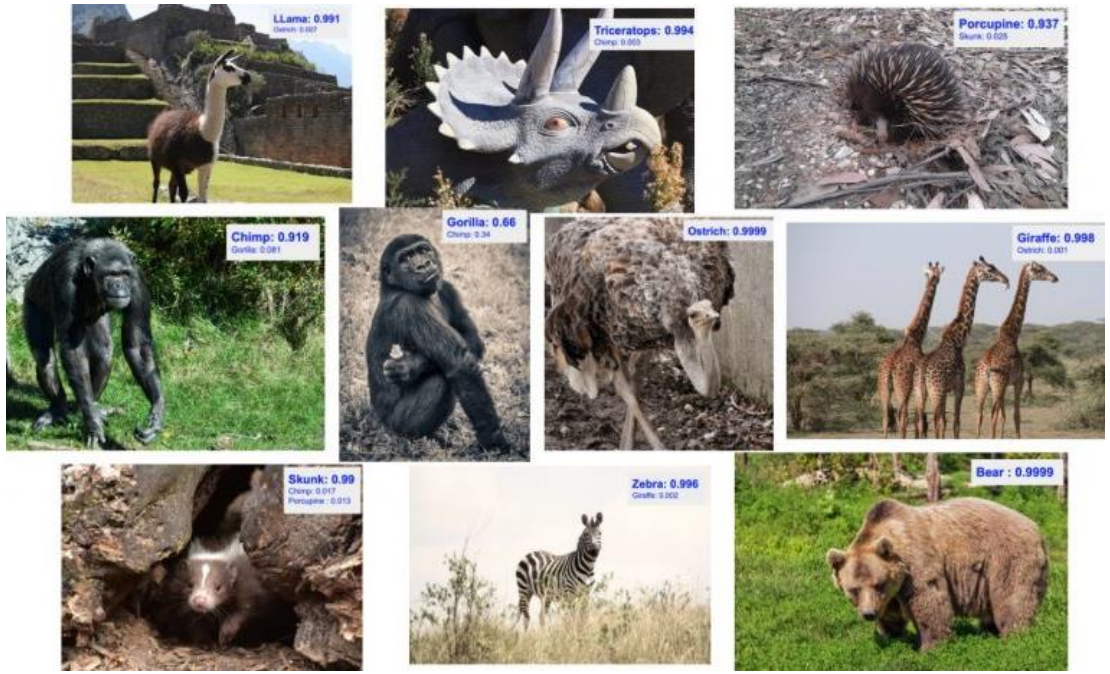
Sistemas de preguntas y respuesta

WORD EMBEDDING (texto)

