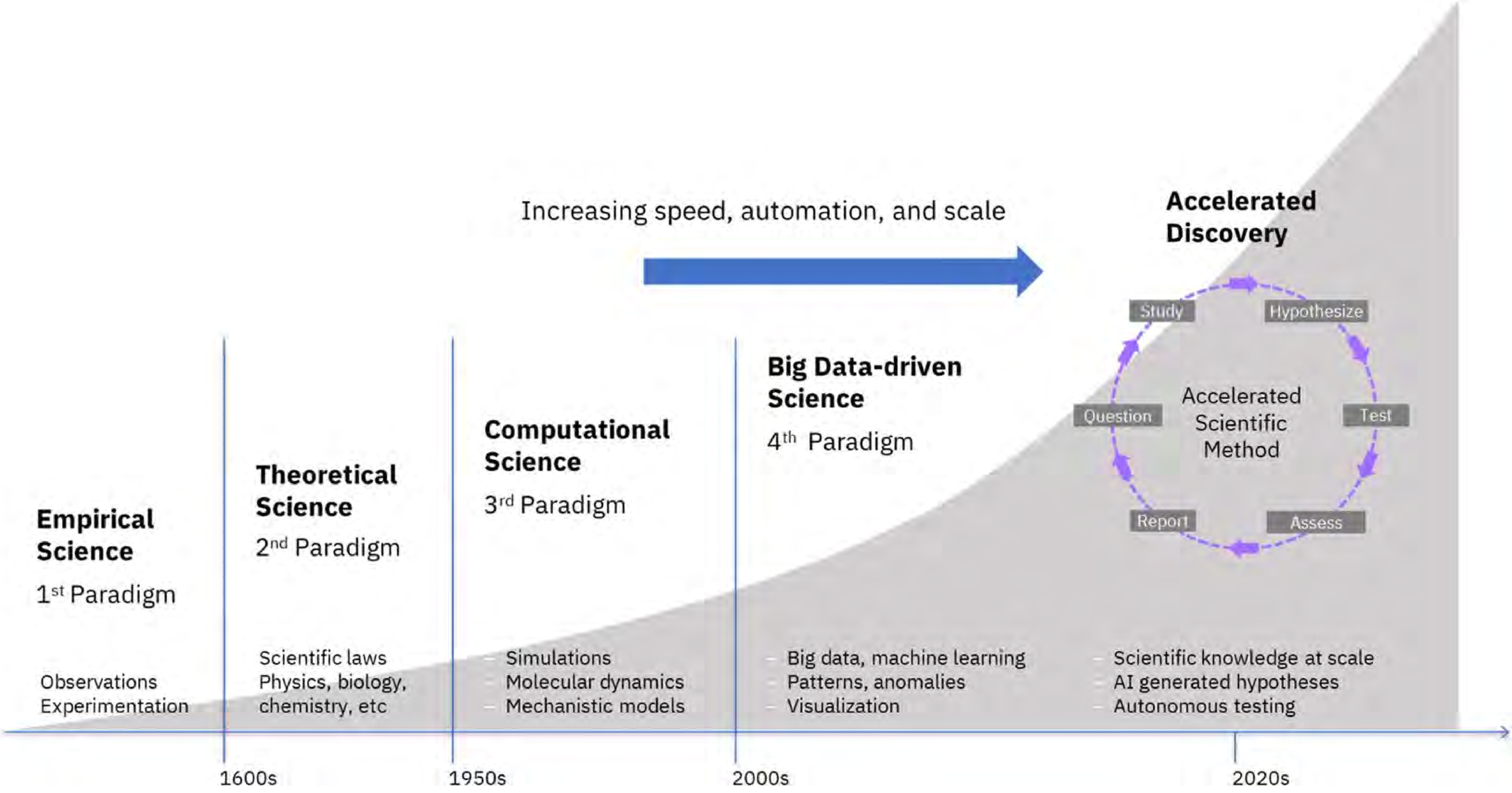
SLU

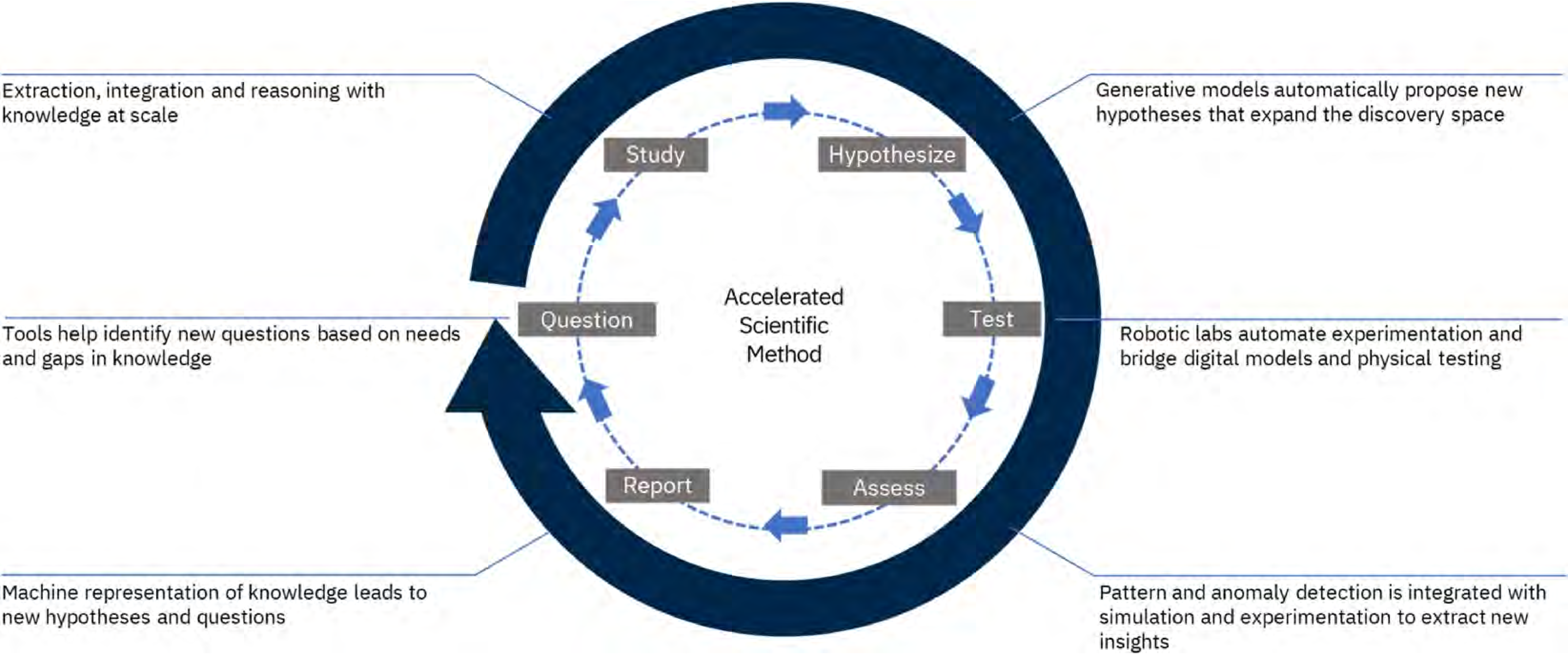
SCIENCE AND EDUCATION FOR SUSTAINABLE LIFE

Importance of scalability of computer vision models within the livestock sector: how to achieve a painless transition from research into practice

