GDIT

MeriTalk
Improving the Outcomes of Government IT

EMERGE 2019

ART OF THE POSSIBLE

April 23, 2019

Transforming Through AI

Anthony Dina

Director of Data Analytics / AI, Dell Technologies

Agenda

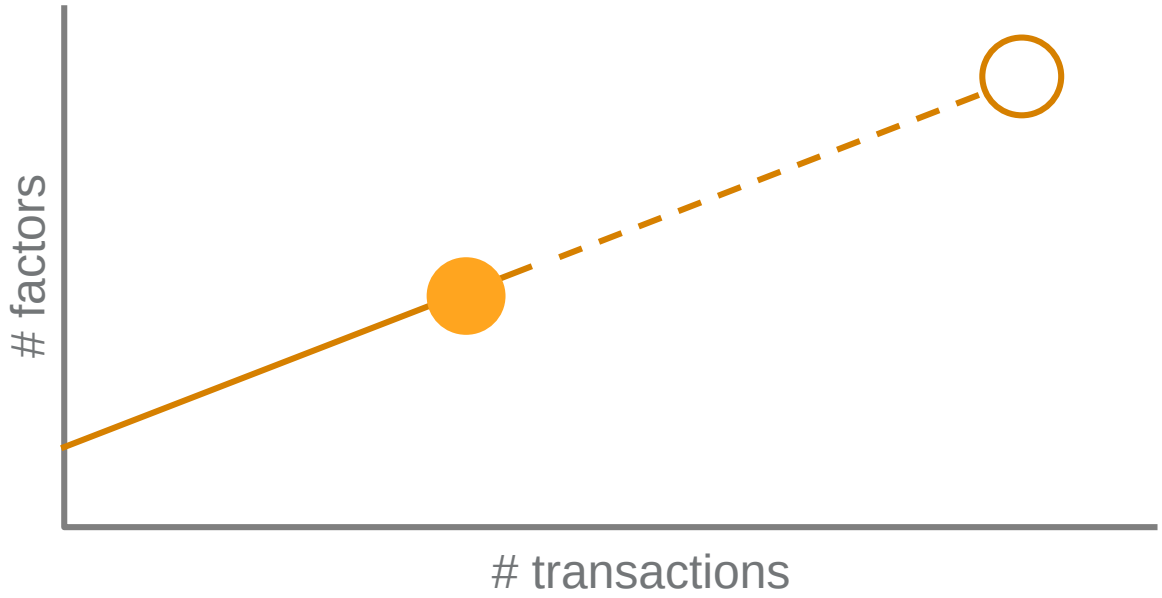
1. The compelling need for AI
2. The infrastructure view of AI
3. AI at the Edge
4. What's next

Nick Curcuru
Vice President of Data Analytics and Cyber Security, Master Card

BUSINESS PROBLEMS PUSH US INTO AI

Fraud is estimated $24.26 billion in losses in 2019 - is projected to rise to $34.66 billion in 2022 [1]

Card-not-present fraud is now 81% more likely than card-present [2]