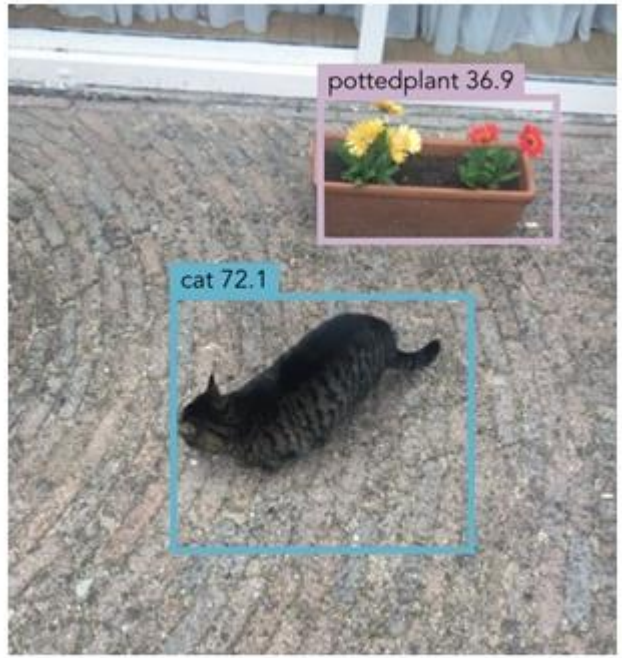


CA SO 3
IMÁGENES: DETECCIÓN
ETOS Y ASIFICACIÓN

Clasificación y detección de objetos



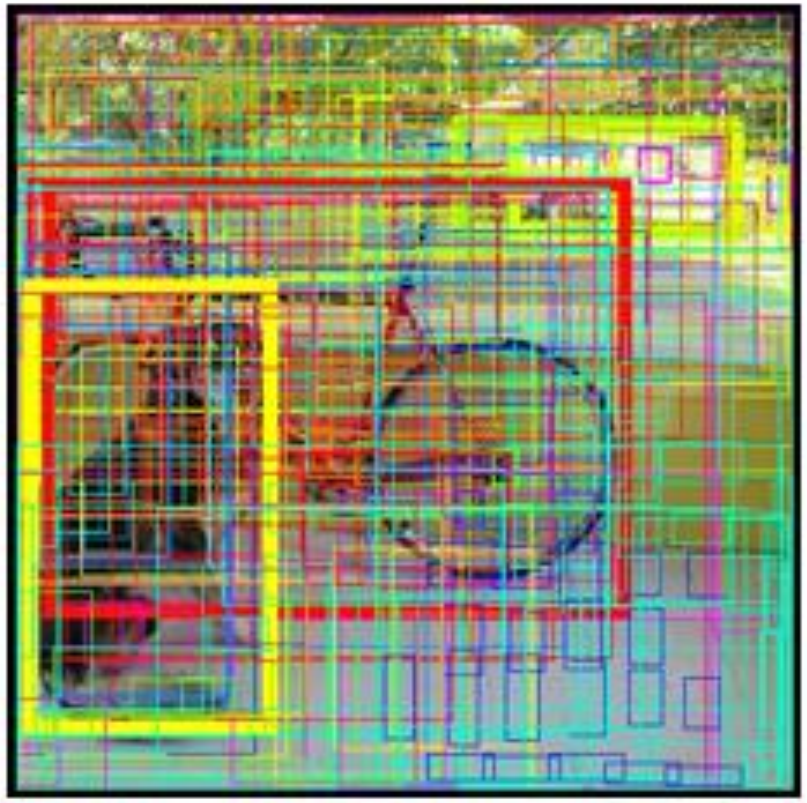
Classification

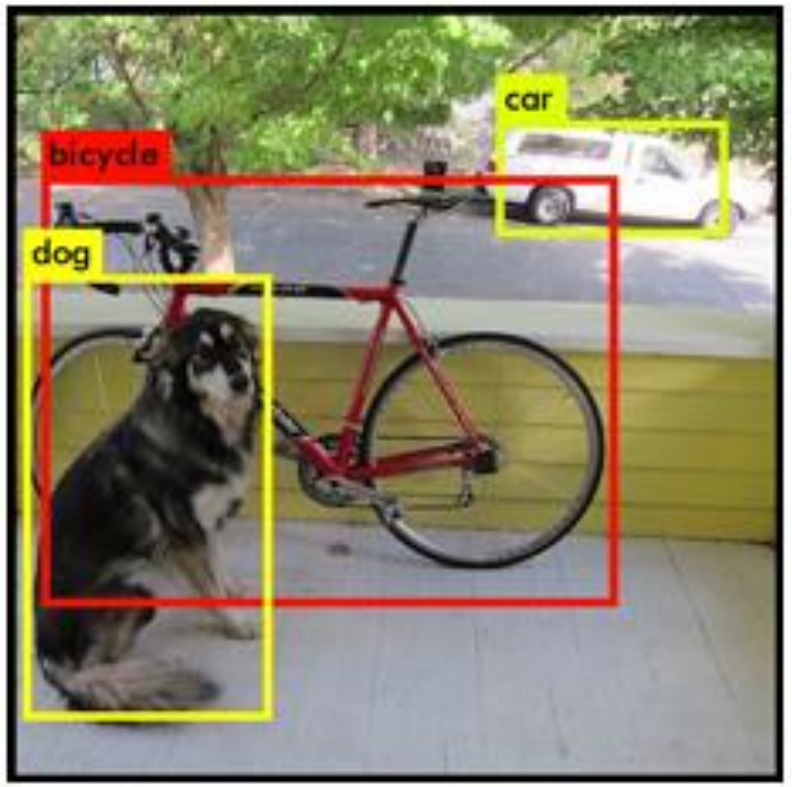


Object detection

Detección de objetos

Detección de objetos





Sistema de Visual Sensing

EN QUÉ CONSISTE

PASO 1. Detección de persona

Sistema de Visual Sensing

EN QUÉ CONSISTE

PASO 2. Obtener firma foto

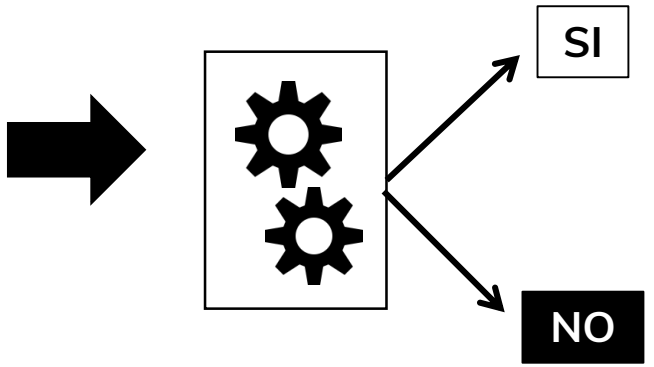
