

Generación de Lenguaje Natural y Síntesis de Voz

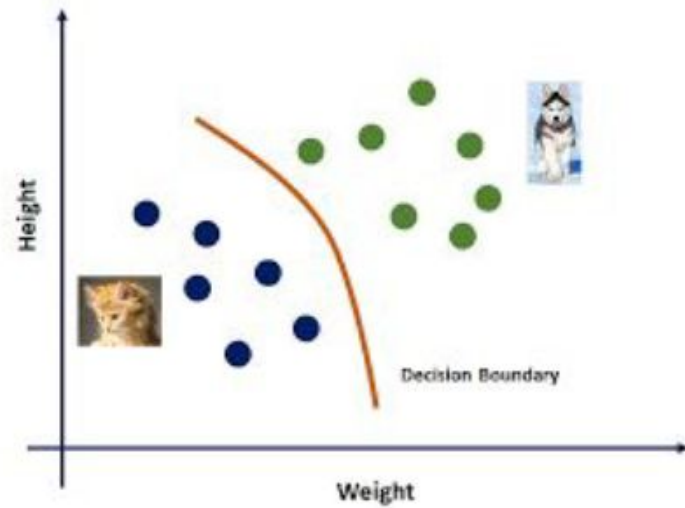
OBS Business School

Introducción

- Modelos generativos
- Embeddings e IA multimodal
- Generación de sonido
 - Síntesis de voz
 - Generación de música

Modelos generativos

Discriminative

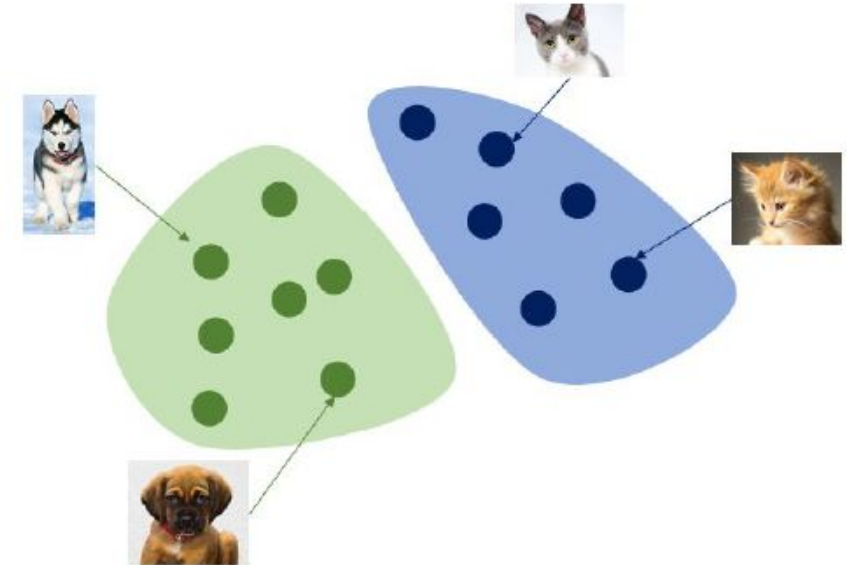


Model of the conditional probability of the target Y , given an observation x

Features $X \rightarrow Y$ Target

$$P(Y|X)$$

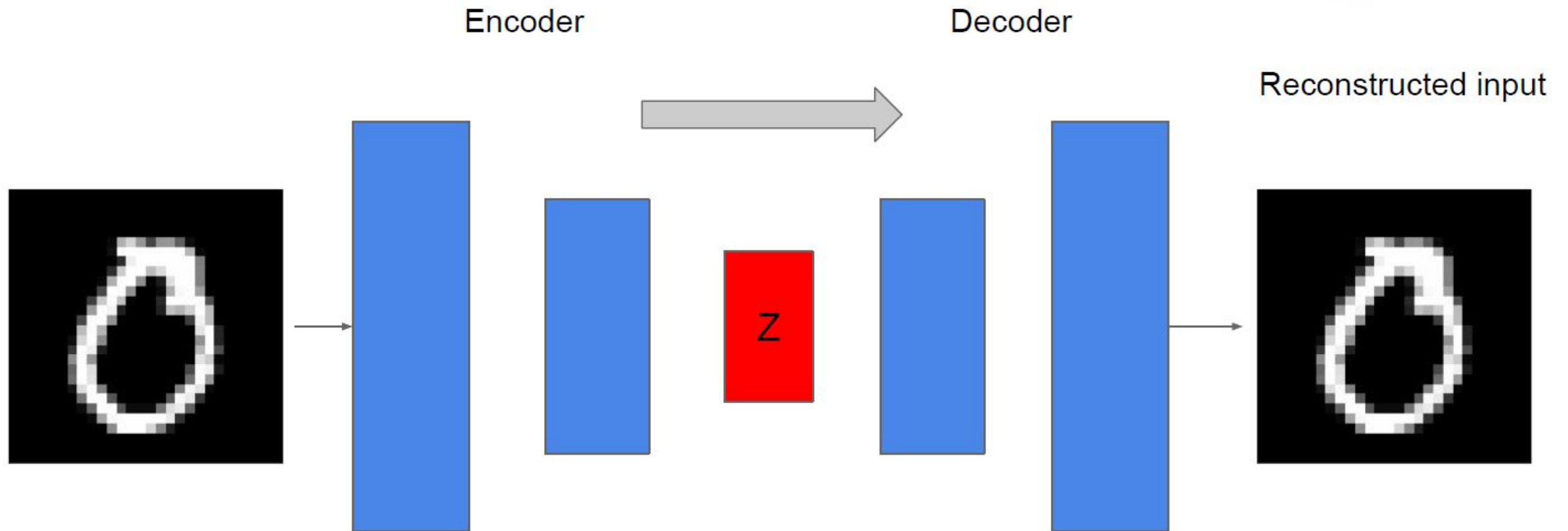
Generative



Target $Y \rightarrow X$ Features

$$P(X|Y)$$

Un autoencoder es una red neuronal que aprende a comprimir datos a una representación más pequeña y luego reconstruirlos a su forma original.



2D latent space



7	2	1	0	9	1	9	9	8	9
0	6	9	0	1	5	9	7	3	4
9	6	6	5	9	0	7	9	0	1
3	1	3	0	7	2	7	1	2	1
1	7	4	2	3	5	1	2	9	4
6	3	5	5	6	0	4	1	9	5
7	8	9	2	9	9	6	4	3	0
7	0	2	9	1	9	3	2	9	7
9	6	2	7	3	9	7	3	6	1
3	6	9	3	1	4	1	7	6	9

5D latent space



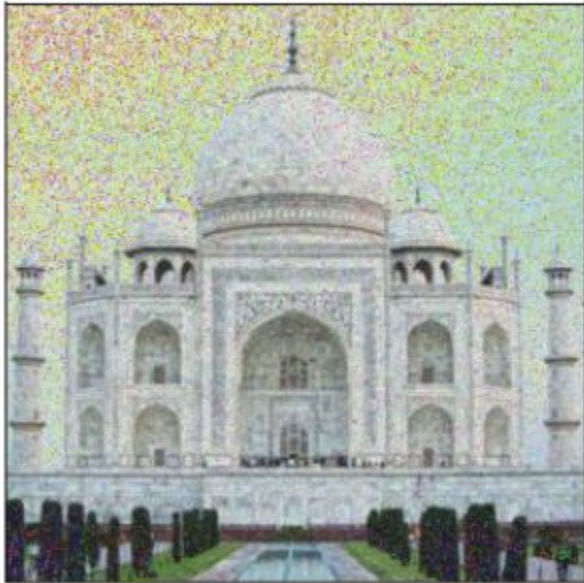
7	2	1	0	9	1	4	9	9	9
0	6	9	0	1	5	9	7	3	4
9	6	6	5	9	0	7	4	0	1
3	1	3	6	7	2	7	1	2	1
1	7	4	2	3	5	1	2	9	4
6	3	5	5	6	0	4	1	9	5
7	8	9	2	9	9	6	4	3	0
7	0	2	9	1	9	3	2	9	7
9	6	2	7	3	9	7	3	6	1
3	6	9	3	1	4	1	7	6	9

Ground Truth



7	2	1	0	9	1	4	9	9	9
0	6	9	0	1	5	9	7	3	4
9	6	6	5	9	0	7	4	0	1
3	1	3	6	7	2	7	1	2	1
1	7	4	2	3	5	1	2	9	4
6	3	5	5	6	0	4	1	9	5
7	8	9	2	9	9	6	4	3	0
7	0	2	9	1	9	3	2	9	7
9	6	2	7	3	9	7	3	6	1
3	6	9	3	1	4	1	7	6	9

