Daniel Fernández Castro / March 2019

DIGITAL MINDS MASTERS / BARREIRA

VOICE INTERFACE DESIGN

PART 0:

WHO ARE YOU AND WHAT DO YOU WANT FROM ME

COURSE INTRODUCTION

WHO ARE YOU AGAIN?

ESPAÑA DOCUMENTO NACIONAL DE IDENTIDAD

APELLIDOS
FERNÁNDEZ
CASTRO

NOMBRE
DANIEL

SEXO M

NACIONALIDAD ESP

FECHA DE NACIMIENTO
04 05 1989

NUM SOPORT [redacted]

VALIDEZ [redacted]

[redacted]

DNI [redacted]

[redacted]

This guy, apparently.

COURSE INTRODUCTION

WHO ARE YOU AGAIN?

- Studied Business Management
    - **Universidad de Valencia** (Spain)
    - **Hochschule Bremen** (Germany)

- Work in tech since the very beginning
    - Sales @ **Microsoft** (Germany, 2011)

    - **Amazon**
        - Site Merchandising Movies & TV (Spain, 2012-14)
        - Site Merchandising/Associate Vendor Manager Music (Spain, 2014-16)
        - Program Manager Alexa (UK, 2016-18)

    - Product Manager/Client Solutions @ **Snips** (France, 2018-)

- Interested in AI & The Great Automation

COURSE INTRODUCTION

WHO ARE YOU AGAIN?

COURSE INTRODUCTION

WHO ARE YOU AGAIN?

www.linkedin.com/in/fernandezcastrodaniel

www.facebook.com/fernandezcastrodaniel

Instagram: @dfercastro

WHAT DO YOU WANT FROM ME?
LEARNING GOALS

Intro: What is AI, and what isn't

Core topic: **Voice User Interfaces**.

- What they are
- How many people use them and what for
- How they work
- What makes them special
- How to design them, and what to optimise for
- The future of voice

WHAT DO YOU WANT FROM ME?
POSSIBLE APPLICATIONS

- Designing VUIs within companies who own the interfaces
- Developing skills/actions/etc for brands that work on those VUI frameworks
- Developing VUIs that are independent of these platforms, based on some of the tech developed by them or other smaller companies. E.g. Mary Poppins ‘talkable’ ad.

COURSE INTRODUCTION

RECOMMENDED READING

Articles and sources:

- https://www.recode.net/2018/11/12/17765390/voice-alexa-siri-assistant-amazon-echo-google-assistant
- Voicebot.ai
- https://medium.com/screenmedia-lab/utterances-slots-and-skills-the-new-vocabulary-needed-to-develop-for-voice-7428bff4ed79
- https://www.theguardian.com/technology/2019/mar/26/smart-talking-are-our-devices-threatening-our-privacy

Books:

- Designing Voice User Interfaces: Principles of Conversational Experiences (Cathy Pearl, O'Reilly, 2016)
- Talk To Me: How Voice Computing Will Transform the Way We Live, Work and Think (James Vlahos, hRandom House Penguin, 2019)

PART 1:
THE (IN)FAMOUS AI

ARTIFICIAL INTELLIGENCE

THE (IN)FAMOUS AI

DEFINITION(S)

Definition 1

The field of computer science dedicated to **solving cognitive problems** commonly associated with human intelligence, such as **learning**, **problem solving**, and **pattern recognition**.

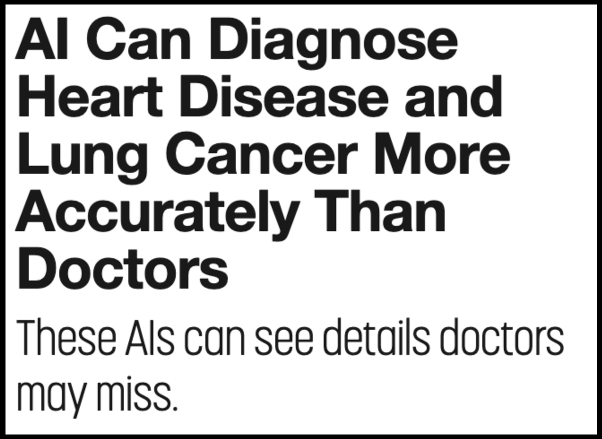
Definition 2

The capability of a machine to imitate intelligent human behaviour *(whatever the means)*

THE (IN)FAMOUS AI

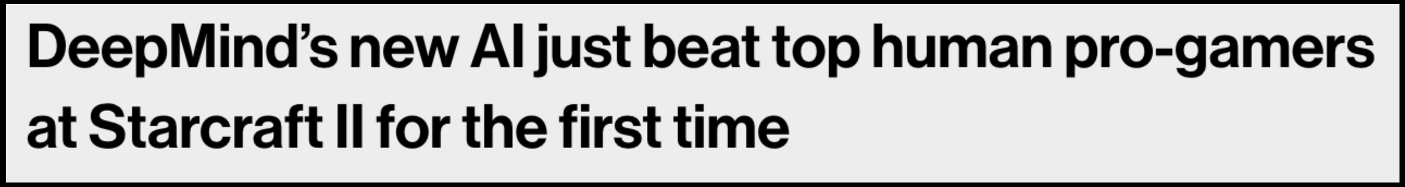
DEFINITION(S)

Definition 1 (field of computer science)

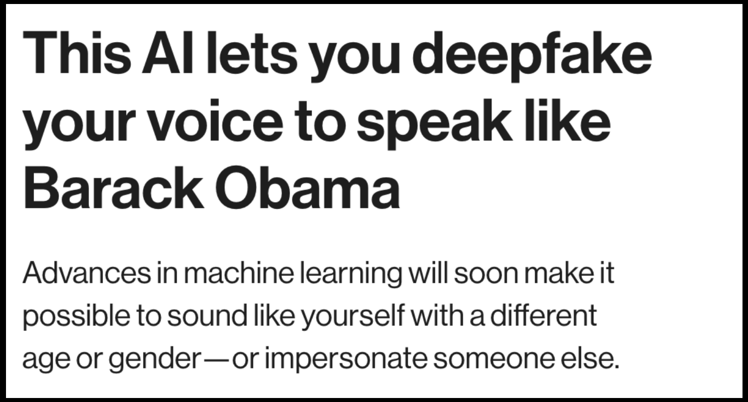


AI Can Diagnose Heart Disease and Lung Cancer More Accurately Than Doctors

These AIs can see details doctors may miss.

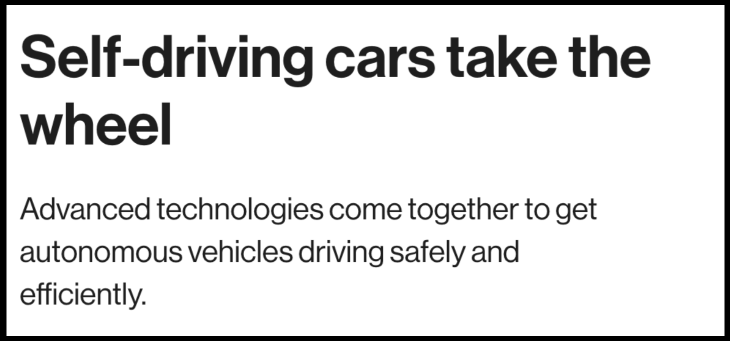


DeepMind's new AI just beat top human pro-gamers at Starcraft II for the first time



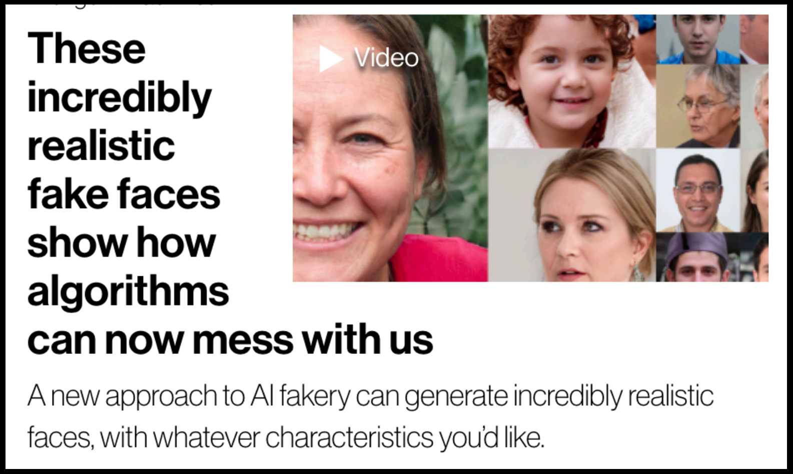
This AI lets you deepfake your voice to speak like Barack Obama

Advances in machine learning will soon make it possible to sound like yourself with a different age or gender—or impersonate someone else.



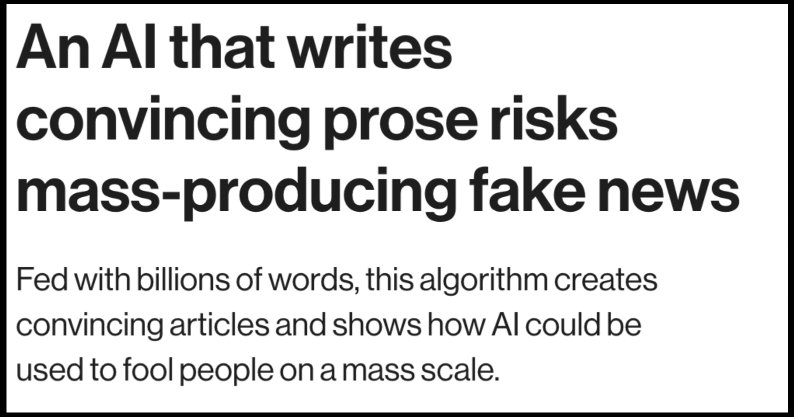
Self-driving cars take the wheel

Advanced technologies come together to get autonomous vehicles driving safely and efficiently.



These incredibly realistic fake faces show how algorithms can now mess with us

A new approach to AI fakery can generate incredibly realistic faces, with whatever characteristics you'd like.

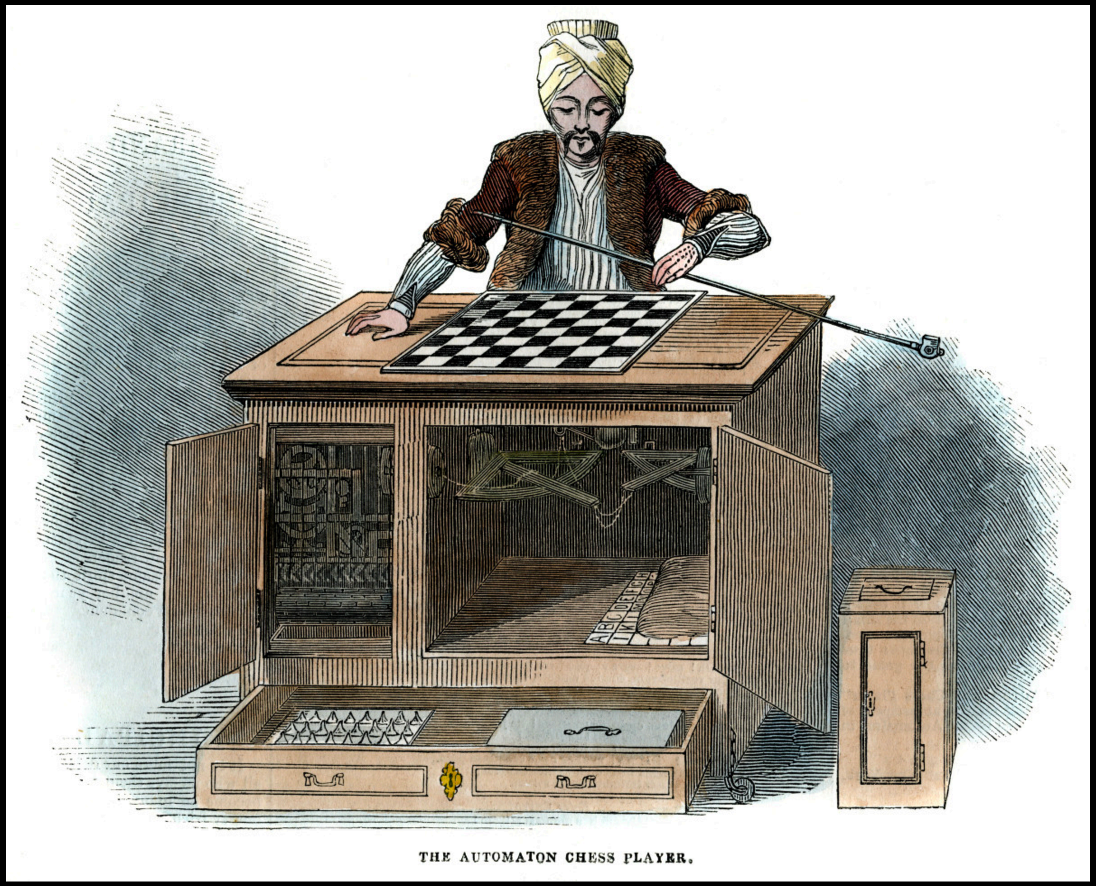


An AI that writes convincing prose risks mass-producing fake news

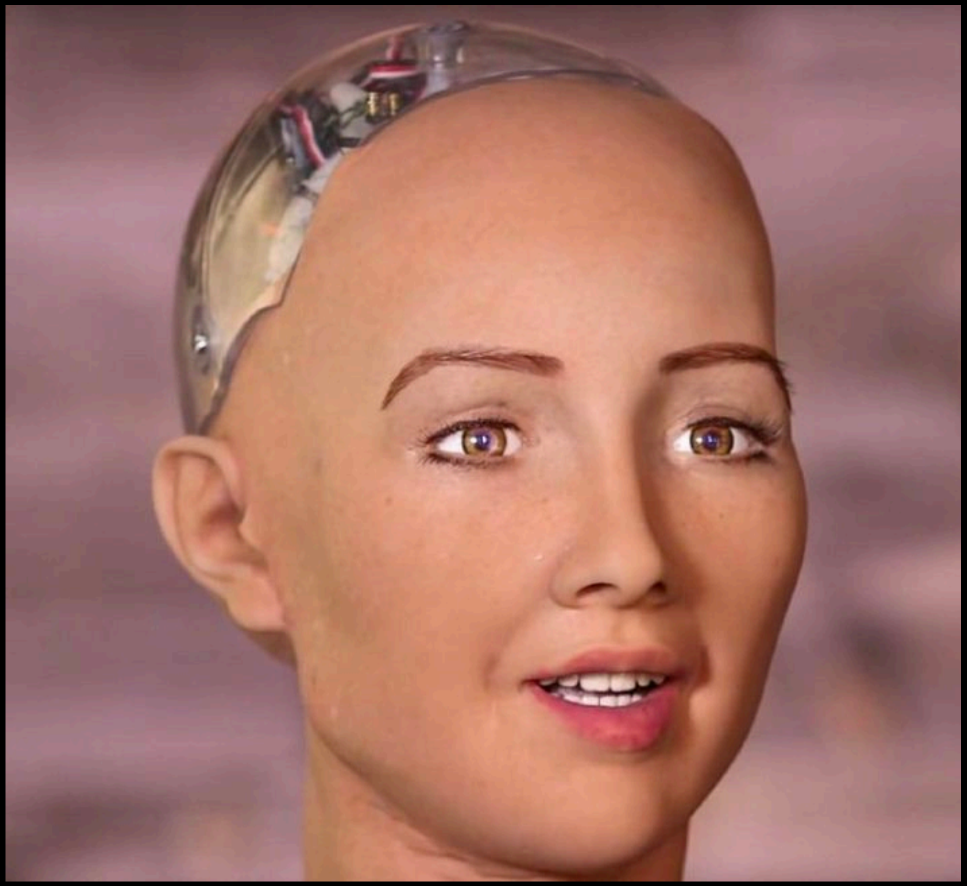
Fed with billions of words, this algorithm creates convincing articles and shows how AI could be used to fool people on a mass scale.

DEFINITION(S)

Definition 2 (machines imitating intelligence)



This one is a bit of a fraud.



This one too. A bit less, though.

AI: DEFINITION(S)

✓

Definition 1

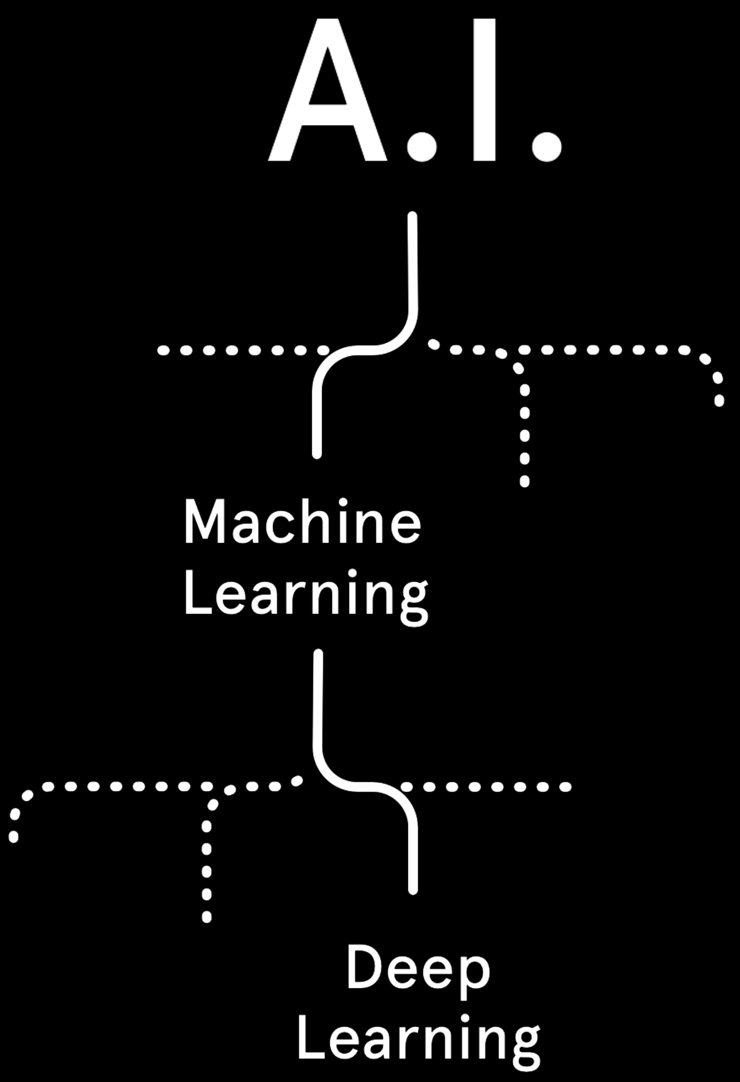
The field of computer science dedicated to **solving cognitive problems** commonly associated with human intelligence, such as **learning**, **problem solving**, and **pattern recognition**.

✗

Definition 2

The capability of a machine to imitate intelligent human behaviour *(whatever the means)*

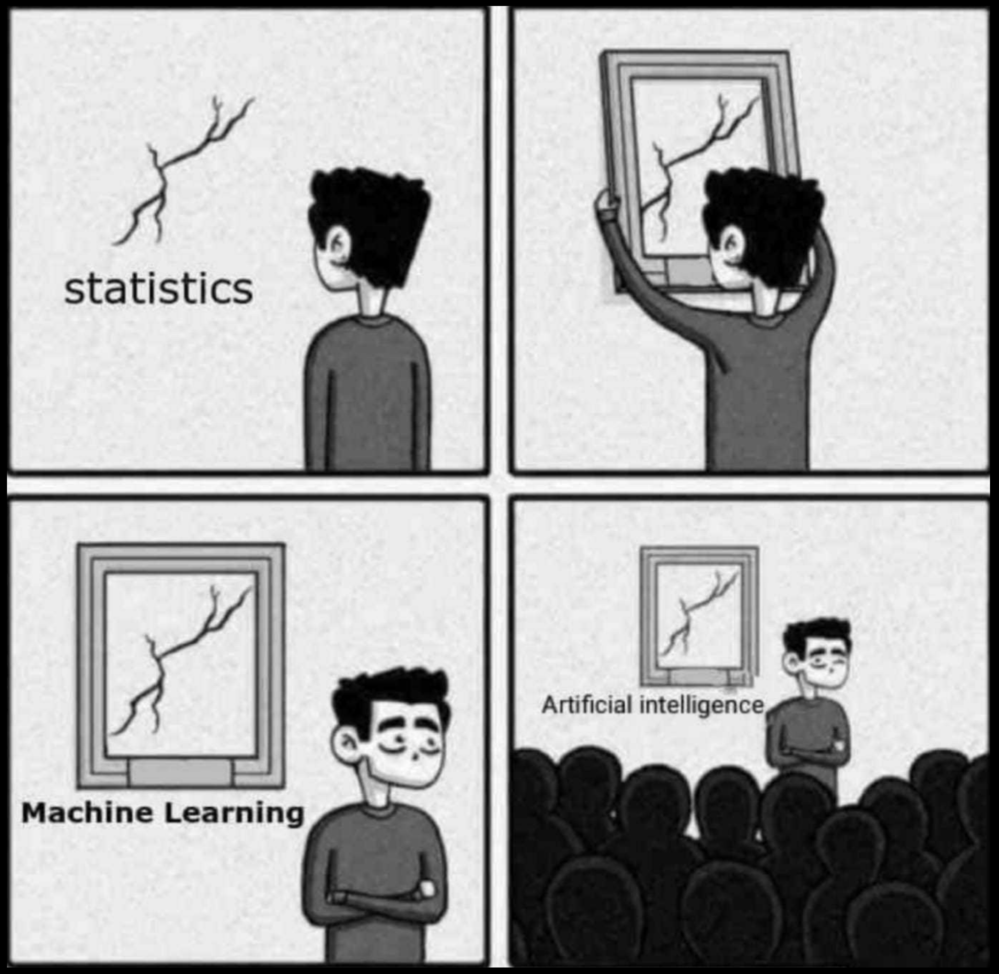
AI: DEFINITION(S)



A.I.

Machine Learning

Deep Learning



TODAY: WEAK AI

Weak AI:

Mastering **Specific Tasks** at superhuman level

- Driving a car
- Picking music
- Understanding human language (kind of)
- Diagnosing cancer
- …

TOMORROW*: STRONG AI / AGI

The intelligence of a machine that could successfully perform **any** intellectual task that a human being can.

How will we know? Wozniak Test, Turing Test…

**in 70-100 years, maybe. MAYBE.*

MACHINE LEARNING

WHAT IS MACHINE LEARNING?

Ability for an algorithm to **learn from prior data**

to produce a **behaviour in an unforeseen situation**,

without a human pre-programming all scenarios and decisions.

WHAT IS MACHINE LEARNING?

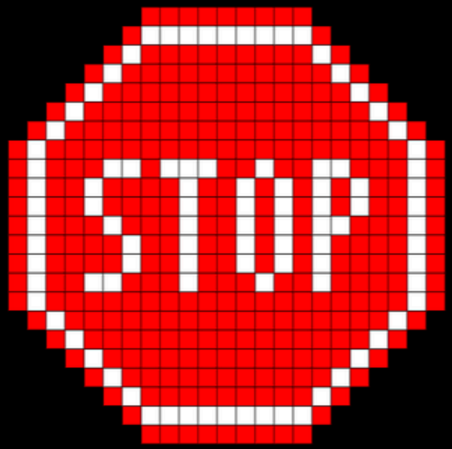
Example: image recognition

Before:

“If there is a red pixel and that red pixel has a white pixel above it,

and […], and […], and […], and […], and […],

…then it’s a STOP sign”.



THE (IN)FAMOUS AI

WHAT IS MACHINE LEARNING?

After:

“Here, machine:

1. take a look at this 2 million images of STOP signs from all angles, distances and with different lighting levels,
2. plus some others that DON’T contain any (flagged as such) and
3. tell me if this other image has one STOP sign in it”.

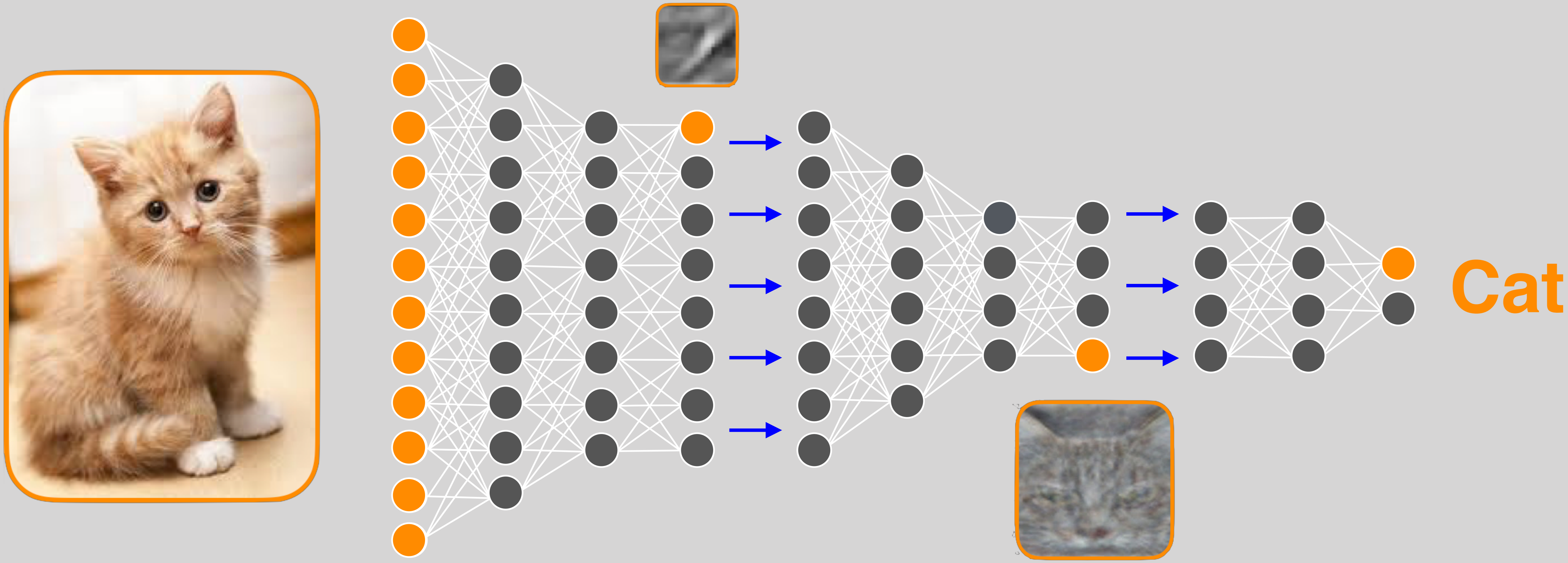
DEEP LEARNING

WHAT IS DEEP LEARNING?

Deep learning is a branch of machine learning.

- Artificial neural networks — *inspired by neurones* — find patterns in raw data by combining multiple layers of artificial neurones.
- As the layers increase (more depth), so does the neural network's ability to learn increasingly abstract concepts.

DEEP LEARNING



WHY NOW?

WHY NOW?

- We finally have the **computational power** to do this cheaply.
- Thanks to the internet, there's a ridiculous **amount of data*, most of it correctly labeled,** available for algorithms to train.
- That's why financial markets got the benefits of AI first. Data had been available electronically for decades.

*Some sources of data useful for AIs:

- You, choosing a song on Spotify
- You, liking a page on Facebook
- You, watching a movie on Netflix

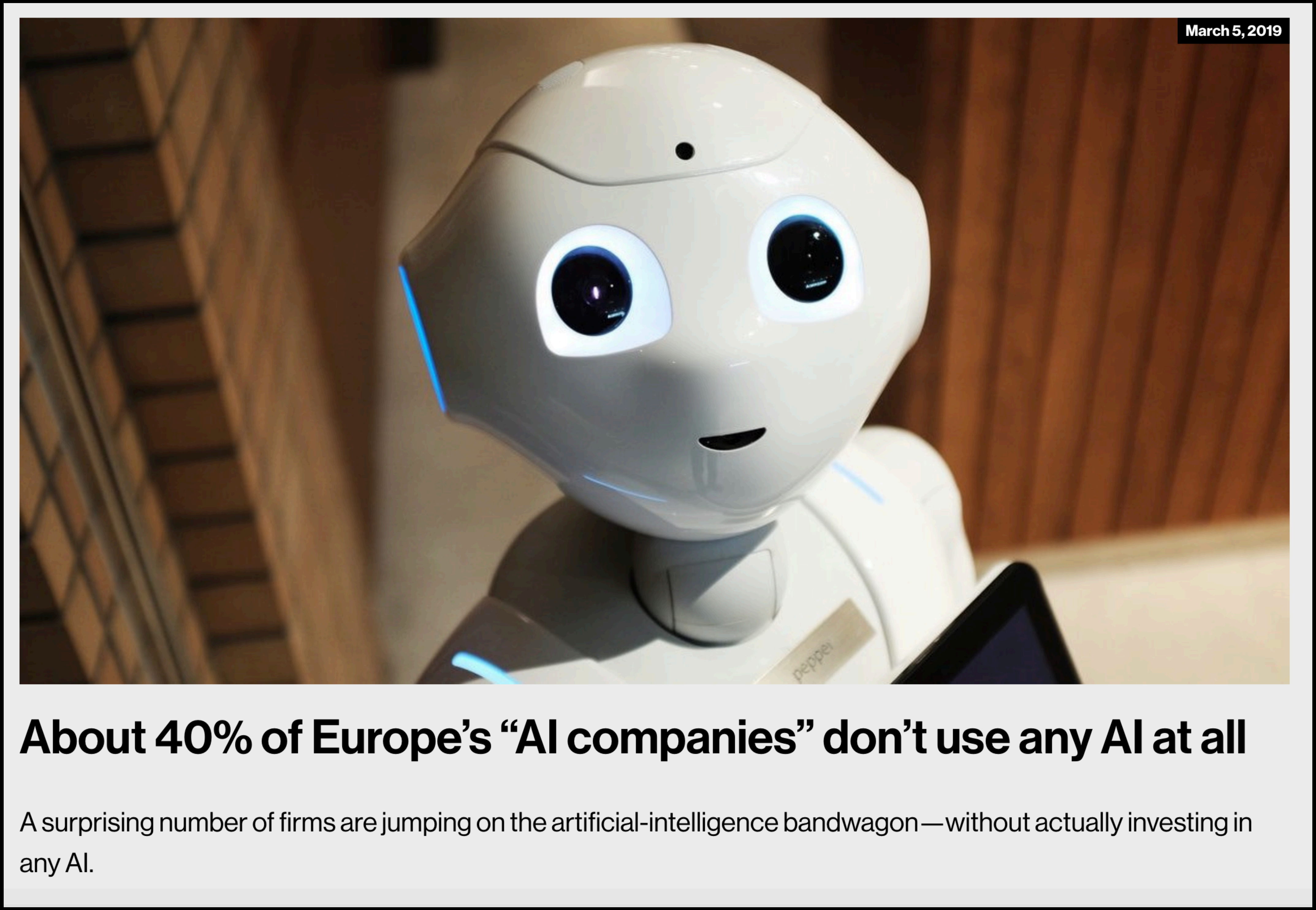
WHERE DO VOICE INTERFACES COME IN THIS?

We would have never been able to understand natural human language* without machine learning.

* Natural language = being able to speak like you would to a human, i.e. saying "will I need an umbrella tonight?" to ask for the weather.

THE (IN)FAMOUS AI

IS THIS AI?