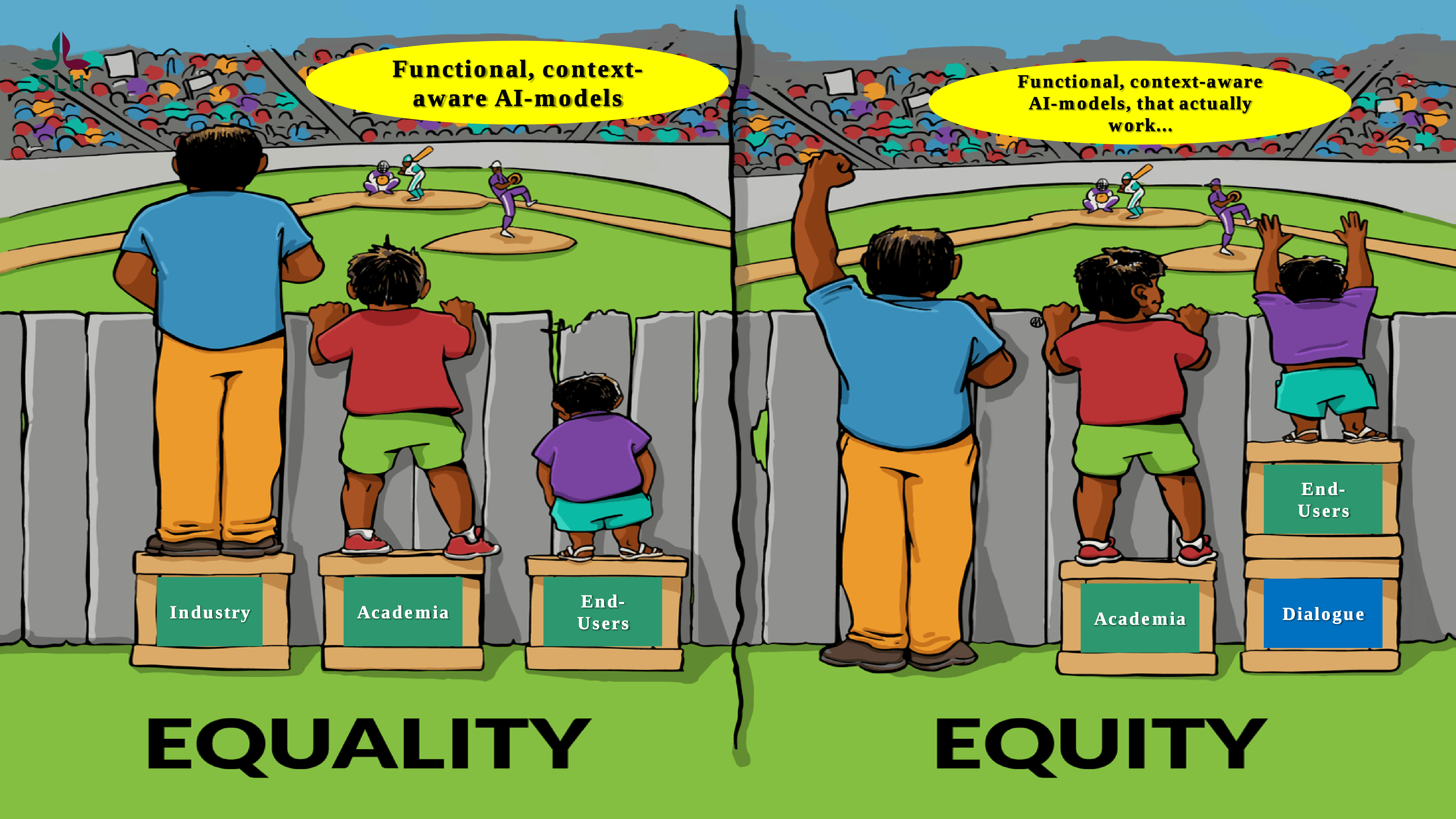
Dr. Oleksiy "Alex" Guzhva,
Assistant Professor,
Dept. of Biosystems and Technology, SLU, Sweden

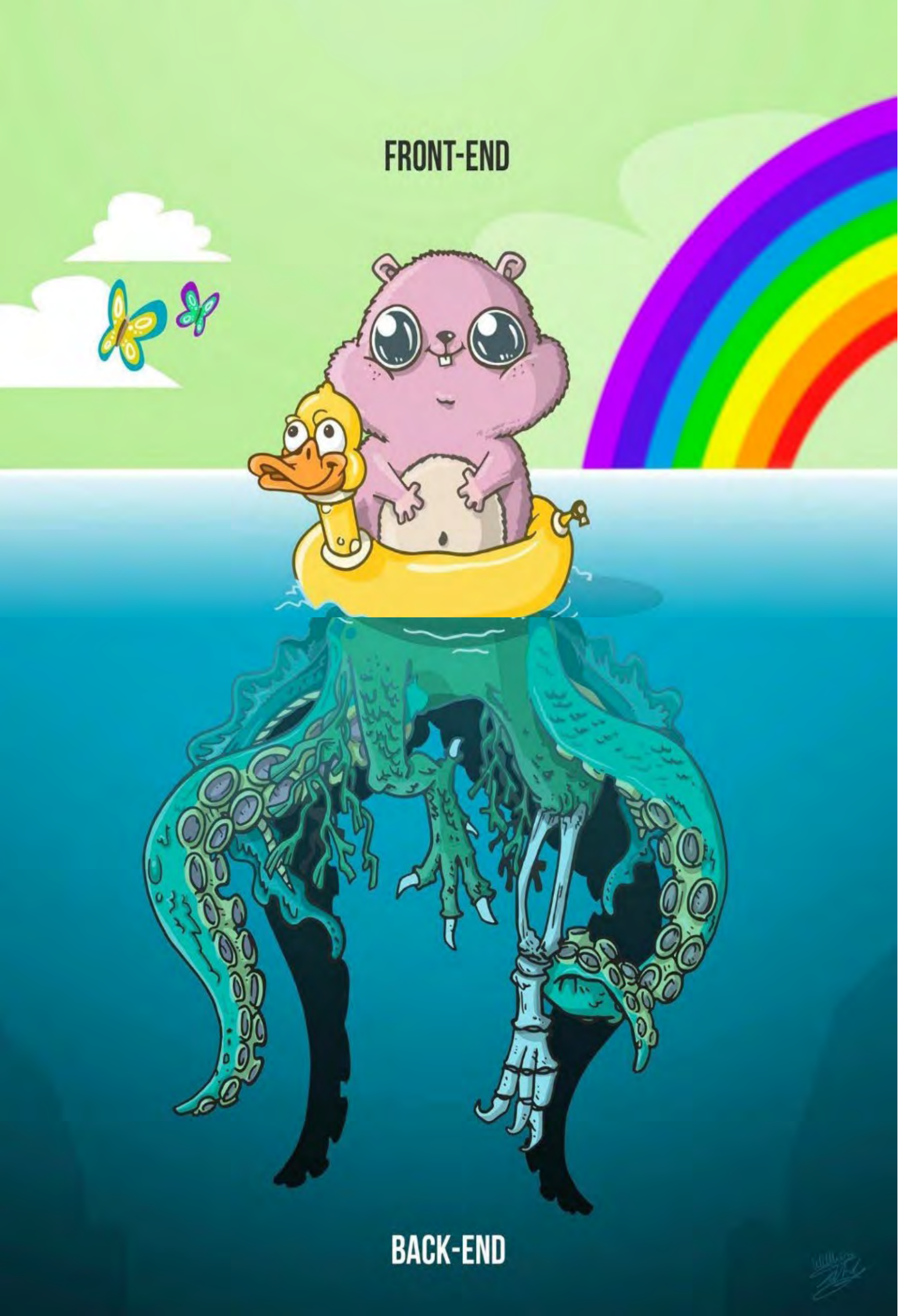


Pyzer-Knapp, E.O., Pitera, J.W., Staar, P.W.J. *et al.* Accelerating materials discovery using artificial intelligence, high performance computing and robotics. *npj Comput Mater* **8**, 84 (2022).



Pyzer-Knapp, E.O., Pitera, J.W., Staar, P.W.J. *et al.* Accelerating materials discovery using artificial intelligence, high performance computing and robotics. *npj Comput Mater* **8**, 84 (2022).


Functional, context-aware AI-models
Industry
Academia
End-Users
EQUALITY
Functional, context-aware AI-models, that actually work...
End-Users
Academia
Dialogue
EQUITY



Innovation process:

- ❑ New tech/AI-model
- ❑ We have a market, let’s sell!
- ❑ First prototype
- ❑ Integration and compatibility
- ❑ Cybersecurity and GDPR
- ❑ Laws and E-Governance
- ❑ Tough production environment
- ❑ Price or value?
- ❑ End-User perspective?