Noisy Image

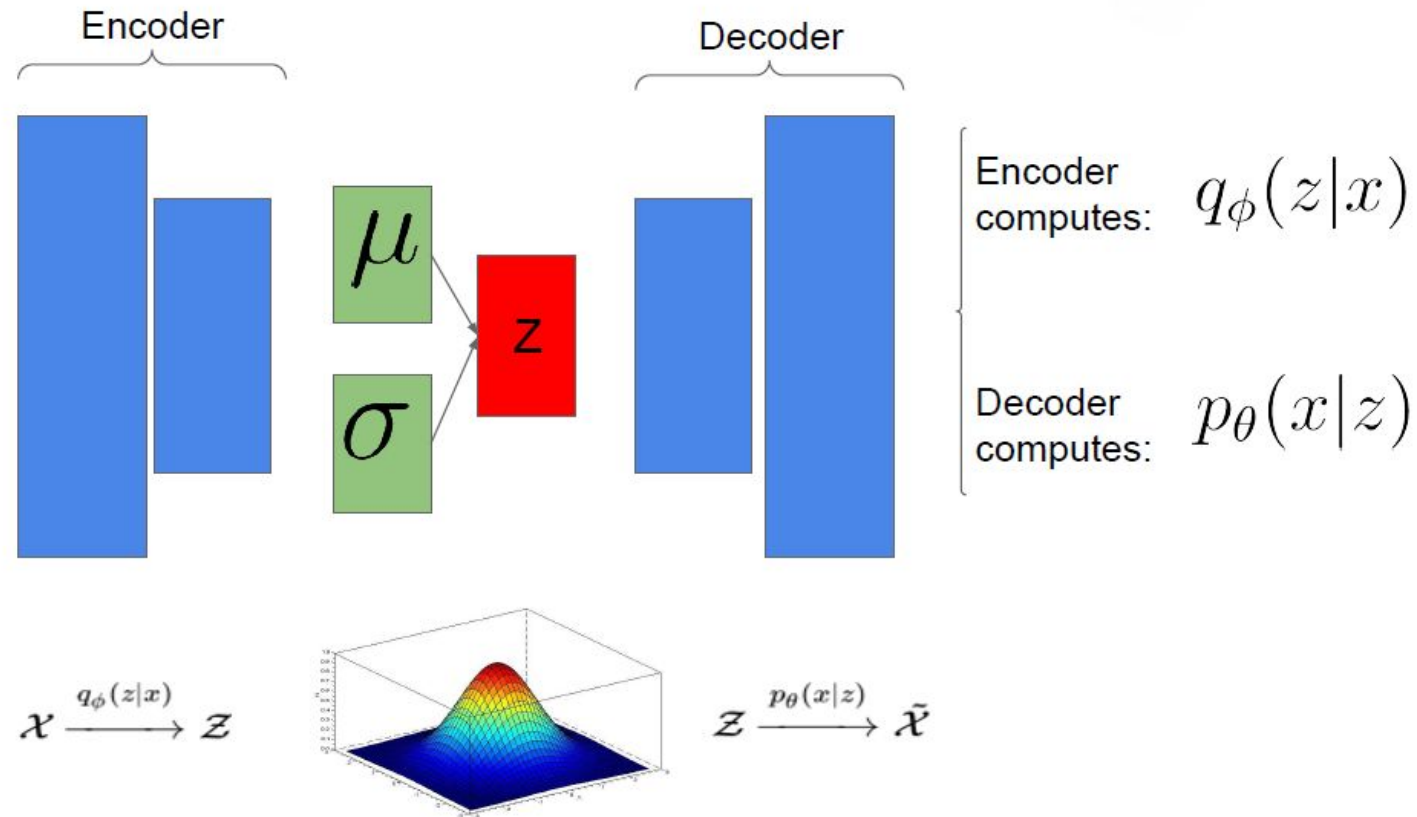


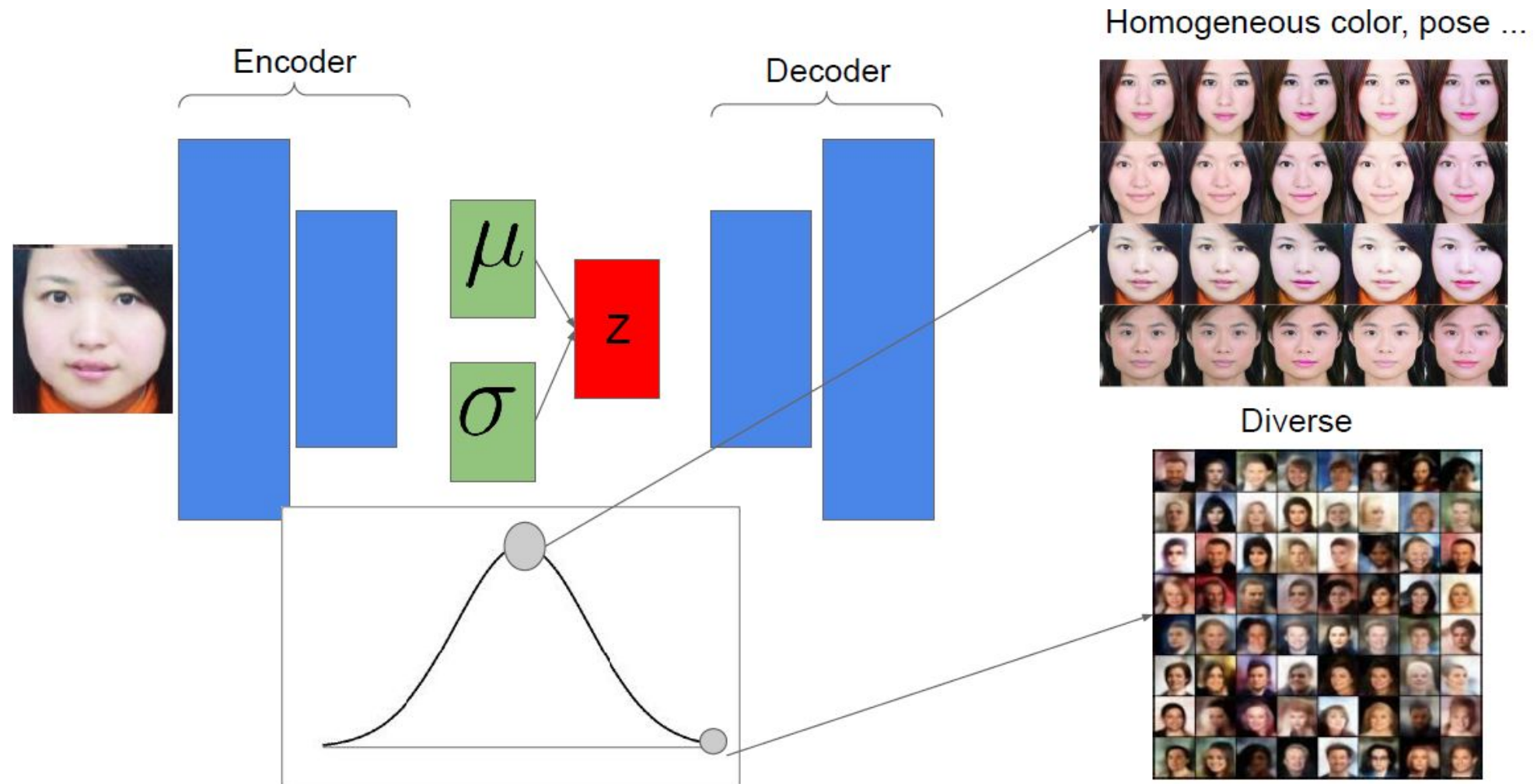
Denoised Image





Un Variational Autoencoder es una variante de los autoencoders que introduce un componente probabilístico. En lugar de codificar los datos a un punto fijo en el espacio latente, se codifican como distribuciones probabilísticas, lo que permite generar nuevos datos similares a los originales mediante el muestreo de estas distribuciones.









Las GAN son un tipo de red neuronal compuesta por dos sub-redes:

- Generador: El generador crea datos falsos similares a los datos reales.
- Discriminador: Intenta distinguir entre los datos reales y los falsos.

Estas redes se entrenan de manera competitiva, mejorando continuamente hasta que el generador produce datos tan realistas que el discriminador no puede diferenciarlos de los datos reales.