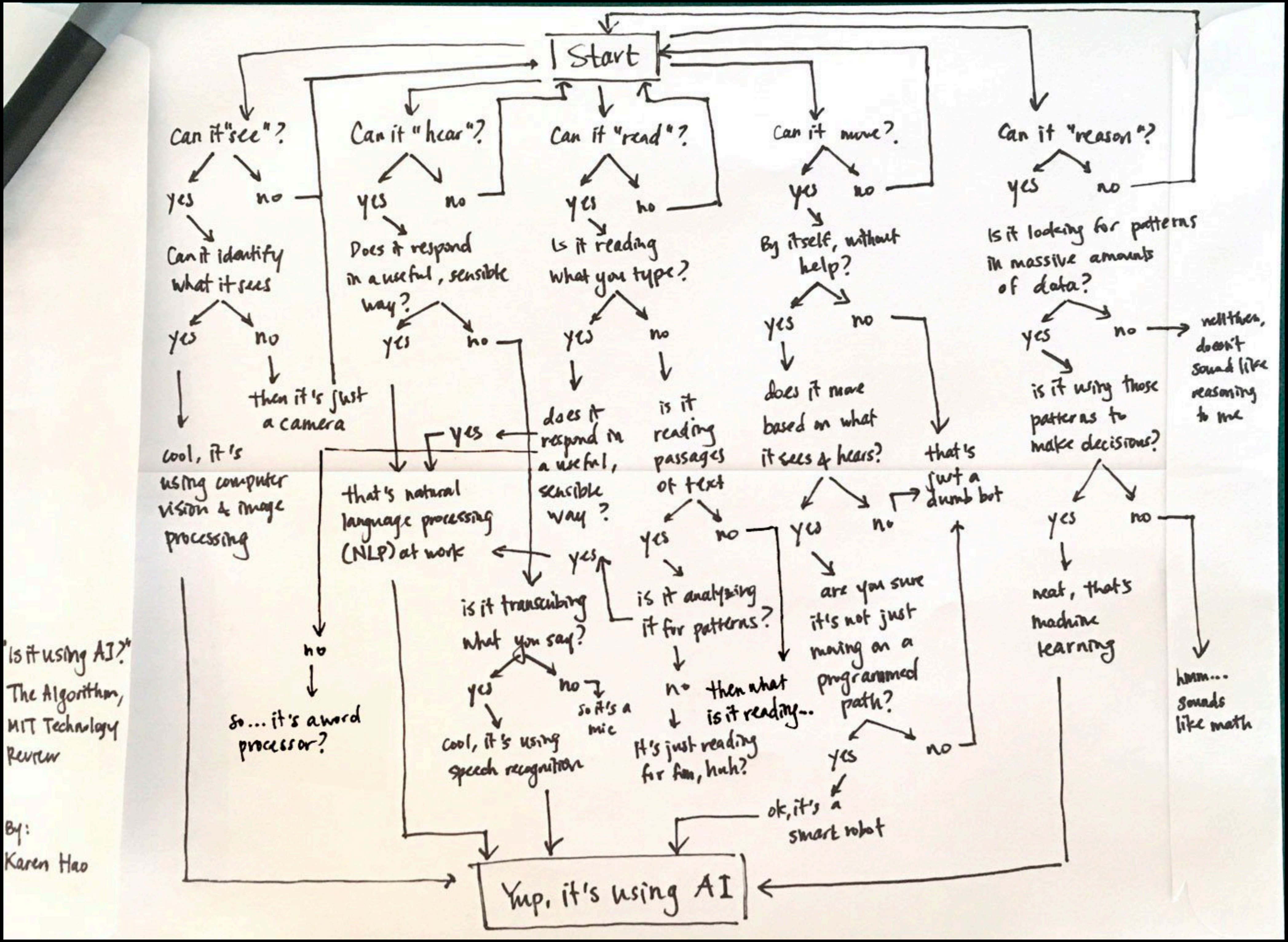
March 5, 2019

About 40% of Europe’s “AI companies” don’t use any AI at all

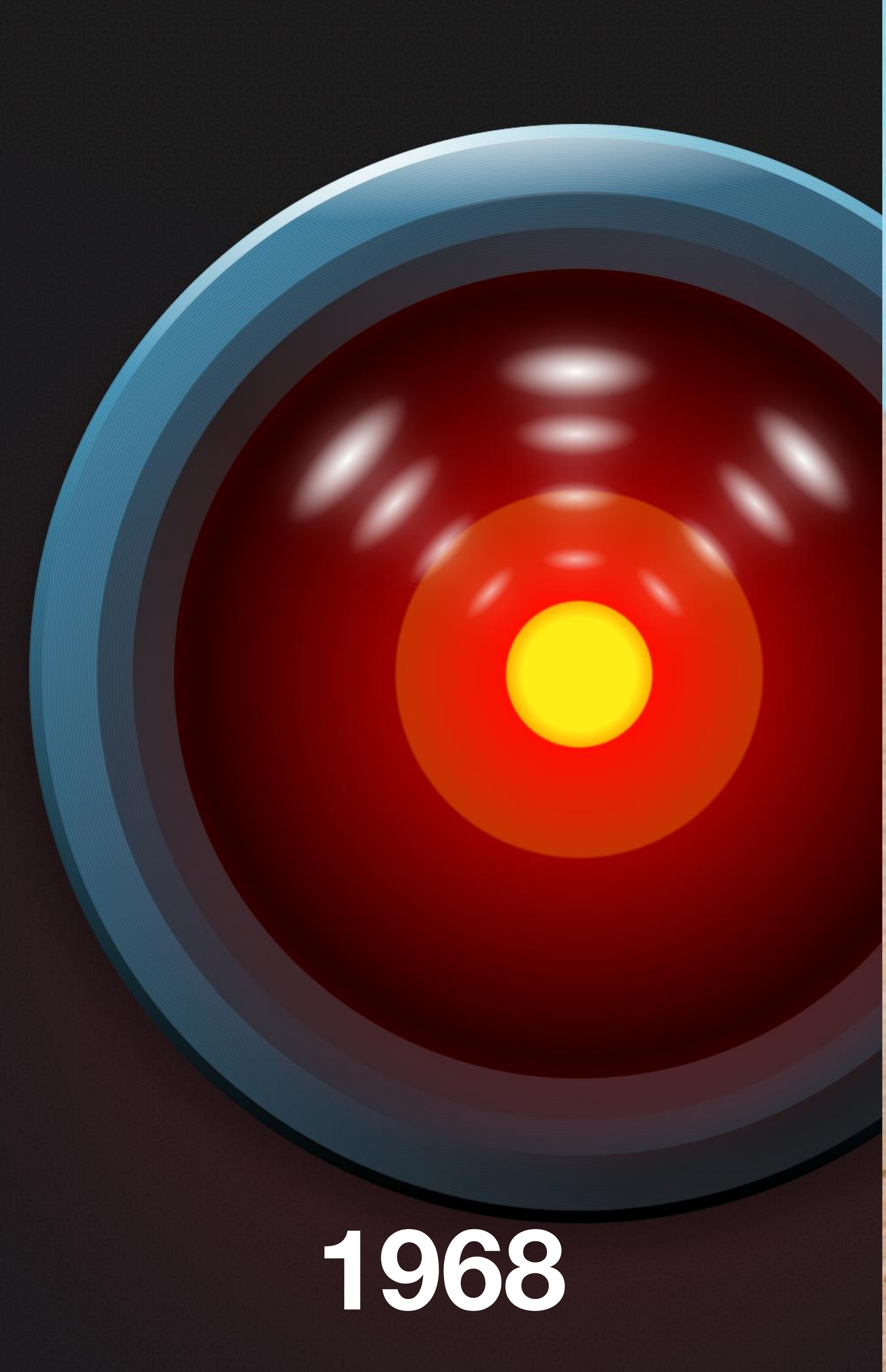
A surprising number of firms are jumping on the artificial-intelligence bandwagon—without actually investing in any AI.

THE (IN)FAMOUS AI

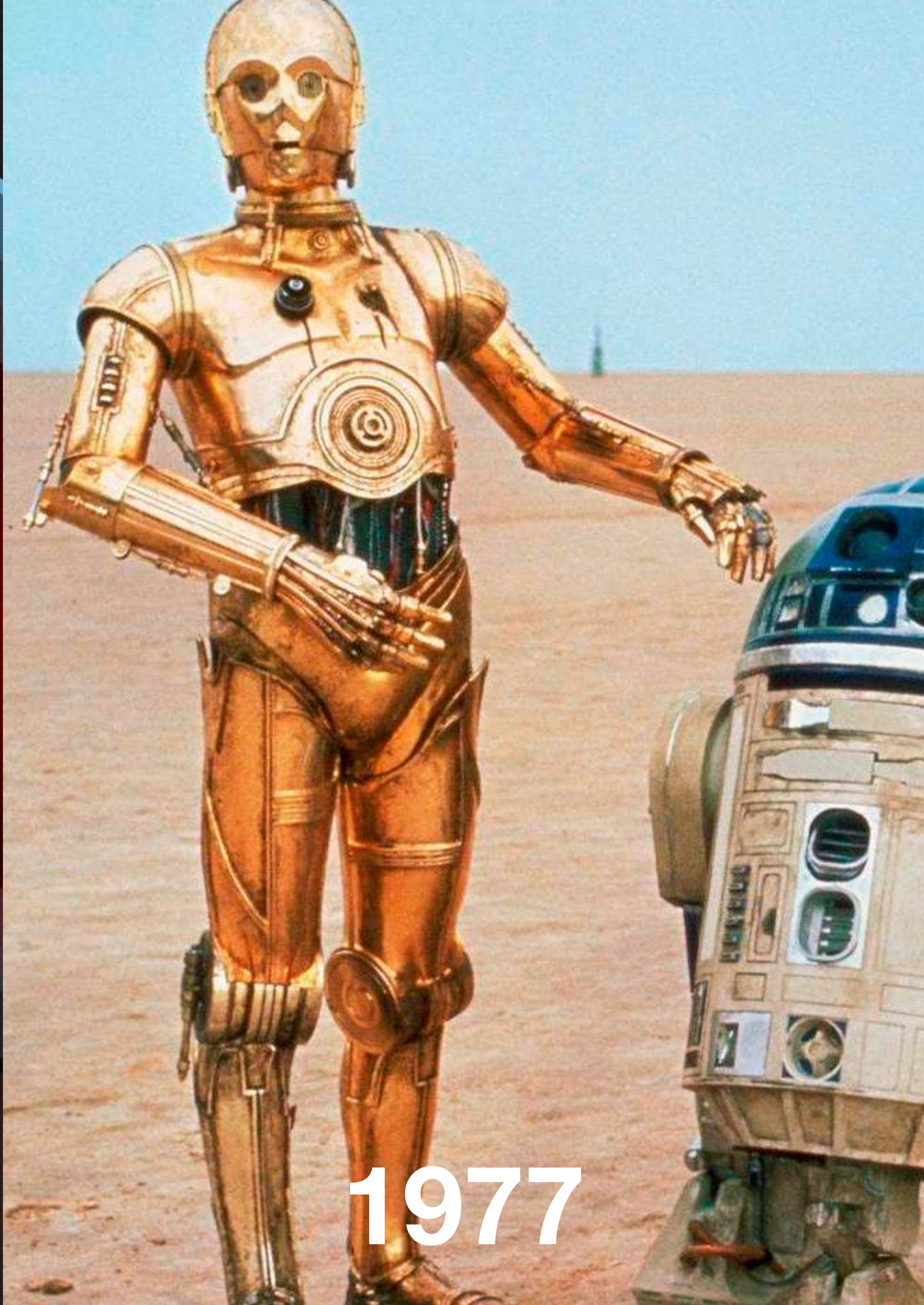
IS THIS AI?


Start
Can it "see"?
yes
no
Can it identify what it sees
yes
no
then it's just a camera
cool, it's using computer vision & image processing
Can it "hear"?
yes
no
Does it respond in a useful, sensible way?
yes
no
that's natural language processing (NLP) at work
no
So... it's a word processor?
Can it "read"?
yes
no
Is it reading what you type?
yes
no
does it respond in a useful, sensible way?
yes
is it transcribing what you say?
yes
no
cool, it's using speech recognition
so it's a mic
is it reading passages of text
yes
no
is it analyzing it for patterns?
no
then what is it reading...
It's just reading for fun, huh?
Can it move?
yes
no
By itself, without help?
yes
no
does it move based on what it sees & hears?
yes
no
that's just a dumb bot
are you sure it's not just moving on a programmed path?
yes
no
ok, it's a smart robot
Can it "reason"?
yes
no
Is it looking for patterns in massive amounts of data?
yes
no
well then, doesn't sound like reasoning to me
is it using those patterns to make decisions?
yes
no
neat, that's machine learning
hmm... sounds like math
Yup, it's using AI
"Is it using AI?" The Algorithm, MIT Technology Review
By: Karen Hao

PART 2:
PAST AND PRESENT OF VOICE INTERFACES



1968



1977

2016

BUT… WHY VOICE?

BUT… WHY VOICE?

1. IT’S EASY
2. IT’S HANDS- AND EYES-FREE
3. WE’RE GONNA NEED IT WHERE WE’RE GOING

1. IT’S EASY

BUT… WHY VOICE?

IT'S EASY

Do you really need a separate app for…

… Reminders?

… Alarms?

… Shopping lists?

… Weather?

… etc?

Sure, you want to learn the UI of 90 apps?

Monthly average amount of apps per smartphone*:

- Used: ~30
- Installed: ~90

And this is just for **smartphones** (add the Web, computers…)
And not even counting the **regular UI updates**.

*Report: Smartphone owners are using 9 apps per day, 30 per month (**TechCrunch**)

IT'S EASY

Speaking is among the first things we learn

+

Learning new UIs all the time is exhausting

=

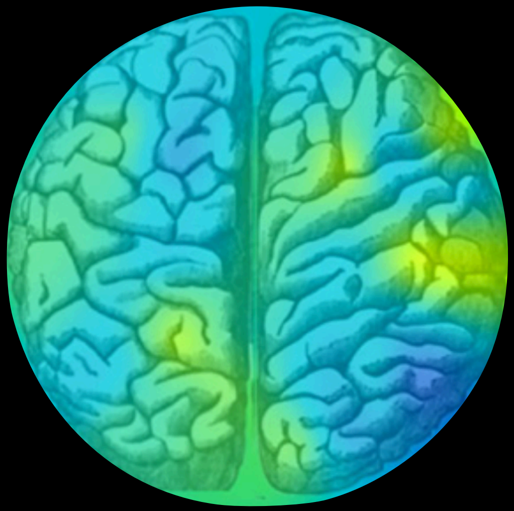
Just speak!

(let us do the design work behind)

BUT… WHY VOICE?

IT'S EASY

Speaking to a device is 2x easier than typing on it



Low mental activity

High mental activity

> Imagine versus other interfaces!

2017 Neuro-Insight study

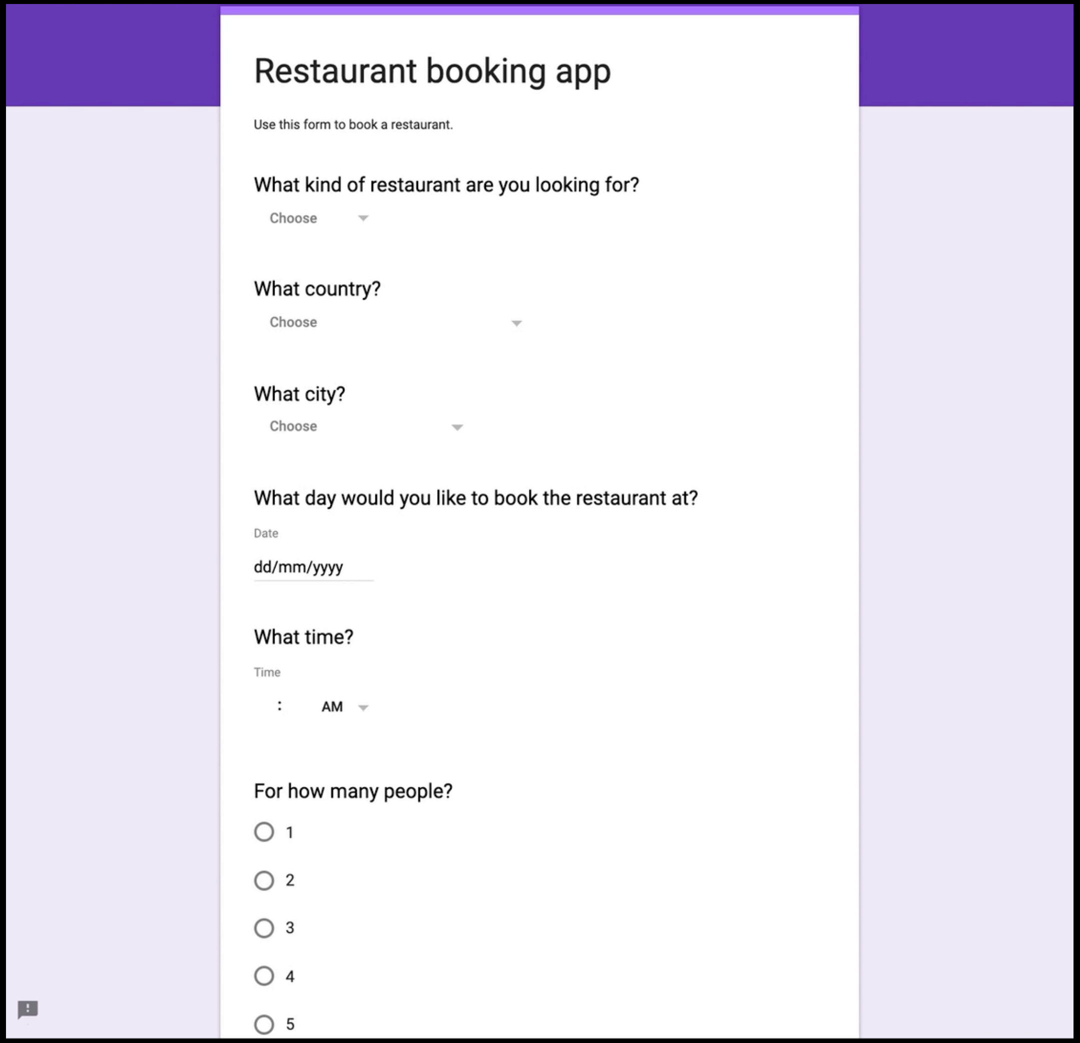
BUT… WHY VOICE?

IT'S EASY

Well, no need to imagine:

*"Hey Snips,
Book me a table in an Italian restaurant in Paris for 8 people on Friday, 10pm"*

VS



Restaurant booking app

Use this form to book a restaurant.

What kind of restaurant are you looking for?
Choose

What country?
Choose

What city?
Choose

What day would you like to book the restaurant at?
Date
dd/mm/yyyy

What time?
Time
: AM

For how many people?
○ 1
○ 2
○ 3
○ 4
○ 5

<6s

>40s

2. IT’S HANDS- AND EYES-FREE

BUT… WHY VOICE?

IT'S HANDS- AND EYES-FREE

Times you don't want to use your eyes…

- Driving or using heavy machinery
- Being concentrated on your book or screen
- Other?

BUT... WHY VOICE?

IT'S HANDS- AND EYES-FREE

Sometimes you don't want to use your hands...

- Cooking
- Writing
- Other?

BUT… WHY VOICE?

IT'S HANDS- AND EYES-FREE

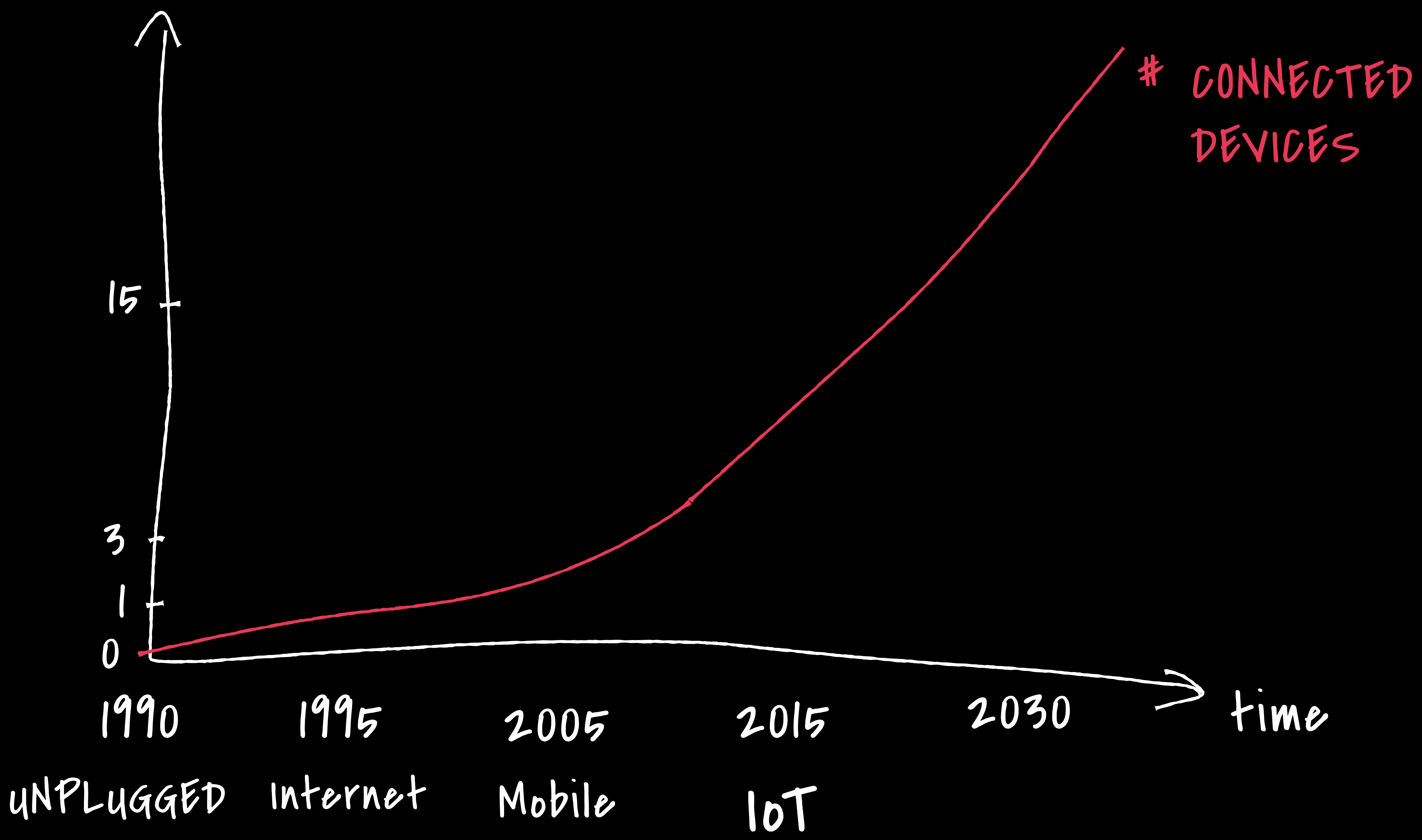
Times you don't want to use your hands OR your eyes…

- Running, swimming, working out
- Getting "close" to your partner
- Other?

3. WE'RE GONNA NEED IT WHERE WE'RE GOING

BUT… WHY VOICE?

WE'RE GONNA NEED IT WHERE WE'RE GOING



BUT… WHY VOICE?

WE’RE GONNA NEED IT WHERE WE’RE GOING

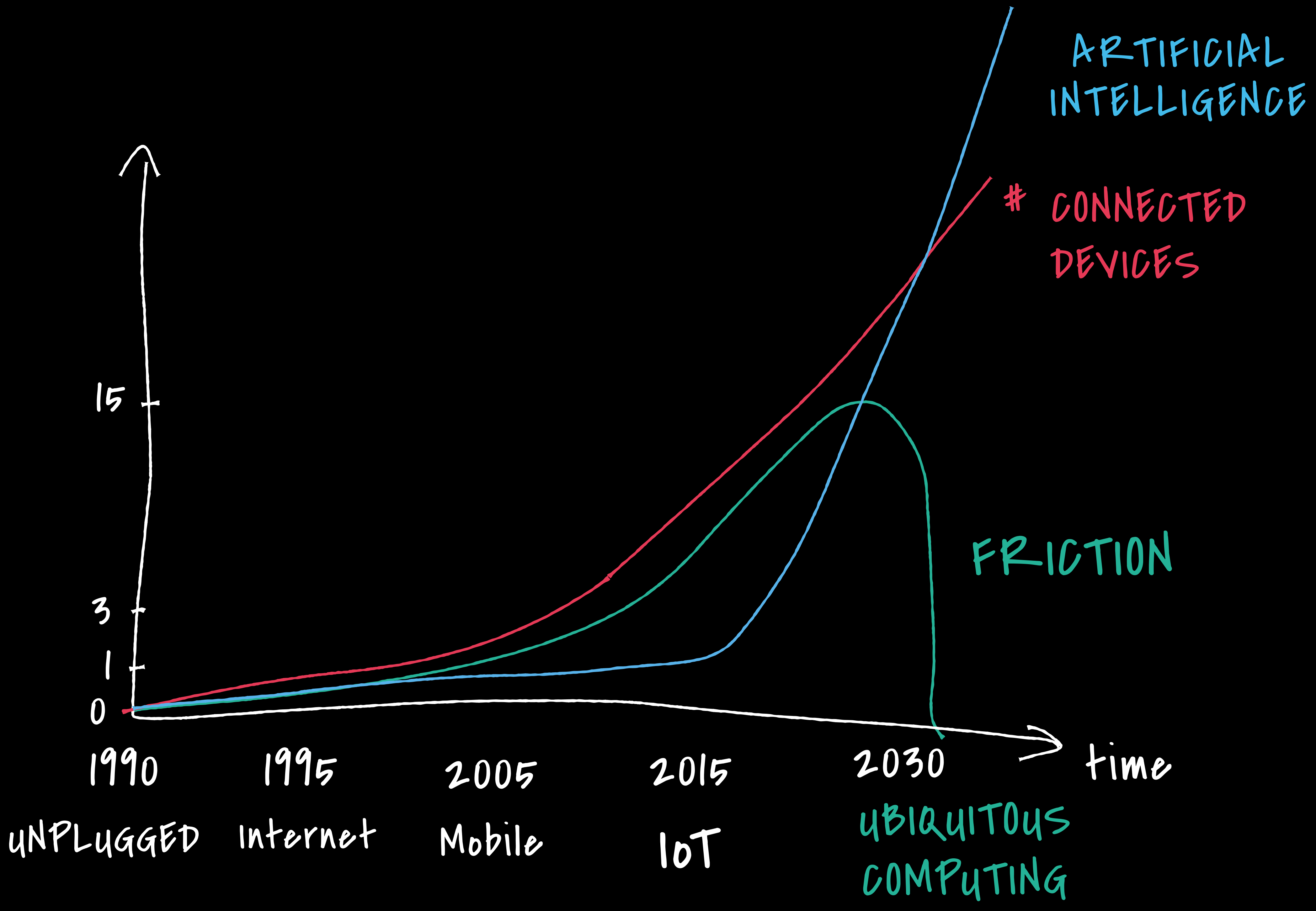
How many total devices?

How many different apps per device?

How many notifications/actions per app?

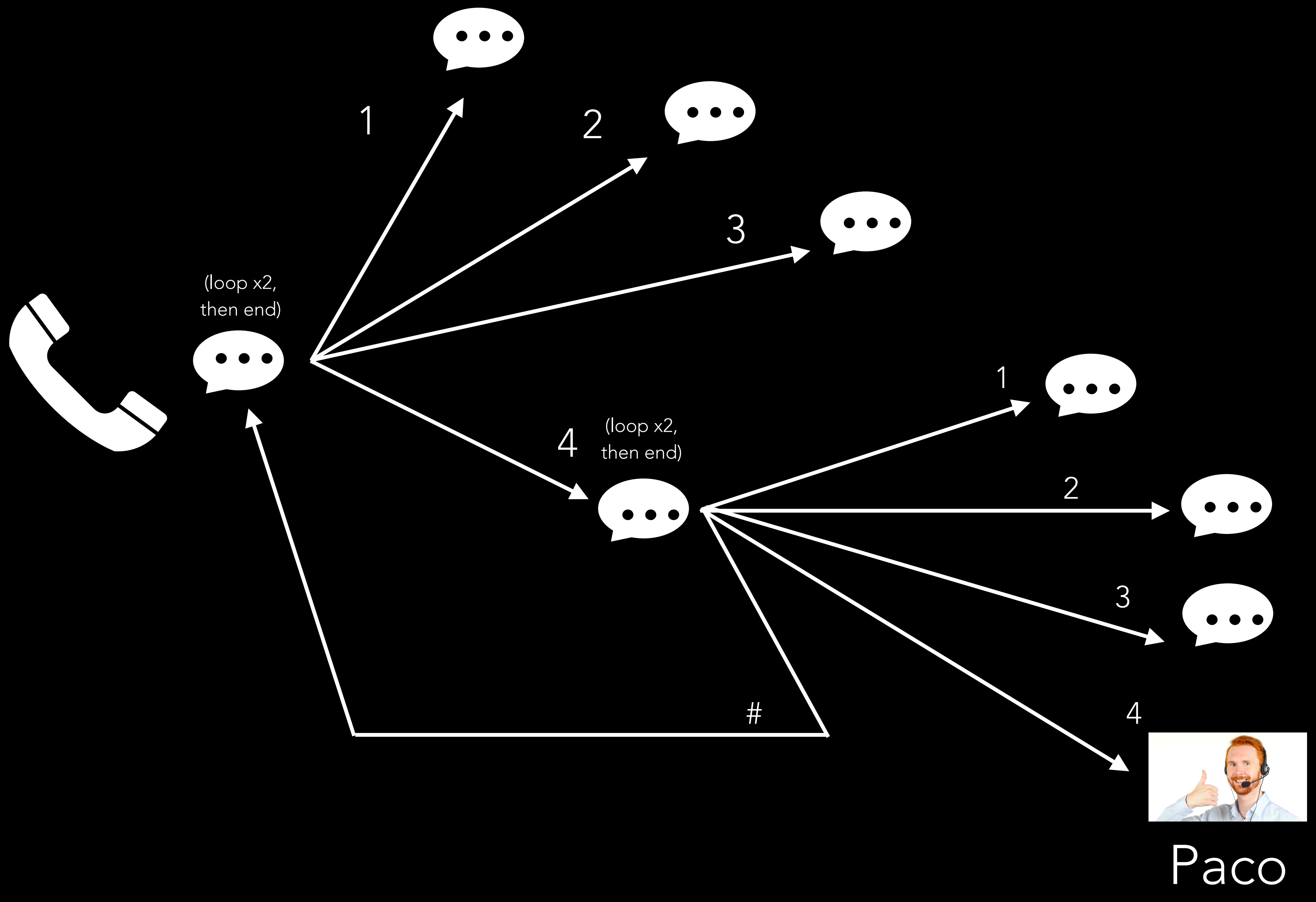
BUT... WHY VOICE?

WE'RE GONNA NEED IT WHERE WE'RE GOING



A BRIEF HISTORY OF VOICE INTERFACES

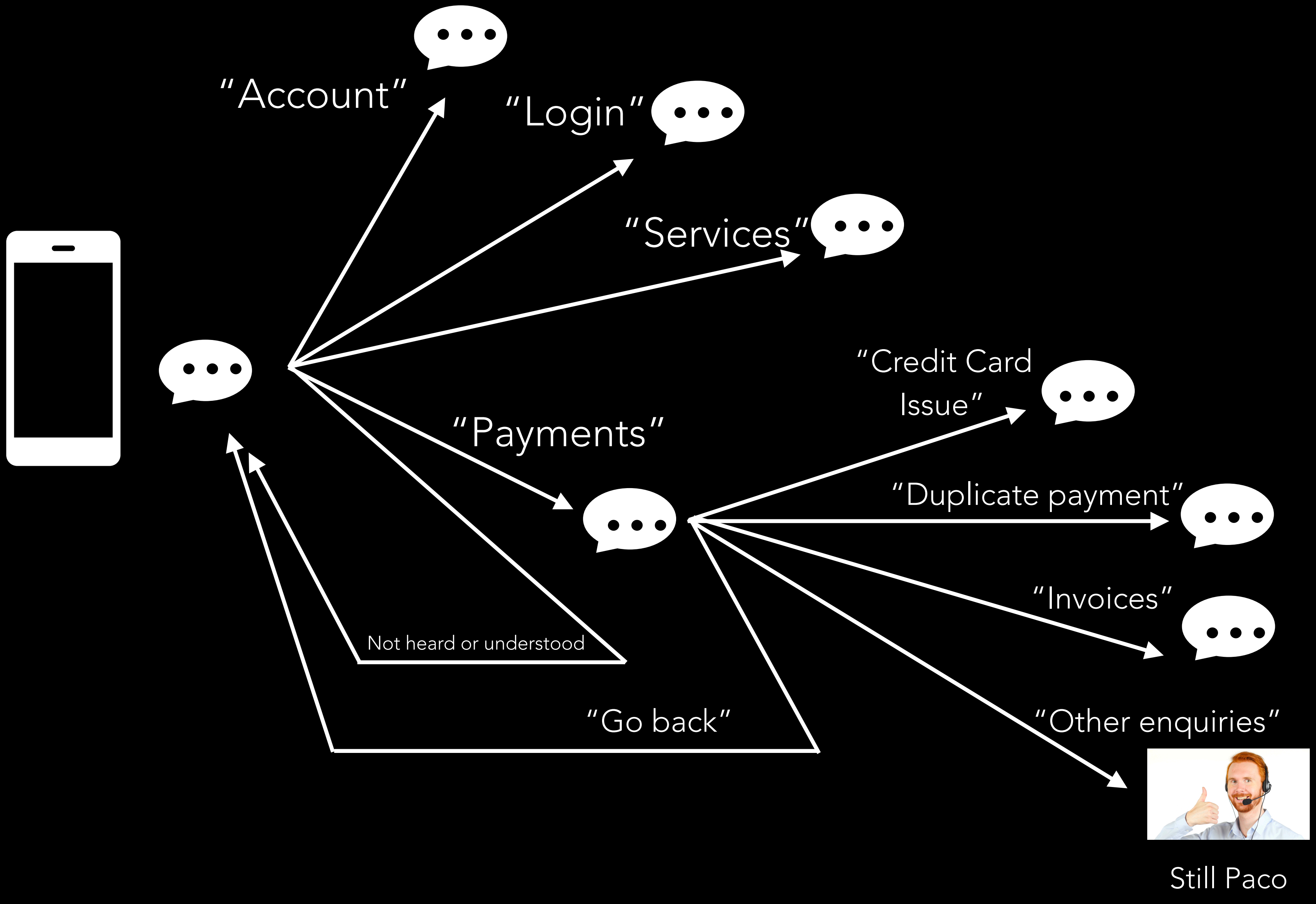
INTERACTIVE VOICE RESPONSES / DUMB IVRS



1980s

A BRIEF HISTORY OF VOICE INTERFACES

INTERACTIVE VOICE RESPONSES / OKAYISH IVRS



Early 2000s

A BRIEF HISTORY OF VOICE INTERFACES

DIFFERENCES BETWEEN IVRS AND "PROPER" VUIS

IVR	Other VUIs
Single-purpose	Several domains
Restricted flow (decision tree)	Flexible / Free flow (ask whatever, whenever)
Essentially *"voice clicks"*	Focus on understanding intention from natural speech

DIFFERENCES BETWEEN CHATBOTS AND VUIS

Chatbots (asynchronous & visual - like messaging)	**VUIs** (synchronous & auditive - like talking)
Text is clear and perfect in format (Text doesn't have noise, doesn't depend on microphone quality or someone's accent)	Voice needs additional processing (Voice includes background noise, accents, pronunciations…)
No need to memorise info (history of conversation and rich content available to check before formulating answer)	*"Sorry, what were we talking about?"* (human brain isn't necessarily good at memorising stuff from hearing)
Bilateral certainty about duration (both players know when they're supposed to speak and when they're supposed to listen)	Bilateral **un**certainty about duration (both players have to figure out when they're supposed to speak and when they're supposed to listen)
Users can think their input through (and only send when ready)	Users express their intent naturally (including doubting, corrections on the fly…)

THE STATE OF VOICE INTERFACES TODAY

THE STATE OF VOICE INTERFACES TODAY

HOW MANY

Between 1 and 3.2 billion voice assistants in use today*.