So, where and how do we start?

Research-Only

Industrial R&D

Low/High TRL

ABM/PLF

One/Multi-Species

Behavior/Health/Production

Data Availability

Model

Transfer Learning

Expected Outcome

Resources

Real-World Performance

Interpretability

Integration Potential

E-Governance

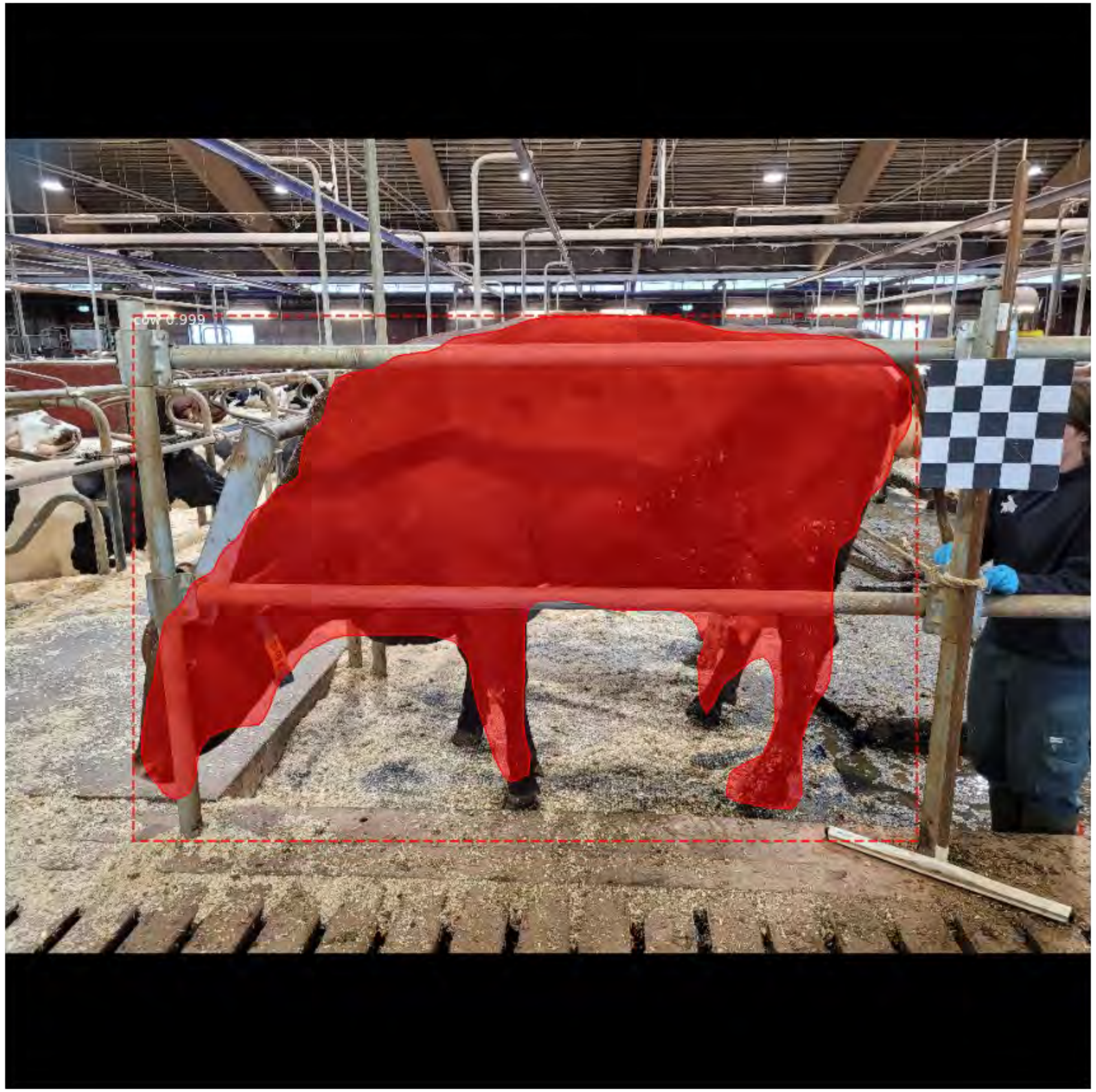
Public Opinion

Digital 3R

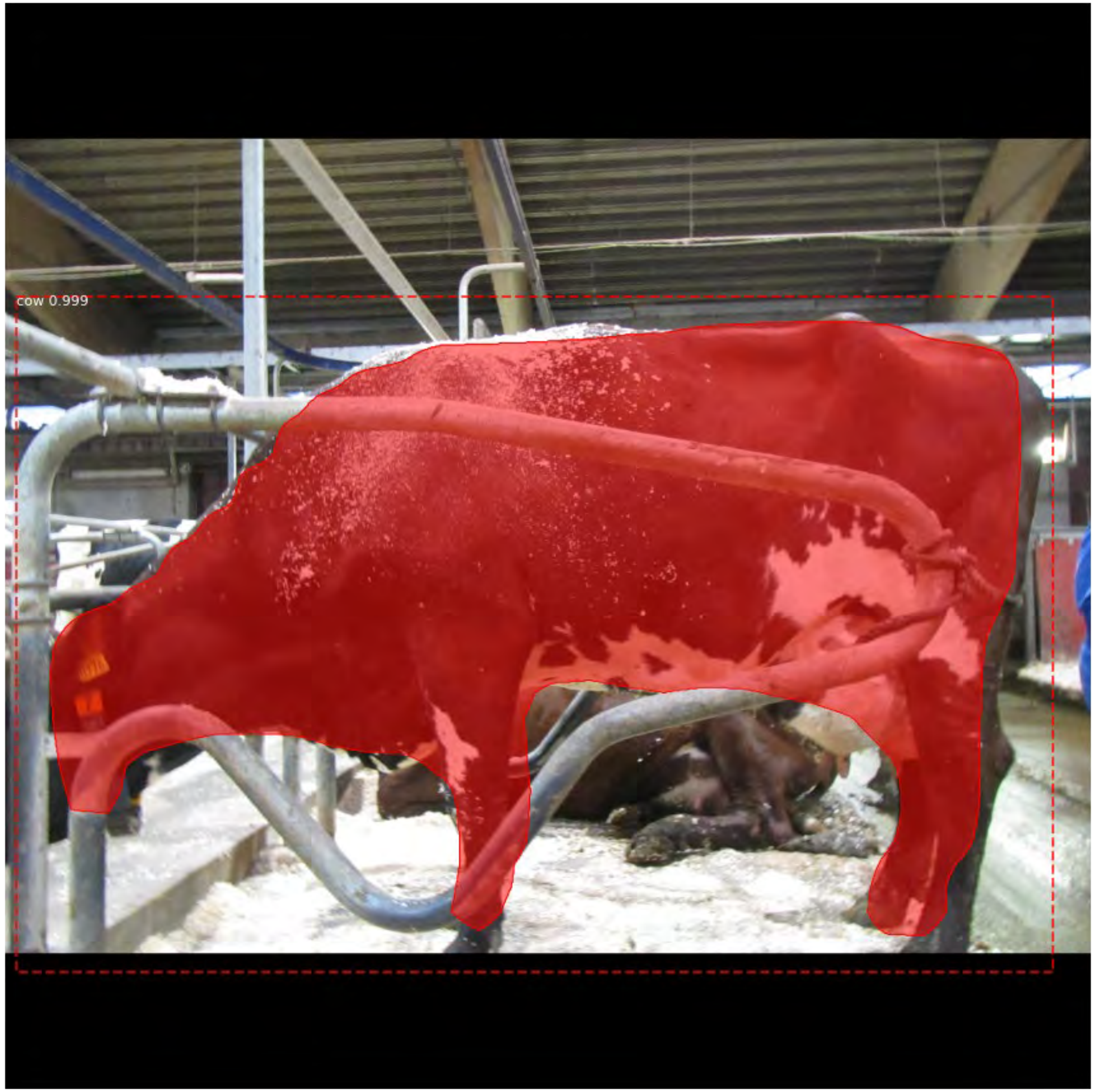
Security

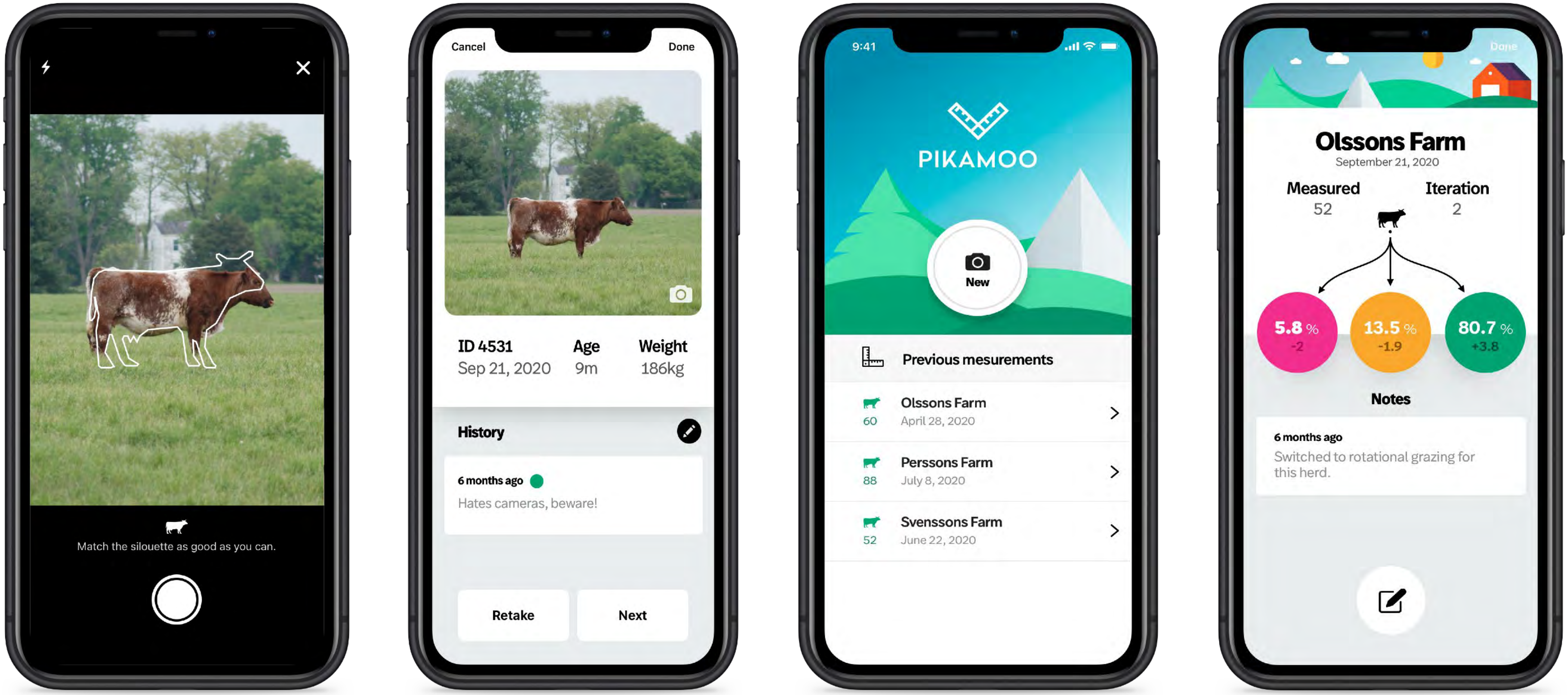

SLU
What can we see with help from
Computer Vision/AI?

Predictions

