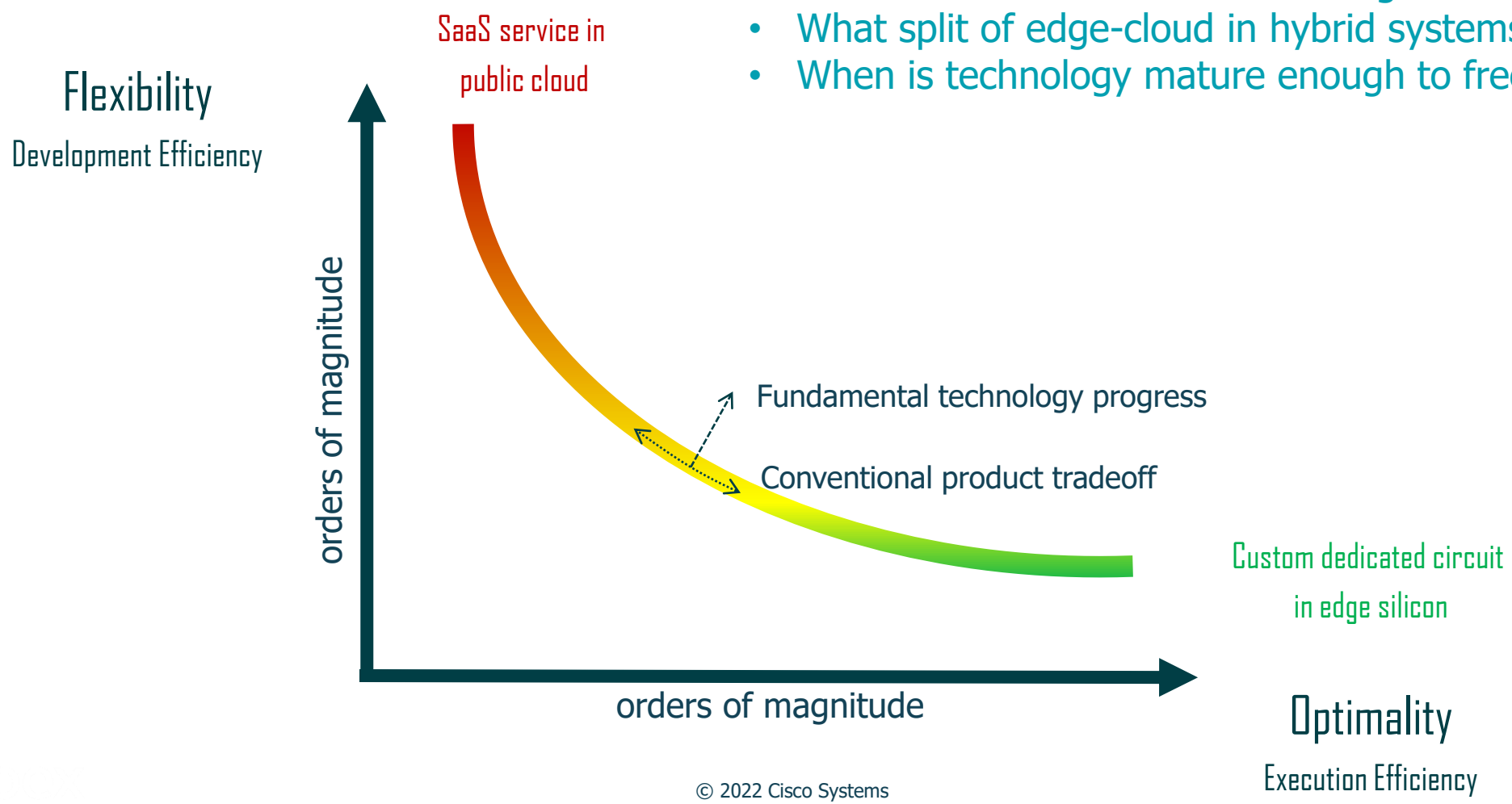
System Imperatives for Audio and Video AI at the Edge