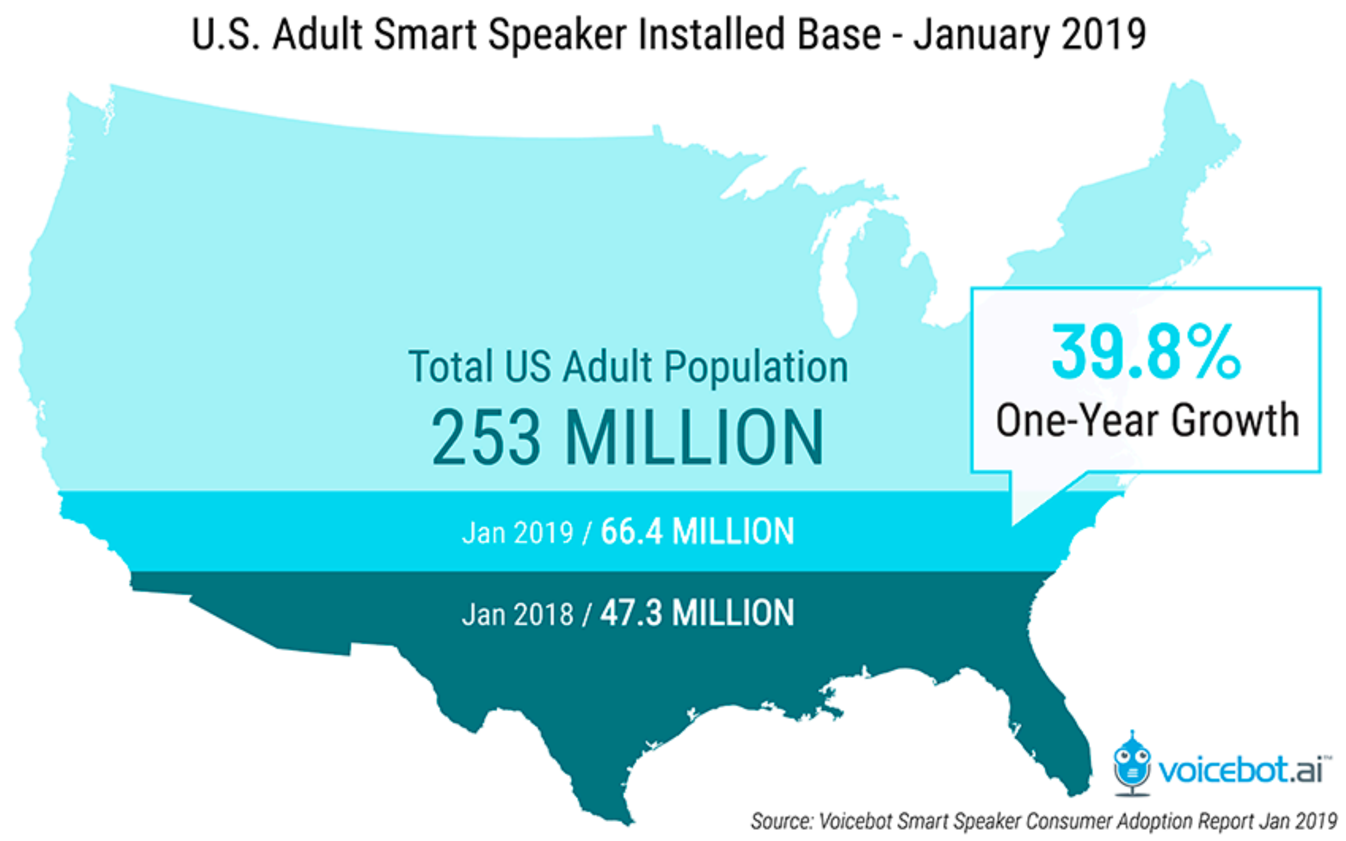
Leaders:

- Google Assistant (Android phones, Google Home devices)
- Apple Siri (iPhones, Mac computers, HomePods)
- Amazon Alexa (Echo devices)
- Microsoft Cortana (PCs, I guess?)

*Source: Juniper Research, 2018

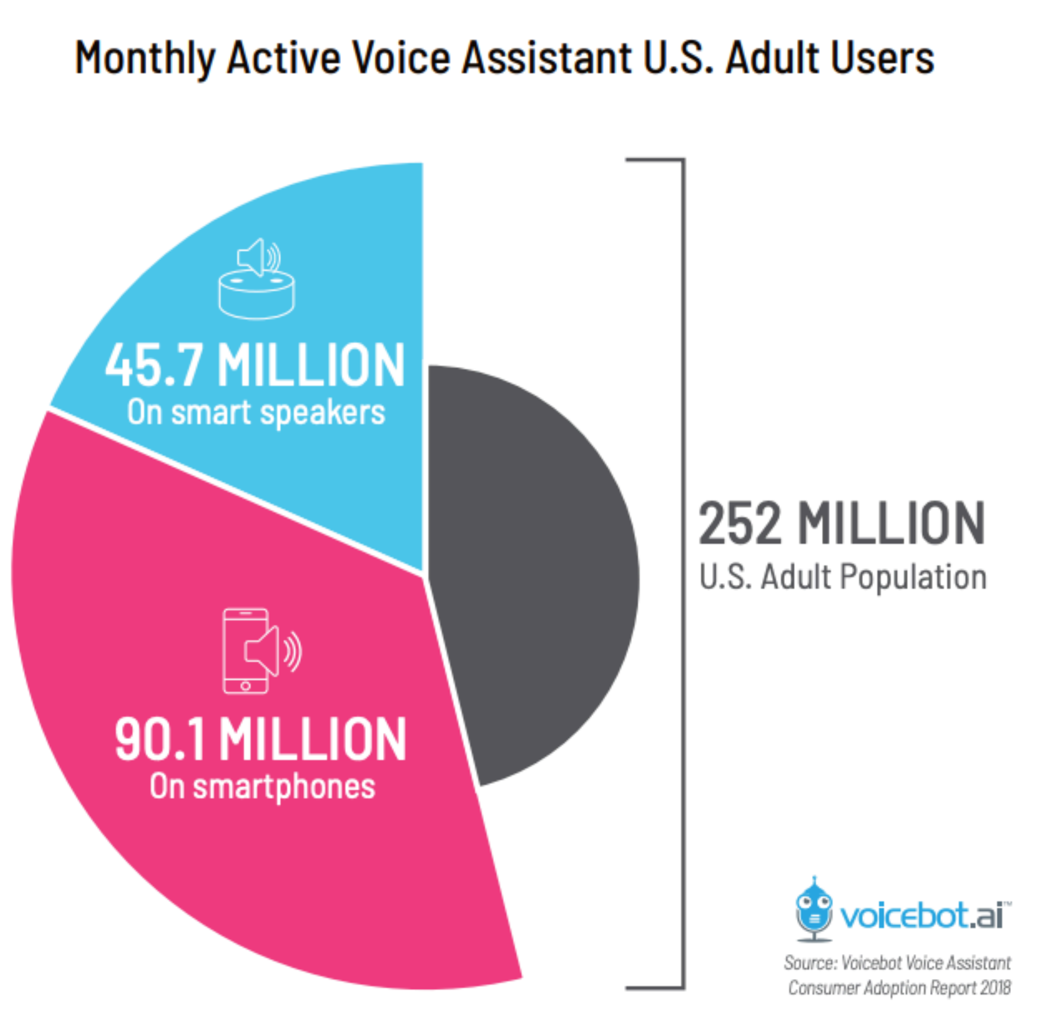
THE STATE OF VOICE INTERFACES TODAY

HOW MANY



Source: Voicebot Smart Speaker Consumer Adoption Report Jan 2019

1 in 4 US households have a smart speaker



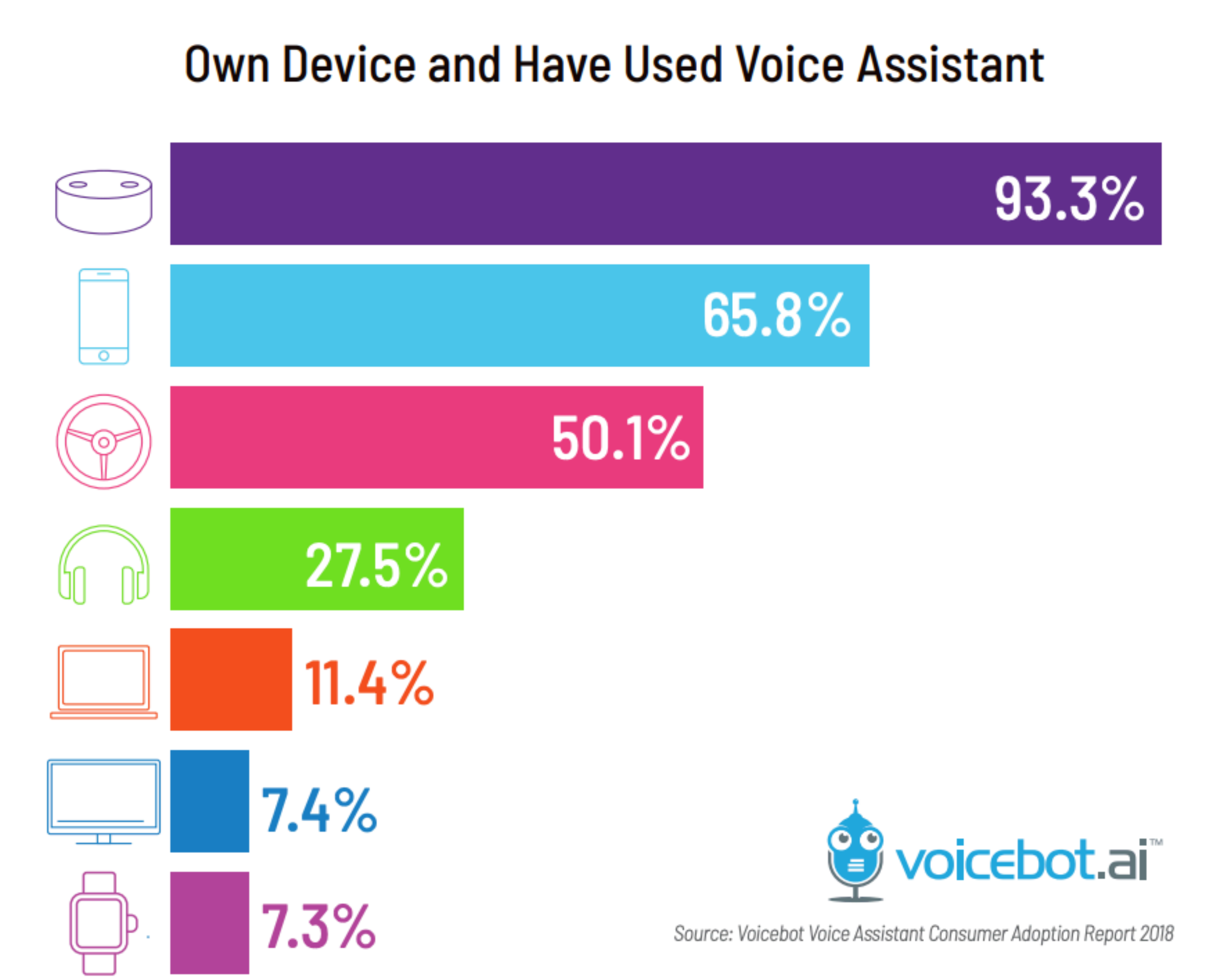
Source: Voicebot Voice Assistant Consumer Adoption Report 2018

54% of US households use a voice assistant on a monthly basis

*Source: Voicebot.ai, 2018

THE STATE OF VOICE INTERFACES TODAY

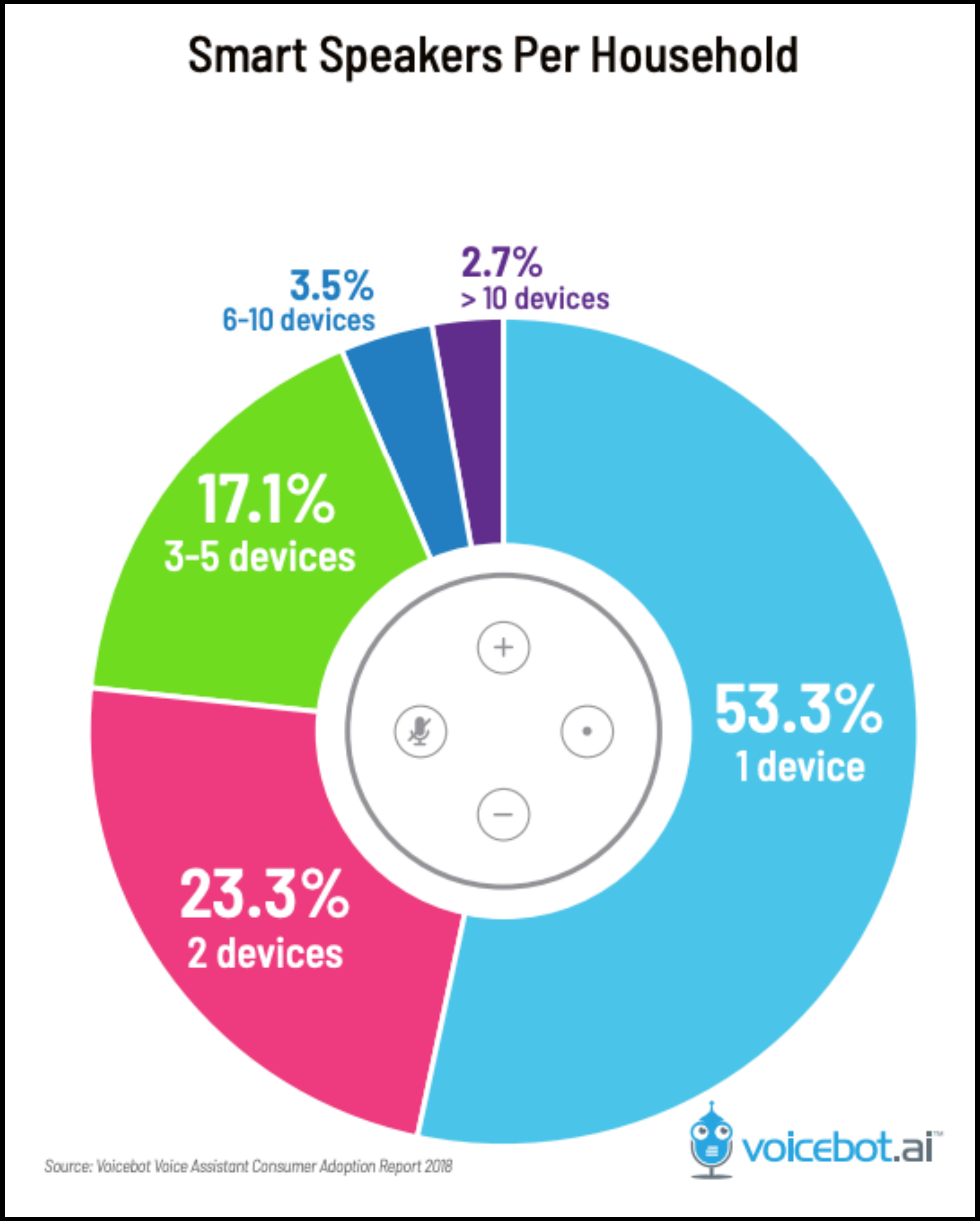
HOW MANY (II)



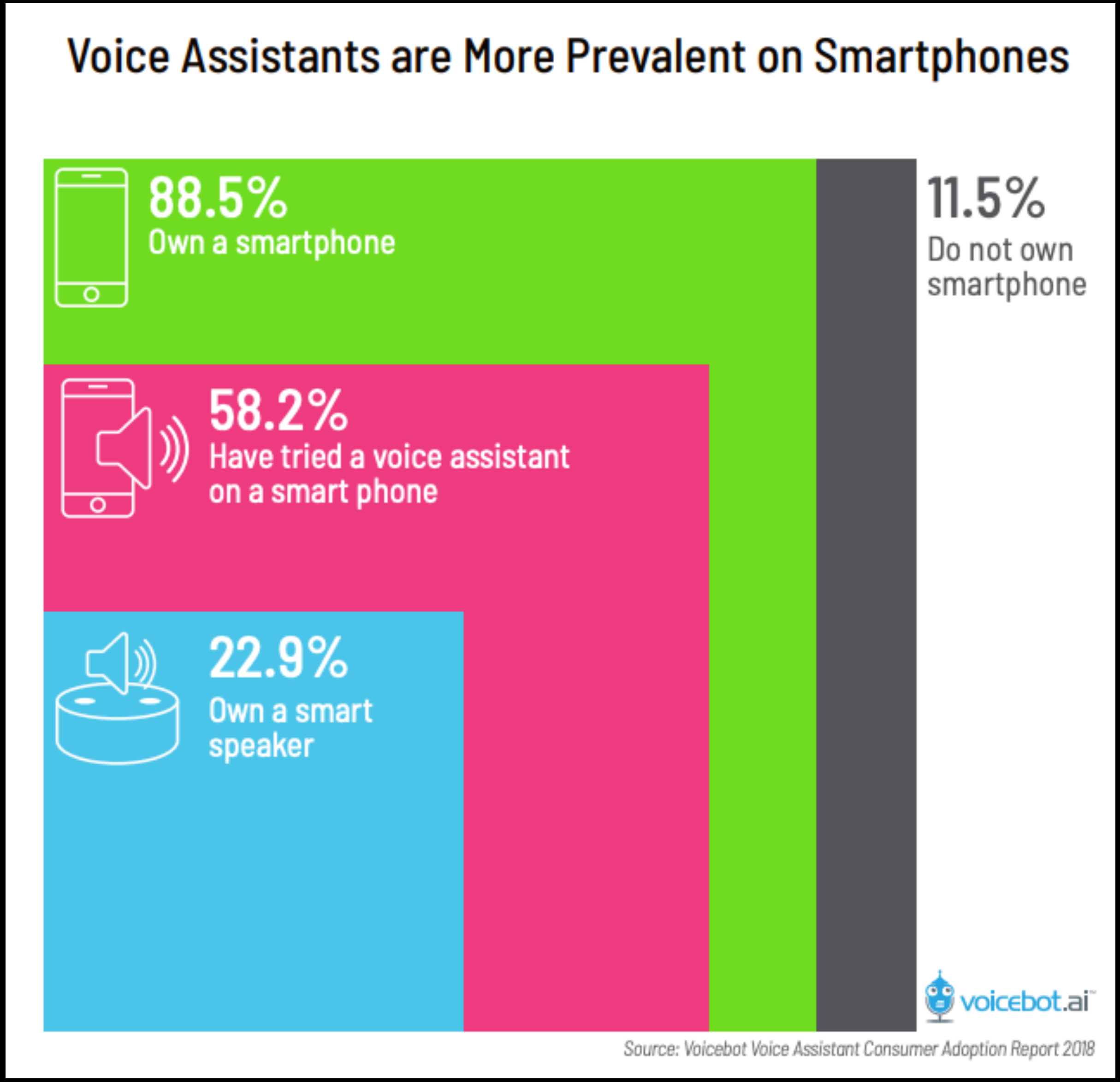
Source: Voicebot Voice Assistant Consumer Adoption Report 2018

*Source: Voicebot.ai, 2018

HOW MANY (II)



Source: Voicebot Voice Assistant Consumer Adoption Report 2018



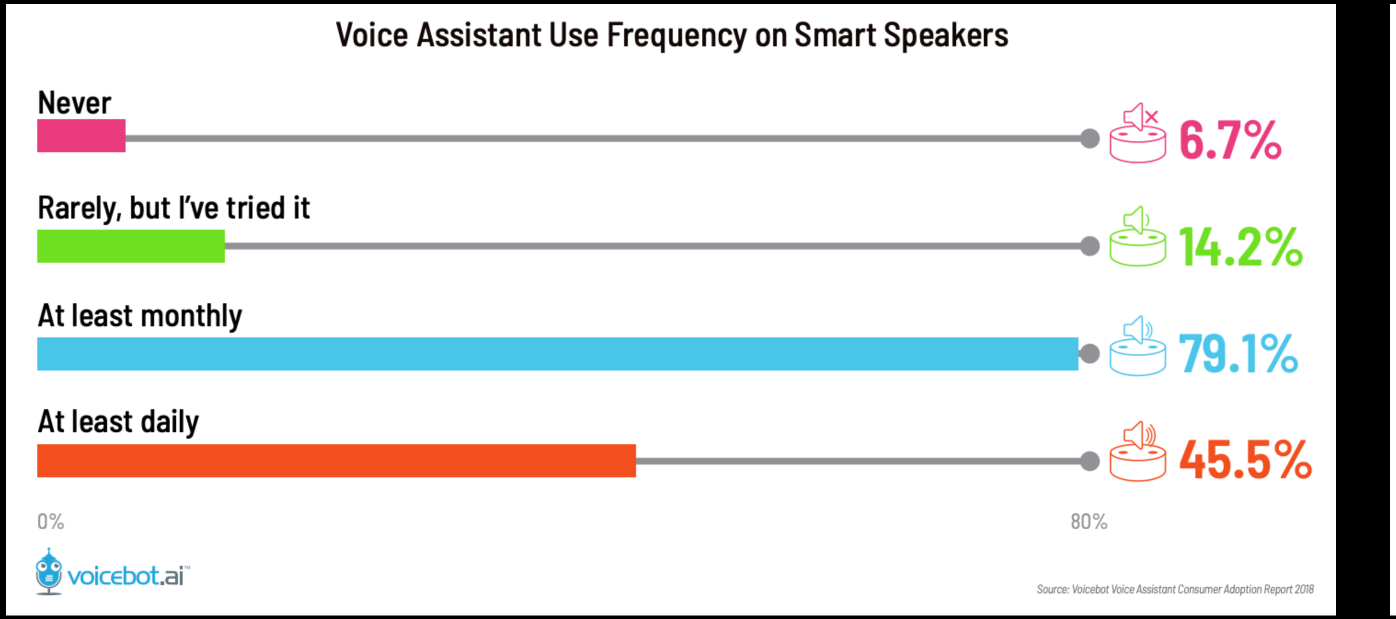
Source: Voicebot Voice Assistant Consumer Adoption Report 2018

*Source: Voicebot.ai, 2018

THE STATE OF VOICE INTERFACES TODAY

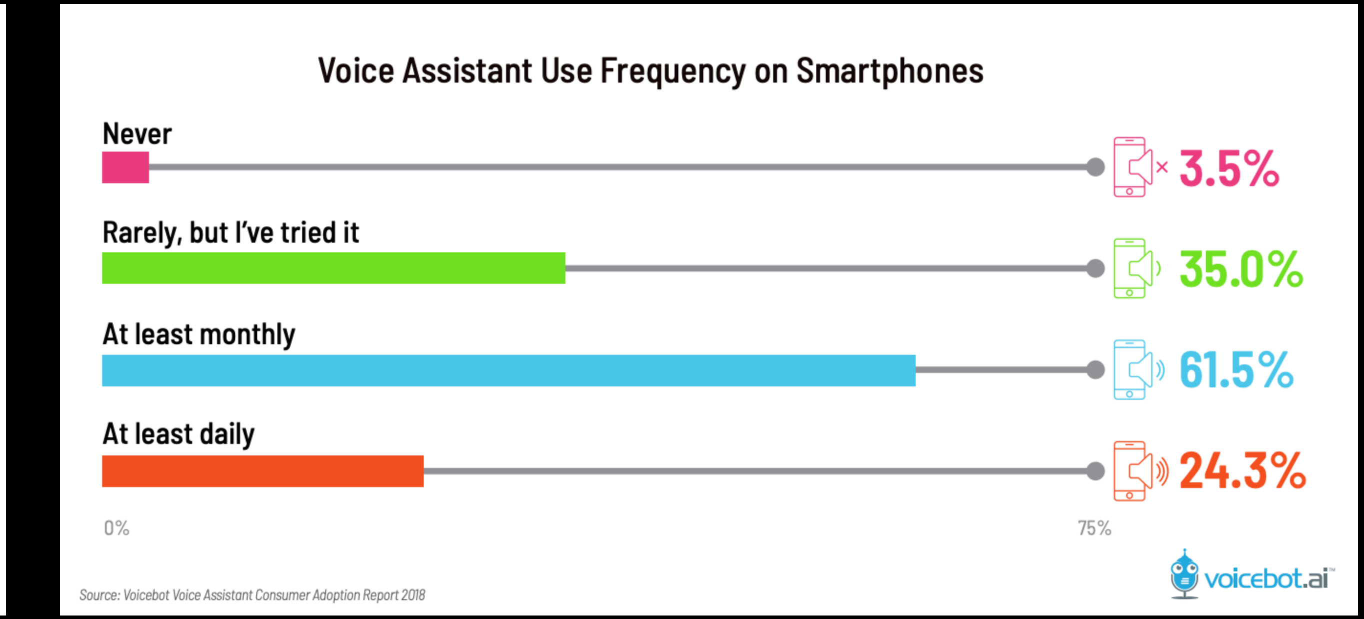
HOW OFTEN

Smart Speakers



Source: Voicebot Voice Assistant Consumer Adoption Report 2018

Smartphones

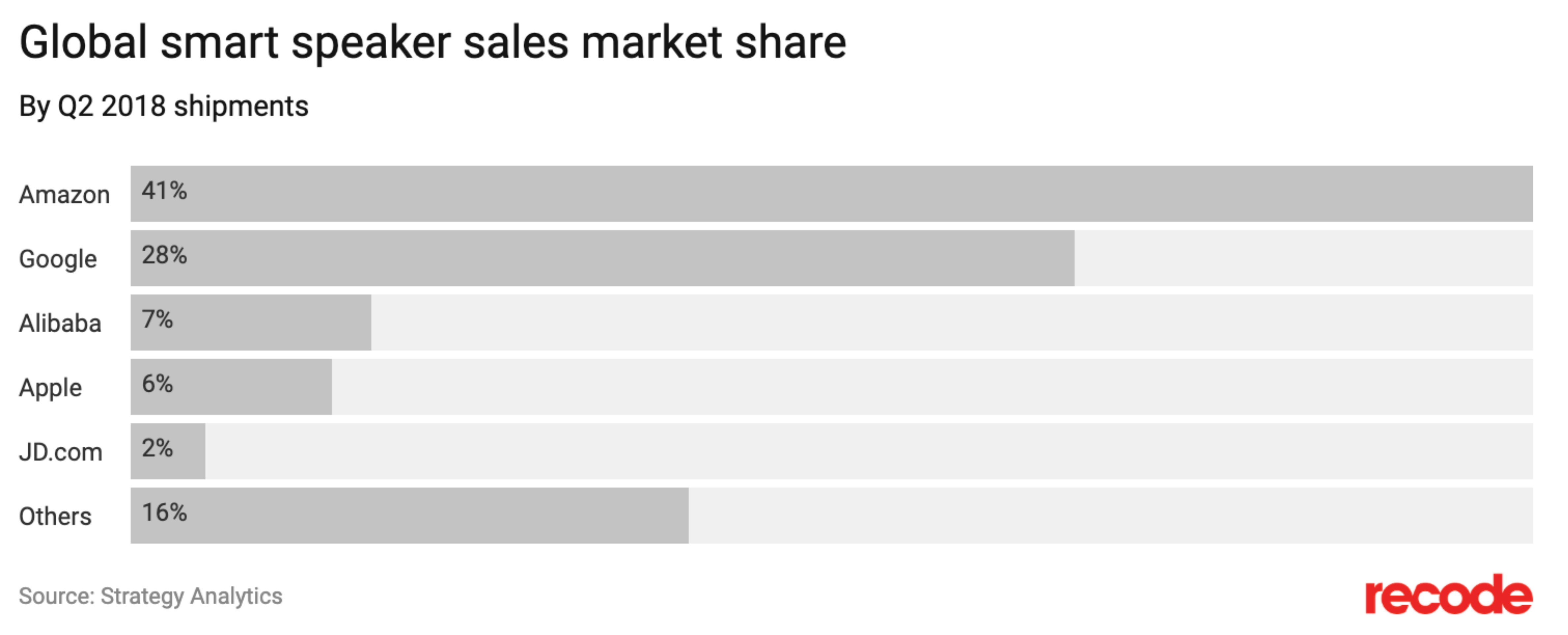


Source: Voicebot Voice Assistant Consumer Adoption Report 2018

*Source: Voicebot.ai, 2018

THE STATE OF VOICE INTERFACES TODAY

WHO

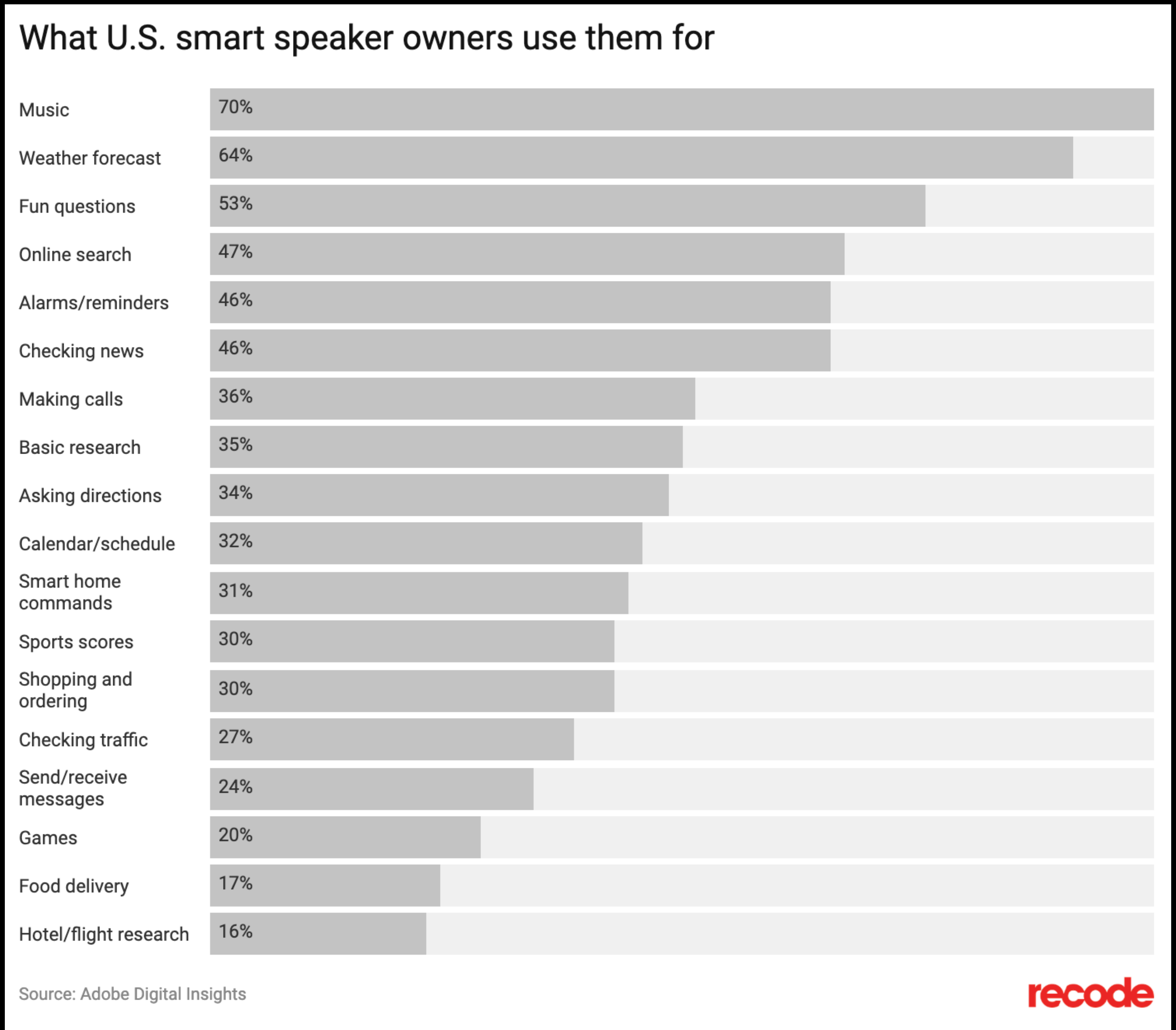


Source: Recode, 2018

THE STATE OF VOICE INTERFACES TODAY

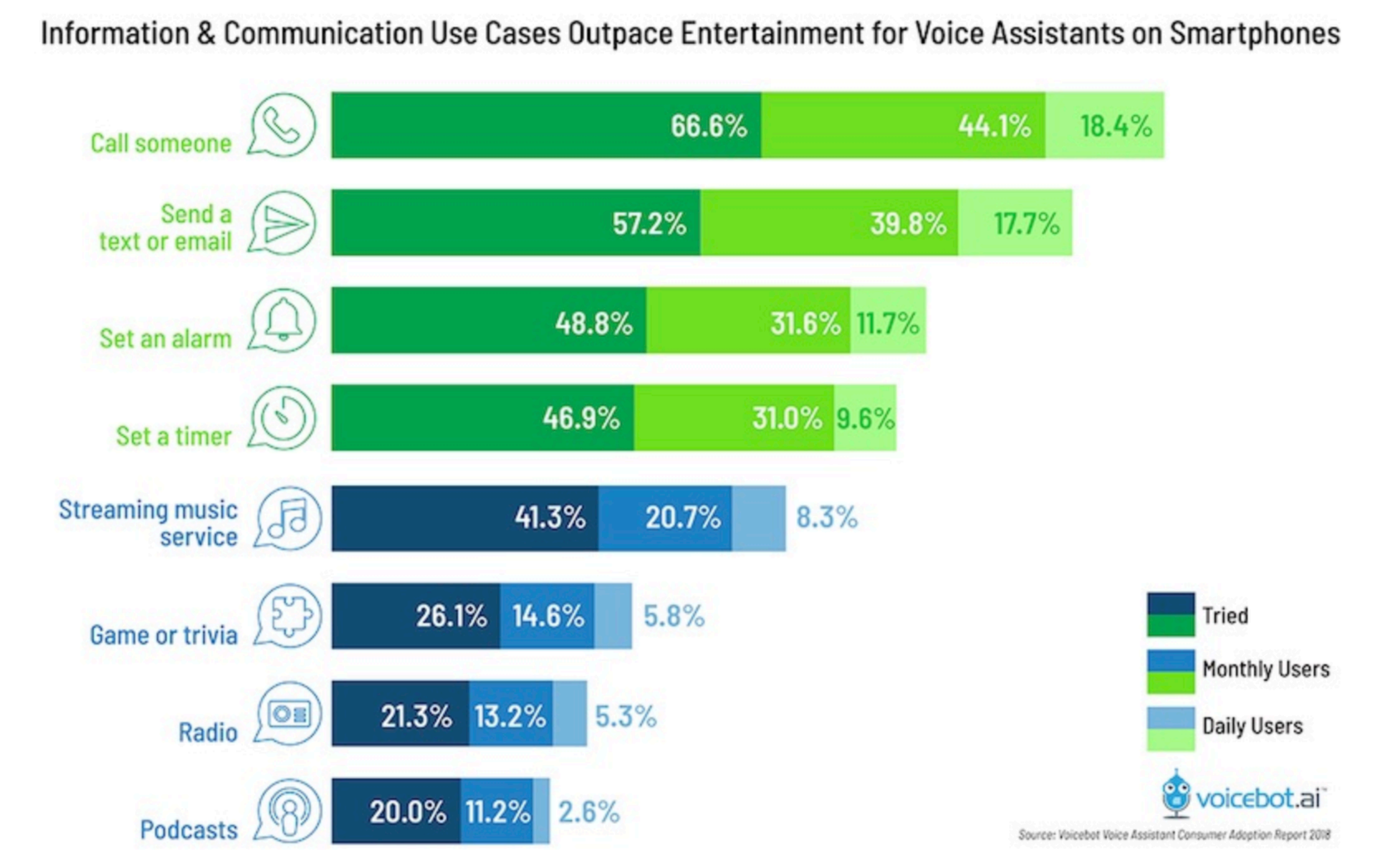
WHAT FOR

Smart Speakers


What U.S. smart speaker owners use them for

Source: Adobe Digital Insights

Smartphones


Information & Communication Use Cases Outpace Entertainment for Voice Assistants on Smartphones