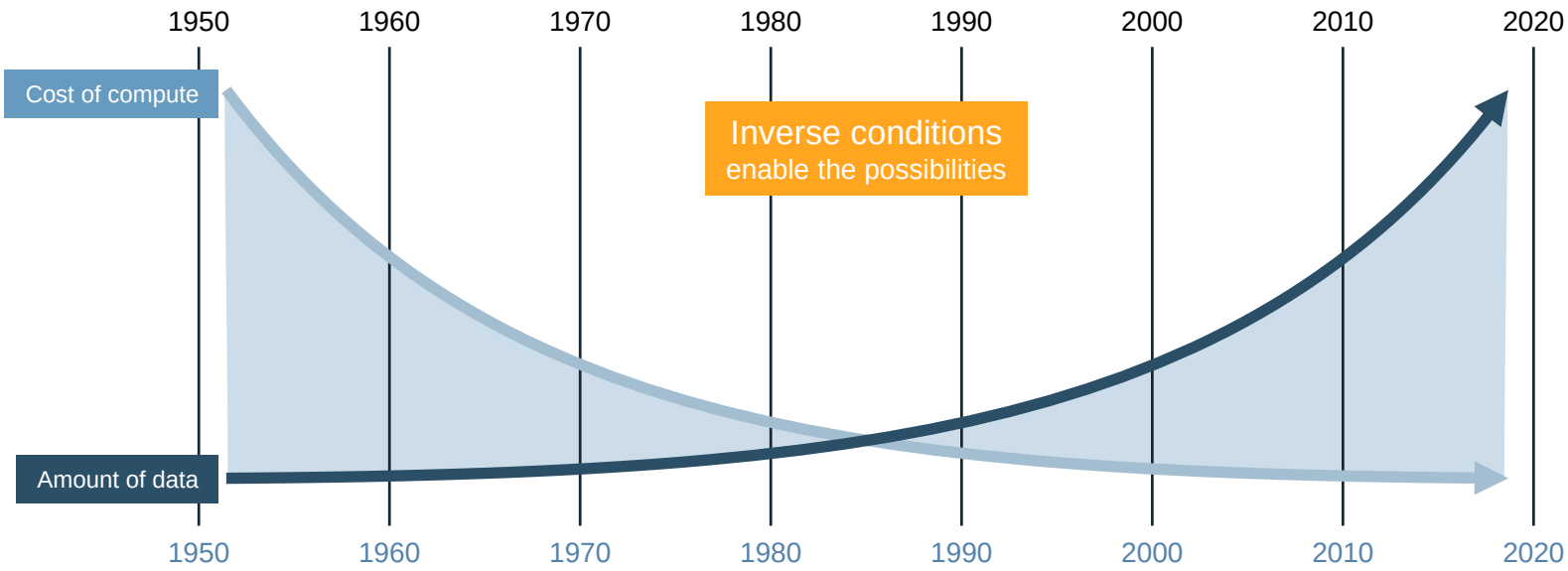


[1] Nilson Report, a newsletter
[2] Javelin Strategy & Research

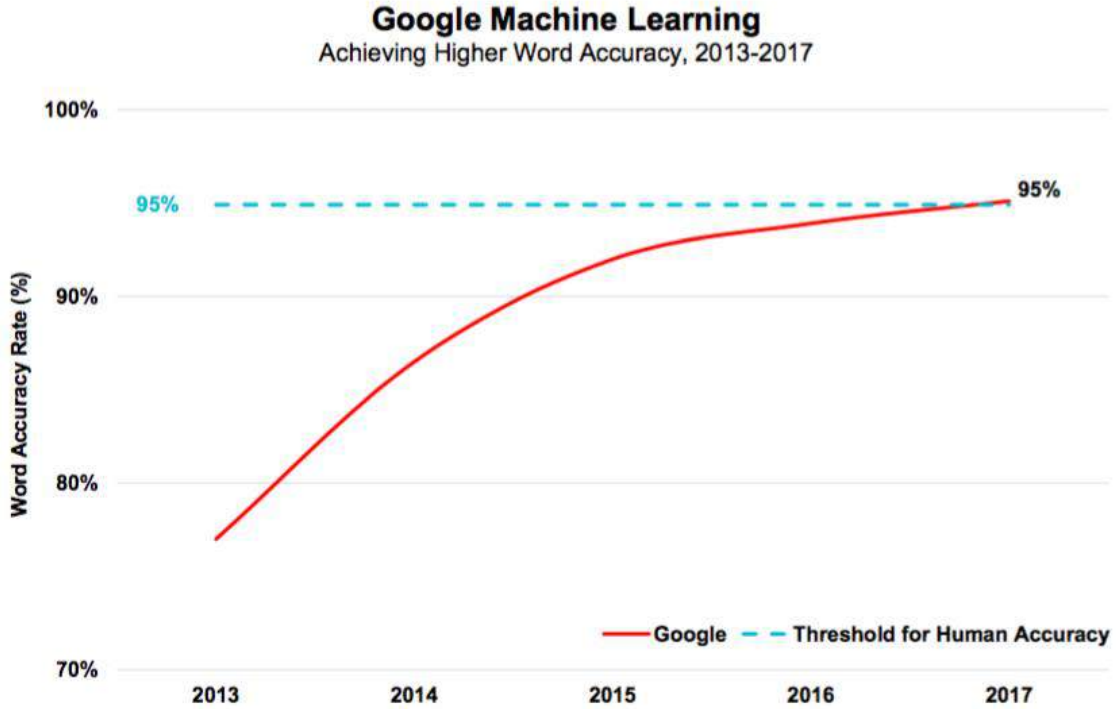
AI FITS WITH FEDERAL NEEDS AS WELL

Healthcare	Financial	National Security	Emergency Management
Case Automation	Tax Audit	Facial recognition	Emergency Services
Fraud Detection/Prevention	Fraud Detection/Prevention	Video surveillance	Event Prediction
	Investment predictions	Cyber security	
	Customer service	Satellite imagery	
	Network security		

