tinyML® Talks

Enabling Ultra-low Power Machine Learning at the Edge

“Making ML work in the real world”

Dominic Binks - Audio Analytic

UK Area Group – February 23, 2021

www.tinyML.org

tinyML Talks Sponsors

tinyML Strategic Partner

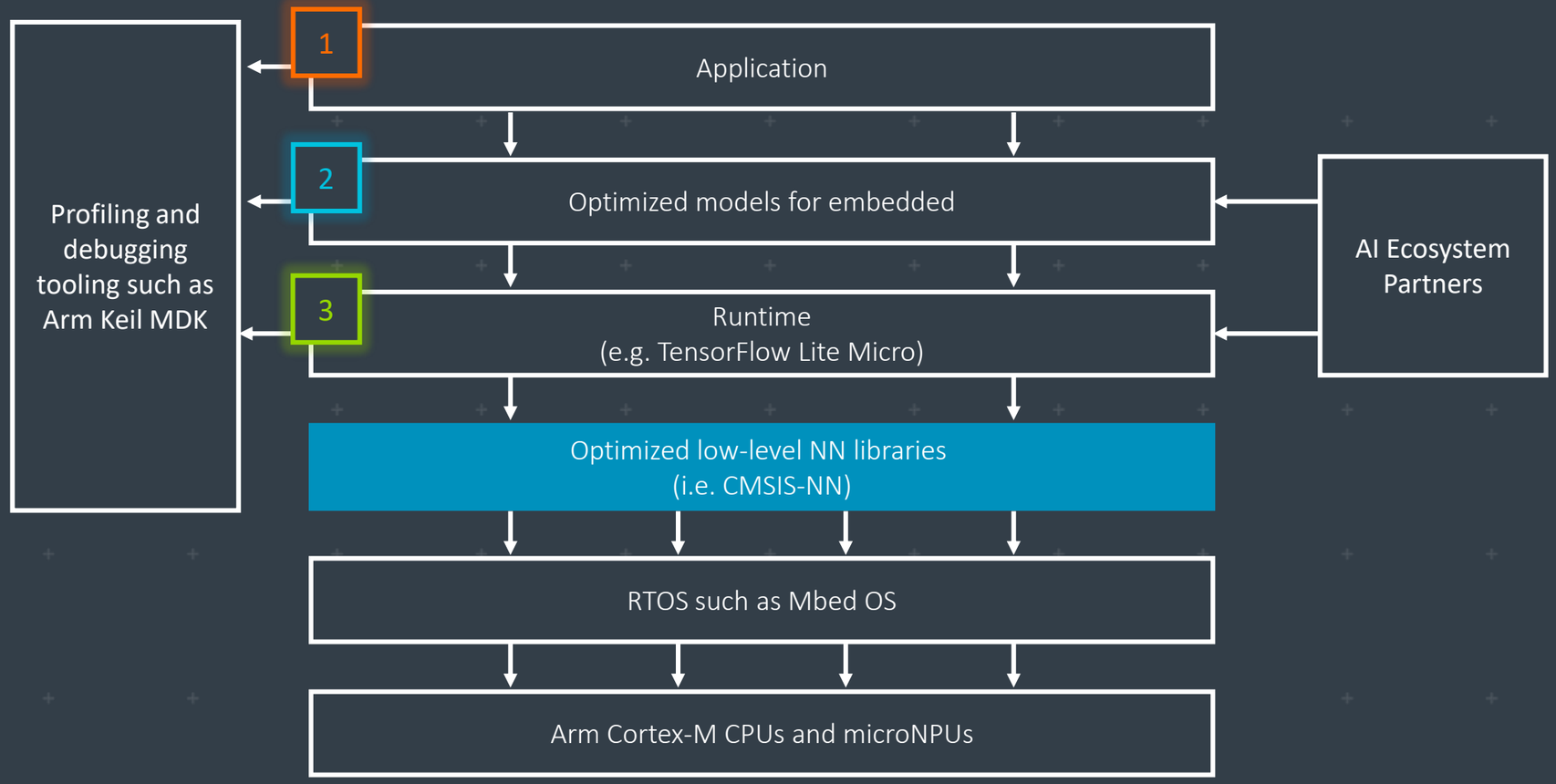
Additional Sponsorships available – contact Bette@tinyML.org for info

Arm: The Software and Hardware Foundation for tinyML

1 Connect to high-level frameworks

2 Supported by end-to-end tooling

3 Connect to Runtime



Stay Connected

@ArmSoftwareDevelopers

@ArmSoftwareDev

Resources: developer.arm.com/solutions/machine-learning-on-arm

arm

Deeplite

WE USE AI TO MAKE OTHER AI FASTER, SMALLER AND MORE POWER EFFICIENT

Automatically compress SOTA models like MobileNet to <200KB with **little to no drop in accuracy** for inference on resource-limited MCUs

Reduce model optimization trial & error from weeks to days using Deeplite's **design space exploration**

Deploy more models to your device without sacrificing performance or battery life with our **easy-to-use software**

BECOME BETA USER **bit.ly/testdeeplite**

mobilityXlab arm AI 100 CBINSIGHTS

TinyML for all developers

C++ library

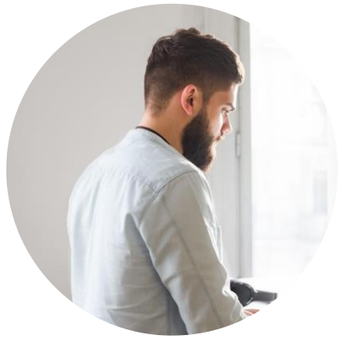
Arduino library

WebAssembly

Dataset

Acquire valuable training data securely

Enrich data and train ML algorithms



Edge Device

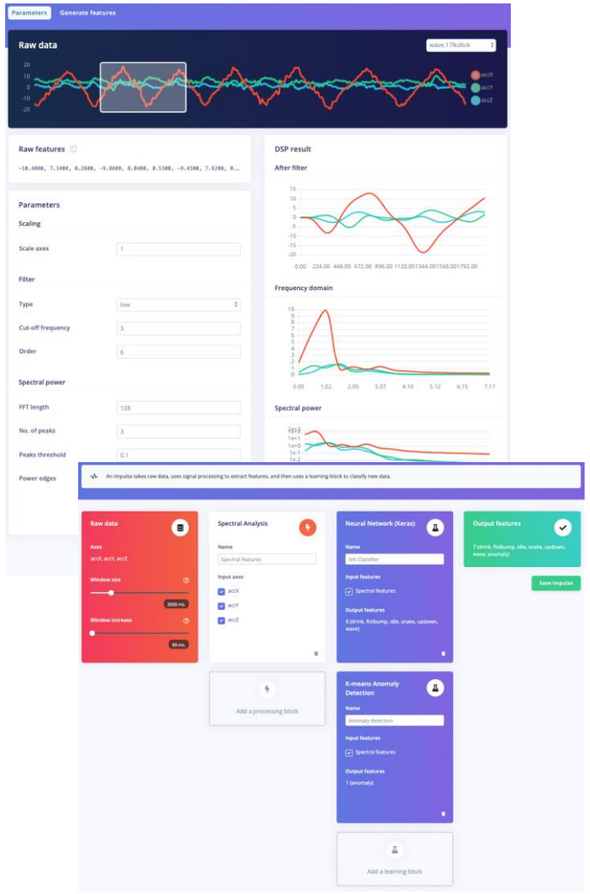
Real sensors in real time
Open source SDK