0.250, 0.328, 0.432, 0.378, ...

Sistema de Visual Sensing

EN QUÉ CONSISTE

PASO 3. Generar clasificador

¿Lleva chaleco?

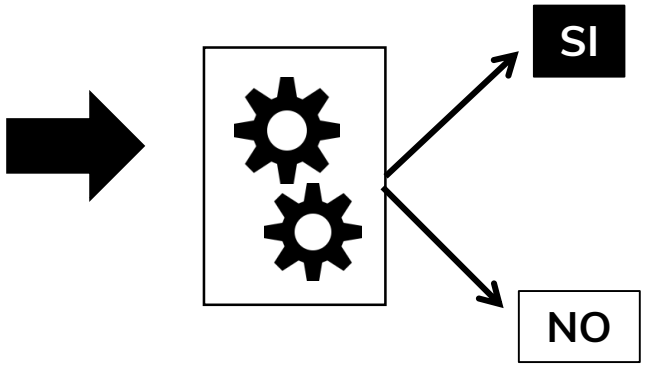


Sistema de Visual Sensing

EN QUÉ CONSISTE

PASO 3. Generar clasificador

¿Lleva chaleco?



Sistema de Visual Sensing

- Sistema Visual Sensing: detección de personas y medir tiempo en escena: https://youtu.be/V09eFqPVFcQ
- Sistema Visual Sensing: detectar si una persona lleva EPI: https://youtu.be/fFG6nM45t5Y