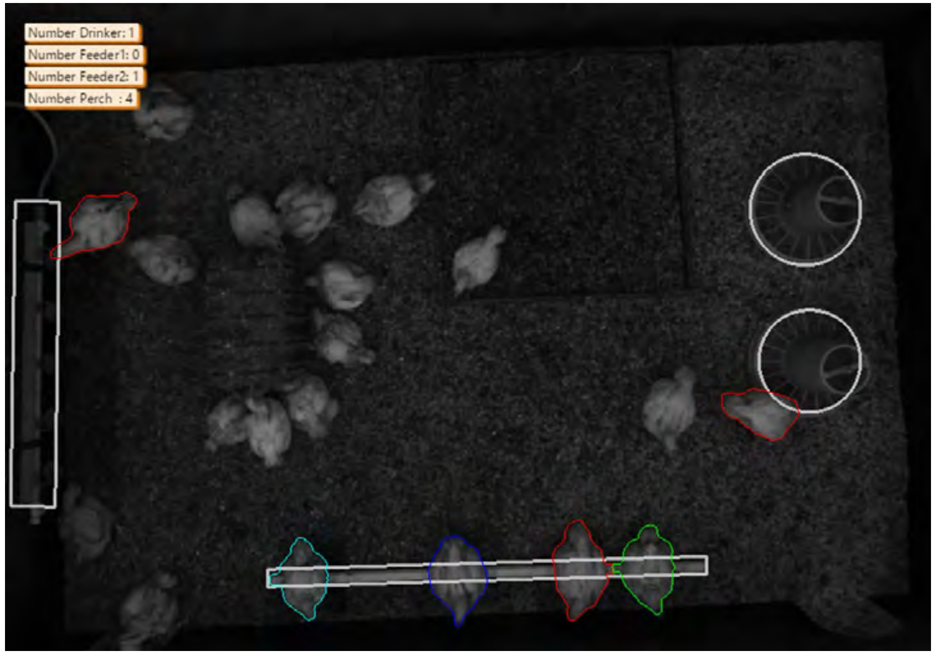


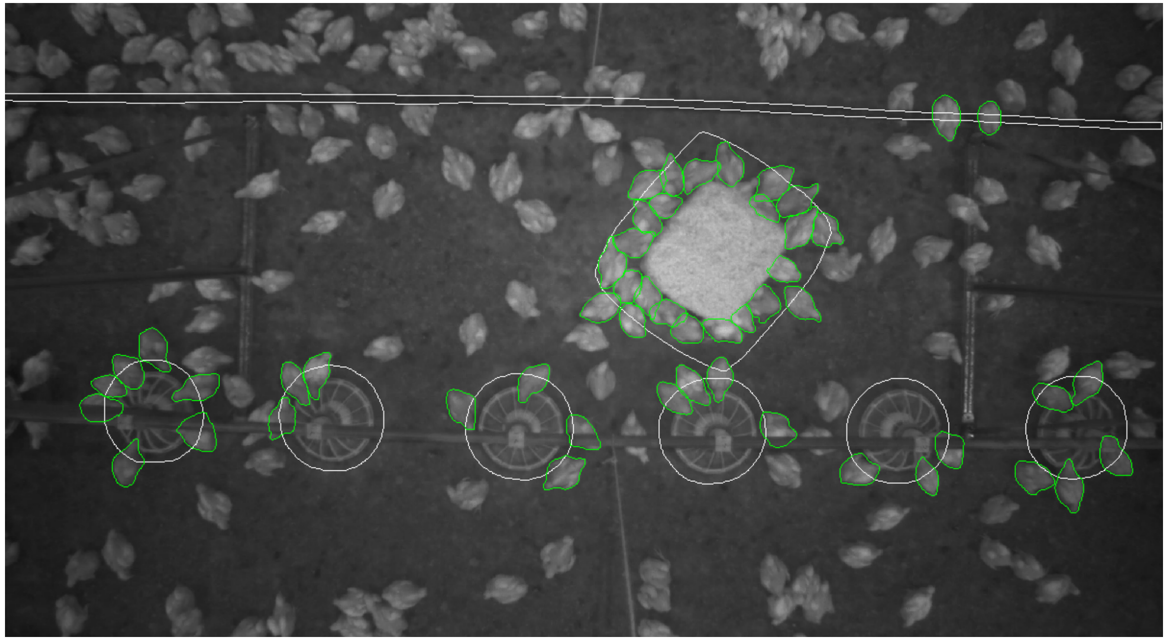
Predictions




Match the silouette as good as you can.
Cancel
Done
ID 4531
Sep 21, 2020
Age
9m
Weight
186kg
History
6 months ago
Hates cameras, beware!
Retake
Next
9:41
PIKAMOO
New
Previous mesurements
Olssons Farm
60
April 28, 2020
Perssons Farm
88
July 8, 2020
Svenssons Farm
52
June 22, 2020
Done
Olssons Farm
September 21, 2020
Measured
52
Iteration
2
5.8 %
-2
13.5 %
-1.9
80.7 %
+3.8
Notes
6 months ago
Switched to rotational grazing for this herd.

