

*IA Conversacional: más allá del  
“Perdona, no te he entendido”*

**Nieves Ábalos (@nieves\_as)**

Founder & Chief Product Officer, Monoceros Labs

# ¡Hola!



**Nieves Ábalos Serrano**

Chief Product Officer &  
Cofundadora de [Monoceros Labs](#)

Twitter: [@nieves\\_as](#)

LinkedIn: [/in/nievesabalosserrano](#)

- **UGR:** Ingeniera Software, Máster y PhD sin terminar en sistemas de diálogo.
- **Dpto Innovación BEEVA (BBVA Next Technologies):** I+D en conversacional y NLP.
- **Monoceros Labs:** emprendedora, estrategia y gestión de producto, diseño conversacional y desarrollo conversacional (lo que toque).
- **Alexa Champion.**
- **Women in Voice Spain (Lead)**
- **Comunidad de Alexa en español.**

# Perdona, no te he entendido



# IA conversacional

---

# IA Conversacional

The screenshot shows the top navigation bar of the Amazon Alexa website. It includes the logo "amazon alexa" with a blue circle icon, followed by links for "Alexa Skills Kit", "Alexa Voice Service", "Dispositivos conectados", "Programas de Alexa", "Docs", and "Blog de Alexa". On the right side, there are links for "Mis consolas Alexa", "Identificarse", a help icon with a question mark, and a search icon.

## Conversational AI: Computers That Talk

Conversational AI systems are computers that people can interact with simply by having a conversation, our most natural form of interaction. In short, it is what allows us to talk to voice-driven technologies like Amazon Alexa and ask about the weather, order products online, and even call a cab, simply by using the language we already know.

With conversational AI, voice-enabled devices like Amazon Echo are finally starting to enable the sort of magical interactions we've dreamed of for decades (think: Star Trek computer). Through a [voice user interface \(VUI\)](#), voice services like Alexa can communicate with people in ways that feel effortless, solve problems, and get smarter over time.



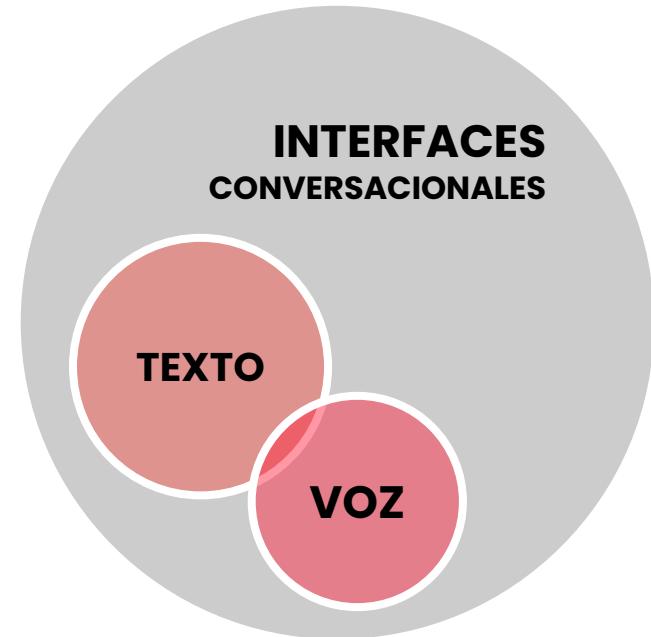
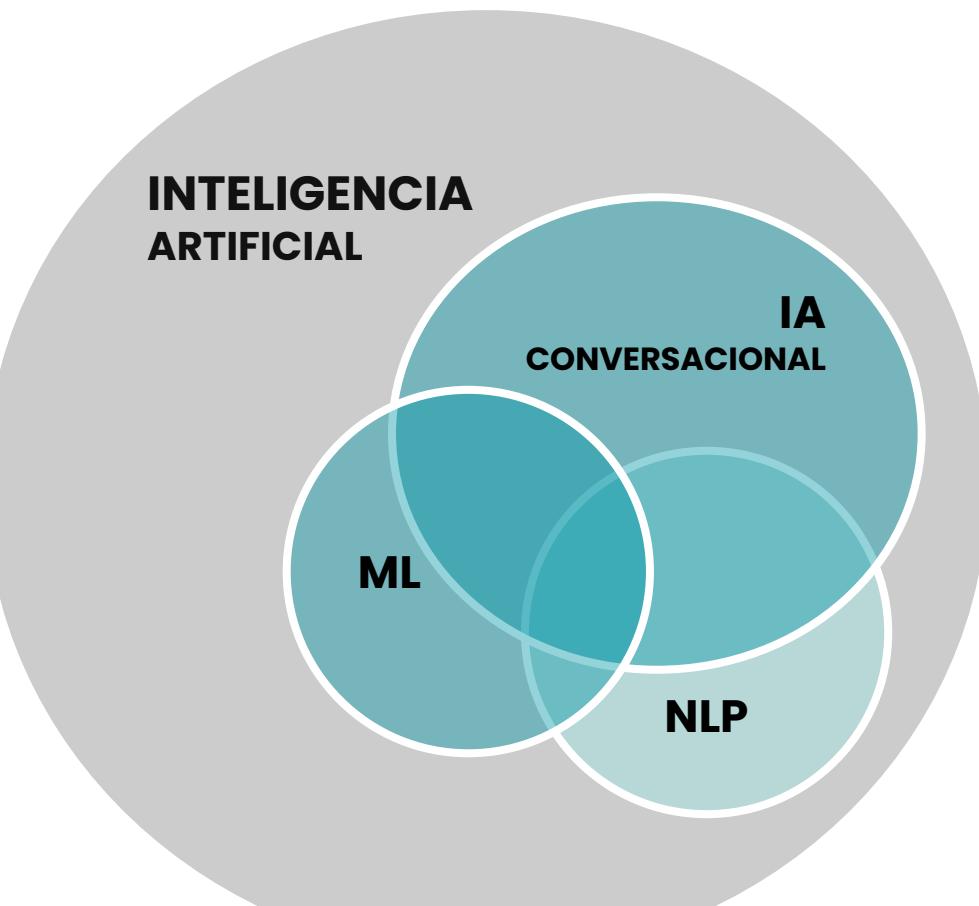
## Teaching Computers to Converse

Conversations can be conceptually and emotionally complex; they entail much more than simple input and output. When we talk to each other, how we say something also matters as much as *what* we say.

Computers can't grasp these nuances, and that's where voice design comes in. A well-designed VUI is flexible, and takes into account these unwritten rules of conversation. It enables computers to think and talk as we do, and not like a robot.