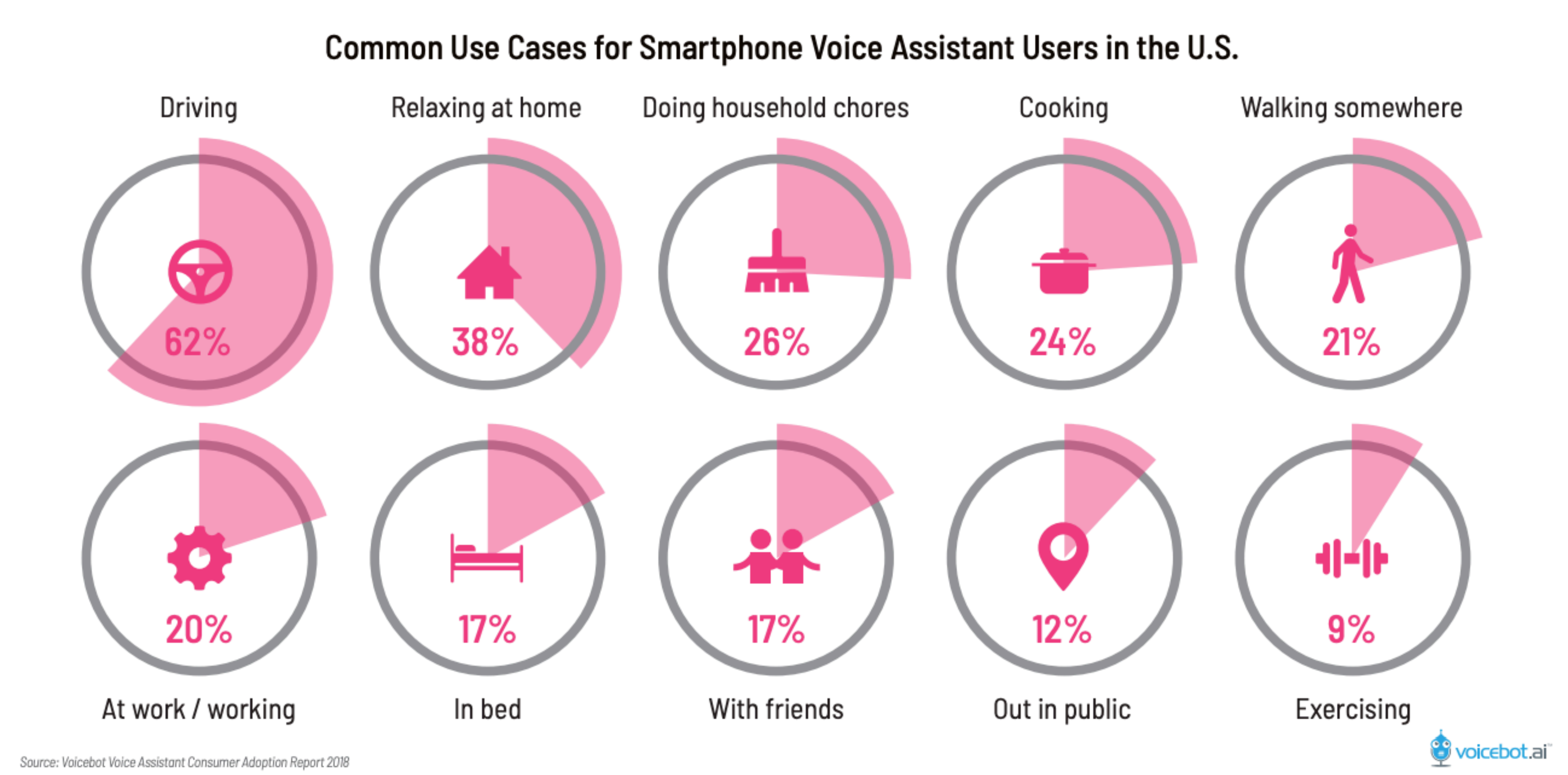
Source: Voicebot Voice Assistant Consumer Adoption Report 2018

Almost opposite!

*Sources: Voicebot.ai, 2018, Recode, 2018

THE STATE OF VOICE INTERFACES TODAY

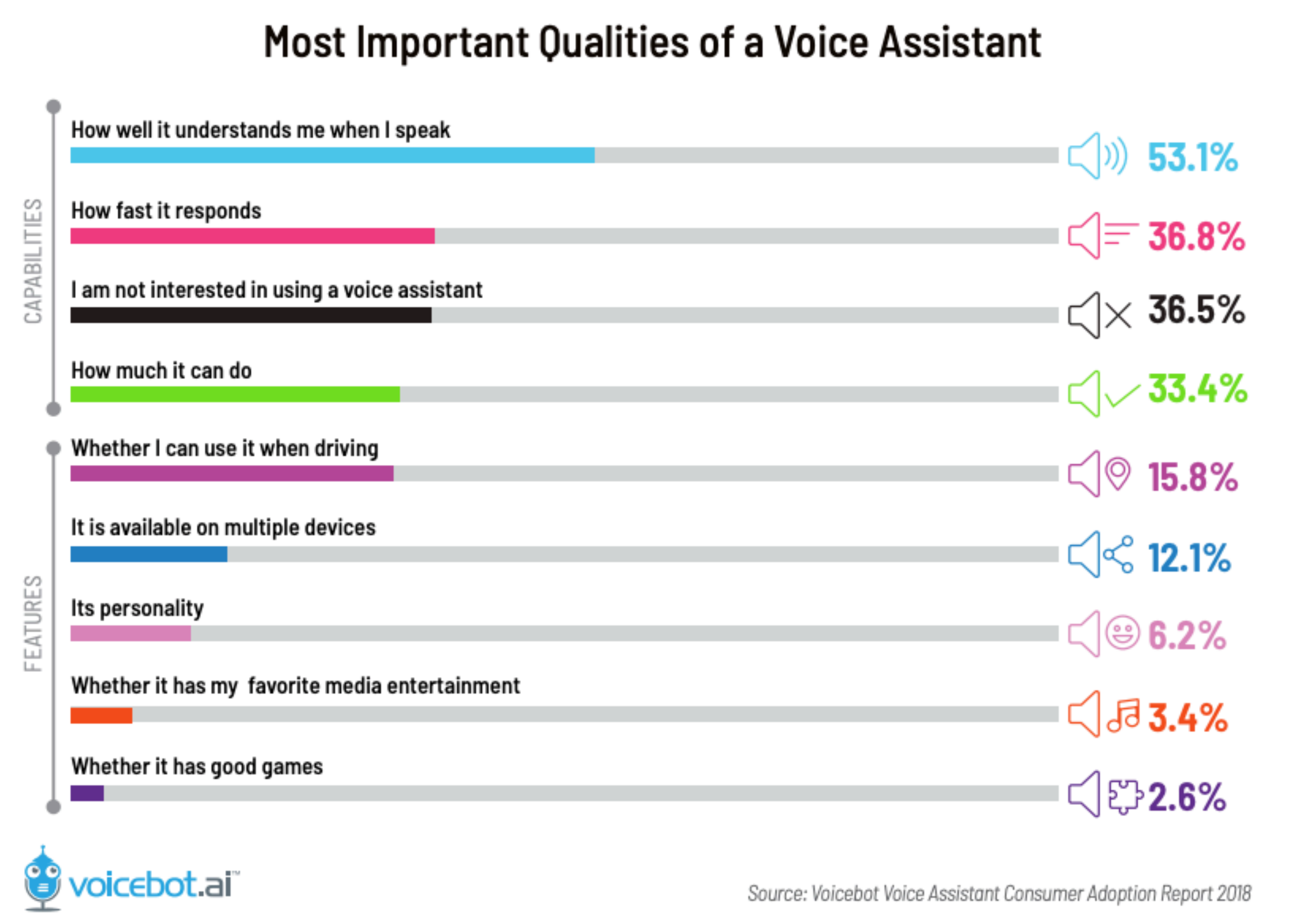
WHERE



*Source: Voicebot.ai, 2018

THE STATE OF VOICE INTERFACES TODAY

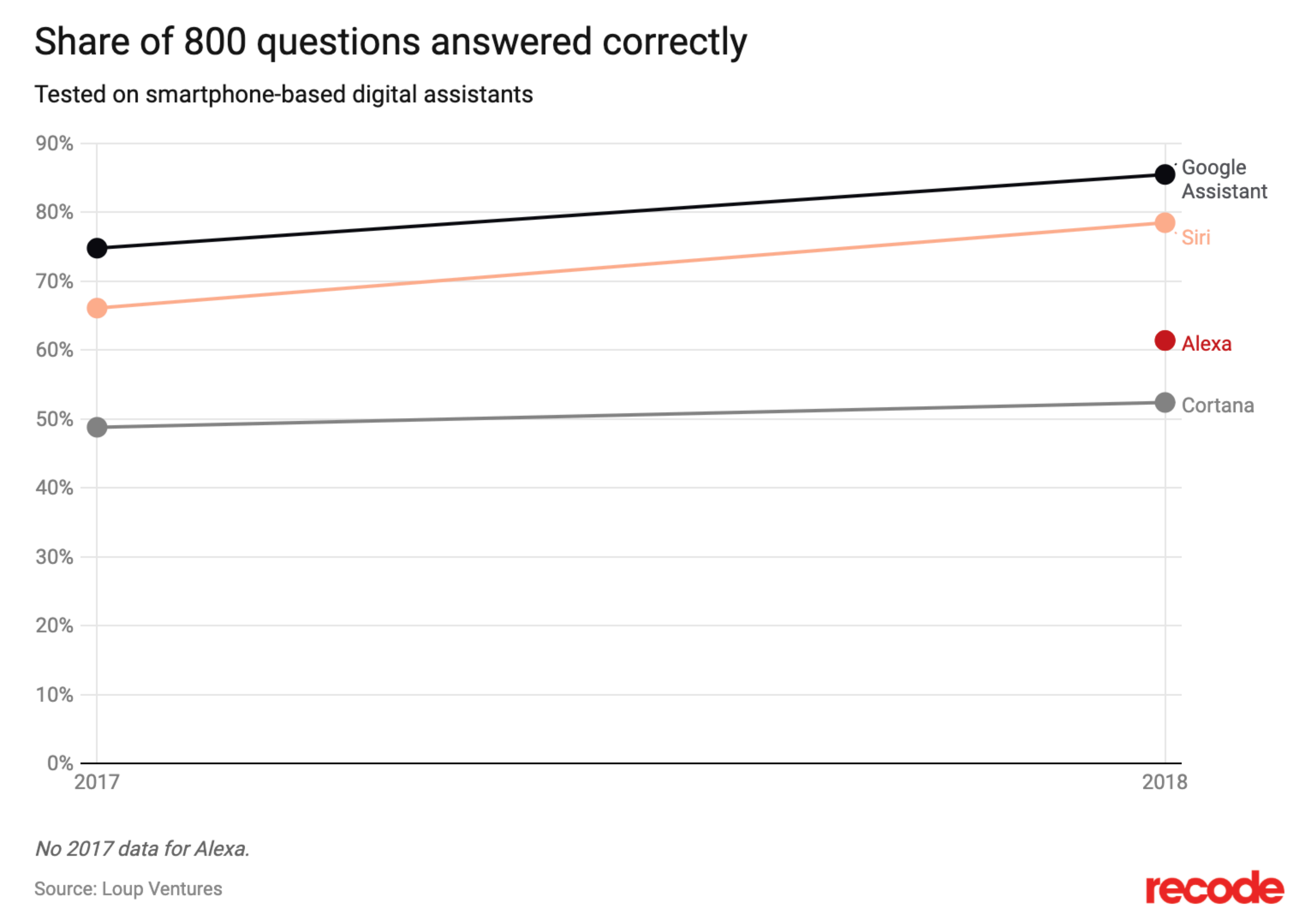
WHY



*Source: Voicebot.ai, 2018

THE STATE OF VOICE INTERFACES TODAY

WHY



*Source: Recode, 2018

THE STATE OF VOICE INTERFACES TODAY

MARKET EVOLUTION: CONCLUSIONS

- **Currently**: between **1 and 3.2 billion voice assistants in use**.
- **Estimates**: two-figure yearly growth (+20-40%) to **8 billion voice assistants in 2023***.
- **Smart speaker base growing fast** in US (25% of population, +40% YoY), UK, EU, CN and AU. Rest of world: main access through smartphones.
- **54% of US households use** a voice assistant on a **monthly basis**
- **61% of US smartphone owners** use voice assistants **monthly**
- **Main players: Google and Amazon** (70% of aggregated share)
- **Main use cases:**
    - Smart Speakers: Music (70%), Weather (64%), Q&A (~50%)
    - Smartphones: Calling (44%), Texting/Email 40%), Alarms (32%)

*Source: Juniper Research, 2018

PART 3: VOICE INTERFACE BASICS

Part 3: Voice interface basics

- What is a voice interface
- Voice interfaces in real life
- The tech behind voice
- Structure of a voice interface

WHAT IS A VOICE INTERFACE

WHAT IS A VOICE INTERFACE

DEFINITION (REMINDER I)

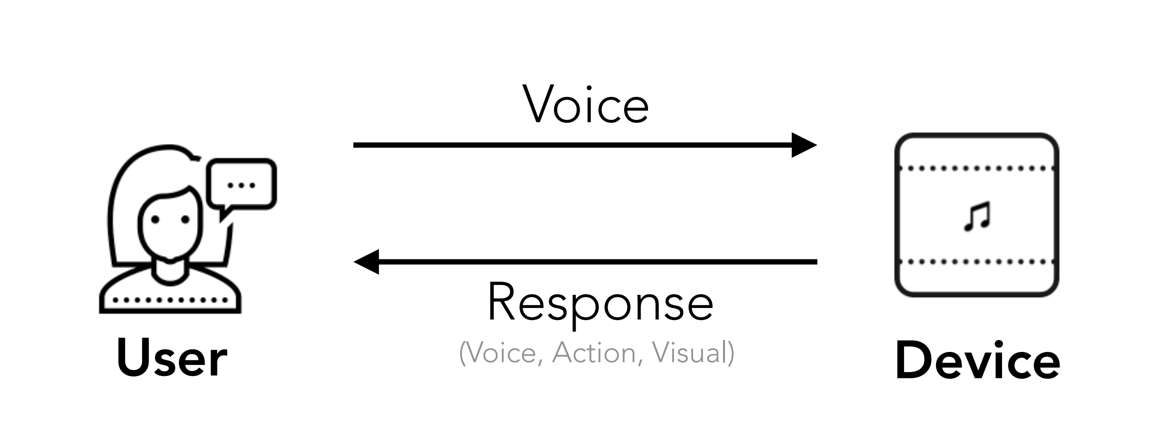
A **voice-user interface** (**VUI**):

allows for **spoken interaction** with computers using **speech recognition** to understand spoken commands/questions,

and typically **text to speech** to play a reply,

and/or other **actions,**

and/or **visual responses**.



WHAT IS A VOICE INTERFACE

WHAT IS NOT A VOICE INTERFACE

Amazon **Echo**

Google **Home**

Your phone

Your TV remote

ARE NOT voice interfaces (they are devices)**.**

Amazon **Alexa**

Google **Assistant**

Siri

ARE voice interfaces.

VOICE INTERFACES IN REAL LIFE

VOICE INTERFACES IN REAL LIFE

EXAMPLES (I)

"Alexa, play some music by The Strokes."

"Shuffling your songs by The Strokes, from Amazon Music."

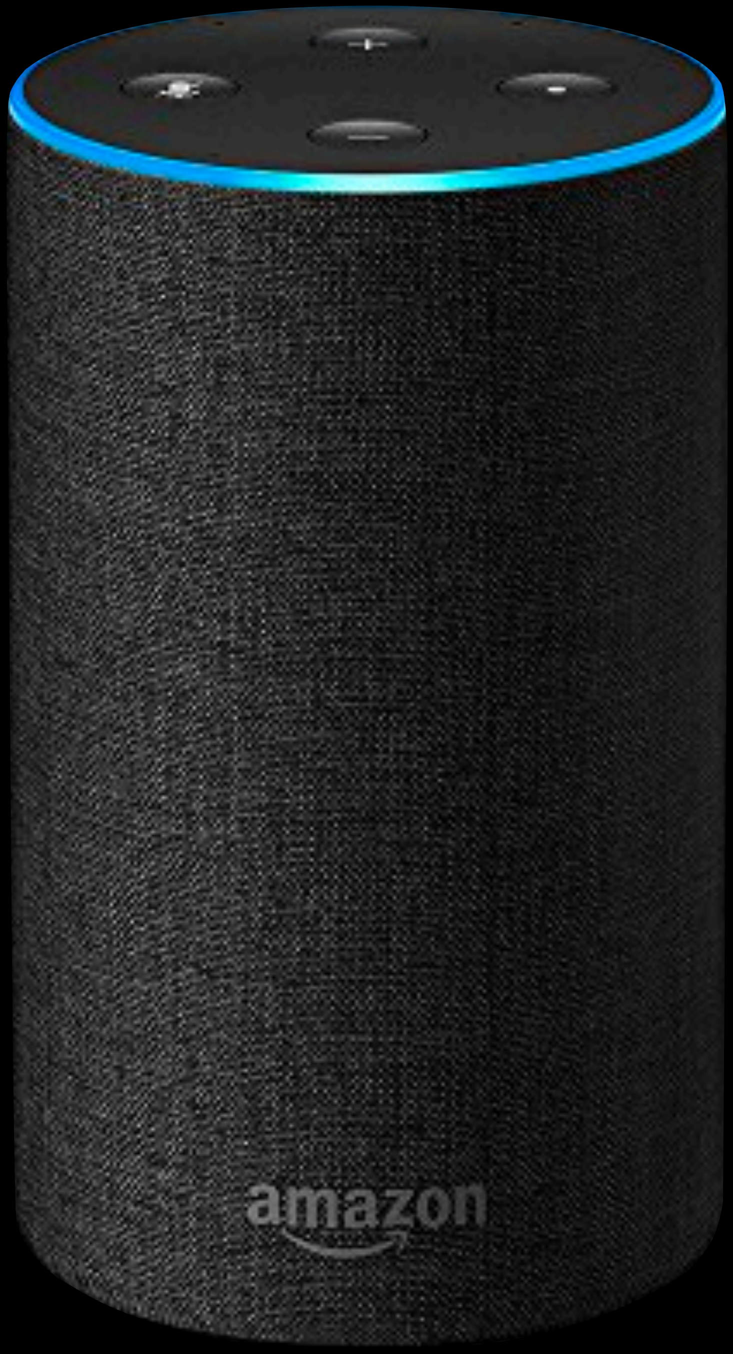
[Actually plays songs by The Strokes, in order of popularity]

"Alexa, it's too loud."

[Lowers volume]

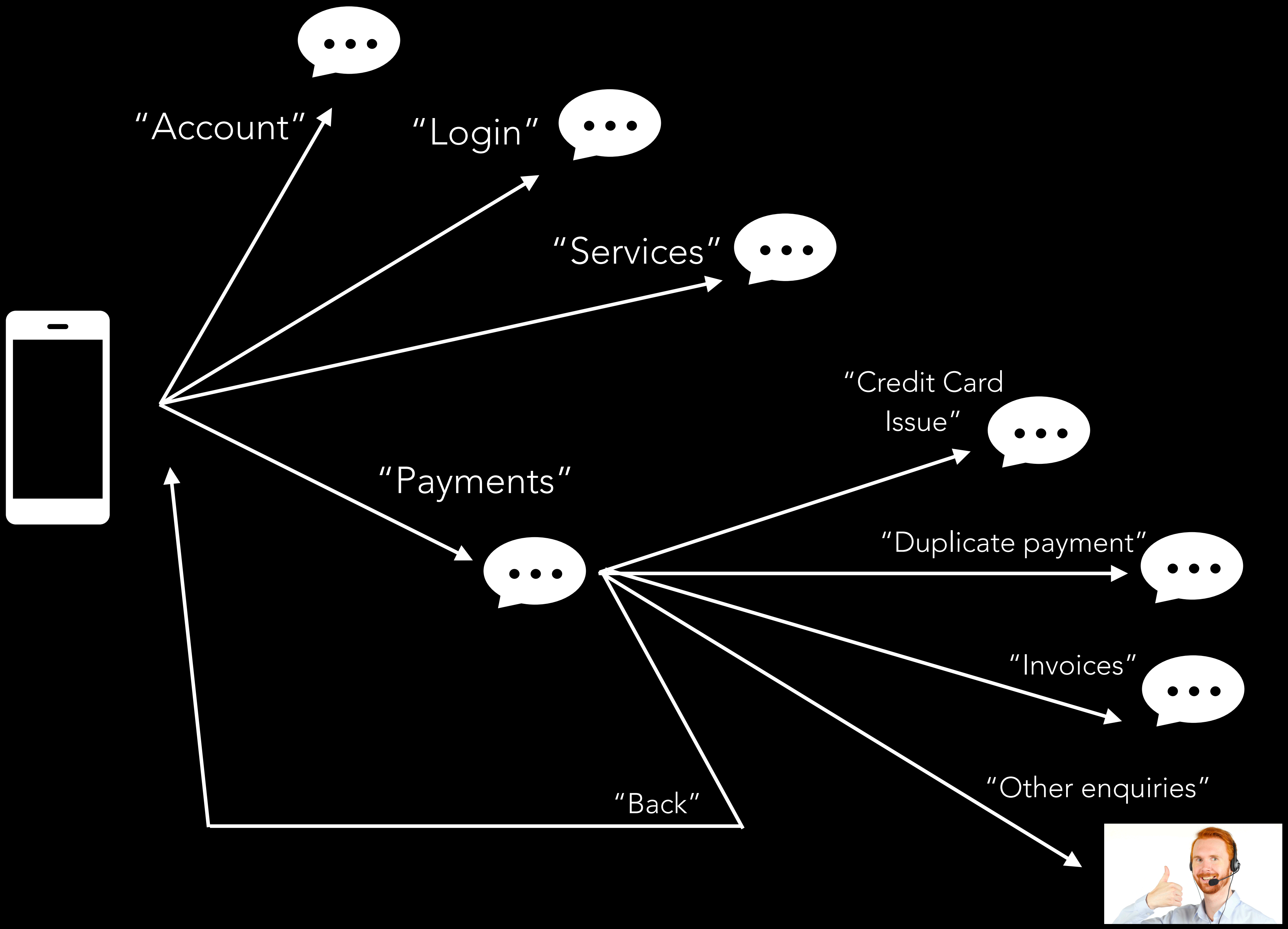
"Alexa, I don't really like this song."

[Skips song]



VOICE INTERFACES IN REAL LIFE

EXAMPLES (II)



Remember Paco?

THE TECH BEHIND VOICE

DEFINITION (REMINDER I)

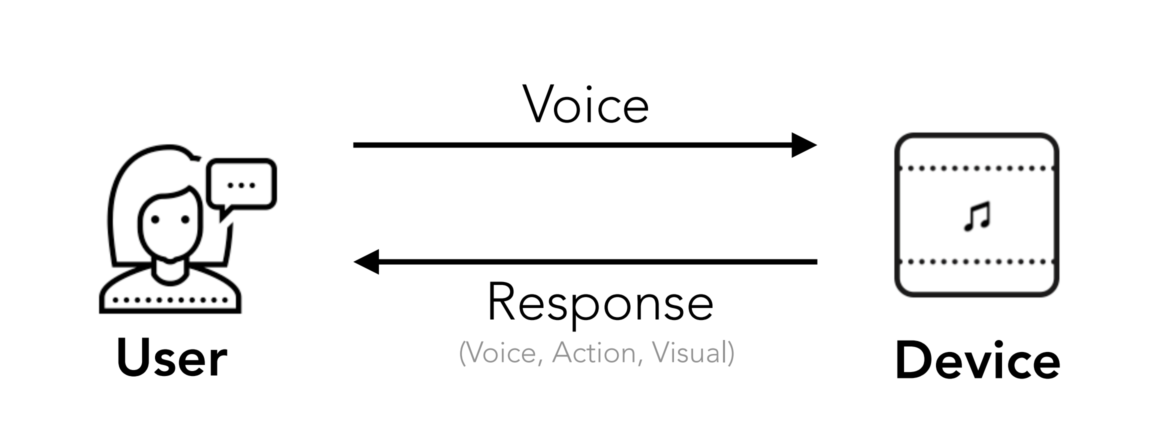
A **voice-user interface** (**VUI**):

allows for **spoken interaction** with computers using **speech recognition** to understand spoken commands/questions,

and typically **text to speech** to play a reply,

and/or other **actions,**

and/or **visual responses**.

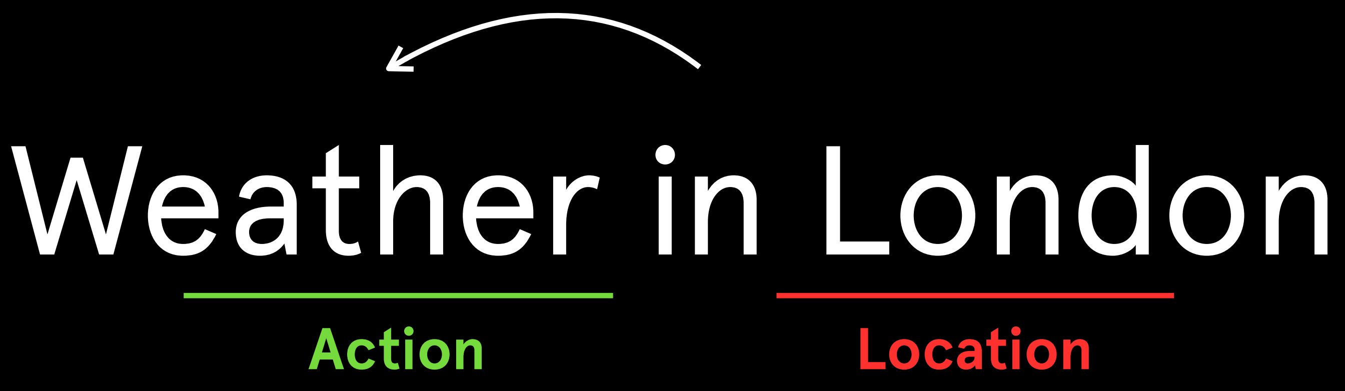


Daniel Fernández Castro / March 2019

Why is it harder than getting to space?

THE TECH BEHIND VOICE INTERFACES

WHY IS IT SO COMPLEX? (II)



Weather in London

Action

Location

This one’s actually a piece of cake.

THE TECH BEHIND VOICE INTERFACES

WHY IS IT SO COMPLEX? (II)

Airbnb HQ? Or where?

Is "me" the object?

Find me an Italian near my Airbnb

Action

Object

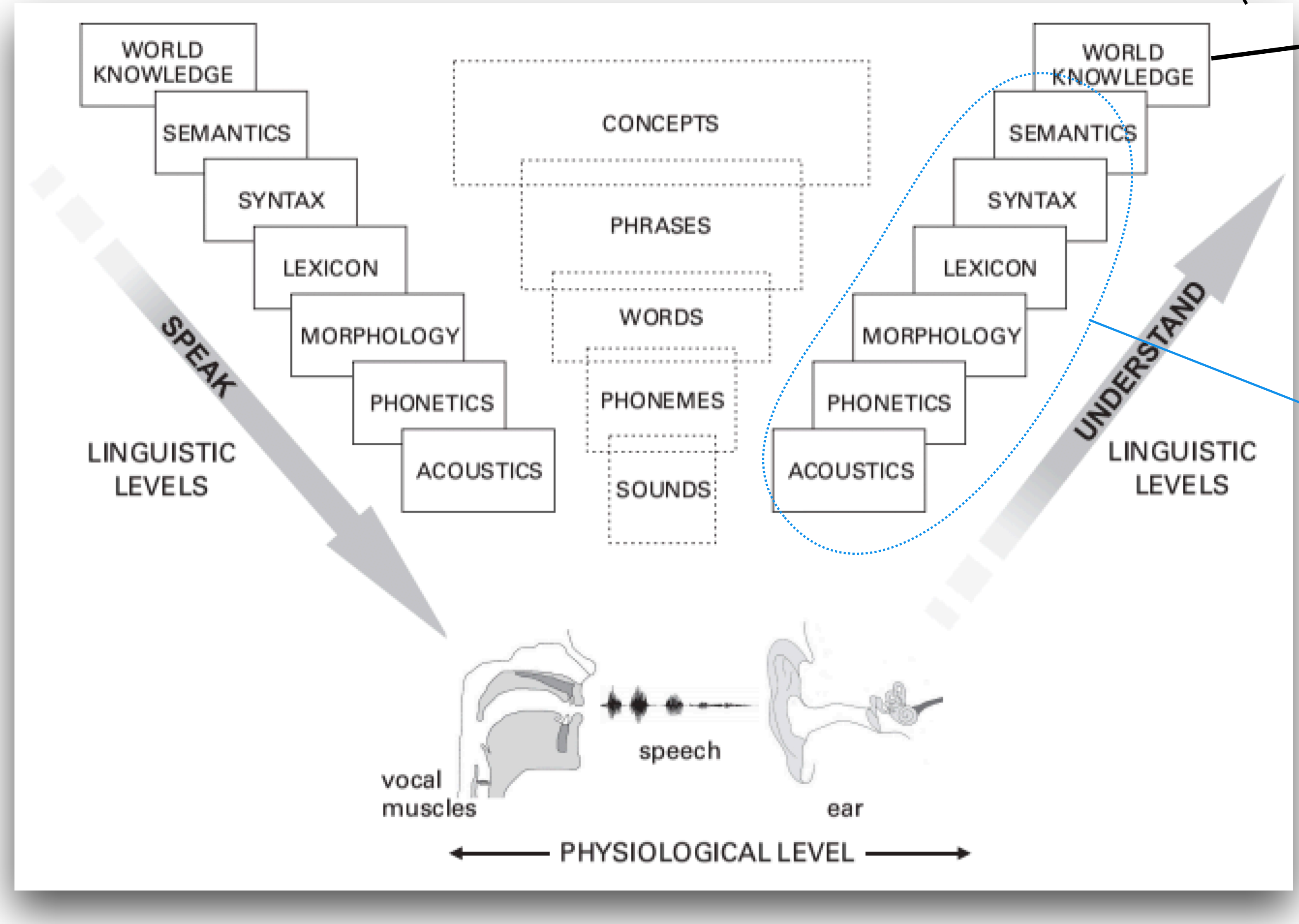
Location

An Italian what? Person?