Impulse

Embedded and edge compute deployment options

Test impulse with real-time device data flows

Test



Get your free account at http://edgeimpulse.com

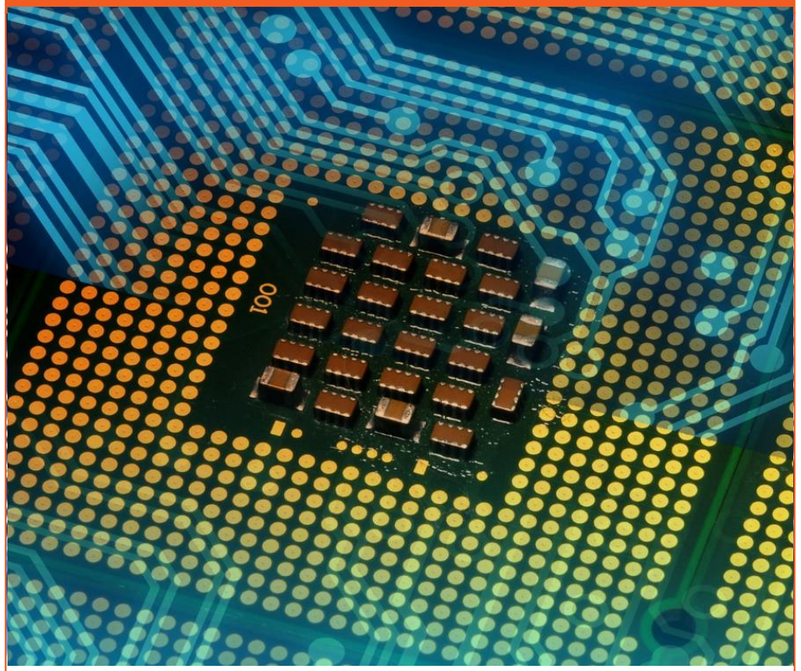
Maxim Integrated: Enabling Edge Intelligence

www.maximintegrated.com/ai

Sensors and Signal Conditioning

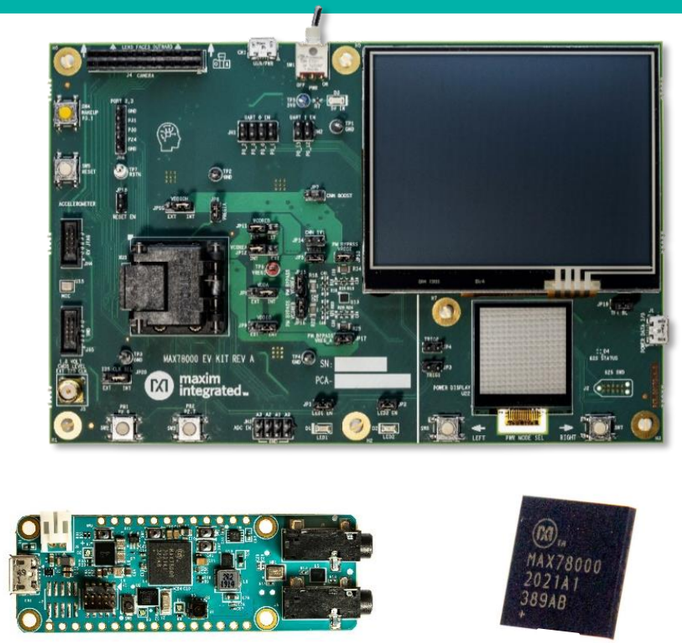
Health sensors measure PPG and ECG signals critical to understanding vital signs. Signal chain products enable measuring even the most sensitive signals.

Low Power Cortex M4 Micros



The biggest (3MB flash and 1MB SRAM) and the smallest (256KB flash and 96KB SRAM) Cortex M4 microcontrollers enable algorithms and neural networks to run at wearable power levels

Advanced AI Acceleration



The new MAX78000 implements AI inferences at over 100x lower energy than other embedded options. Now the edge can see and hear like never before.

maxim integrated™

Qeexo AutoML for Embedded AI

Automated Machine Learning Platform that builds tinyML solutions for the Edge using sensor data

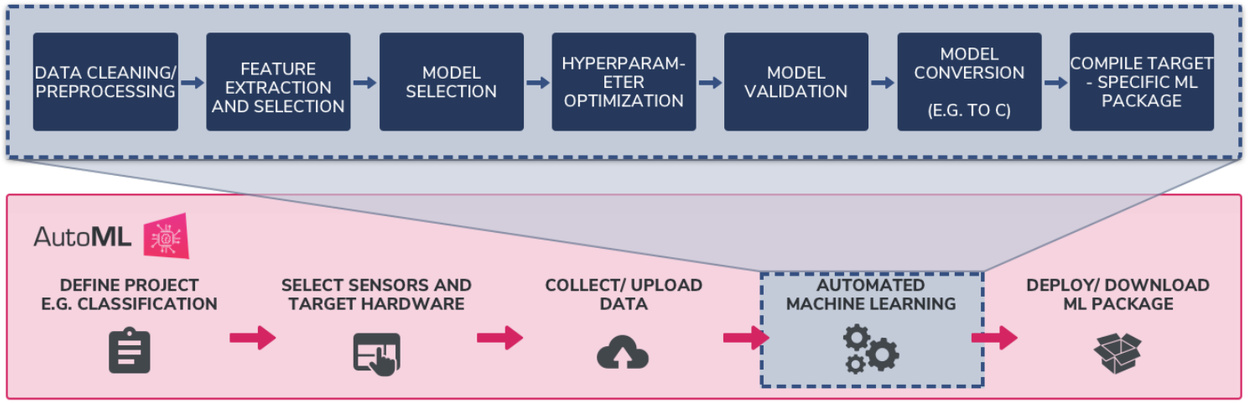
Key Features

- Wide range of ML methods: GBM, XGBoost, Random Forest, Logistic Regression, Decision Tree, SVM, CNN, RNN, CRNN, ANN, Local Outlier Factor, and Isolation Forest
- Easy-to-use interface for labeling, recording, validating, and visualizing time-series sensor data
- On-device inference optimized for low latency, low power consumption, and a small memory footprint
- Supports Arm® Cortex™- M0 to M4 class MCUs
- Automates complex and labor-intensive processes of a typical ML workflow – no coding or ML expertise required!

