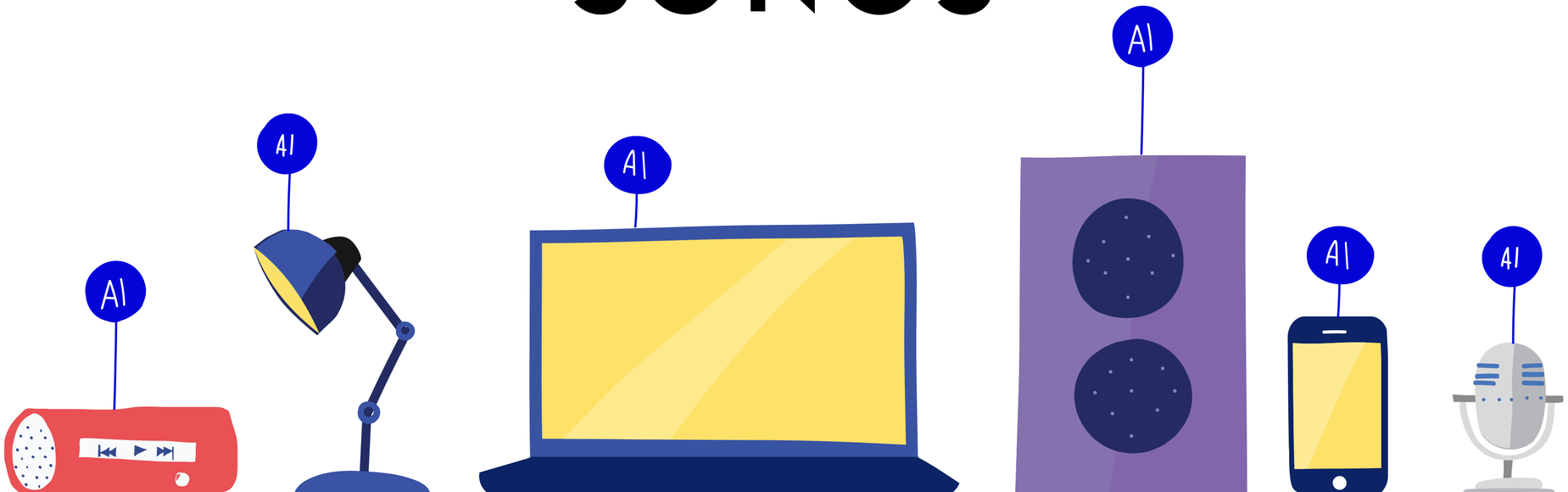
Spoken Language Understanding on the Edge

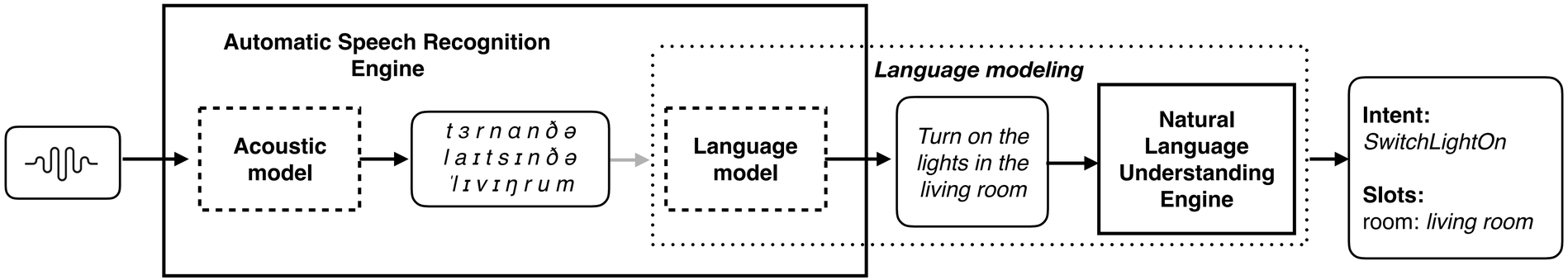
Alaa Saade, Alice Coucke, Alexandre Caulier, Joseph Dureau, Adrien Ball, Théodore Bluche, David Leroy, Clément Doumouro, Thibault Gisselbrecht, Francesco Caltagirone, Thibaut Lavril, Mael Primet
Snips, Paris