YES AI AGAIN, BUT THIS TIME…

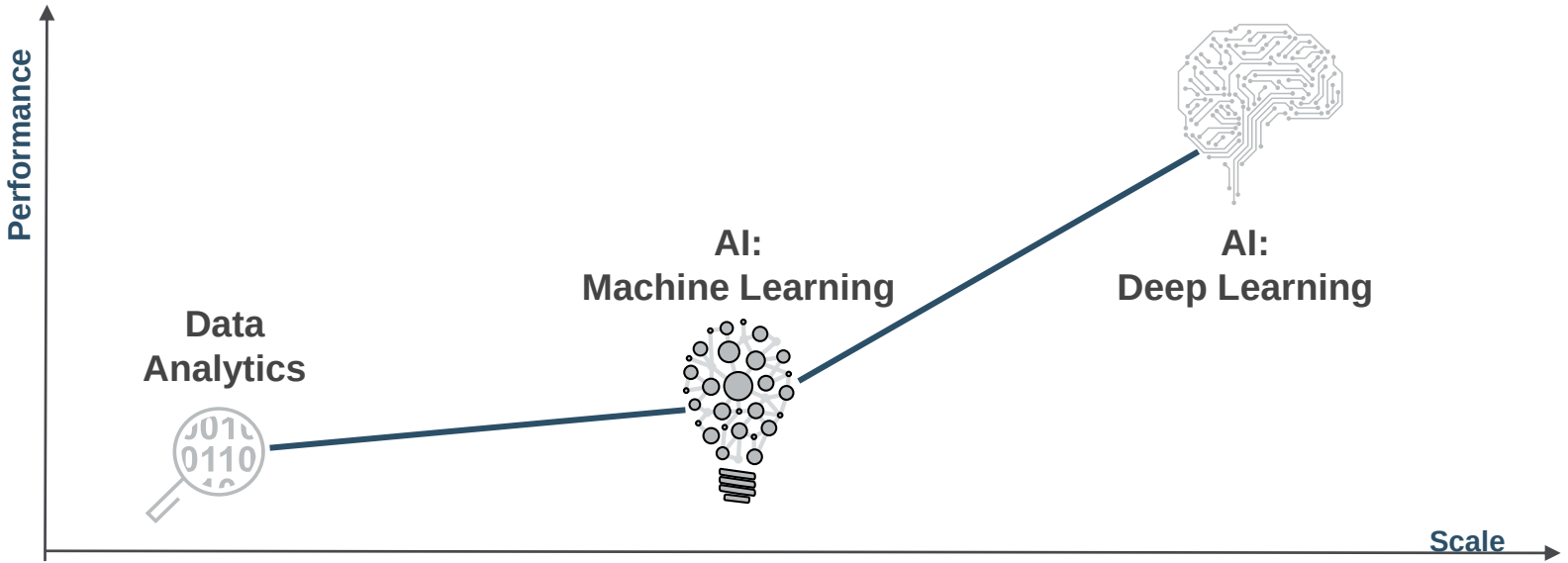


AI IS ABOUT ACCURACY AT SCALE



https://www.recode.net/2017/5/31/15720118/google-understand-language-speech-equivalent-humans-code-conference-mary-meeker

ACCURACY AT SCALE HAS A COST



Analytics Requirements

Compute: **CPU**
Throughput: **MB/s**
Concurrency: **10s**
Scale: **GBs to TBs**
Data Type: **Structured**