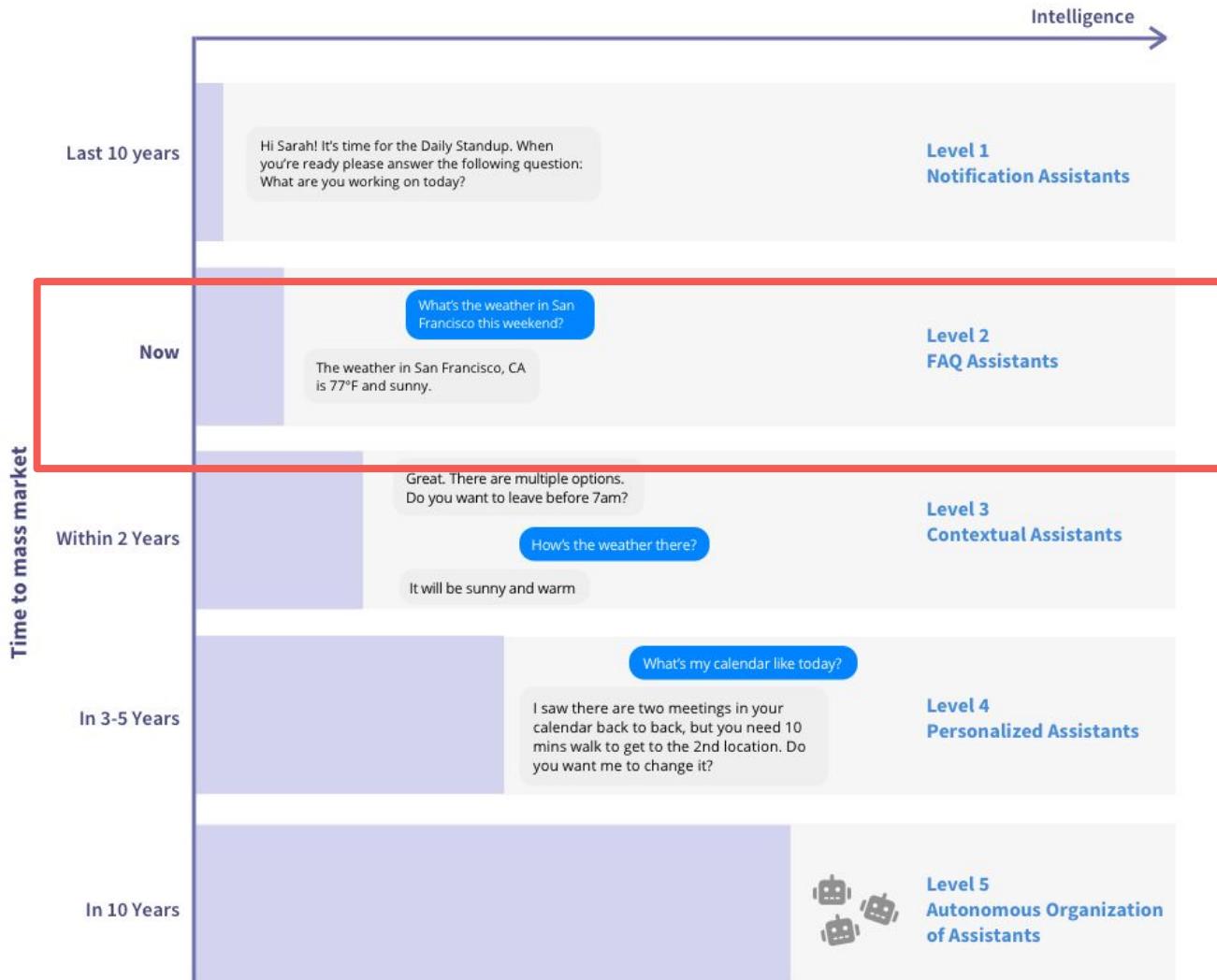
**El objetivo es hacer que las personas interactuemos de manera natural conversando con la tecnología, teniendo en cuenta que las conversaciones pueden ser emocionalmente complejas, además de las reglas no escritas de la comunicación.**

