ML Commons

MLCommons™ GSMA Introduction

June 22nd, 2021

Executive Director: David Kanter

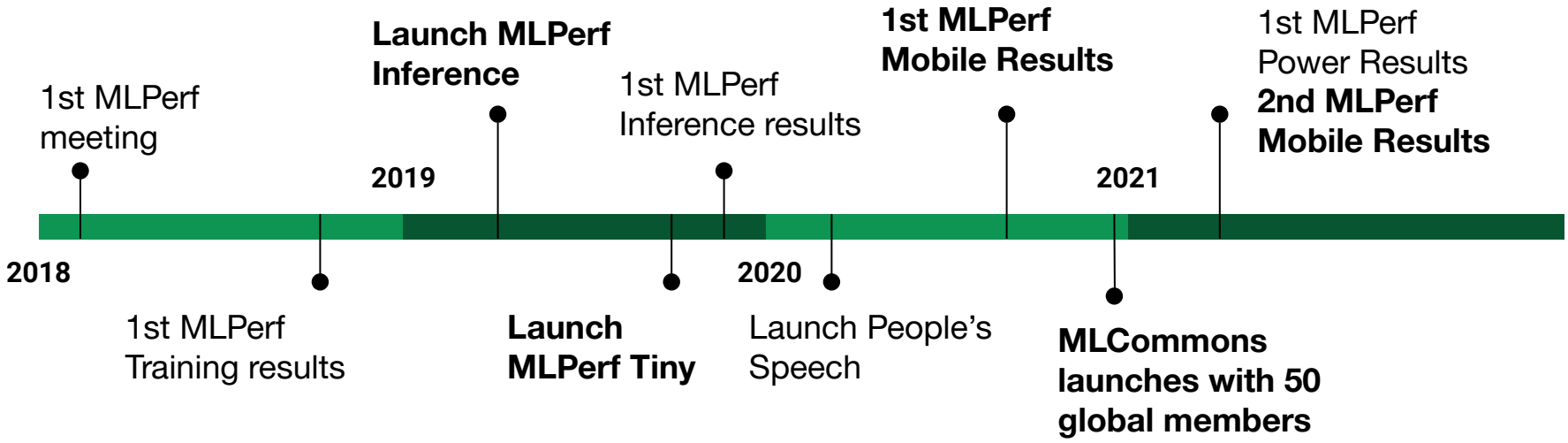
David Kanter is a Founder and the Executive Director of MLCommons where he helps lead the MLPerf benchmarks and other initiatives. He previously led the MLPerf Inference, Mobile, and Power working groups. He has 16+ years of experience in semiconductors, computing, and machine learning. He founded a microprocessor and compiler startup, was an early employee at Aster Data Systems, and has consulted for industry leaders such as Intel, Nvidia, KLA, Applied Materials, Qualcomm, Microsoft and many others. David holds a Bachelor of Science degree with honors in Mathematics with a specialization in Computer Science, and a Bachelor of Arts with honors in Economics from the University of Chicago.

President: Peter Mattson

Peter Mattson co-founded and is President of MLCommons, and co-founded and was General Chair of the MLPerf consortium that preceded it. Previously, he founded the Programming Systems and Applications Group at NVIDIA Research, was VP of software infrastructure for Stream Processors Inc (SPI), and was a managing engineer at Reservoir Labs. His research focuses on understanding machine learning models and data through quantitative metrics and analysis. Peter holds a PhD and MS from Stanford University and a BS from the University of Washington.

From MLPerf™ to MLCommons™

- Started as the unofficial MLPerf consortium
- Grew rapidly into non-profit MLCommons Association