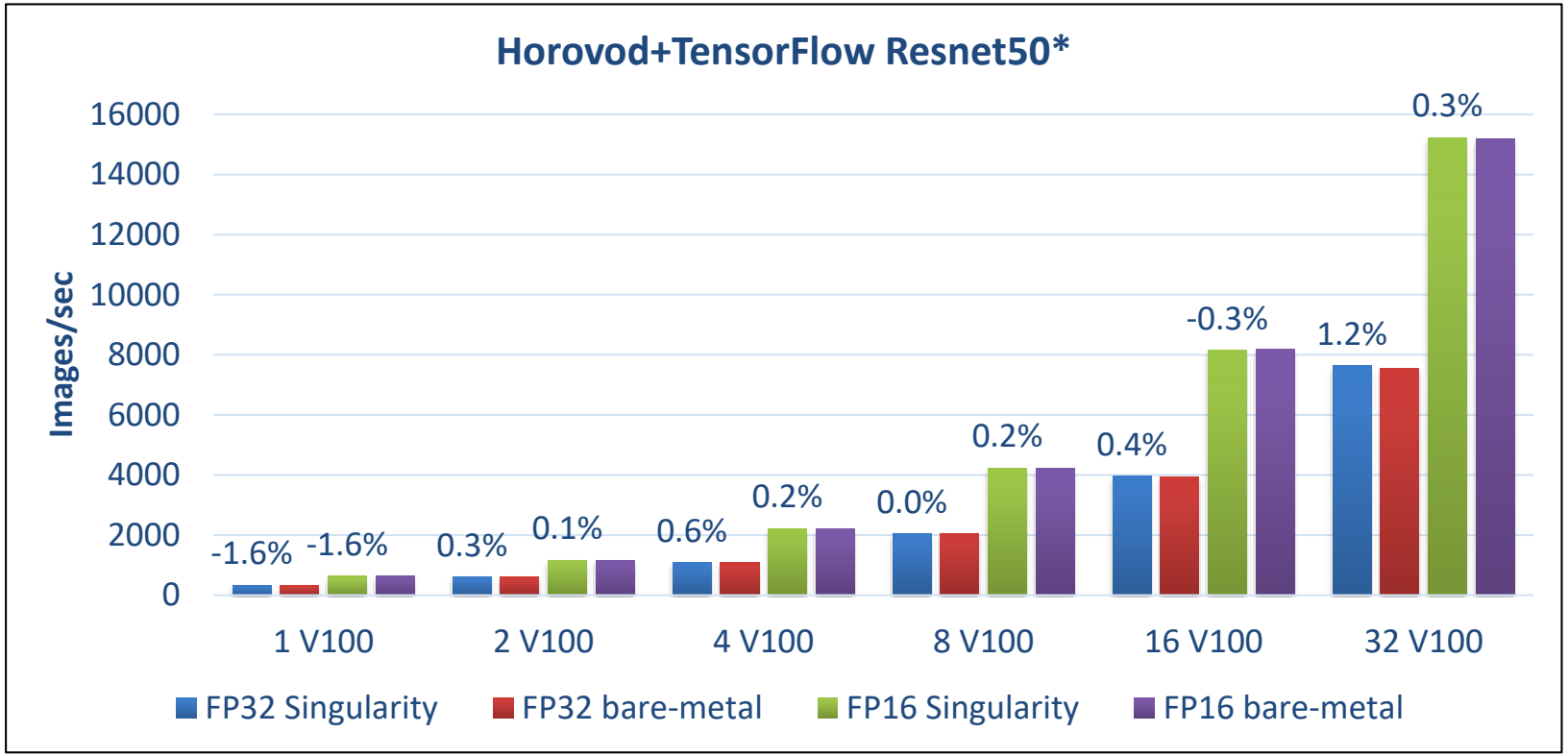
ML Requirements

Compute: **CPU & GPU**
Throughput: **GB/s**
Concurrency: **100s–K's**
Scale: **10s TB's–PB's**
Data Type: **Semi-structured, labeled**

DL Requirements

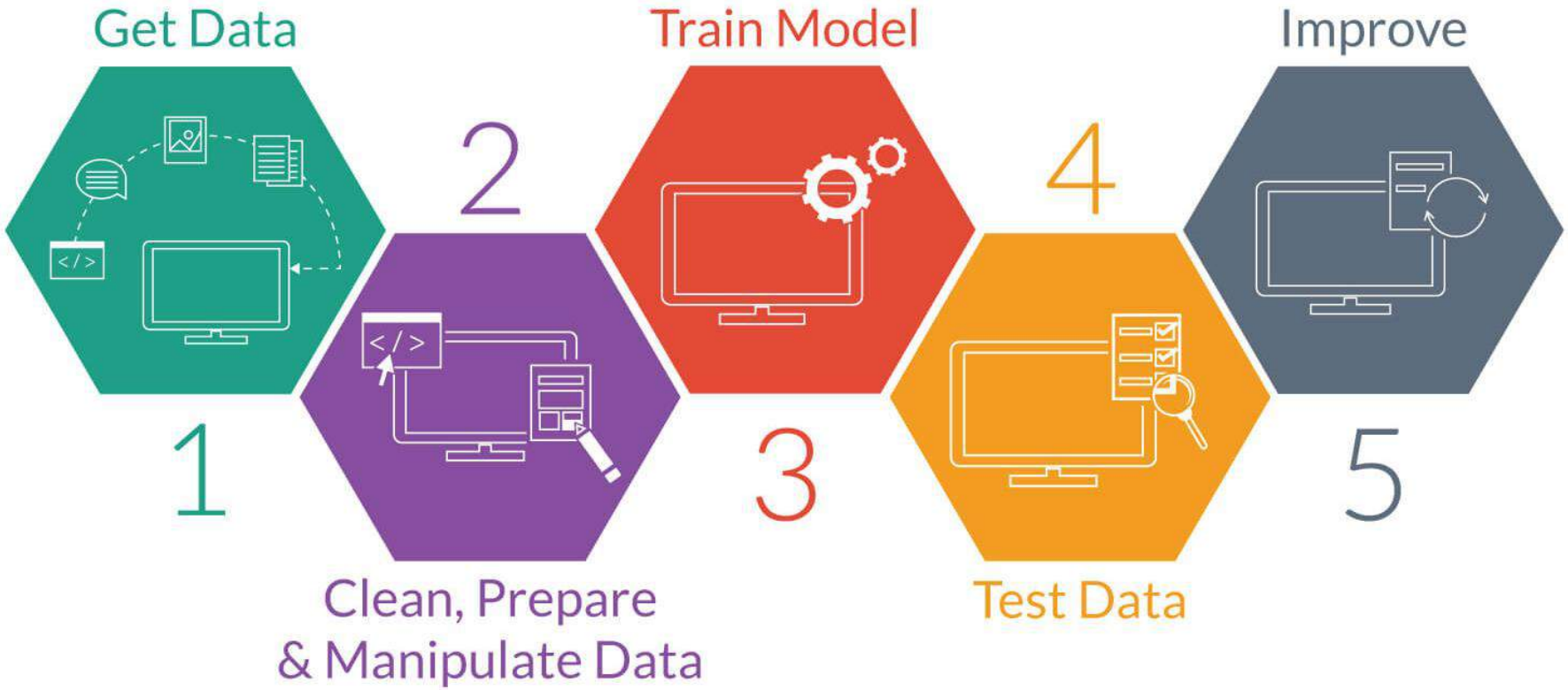
Compute: **10's-1000's GPU, FPGA, etc.**
Throughput: **10's-100's GB/s**
Concurrency: **1000s – M's**
Scale: **100s TB's to 10s PB's**
Data Type: **Unstructured, unlabeled**