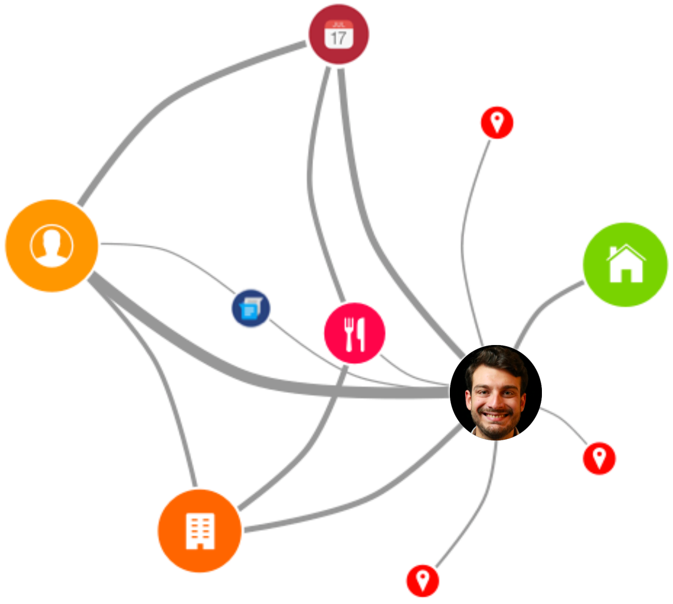
Good luck with this one

THE TECH BEHIND VOICE INTERFACES

WHY IS IT SO COMPLEX? (III)



(for example)

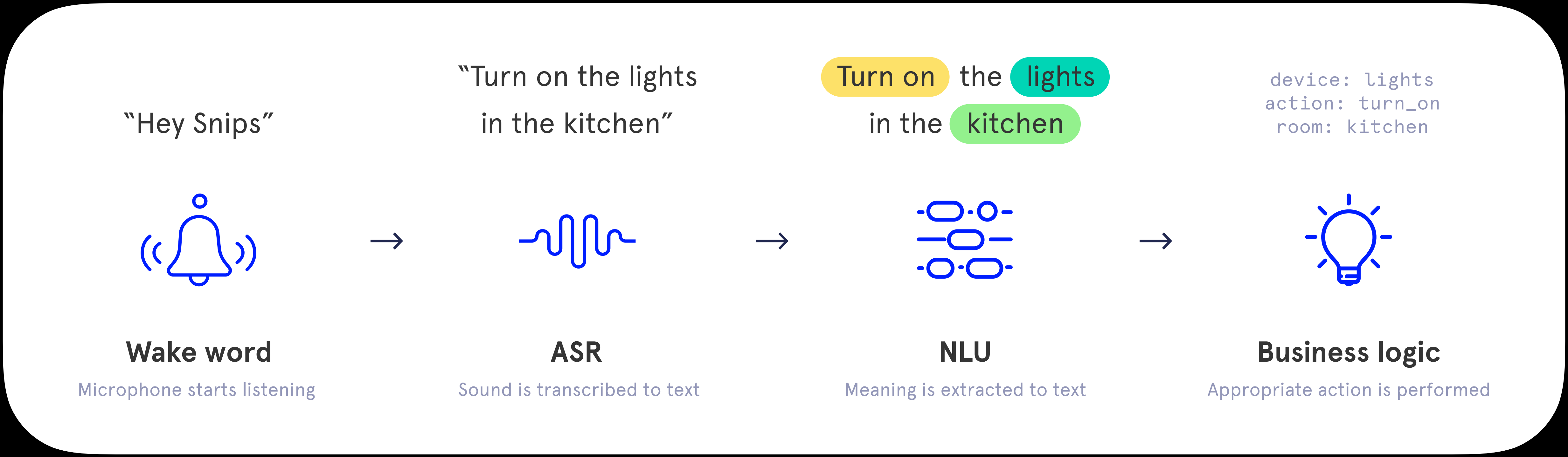


This is where we use machine learning

STRUCTURE OF A VOICE INTERFACE

STRUCTURE OF A VOICE INTERFACE

OVERALL STRUCTURE (I)



"Hey Snips"

Wake word

Microphone starts listening

→

"Turn on the lights in the kitchen"

ASR

Sound is transcribed to text

→

Turn on the lights in the kitchen

NLU

Meaning is extracted to text

→

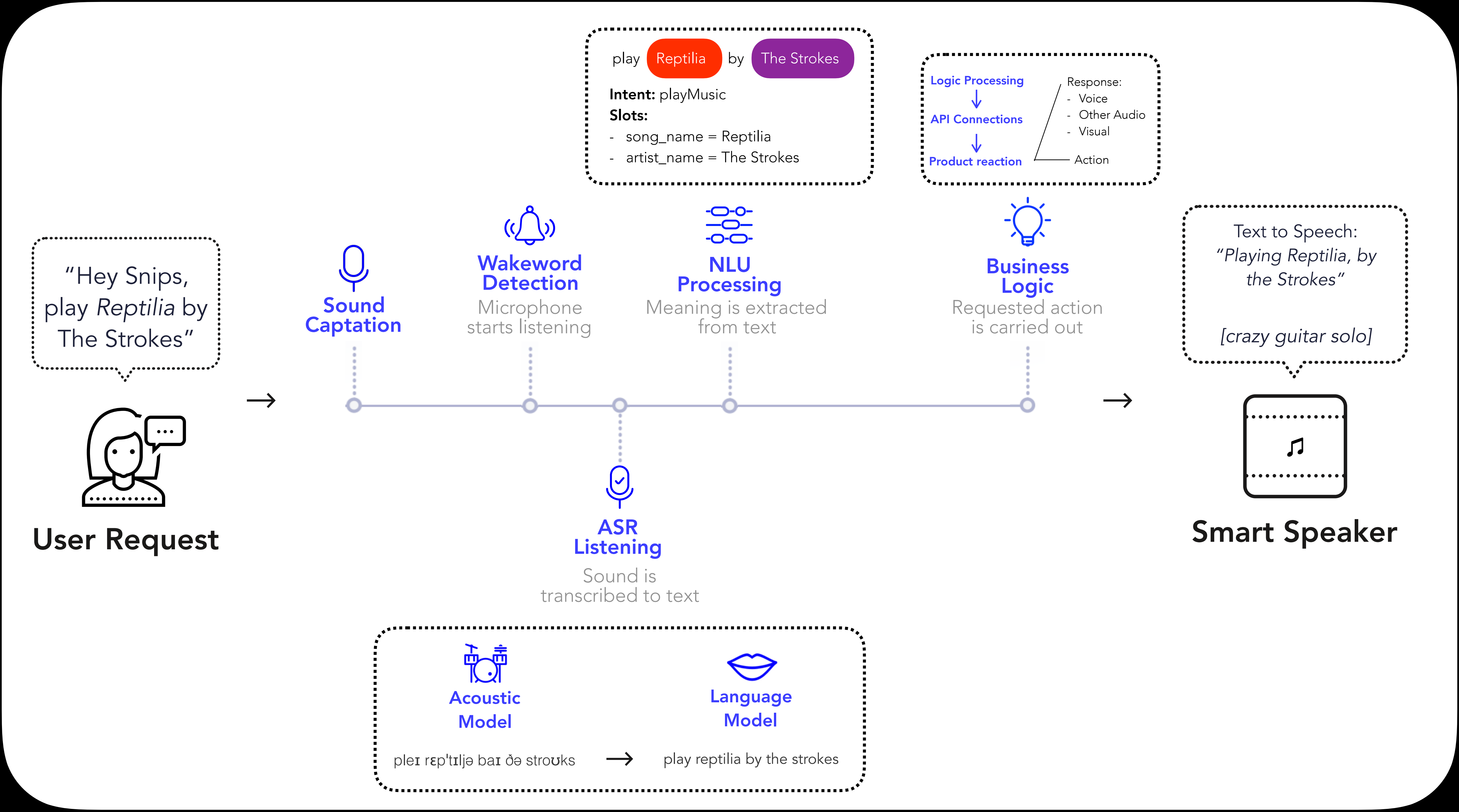
```
device: lights
action: turn_on
room: kitchen
```

Business logic

Appropriate action is performed

STRUCTURE OF A VOICE INTERFACE

OVERALL STRUCTURE (I)



"Hey Snips, play *Reptilia* by The Strokes"

User Request

→

Sound Captation

Wakeword Detection
Microphone starts listening

ASR Listening
Sound is transcribed to text

Acoustic Model

Language Model

pleɪ rɛp'tɪljə baɪ ðə stroʊks → play reptilia by the strokes

NLU Processing
Meaning is extracted from text

play Reptilia by The Strokes

Intent: playMusic
Slots:

- song_name = Reptilia
- artist_name = The Strokes

Business Logic
Requested action is carried out

Logic Processing
↓
API Connections
↓
Product reaction

Response:

- Voice
- Other Audio
- Visual

Action

→

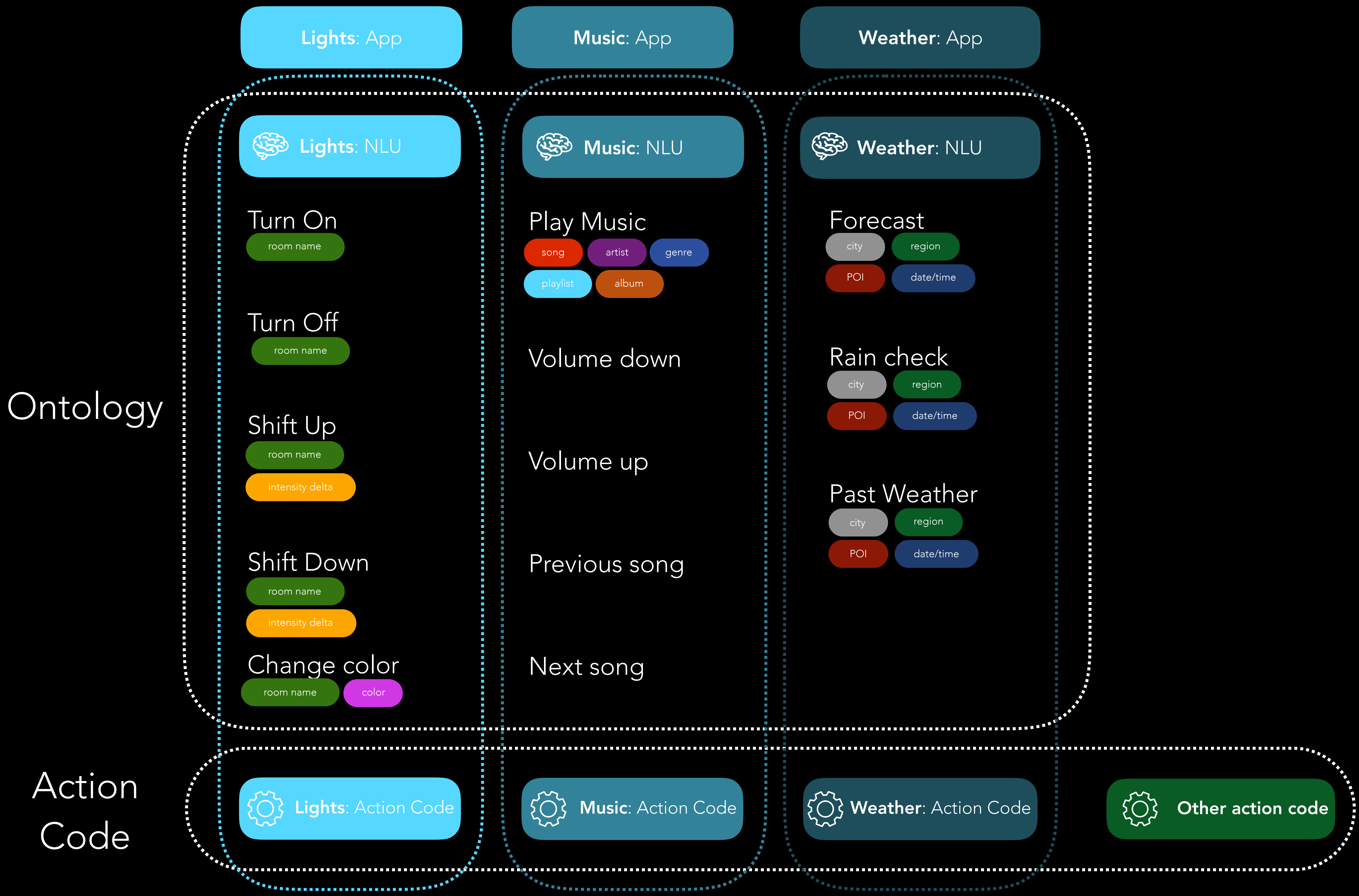
Text to Speech:
"Playing Reptilia, by the Strokes"

[crazy guitar solo]

Smart Speaker

STRUCTURE OF A VOICE INTERFACE

OVERALL STRUCTURE (I)



Lights: App

Music: App

Weather: App

Ontology

Lights: NLU

Turn On
room name

Turn Off
room name

Shift Up
room name
intensity delta

Shift Down
room name
intensity delta

Change color
room name
color

Music: NLU

Play Music
song
artist
genre
playlist
album

Volume down

Volume up

Previous song

Next song

Weather: NLU

Forecast
city
region
POI
date/time

Rain check
city
region
POI
date/time

Past Weather
city
region
POI
date/time

Action Code

Lights: Action Code

Music: Action Code

Weather: Action Code

Other action code

STRUCTURE OF A VOICE INTERFACE

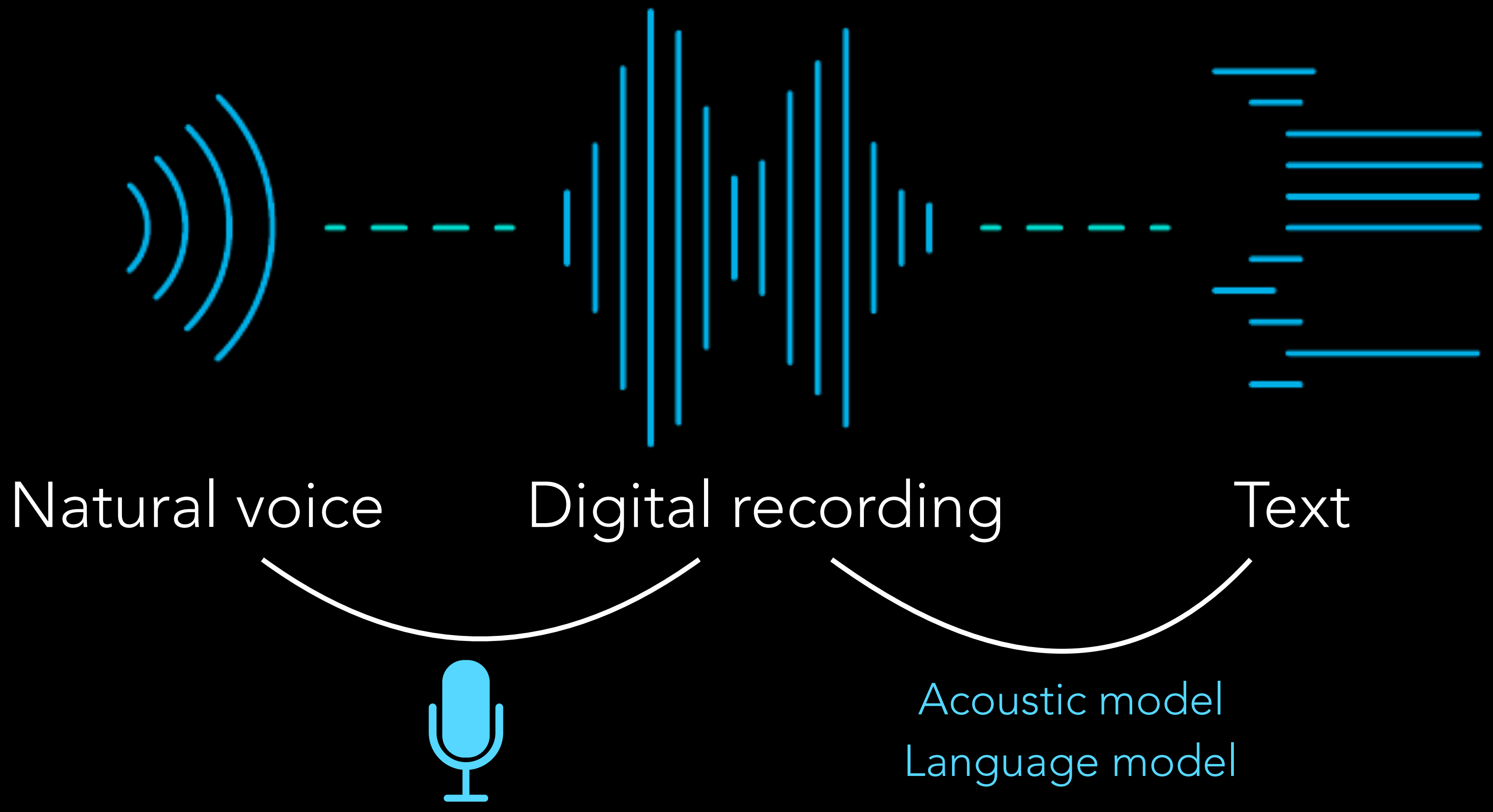
WAKE WORD

- To initiate the conversation
- Assistant constantly listening for it
- When detected, speech recognition is triggered

"Alexa"
"Ok Google"
"Hey Siri"

STRUCTURE OF A VOICE INTERFACE

AUTOMATIC SPEECH RECOGNITION (ASR)



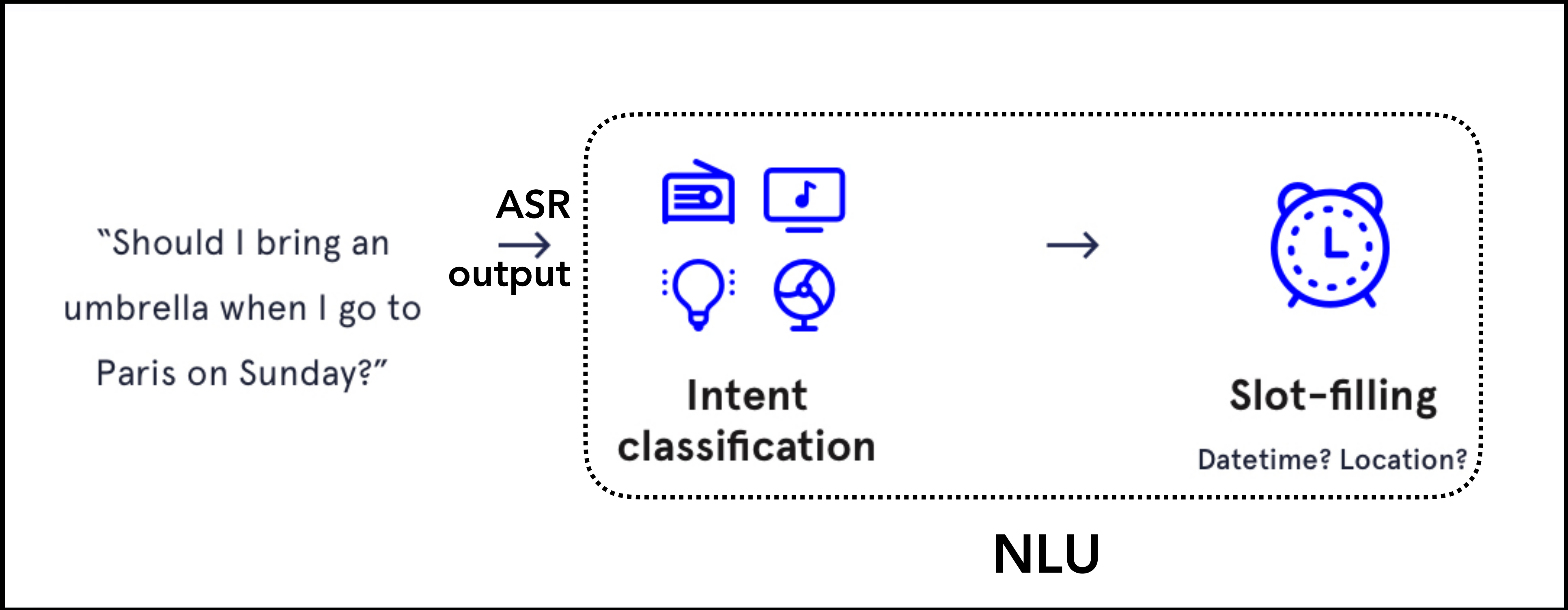
Natural voice

Digital recording

Text

Acoustic model
Language model

NATURAL LANGUAGE UNDERSTANDING (NLU)



What's the user's intent with their utterance?

Weather Check

What are the slots/parameters associated with their query?

City: **Paris**
Date/Time: **Sunday**

NLU - ONTOLOGY (I)

- *“A set of concepts and categories in a subject area or domain that shows their properties and the relations between them”*

- Structure to explain language to a machine

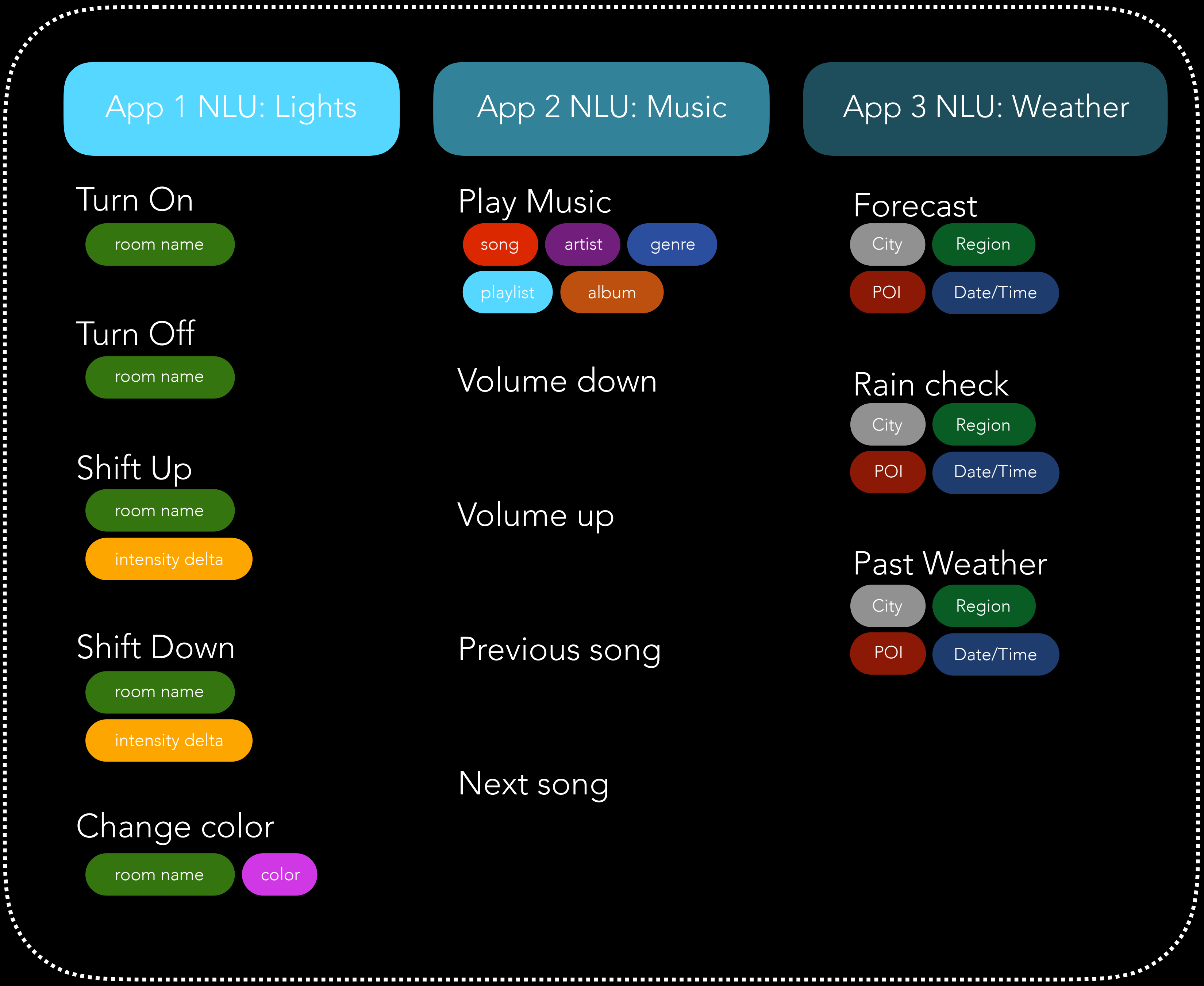
- Deals with questions like:
  - How many **domains** in total?
  - What **intents** and how many different ones?*
  - What **slots** per intent?
  - What **values** can a slot take?
  - How are sentences usually structured for an intent?

**More about this in next sections*

STRUCTURE OF A VOICE INTERFACE

NLU - ONTOLOGY (II)

Assistant ontology



App 1 NLU: Lights

Turn On
- room name

Turn Off
- room name

Shift Up
- room name
- intensity delta

Shift Down
- room name
- intensity delta

Change color
- room name
- color

App 2 NLU: Music

Play Music
- song
- artist
- genre
- playlist
- album

Volume down

Volume up

Previous song

Next song

App 3 NLU: Weather

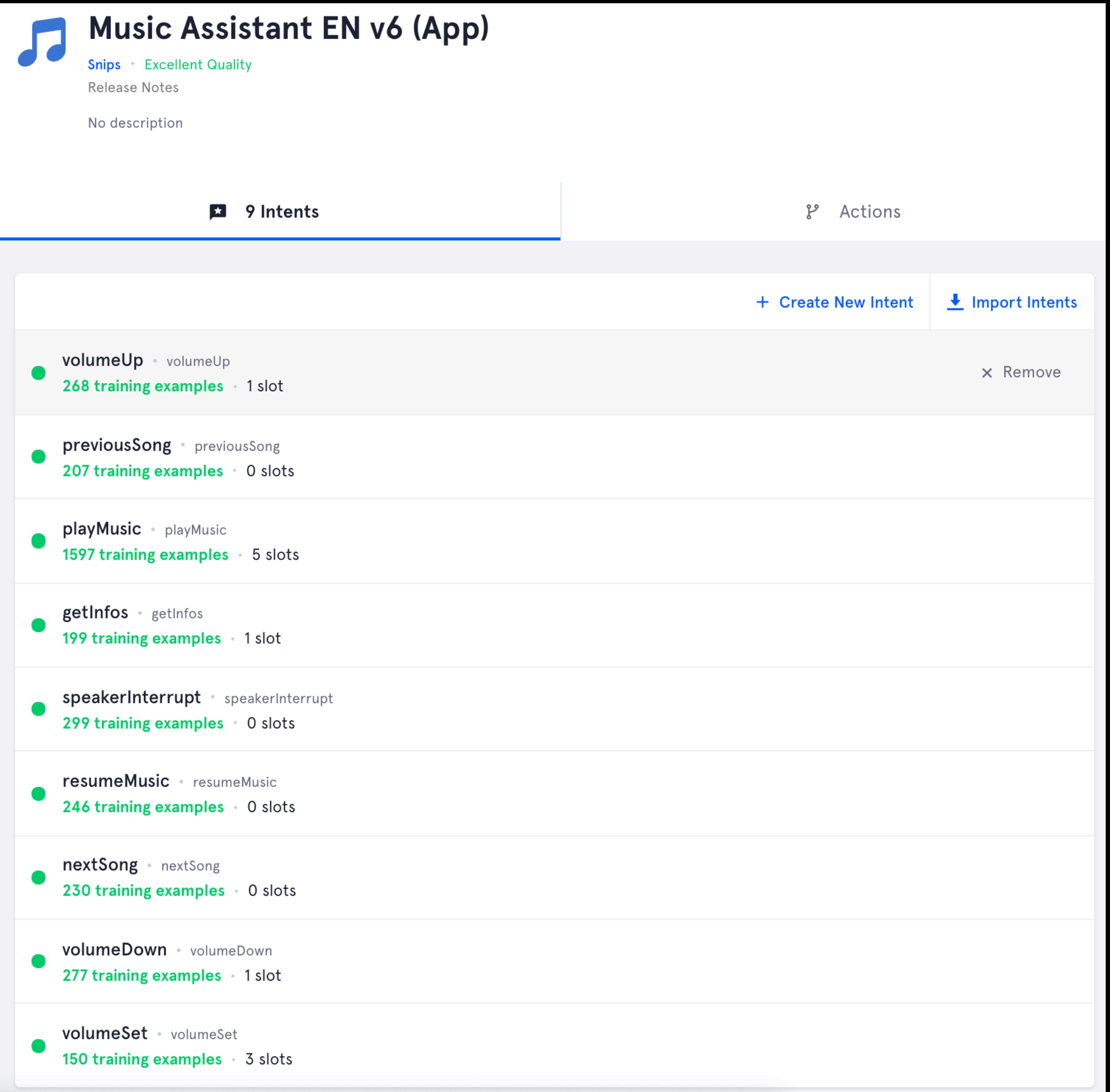
Forecast
- City
- Region
- POI
- Date/Time

Rain check
- City
- Region
- POI
- Date/Time

Past Weather
- City
- Region
- POI
- Date/Time

STRUCTURE OF A VOICE INTERFACE

NLU - ONTOLOGY (III)



Music Assistant EN v6 (App)

Snips · Excellent Quality

Release Notes

No description

● volumeUp · volumeUp
268 training examples · 1 slot

● previousSong · previousSong
207 training examples · 0 slots

● playMusic · playMusic
1597 training examples · 5 slots

● getInfos · getInfos
199 training examples · 1 slot

● speakerInterrupt · speakerInterrupt
299 training examples · 0 slots

● resumeMusic · resumeMusic
246 training examples · 0 slots

● nextSong · nextSong
230 training examples · 0 slots

● volumeDown · volumeDown
277 training examples · 1 slot

● volumeSet · volumeSet
150 training examples · 3 slots

STRUCTURE OF A VOICE INTERFACE

NLU - ONTOLOGY - DOMAINS

A **group of intents** within the **same topic**.

E.g.:

- **Light controls** (Turn on, Turn off, Change color…)
- **Music** (Play music, Volume down, Next song…)
- **Weather** (Forecast, Past weather, Rain check…)

STRUCTURE OF A VOICE INTERFACE

NLU - ONTOLOGY - INTENTS

Represents the **intention** behind a sentence

Can be **said in different ways**

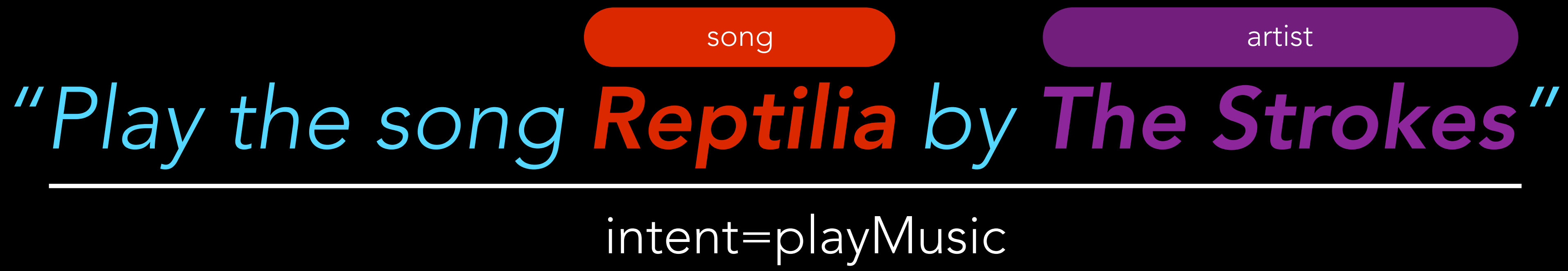
"It's dark in here" = *"Turn on the light"*

intent=lightsTurnOn   intent=lightsTurnOn

STRUCTURE OF A VOICE INTERFACE

NLU - ONTOLOGY - SLOTS

Slots: Parameters that add information to the intent.



song

artist

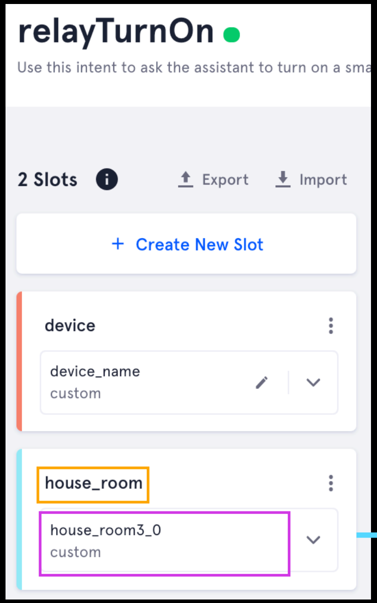
*"Play the song **Reptilia** by **The Strokes**"*

intent=playMusic

STRUCTURE OF A VOICE INTERFACE

NLU - ONTOLOGY - SLOTS (II)

Slots: Parameters that add information to the intent.



relayTurnOn ●

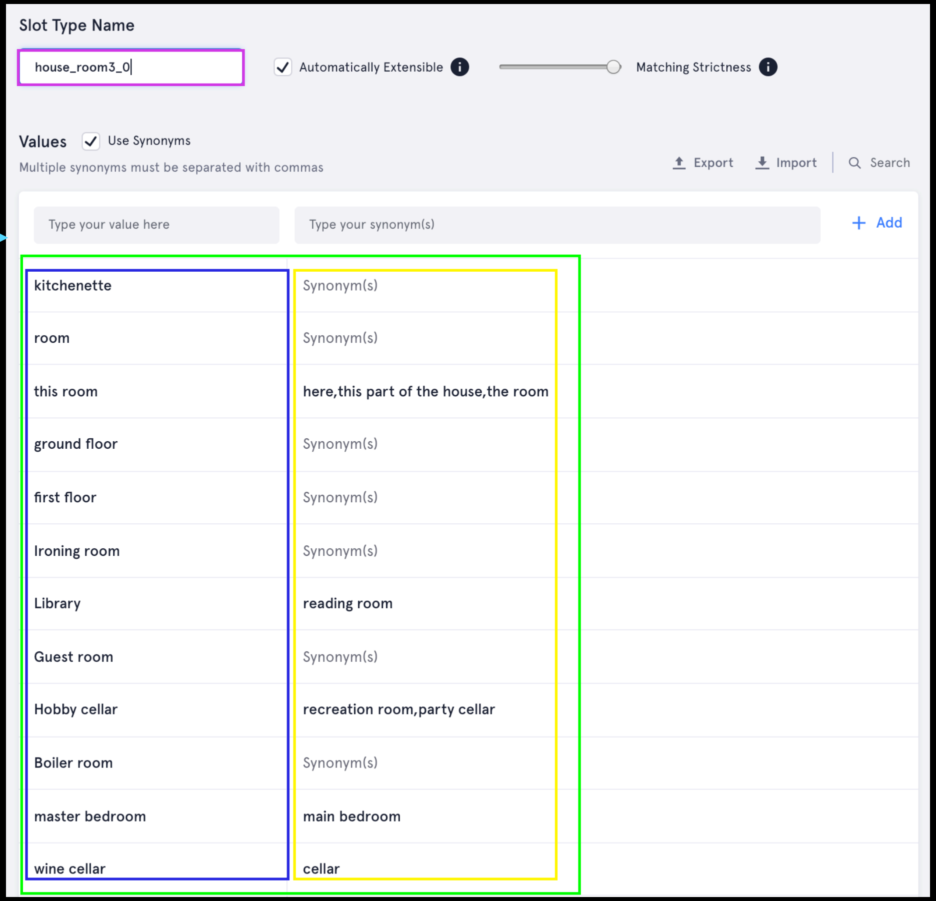
Use this intent to ask the assistant to turn on a sma

device

device_name
custom

house_room

house_room3_0
custom



Slot Type Name: house_room3_0

☑ Automatically Extensible

Matching Strictness

Values ☑ Use Synonyms

Multiple synonyms must be separated with commas

kitchenette	Synonym(s)
room	Synonym(s)
this room	here,this part of the house,the room
ground floor	Synonym(s)
first floor	Synonym(s)
Ironing room	Synonym(s)
Library	reading room
Guest room	Synonym(s)
Hobby cellar	recreation room,party cellar
Boiler room	Synonym(s)
master bedroom	main bedroom
wine cellar	cellar

Slot type:
The definition of a slot in terms of values it can recognise. Can be used in several intents.
(house_room_3_0)

Slot name:
The name a slot gets in a particular intent.
(house_room)

Slot value:
A value the system is able to recognise in a slot.