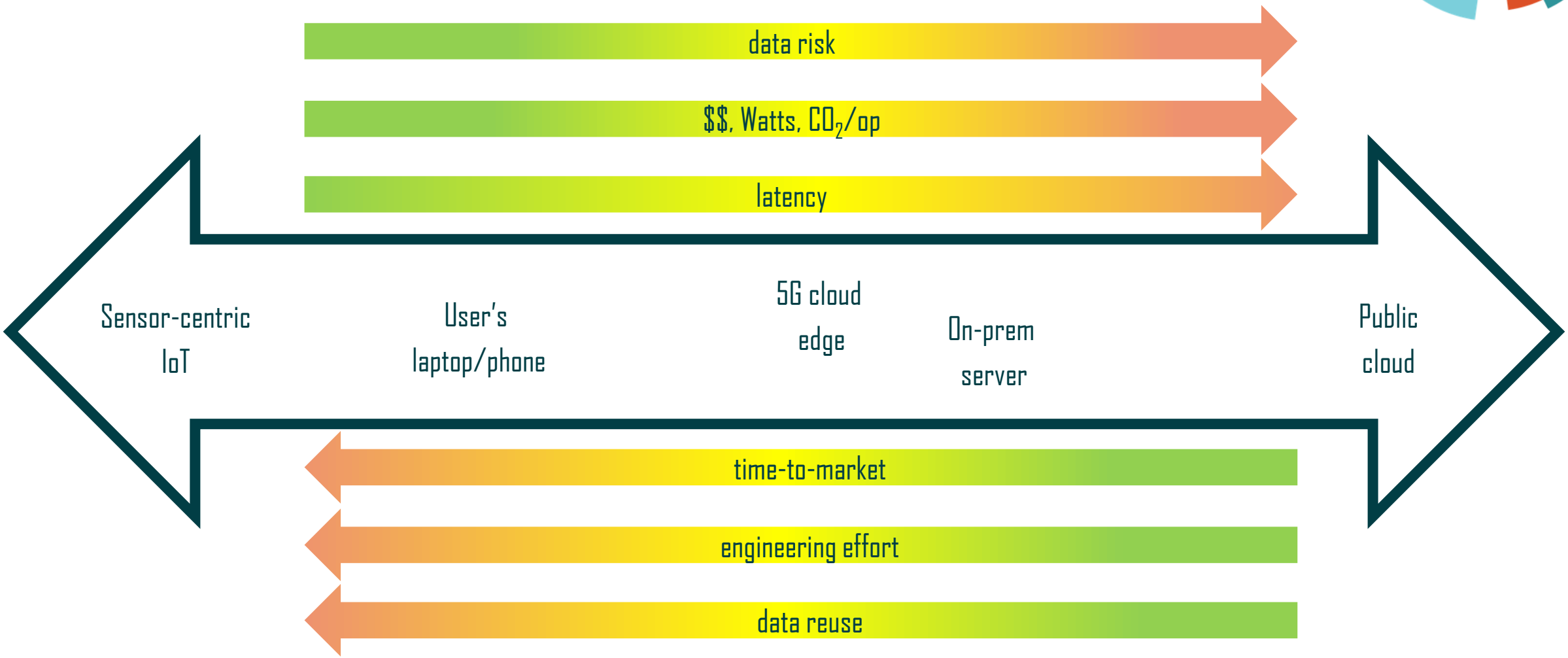
Dr. Chris Rowen
VP of Engineering, Collaboration AI
Cisco Systems

The Grand Tradeoff
The most essential picture in tech

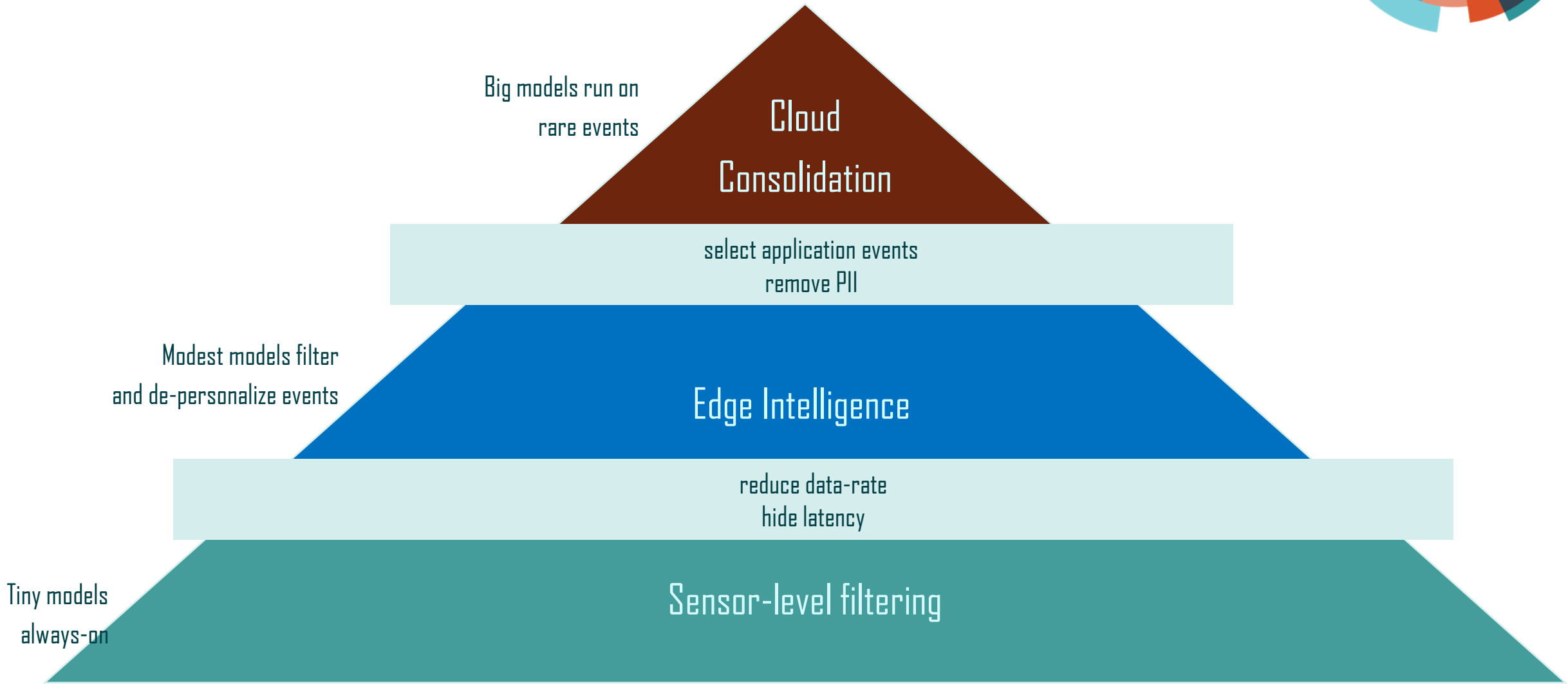
- How much more efficient must edge solutions be?
- What split of edge-cloud in hybrid systems?
- When is technology mature enough to freeze into silicon?