ACCURACY AT SCALE IS POSSIBLE: 32 V100 = 22.4X

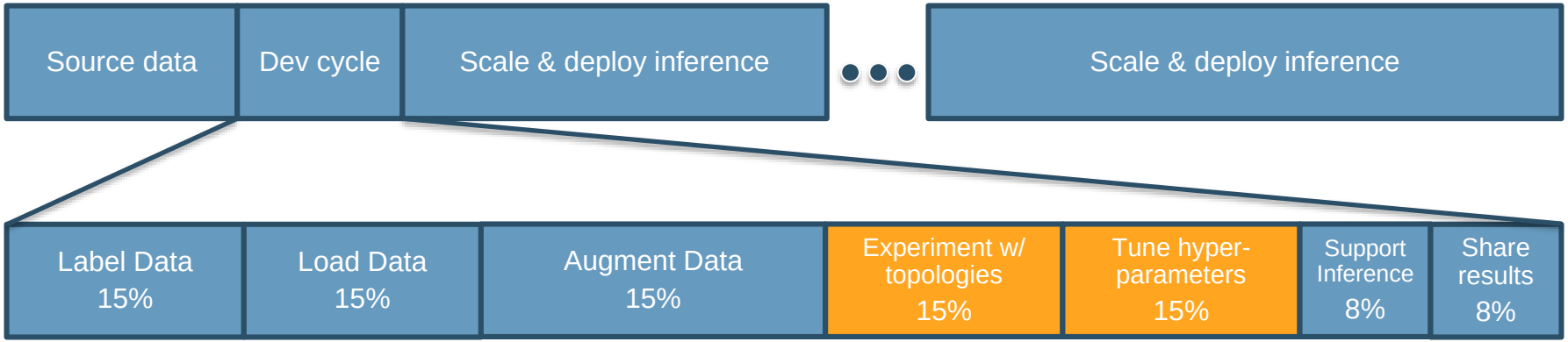


*Note: In FP32 mode, batch size: 128 per GPU; In FP16 mode, batch size: 256 per GPU
Note: MPI used for multi-node communication

AI PROCESS


Get Data
1
2
Clean, Prepare
& Manipulate Data
Train Model
3
4
Test Data
Improve
5

AI LIFECYCLE



Data Collection

Managing the Stack

Data Cleansing

Training Models

What about AI at the Edge?