MLCommons is a global community

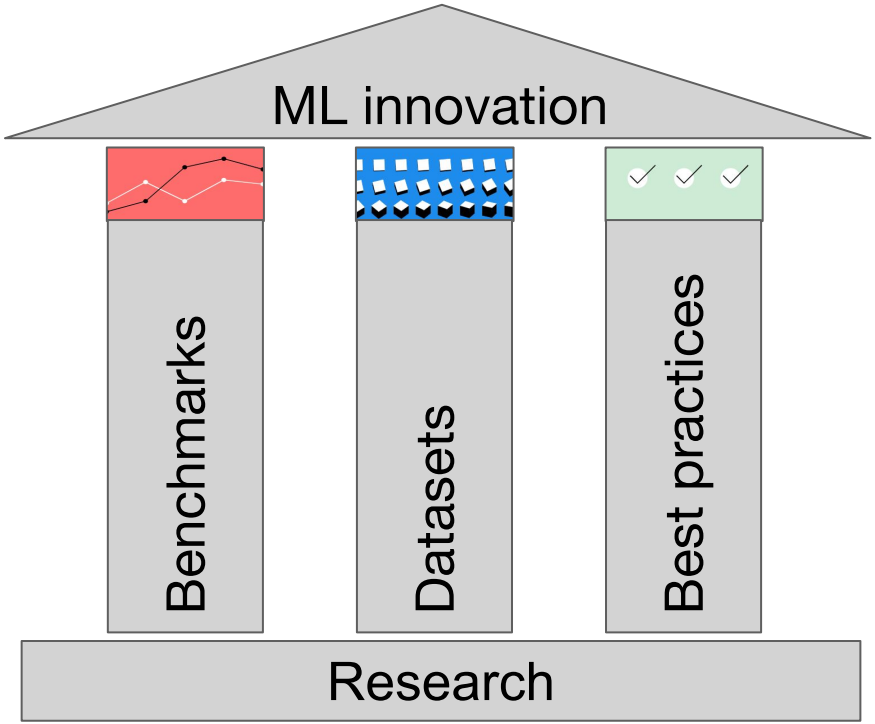
Founding Members

Members

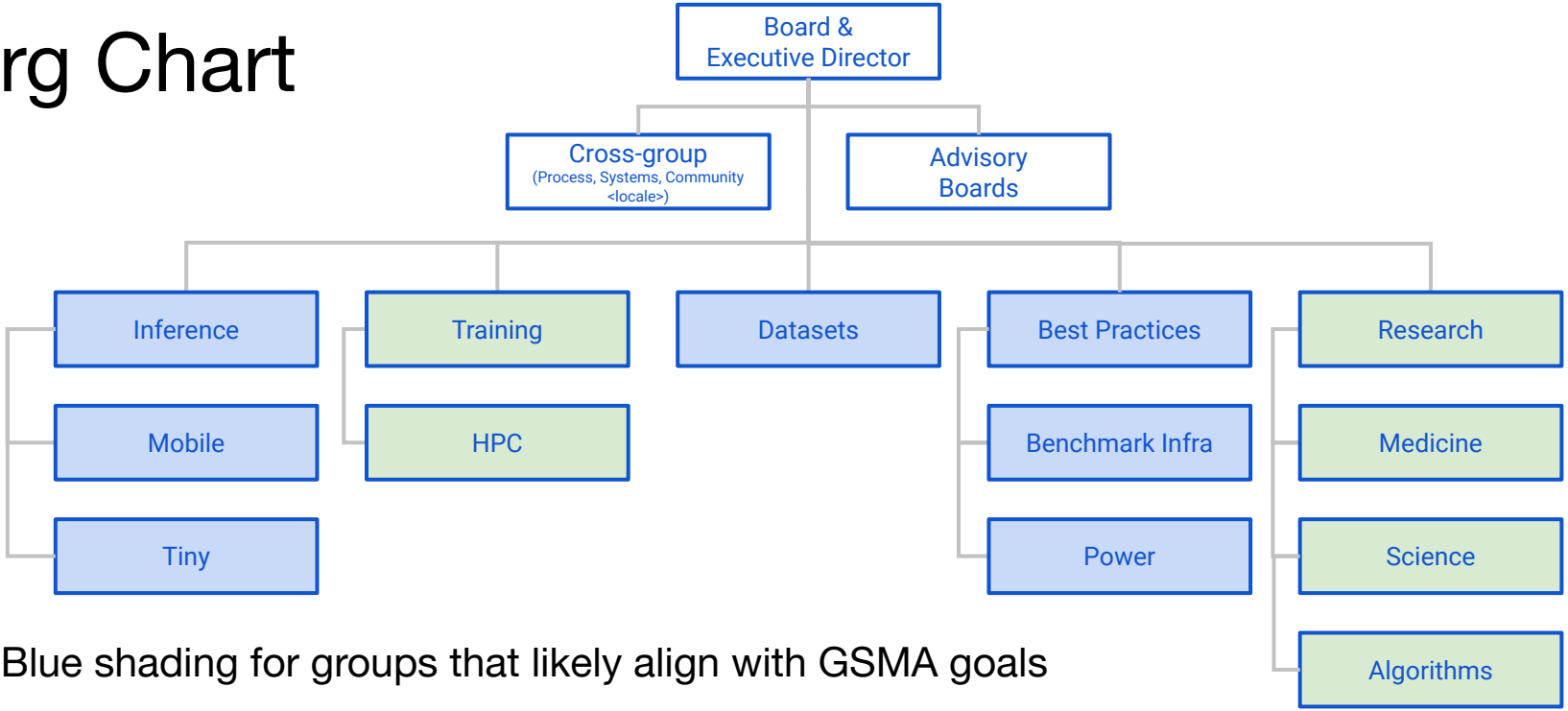
Academics from educational institutions including:

Harvard University
Indiana University
Polytechnique Montreal
Peng Cheng Laboratory
Stanford University
University of California, Berkeley
University of Toronto
University of Tübingen
University of York, United Kingdom
Yonsei University

Mission: Better ML for Everyone



Org Chart



Blue shading for groups that likely align with GSMA goals

Best initial candidates: Mobile, Tiny. Discuss?

MLPerf Benchmark Philosophy

- Goal: fair, useful, and open to help drive the industry forward

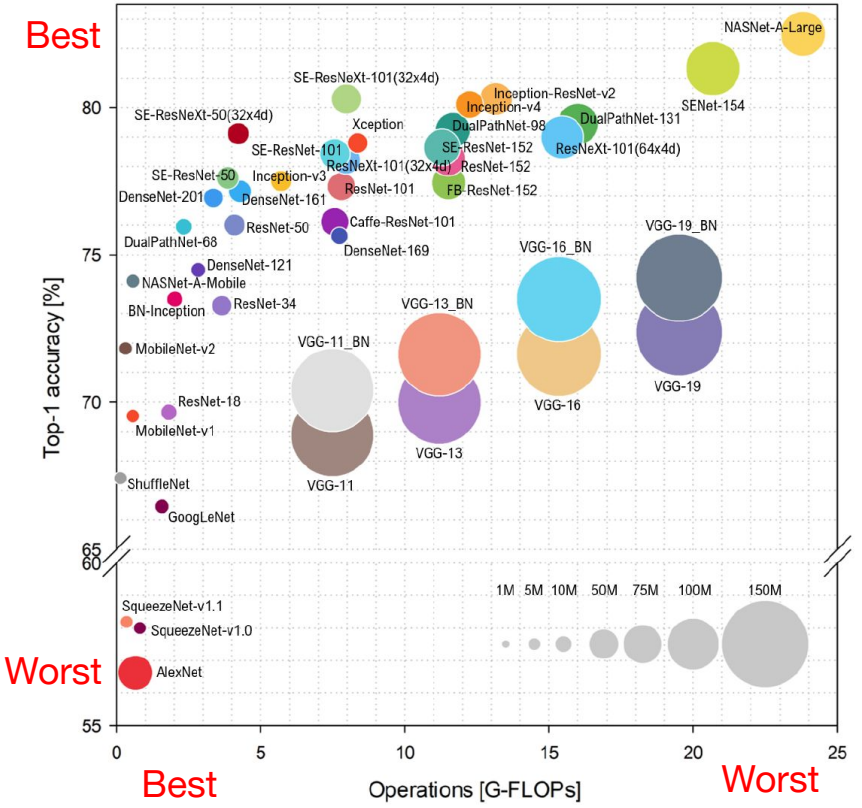
- Machine learning is uniquely challenging
  - Rapidly moving field
  - Dataset challenges around privacy, licensing, size
  - Broad range of applications, e.g., from microwatts to megawatts

- Move fast because ML is rapidly evolving
- Focus on state-of-the-art networks and metrics that measure value to customers

- Great overview from Prof. Dave Patterson: MLPerf: A Benchmark Suite for Machine Learning - David Patterson (UC Berkeley)

