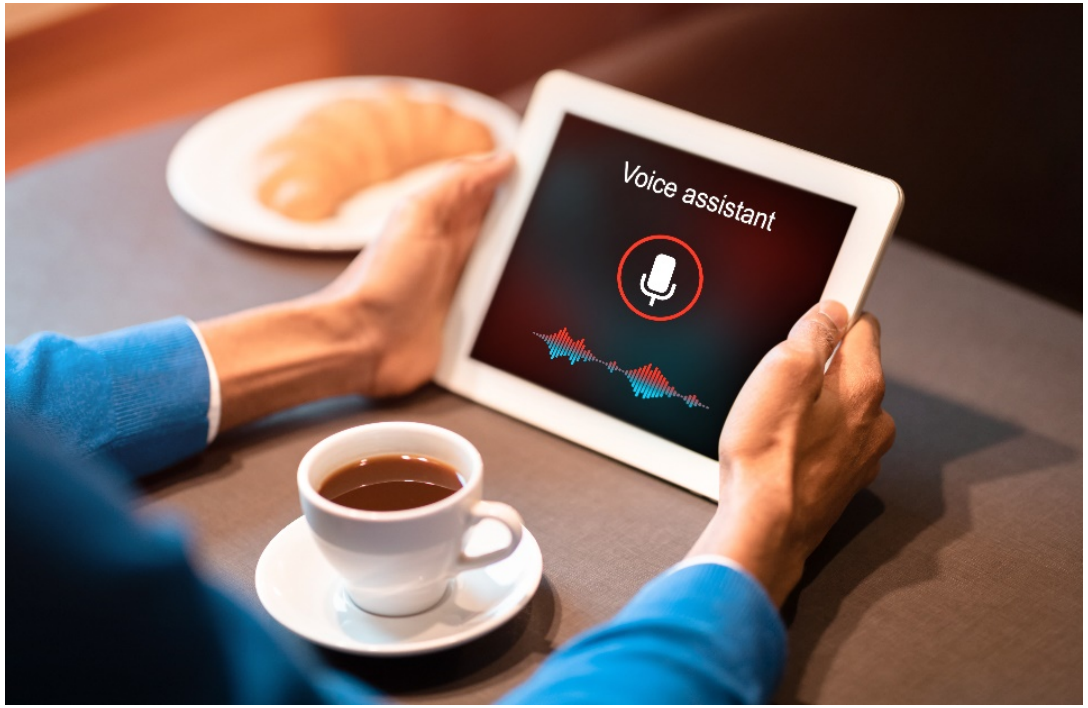
ML school audio track

Keyword Spotting

Dr. Paul Wallbott



Open Questions and exercise feedback

Use Case: Wake word detection

SMALL-FOOTPRINT KEYWORD SPOTTING USING DEEP NEURAL NETWORKS

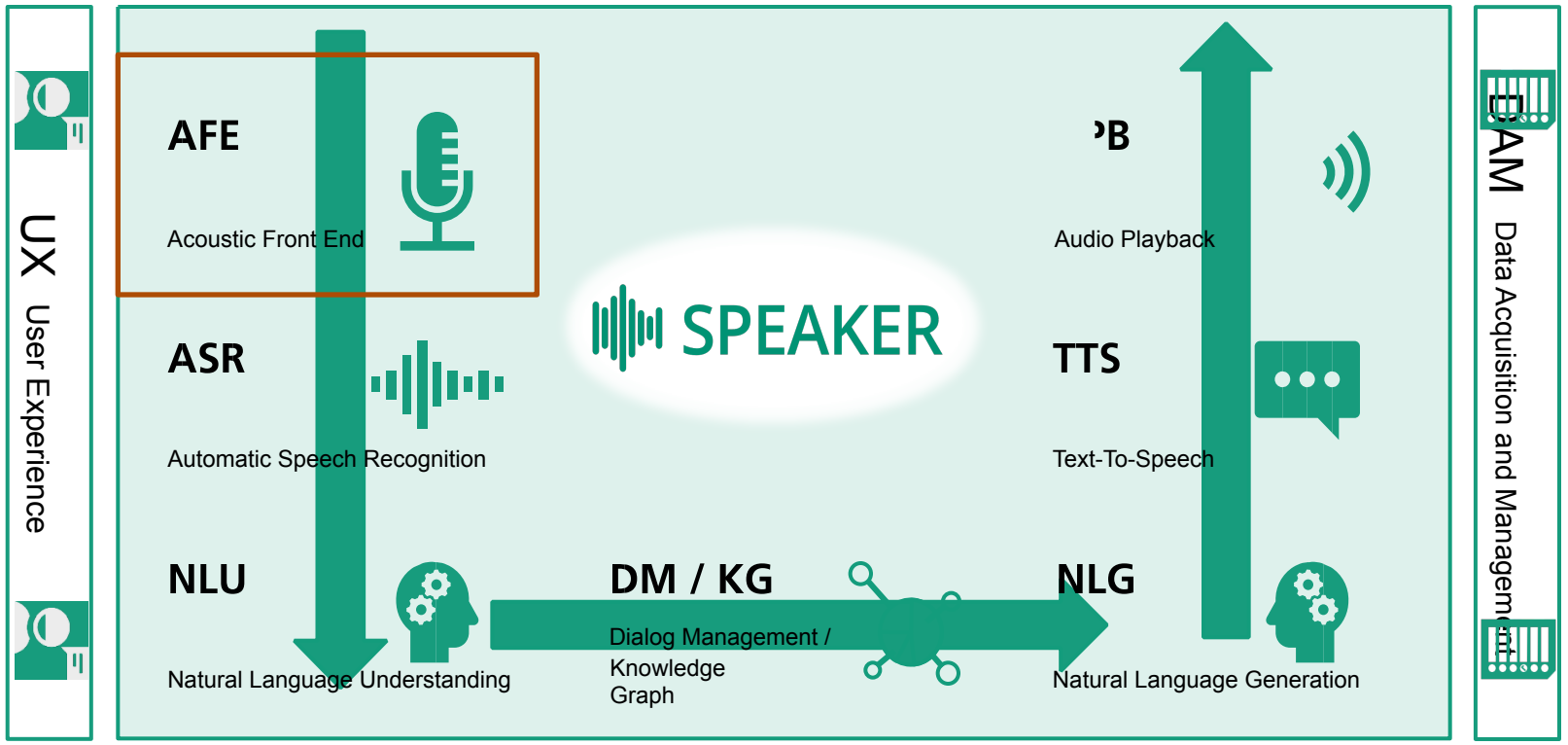
Guoguo Chen[*1] *Carolina Parada*[2] *Georg Heigold*[2]

[1] Center for Language and Speech Processing, Johns Hopkins University, Baltimore, MD
[2] Google Inc., Mountain View, CA
guoguo@jhu.edu carolinap@google.com heigold@google.com

Google research 2014

Siri, ok google 2017

Use Case: Wake word detection



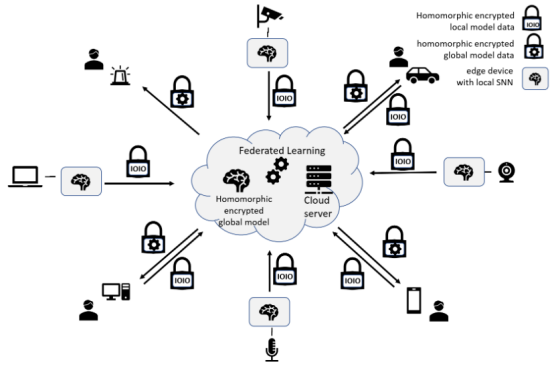
Credits to M. Gref

Other Use Cases

- Command control of smart devices
  - Smart home
  - Robotics
  - …
- KWS advantages vs full fetched ASR
  - Much lower energy footprint
  - On device always possible
- progress in neuromorphic hardware + ASR algorithms -> on device ASR
  - Eventually privacy issue might decouple

Small detour: privacy

- Privacy is a major issue in many domains:
  - Personal data (medical, home assistants, …)
  - Corporate data (competitive advantage)



- Corporate data
  - Fraunhofer speech assistant as privacy preserving business solution
  - European initiative GAIA-X sets a frame for bringing cloud providers, model builders and data owners together

- Personal data
  - Growing consumer interest for data privacy, companies adapt to it
  - Tension between privacy and data gathering causes emergence of research activity (federated learning, homomorphic encryption etc.)