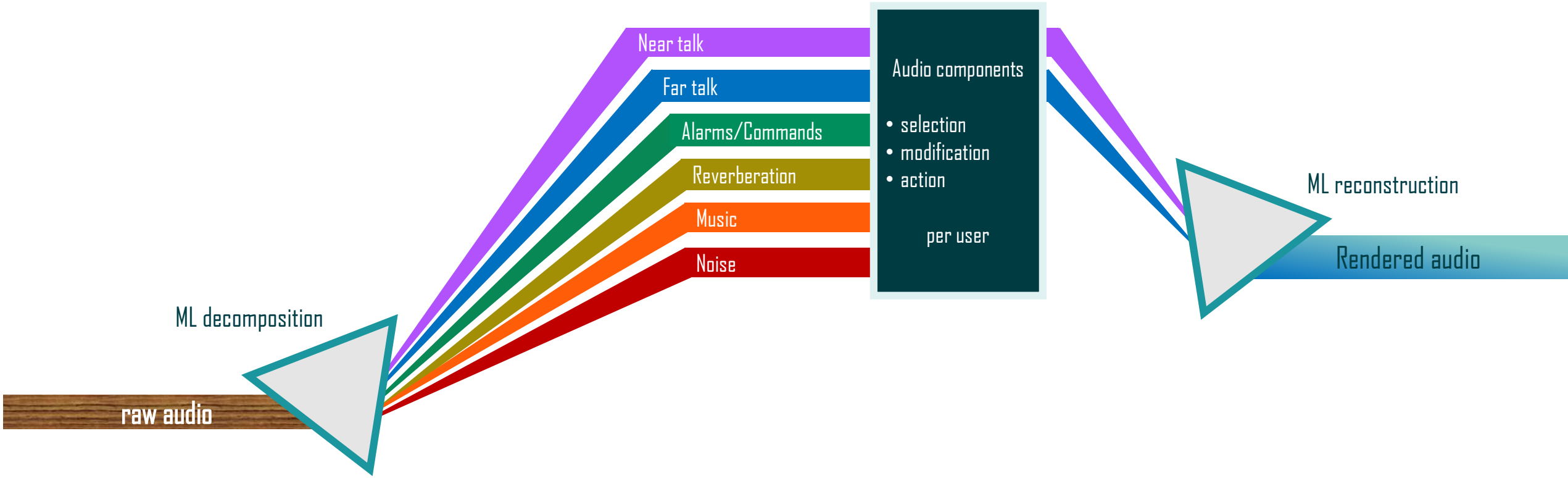
Where to Compute



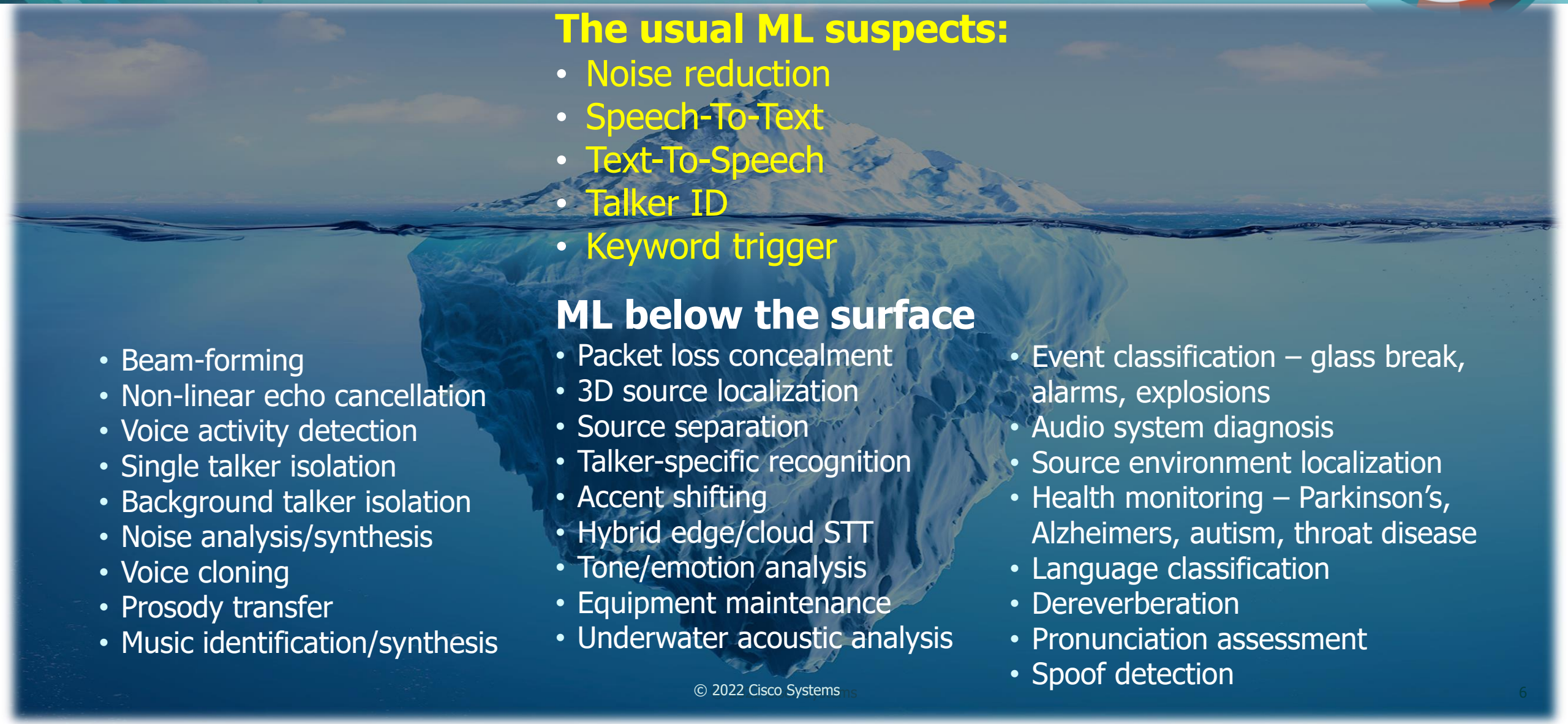
The Cognitive Hierarchy



Rowen's Prism
Decompose-Analyze-Reconstruct Audio



The Audio Iceberg



The usual ML suspects:

- Noise reduction
- Speech-To-Text
- Text-To-Speech
- Talker ID
- Keyword trigger

ML below the surface

- Beam-forming
- Non-linear echo cancellation
- Voice activity detection
- Single talker isolation
- Background talker isolation
- Noise analysis/synthesis
- Voice cloning
- Prosody transfer
- Music identification/synthesis