CASO 4

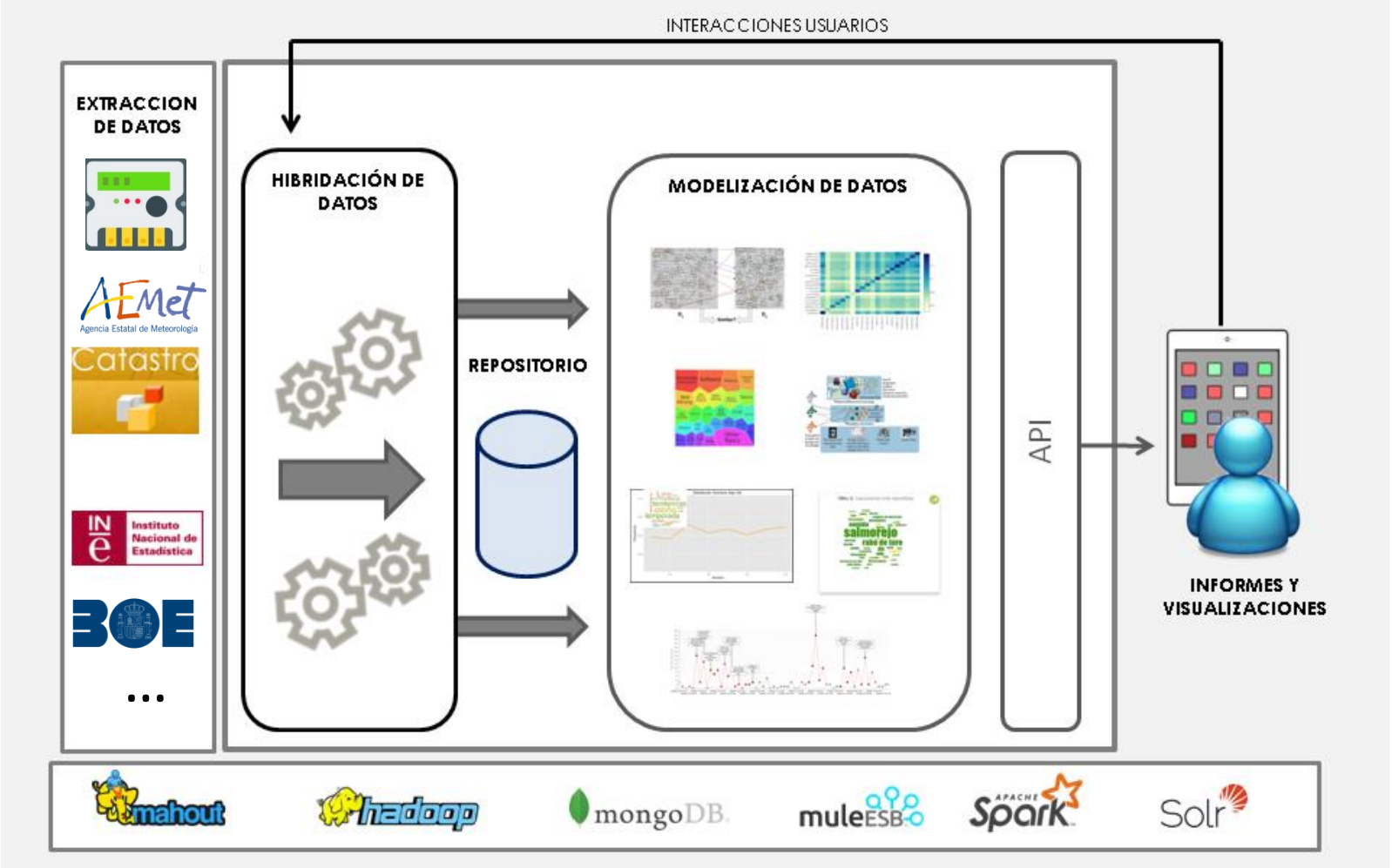
PLATAFORMA PLICACIÓN

NSUMOS ERGÉTICOS

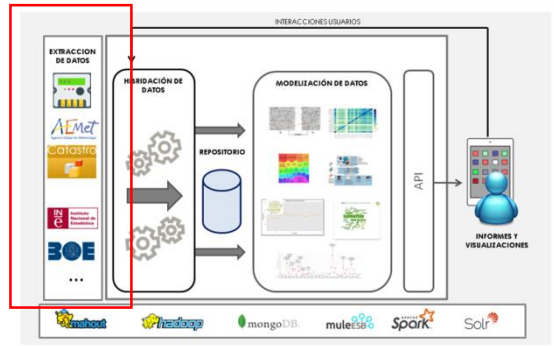
Contadores inteligentes

Objetivo: mejorar el conocimiento del consumo eléctrico

Desarrollo técnico

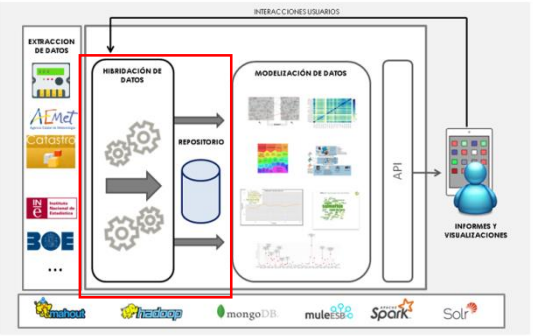


1. Extracción de datos



- FICHEROS: INE, BOE, …

2. Hibridación de datos

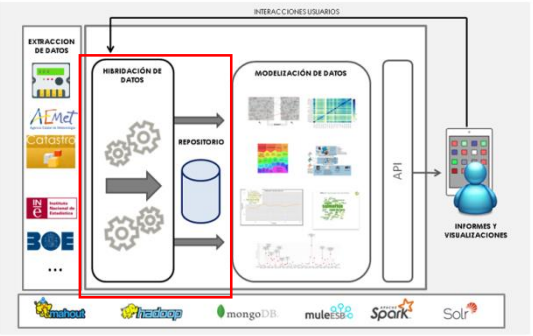


Estructuración de datos

María S., Auxiliadora, 25
Pto. Sta. María, Cádiz

- **Provincia:** Cádiz
- **Población:** El Puerto de Santa María
- **Código Postal:** 11500
- **Calle:** Auxiliadora
- **Número:** 25