EMC2 Workshop @ Neurips 2019
November 13
Alexandre Caulier

SONOS



Spoken language understanding system



Features

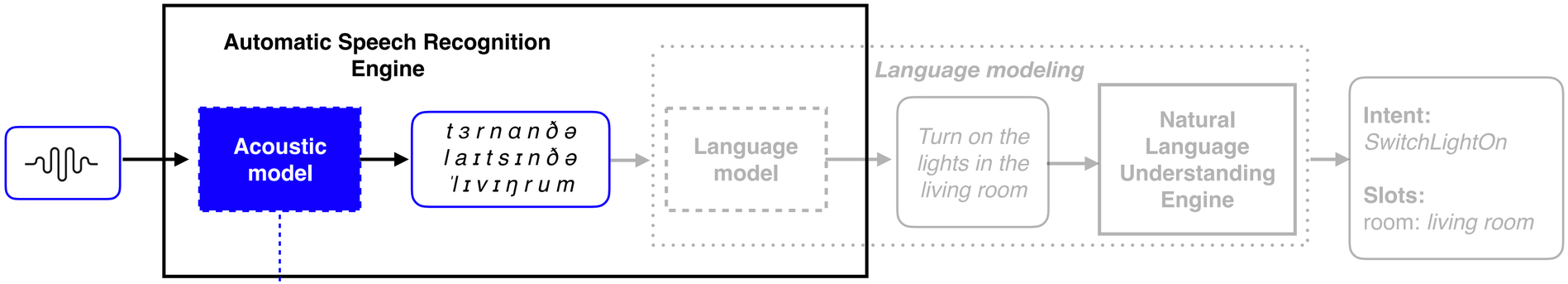
- **Cloud independent -** no remote processing
- **Private by Design -** no user data can be collected
- **Accurate -** on-par with cloud-based solutions

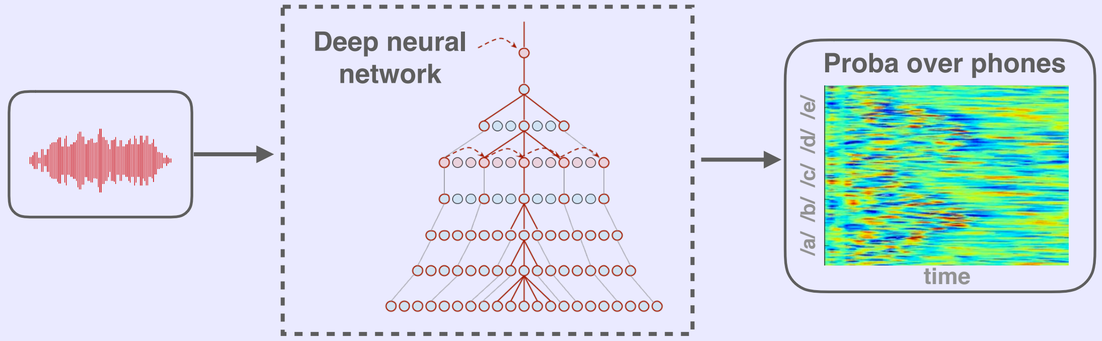
Tested and certified to run on

1GB RAM
1.4GHz CPU

- Raspberry Pi 3 (BCM2837)
- Samsung Artik™ 530
  Samsung Artik™ 710
  Exynos5422
- NXP® i.MX 6DLite
  NXP® i.MX 7D
  NXP® i.MX 8M
- Intel Atom x7
- Broadcom® BCM7271
- Allwinner® A64
- nVidia® Jetson TX2
- Actions® S500
- Snapdragon™ 410
- Respeaker core v2
- iOS