Target Markets/Applications

- Industrial Predictive Maintenance
- Smart Home
- Wearables
- Automotive
- Mobile
- IoT

QEEXO AUTOML: END-TO-END MACHINE LEARNING PLATFORM



For a limited time, sign up to use Qeexo AutoML at automl.qeexo.com for FREE to bring intelligence to your devices!

Reality AI®

is for building products

https://reality.ai info@reality.ai @SensorAI Reality AI

Reality AI Tools® software

- Automated Feature Exploration and Model Generation
- Bill-of-Materials Optimization
- Automated Data Assessment
- Edge AI / TinyML code for the smallest MCUs

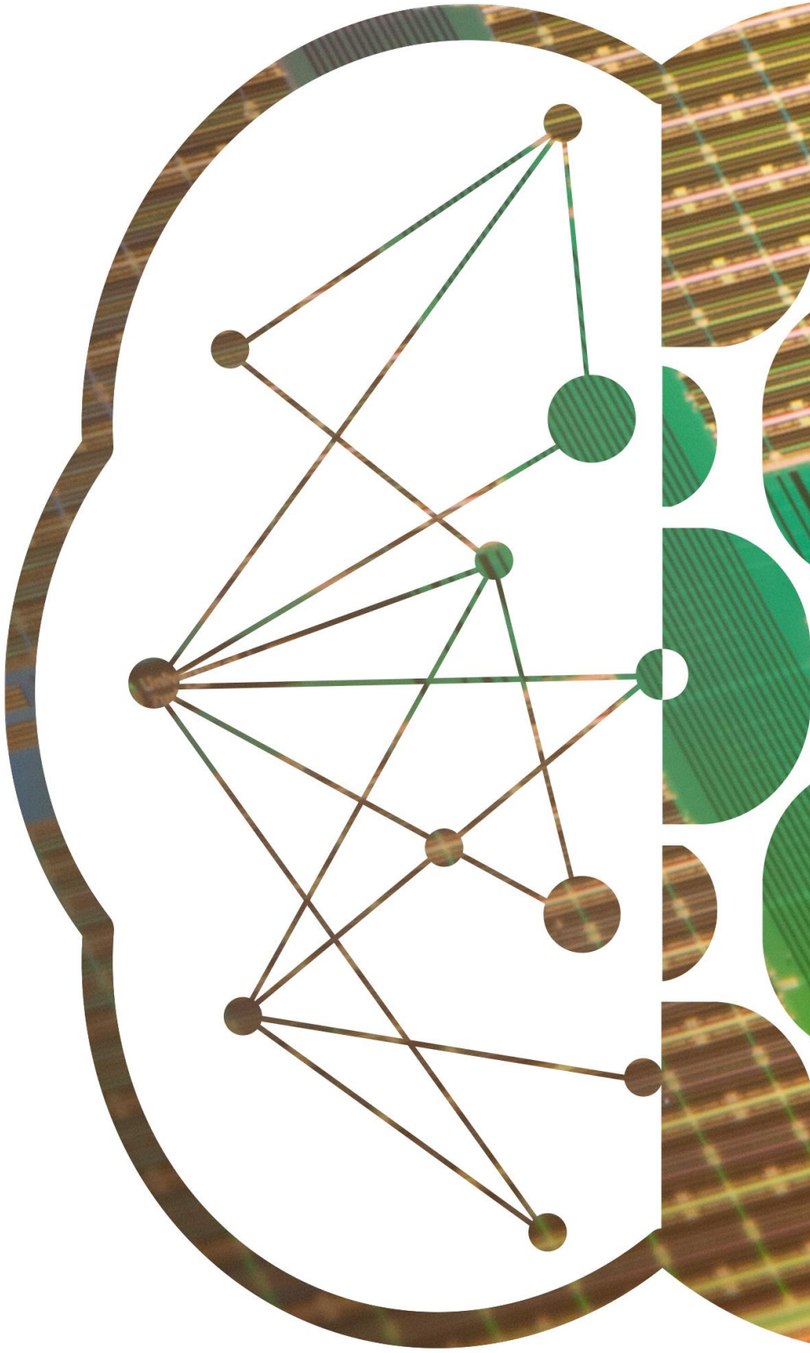
Reality AI solutions

- Automotive sound recognition & localization
- Indoor/outdoor sound event recognition
- RealityCheck™ voice anti-spoofing

SynSense

SynSense builds **ultra-low-power** (sub-mW) **sensing and inference** hardware for **embedded, mobile and edge** devices. We design systems for **real-time always-on smart sensing**, for audio, vision, IMUs, bio-signals and more.

https://SynSense.ai



Next tinyML Talks

Date	Presenter	Topic / Title
Tuesday, March 2	**Eben Upton** founder of the Raspberry Pi Foundation	Inference with Raspberry Pi Pico and RP2040
tinyML UK Tuesday, April 20	**Arduino**	Talk on ML on Arduino platforms

Webcast start time is 8 am Pacific time

Please contact talks@tinyml.org if you are interested in presenting

Announcement

www.tinyML.org/summit2021

Highlights:

- Keywords: Premier Quality, Interactive, LIVE … and FREE
- 5 days, 50+ presentations
- 4 Tutorials
- 2 Panel discussions: (i) VC and (ii) tinyML toolchains
- tinyML Research Symposium
- Late Breaking News
- 3 Best tinyML Awards (Paper, Product, Innovation)
- 10+ Breakout sessions on various topics
- tinyML Partner sessions
- tinyAI for (Good) Life
- LIVE coverage, starting at 8am Pacific time