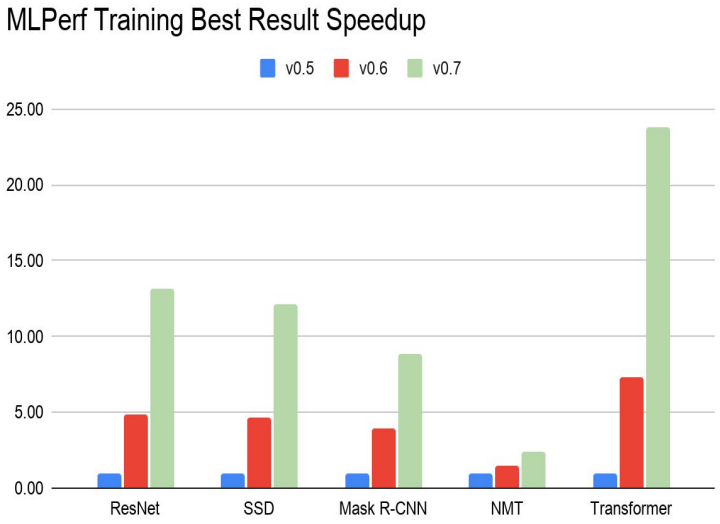
Keeping up with Rapid Evolution

- Pick task and networks that are used widely
  - Ensures relevance
- Require state-of-the-art accuracy
  - Accuracy drives value
  - E.g., ~75% top-1 image classification
- Focus on metrics that convey customer value
  - E.g., inference latency, not FLOP/s



Source: Benchmark Analysis of Representative Deep Neural Network Architectures

Industry standards drive innovation



AI Accelerators: TOPS is Not the Whole Story - EETimes
EE Times • 2 days ago

Intel unveils next-gen Movidius VPU, codenamed Keem Bay
ZDNet • Last month

Centaur Unveils an x86 SoC with Integrated AI Coprocessor
CNX Software • Last month

The MLPerf Consortium, with Members like ARM & Google, have introduced Tech Industry's First Standard ML Benchmark Suit
Patently Apple • Jun 26

MLPerf benchmark results showcase Nvidia's top AI training times
ZDNet

Google Cloud and Nvidia Tesla set new AI training records with MLPerf benchmark results
Packt Hub • Jul 15

Who's Winning the AI Inference Race?
Eetasia.com • Last month

AI Gets Inference Benchmarks
EE Times • Jun 24

Intel, GraphCore And Groq: Let The AI Cambrian Explosion Begin
Forbes • Last month

Centaur announces new SoC featuring an 8-core server-class x86 CPU with AVX512 support and an integrated 20 TOPS AI co-processor
Notebookcheck.net • Last month

MLPerf Releases Five Benchmarks
EE Times India • Jun 26

NVIDIA Corp (NVDA) Q3 2019 Earnings Call Transcript
The Motley Fool • Last month

Twitter wants help with deepfakes, and Microsoft Azure will rent out new AI chips for its cloud users, and more
The Register • Last month

Embedded Benchmark Calls for Support
EE Times • Jun 12

Startup Runs AI in Novel SRAM
EE Times • Jul 22

MLPerf Releases v0.6 Training Results
HPCwire • Jul 10

MLPerf To Provide Much Needed Clarity In The Field Of Machine Learning
Forbes • Jun 25

Digging into MLPerf Benchmark Suite to Inform AI Infrastructure Decisions
HPCwire • Apr 9

MLPerf Is Changing the AI Hardware Performance Conversation. Here's how
Data Center Knowledge • Aug 1