Acoustic modeling





Challenges

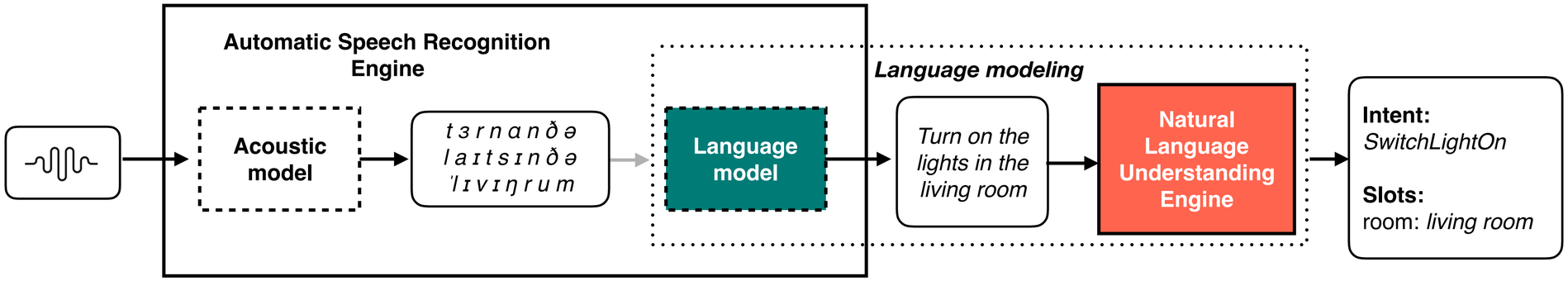
Large deep learning models

Computationally & **memory** intensive

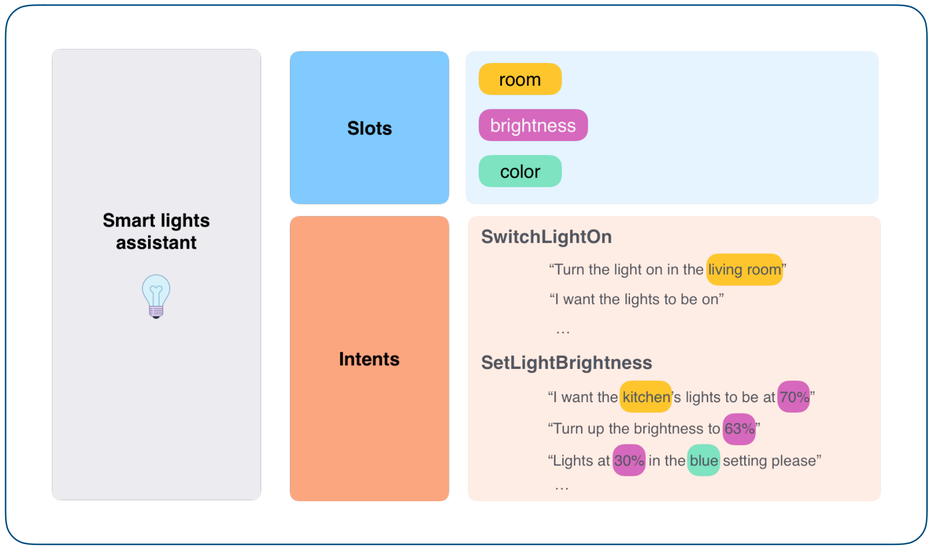
Training data: **10K+ hours** of **in-domain** audio with transcript per language

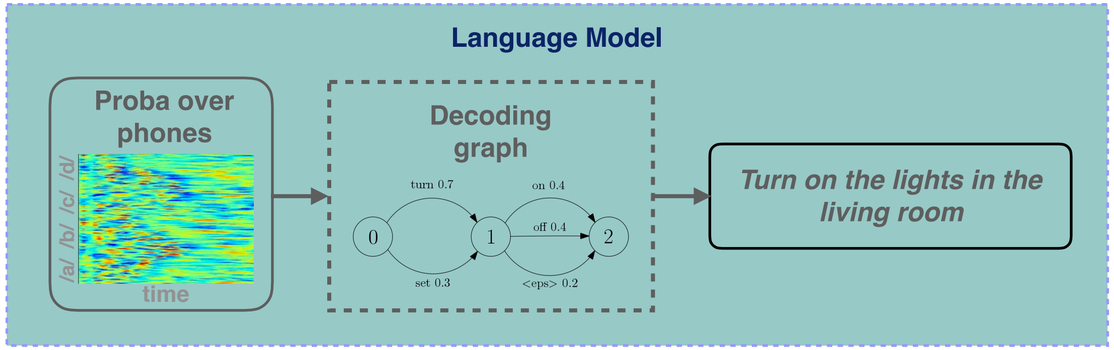
- **Trade-off** between accuracy & computational efficiency
- **Reduced** model size (~10MB)
- **Few K hours** of training data