IS THIS THE EDGE?

https://www.thedrum.com/news/2017/02/06/the-ai-doctor-will-see-you-now-how-babylon-getting-people-trust-chatbot

IS THIS THE EDGE?

https://latinousa.org/2017/02/01/us-customs-border-protection-publishes-specific-qa-trump-immigration-executive-order/

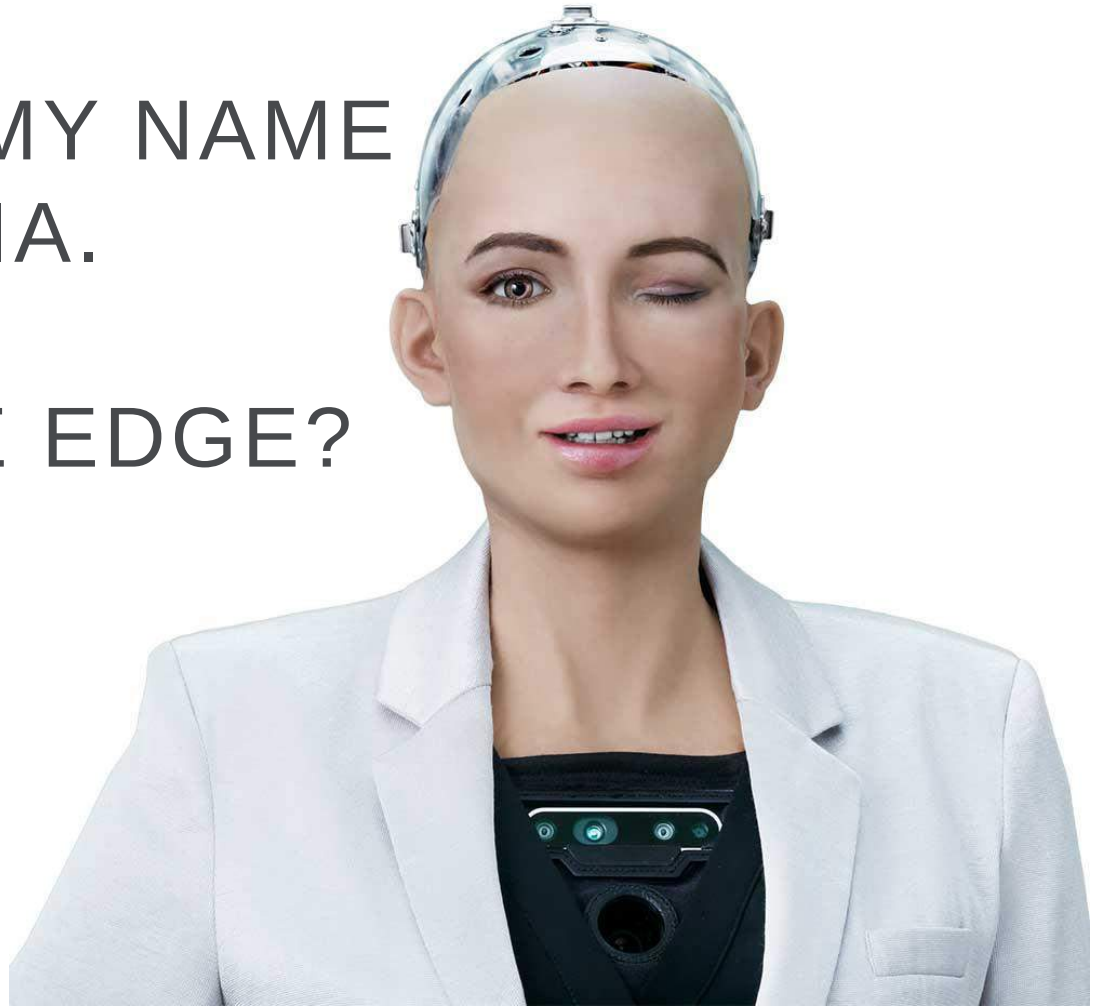
IS THIS THE EDGE?

https://www.rferl.org/a/u-s-warship-donald-cook-heads-to-black-sea-for-security-operations-/29720003.html