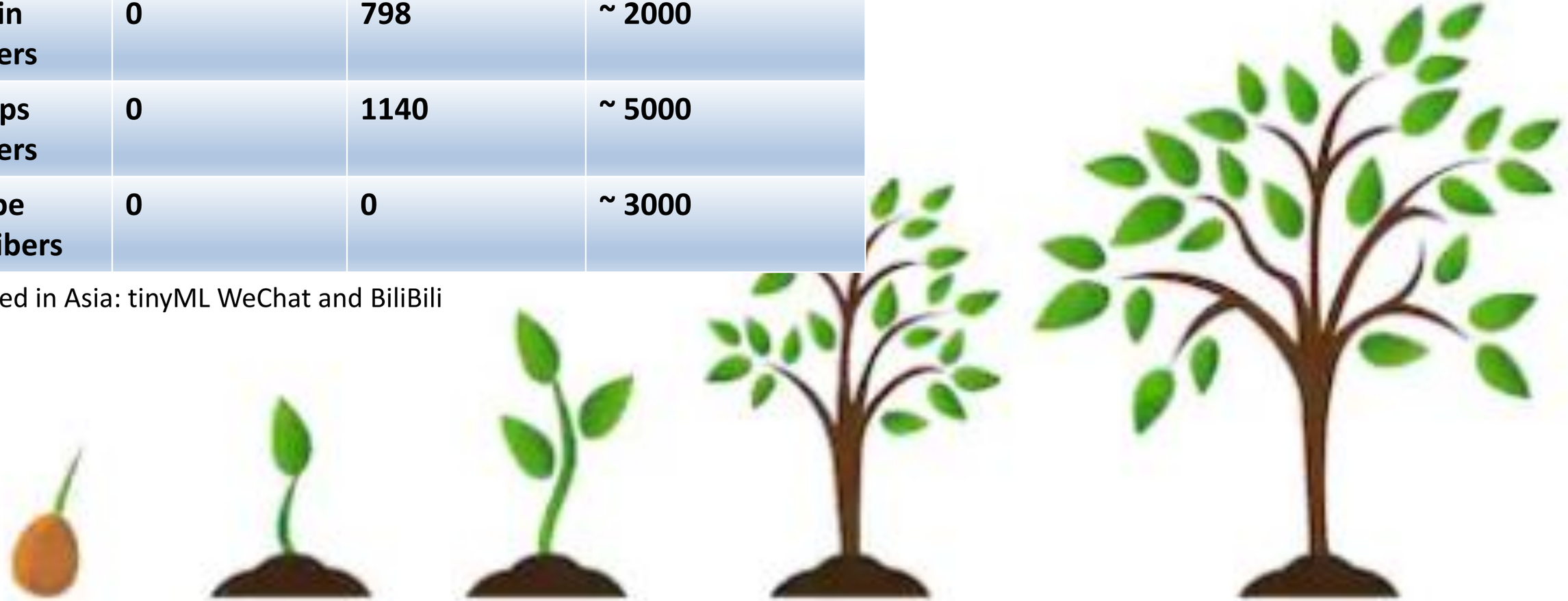
What should I do about it:

- Check out the program – you will be impressed
- Register on-line (takes 5 min)
- If interested: Submit nominations for Best Awards and/or Late News – February 28 deadline
- Block out your calendar: March 22-26
- Become a sponsor (sponsorships@tinyML.org)
- Actively participate at the Summit
- Provide your feedback – we listen !
- Don’t worry about missing some talks – all videos will be posted on YouTube.com/tinyML

tinyML is growing fast

	2019 Summit *(March 2019)*	2020 Summit *(Feb 2020)*	2021 Summit *(March 2021), expected*
Attendees	**160**	**400+**	**3000+**
Companies	**90**	**172**	**300+ (?)**
Linkedin members	**0**	**798**	**~ 2000**
Meetups members	**0**	**1140**	**~ 5000**
YouTube subscribers	**0**	**0**	**~ 3000**

also started in Asia: tinyML WeChat and BiliBili



Summit Sponsors

(as of Feb 15, 2021)

Contact: sponsorships@tinyML.org

multiple levels and benefits available
(also check www.tinyML.org)

TINY ML TALKS webcast

tinyML UK Committee

Alessandro Grande

Developer advocate & ecosystem manager, Arm

Dominic Binks

VP Technology, Audio Analytic

Gian Marco Iodice

ML Techlead, Arm

Neil Cooper

VP Marketing, Audio Analytic

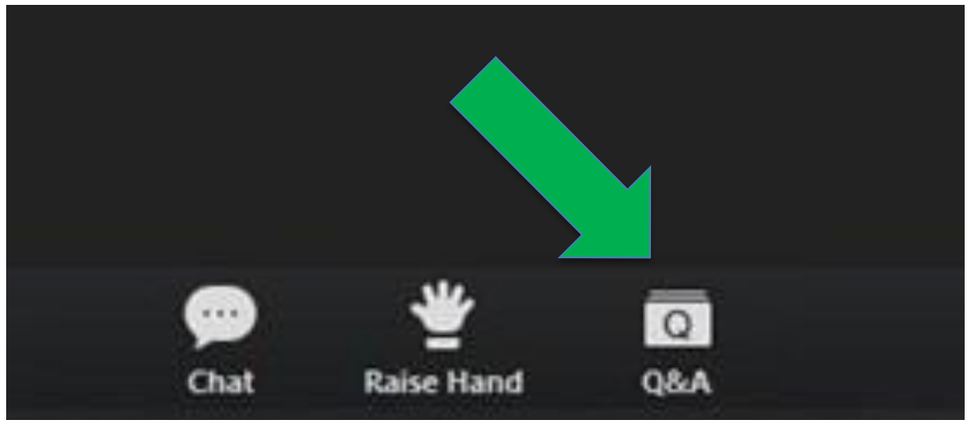
Reminders

Slides & Videos will be posted tomorrow

tinyml.org/forums

youtube.com/tinyml

Please use the Q&A window for your questions



TINY ML TALKS webcast

Dominic Binks

Dr Dominic Binks was previously a Staff Engineer at Qualcomm working in a variety of different software roles prior to joining Audio Analytic. At Qualcomm in Cambridge, he worked on mShop, a BREW-based shopping application and Vuforia, Qualcomm's cross-platform augmented reality SDK. In addition Dominic spent time in San Diego working on Qualcomm's core Android porting team with responsibility for the build and release team. Prior to Qualcomm, Dominic worked in technical presales at SavaJe, Android's forerunner. Before joining SavaJe Dominic worked as a technical consultant at Scientific Generics (now Sagentia) and prior to this he worked on pre-paid calling platforms deployed to a number of mobile operators worldwide. Dominic's PhD investigated techniques for automating fault finding (debugging) in pieces of software.

Making ML Work in the Real World

Dr Dominic Binks

VP Technology

February 2021

Giving machines a sense of hearing

“Like a Shazam for real-world sounds”

Bloomberg

as featured in

THE TIMES

The Economist

WIRED

FAST COMPANY

Forbes

BBC

IEEE SPECTRUM

engadget

techradar.

Gartner

Machine Learning in the real world is hard

Artificial intelligence / Machine learning

Google's medical AI was super accurate in a lab. Real life was a different story.