# Inteligencia Artificial + Canal / Modo



# Interfaces conversacionales hoy

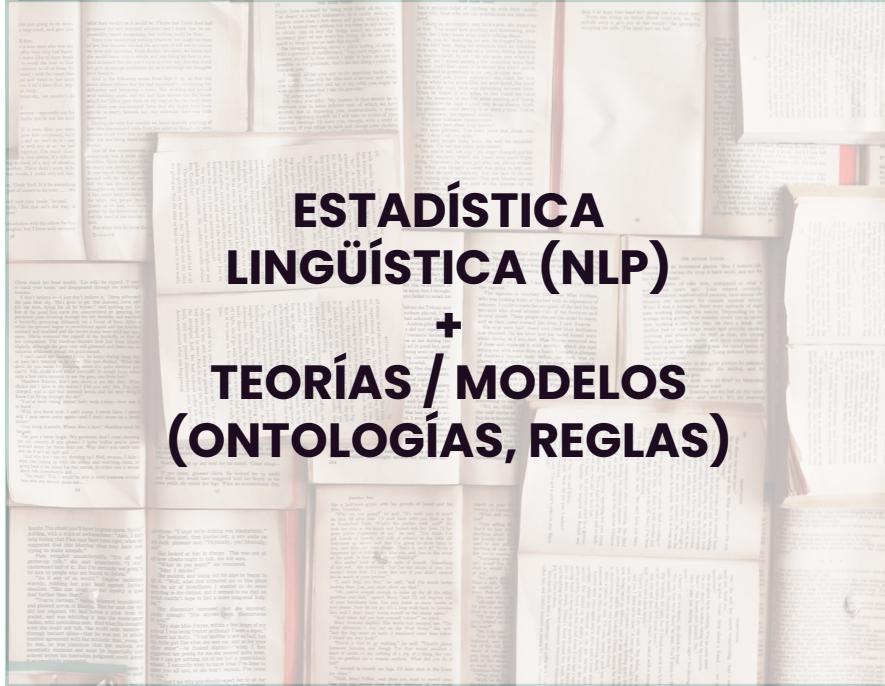
---



**¿Dónde estamos entonces?**  
**Inteligencia vs Tiempo hasta llegar al mercado**

Source: Alan Nichols, August 2018,  
O'Reilly

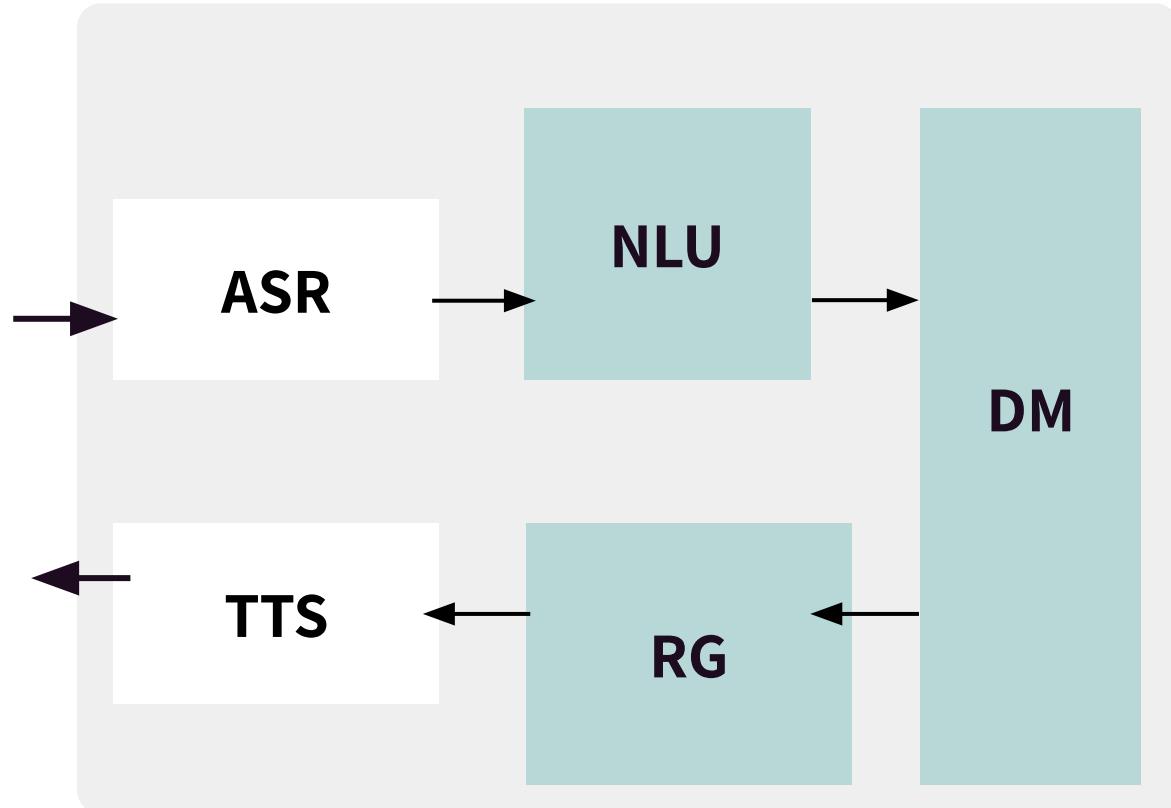
# Tecnologías: ayer y hoy

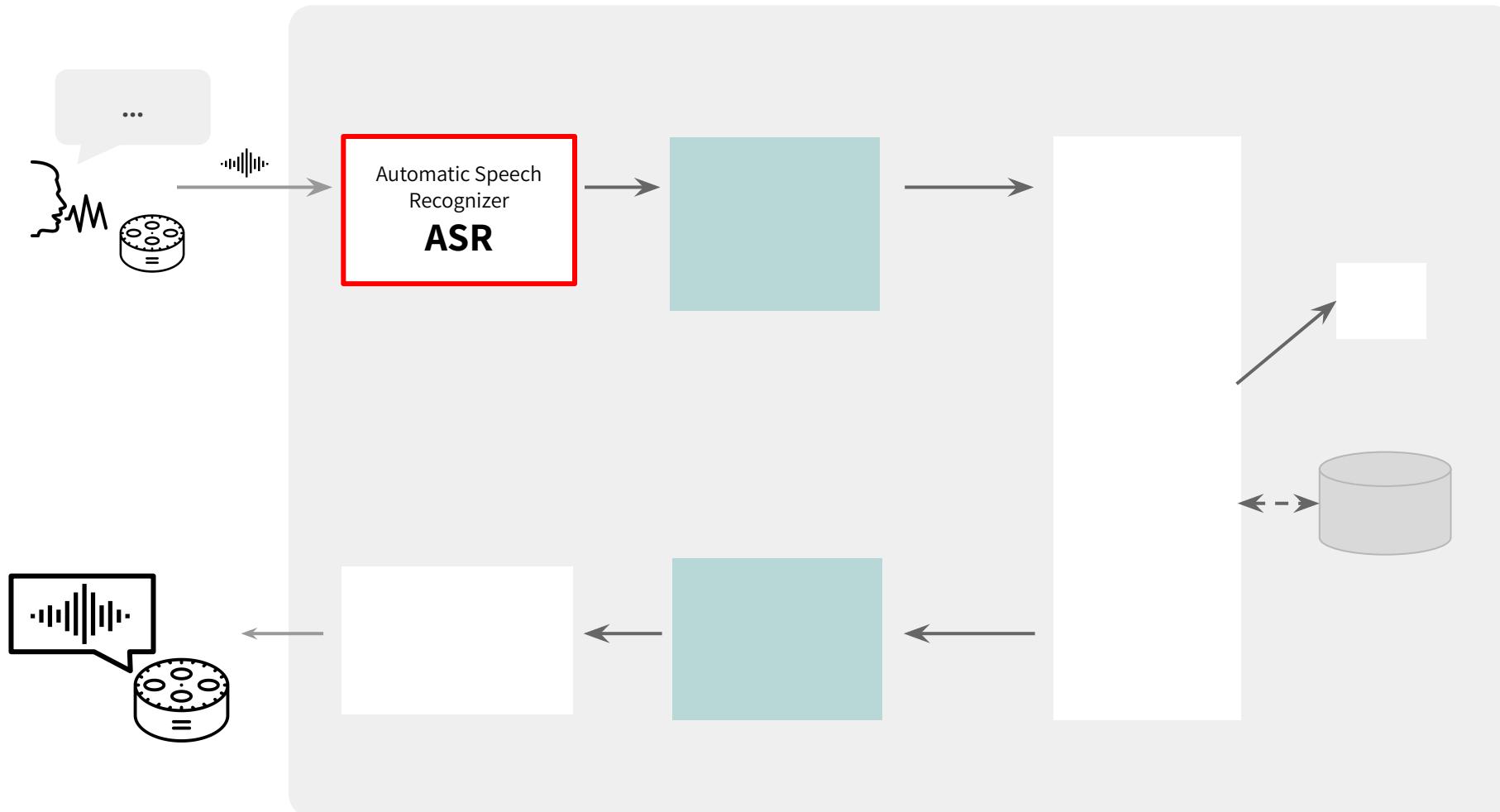


**REDES NEURONALES**  
+  
**CORPUS DE CONVERSACIONES  
(MUCHOS DATOS)**

# ¿Es “inteligente” un asistente de voz?

“Quiero viajar a  
México?”





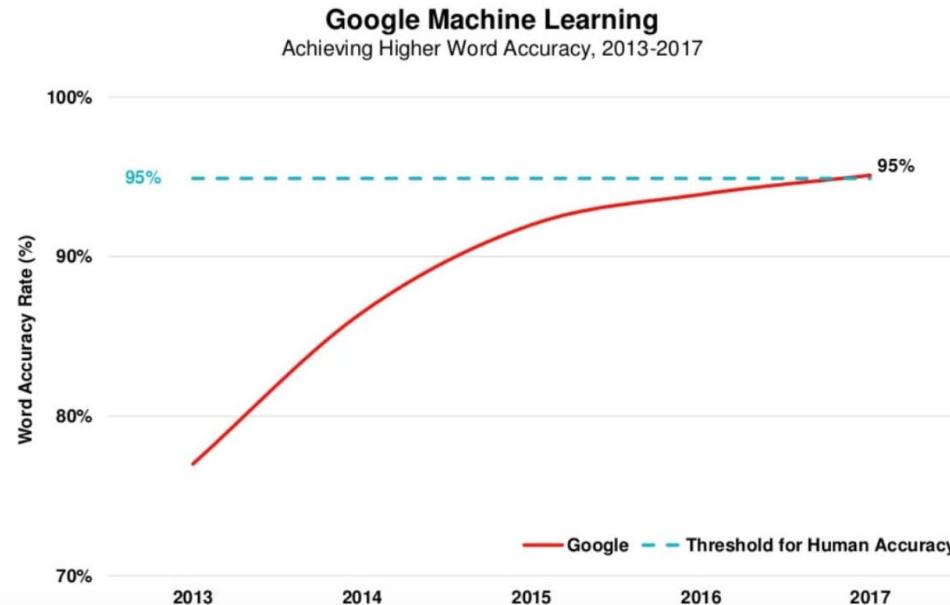
# 2013... ¡ASR en el umbral de la capacidad humana!



La mejora en la capacidad de cómputo (nube) **redujo el tiempo** de procesamiento de **enormes cantidades de datos** (speech) >> mejores modelos de ASR



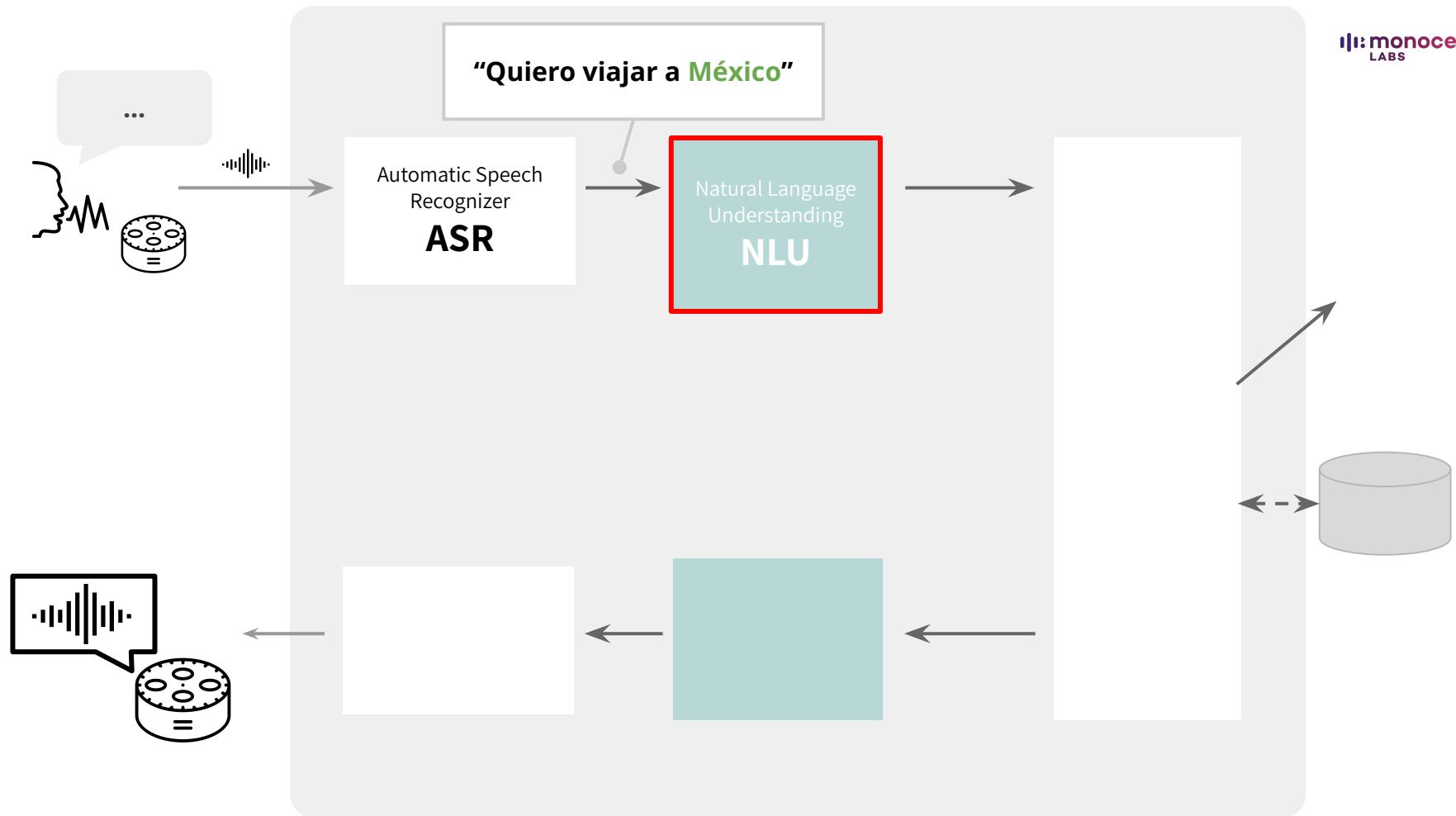
Computing power and artificial intelligence are largely behind the advances in this space. With massive amounts of speech data combined with faster processing, speech recognition has hit an inflection point where its capabilities are roughly on par with humans. The graph below is from [Mary Meeker's 2017 Internet Trends report](#). It plots Google's word accuracy rate which recently broke the 95% threshold for human accuracy.