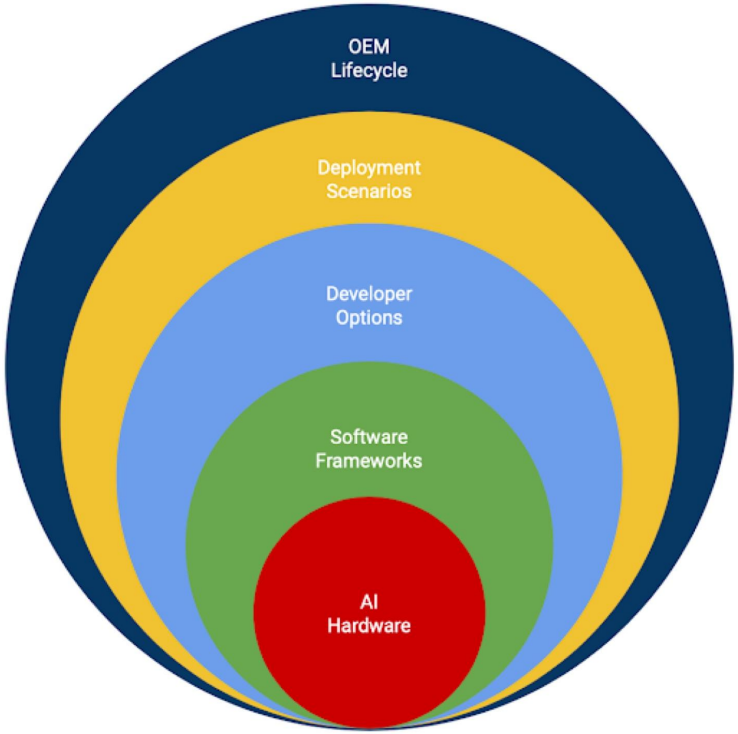
GPUs Continue to Dominate the AI Accelerator Market for Now
InformationWeek • Last month

Nvidia tops AI inference benchmarks, also announces 'world's smallest supercomputer' chip for AI tasks
Firstpost • Last month

Why I joined MLPerf
EE Times • Mar 20

Mobile AI Performance Analysis is Hard

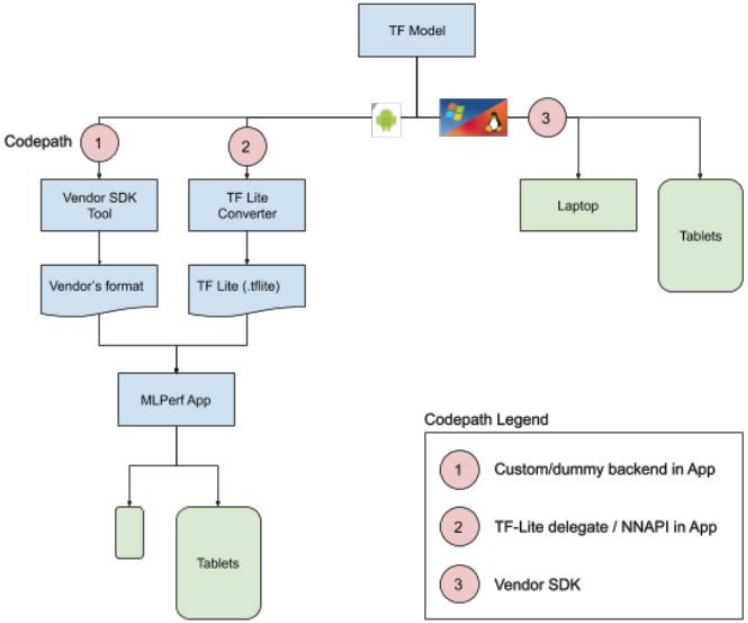
- Multiple different hardware engines
  - CPUs, GPU, DSP, NNAs
- Many different software frameworks
  - E.g., Caffe, TensorFlow
- Lots of different developer options
  - E.g., SDK, platform API
- Lots of different deployment scenarios
  - Known vs. unknown devices
- Controlled by OEM lifecycle management
  - Variable update timing



Mobile Measurement Challenges

AI Hardware x SW Frameworks x Dev. Options x Deployment Scenarios x OEM Lifecycle =