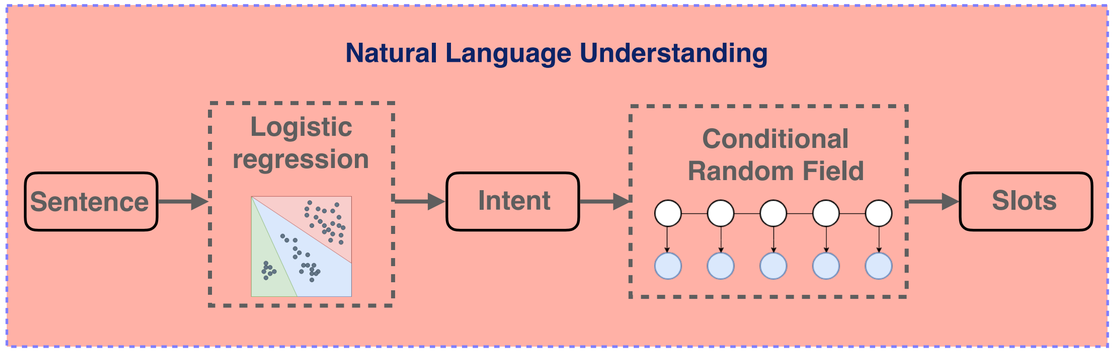
Assistant Contextualization



Approach : LM and NLU are **consistent** and **contextualized**







✓ **Lightweight** models

✓ **Out of vocabulary** management

✓ **On-device personalization**

Benchmarks - Datasets Open Sourcing

Experimental setting

Datasets

Audio utterances with transcripts & supervision

Recorded in close and far-field

💡 Smart Lights Assistant

1.8K utterances

400 word pronunciations

🎵 Music Assistant

3K utterances

178K word pronunciations

Method

snips

Specialized for 💡 & 🎵

<100MB, real time on a Raspberry Pi 3

Google

Google Speech-to-Text cloud services

One-size-fits-all engine

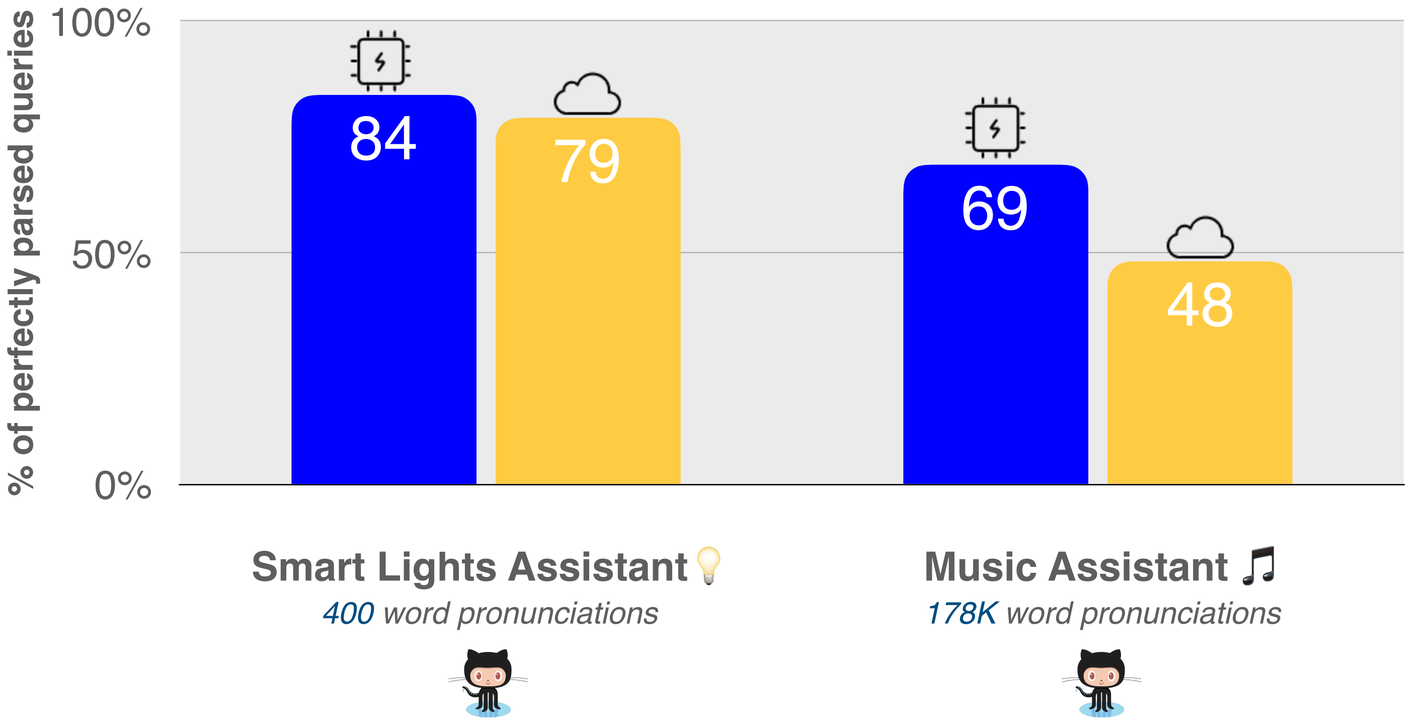
Metrics

End-to-end score

% of perfectly parsed queries

Benchmarks

End-to-End performance



snips — Contextualized for 💡 & 🎵
<100MB, real time on a Raspberry Pi 3

Google — *STT cloud service*
One-size-fits-all engine

🎵

	Tier 1 Artists 1-1k	**Tier 2 Artists** 4.5k-5.5k	**Tier 3 Artists** 9k-10k
Snips	**71 %**	**68 %**	**67 %**
Google	**69 %**	**38 %**	**37 %**

Questions ?