If AI is really going to make a difference to patients we need to know how it works when real humans get their hands on it, in real situations.

https://www.technologyreview.com/2020/04/27/1000658/google-medical-ai-accurate-lab-real-life-clinic-covid-diabetes-retina-disease/

Underspecification Presents Challenges for Credibility in Modern Machine Learning

Alexander D'Amour*	ALEXDAMOUR@GOOGLE.COM
Katherine Heller*	KHELLER@GOOGLE.COM
Dan Moldovan*	MDAN@GOOGLE.COM
Ben Adlam	ADLAM@GOOGLE.COM
Babak Alipanahi	BABAKA@GOOGLE.COM
Alex Beutel	ALEXBEUTEL@GOOGLE.COM
Christina Chen	CHRISTINIUM@GOOGLE.COM
Jonathan Deaton	JDEATON@GOOGLE.COM
Jacob Eisenstein	JEISENSTEIN@GOOGLE.COM
Matthew D. Hoffman	MHOFFMAN@GOOGLE.COM
Farhad Hormozdiari	FHORMOZ@GOOGLE.COM
Neil Houlsby	NEILHOULSBY@GOOGLE.COM
Shaobo Hou	SHAOBOHOU@GOOGLE.COM
Ghassen Jerfel	GHASSEN@GOOGLE.COM
Alan Karthikesalingam	ALANKARTHI@GOOGLE.COM
Mario Lucic	LUCIC@GOOGLE.COM
Yian Ma	YIANMA@UCSD.EDU
Cory McLean	CYM@GOOGLE.COM
Diana Mincu	DMINCU@GOOGLE.COM
Akinori Mitani	AMITANI@GOOGLE.COM
Andrea Montanari	MONTANARI@STANFORD.EDU
Zachary Nado	ZNADO@GOOGLE.COM
Vivek Natarajan	NATVIV@GOOGLE.COM
Christopher Nielson†	CHRISTOPHER.NIELSON@VA.GOV
Thomas F. Osborne†	THOMAS.OSBORNE@VA.GOV
Rajiv Raman	DRRRN@SNMAIL.ORG
Kim Ramasamy	KIM@ARAVIND.ORG
Rory Sayres	SAYRES@GOOGLE.COM
Jessica Schrouff	SCHROUFF@GOOGLE.COM
Martin Seneviratne	MARTSEN@GOOGLE.COM
Shannon Sequeira	SHNNN@GOOGLE.COM
Harini Suresh	HSURESH@MIT.EDU
Victor Veitch	VICTORVEITCH@GOOGLE.COM
Max Vladymyrov	MXV@GOOGLE.COM
Xuezhi Wang	XUEZHIW@GOOGLE.COM
Kellie Webster	WEBSTERK@GOOGLE.COM
Steve Yadlowsky	YADLOWSKY@GOOGLE.COM
Taedong Yun	TEDYUN@GOOGLE.COM
Xiaohua Zhai	XZHAI@GOOGLE.COM
D. Sculley	DSCULLEY@GOOGLE.COM

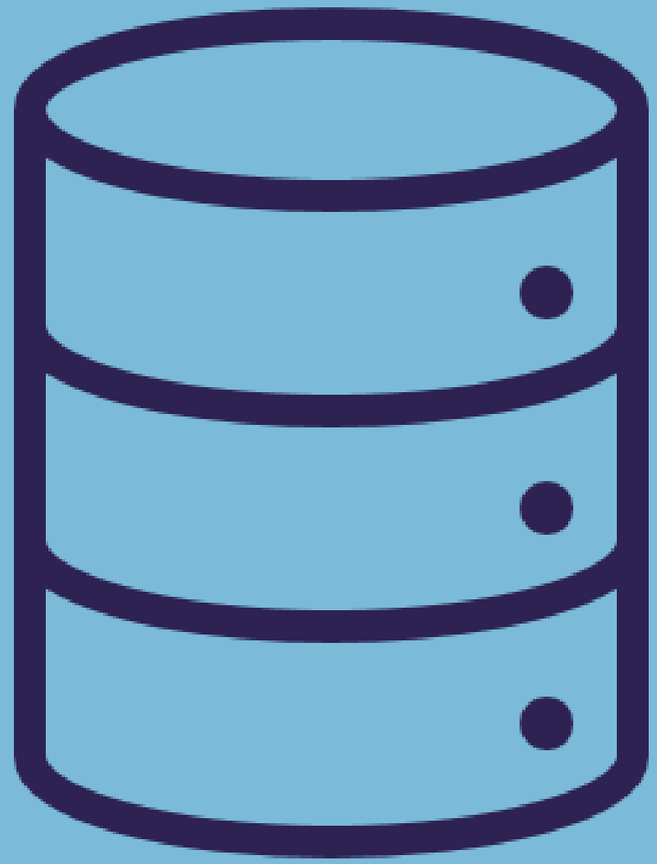
https://arxiv.org/pdf/2011.03395.pdf

Why ML is hard in the real world – some causes

- Unobserved differences
  - Images from one smartphone camera being of very different quality to those from another
- Improbable inputs
  - Share price changes very rapidly
- Unforeseen inputs
  - Parrots in France?
- Combinations
  - Leaves from a tree moving in a camera image
- Over simplification
  - Alexa (the person) verses Alexa (the digital assistant)

- Essentially all examples of under specification

“We say that an ML pipeline is *underspecified* if there are many predictors f that a pipeline could return with similar predictive risk”



How do we mitigate underspecification?

Understanding use case reduces under specification

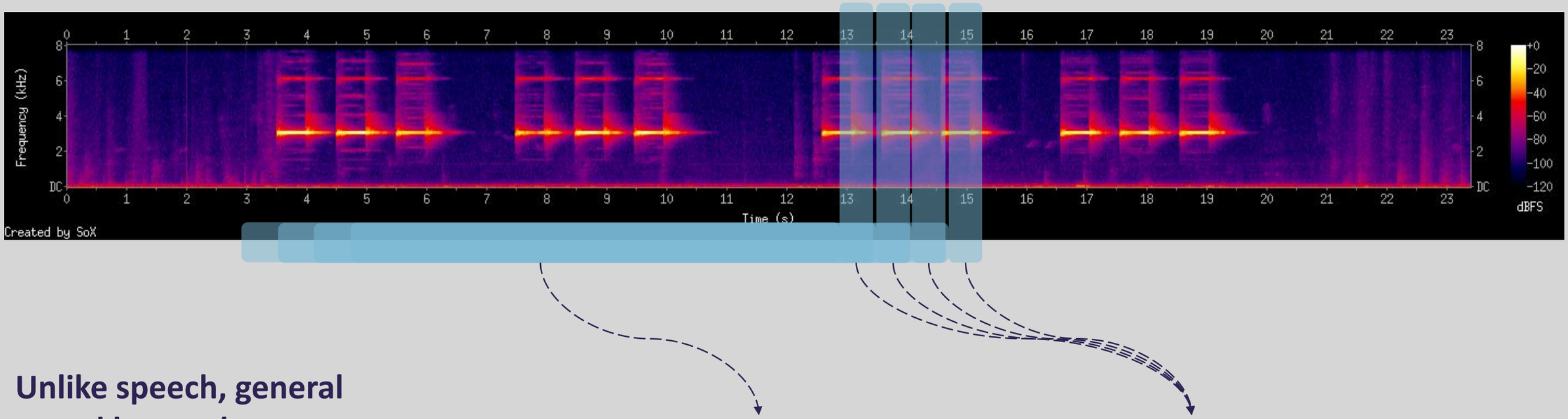
A product is defined by how it is to be *used*

Smoke alarm detection – a use case approach

- What constitutes a correct detection?
- Where will it work?
- When will it work?
- What does it mean to work correctly?
- What’s the budget for processing?
- What’s the desirable latency?
- Who will use it?
- How will a user interact with it?