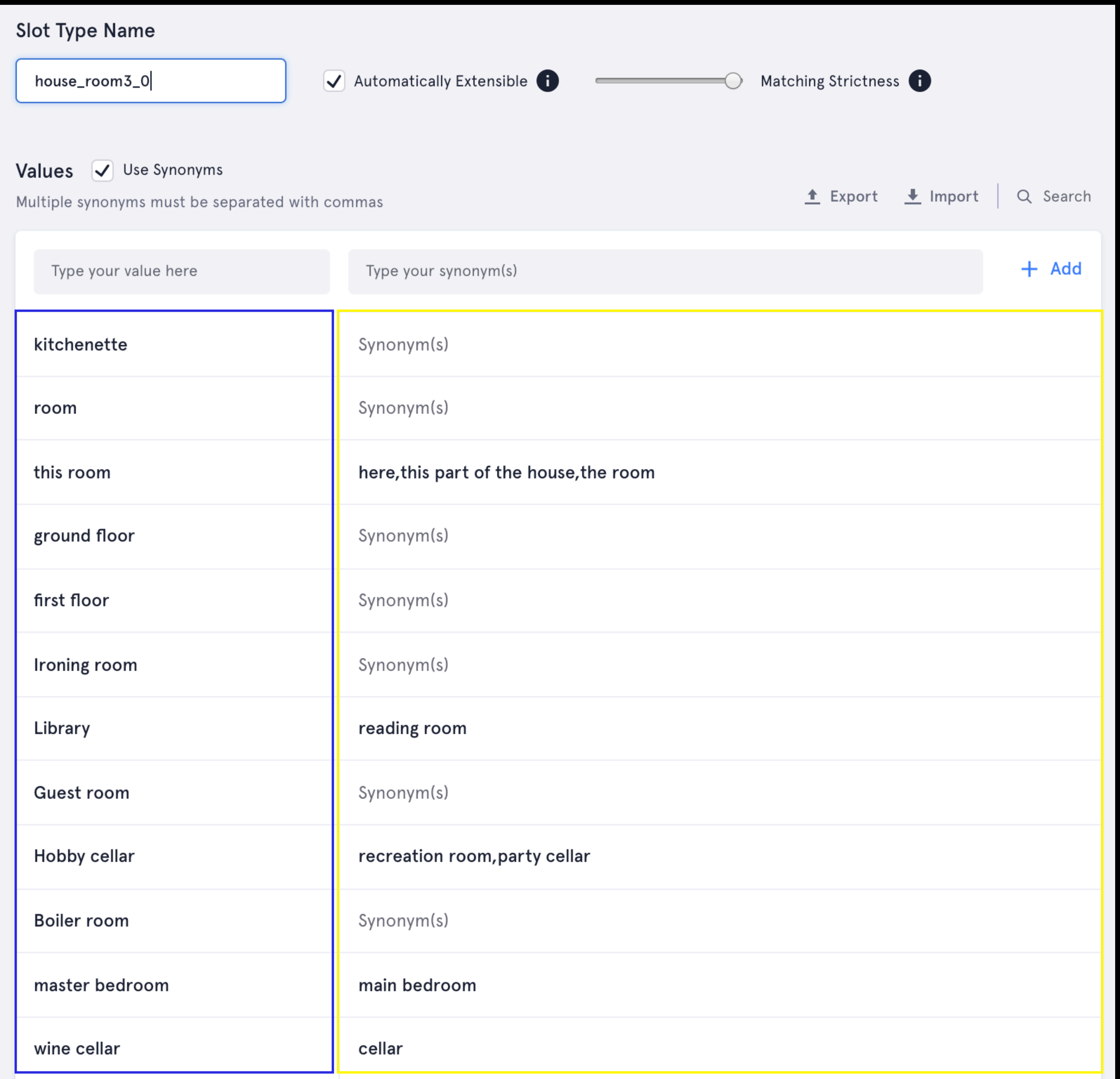
- **Main values:**
  The value that will be sent to the next step.
  ("kitchenette", "first floor", "library")

- **Synonyms:**
  Other values the system understands, and associates to main values.
  ("reading room")

STRUCTURE OF A VOICE INTERFACE

NLU - ONTOLOGY - SLOTS (III)

Synonyms: whichever gets picked up, the main one would be passed on to the next step



Slot Type Name

house_room3_0

☑ Automatically Extensible

Matching Strictness

Values ☑ Use Synonyms

Multiple synonyms must be separated with commas

kitchenette	Synonym(s)
room	Synonym(s)
this room	here,this part of the house,the room
ground floor	Synonym(s)
first floor	Synonym(s)
Ironing room	Synonym(s)
Library	reading room
Guest room	Synonym(s)
Hobby cellar	recreation room,party cellar
Boiler room	Synonym(s)
master bedroom	main bedroom
wine cellar	cellar

"Turn on the lights in ***this room****"*

Intent: lightsTurnOn

Slot name: house_room

Slot type: house_room3_0

Slot value: "**this room**"

"Turn on the lights ***here****"*

Intent: lightsTurnOn

Slot name: house_room

Slot type: house_room3_0

Slot value: "**this room**"

STRUCTURE OF A VOICE INTERFACE

NLU - ONTOLOGY - SLOTS (IV)

Extensibility/Free-text: the ability for the system to pick up slot values that haven't been specified, based on patterns picked up in examples.

Pros: richer list of slot values

Cons: problems down the line in the action code

Utterance:	***Recognised slot name and value:***
*"Turn on the lights in **this room**"*	*house_room/"**this room**"* (in list - main)
*"I want light in the **kitchen**"*	*house_room/"**kitchen**"* (in list - main)
*"It's too dark in the **reading room**"*	*house_room/"**library**"* (in list - synonym)
*"Turn the lights on in the **loo**"*	*house_room/"**loo**" (**not** in list, but looks like a slot value)*

NLU - ONTOLOGY - SLOTS (V)

Matching strictness: the ability for the system to pick up slots even if they're only partially present in the sentence (expressed in %)

Pros: easier to understand user
Cons: easier to **misunderstand** user

Utterance:
*"Play a **Beatles** song"*

***With** partial matching (50%):*
*artist_name/"**The Beatles**"* (in list - main)

***Without** partial matching:*
(slot not recognised)

NLU - ONTOLOGY - TRAINING EXAMPLES

Definition: Human-generated data to train the model. Base for system to find patterns and make inferences about intents and slots.

- **Use of words**
  - What slots are included in examples? How many per type and sentence?
  - What other words are used? How are they combined?

- **Formulation patterns**
  - Order of words
  - How are slots connected to rest of sentence and/or to other slots?

"I want the light in ***this room*** *turned up to* ***60%****"* VS *"I want to turn up the light in* ***this room*** *by* ***60%****"*

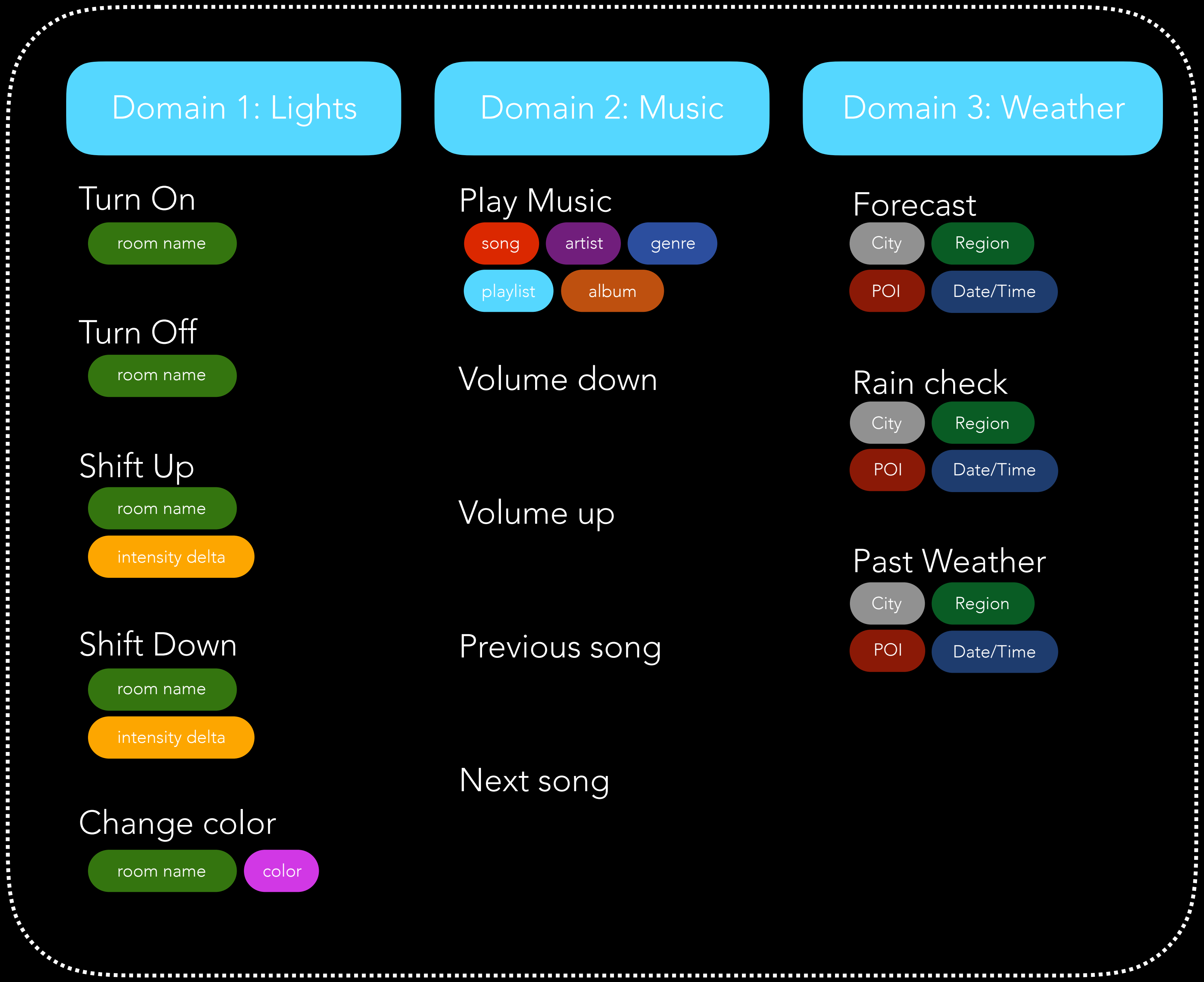
intent = **lightsSet**
house_room = "**this room**"
target_brightness = "**60%**"

intent = **lightsShiftUp**
house_room = "**this room**"
delta_brightness = "**60%**"

STRUCTURE OF A VOICE INTERFACE

NLU ONTOLOGY - SUMMARY

Assistant ontology



Domain 1: Lights

Turn On
- room name

Turn Off
- room name

Shift Up
- room name
- intensity delta

Shift Down
- room name
- intensity delta

Change color
- room name
- color

Domain 2: Music

Play Music
- song
- artist
- genre
- playlist
- album

Volume down

Volume up

Previous song

Next song

Domain 3: Weather

Forecast
- City
- Region
- POI
- Date/Time

Rain check
- City
- Region
- POI
- Date/Time

Past Weather
- City
- Region
- POI
- Date/Time

STRUCTURE OF A VOICE INTERFACE

DIALOGS

Definition: Any response by the assistant requesting further interaction from the user's side.

Two types:

- Follow-up / slot-filling dialogs

  User: *"Hey Snips, turn off the lights"*
  Assistant: *"Sorry, for what room?"*
  User: *"The kitchen"*
  Assistant: *"Ok."*

  This can be added as an extension of the NLU

- Conversational / multi-turn dialogs

  User: *"Hey Snips, start the lesson with 4 questions on table 9"*
  Assistant: *"What's 5 x 9?"*
  User: *"Hmm the answer is 45 isn't it?"*
  Assistant: *"That's it, well done ! What's 4 x 9 ?"*
  User: *"I think it's 33"*
  Assistant: *"Oh no, wrong answer, 4 x 9 = 36. What's 3 x 9 ?"*
  User: *"I don't want to play anymore"*
  Assistant: *"Ok, the game is over."*

  This is part of the UX/action code of this specific app

STRUCTURE OF A VOICE INTERFACE

ACTION CODE

Definition: What we do after understanding the user

Some examples:

- Playing music and announcing what's playing
- Fetching info from the web and responding via synthesised voice
- Turning on the lights
- Increasing the volume and showing a visual cue
- Recognising that we didn't understand the command

STRUCTURE OF A VOICE INTERFACE

ACTION CODE (II)

Usually split in a few parts:

1. Logic processing
2. API connection

(I won't bore you with details though)

3. **Product reaction** — *(Let's focus on this)*

STRUCTURE OF A VOICE INTERFACE

PRODUCT REACTION

What can a device (visibly) do?

- **Assistant device:**
  - Lights
  - Sound cues
  - Verbal response (talk, but also whisper and sing)
  - Other sound (music, video)
  - Show something on a screen

- **3rd party devices** controlled by assistant:
  - Turn on/off
  - Change color
  - Open/close
  - Vibrate
  - **Basically anything, really.**

STRUCTURE OF A VOICE INTERFACE

ACTION CODE - VERBAL RESPONSE

- Part of the action code
- Useful to give answers (hands- and eyes-free!)
- Recorded or synthesised

VERBAL RESPONSE - TEXT-TO-SPEECH (TTS)

- Artificial production of human speech
- Responses are preprogrammed in text
- TTS provider (don’t reinvent the wheel)

WHAT IS A VOICE INTERFACE

DEFINITION (REMINDER II)

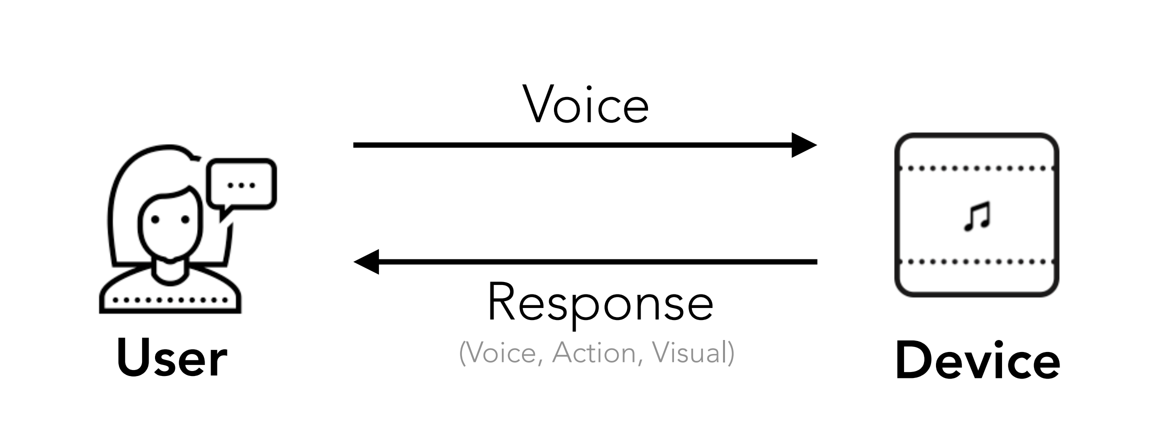
A **voice-user interface (VUI)**:

allows for **spoken interaction** with computers using **speech recognition** to understand spoken commands/questions,

and typically **text to speech** to play a reply,

and/or other **actions,**

and/or **visual responses**.



PRODUCT REACTION

Examples

(a bit simplified)

WHAT IS A VOICE INTERFACE

PRODUCT REACTION / EXAMPLES (I)

"Alexa, […]

[…] what's the weather in Paris?"

Listening mode:
[beeps, blue ring points at user]

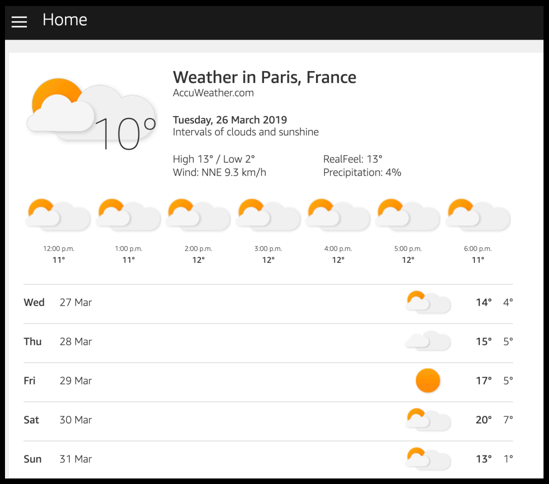
Wake word feedback beep
Simple visual

Processing mode:
[blue ring spins]

Simple visual

"Right now in Paris it's 10°C with mostly cloudy skies […]"

TTS



Home

Weather in Paris, France
AccuWeather.com

10°

Tuesday, 26 March 2019
Intervals of clouds and sunshine

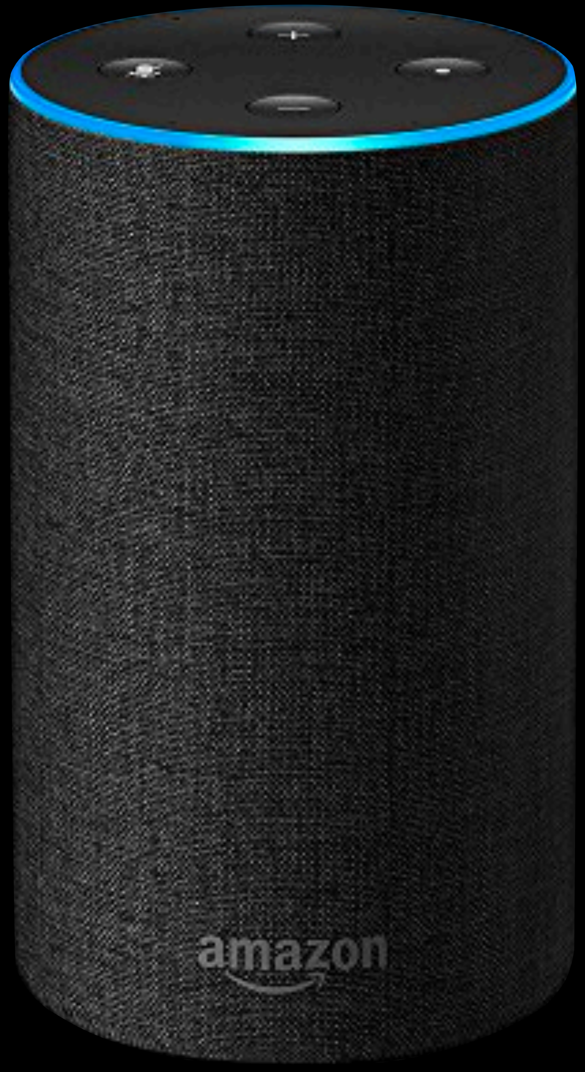
High 13° / Low 2°
Wind: NNE 9.3 km/h

RealFeel: 13°
Precipitation: 4%

12:00 p.m.	1:00 p.m.	2:00 p.m.	3:00 p.m.	4:00 p.m.	5:00 p.m.	6:00 p.m.
11°	11°	12°	12°	12°	12°	11°

Wed	27 Mar	14°	4°
Thu	28 Mar	15°	5°
Fri	29 Mar	17°	5°
Sat	30 Mar	20°	7°
Sun	31 Mar	13°	1°

Rich visual
(app, browser or device screen)



WHAT IS A VOICE INTERFACE

PRODUCT REACTION / EXAMPLES (II)

"Alexa, […]

[…] turn on the living room lights'

Listening mode:
[beeps, blue ring points at user]

Wake word feedback beep
Simple visual

Processing mode:
[blue ring spins]

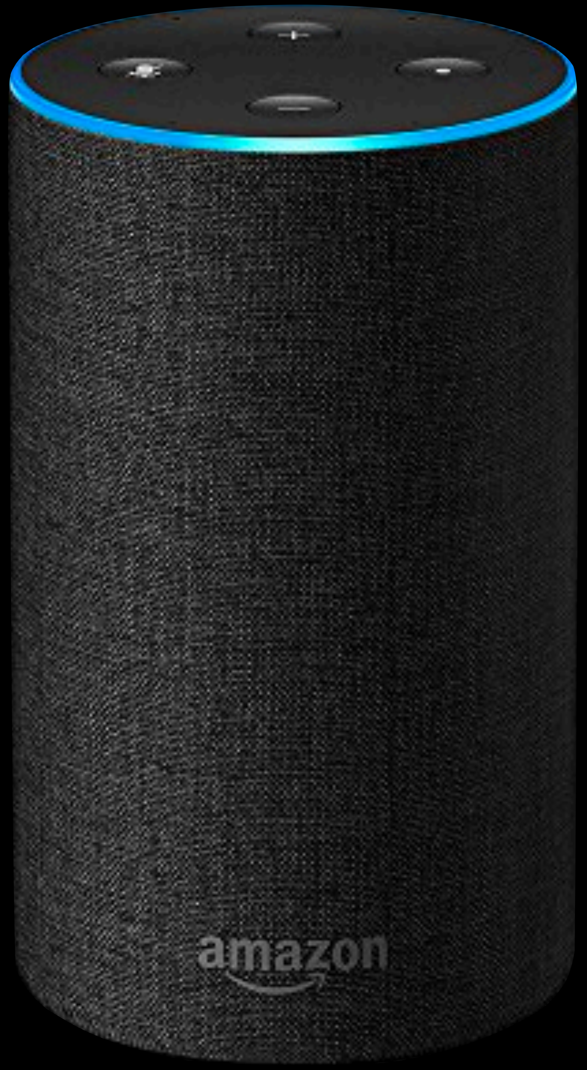
Simple visual

[Philips HUE bulb turns on]

3rd party hardware trigger

"OK"

TTS



WHAT IS A VOICE INTERFACE

PRODUCT REACTION / EXAMPLES (III)

"Alexa, […]

Listening mode:
[beeps, blue ring points at user]

Wake word feedback beep
Simple visual

[…] play music by The Strokes"

Processing mode:
[blue ring spins]

Simple visual

"Shuffling songs by The Strokes, from Spotify."

TTS

[Plays songs by The Strokes in order of popularity]

Sound from 3rd party service

"Alexa, […]

Listening mode:
[beeps, blue ring points at user]

Wake word feedback beep
Simple visual

[…] it's too loud"

Processing mode:
[blue ring spins]

Simple visual

*[Lowers volume,
light circle on top of device shows decrease]*

Device setting change

Simple visual

PRODUCT REACTION / EXAMPLES (IV)

"Alexa, […]

Listening mode:
[beeps, blue ring points at user]

Wake word feedback beep
Simple visual

[…], set a reminder"

Processing mode:
[beeps, blue ring spins]

Simple visual

What's the reminder for?

TTS

Listening mode:
[beeps, blue ring points at user]

Simple visual

"Buy cookies"

Processing mode:
[beeps, blue ring spins]

Simple visual

When should I remind you?

TTS

Listening mode:
[beeps, blue ring points at user]

Simple visual

"Tomorrow at 10 a.m."

Ok, I will remind you tomorrow at 10 a.m.

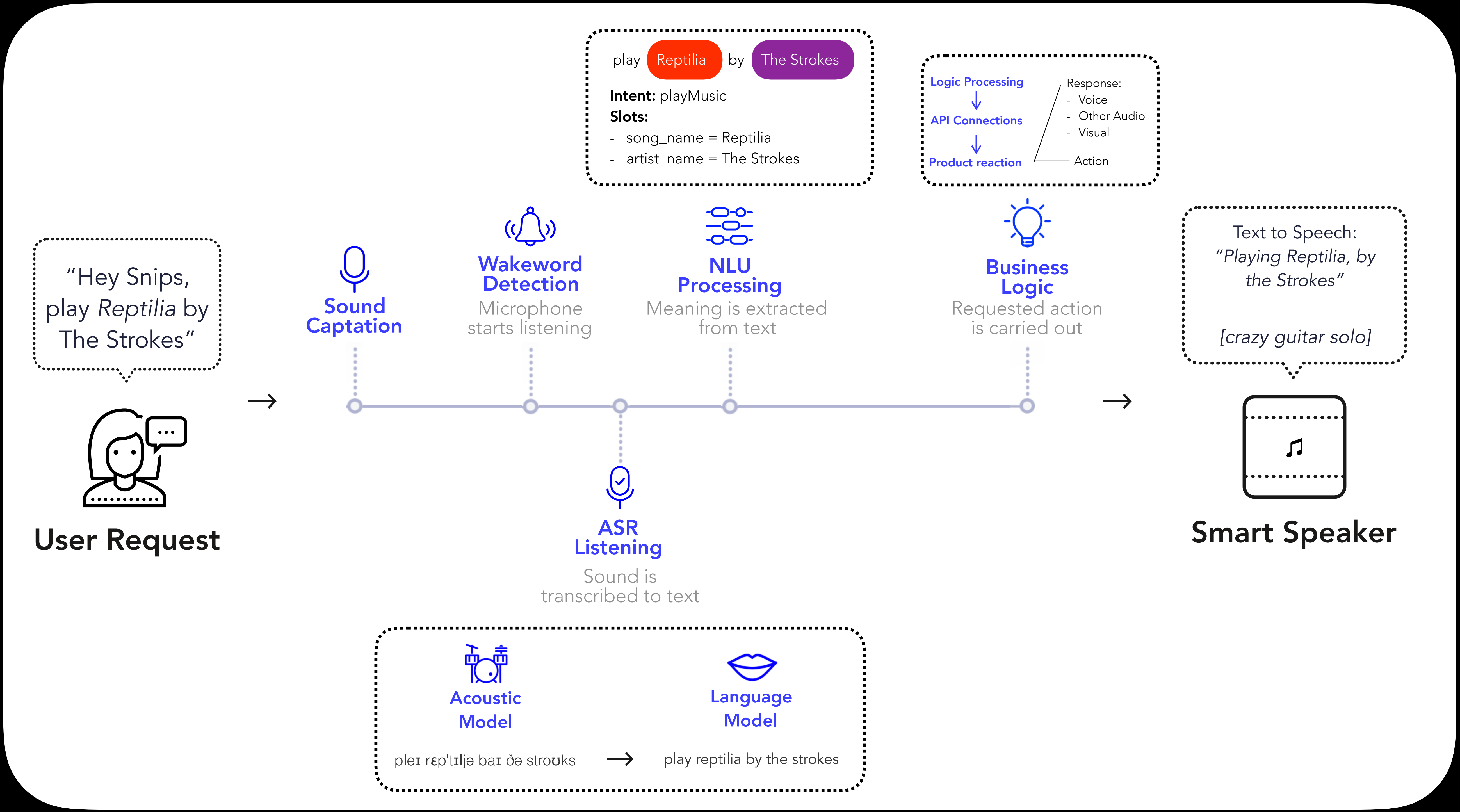
TTS

Processing mode:
[beeps, blue ring spins]

Simple visual

STRUCTURE OF A VOICE INTERFACE

OVERALL STRUCTURE (REMINDER)



"Hey Snips, play *Reptilia* by The Strokes"

User Request

→

Sound Captation

Wakeword Detection
Microphone starts listening

ASR Listening
Sound is transcribed to text

Acoustic Model

pleɪ rɛp'tɪljə baɪ ðə stroʊks

→

Language Model

play reptilia by the strokes

NLU Processing
Meaning is extracted from text

play Reptilia by The Strokes

Intent: playMusic
Slots:
- song_name = Reptilia
- artist_name = The Strokes

Business Logic
Requested action is carried out

Logic Processing
↓
API Connections
↓
Product reaction

Response:
- Voice
- Other Audio
- Visual

Action

→

Text to Speech:
"Playing Reptilia, by the Strokes"

[crazy guitar solo]

♫

Smart Speaker

PART 4:
DESIGNING FOR VOICE

MAIN CHARACTERISTICS OF VOICE

MAIN CHARACTERISTICS OF VOICE

HOW DO WE PROCESS INFORMATION?

- **Short auditive memory**. Humans don’t like long lists or a lot of auditive content, we get overwhelmed (that’s why numbers get repeated)

- **We hesitate when we speak**. Machines struggle with this (is this silence the end of it?)

- **Lack of thoughtfulness.** We say the first thing that comes to mind (vs writing).

- **Uncertainty about duration.** For how long is this assistant(/human) going to talk?

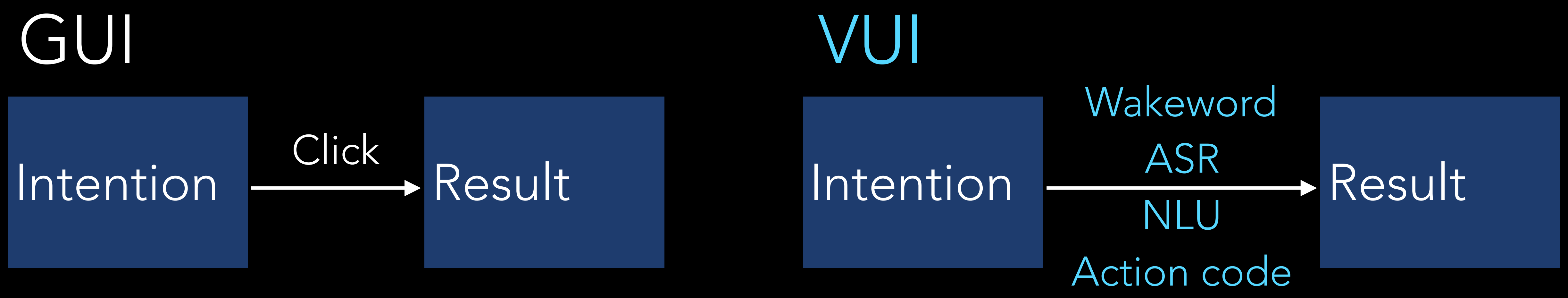
- **Uncertainty about expected input**. What am I supposed to respond to this assistant(/human)?

MAIN CHARACTERISTICS OF VOICE

WHAT IS VOICE GOOD FOR?

Main difference with other interfaces:

There are additional steps between the user decision and the trigger of the action: understanding!



GUI

Intention → Click → Result

VUI

Intention → Wakeword, ASR, NLU, Action code → Result

WHEN TO USE VOICE (VS NOT)

WHAT IS VOICE GOOD FOR?

1. It’s the most natural form of communication we know

You don’t need to learn how to communicate to navigate a UI.
No new languages or rules. Just speak!

Assistants should learn how users speak (and not the other way around)

WHAT IS VOICE GOOD FOR?

2. Voice reduces the effort to navigate through options

Visual interfaces can be great for some stuff,
but sometimes you can do more with voice, faster and with less friction

WHEN TO USE VOICE (VS NOT)

WHAT IS VOICE GOOD FOR?

3. You can add features without adding complexity

Keep on adding things your users will find useful, without them having to learn new stuff all the time.

WHEN TO USE VOICE (VS NOT)

WHAT IS VOICE GOOD FOR?

4. Eyes- and hands-free experience

Visual-only interfaces require users to perform tasks with hands and eyes.

Probably not the best idea while cooking, running or driving.

WHAT IS VOICE GOOD FOR?

5. Voice interactions can be shared

When interacting to a screen, it's the screen and you.

With voice, multiple people can take part on the request and the response.

WHAT DOES VOICE STRUGGLE WITH

1. Bad for public places

You don't want to check your bank account balance in the bus.

Or an office full of people talking to machines all the time.

Plus, recognition suffers with background noise.

Solution:

Choose your fights carefully

WHEN TO USE VOICE (VS NOT)

WHAT DOES VOICE STRUGGLE WITH

2. Feature limitations are not evident/visible

Unlike in GUIs, users might not know what works… until they try.

Users won’t know why something is failing (lots of layers involved)

Solution:

Be explicit and proactive about the limitations of your VUI

WHAT DOES VOICE STRUGGLE WITH

3. Too much voice creates a cognitive overload

Giving too much info might be overwhelming.

Consumption of large amounts of info is done best with your eyes.

With your eyes, you can:check highlights, browse, compare, re-read…

Solution:

Answer succinctly. Accompany by visual media where possible.

WHAT DOES VOICE STRUGGLE WITH

4. Voice is ambiguous

Different intonations, context, cultural differences
can make a word mean different things.

Solution:
Keep responses neutral & useful until tech figures out emotions.

VUI DESIGN GUIDELINES AND TIPS

GENERAL DESIGN GUIDELINES

Be transparent about its possibilities and limitations

- **You're not designing to pass the Turing test**, but to help users with tasks. Don't overpromise. A few failed interactions and they'll never use your VUI again.

- **It's more difficult to know what can be done and what can't. Try and be explicit about it** (e.g. in VUI intro, in skill description, in help pages…)

GENERAL DESIGN GUIDELINES

Don’t think about features, think about tasks

- Be clear about what **success means** for your users’ **task at hand.** It doesn’t have to mean presenting all the information, but the appropriate one.

- Many times, even if outcome isn’t perfect, it does help the client. Plan your fallbacks smartly.

GENERAL DESIGN GUIDELINES

Think more broadly than the happy path you designed on the whiteboard.

- **Think of the most usual scenarios, then think again.** Your voice interface will be on its own, so plan ahead so it'll be ready.

- **Prepare for users saying things you didn't expect,** either as a starting command or intermediate responses.

- Prepare 'gracious fallbacks' for unexpected turns in the conversation.

- Test, test, test!

GENERAL DESIGN GUIDELINES

Be aware of the benefits of voice… and its challenges

- Don’t replicate a GUI into a VUI.

- Avoid tasks with complex input and high ambiguity. Don’t try to create a voice application controlling MS Excel. Instead, focus on the task that a user might want to solve with such a program.

GENERAL DESIGN GUIDELINES

Customise the interaction to your user

- **Randomise responses.** We prefer consistent GUIs but have low tolerance for repetitive voice interactions.

- In some cases your user will use the same feature hundreds of times - don’t give the same level of intro information each time.

- **Use your knowledge about the user** to offer relevant experiences (music taste, usual choice of pizza, audiobook chapter)

VUI DESIGN GUIDELINES AND TIPS

GENERAL DESIGN GUIDELINES

Mimic human-to-human conversations

- **Show understanding**. Anchor sentences to previous statements without explicitly citing words.

- **Let users talk like a person** and do not force them into reciting a narrow set of rigidly defined prompts.

- Assume that your user will **answer more than asked for**, or something completely unrelated. Save all provided info to answer future questions in the session.

GENERAL DESIGN GUIDELINES

Don't think voice-only, think voice-first

- Many devices already incorporate screens and other GUIs
- When you develop your app, think of how other circumstances surrounding the user might help them achieve the task at hand.