# El ruido y *The Cocktail Party Problem*

El reto del ASR: tratar de identificar al hablante y separar lo que dice, en entornos de mucho ruido (como una fiesta).





# Arquitectura: NLU (Entrenando el modelo)

frase o *utterance*:

Quiero viajar a **México**

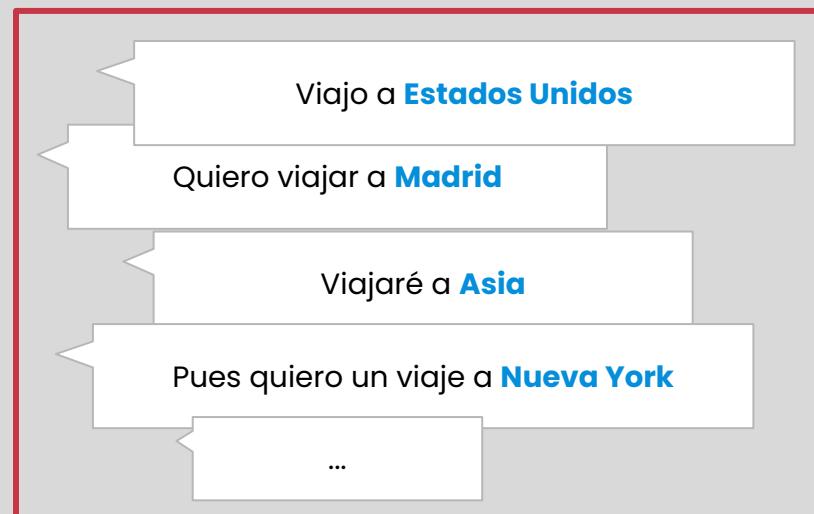
valor  
entidad o slot  
lugar

Clasificamos diferentes frases según la intención del usuario:

- aportamos **semántica**,
- indicamos la información relevante (entidades).

**viajarLugar**

intención del usuario o *intent*



# Arquitectura: NLU (Entrenando el modelo)

frase o *utterance*:

Quiero viajar a {México}

valor  
entidad o  
slot  
ciudad

viajarCiudad

intención del usuario o *intent*

Viajo a Newark

Quiero viajar a Madrid

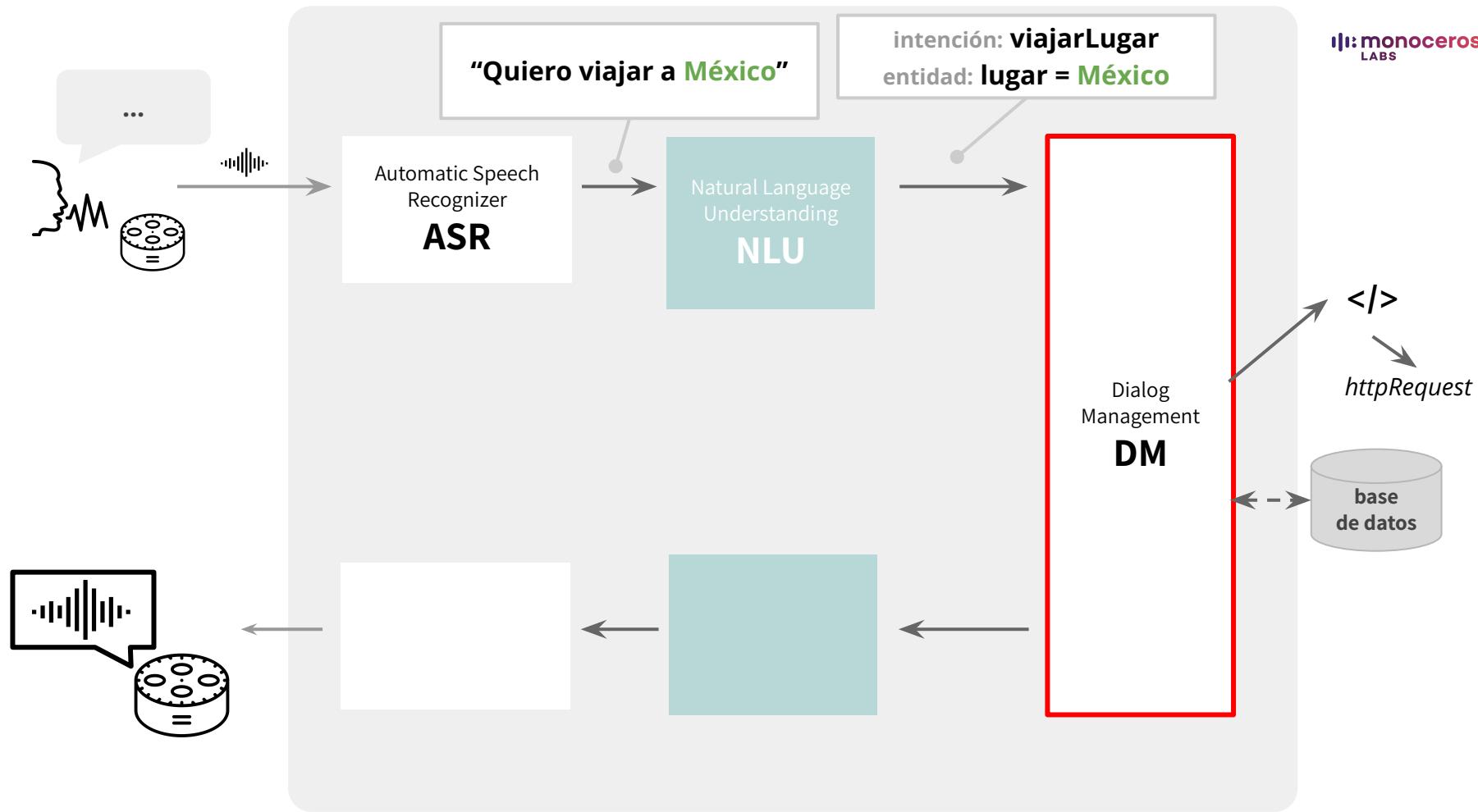
Viajaré a Bangkok

Pues quiero un viaje a Nueva York

...

Clasificamos diferentes frases según la intención del usuario:

- aportamos **semántica**,
- indicamos la información relevante (entidades).



# Arquitectura: DM > Gestión de errores

- Slot filling

Quiero viajar a



viajarLugar

necesita siempre ir acompañado de una serie de entidades: lugar

viajarLugar

lugar = ??



ACCIÓN: Preguntar al Usuario por un LUGAR

# Arquitectura: DM > Gestión de errores

- Slot filling



viajarLugar lugar = ??

¡KO!

ACCIÓN: Preguntar al Usuario por un LUGAR

viajarLugar lugar = México

¡OK!