How could we go about building a smoke alarm detector?

Sound Recognising as an ML Task



Unlike speech, general sound has no *language* model – there is no existing corpus of data with which to train how sound occurs in time

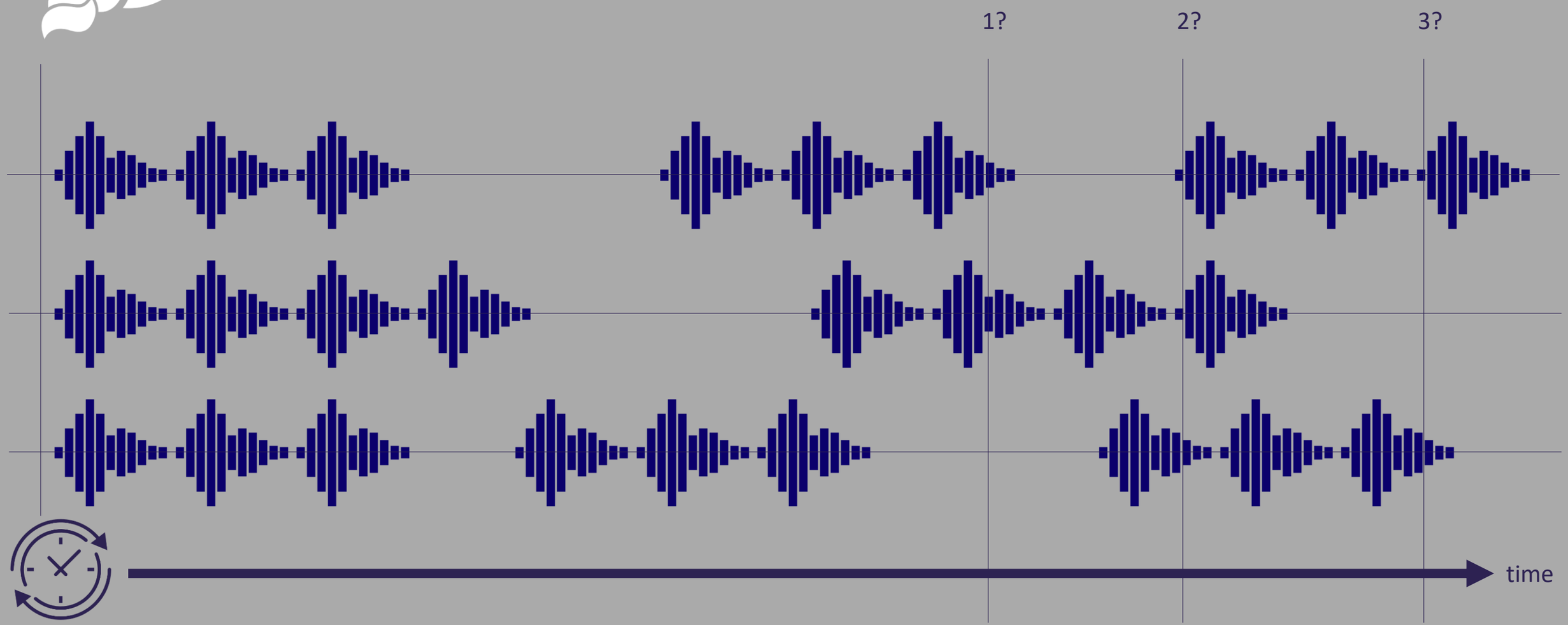
Temporal modelling

Predominantly occurrence of frequency components over longer time window (e.g. sequences of pitch elements, repetitions etc.)

Acoustic modelling

Predominantly frequency composition of a short time window

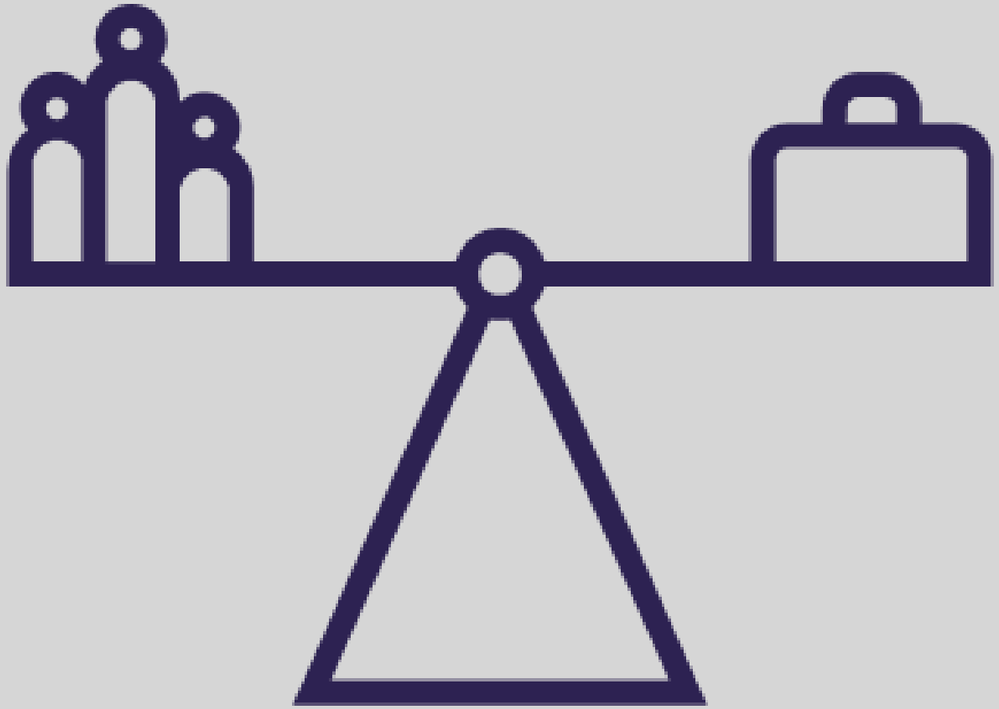
When is a smoke alarm sounding?