GENERAL DESIGN GUIDELINES

Use other cues (visual, sound) to show the user what's happening

- Many devices have screens or LED lights - make use of them (different colors or intensities or direction/speed of movement) to tell the user when to talk, when you're processing, etc.
- Many times, you won't have any visual resources (or the user will be looking elsewhere), so also think about audio cues.

USING VOICE INPUT

USING VOICE INPUT

Do your best with what you have

Avoid correcting parts of sentences that weren't recognised or solving every ambiguity.
Go with what you have and plan to deal with unexpected responses.
Good practice: go for useful default or personalised responses.

Do

"Alexa, play some music"

"Resuming music from Spotify"

or

"Alexa, play some music"

"Here's a playlist you might like: Indie-Rock from the 2000s"

Don't

"Alexa, play some music"

"What music do you want to hear?"

"Whatever works!"

"I don't find any music with the name 'whatever'. Please try again."

"OK, play some rock"

"We have more than 14,000 playlists with the word 'rock' in it. Do you want classic rock, indie rock, progressive rock, …"

"Alexa, shut the f**k up"

STRUCTURE OF A VOICE INTERFACE

USING VOICE INPUT

Kill those menus

While in GUIs you want everything to follow a hierarchical structure…
… in VUIs you want the exact opposite. You don't want to navigate through endless options.
Let people access the end of those trees' branches straight with one command.

Do

"Alexa, I want pizza"

"Would that be a middle-sized Carbonara and a can of Coke from Telepizza, like your last order?"

"Yep, that works"

"The price is 14.95€ and the estimated delivery date is in 25 minutes, to your home address. Can you confirm?"

"Yes"

"Great. The order should be ready soon. You can follow the progress or your order live on your UberEATS app"

Don't

"Alexa, I want pizza"

"What provider do you want it from?"

"Telepizza"

"OK. What pizza do you want?"

"I want a Carbonara."

"OK, what size?"

"Medium"

"Any additional ingredient?

"Pepperoni and bacon"

"Sorry, I don't find an ingredient called 'pepperoni and bacon' for this vendor. Try again"

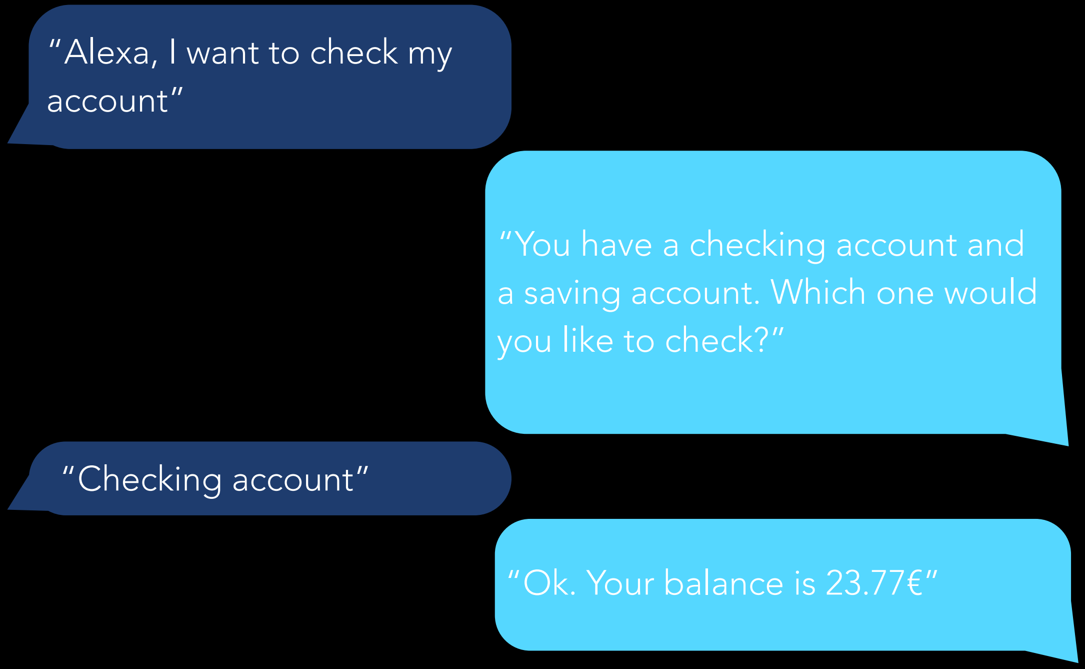
[Throws device out of the window]

USING VOICE INPUT

Leave the prompt for the end

If you ask the question and then give the options, you'll have the users jump in the middle of the sentence to answer. And you won't hear it.

Do



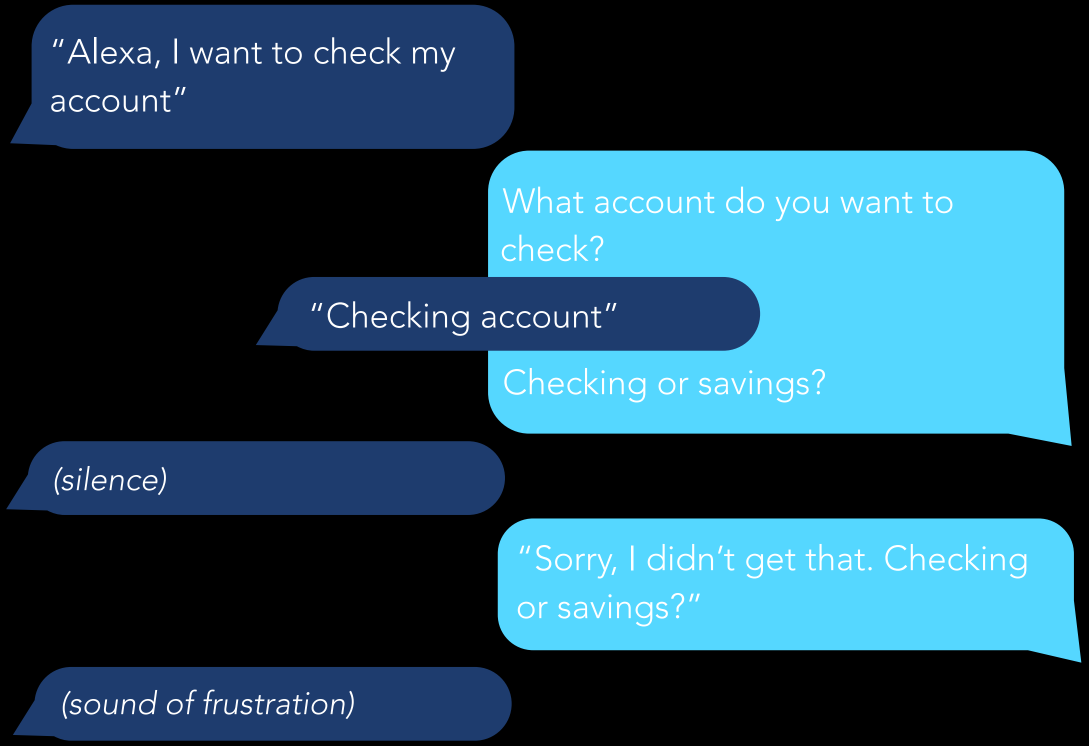
"Alexa, I want to check my account"

"You have a checking account and a saving account. Which one would you like to check?"

"Checking account"

"Ok. Your balance is 23.77€"

Don't



"Alexa, I want to check my account"

What account do you want to check?

"Checking account"

Checking or savings?

(silence)

"Sorry, I didn't get that. Checking or savings?"

(sound of frustration)

USING VOICE INPUT

Offer users a well-defined, simple set of options

Many times, users won't know the available options, so make them aware of them.
Other times, letting them know what information or format you need will reduce churn.

Do

"Alexa, I want to check my account"

"You have a checking account and a saving account. Which one would you like to check?"

"Checking account"

"Ok. Your balance is 23.77€"

Don't

"Alexa, I want to check my account"

"What account do you want to check?"

"Ehm. I don't know, MY account?"

"Sorry, there is no account called 'my account'. Please try again"

(gunshots)

VUI DESIGN GUIDELINES AND TIPS

USING VOICE INPUT

When possible, allow for barge-in

Many times, you'll screw up with your voice interaction, and the user will not want to hear an entire non-relevant message.

Allow the user to catch your attention again and try again mid-sentence.

DESIGNING VOICE OUTPUT

VUI DESIGN GUIDELINES AND TIPS

DESIGNING VOICE OUTPUT

Be concise

Avoid overwhelming users with information. Offer a preview and give additional options (voice or visual) for further information.

Do

"Alexa, what are some of the U.S. presidents?"

"Some of the recent U.S. Presidents include Donald Trump, Barack Obama and George W. Bush. Do you want to hear more?"

"Yes"

"Ok, there was also Bill Clinton, George Bush…"

Don't

"Alexa, what are some of the U.S. presidents?"

Until today, 44 people have won into office as President of the United States. From last to first, there is Donald Trump, Barack Obama, George W. Bush, Bill Clinton, George Bush, Ronald Reagan, Jimmy Carter, Gerald Ford, Richard Nixon, Lyndon B. Johnson, John F. Kennedy, Dwight D. Eisenhower…"

Alexa, stop"

VUI DESIGN GUIDELINES AND TIPS

DESIGNING VOICE OUTPUT

Default to simple, functional responses for ambiguous/ generic queries.

Do

"Alexa, play some music"

"Resuming music from Spotify"

or

"Alexa, play some music"

"Here's a playlist you might like: Indie-Rock from the 2000s"

Don't

"Alexa, play some music"

"What music do you want to hear?"

"Whatever works!"

"I don't find any music with the name 'whatever'. Please try again."

"OK, play some rock"

"We have more than 14,000 playlists with the word 'rock' in it. Do you want classic rock, indie rock, progressive rock, …"

"Alexa, shut the f**k up"

DESIGNING VOICE OUTPUT

In lists, mention their amount of items first

Especially when users ask about the quantity, but also when they ask for the items in a list specifically, announce how many you're going to mention and ask if they're interested. Restrict number of items in individual enumerations to 2-5 and check back on interest.

Do

"Alexa, do I still have many things on my shopping list?"

"There are 19 items in your shopping list. Do you want to hear them?"

"Yes"

"There are: onions, tomatoes, pasta. Do you want to hear more?"

"No, that'd do for the moment"

Don't

"Alexa, do I still have many things on my shopping list?"

"The following items are in your shopping list: onions, tomatoes, pasta, soy sauce, surface cleaning product, toilet paper…"

"Alexa, stop"

DESIGNING VOICE OUTPUT

Whenever you use visuals, explain what they are

Don't just put out the visual material there and expect the user to know what they're seeing. They might not even see it at the time you propose them to.

Do

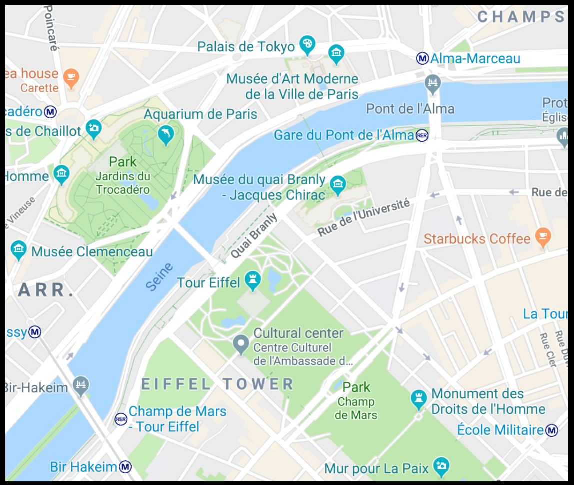
"Alexa, where's the Eiffel Tower"

"The Eiffel Tower is located in the 7th arrondissement of Paris, France."

Eiffel Tower
Landmark 324m-high 19th-century tower
Palais de Tokyo
Alma-Marceau
Musée d'Art Moderne de la Ville de Paris
Tea house Carette
Trocadéro
Palais de Chaillot
Aquarium de Paris
Pont de l'Alma
Gare du Pont de l'Alma
de l'Homme
Park Jardins du Trocadéro
Musée du quai Branly - Jacques Chirac
Rue de l'Université
Starbucks Coffee
Musée Clemenceau
Seine
TH ARR.
Passy
Cultural center Centre Culturel de l'Ambassade d...
EIFFEL TOWER
Park Champ de Mars
Monument des Droits de l'Homme
École Militaire
Champ de Mars - Tour Eiffel
Bir Hakeim
Mur pour La Paix
CHAMPS

Don't

"Alexa, where's the Eiffel Tower"



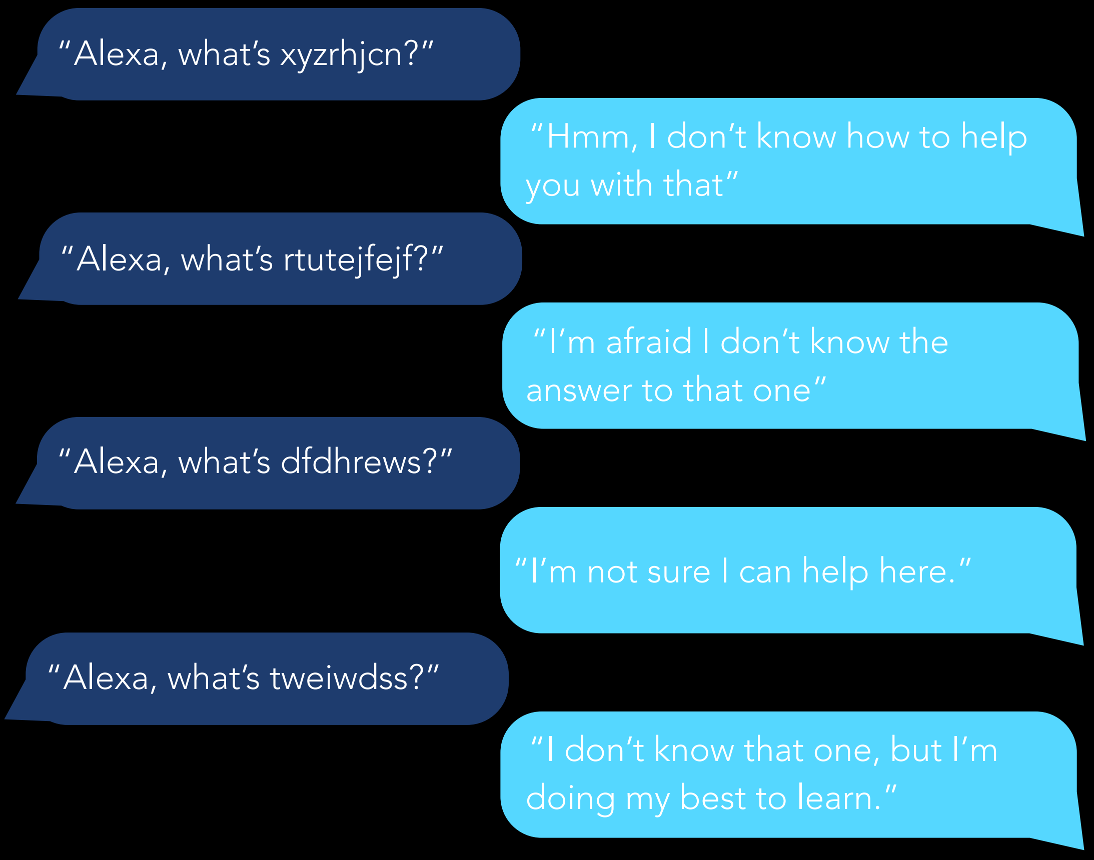
"Huh?"

DESIGNING VOICE OUTPUT

Randomize your TTS responses for similar lines

Listen to your TTS often, over a long time. Only then you'll understand how frustrating it can get to users hearing the same all the time - again, randomise and customise!

Do



"Alexa, what's xyzrhjcn?"

"Hmm, I don't know how to help you with that"

"Alexa, what's rtutejfejf?"

"I'm afraid I don't know the answer to that one"

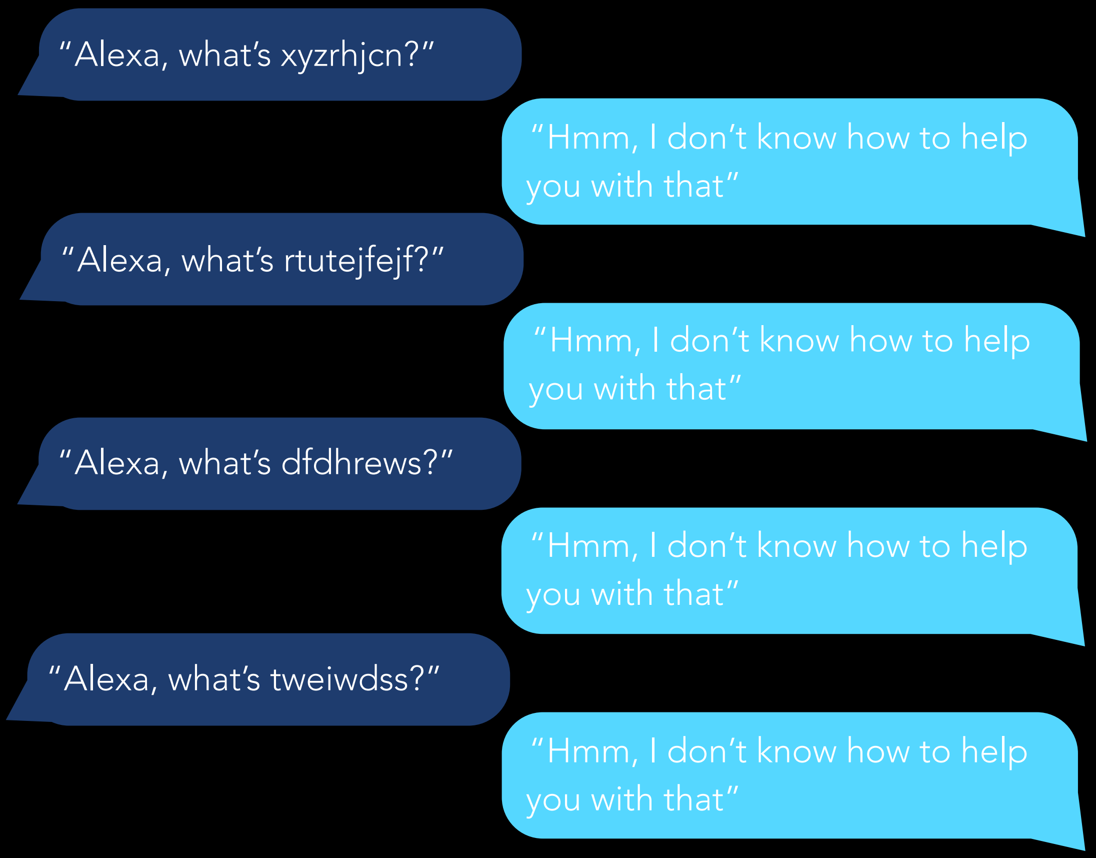
"Alexa, what's dfdhrews?"

"I'm not sure I can help here."

"Alexa, what's tweiwdss?"

"I don't know that one, but I'm doing my best to learn."

Don't



"Alexa, what's xyzrhjcn?"

"Hmm, I don't know how to help you with that"

"Alexa, what's rtutejfejf?"

"Hmm, I don't know how to help you with that"

"Alexa, what's dfdhrews?"

"Hmm, I don't know how to help you with that"

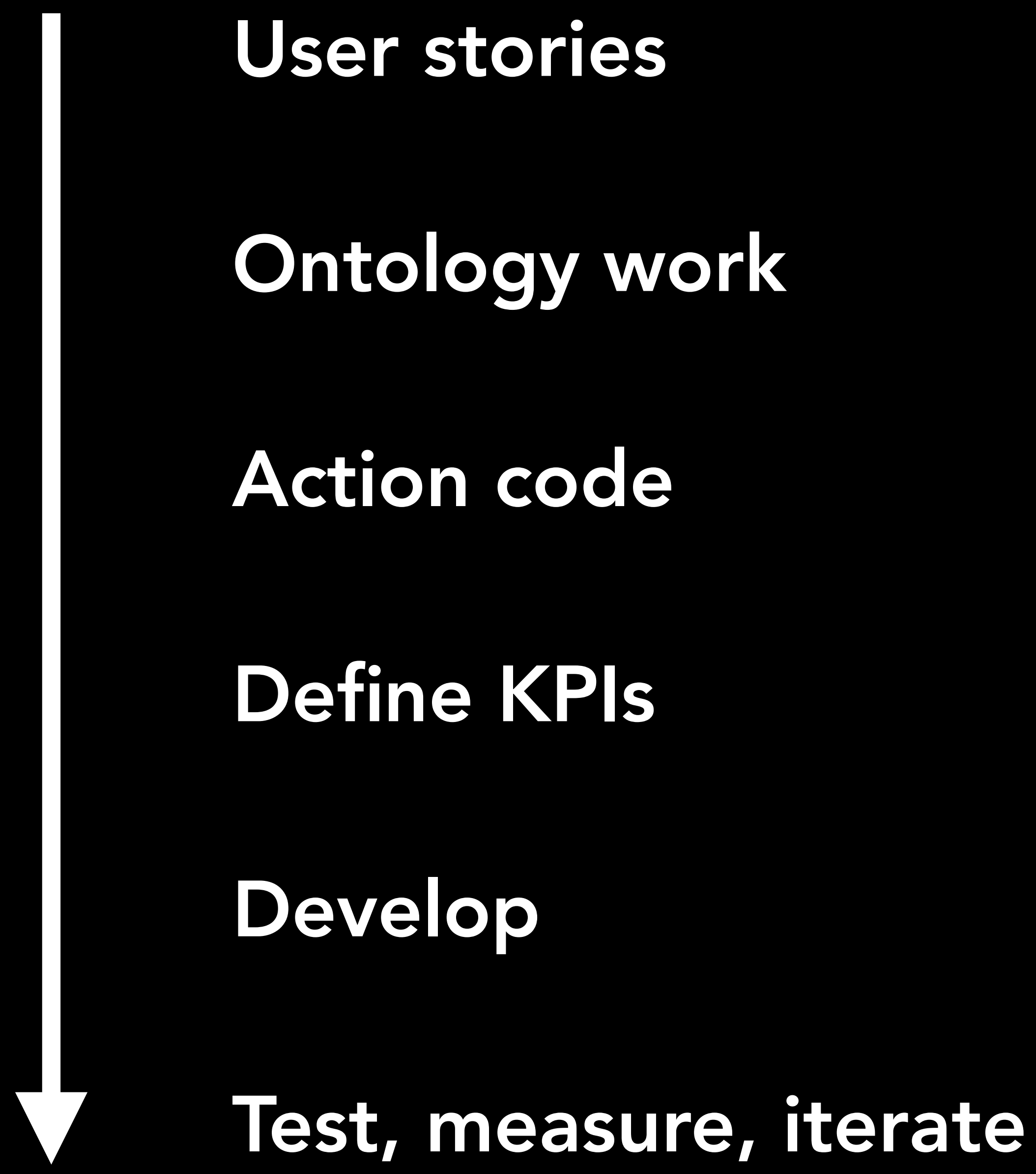
"Alexa, what's tweiwdss?"

"Hmm, I don't know how to help you with that"

HANDS-ON DESIGN PROCESS

HANDS-ON DESIGN PROCESS

GETTING STARTED



User stories

Ontology work

Action code

Define KPIs

Develop

Test, measure, iterate

HANDS-ON DESIGN PROCESS

GETTING STARTED

- **Start from the use case, work backwards.** Don't try to use the tech *because you can.* Does it make sense to solve a user problem?

- **What's so cool about your product?** Why would people use it and encourage others to do so?

- **How many domains** do you want to cover? **How widely** do you want to cover each of them?

- Start listing user stories and prioritise (P1-must, P2-wish, P3-long term)

DESIGNING AN ONTOLOGY: INTENTS

Step 1. List the different intents you'd like to cover

For this step, don't think too much. Don't get to specific sentences, just think of overall actions. They'll later turn into proper intents, but for the moment you're just thinking top-level actions/tasks you want to help a user with.

E.g.: "Turn on lights", "Turn off lights", "Heat food", "Open door"

DESIGN AN ONTOLOGY: INTENTS

Step 2. Draft the user flows

Point is to check if there's anything you're missing with the ontology design. This should be very straightforward if your assistant-to-be doesn't have multi-turn dialogs.

DESIGN AN ONTOLOGY: INTENTS

Step 3. List a few examples per action

Think of 3-5 example sentences per intent, to verbalise them and validate whether they make sense, in general, together and within that intent. Make them as different as you can, keeping the task at hand in mind.

If there's something that feels odd with the structure after developing examples, feel free to shuffle things a little. That's what this exercise is good for :)

E.g.: for "turn on lights":

- *Turn on the kitchen lights*
- *Can you make sure the lights are on*
- *It's pitch-black in this room, light it up*
- *I need light in here*
- *Turn on living room*

DESIGN AN ONTOLOGY: INTENTS

Step 4. Take note of the first slots

Based on the 3-5 examples per intent from the previous exercise, think of what parameters/ interchangeable parts each intent will have. You will most definitely find more later, but it's good for a start.

Remember: not all sentences need to have a slot. If they all do, that probably points at that slot being required (more about that later)

E.g.: for "turn on lights":

- *Turn on the* ***kitchen*** *lights*
- *Can you make sure the lights are on*
- *It's pitch-black in this room, light it up*
- *I need light in here*
- *Turn on* ***living room***

DESIGN AN ONTOLOGY: INTENTS

Step 5. Diamond technique to ensure MECE-ness

5.1. Decompose more!

Take the list of intents, **talk to other people (4-6)**, check how they'd say it.
With that, try and **divide them into sub-intents** (imaginary entity) even further than what it'd make sense to. See if it makes sense to decompose them even further.

5.2. Check exhaustiveness

With the list of intents, try and get to 10 examples per intent, with your colleagues' inputs and your own. Does the sum of all of them represent the totality of things you want to do (even if there are overlaps)?

5.3. Summarise and merge

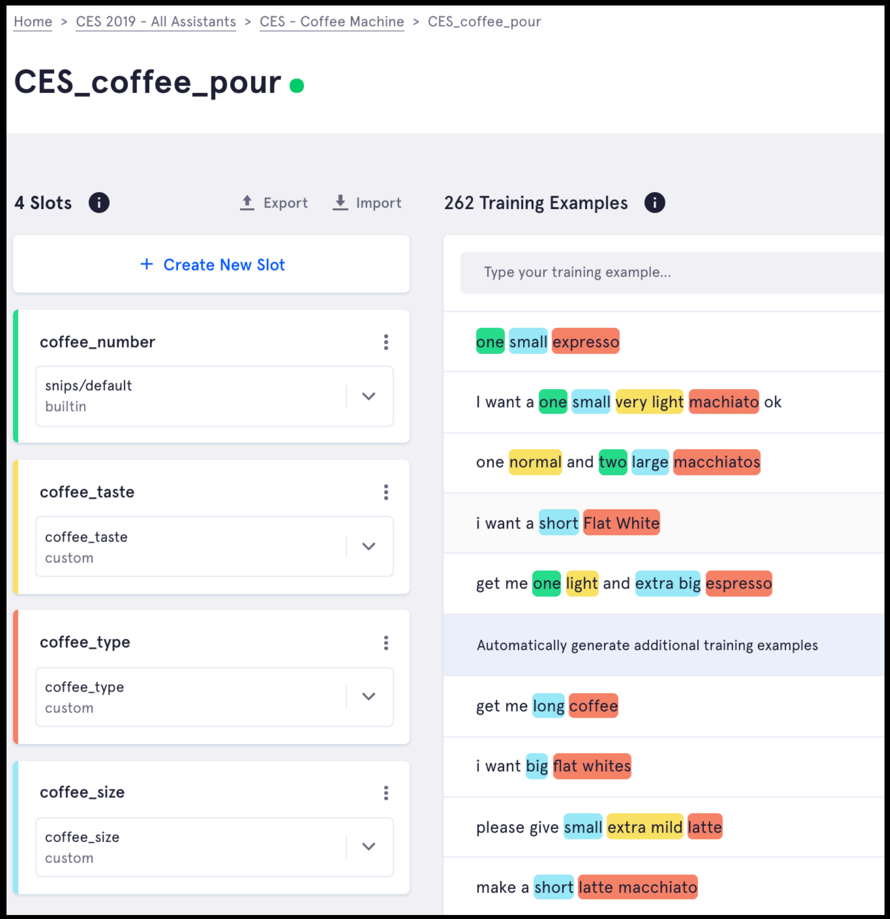
Think of a unique, simple sentence to describe each intent to someone who's never heard of it before, trying to avoid conditioning them into thinking of a specific sentence. E.g.:"You're asking an assistant about the weather conditions".

With the above, you'll realise that several intents might have equal or very similar explanation sentences. Those sub-intents/intents are meant to be together :)

DESIGN AN ONTOLOGY: SLOTS

Things to consider for your slots:

- Add as many **synonyms** to each slot value as you can. BUT make sure synonyms don't appear in more than one slot value. What's the official name of something? How do people call it? How is it written in your database?
- Identify which **slots are required**. On its own? Either one or the other slot? You'll be able to ask the user as a follow-up question.
- **Extensible slots are dangerous**: your action code might not recognise them.
- **So is partial matching**: handle with care and only when you're completely sure there won't be overlaps.


Home > CES 2019 - All Assistants > CES - Coffee Machine > CES_coffee_pour
Export
Import
+ Create New Slot
Type your training example...

CES_coffee_pour ●

4 Slots

coffee_number
snips/default
builtin

coffee_taste
coffee_taste
custom

coffee_type
coffee_type
custom

coffee_size
coffee_size
custom

262 Training Examples

one small expresso

I want a one small very light machiato ok

one normal and two large macchiatos

i want a short Flat White

get me one light and extra big espresso

Automatically generate additional training examples

get me long coffee

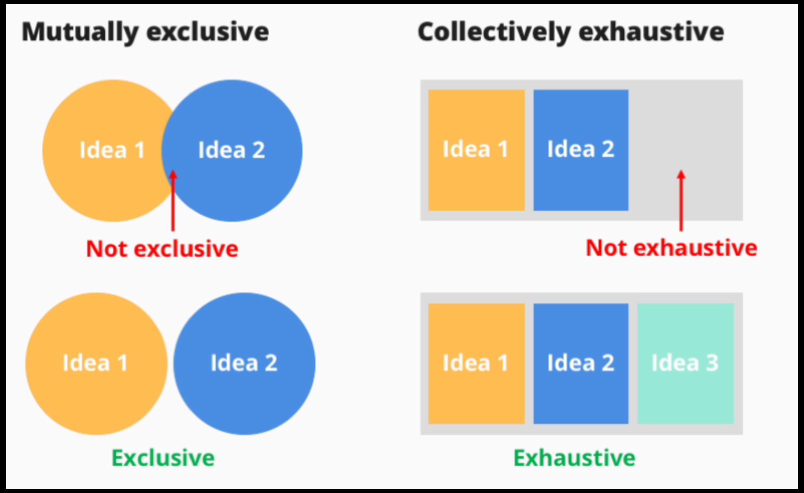
i want big flat whites

please give small extra mild latte

make a short latte macchiato

DESIGN AN ONTOLOGY: BEST PRACTICES

Do	Don't
Assign an intent for each use case. E.g.: • 1 intent for turning on lights • 1 intent for turning off lights • 1 intent for adjusting brightness of lights	Use the same intent for different use cases and channel different use cases via different slots. E.g.: • 1 intent for everything that has to do with lights, setting the different use cases with different slot values ('on', 'off', 'down'…)
Create **one intent per action**, NOT per way of commanding that action.	Get crazy creating intents for each way of saying something. One for "Open lights", one for "Turn on lights", one for "Switch on lights"…
Collective exhaustiveness: Have your assistant cover all the possible cases	Leave relevant use cases behind
Mutual exclusiveness: Have each intent cover one single general use case, avoid overlaps	Cover the same or a similar use case in two or more different intents



STRUCTURE OF A VOICE INTERFACE

ACTION CODE / FLOW DESIGN (I)

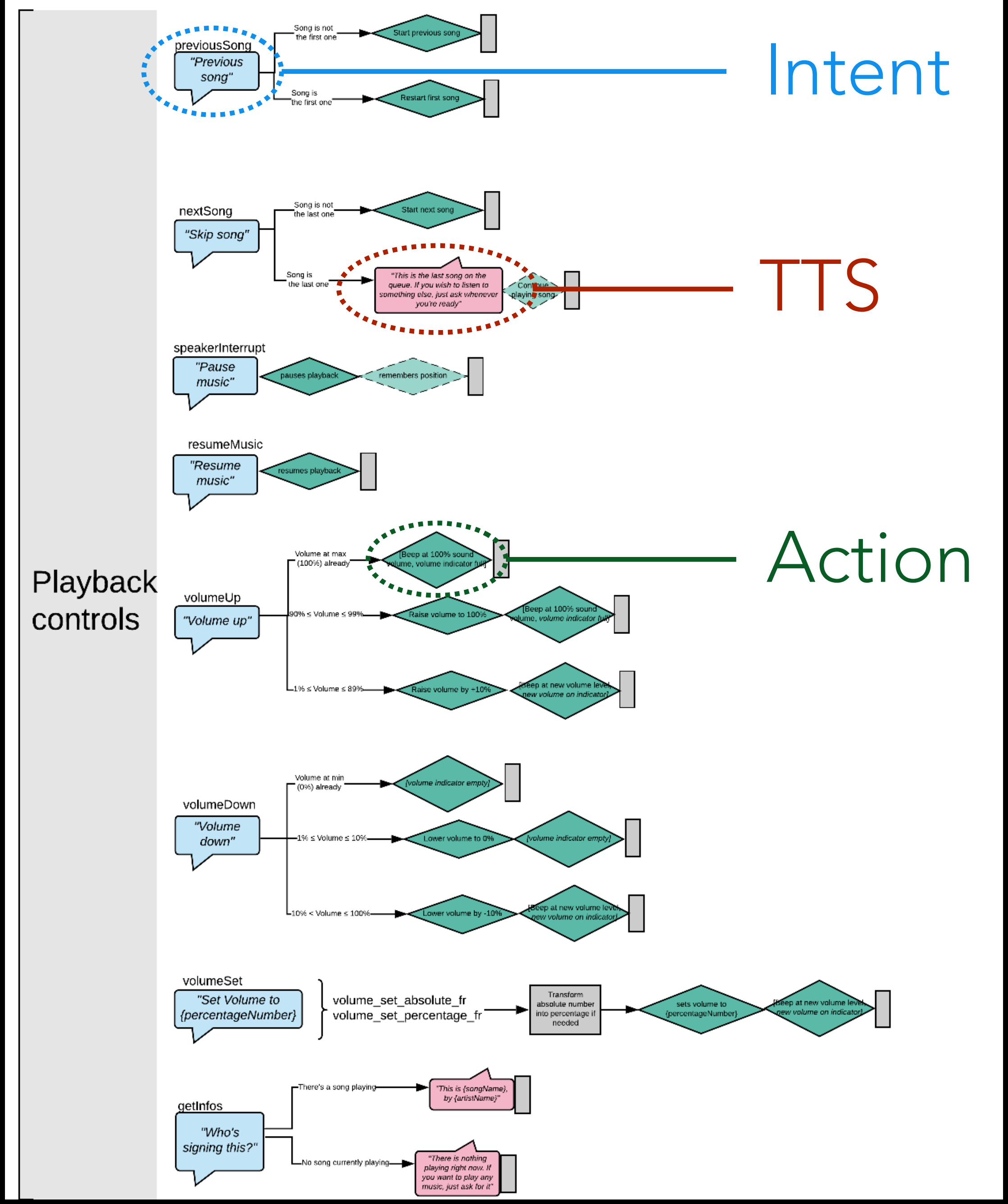
What should we do once we understand the user?