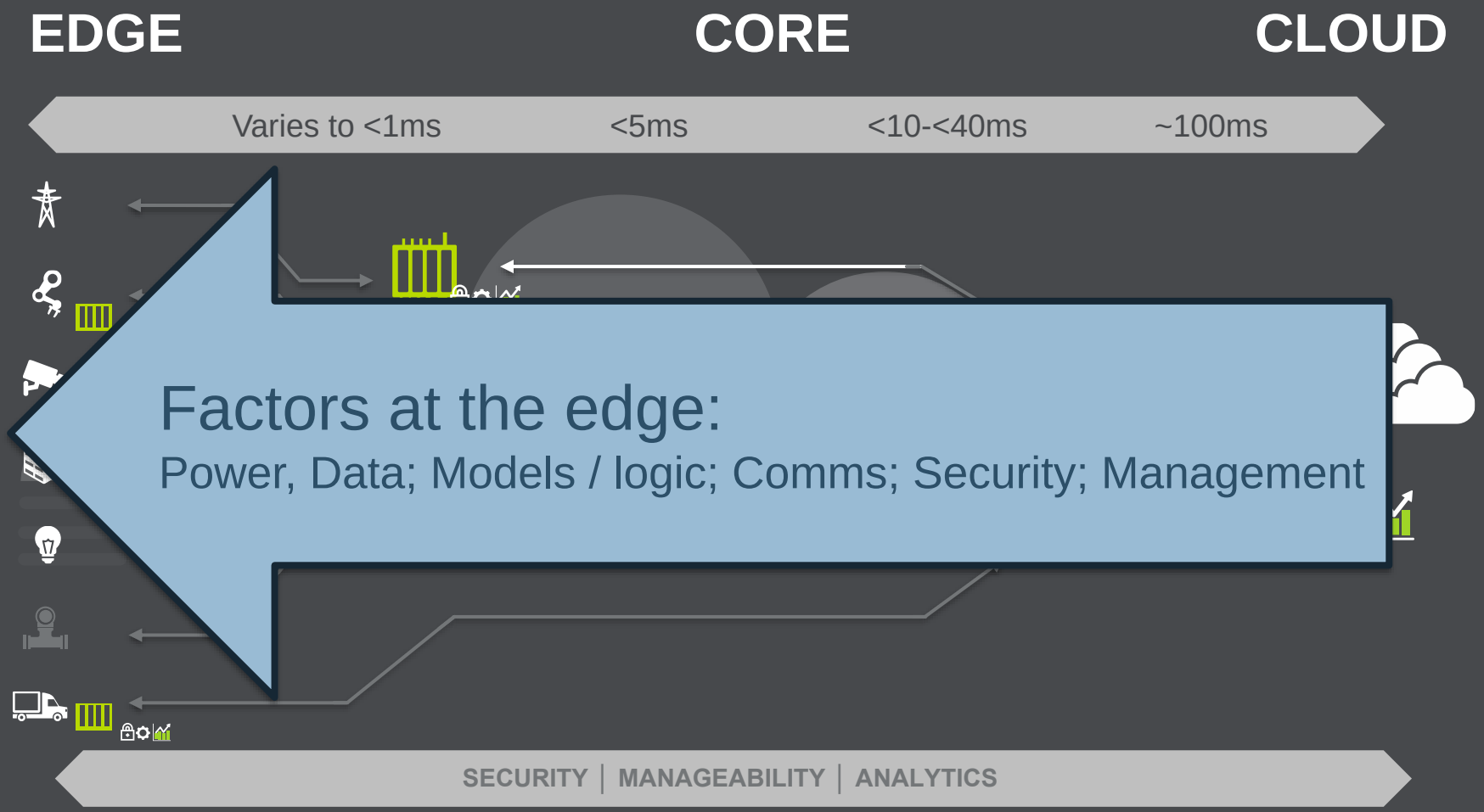
U.S. Navy (EPA-EFE)

HELLO. MY NAME IS SOPHIA.