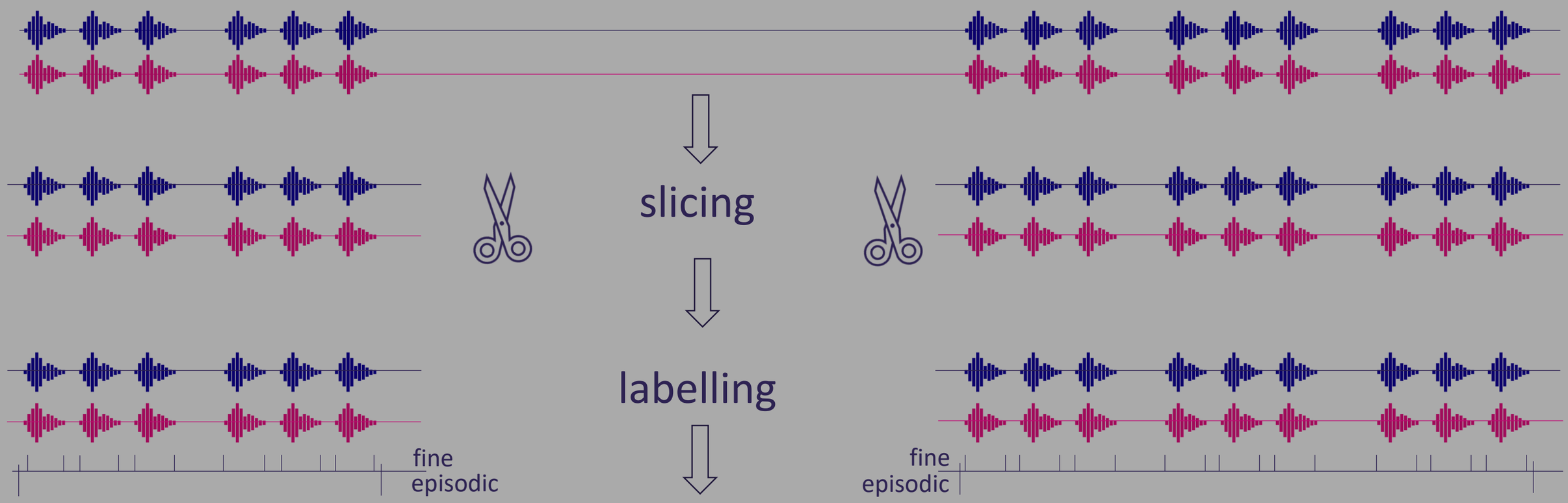
How many smoke alarms do you need?

Where?

