# Riding the Deep Learning Tsunami

Garth A. Gibson

Vector Institute for Artificial Intelligence
Carnegie Mellon University, University of Toronto

Algorithms Architecture Applications

40TH ANNIVERSARY LECTURE SERIES

# HPC and GPUs

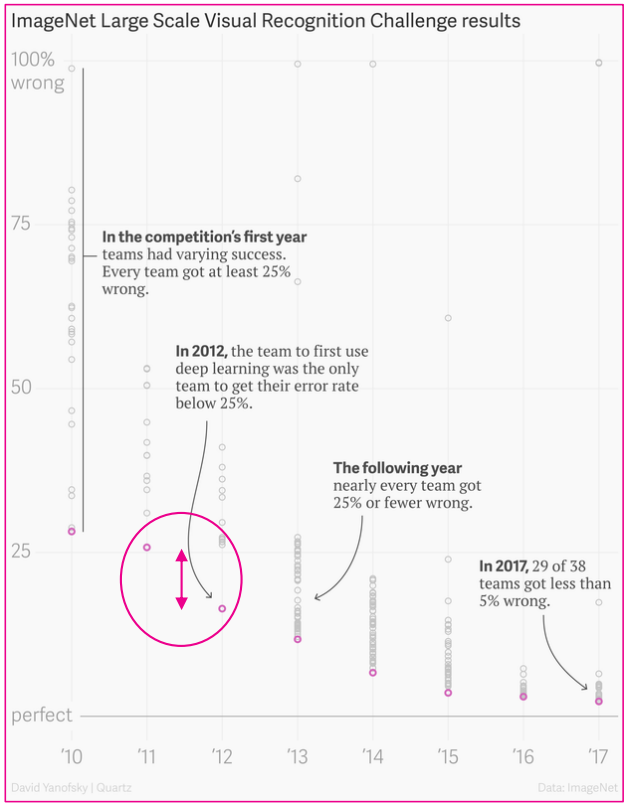
- 2008 IBM & LANL's Roadrunner exceeded 1 PetaFLOP/s using 13000 PowerXCell 8i coprocessors, derived from the same architecture as was used in Sony's 2006 Playstation 3 game console


- 2018 IBM & ORNL's Summit exceed 1 Exa-calcs/s using 27000+ Nvidia V100 tensor core matrix multiply function units for CoMet genomic analysis code
  - V100 Volta Tensor Cores developed to optimize AI's Deep Learning

# AI & GPUs

- Deep Learning's ability to compute a gradient for a model's parameters based on the error observed by a data sample is Back Propagation, a computationally intensive technique developed in the 1980s
- 2012 University of Toronto researchers code DL into a gaming GPU, delivering a big improvement in image classification