SLU


Number Drinker: 1
Number Feeder1: 0
Number Feeder2: 1
Number Perch : 4



Living organisms are ...

...individually different

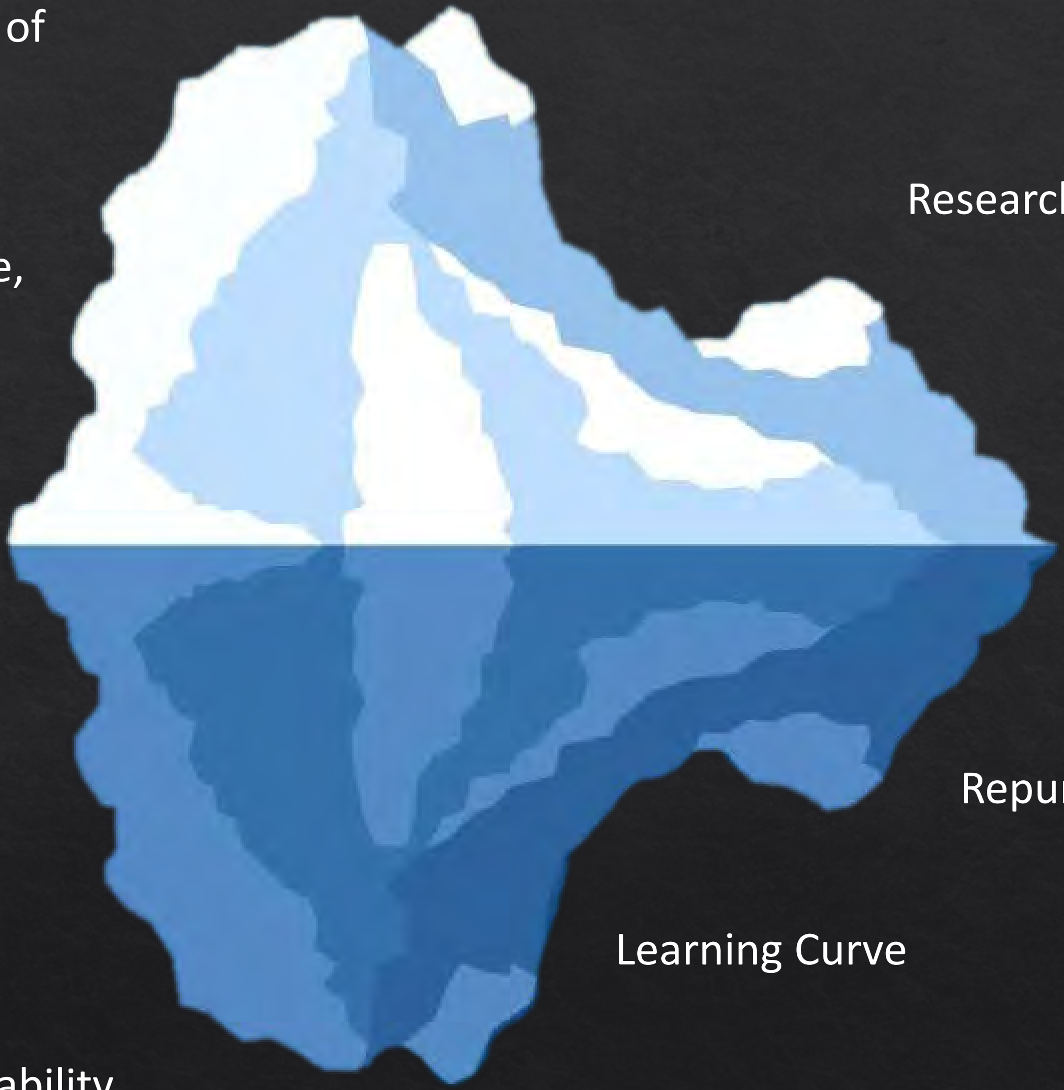
Important: to come as close to an individual animal as possible!

What:
Health, welfare, management, all of it..?

How:
Local, Cloud, Federate Learning..?

When:
Real-Time, Near Real-Time, Offline..?

Why:
Research-Only, R&D, Commercial Application..?



Resource Efficiency

Clarity

Elegance

Repurposing

Adjustability

Learning Curve

Scalability

Since a large part of machine learning is feeding data to an algorithm that performs heavy computations iteratively, the choice of hardware also plays a significant role in scalability.

Scaling activities for computations in machine learning (specifically deep learning) should be concerned about executing matrix multiplications as fast as possible with less power consumption (because of cost!).