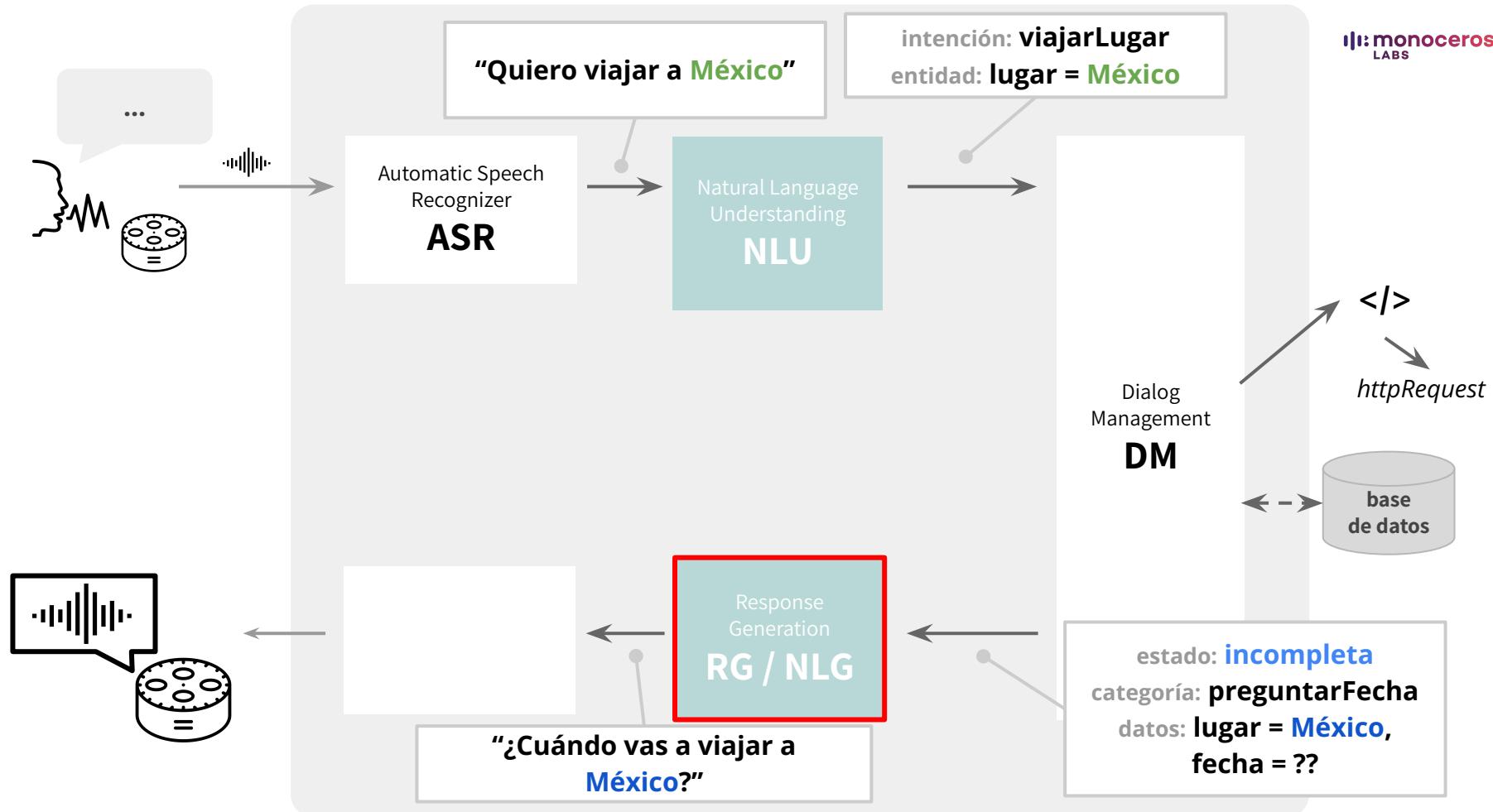
ACCIÓN: Voy a consultar en nuestro sistema fechas disponibles para LUGAR, después pregunta por FECHA

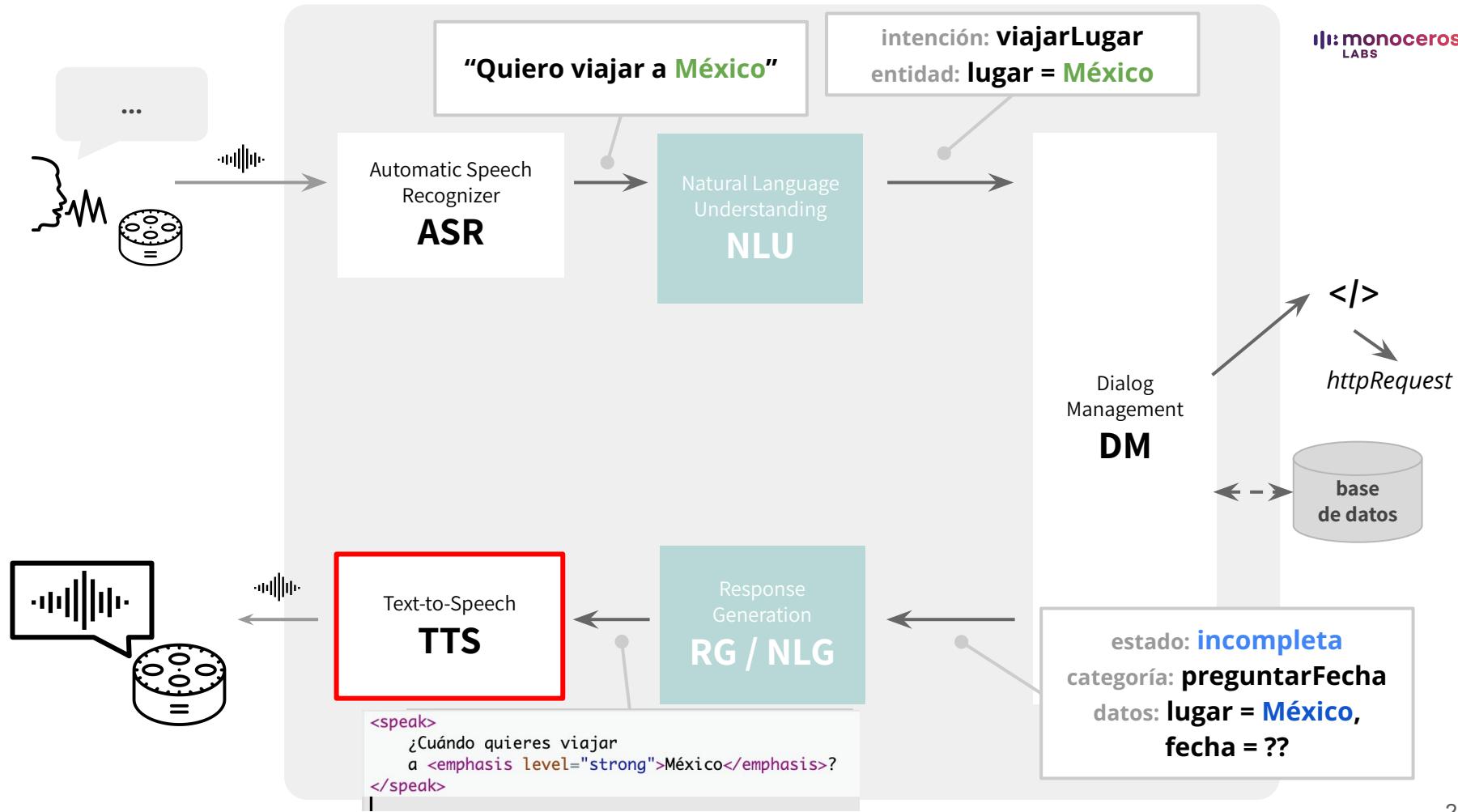
fecha = ...