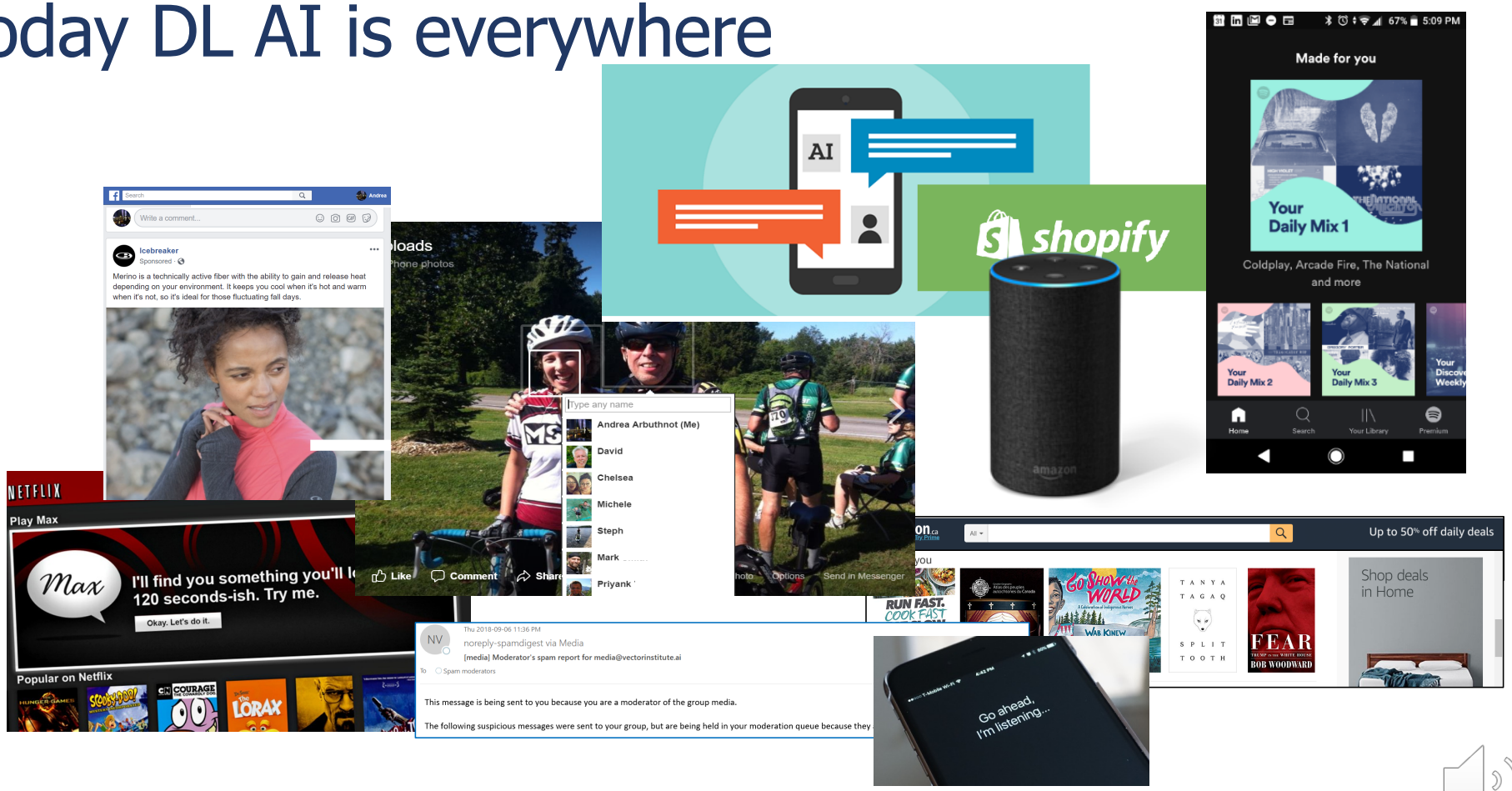
*Ilya Sutskever, Alex Krizhevsky and University Professor Geoffrey Hinton of the University of Toronto's Department of Computer Science (photo by John Guatto)*




https://qz.com/1034972/the-data-that-changed-the-direction-of-ai-research-and-possibly-the-world/

"Indeed, **if the artificial intelligence boom we see today could be attributed to a single event**, it would be the announcement of the 2012 ImageNet challenge result…" – Quartz magazine

# Today DL AI is everywhere



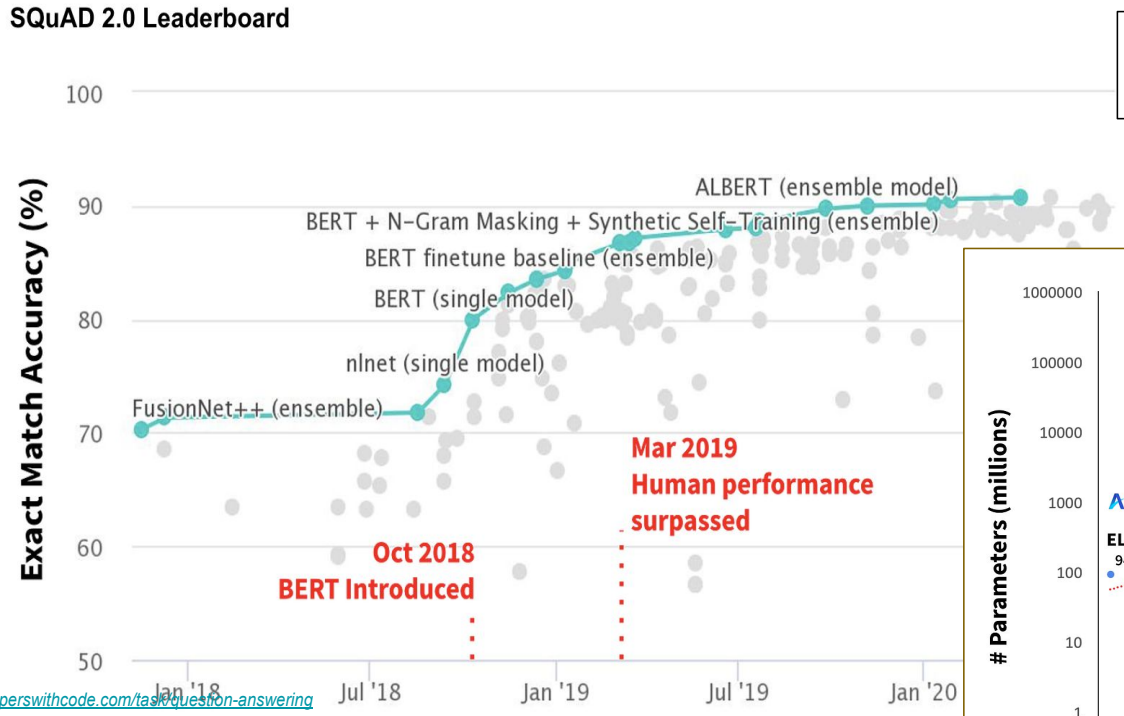
# 2018 ACM A.M. Turing Award goes to Hinton, LeCun, Bengio