Generator

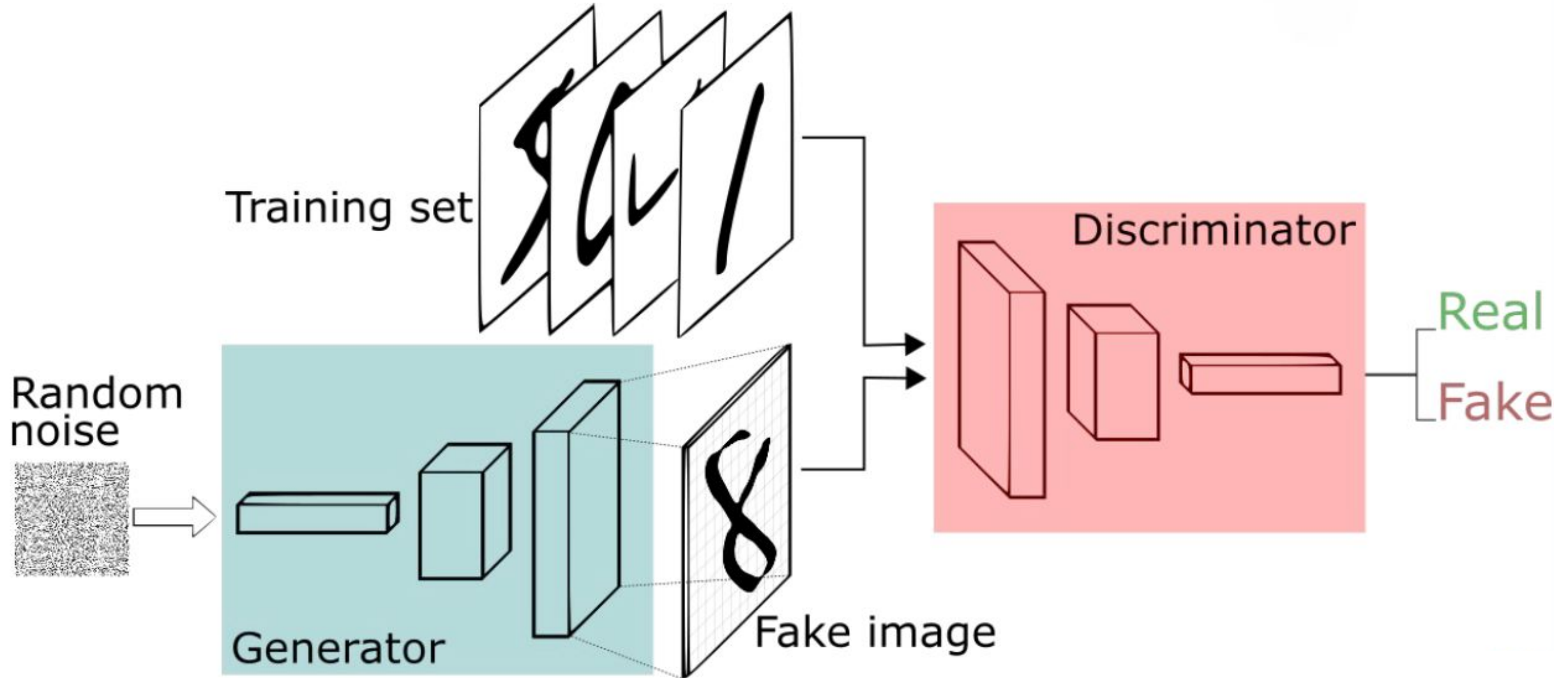
Learn to sample **fake** data, that **looks real**.

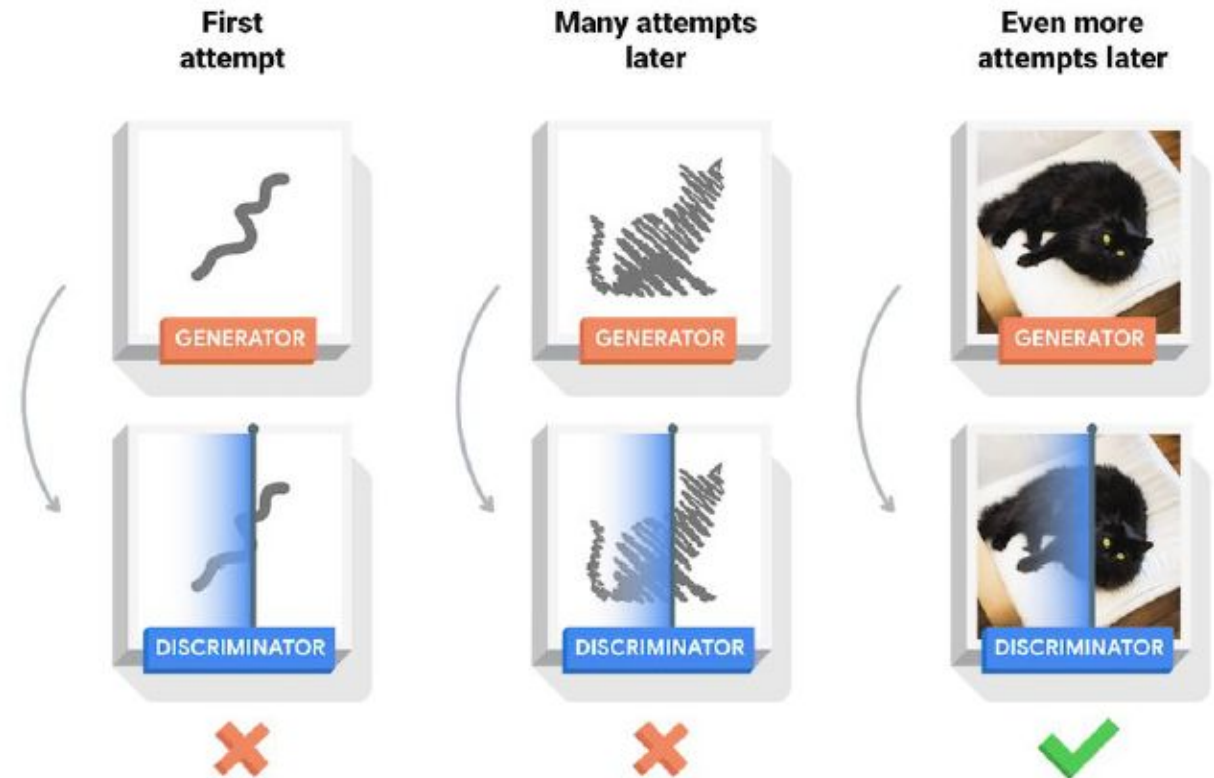
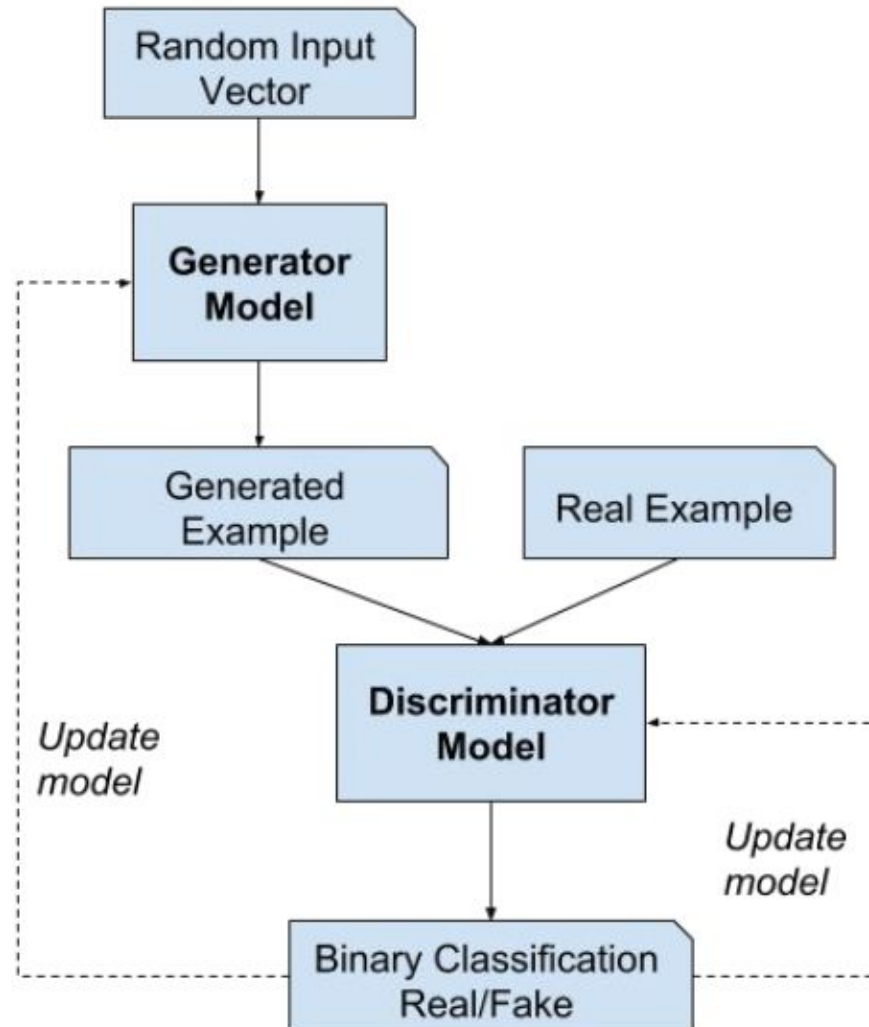


Discriminator

Learns to **distinguish** real from fake data.







Generator and discriminator, are trained together. The generator generates a batch of samples, and combined with real data are provided to the discriminator and classified as real or fake.