f(x){...}  
API -> GET  
FECHAS



Nieves,  
{"lugar": "México"}

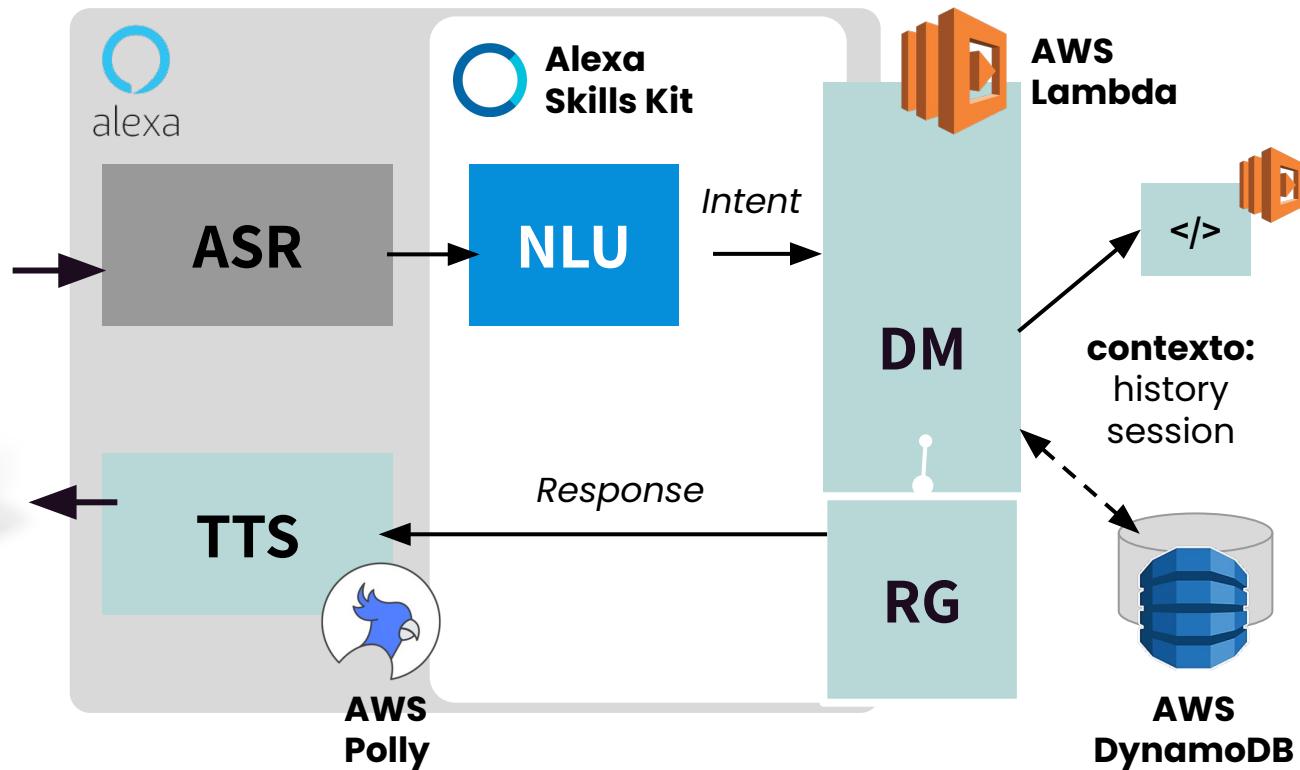




# Tecnologías subyacentes (Amazon Alexa)

Digo:

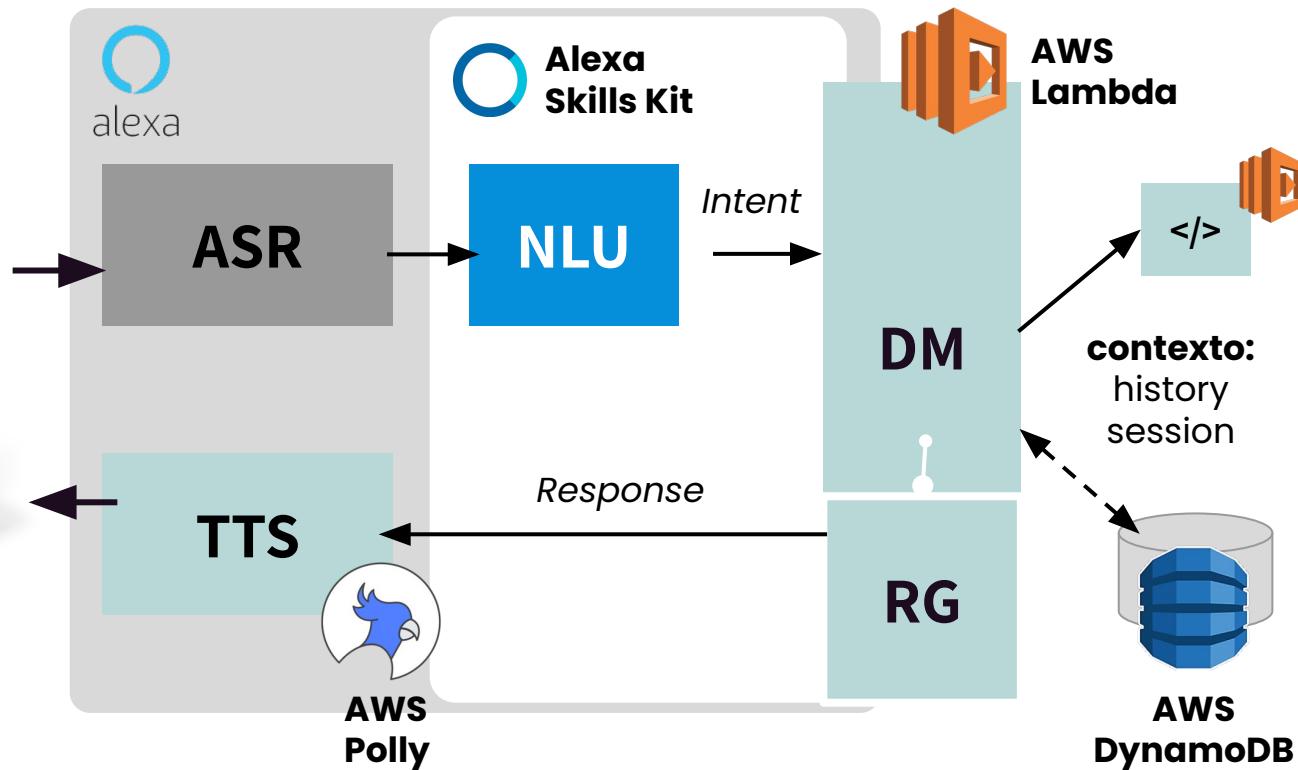
"Alexa,  
abre Veo Veo"



# Tecnologías subyacentes (Amazon Alexa)

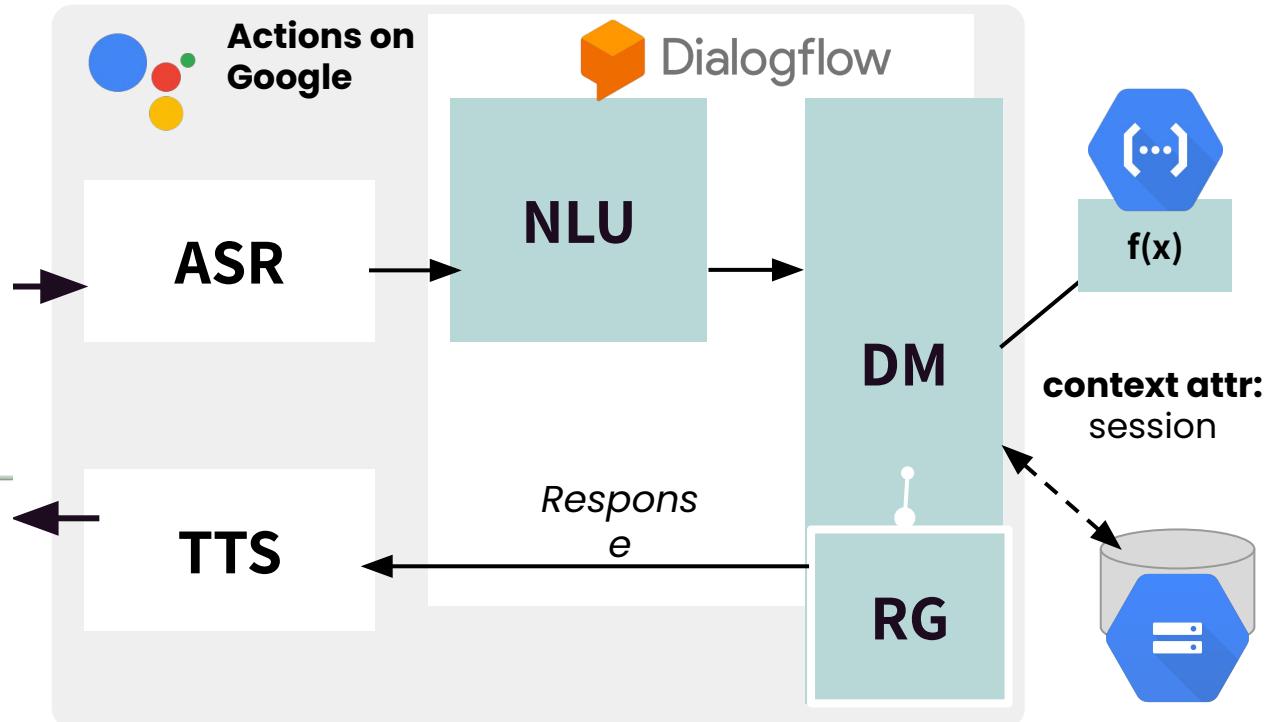
Digo:

"Alexa,  
abre Veo Veo"



# Tecnologías subyacentes (Google Assistant)

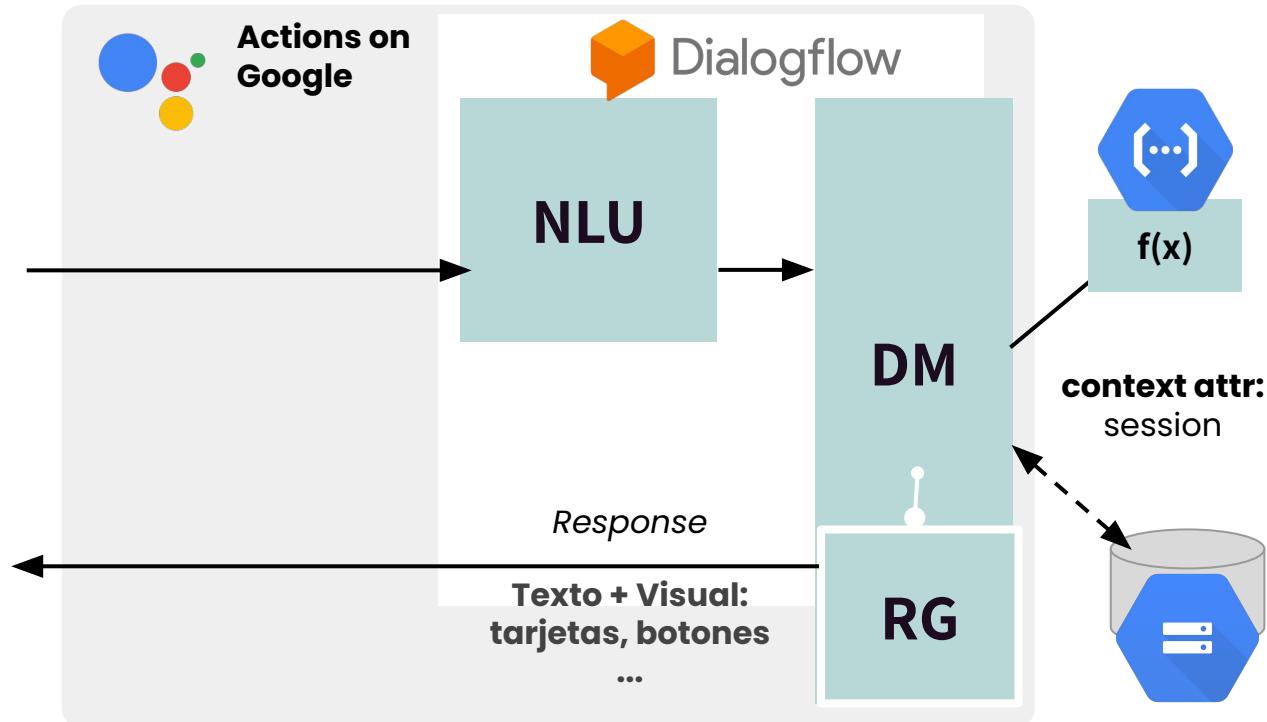
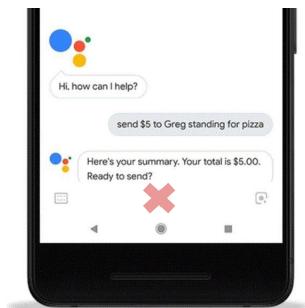
Digo:  
“OK Google,  
abre Juego de  
Tronos”



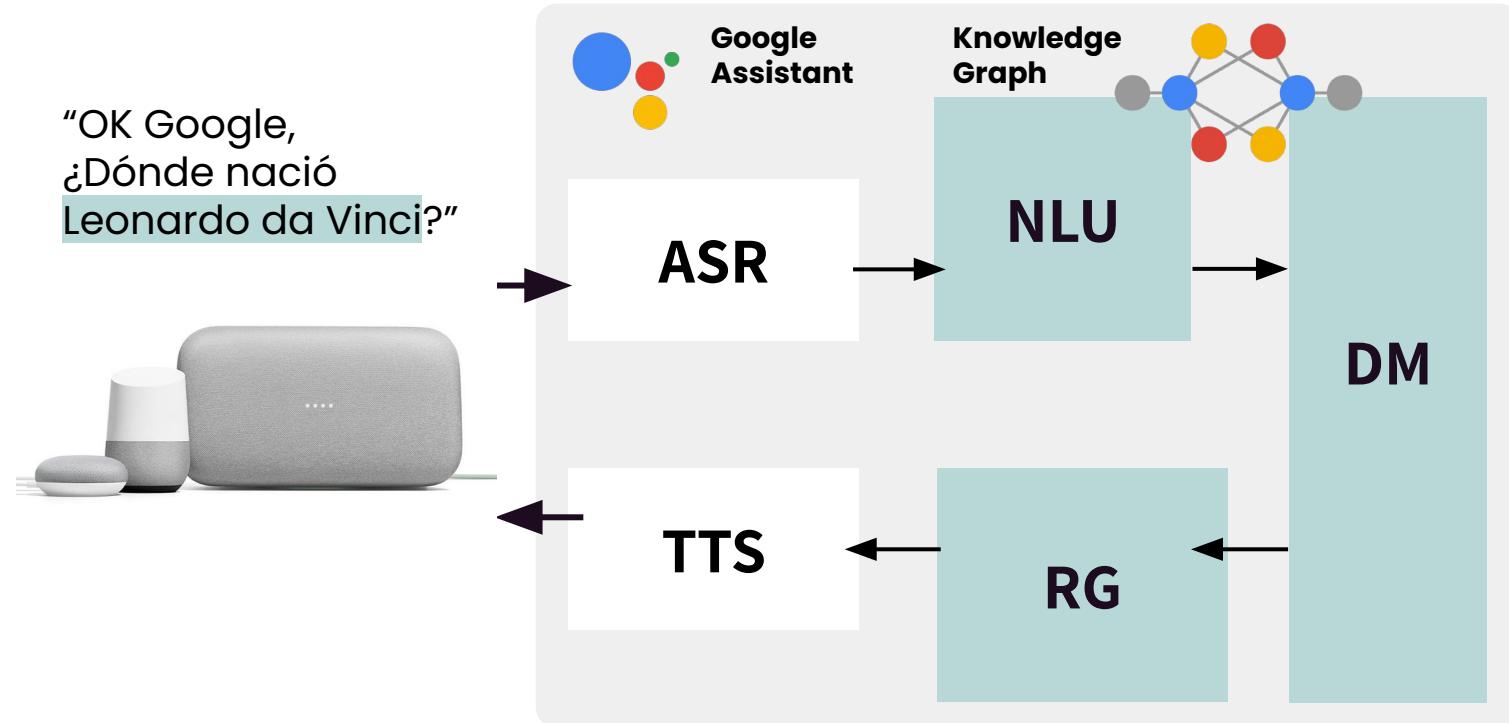
# Tecnologías subyacentes (Google Assistant)

**Escribo:**

"abre Juego de Tronos"



# Tecnologías subyacentes (Google Assistant)



... y mañana?