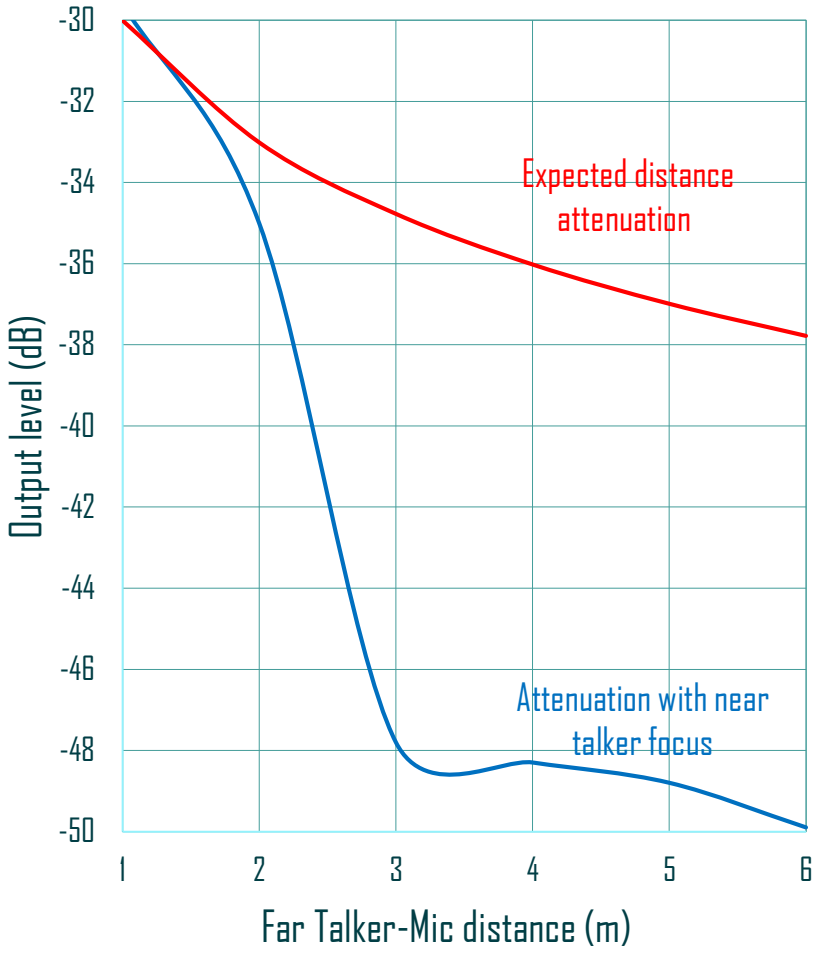
- Packet loss concealment
- 3D source localization
- Source separation
- Talker-specific recognition
- Accent shifting
- Hybrid edge/cloud STT
- Tone/emotion analysis
- Equipment maintenance
- Underwater acoustic analysis

- Event classification – glass break, alarms, explosions
- Audio system diagnosis
- Source environment localization
- Health monitoring – Parkinson's, Alzheimers, autism, throat disease
- Language classification
- Dereverberation
- Pronunciation assessment
- Spoof detection