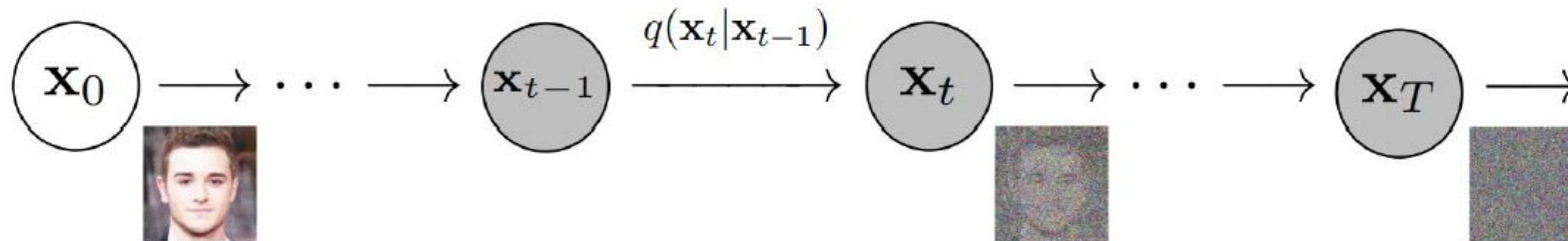


Link: <https://thispersondoesnotexist.com/>

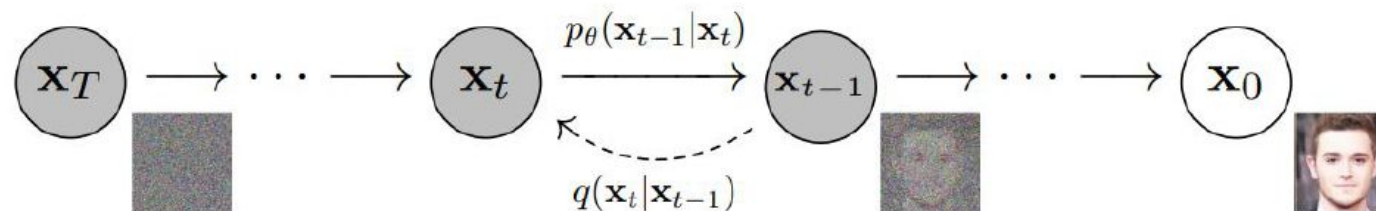
Los modelos de difusión son una clase de modelos generativos que aprenden a generar datos donde se añade ruido de forma incremental a los datos hasta que se convierten en ruido puro. Durante el entrenamiento, el modelo aprende a revertir este proceso, eliminando el ruido paso a paso para generar nuevos datos que se asemejan a los datos originales.

The Diffusion Process

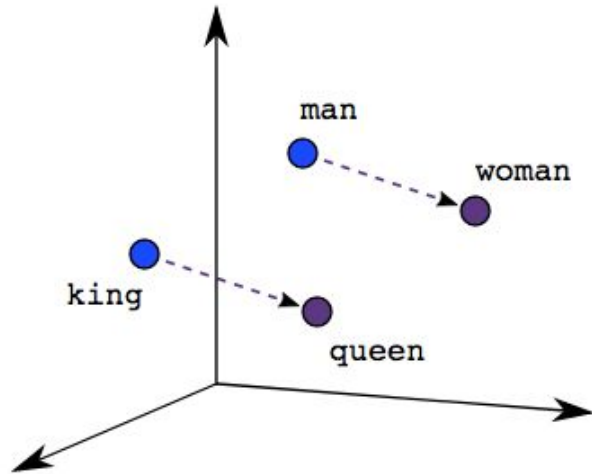
- Diffusion process: original data sample is progressively corrupted by adding noise.



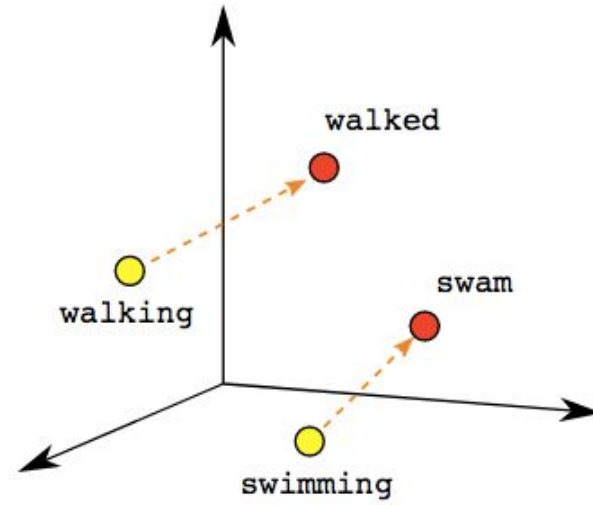
- Reverse process (denoising): learn to recover the data by reversing this noising process.



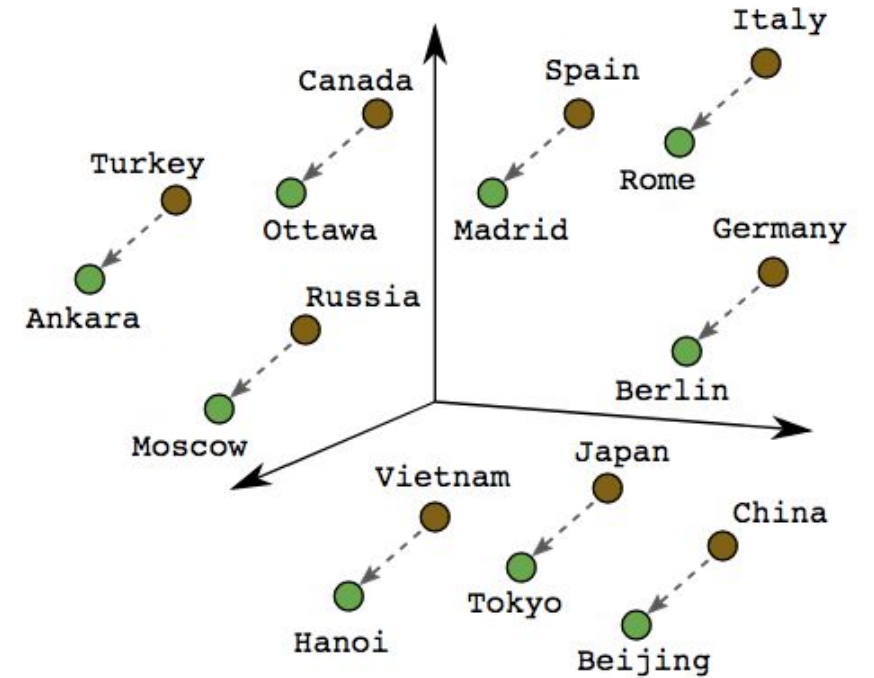
Embeddings e IA multimodal



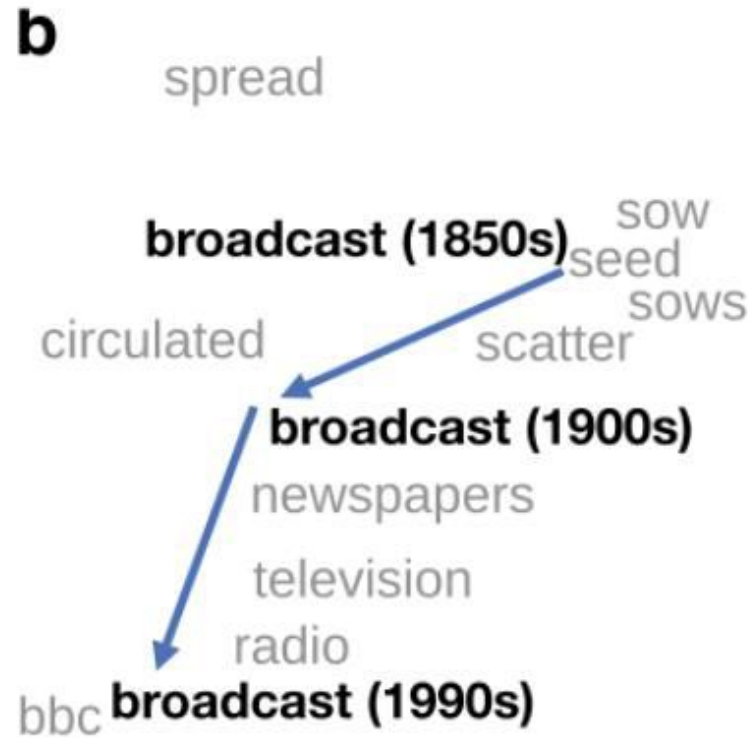
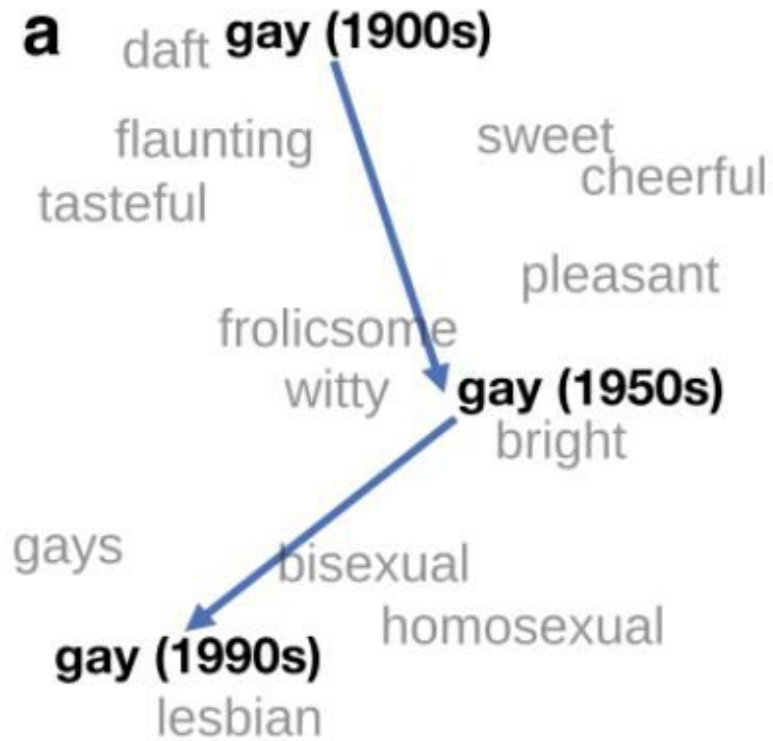
Male-Female