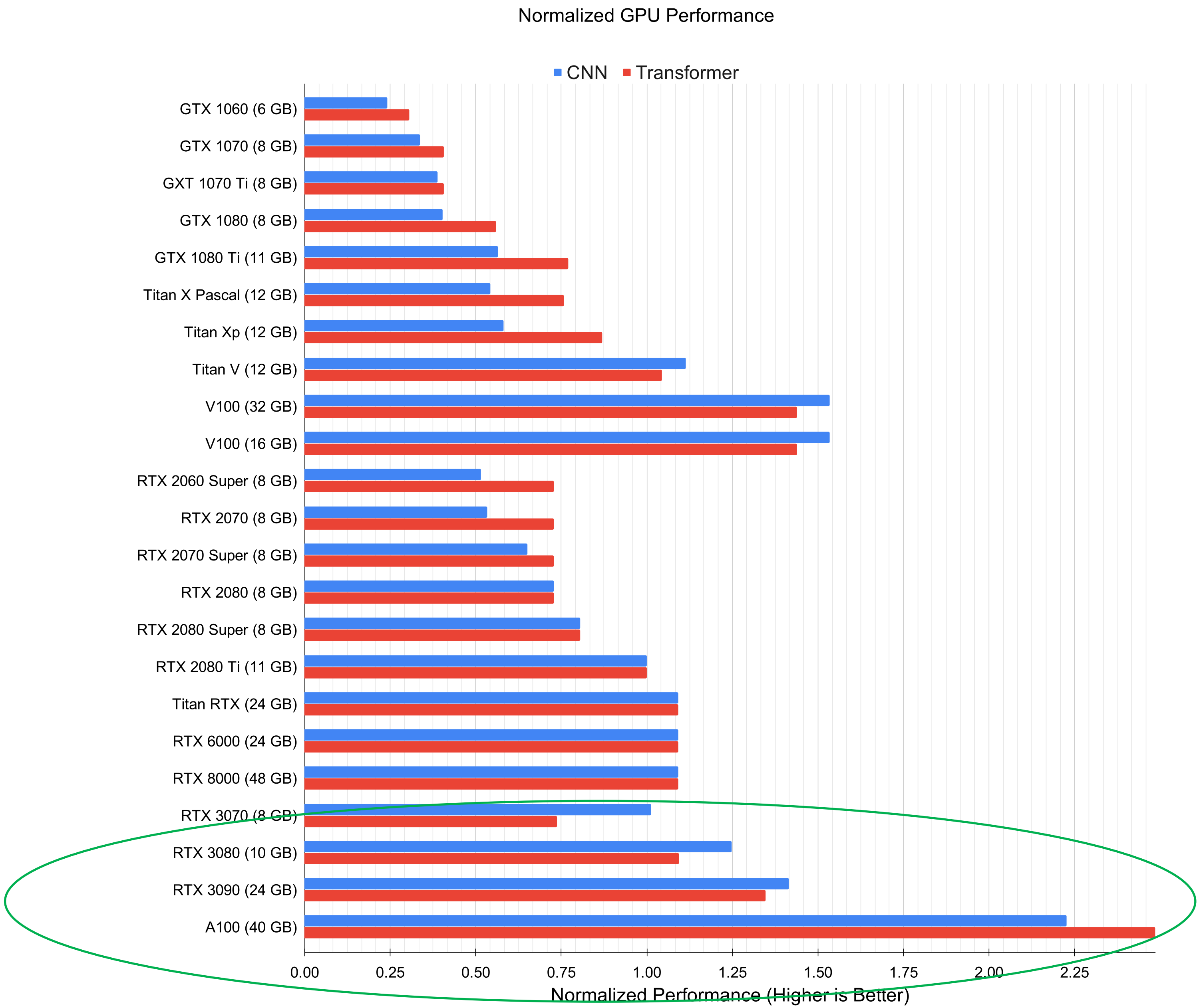
Tools for development

200 000+ USD

1500+ USD

70+ USD

As well as their cost and efficiency…



Normalized GPU Performance
CNN
Transformer
GTX 1060 (6 GB)
GTX 1070 (8 GB)
GXT 1070 Ti (8 GB)
GTX 1080 (8 GB)
GTX 1080 Ti (11 GB)
Titan X Pascal (12 GB)
Titan Xp (12 GB)
Titan V (12 GB)
V100 (32 GB)
V100 (16 GB)
RTX 2060 Super (8 GB)
RTX 2070 (8 GB)
RTX 2070 Super (8 GB)
RTX 2080 (8 GB)
RTX 2080 Super (8 GB)
RTX 2080 Ti (11 GB)
Titan RTX (24 GB)
RTX 6000 (24 GB)
RTX 8000 (48 GB)
RTX 3070 (8 GB)
RTX 3080 (10 GB)
RTX 3090 (24 GB)
A100 (40 GB)
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
2.00
2.25
Normalized Performance (Higher is Better)

Aren't we being ridiculous?

NVIDIA H100 Tensor Core GPU

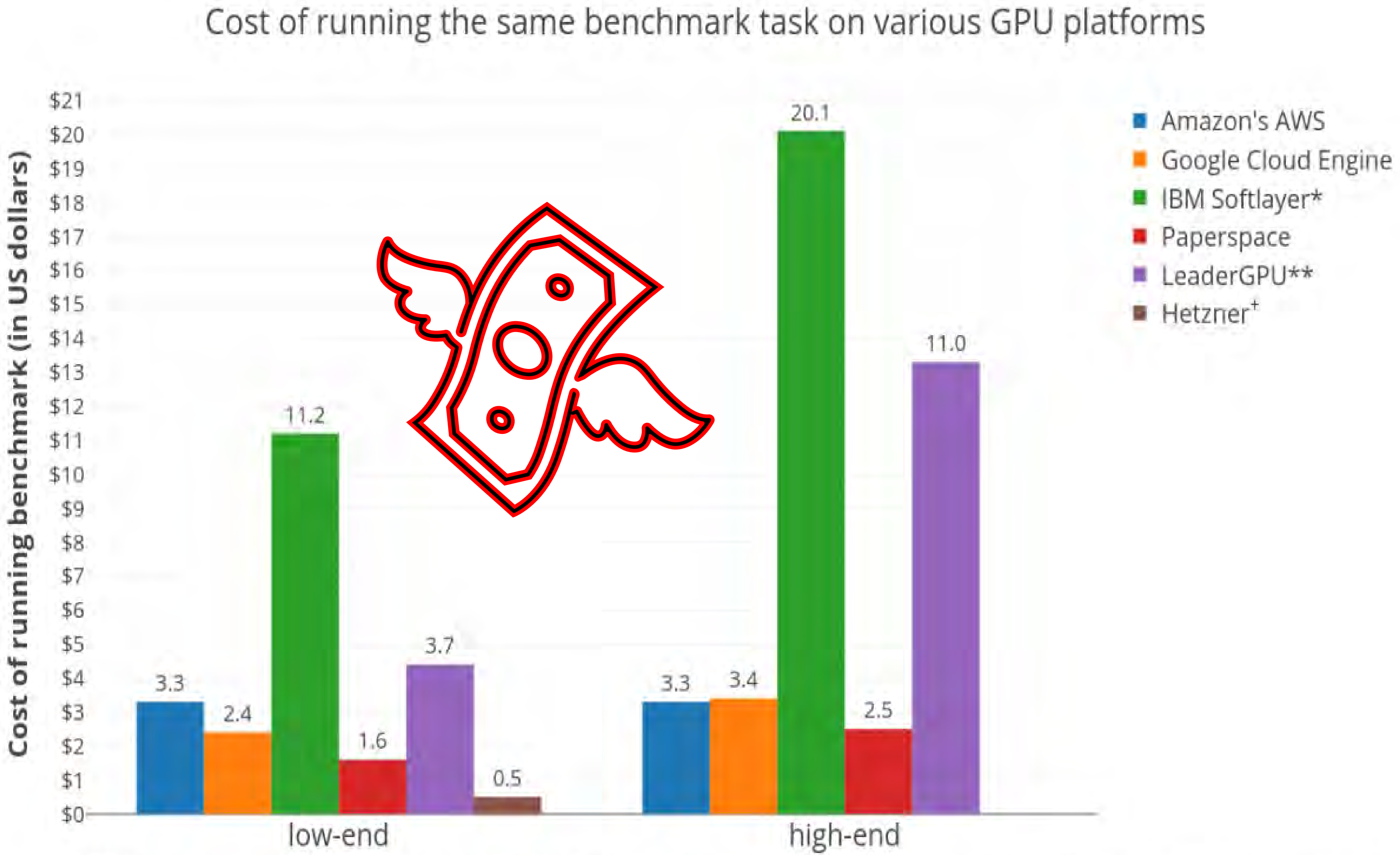
Unprecedented performance, scalability, and security for every data center.

Learn More

An order-of-magnitude leap for accelerated computing.

Tap into unprecedented performance, scalability, and security for every workload with the NVIDIA H100 Tensor Core GPU. With NVIDIA® NVLink® Switch System, up to 256 H100s can be connected to accelerate exascale workloads, along with a dedicated Transformer Engine to solve trillion-parameter language models. H100's combined technology innovations can speed up large language models by an incredible 30X over the previous generation to deliver industry-leading conversational AI.