Webex Audio Demo:
Noise Removal & Talker Selection

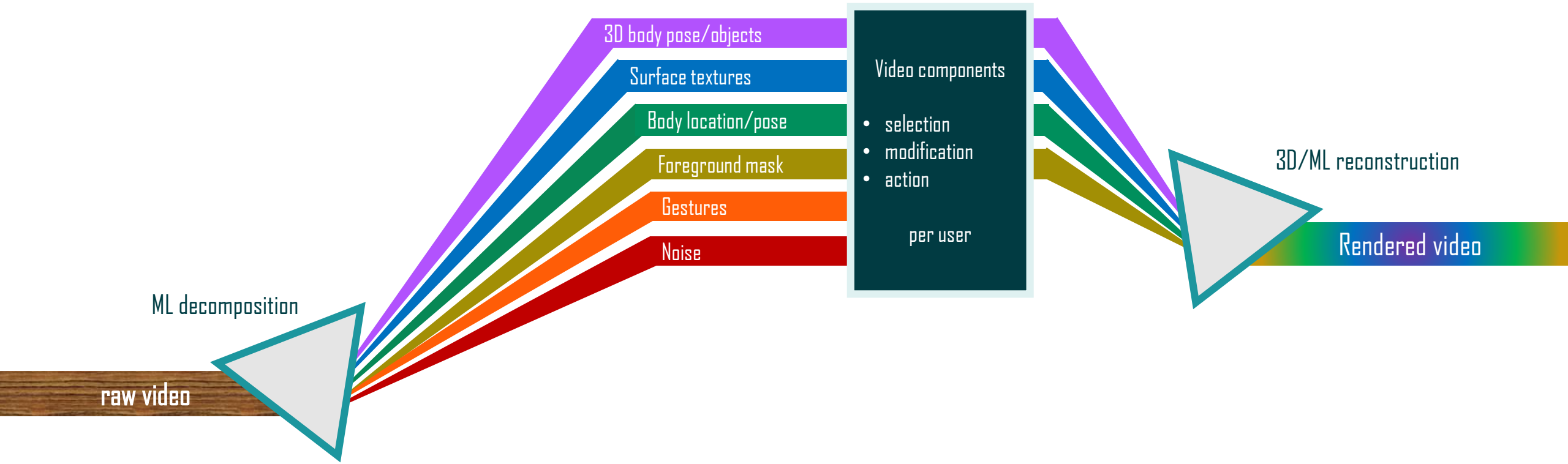
Noise removal (near-talker focus) and speech normalization use-cases

"Optimize for my voice"

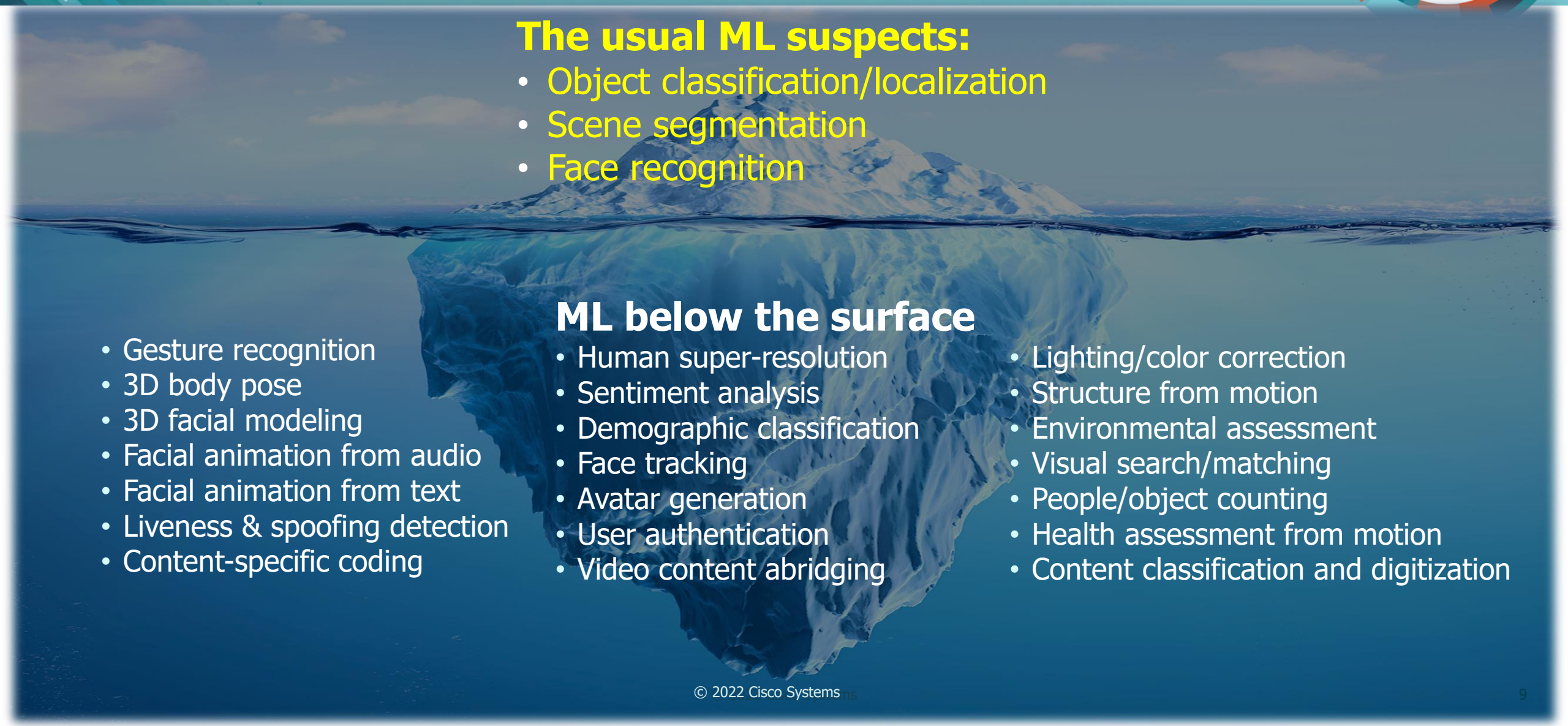


Live measurement on Android/Samsung Galaxy S8

Rowen's Prism
Decompose-Analyze-Reconstruct Video



The Video Iceberg



The usual ML suspects:

- Object classification/localization
- Scene segmentation
- Face recognition

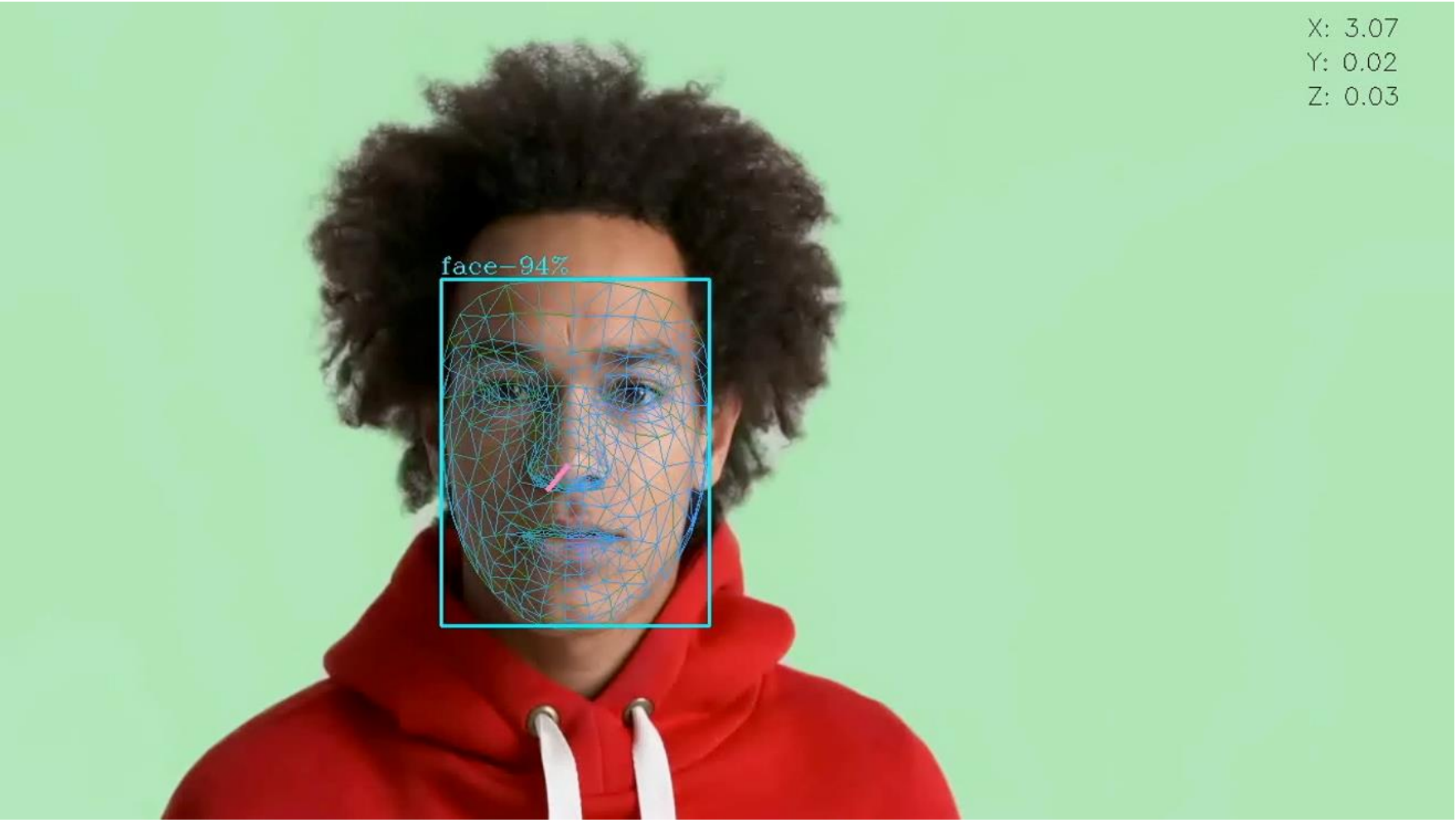
ML below the surface

- Gesture recognition
- 3D body pose
- 3D facial modeling
- Facial animation from audio
- Facial animation from text
- Liveness & spoofing detection
- Content-specific coding

- Human super-resolution
- Sentiment analysis
- Demographic classification
- Face tracking
- Avatar generation
- User authentication
- Video content abridging

- Lighting/color correction
- Structure from motion
- Environmental assessment
- Visual search/matching
- People/object counting
- Health assessment from motion
- Content classification and digitization

Dealing with Overlapping ML Models



Complex systems run multiple parallel ML models

Webex example:

- background segmentation
- rich gestures
- face localization
- 3D model

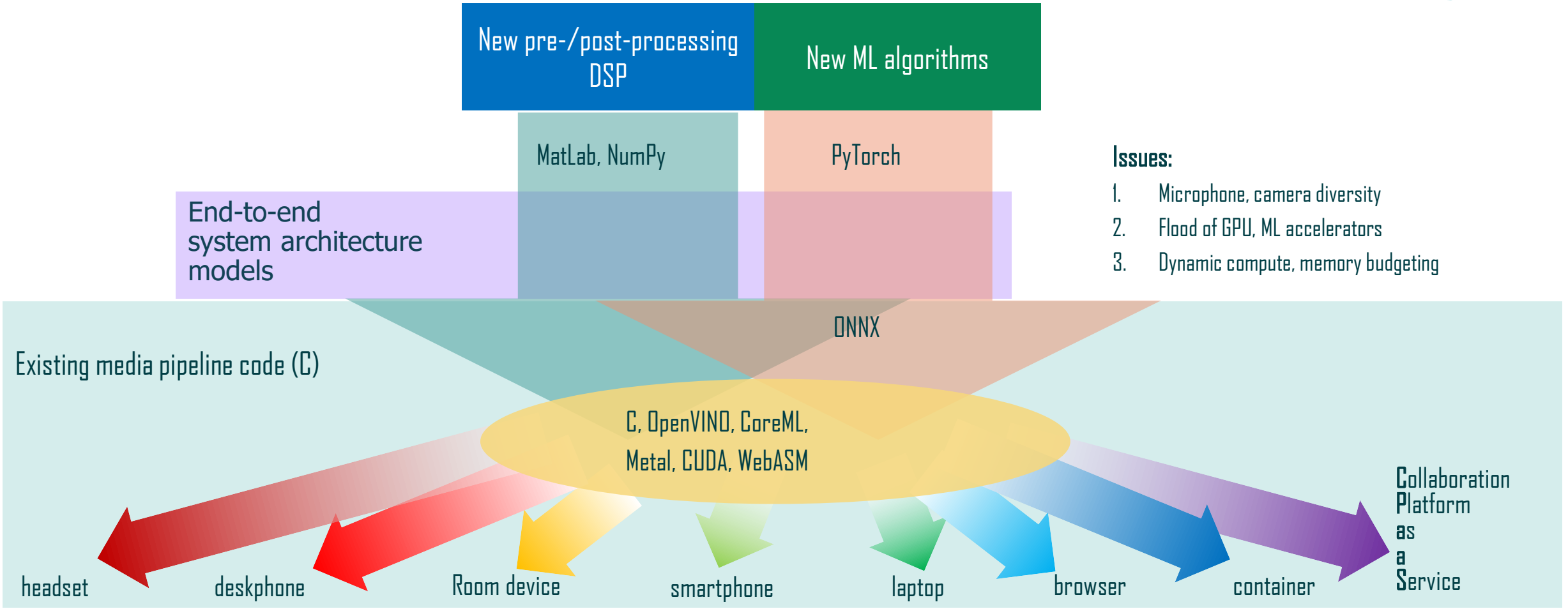
Compete for compute

Challenges in both unified and independent models

When to use ML methods over conventional

1. **Accuracy** matters
2. **Complex** scenarios – "I can't define it but I know it when I see it"
3. Compute/memory **footprint available**: typically > 100 MULs/sample
4. Sufficient **data available**. More data ➔ smaller model
5. **Non-linear** transformation is OK – not feeding 3rd-party ML model

Grand Challenge:
Heterogeneous Media ML Deployment



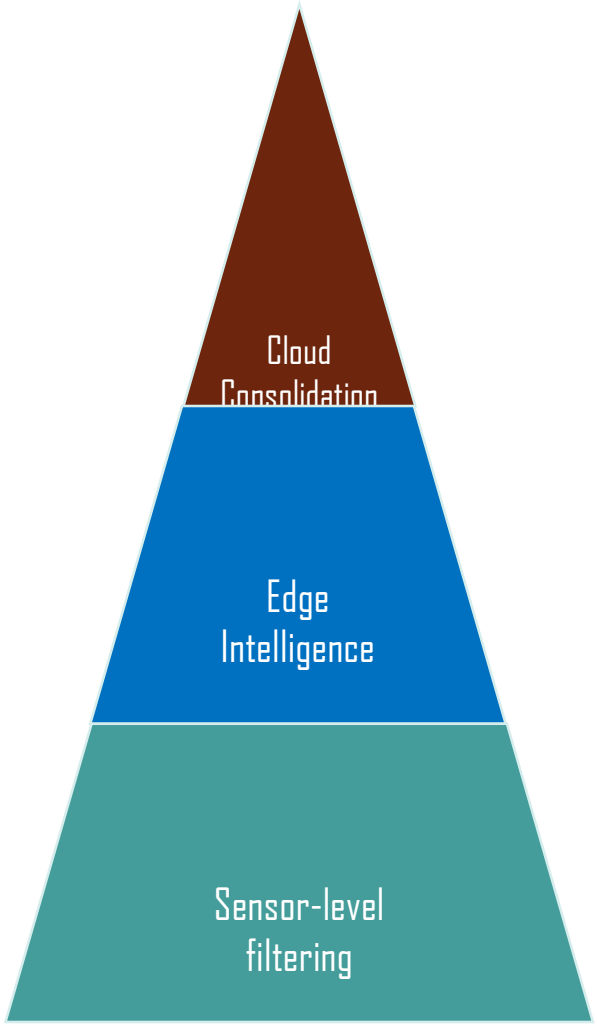
Issues:

1. Microphone, camera diversity
2. Flood of GPU, ML accelerators
3. Dynamic compute, memory budgeting

Interfaces and ML

Stable, exposed interfaces:

- Improve development partitioning and evolution
- May **degrade** cost, power, size, security



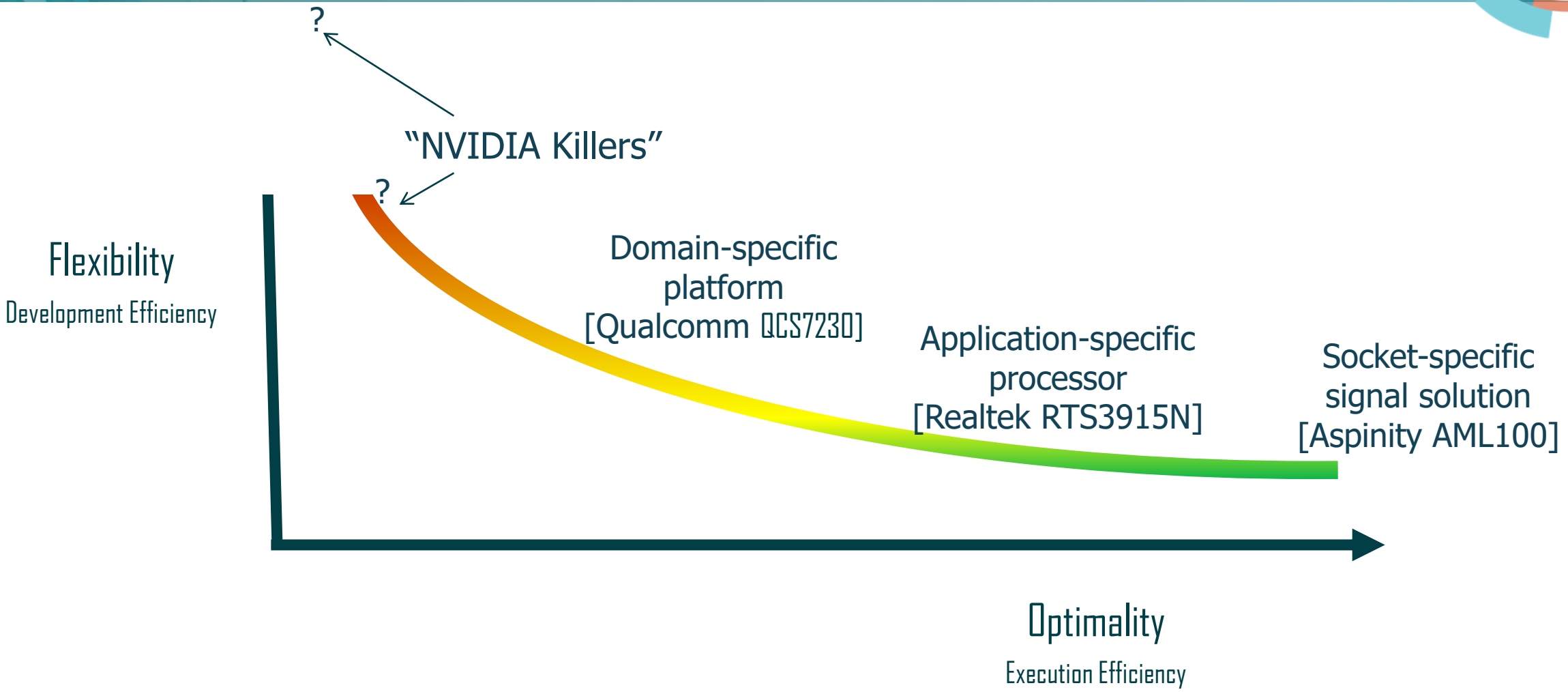
- New database and insight sharing models
- Service federation for regional data compliance (e.g. EU GDPR)

- Improved filtering to reduce cloud bandwidth and compute
- More data de-identification for stricter privacy compliance
- Model improvement within footprint

- Easy sensor device mix-and-match
- Tuning on deployment data
- Adapt to evolving up-link and security profiles

webex

Where Does ML Silicon Fit In?



Guidance

1. Know thy application – accuracy, data, footprint, latency, use-cases
2. Understand tradeoff between development and execution efficiency
   - ➢ Don't freeze a sub-optimal algorithm
3. Better data beats a bigger network
4. Design application hierarchy to move as little data as possible
5. ML Responsibly: Fairness + Transparency + Privacy + Security

Some Resources

- My recent blogs on AI in collaboration: https://blog.webex.com/author/crowen/
- An earlier talk on audio/video ML startups: https://youtu.be/McFCQGO-SoQ
- Cisco's Responsible AI manifesto: https://blogs.cisco.com/security/introducing-cisco-responsible-ai-enhancing-technology-transparency-and-customer-trust
- Pushing ML to ultra-low-power – TinyML: https://www.tinyml.org/about/
- ONNX Tutorials: https://github.com/onnx/tutorials
- Audio ML with Python: https://opensource.com/article/19/9/audio-processing-machine-learning-python
- Video ML with Python: https://www.analyticsvidhya.com/blog/2018/09/deep-learning-video-classification-python/
- Recent funding in AI chip startups: https://www.wsj.com/articles/ai-chip-startups-pull-in-funding-as-they-navigate-supply-constraints-11647338402
- 95 AI chip startups: https://github.com/aolofsson/awesome-semiconductor-startups