The Sec-Learn project

- Fraunhofer project
- Neuromorphic hardware for on device computations
- Privacy preserving federated learning
  - On device: training, inference
  - Cloud: model aggregation

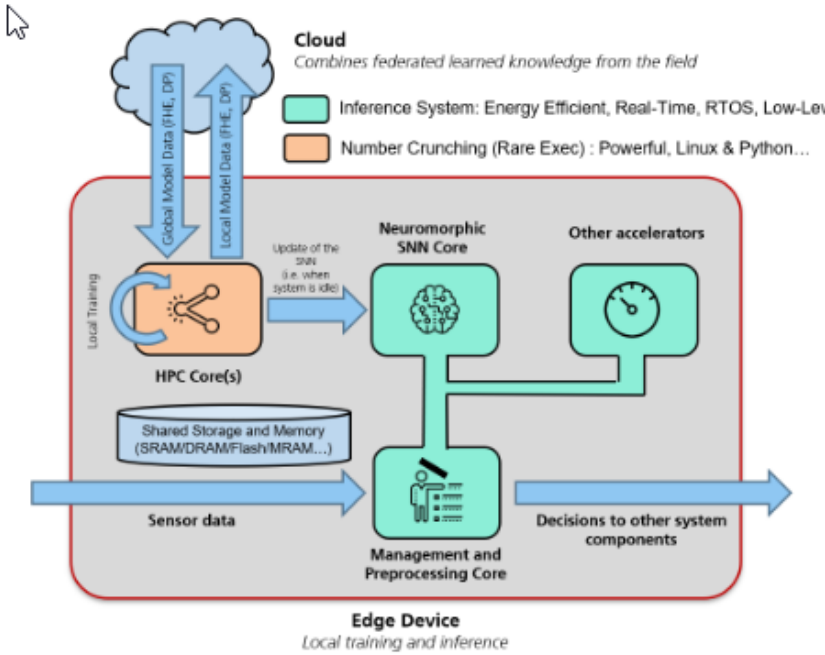


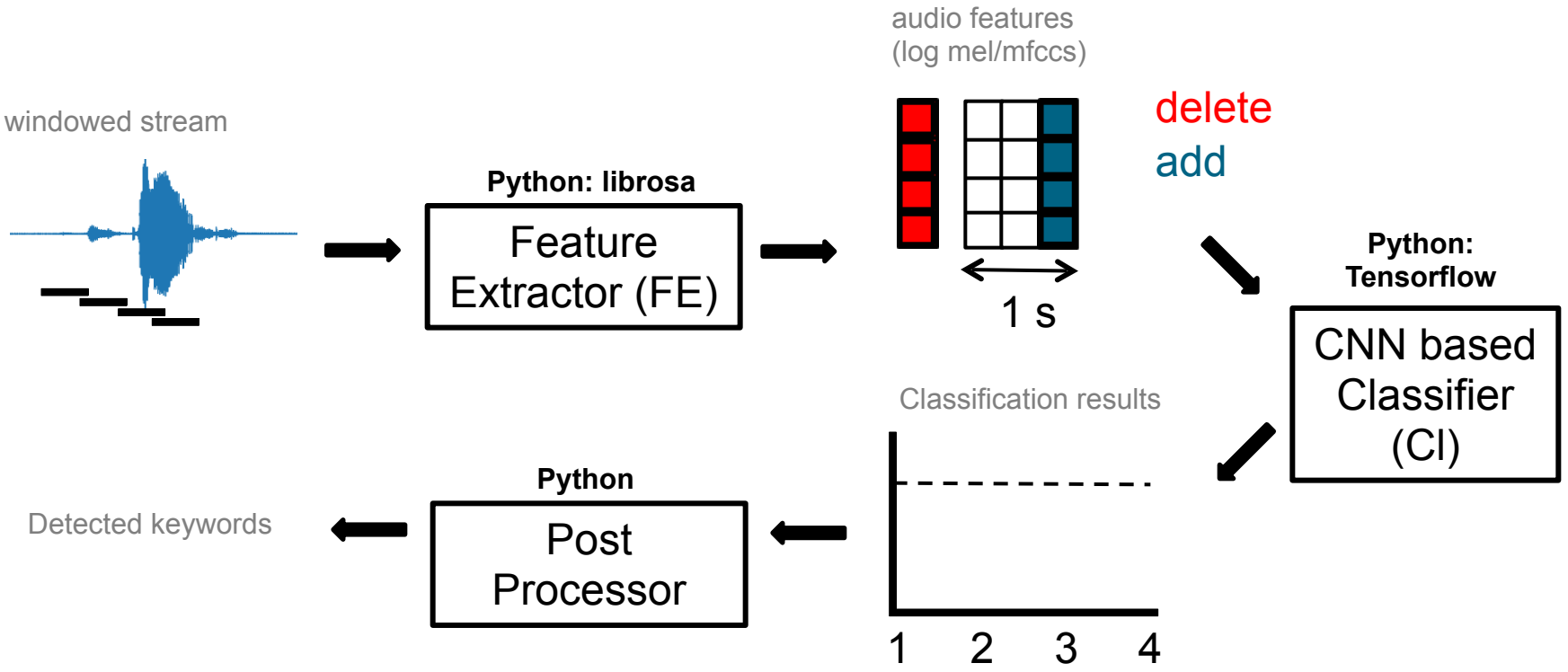
Fig. 3: SEC-Learn Ecosystem

https://samos-conference.com/wp/wp-content/uploads/2021/07/NGC2_59_PDF.pdf

We are almost there…

https://pixabay.com/de/photos/container-containerschiff-hafen-1638068/

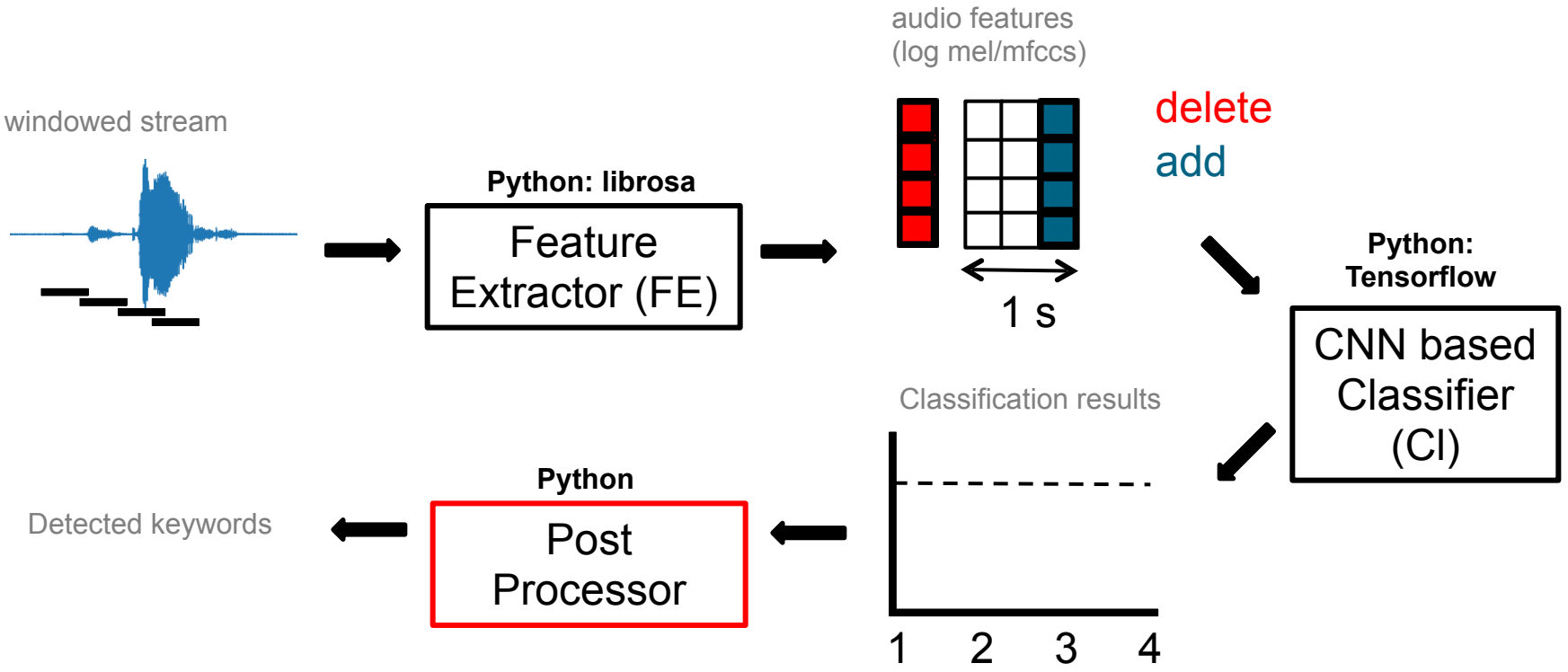
KWS streaming



Introduction: word-based kws

- Feature extractor
  - Speech can be seen as characteristic frequencies being present in the audio signal
  - Convert audio signal to frequency domain (spectrogram, mfccs)
- NN classifier (1s audio -> class probabilities)
  - Neural network that can predict the likelihood of a set of keywords being part of the 1s audio chunk
- Post processor
  - Rule that allows to say which keyword is currently present (or silence or unknown)
  - Threshold or smoothing over the last N probabilities (reduce false positives)

Session 3: streaming



Session 3: Streaming

- Exercise 04: putting it all together!
  - In exercise_04 we put the pieces together and test our keyword spotter in practice
  - Record a 10s long audio with some keywords and try it out with that one!