Now it's the time to think through all possible user scenarios.

Stay **top-level**, but do it in **close contact with your developers**: they'll know better what can be done and what can't, what database endpoints the app is hitting, etc.

Start with the easy ones - what happens when someone actually does what they're supposed to?

Then add complexity. What happens when…
… a slot isn't understood?
… two slots collide with each other?
… user is requesting an action that can't be performed?
…


Playback controls
previousSong
"Previous song"
Song is not the first one
Start previous song
Song is the first one
Restart first song
Intent
nextSong
"Skip song"
Song is not the last one
Start next song
Song is the last one
"This is the last song on the queue. If you wish to listen to something else, just ask whenever you're ready"
Continue playing song
TTS
speakerInterrupt
"Pause music"
pauses playback
remembers position
resumeMusic
"Resume music"
resumes playback
volumeUp
"Volume up"
Volume at max (100%) already
[Beep at 100% sound volume, volume indicator full]
Action
90% ≤ Volume ≤ 99%
Raise volume to 100%
[Beep at 100% sound volume, volume indicator full]
1% ≤ Volume ≤ 89%
Raise volume by +10%
[Beep at new volume level, new volume on indicator]
volumeDown
"Volume down"
Volume at min (0%) already
[volume indicator empty]
1% ≤ Volume ≤ 10%
Lower volume to 0%
[volume indicator empty]
10% < Volume ≤ 100%
Lower volume by -10%
[Beep at new volume level, new volume on indicator]
volumeSet
"Set Volume to {percentageNumber}
volume_set_absolute_fr
volume_set_percentage_fr
Transform absolute number into percentage if needed
sets volume to {percentageNumber}
[Beep at new volume level, new volume on indicator]
getInfos
"Who's signing this?"
There's a song playing
"This is {songName}, by {artistName}"
No song currently playing
"There is nothing playing right now. If you want to play any music, just ask for it"

DIALOGS / USUAL TOOLS

- **Slot-filling/follow-up dialog:**

The most usual kind of dialog. Follows up with user to ask for missing information required for task.

- **Multi-turn dialogs:**

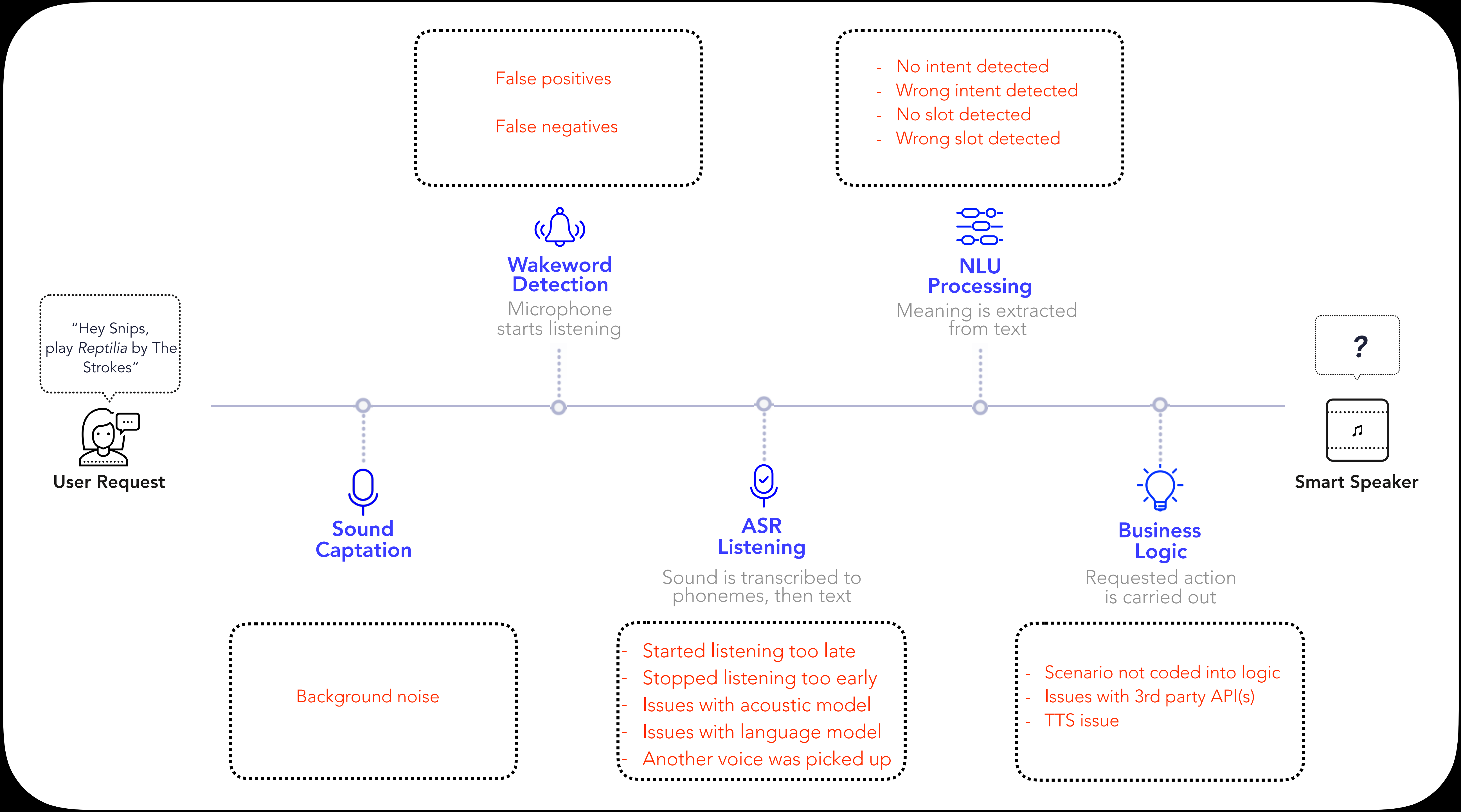
Dialogs that include several back-and-forths with user.
In some parts of it: **use intent filtering!**
(We might only want to enable some intents ("Cancel", "Confirm", "Select number three") in particular points of the interaction, and not by default.)

WHAT CAN GO WRONG?

Short answer: **everything**.

KPIS AND QUALITY ASSURANCE

POTENTIAL ISSUES (SOME OF MY FAVOURITES)



"Hey Snips, play *Reptilia* by The Strokes"

User Request

Sound Captation

Background noise

Wakeword Detection

Microphone starts listening

False positives

False negatives

ASR Listening

Sound is transcribed to phonemes, then text

- Started listening too late
- Stopped listening too early
- Issues with acoustic model
- Issues with language model
- Another voice was picked up

NLU Processing

Meaning is extracted from text

- No intent detected
- Wrong intent detected
- No slot detected
- Wrong slot detected

Business Logic

Requested action is carried out

- Scenario not coded into logic
- Issues with 3rd party API(s)
- TTS issue

?

Smart Speaker

POTENTIAL ISSUES

For each layer of the system, I will be assuming that everything went fine until that point.

But problems usually accumulate!

KPIS AND QUALITY ASSURANCE

WAKE WORD ISSUES

Issue	How to recognise it	What causes it	BRACE!	Repair
False positives	Wake word triggers when people speak about something else	- Wake word too sensitive - Phonemes too generic - Model lacking negative data	*N/A*	- Look for a better WW - Reduce sensitivity - Improve model
False negatives	Wake word doesn't trigger when summoned	- Wake word not sensitive enough - Model not well trained	*N/A*	- Improve model - Increase sensitivity

KPIS AND QUALITY ASSURANCE

ASR ISSUES

Issue	How to recognise it	What causes it	BRACE!	Repair
Started listening too late	Recognised words starts mid-sentence	ASR latency too high, sound feedback interfering	*"Sorry, I'm not sure I understood"*	Accelerate ASR processing or remove influence of feedback via AEC
Stopped listening too late	Recognised words end mid-sentence	ASR endpointing too sensitive to silence or request too slow	*"Sorry, I'm not sure I understood"*	Reduce ASR endpointing sensitivity
Problem with acoustic model	Words get mistaken by others or not recognised at all	- Pronunciation - Foreign words - Sound quality	*"Sorry, I'm not sure I understood"*	Add more relevant phonemes per word
Problem with language model	Words get mistaken by others or not recognised at all	Word not in library	*"Sorry, I'm not sure I understood"*	Make sure words are present in ASR library
Another voice was picked up	Other voices in the room get picked up	Microphone not beam forming	*N/A*	Improve microphone or Digital Signal Processor/Audio Frontend

KPIS AND QUALITY ASSURANCE

NLU ISSUES (I)

Issue	What causes it	How to recognise it	BRACE!	Repair
No intent detected (false negative)	Not enough (relevant) training examples for intent	ASR picks up sentence but NLU doesn't know what to do with it	*"Sorry, I don't know how to help you with that"*	Enrich training data set with different sentence structures and patterns
No intent detected (true negative)	Question is out of domain	ASR picks up sentence but NLU doesn't know what to do with it	*"Sorry, I don't know how to help you with that"*	*N/A, all good*
Wrong intent detected	- Not enough (relevant) training examples - Two or more intents too close to each other or overlapping	NLU picks up the wrong intent	*N/A*	- Enrich training data for existing intents - Remove bad training data - Review ontology (merge?)

KPIS AND QUALITY ASSURANCE

NLU ISSUES (II)

Issue	What causes it	How to recognise it	BRACE!	Repair
No slot detected (false negative)	Slot value not in list	Slot will be ignored	- *If slot required: "Sorry, in what room do you want to turn on the lights?"* - *Else: N/A*	Make sure value is part of the slot.
No slot detected (true negative)	People asking for nonsensical stuff	Slot will be ignored	- *If slot required: "Sorry, in what room do you want to turn on the lights?"* - *Else: N/A*	*N/A, all good*
Wrong slot detected	- Slot value not in list - Slot value duplicated or too close to others	Wrong slot value will be picked	*N/A :(*	- Enrich list of slot values, remove bad data

KPIS AND QUALITY ASSURANCE

ACTION CODE ISSUES

Issue	What causes it	How to recognise it	BRACE!	Repair
Action code not responding to NLU output	Intent/slot combination not mapped to any action	The thing will crash or won't do anything	- *"I'm afraid something went wrong"* - Do whatever you can with what you have	Review decision tree, make sure it includes handling edge cases
(Any other issue with APIs, DBs)	(Any other issue with APIs, DBs)	The thing will crash or won't do anything	- *"I'm afraid something went wrong"* - Do whatever you can with what you have	Check what the heck is wrong with APIs or DBs

KPIS AND QUALITY ASSURANCE

KPIS AND QUALITY ASSURANCE

Point is not to understand everything 100%.

Point is to help a user solve a task.

Everything else is a means to an end.

KPIS AND QUALITY ASSURANCE

KPIS AND QUALITY ASSURANCE

General KPI: Global Success Rate (GSR%)

- Includes all phases from wake word recognition to the execution of the action (base: your perfect scenario)
- Golden rule: Don’t launch until you have >75% (but strive for 95%). If voice interactions work less than 3/4 of times, people won’t use them anymore.
- Not everything is black or white, though. How many interactions ended up solving a user problem in one way or the other? Rely on qualitative data from your users.

KPIS AND QUALITY ASSURANCE

Other things to look for:

- People stopping/barging in halfway through your beautifully designed process
- Actual negative feedback on a given feature - access logs!

OTHER THINGS TO CONSIDER

OTHER THINGS TO CONSIDER

ONE-SHOT REQUESTS OR CONVERSATIONS?

Is your assistant wide but shallow?

- Alexa, Google Assistant, Siri…
- Better for helping users deal with particular tasks
- Larger range of things it can help with

Or narrow but deep?

- Better for stories, etc
- Focus on the script and the TTS more than the actions behind

OTHER THINGS TO CONSIDER

WHAT TTS SHOULD I GO FOR?

Is it better to have a male, female or neutral voice?

Do people mind about TTS with different accents?

What service provides with the best intonation/contextual adaptation?

OTHER THINGS TO CONSIDER

VOICE ID

Does it add value to recognise who's speaking to the assistant?

Is it more useful than it is intrusive?

Can it make sense as a way to authenticate a user (like a password)?

OTHER THINGS TO CONSIDER

PERSONAS

- **Are you even able to create a separate persona?** Or you rely on/be devoured by the larger assistant brand?
- **Some variables to make brands sound audibly distinct:** Music, jingles, earcons, words, phrasing, length, volume, pronunciation, tempo.
- **Optimise for the channel.** Don't regurgitating the copy from your website. Written text won't translate well into audio communication.

OTHER THINGS TO CONSIDER

MULTIMODAL

- **Voice-first doesn’t mean voice-only**. It means you need to design for voice first because it imposes the **most constraints**.

- **Multimodal: voice and visual work together.** Don’t just replicate a side-by-side voice-only and visual-only experience to be run simultaneously.

- **Give users a choice** of how they prefer to interact.

- **Multimodal isn’t just about images and video.** Light rings, spinning lights, avatars, text, images on smartphones or screens… there’s a myriad of ways to design for multimodal.

- Even if most devices are voice-only, have multimodal in mind. **You’ll have to consider both.**

OTHER THINGS TO CONSIDER

ADDITIONAL TOPICS

- Designing VUIs for kids
- VUIs and privacy
- …

RESOURCES

RESOURCES

TOOLS TO DESIGN VUIS

- **To design end-to-end interfaces:**
  - Alexa Skills console
  - BotSociety.io
  - DialogFlow by Google

- **To work on your ontologies:**
  - (Any of the above)
  - Snips console (Snips.ai)

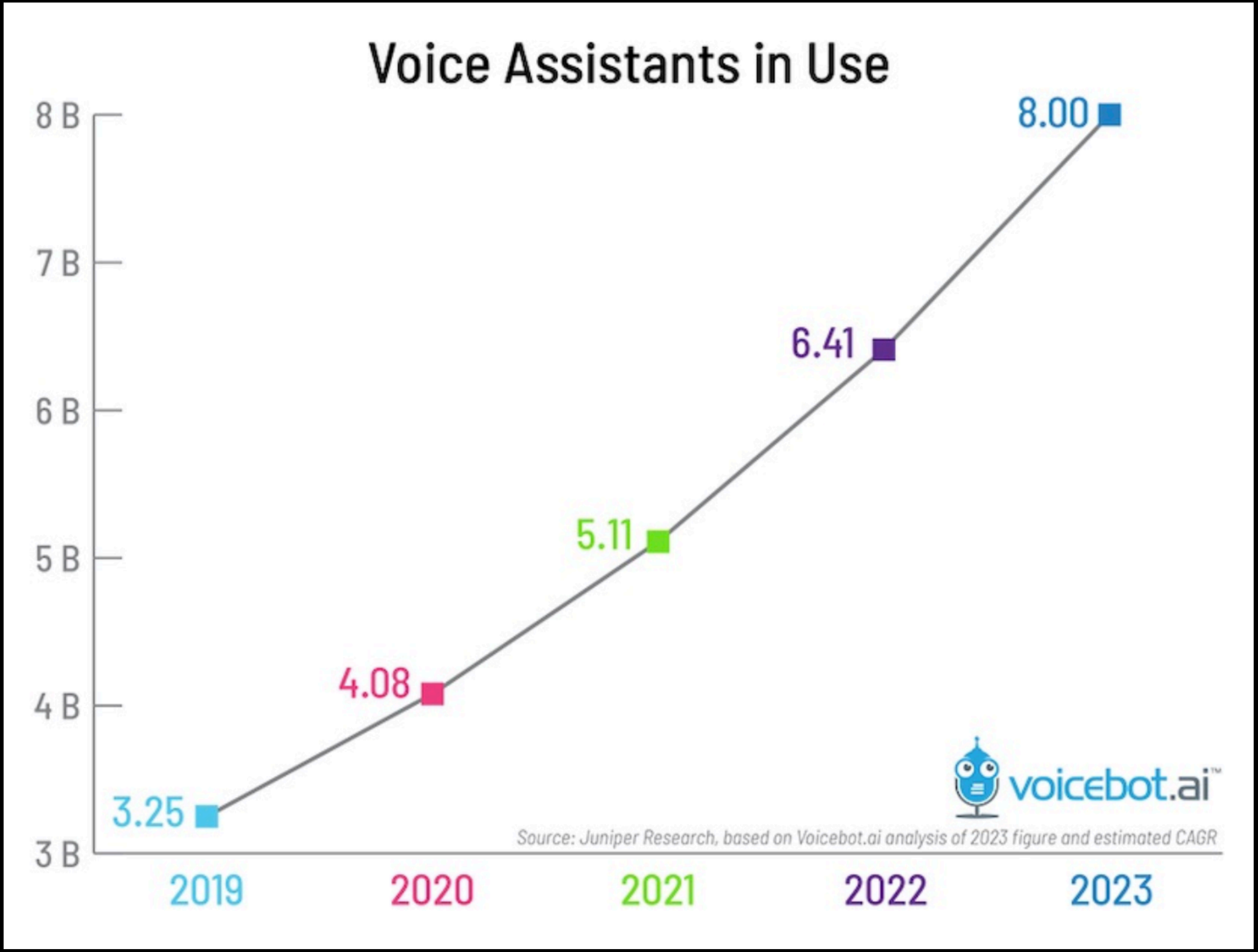
- **To design action code charts:**
  - LucidChart.com

PART 5:
WHAT’S NEXT?

MARKET TRENDS

THE STATE OF VOICE INTERFACES TODAY

MARKET PROJECTIONS



Source: Juniper Research (2018)

WHAT'S NEXT?

MARKET PROJECTIONS

- **Currently**: between **1 and 3.2 billion voice assistants in use**.
- **Estimates**: two-figure yearly growth (+20-40%) to **8 billion voice assistants in 2023** (Juniper Research)
- **Smart speaker base growing fast** in US (25% of population, +40% YoY), UK, EU, CN and AU. Rest of world: main access through smartphones.
- 31 million Installed base of IoT devices worldwide by 2020 (Statista)
- 85% percent of customer interactions that will be managed without a human agent by 2020 (Gartner)

USAGE TRENDS

WHAT'S NEXT?

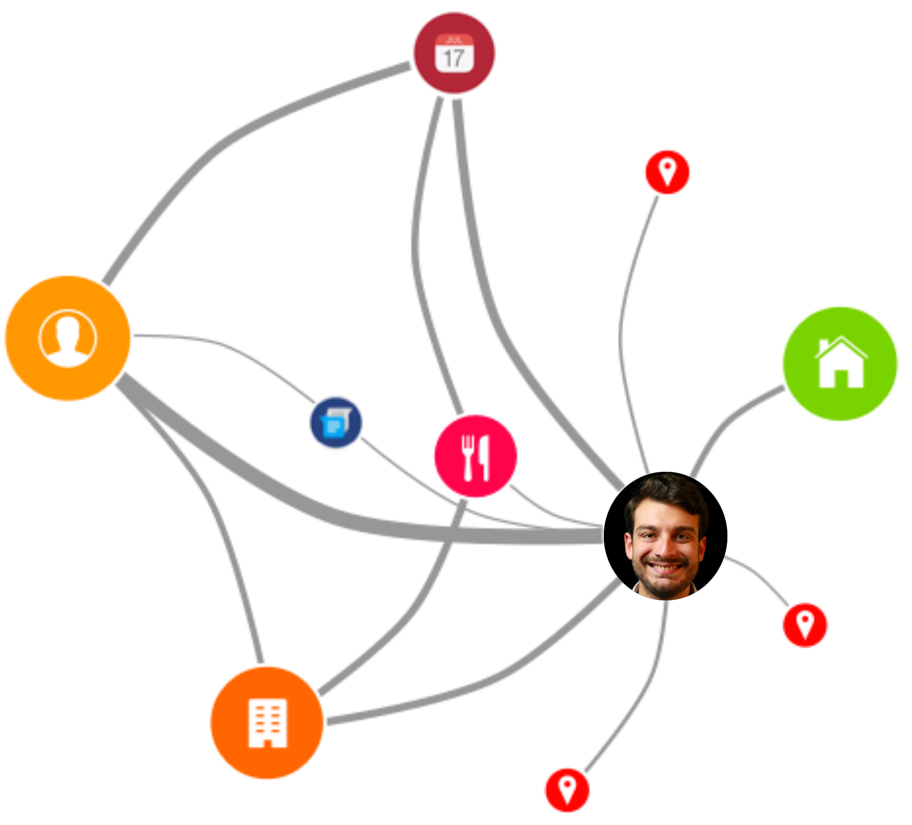
UPCOMING USAGE TRENDS

MULTIMODAL

WHAT'S NEXT?

UPCOMING USAGE TRENDS

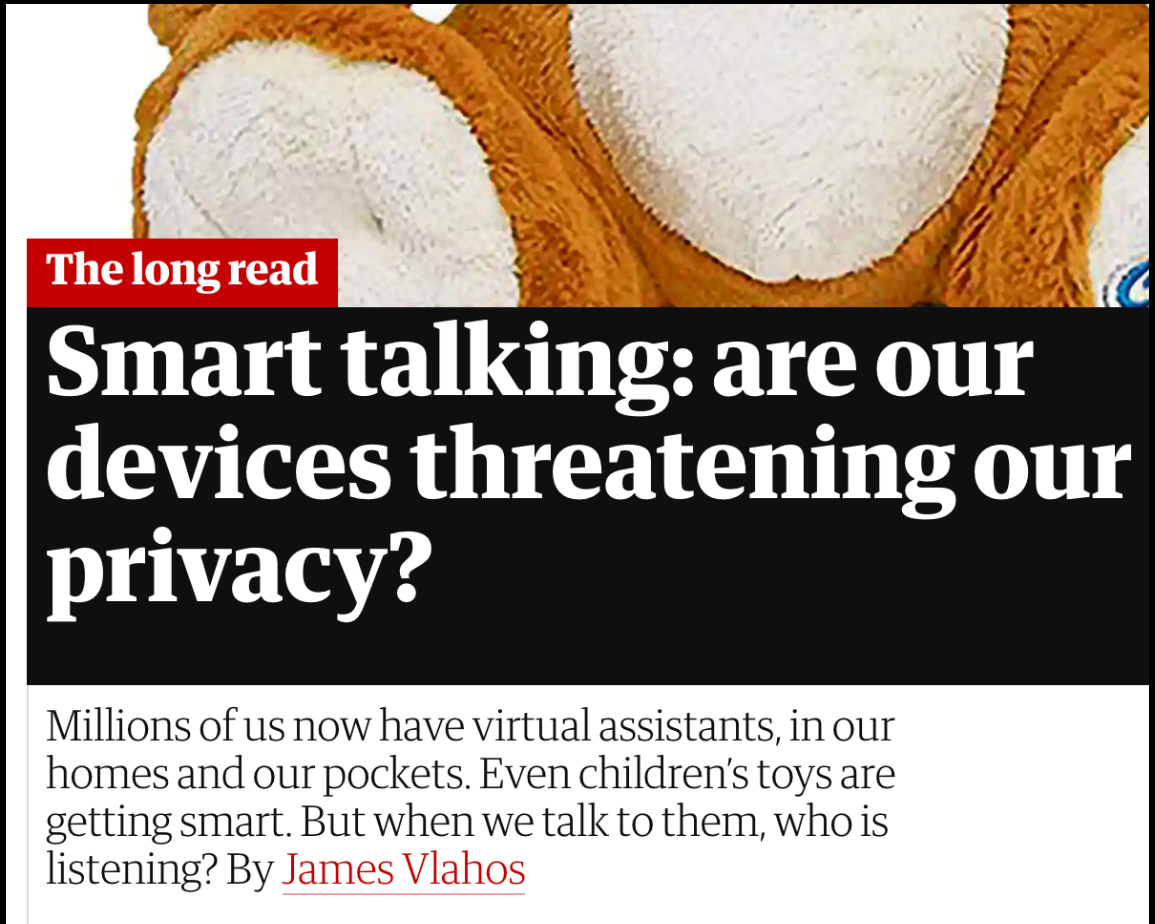
CONTEXTUAL AWARENESS



WHAT'S NEXT?

UPCOMING USAGE TRENDS

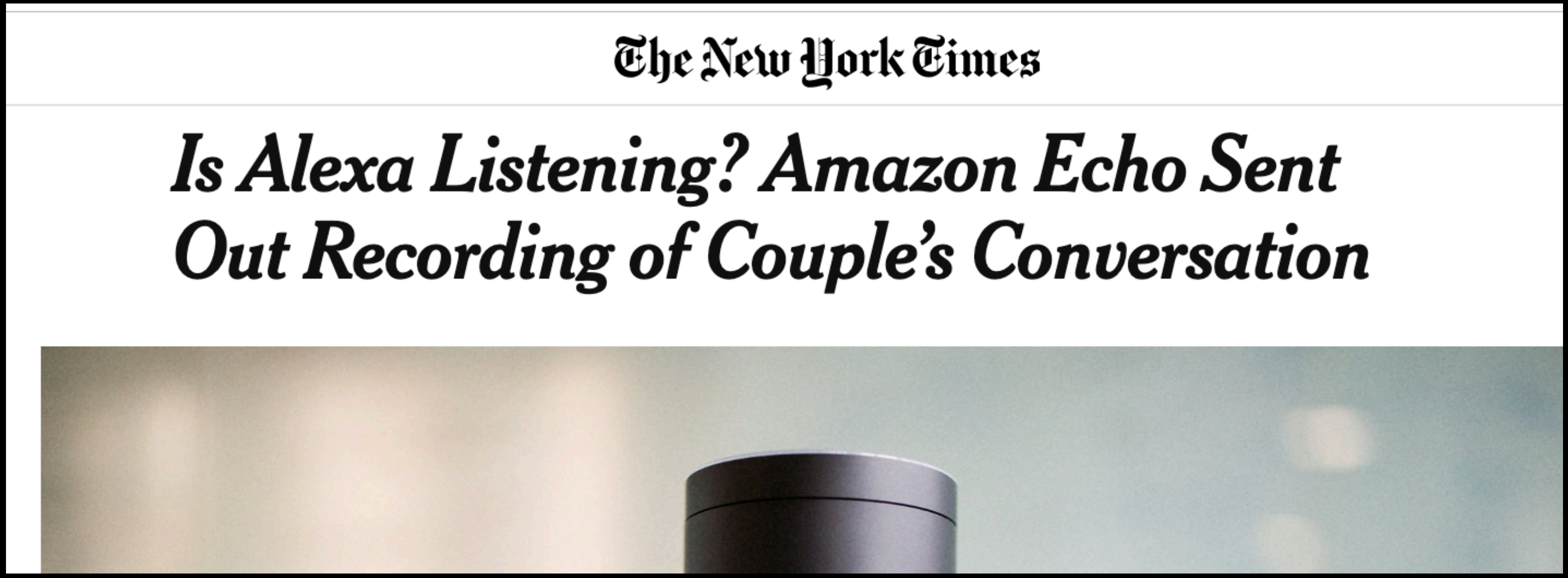
PRIVACY



The long read

Smart talking: are our devices threatening our privacy?

Millions of us now have virtual assistants, in our homes and our pockets. Even children's toys are getting smart. But when we talk to them, who is listening? By James Vlahos



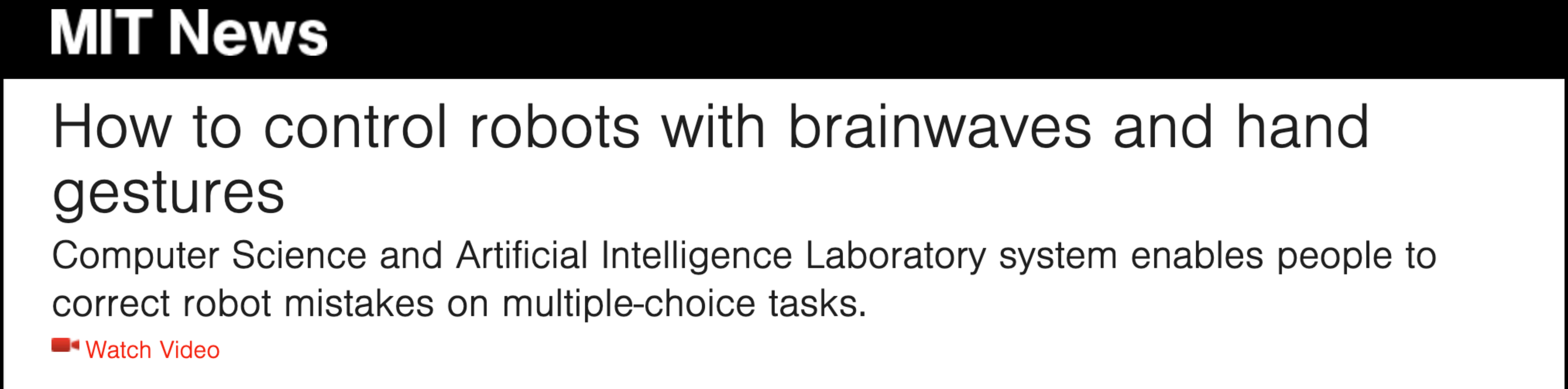
The New York Times

Is Alexa Listening? Amazon Echo Sent Out Recording of Couple's Conversation

WHAT'S NEXT?

UPCOMING USAGE TRENDS

OTHER INPUTS



MIT News

How to control robots with brainwaves and hand gestures

Computer Science and Artificial Intelligence Laboratory system enables people to correct robot mistakes on multiple-choice tasks.

Watch Video

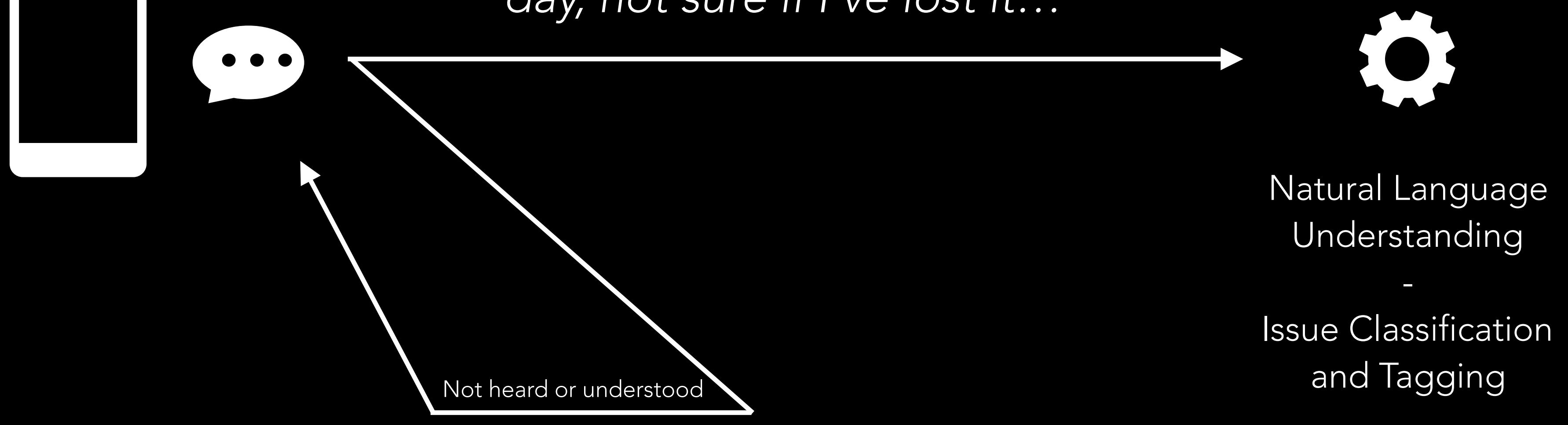
WHAT'S NEXT?

UPCOMING USAGE TRENDS

Even IVRs will be useful!

e.g. Google Duplex

"Ehm, I think I have an issue with my card. There are some weird charges and I might have left it in a restaurant the other day, not sure if I've lost it…"



Natural Language Understanding
-
Issue Classification and Tagging
-
Action Code & TTS

ROBOPACO

PART 6:
PRACTICAL EXERCISES

Exercise 1. (1h30) For a domain (options: a)music, b)local business, c) sports information), observe what each of the three main assistants (Google Assistant, Siri, Alexa) is capable of doing. To do this, use the documentation available on the network and test on the assistants in the corresponding applications.
A. Try to do everything with that assistant on the domain at hand, take note of the interactions and responses, take note of the TTS... (Tip: spreadsheet with columns: interaction/expected answer/answer/comments)
B. Try to see in which intents and slots the assistant is structured in.
C. Extra: Take note of things that work (and find those that don't) and try to understand what may have gone wrong in those that have failed (ASR? NLU? Action code?)

Exercise 2. (2h00) On a domain to choose (options: a) assistant on a screen in an airport, b) assistant for navigator in a car, c) assistant in a sports wearable):
A. Develop ontology (intents, slots) - including, as far as possible (for one or several slots), the accepted values and their characteristics (Required? Synonyms?)
B. Create an approximate flow chart of what has to happen with each interaction, including TTS.
C. Extra: think about what specific problems the specific case may have in terms of ASR, NLU, action code or dialogue.