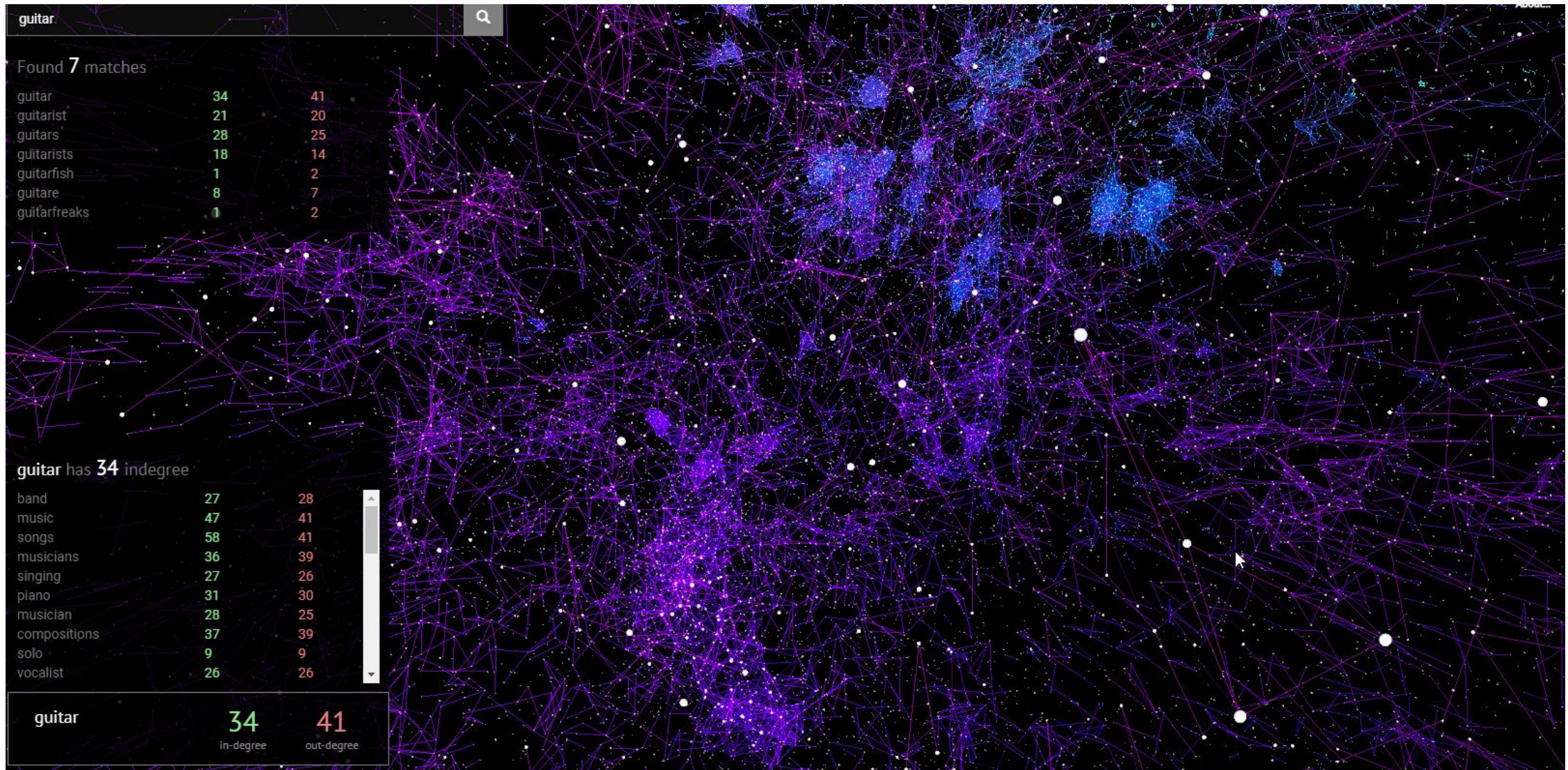


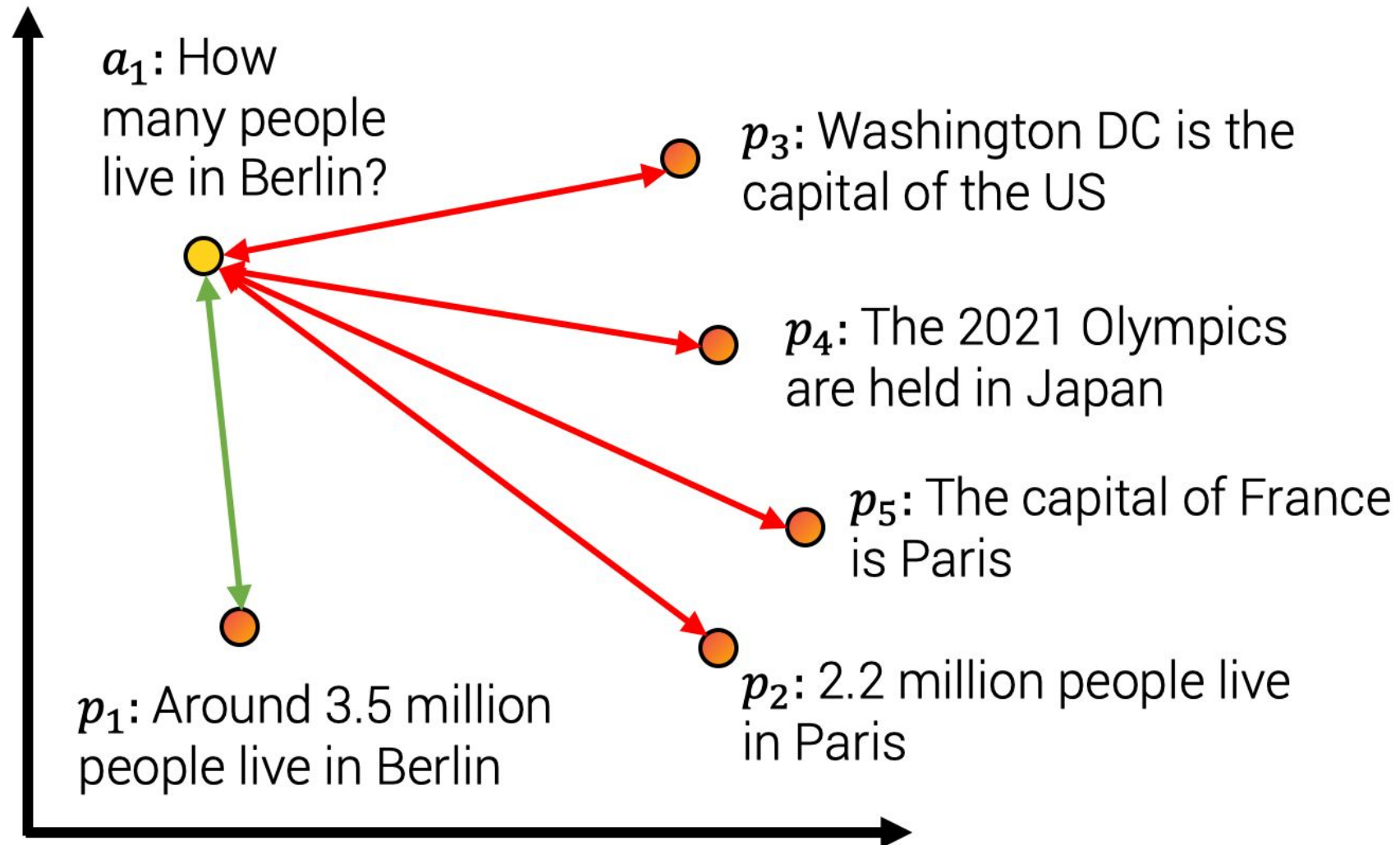
Verb Tense

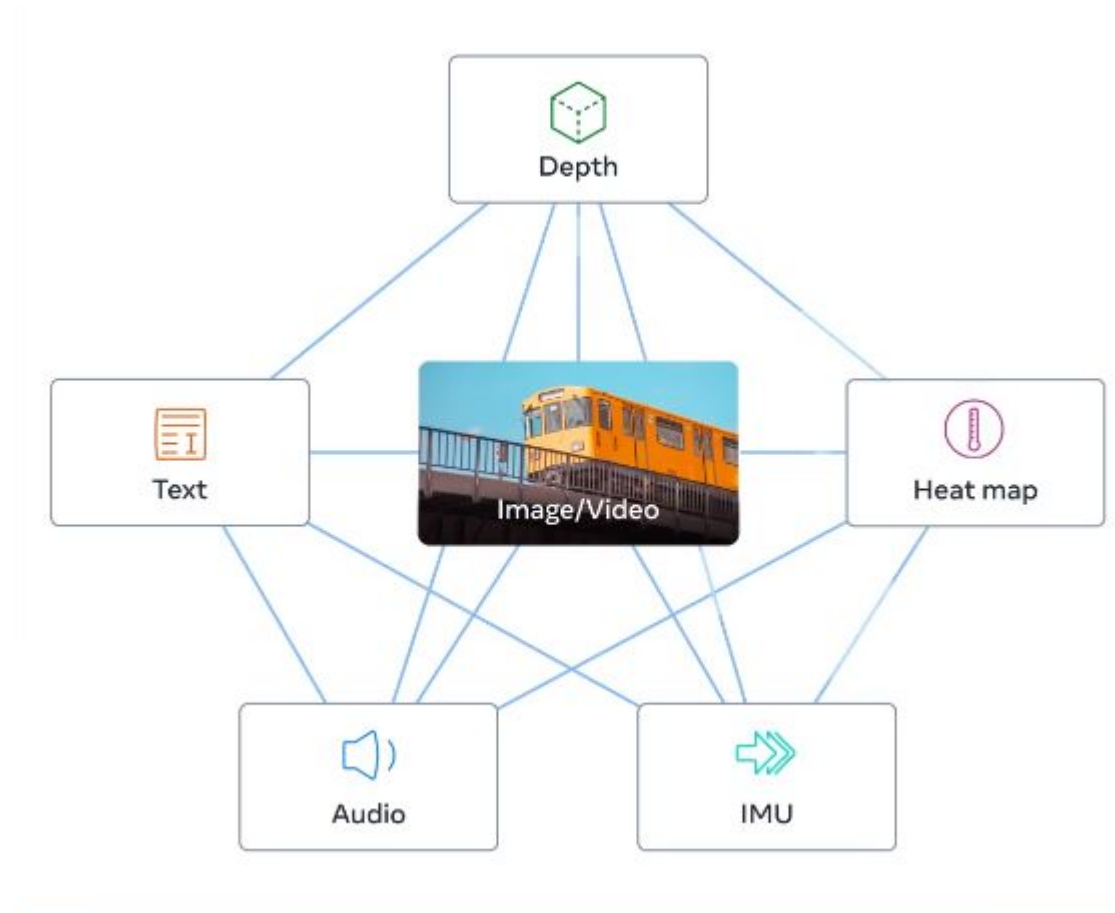


Country-Capital





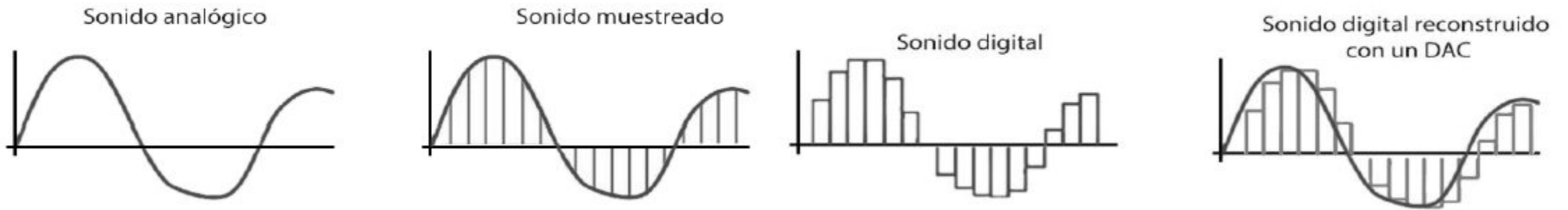




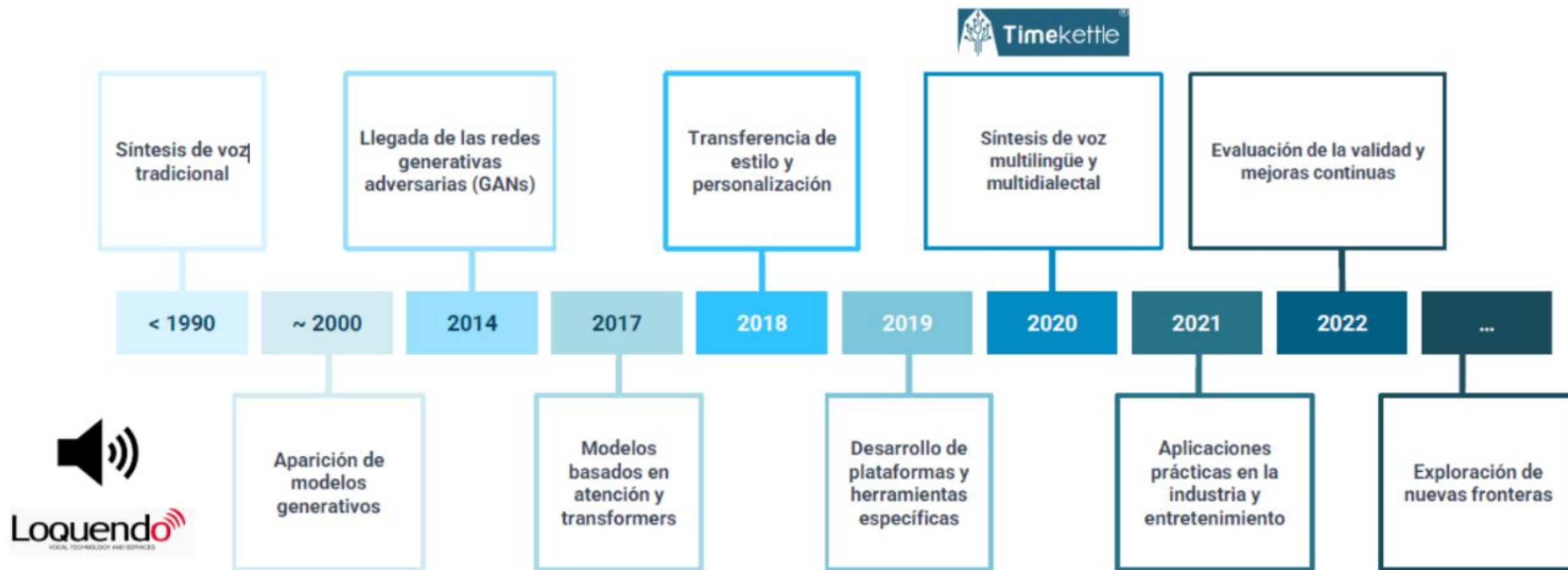
Link: <https://imagebind.metademolab.com/>

Generación de sonido

El sonido es una vibración que se propaga como una onda acústica, a través de un medio de transmisión como un gas, líquido o sólido.



- **Naturaleza de la señal de entrada:**
 - Voz: Énfasis en la coherencia lingüística, entonación y pronunciación.
 - Música: Creación de sonidos más abstractos y complejos, expresando emociones y creatividad artística.
- **Complejidad de la información:**
 - Voz: Detalles específicos del habla.
 - Música: Implica manipular múltiples capas de información, como tonos, ritmos, armonías y dinámicas, con una complejidad potencialmente más elevada.
- **Contexto temporal y secuencial:**
 - Voz: Aborda la naturaleza secuencial del habla.
 - Música: También tiene una dimensión temporal, pero la estructura secuencial puede ser más flexible y depende del género musical específico.
- **Personalización y estilo:**
 - Voz: Se centra en replicar y adaptar características vocales, como acentos y tonos.
 - Música: La transferencia de estilo es crucial para adaptarse a diferentes géneros y estilos artísticos.
- **Aplicaciones prácticas y evaluación de calidad:**
 - Voz: Se aplica en asistentes virtuales y accesibilidad, evaluándose por naturalidad y capacidad de expresar emociones.
 - Música: Se aplica en bandas sonoras y composición automática, evaluándose por autenticidad, coherencia estilística y capacidad para evocar emociones deseadas.