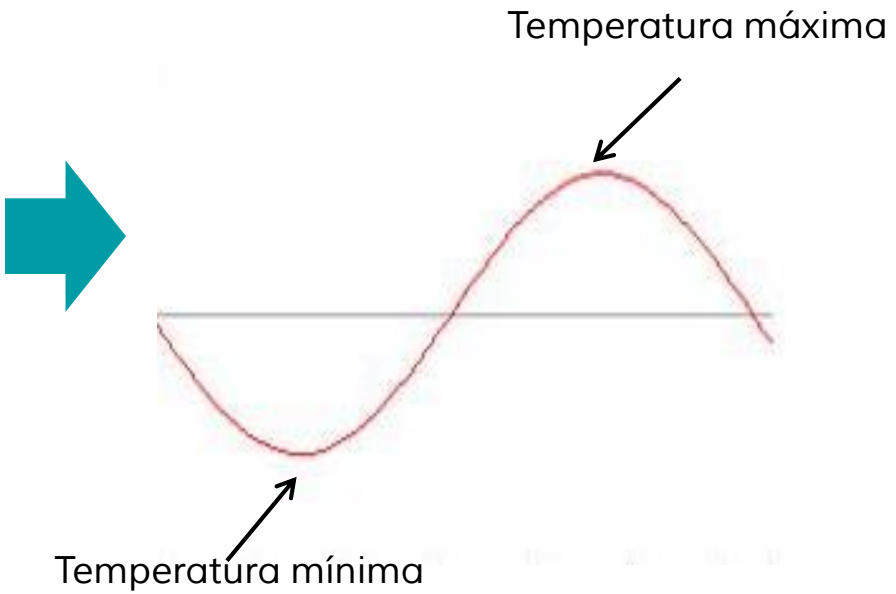
2. Hibridación de datos



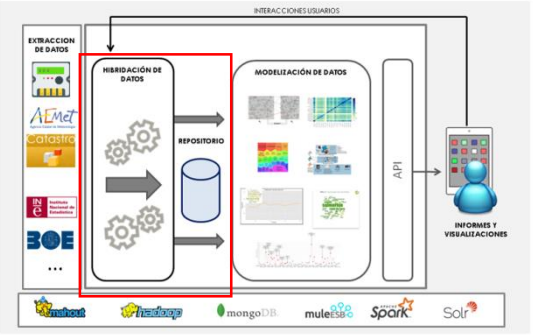
Homogeneización de datos

Tenemos la temperatura mínima y máxima del día

¿Qué hacemos para conocer la temperatura de cada hora del día?