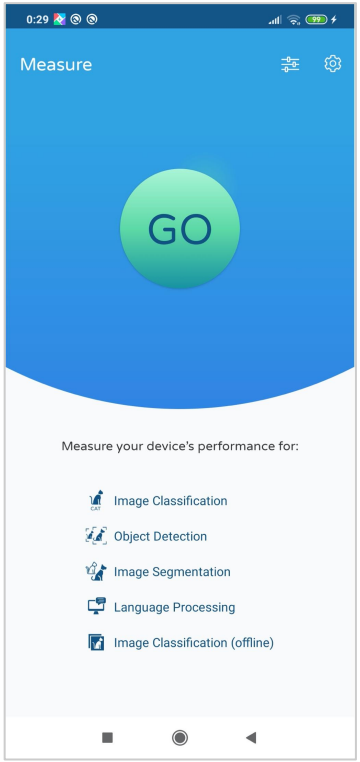
MLPerf Mobile

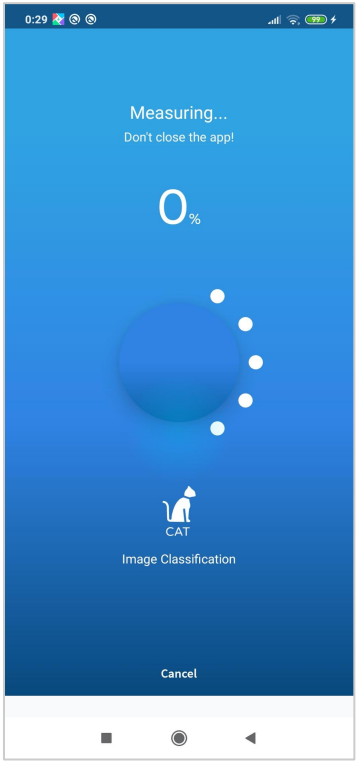


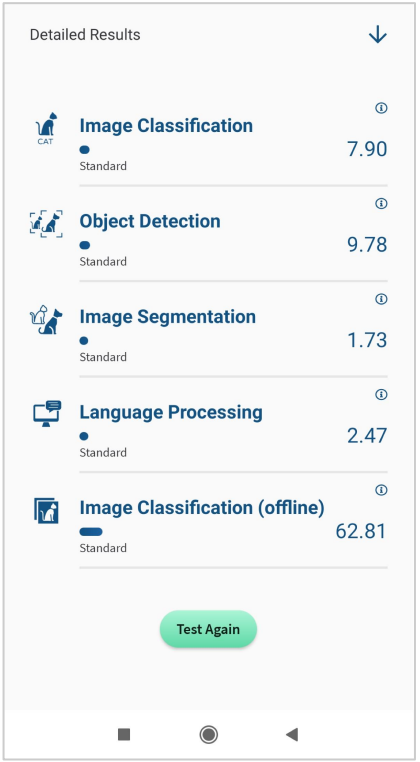
- Initial focus on foundational infrastructure
  - Rules, multi-platform app
  - Small # of use-cases
- Performance is primarily inference latency
  - How fast can I classify an image?
- Collaboration between industry and academia
- **Future evolution?**
  - Open-source in 2021
  - Investigate new tasks / networks
  - Interested in GSMA TS47 usage models

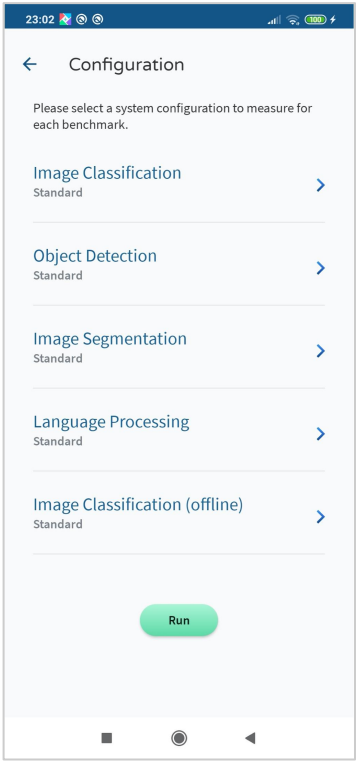
Source: MLPerf Mobile Inference Benchmark

MLPerf Mobile - Already in use









Discussion and next steps?

- What GSMA usage models are important for the future of mobile?
  - TS47 has some nice examples

- Could we adapt MLPerf Inference for machine learning in the network?
  - What about split execution between handset and basestation?

- What would make ML focused benchmarking successful for GSMA?

Backup

Benchmarking breadth: μW to MW

- Agile development approach
- Iterate and build on prior experience
- Add complements, e.g., power/energy metrics

Scale	2018	2019	2020	2021
Training - HPC				
Training				
Inference - Datacenter				
Inference - Edge				
Inference - Mobile				
Inference - Tiny (IoT)				