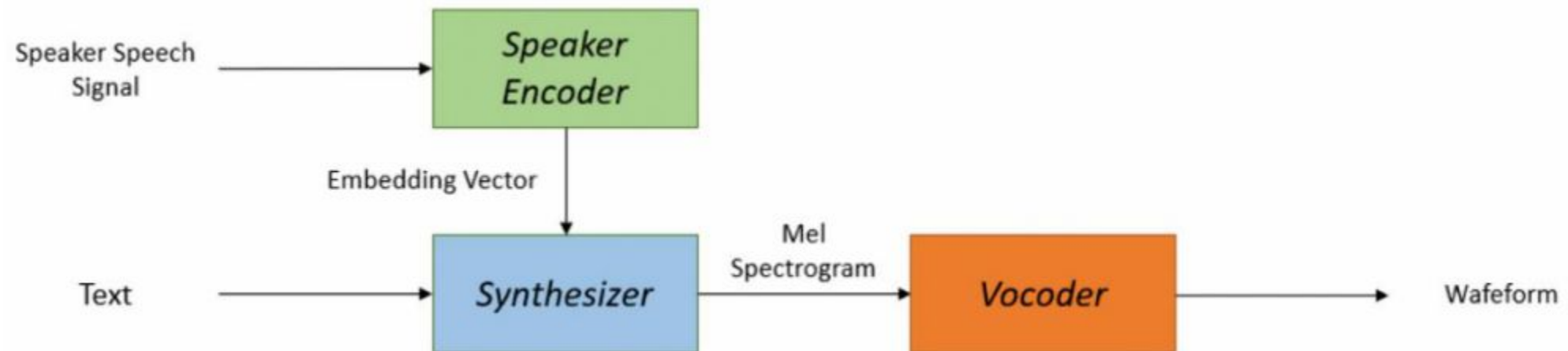
- **Physical Modeling Synthesis:** Modelos matemáticos para simular las propiedades físicas de instrumentos musicales reales. Aunque más tradicional, sigue siendo un método poderoso para generar sonidos realistas de instrumentos.
- **Speech Synthesis Markup Language (SSML):** Un lenguaje de marcado que ayuda a los desarrolladores a controlar aspectos de la síntesis de voz, como la voz, la pronunciación y el tono.
- **Variational Autoencoders (VAE):** Modelos de aprendizaje automático que comprimen datos y generan nuevos sonidos o música que se asemejan estrechamente a las entradas originales.
- **Generative Adversarial Networks (GAN):** Modelos de aprendizaje automático que utilizan dos redes neuronales en competencia para generar nuevos audios de sonido natural basados en datos de entrenamiento.

SSML proporciona una manera estructurada de incorporar información adicional en el texto que será convertido en voz, permitiendo especificar ciertos elementos, como la entonación, el ritmo, la velocidad y otros atributos vocales.

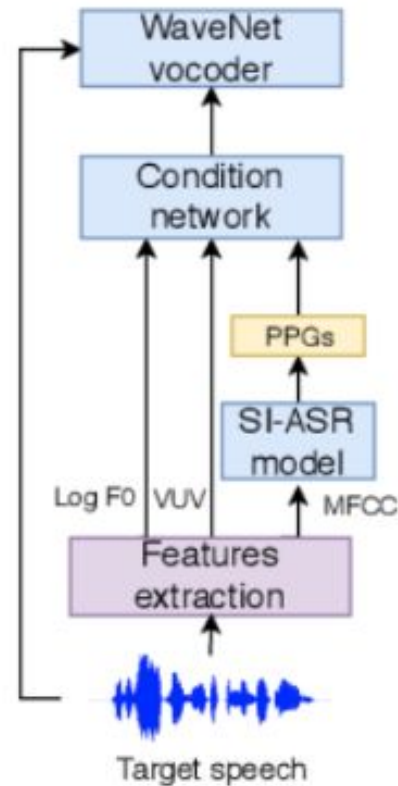
```
< speak>  
¡Hola! < prosody rate="fast">Espero que estés teniendo un buen día.</ prosody> < break time="500ms"/>  
Permíteme mostrarte algo interesante.  
< p>< s>Primero,</ s> vamos a < emphasis level="moderate">enfaticar</ emphasis> esta palabra.</ p>  
< p>También podemos hacer una pausa < break time="1s"/> para efecto dramático.</ p>  
< s>Finalmente, podemos cambiar la velocidad de habla: < prosody rate="slow">más lento</ prosody>.</ s>  
</ speak>
```

La generación de voz a partir de texto se realiza combinando tres neuronales diferentes:

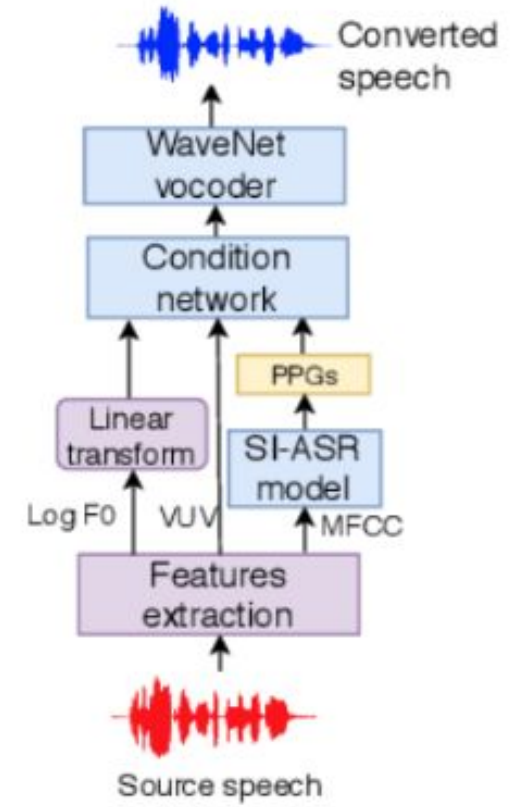
- Codificador de hablante (Speaker Encoder): Modelo de IA que convierte la voz de una persona en una representación numérica, capturando características distintivas de su habla para análisis o síntesis.
- Sintetizador (Synthesizer): Modelo de IA que genera señales de audio o habla artificial a partir de datos, como texto o representaciones numéricas, para crear voces sintéticas de alta calidad.
- Vocoder: Algoritmo o modelo de IA que modifica o sintetiza la voz procesando las características espectrales de una señal de audio, permitiendo la manipulación de la voz para crear efectos o cambiar sus características.



- Introducido en 2016 por DeepMind, WaveNet fue uno de los primeros modelos de IA en generar habla con un sonido natural.
- Orientado a modificar un audio original.
- Modificación del tono, timbre o ritmo de la voz.
- Voz producida de manera más natural.



(a) Training stage



(b) Conversion stage

MusicLM: Generating Music From Text

| paper | dataset |

Andrea Agostinelli, Timo I. Denk, Zoltan Borsos, Jesse Engel, Mauro Verzetti, Antoine Caillon, Qingqing Huang, Aren Jansen, Adam Roberts, Marco Tagliasacchi, Matt Sharifi, Neil Zeghidour, Christian Frank
Google Research

Abstract We introduce MusicLM, a model generating high-fidelity music from text descriptions such as “a calming violin melody backed by a distorted guitar riff”. MusicLM casts the process of conditional music generation as a hierarchical sequence-to-sequence modeling task, and it generates music at 24 kHz that remains consistent over several minutes. Our experiments show that MusicLM outperforms previous systems both in audio quality and adherence to the text description. Moreover, we demonstrate that MusicLM can be conditioned on both text and a melody in that it can transform whistled and hummed melodies according to the style described in a text caption. To support future research, we publicly release MusicCaps, a dataset composed of 5.5k music-text pairs, with rich text descriptions provided by human experts.

Audio Generation From Rich Captions

Caption

Generated audio

The main soundtrack of an arcade game. It is fast-paced and upbeat, with a catchy electric guitar riff. The music is repetitive and easy to remember, but with unexpected sounds, like cymbal crashes or drum rolls.

▶ 0:00 / 0:30 — 🔊 ⋮

A fusion of reggaeton and electronic dance music, with a spacey, otherworldly sound. Induces the experience of being lost in space, and the music would be designed to evoke a sense of wonder and awe, while being danceable.