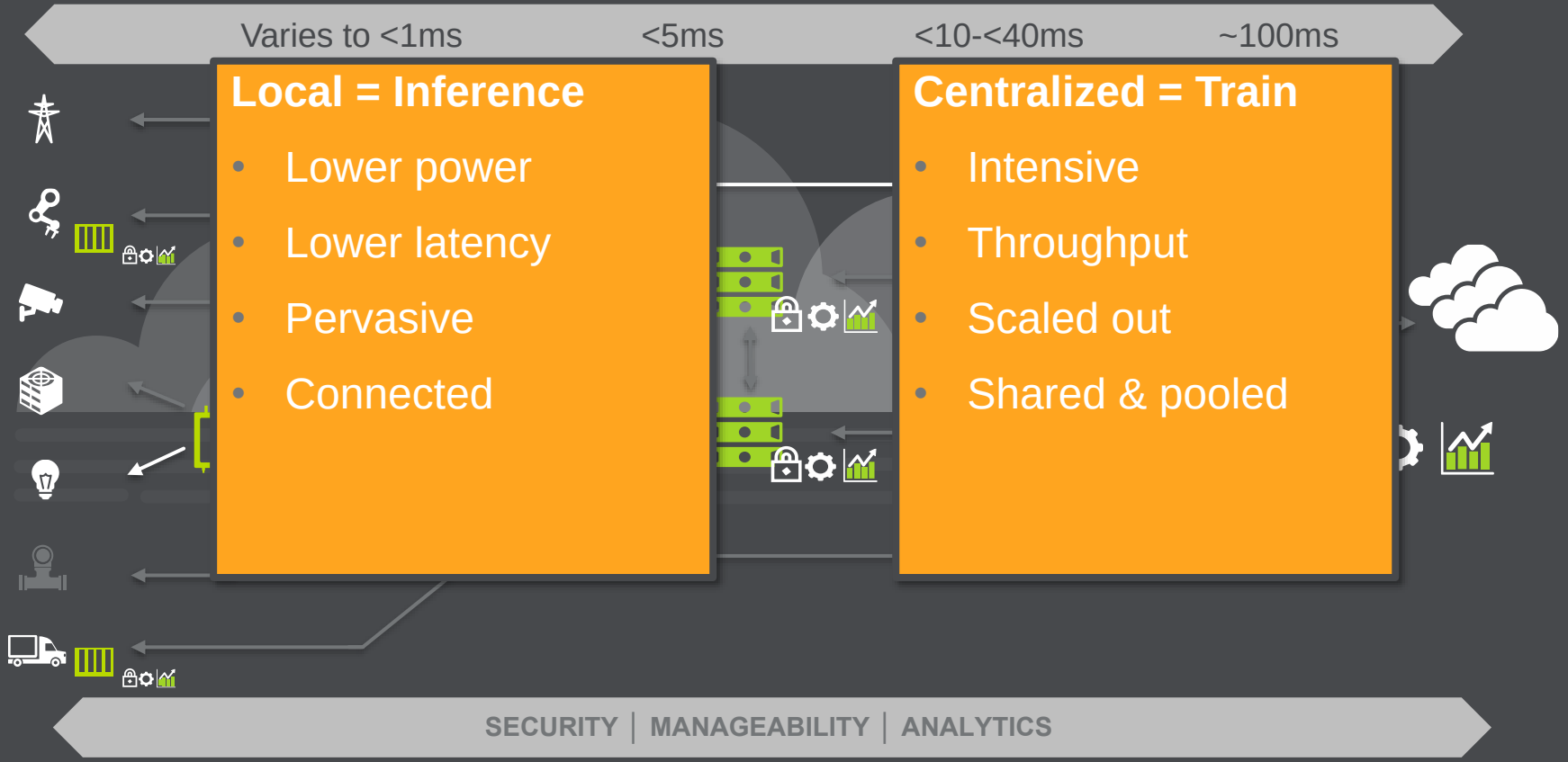
AM I THE EDGE?



https://www.hansonrobotics.com/sophia/

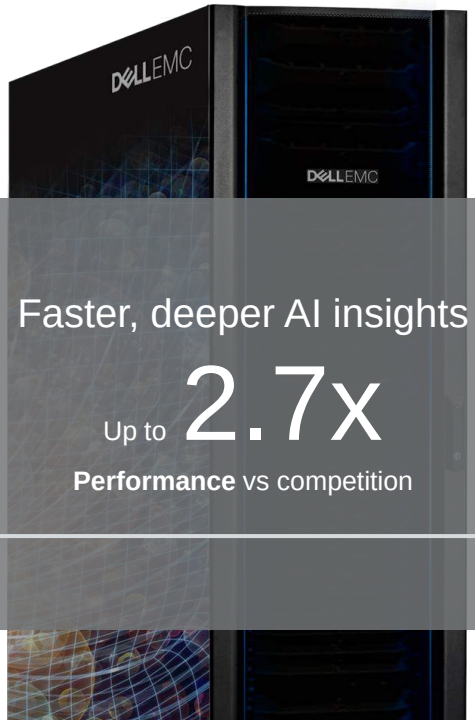

EDGE
CORE
CLOUD
Varies to <1ms
<5ms
<10-<40ms
~100ms
Factors at the edge:
Power, Data; Models / logic; Comms; Security; Management
SECURITY | MANAGEABILITY | ANALYTICS

EDGE CORE CLOUD



Ready Solutions for AI

Validated stack built to handle most demanding AI workloads



Simpler AI experience

30%

Improved data scientist productivity

Faster, deeper AI insights

Up to 2.7x

Performance vs competition

Proven AI expertise

98%

Lower training time

Where do we go from here?

WHERE DO WE GO FROM HERE?

Today

- World of clickstream
- E.g. chatbots; recommendation engines; etc.
- Lambda architectures: speed and batch
- Gateways / regional DC
- Train centrally; infer locally (Models pushed down)

Emerging

- World of smart IoT
- E.g. autonomous vehicle
- Speed layer is the only layer (Kappa)
- Decisions at ingest
- Mobile / Micro DC
- Train locally (Models built at edge then moved centrally)

The Possible

- World of meshed intelligence
- E.g. Swarm drones
- High speed P2P networks
- Continuously mobile DC
- Clusters built at edge
- Autonomous learning (models moved laterally)


DELL EMC