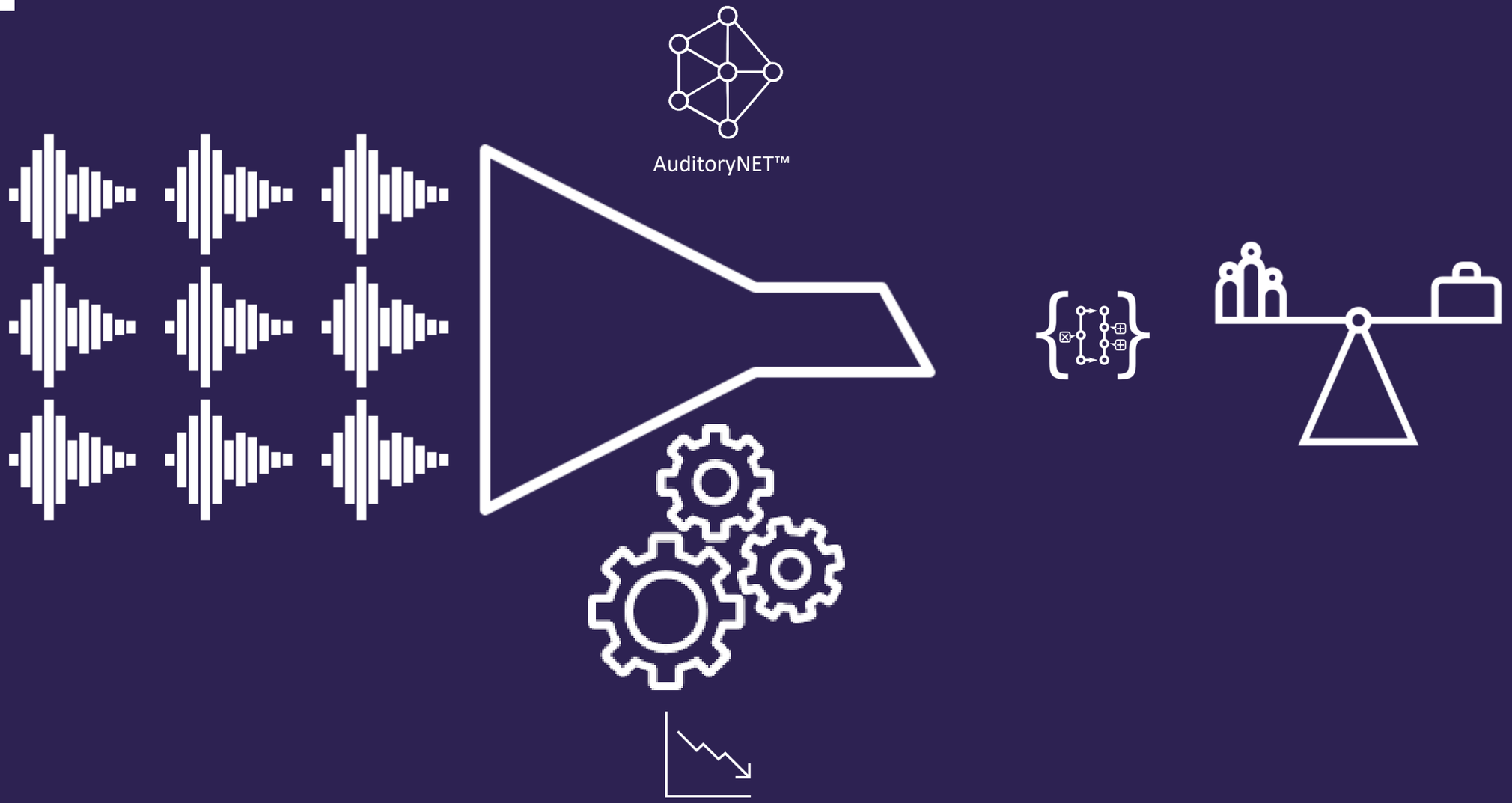
Acceptance



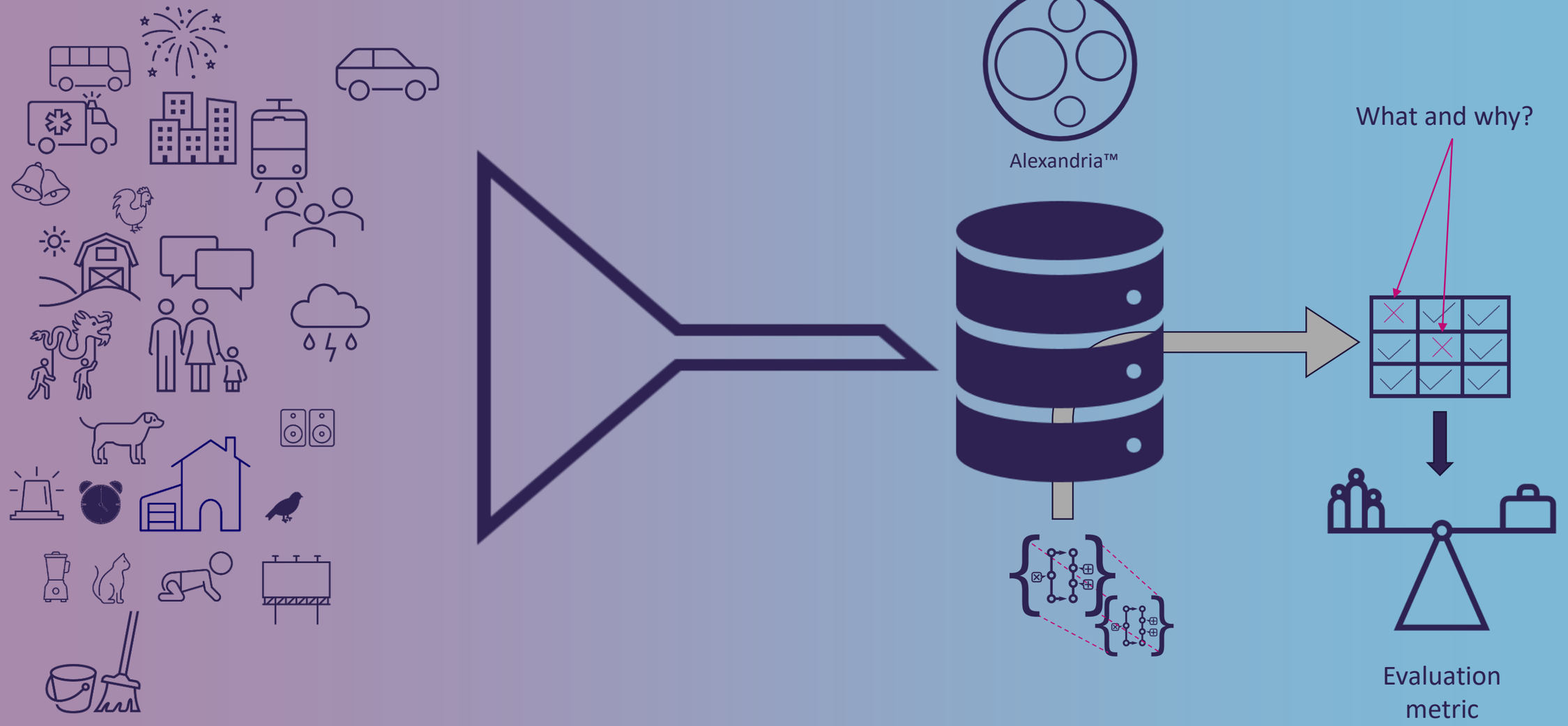
Data Wrangling



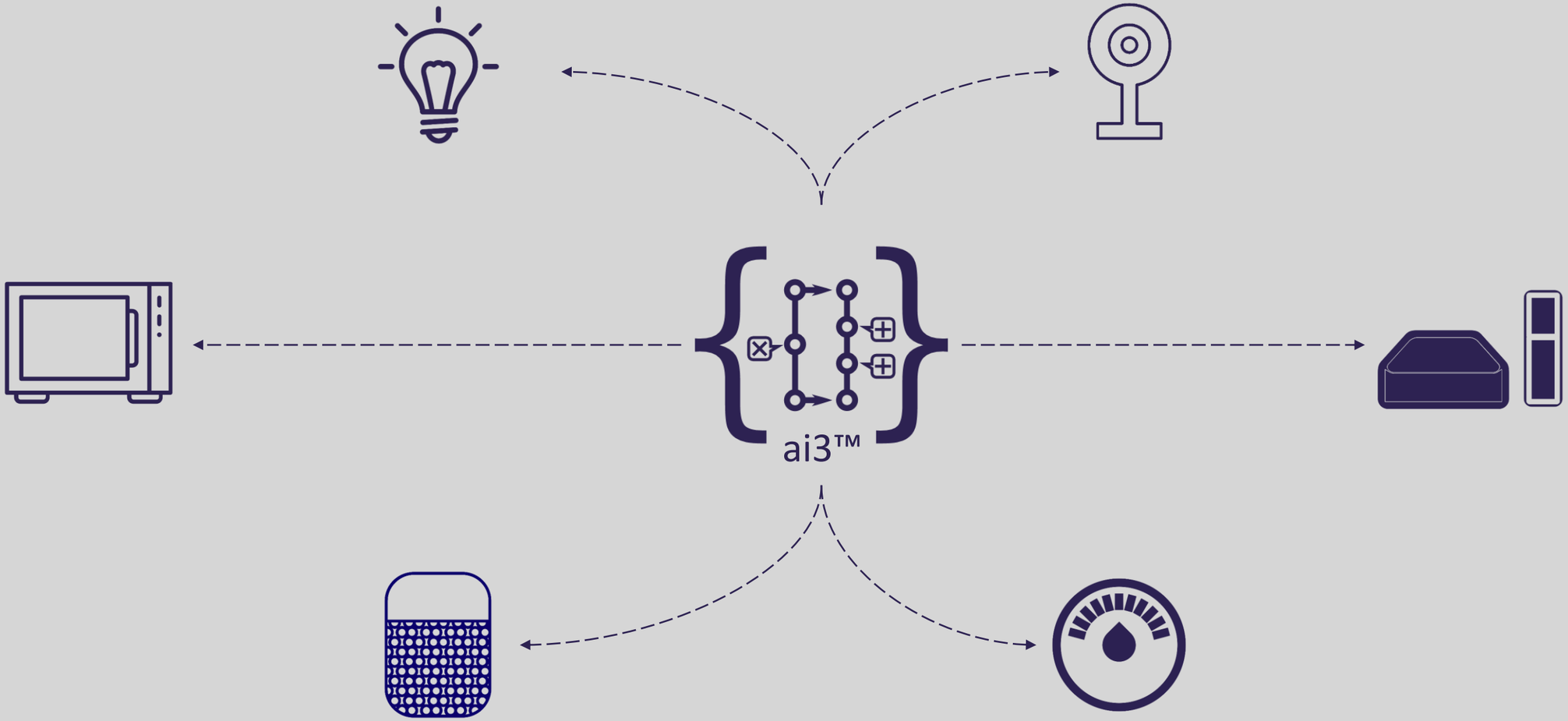
Training



Real World Evaluation



Product



https://youtu.be/ce5y5myvLug

“There are known knowns, things we know that we know; and there are known unknowns, things that we know we don't know. But there are also unknown unknowns, things we do not know we don't know”

Donald Rumsfeld, 2002

Donald Rumsfeld, Hero of ML

	… that we know	… that we don’t know
What we know…	*Assumptions*	*Gaps*
What we don’t know…	*Tacit knowledge*	*Discoveries*

Thank you

info@audioanalytic.com
audioanalytic.com

UK headquarters
2 Quayside
Cambridge
CB5 8AB, UK

Sound Labs
11b Nuffield Road
Cambridge
CB4 1TF, UK

US office
44 Montgomery Street
San Francisco
CA 94104, USA

Copyright Notice

www.tinyML.org