▶ 0:00 / 0:30 — 🔊 ⋮

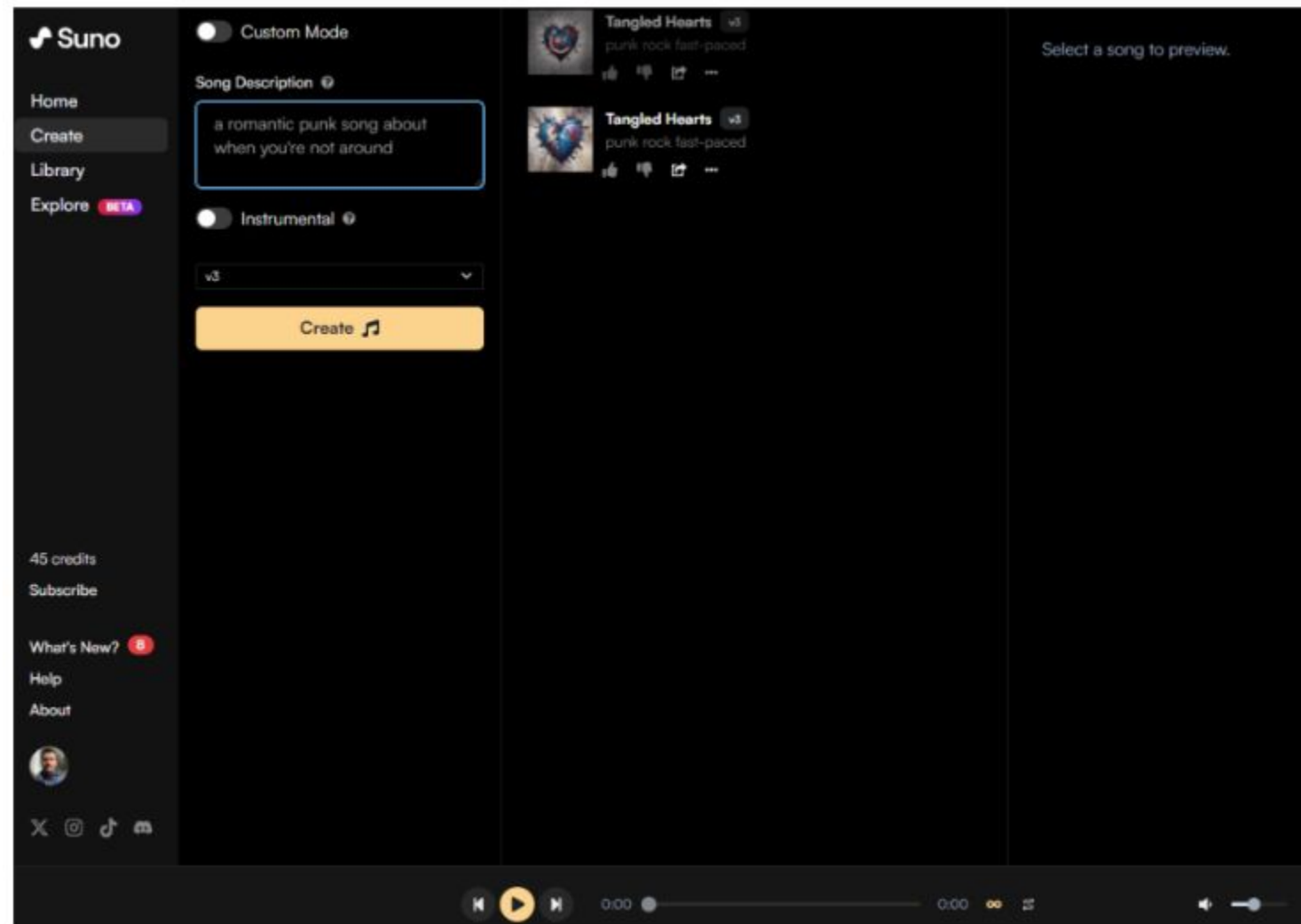
A rising synth is playing an arpeggio with a lot of reverb. It is backed by pads, sub bass line and soft drums. This song is full of synth sounds creating a soothing and adventurous atmosphere. It may be playing at a festival during two songs for a buildup.

▶ 0:00 / 0:30 — 🔊 ⋮

Slow tempo, bass-and-drums-led reggae song. Sustained electric guitar, High-pitched bongos with ringing tones. Vocals are relaxed with a laid-back feel, very expressive.

▶ 0:00 / 0:30 — 🔊 ⋮

1 2 3



Link: <https://suno.com/create>

The screenshot shows the Hugging Face website interface. At the top, there's a navigation bar with links for Models, Datasets, Spaces, Posts, Docs, Solutions, and Pricing. Below this is a search bar and a sidebar menu. The sidebar menu is divided into 'TASK GUIDES' and 'DEVELOPER GUIDES'. Under 'TASK GUIDES', there are links for Natural Language Processing, Audio, Computer Vision, and Multimodal. Under 'DEVELOPER GUIDES', there are links for Use fast tokenizers from Transformers, Run inference with multilingual models, Use model-specific APIs, Share a custom model, Templates for chat models, Trainer, Run training on Amazon SageMaker, and Export to ONNX. The main content area is titled 'Text to speech' and contains a paragraph explaining TTS, a code snippet for generating audio, and a link to an audio course. The right sidebar shows a list of tasks related to 'Text to speech'.

Hugging Face Search models, datasets, users...

Models Datasets Spaces Posts Docs Solutions Pricing

Transformers Search documentation Ctrl+K

V4.41.2 EN 126,832

TASK GUIDES

- NATURAL LANGUAGE PROCESSING
- AUDIO
- COMPUTER VISION
- MULTIMODAL
 - Image captioning
 - Document Question Answering
 - Visual Question Answering
 - Text to speech**

GENERATION

PROMPTING

DEVELOPER GUIDES

- Use fast tokenizers from Transformers
- Run inference with multilingual models
- Use model-specific APIs
- Share a custom model
- Templates for chat models
- Trainer
- Run training on Amazon SageMaker
- Export to ONNX

Text to speech

Text-to-speech (TTS) is the task of creating natural-sounding speech from text, where the speech can be generated in multiple languages and for multiple speakers. Several text-to-speech models are currently available in 🤗 Transformers, such as [Bark](#), [MMS](#), [VITS](#) and [SpeechT5](#).

You can easily generate audio using the "text-to-audio" pipeline (or its alias - "text-to-speech"). Some models, like Bark, can also be conditioned to generate non-verbal communications such as laughing, sighing and crying, or even add music. Here's an example of how you would use the "text-to-speech" pipeline with Bark:

```
>>> from transformers import pipeline

>>> pipe = pipeline("text-to-speech", model="suno/bark-small")
>>> text = "[clears throat] This is a test ... and I just took a long pause."
>>> output = pipe(text)
```

Here's a code snippet you can use to listen to the resulting audio in a notebook:

```
>>> from IPython.display import Audio
>>> Audio(output["audio"], rate=output["sampling_rate"])
```

For more examples on what Bark and other pretrained TTS models can do, refer to our [Audio course](#).

If you are looking to fine-tune a TTS model, the only text-to-speech models currently available in 🤗 Transformers are [SpeechT5](#) and [FastSpeech2Conformer](#), though more will be added in the future. SpeechT5 is pre-trained on a combination of speech-to-text and text-to-speech data, allowing it to learn a unified space of hidden representations shared by both text and speech. This means that the same pre-trained model can be fine-tuned for different tasks. Furthermore, SpeechT5 supports multiple speakers


Text to speech

- Load the dataset
- Preprocess the data
 - Text cleanup for SpeechT5 tokenization
- Speakers
- Speaker embeddings
- Processing the dataset
- Data collator
- Train the model
- Inference
 - Inference with a pipeline
 - Run inference manually

Links: <https://huggingface.co/docs/transformers/tasks/text-to-speech>

MusicGen

This is the demo for [MusicGen](#), a simple and controllable model for music generation presented at: "[Simple and Controllable Music Generation](#)".

 **Duplicate Space** for longer sequences, more control and no queue.

Describe your music

Condition on a melody (optional) File or Mic

☒ file

☐ mic

File

Coloque el audio aquí
- 0 -
Haga click para cargar

Generated Music

Generated Music (wav)

Generate

Examples

Describe your music	File
An 80s driving pop song with heavy drums and synth pads in the background	bach.mp3
A cheerful country song with acoustic guitars	bolero_ravel.mp3
90s rock song with electric guitar and heavy drums	
a light and cheerly EDM track, with syncopated drums, aery pads, and strong emotions bpm: 130	bach.mp3
lofi slow bpm electro chill with organic samples	

Links: <https://huggingface.co/spaces/facebook/MusicGen>

Próximos pasos



Forbes



Ranking Educativo
Innovatec



EL MUNDO



- Repaso y lectura de los conceptos tratados en la 3ª sesión.
- Actividad evaluable:
 - Test multirespuesta temas 2 y 3 (individual).
 - Caso práctico (equipos TFM).
- Próxima sesión: miércoles 5 de junio a las 19:00 (CEST).
- Dudas y preguntas, vía *tablero de discusión* o email.

OBS Business
School



Planeta Formación y Universidades