Accelerator Model	**K80**	**P100**	**P100**	**P100**	**V100**	**V100**	**A100**	**A100**	**A100**	**H100**	**H100**
Bus	PCI-E 3.0	PCI-E 3.0	PCI-E 3.0	SXM	PCI-E 3.0	SXM2/3	PCI-E 4.0	SXM4	SXM4	PCI-E 5.0	SXM5
GPU	2 * GK210B	GP100	GP100	GP100	GV100	GV100	GA100	GA100	GA100	GH100	GH100
FP32 Cores	4,992	3,584	3,584	3,584	5,120	5,120	6,912	6,912	6,912	14,592	16,896
FP64 Cores	-	1,792	1,792	1,792	2,560	2,560	3,456	3,456	3,456	7,296	8,448
Tensor Cores	-	-	-	-	640	640	432	432	432	456	528
RT Cores	-	-	-	-	-	-	-	-	-	-	-
Base Core Clock Speed	560 MHz	1,126 MHz	1,126 MHz	1,328 MHz	937 MHz	1,327 MHz	765 MHz	1,095 MHz	1,095 MHz	***600 MHz***	***1,095 MHz***
GPU Boost Clock Speed	875 MHz	1,303 MHz	1,303 MHz	1,480 MHz	1,312 MHz	1,530 MHz	1,410 MHz	1,410 MHz	1,410 MHz	***1,698 MHz***	***1,833 MHz***
SMs	2 * 13	56	56	56	80	80	108	108	108	114	132
Peak FP8 Tensor Core FP16 or FP32 ACC, Teraflops	-	-	-	-	-	-	-	-	-	1,600/3,200	2,000/4,000
Peak FP16 Tensor Core FP16 ACC, Teraflops	-	-	-	-	112.0	125.0	312/624	312/624	312/624	800/1,600	1,000/2,000
Peak FP16 Tensor Core FP32 ACC, Teraflops	-	-	-	-	112.0	125.0	312/624	312/624	312/624	800/1,600	1,000/2,000
Peak BF16 Tensor Core FP32 ACC, Teraflops	-	-	-	-	-	-	312/624	312/624	312/624	800/1,600	1,000/2,000
Peak TF32 Tensor Core, Teraflops	-	-	-	-	-	-	156/312	156/312	156/312	400/800	500/1,000
Peak FP64 Tensor Core, Teraflops	-	-	-	-	-	-	19.5	19.5	19.5	48.0	60.0
Peak INT8 Tensor Core, Teraops	-	-	-	-	-	-	624/1,248	624/1,248	624/1,248	1,600/3,200	2,000/4,000
Peak INT4 Tensor Core, Teraops	-	-	-	-	-	-	1,248/2,496	1,248/2,496	1,248/2,496	-	-
Peak INT8, Teraops	-	-	-	-	56.0	62.8	-	-	-	-	-
Peak INT4, Teraops	-	-	-	-	28.0	31.2	-	-	-	-	-
Peak FP16, Teraflops	-	18.7	18.7	21.2	28.0	31.4	78.0	78.0	78.0	96.0	120.0
Peak BF16, Teraflops	-	18.7	18.7	21.2	14.0	15.6	39.0	39.0	39.0	96.0	120.0
Peak FP32, Teraflops	8.7	9.3	9.3	10.6	14.0	15.7	19.5	19.5	19.5	48.0	60.0
Peak FP64, Teraflops	2.91	4.70	4.70	5.30	7.00	7.80	9.70	9.70	9.70	24.00	30.00
Peak INT32, Teraops	-	-	-	-	14.00	15.70	19.50	19.50	19.50	24.00	30.00
Peak RT Core, Teraflops	-	-	-	-	-	-	-	-	-	-	-
GDDR5 or GDDR6/**HBM2** Memory	**24 GB**	**12 GB**	**16 GB**	**16 GB**	**16/32 GB**	**16/32 GB**	**40 GB**	**40 GB**	**80 GB**	**80 GB**	**80 GB**
Memory Clock Speed	2.5 GHz	703 MHz	703 MHz	703 MHz	877.5 MHz	877.5 MHz	1,215 MHz	1,215 MHz	1,593 MHz	1,658 MHz	2,072 MHz
Memory Bandwidth	480 GB/sec	540 GB/sec	720 GB/sec	720 GB/sec	900 GB/sec	900 GB/sec	1,555 GB/sec	1,555 GB/sec	2,039 GB/sec	2,000 GB/sec	3,000 GB/sec
Power Draw	300 W	250 W	250 W	300 W	250 W	300/350 W	300 W	400 W	400 W	350 W	700 W

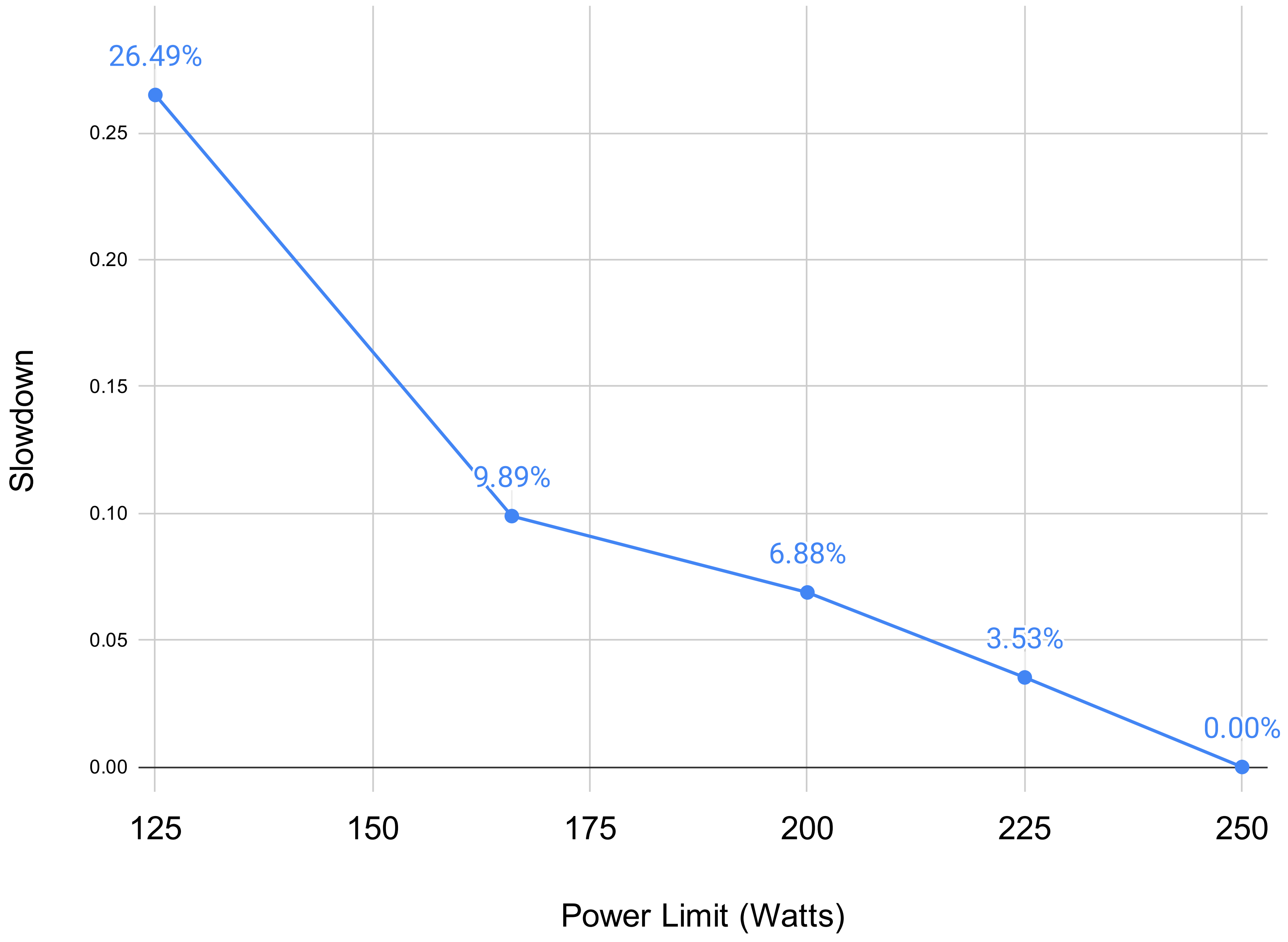


*Softlayer does not provide single GPU machines. The cost figures are based on multi GPU training using Keras's suboptimal multi GPU implementation.
LeaderGPU also does not provide single GPU servers. The cost figures are based on training using only **one of multiple available GPUs
+Hetzner only provides dedicated servers on monthly basis. Figures here reflect hourly prorated costs

Energy is sustainable if it meets the needs of the present without compromising the ability of future generations to meet their own needs…