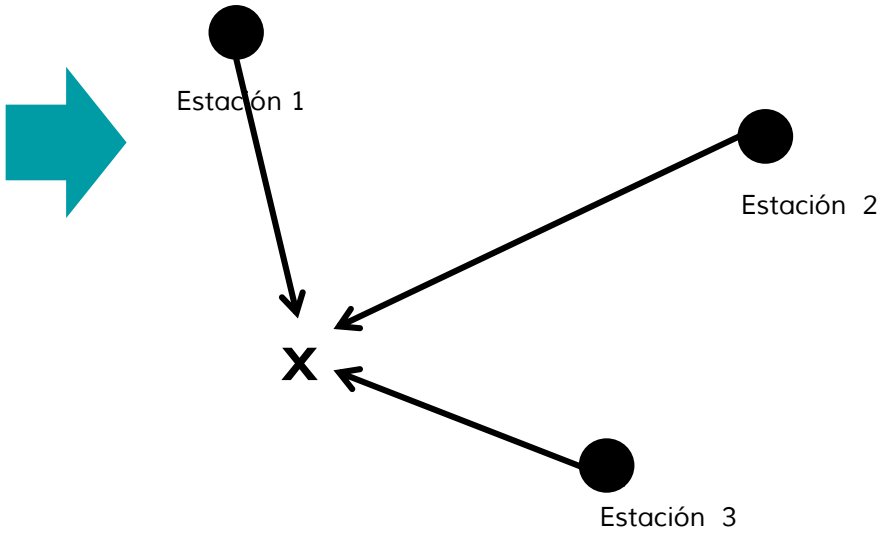
2. Hibridación de datos



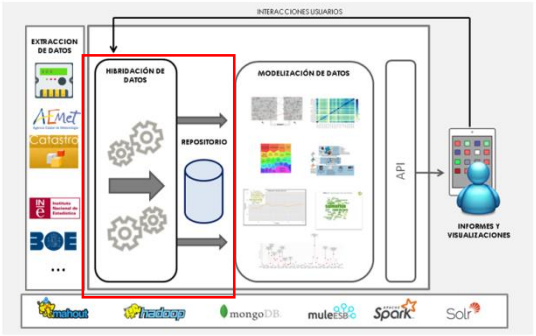
Homogeneización de datos

La AEMET nos da datos de unas 300 estaciones meteorológicas

¿Qué hacemos si queremos los datos de una coordenada?



2. Hibridación de datos



Armonización de datos

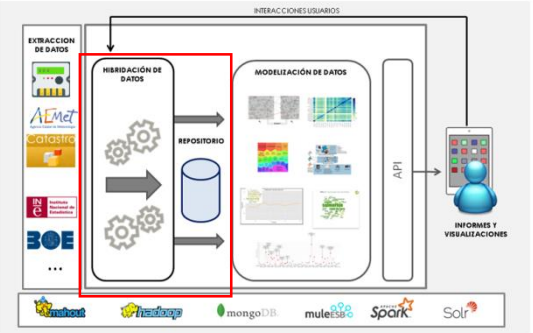
Festivo
Las Palmas, 5 marzo

INE

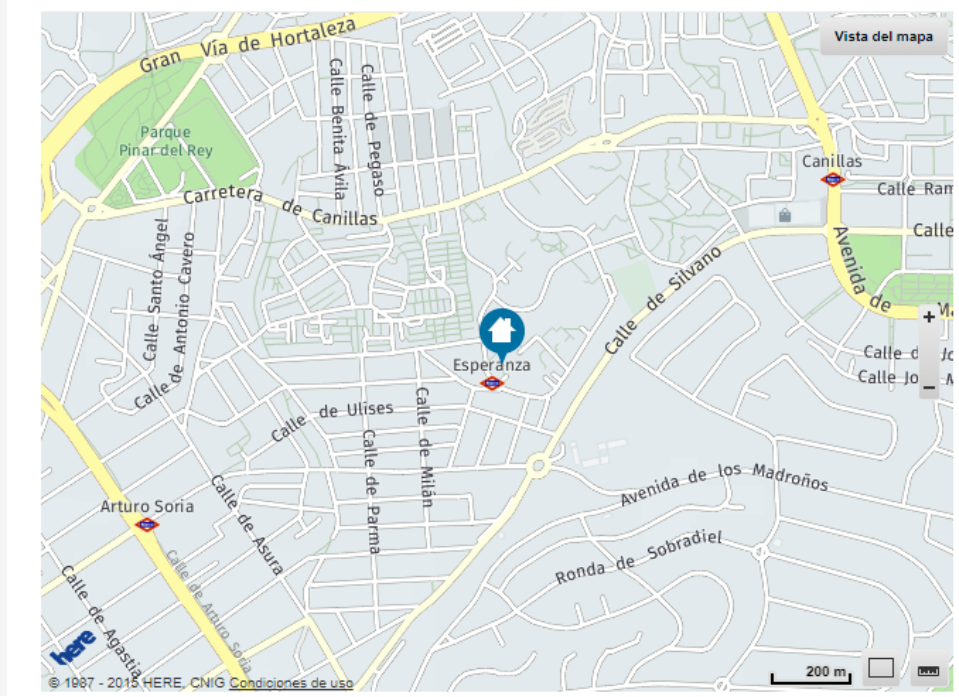
35 0167 Palmas de Gran Canaria (Las)