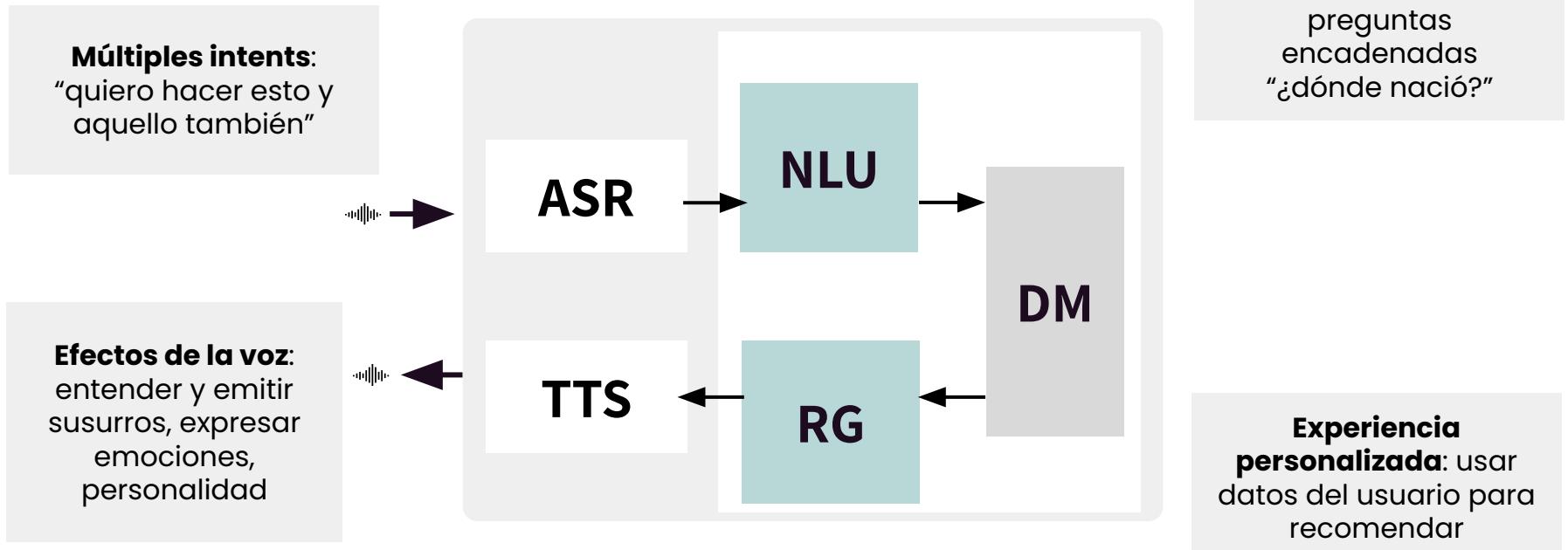
---

# ¿Tecnología en fase temprana?

**Advanced features** (*state-of-the-art*, disponibles en algunos asistentes, o en inglés) que van orientadas a buscar la naturalidad en la conversación.

# ¿Tecnología en fase temprana?

**Advanced features** (*state-of-the-art*, disponibles en algunos asistentes, o en inglés) que van orientadas a buscar la naturalidad en la conversación.



# ¿Tecnología en fase temprana?

**Advanced features** (*state-of-the-art*, disponibles en algunos asistentes, o en inglés) que van orientadas a buscar la naturalidad en la conversación.

The screenshot shows a web browser displaying the Alexa developer documentation. The URL in the address bar is <https://developer.amazon.com/alexa/skill-kit/docs/conversations/about>. The page title is "About Alexa Conversations". The left sidebar has a tree view with "About Alexa Conversations" selected. The main content area starts with a note about supported locales: "(GA) en-US" and "(Beta) en-AU, en-CA, en-IN, en-GB, de-DE, ja-JP, es-ES, es-US". Below this, a paragraph explains what Alexa Conversations is: "Alexa Conversations is a deep learning-based approach to dialog management that enables you to create natural, human-like voice experiences on Alexa. Alexa Conversations helps skills respond to a wide range of phrases and unexpected conversational flows, and gives skills the conversational memory to sustain long, two-way interactions between Alexa and the user." A second paragraph continues: "You can create a skill that uses Alexa Conversations to manage the entire skill experience, or you can extend an existing skill with Alexa Conversations. For example, your skill can use your existing code to handle simple interactions. Then, your skill can delegate dialog management to Alexa Conversations for tasks that involve many two-way conversations with the user."

# ¿Tecnología en fase temprana?

**Advanced features** (*state-of-the-art*, disponibles en algunos asistentes, o en inglés) que van orientadas a buscar la naturalidad en la conversación.

Use New Alexa Emotions and Speaking Styles to Create a More Natural and Intuitive Voice Experience

Catherine Gao Nov 26, 2019

Share: [f](#) [in](#) [t](#)

Game Skills Content Skills Design News



We're excited to introduce two new Alexa capabilities that will help create a more natural and intuitive voice experience for your customers. Starting today, you can enable Alexa to respond with either a happy/excited or a disappointed/empathetic tone in the US. Emotional responses are particularly relevant to skills in the gaming and sports categories. Additionally, you can have Alexa respond in a speaking style that is more suited for a specific type of content, starting with news and music. Speaking styles are curated text-to-speech voices designed to create a more delightful customer experience for

# ¿Tecnología en fase temprana?

**Advanced features** (*state-of-the-art*, disponibles en algunos asistentes, o en inglés) que van orientadas a buscar la naturalidad en la conversación.

TECH \ AMAZON \ AMAZON ALEXA

**All the new features coming to Alexa, including a new voice, frustration mode, and Samuel L. Jackson**

*Also new: bilingual support, and more troubleshooting features like 'frustration detection'*

By Chaim Gartenberg | @cgartenberg | Sep 25, 2019, 1:26pm EDT

f t SHARE



# ¿Tecnología en fase temprana?

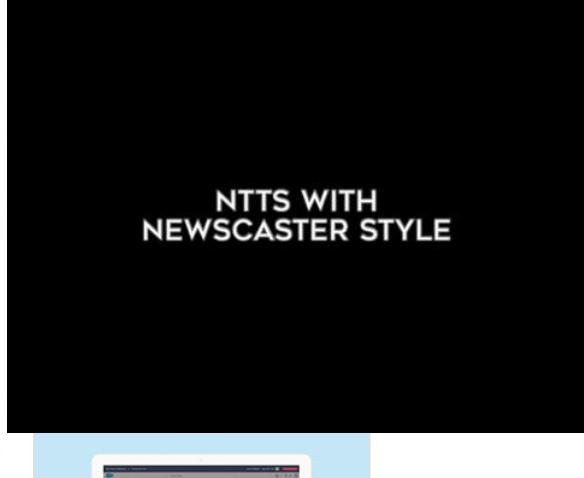
**Advanced features** (*state-of-the-art*, disponibles en algunos asistentes, o en inglés) que van orientadas a buscar la naturalidad en la conversación.

Amazon's neural TTS can model speaking styles with only a few hours of recordings

KYLE WIGGERS @KYLE\_L\_WIGGERS NOVEMBER 19, 2018 6:55 AM



Above: Amazon Echo Sub



# Y lo que vendrá

**Advanced features** (*state-of-the-art*, aún no disponibles en asistentes).

The screenshot shows a web page from Amazon Science. At the top, there's a navigation bar with links for Research areas, Blog, News and features, Publications, Conferences, Collaborations, Careers, Feedback, and a search icon. Below the navigation is a large, abstract geometric diagram composed of overlapping triangles forming a hexagonal pattern. To the right of the diagram, the word "PUBLICATION" is written in capital letters. Below it is the title of a paper: "Contrastive unsupervised learning for speech emotion recognition". Underneath the title, it says "By Mao Li, Bo Yang, Joshua Levy, Andreas Stolcke, Viktor Rozgic, Spyros Matsoukas, Constantinos Papayiannis, Daniel Bone, Chao Wang 2021". A blue button labeled "Download" with a downward arrow is visible. At the bottom left, there's a section titled "AUTHORS" with a list of names and their titles.

**AUTHORS**

- Mao Li
- Bo Yang
- Joshua Levy
- Andreas Stolcke**  
Senior Principal Scientist
- Viktor Rozgic  
Senior applied scientist
- Spyros Matsoukas
- Constantinos Papayiannis
- Daniel Bone
- Chao Wang

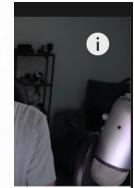
## Abstract

Speech emotion recognition (SER) is a key technology to enable more natural human-machine communication. However, SER has long suffered from a lack of public large-scale labeled datasets. To circumvent this problem, we investigate how unsupervised representation learning on unlabeled datasets can benefit SER. We show that the contrastive predictive coding (CPC) method can learn salient representations from unlabeled datasets, which improves emotion recognition performance. In our experiments, this method achieved state-of-the-art concordance correlation coefficient (CCC) performance for all emotion primitives (activation, valence, and dominance) on IEMOCAP. Additionally, on the MSPodcast dataset, our method obtained considerable performance improvements compared to baselines.

# Y lo que vendrá (... y sus retos)

to overc  
from da  
most co  
explicit

The fiel  
time. Bi  
experie  
Amazo  
convers  
remain:



*Gracias por vuestra atención  
¿Preguntas?*

**Nieves Ábalos (@nieves\_as)**

Founder & Chief Product Officer, Monoceros Labs