- Kutscher, Milford & Kreith 2019

RTX 2080 Ti Slowdown vs Power Limit



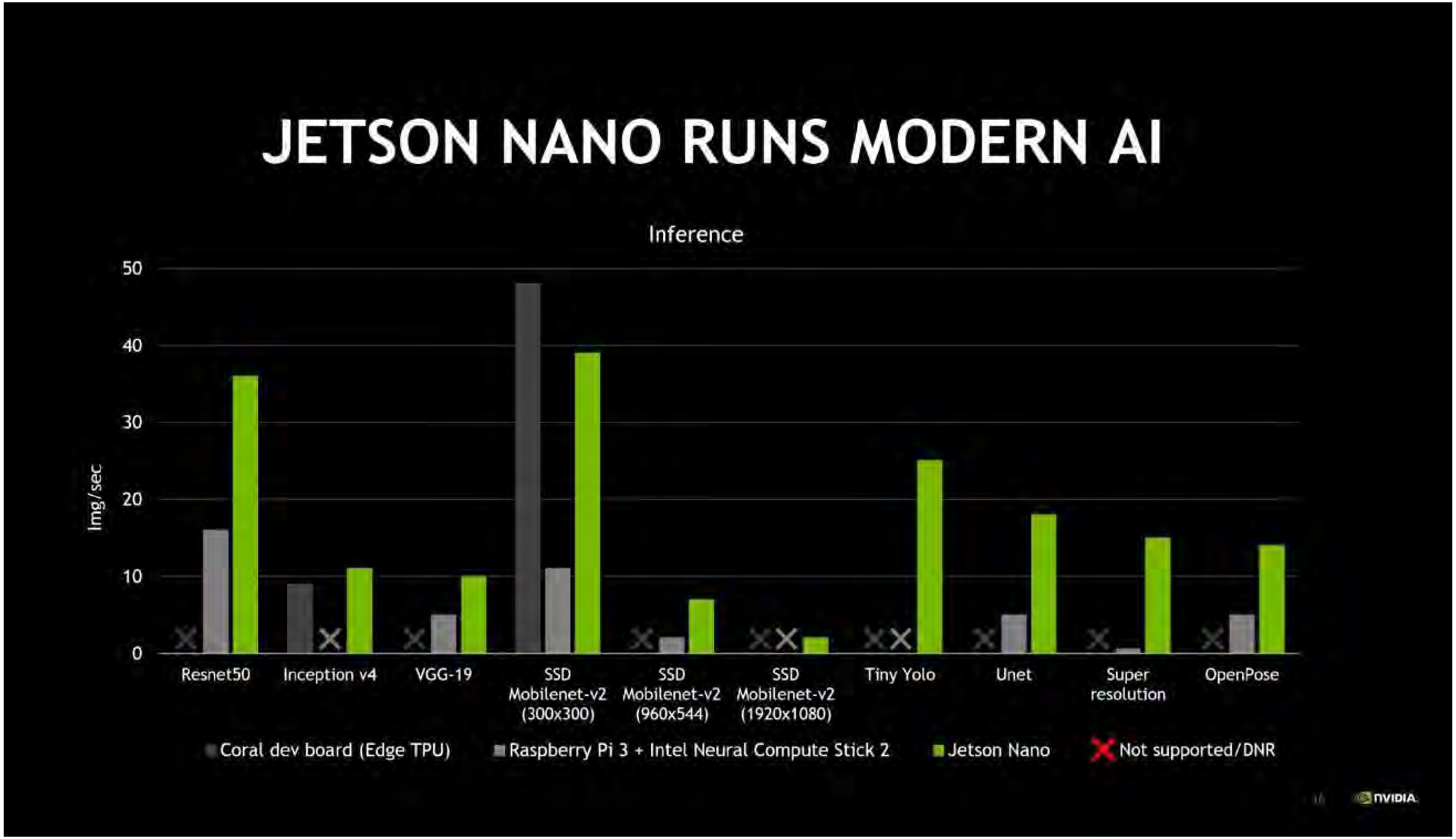
SLU

No Power Limit

```
+-----------------------------------------------------------------------------+
| NVIDIA-SMI 440.100      Driver Version: 440.100      CUDA Version: 10.2     |
|-------------------------------+----------------------+----------------------+
| GPU  Name        Persistence-M| Bus-Id        Disp.A | Volatile Uncorr. ECC |
| Fan  Temp  Perf  Pwr:Usage/Cap|         Memory-Usage | GPU-Util  Compute M. |
|===============================+======================+======================|
|   0  GeForce RTX 208...  Off  | 00000000:09:00.0 Off |                  N/A |
| 50%   71C    P2   255W / 260W |  10436MiB / 11019MiB |    100%      Default |
+-------------------------------+----------------------+----------------------+
|   1  GeForce RTX 208...  Off  | 00000000:0B:00.0 Off |                  N/A |
| 51%   87C    P2   222W / 250W |   9907MiB / 11019MiB |    100%      Default |
+-------------------------------+----------------------+----------------------+
|   2  GeForce RTX 208...  Off  | 00000000:43:00.0 Off |                  N/A |
| 49%   72C    P2   258W / 260W |   9907MiB / 11019MiB |    100%      Default |
+-------------------------------+----------------------+----------------------+
|   3  GeForce RTX 208...  Off  | 00000000:45:00.0 Off |                  N/A |
| 48%   87C    P2   247W / 250W |   9924MiB / 11016MiB |    100%      Default |
+-------------------------------+----------------------+----------------------+
```

200W Power Limit

```
+-----------------------------------------------------------------------------+
| NVIDIA-SMI 440.100      Driver Version: 440.100      CUDA Version: 10.2     |
|-------------------------------+----------------------+----------------------+
| GPU  Name        Persistence-M| Bus-Id        Disp.A | Volatile Uncorr. ECC |
| Fan  Temp  Perf  Pwr:Usage/Cap|         Memory-Usage | GPU-Util  Compute M. |
|===============================+======================+======================|
|   0  GeForce RTX 208...  Off  | 00000000:09:00.0 Off |                  N/A |
| 44%   66C    P2   202W / 200W |  10436MiB / 11019MiB |    100%      Default |
+-------------------------------+----------------------+----------------------+
|   1  GeForce RTX 208...  Off  | 00000000:0B:00.0 Off |                  N/A |
| 43%   84C    P2   196W / 200W |   9907MiB / 11019MiB |    100%      Default |
+-------------------------------+----------------------+----------------------+
|   2  GeForce RTX 208...  Off  | 00000000:43:00.0 Off |                  N/A |
| 43%   66C    P2   196W / 200W |   9907MiB / 11019MiB |    100%      Default |
+-------------------------------+----------------------+----------------------+
|   3  GeForce RTX 208...  Off  | 00000000:45:00.0 Off |                  N/A |
| 42%   82C    P2   190W / 200W |   9924MiB / 11016MiB |    100%      Default |
+-------------------------------+----------------------+----------------------+
```


JETSON NANO RUNS MODERN AI
Inference
50
40
30
20
10
0
Img/sec
Resnet50
Inception v4
VGG-19
SSD Mobilenet-v2 (300x300)
SSD Mobilenet-v2 (960x544)
SSD Mobilenet-v2 (1920x1080)
Tiny Yolo
Unet
Super resolution
OpenPose
Coral dev board (Edge TPU)
Raspberry Pi 3 + Intel Neural Compute Stick 2
Jetson Nano
Not supported/DNR
NVIDIA

How do we contribute to sustainable AI-development?

EDGE-computing

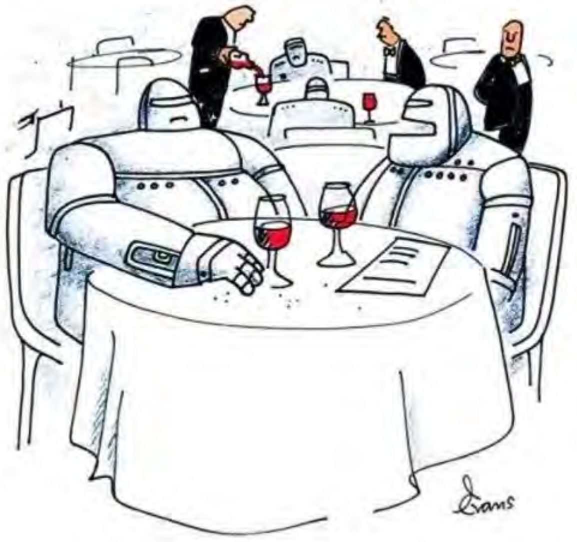
TinyML

“Building Blocks”-approach


Simple models and a lot of data trump more elaborate models based on less data.
More data beats clever algorithms, but better data beats more data.
-
Peter Norvig

Want to chat about AI? Collaborations? Virtual coffee?

oleksiy.guzhva@slu.se



'I can't imagine why they ever thought we'd take their jobs away.'