35 0167
5 marzo

2. Hibridación de datos



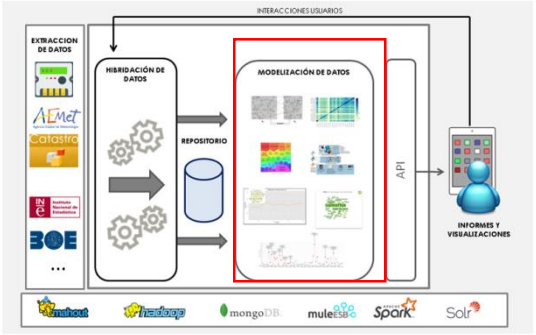
Enriquecimiento de los datos



DATOS DE OTRAS FUENTES

DEMOGRAFIA DE LA ZONA

3. Modelización



Analítica descriptiva

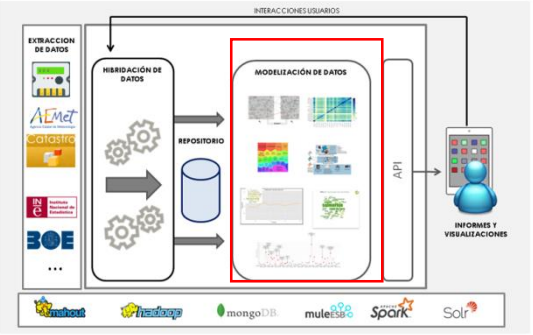
FACTORES

- Temperatura
- Precipitación
- Orto
- Festivos
- Día de la semana
- ....

MODELO APRENDIZAJE SUPERVISADO

CONSUMO

3. Modelización



Modelo de Procesamiento de Lenguaje Natural

INCREMENTO CONSUMO 10%

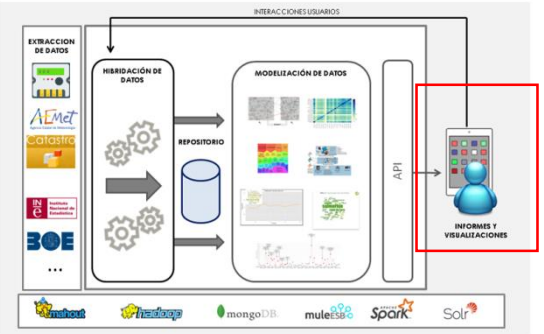
FACTORES

- Temperatura: 6%
- Precipitación: 2%
- Fin de semana: 1%
- Indeterminados: 2%

MODELO NLG

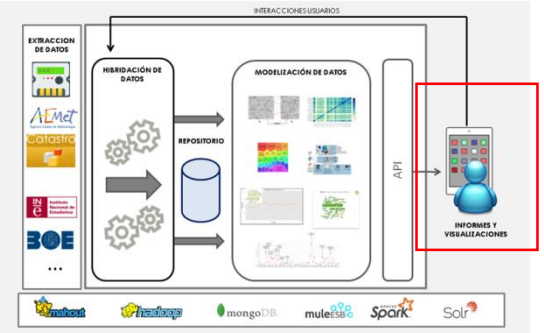
Aunque su consumo se ha incrementado (10%) respecto al mismo mes del año anterior, esto parece obedecer a que este mes ha sido más frío y lluvioso que el año anterior y a que ha habido un domingo más

4. Consumo del resultado



WEB SERVICE (API)

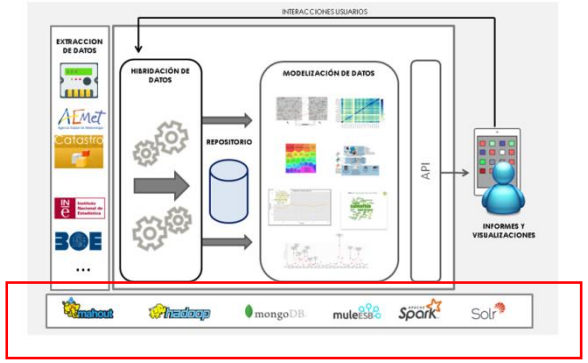
4. Consumo del resultado



Plataforma de datos para la explicación de consumos energético basada en Natural Language Generation y generación de alertas inteligentes con Alexa:

https://youtu.be/dRzRW1D_gk0?list=PLse23TZTsnNFP_qKlwkTKHSCO6UvdBkFk

5. Arquitectura de soporte



amazon web services™

Solr

mongoDB

mahout

Google Maps

¡Gracias!

insights analytics i+a España

LA COMUNIDAD DE MARKET RESEARCH Y DATA SCIENCE

INSIGHTS + ANALYTICS ESPAÑA
C/ Alberto Bosch 13 – 4ª planta, 28014 Madrid
Telf: 91 330 07 19 - secretaria@ia-espana.org
www.ia-espana.es