Dr. Geoffrey Hinton, received the 2018 ACM A.M. Turing Award – often referred to as the 'Nobel Prize' of computer science – for his foundational research in deep learning and neural networks. Dr. Hinton is Chief Scientific Advisor, Vector Institute; VP and Engineering Fellow, Google; and Emeritus Professor, University of Toronto.

Along with Dr. Hinton, the 2018 winners also include fellow Canadian, Dr. Yoshua Bengio, and New York-based Dr. Yann LeCun, the latter of whom did his postdoctoral research with Dr. Hinton at the University of Toronto.

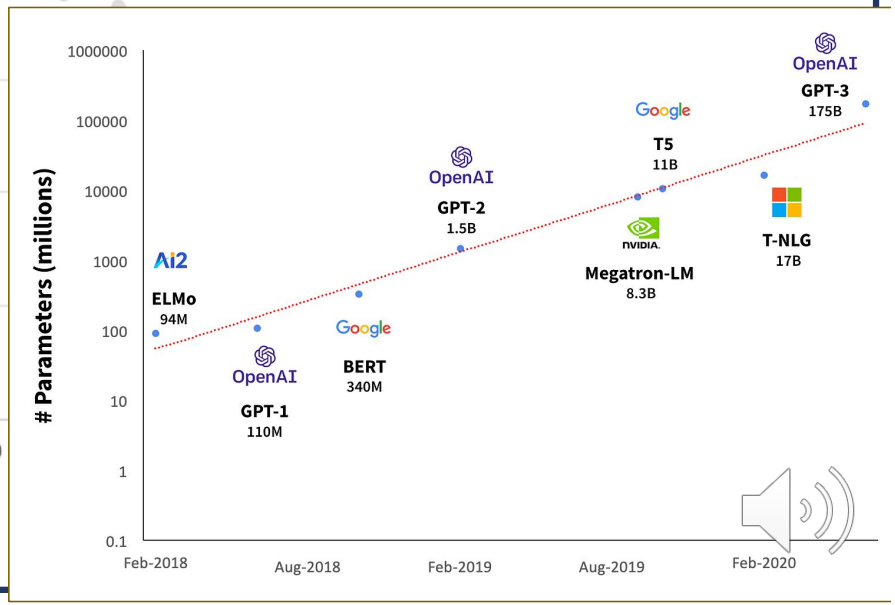
Alan Turing laid foundations of modern computing and inspired the AI revolution by asking "Can machines think?" This year's winners of his namesake award have transformed the field with artificial neural networks.

# DL ”transforms” NLP Question Answering – Deep Transformer Models surpass human quality




Source: https://paperswithcode.com/task/question-answering

Google’s 2021 Mixture-of-experts parallel models run on Trillion parameters



# Recent NLP Question Answering – Deep Transformer Models surpass human quality

- Recent work by OpenAI on NLP scaling shows
  - Additional iterations (=more compute) diminishing returns for fixed data or model size
  - Given more data and larger models (meaning larger memory), more compute improves model prediction accuracy
  - Optimal input data sizes scale as (Compute)^.27
  - Optimal model size scales as (Compute)^.73
  -

$$N \propto C_{\min}^{0.73}, \quad B \propto C_{\min}^{0.24}, \quad S \propto C_{\min}^{0.03}, \quad D = B \cdot S$$

Optimal model size | optimal batch size | optimal training steps | **optimal dataset size**

*Source: Scaling Laws for Neural Language Models* https://arxiv.org/pdf/2001.08361.pdf

# Scalable NLP Available

- E.g. Nvidia's Megatron-LM scales GPT-2 to 8.3B parameters, delivers SOTA accuracy
- Open Source, small mods to PyTorch Transformer [easy to use]
- Combines model parallel (partitioned parameters) & data parallel (partitioned data batches)
- Top systems teams competing to decrease dependence on uniform hardware, bisection BW

**Megatron-LM: Training Multi-Billion Parameter**

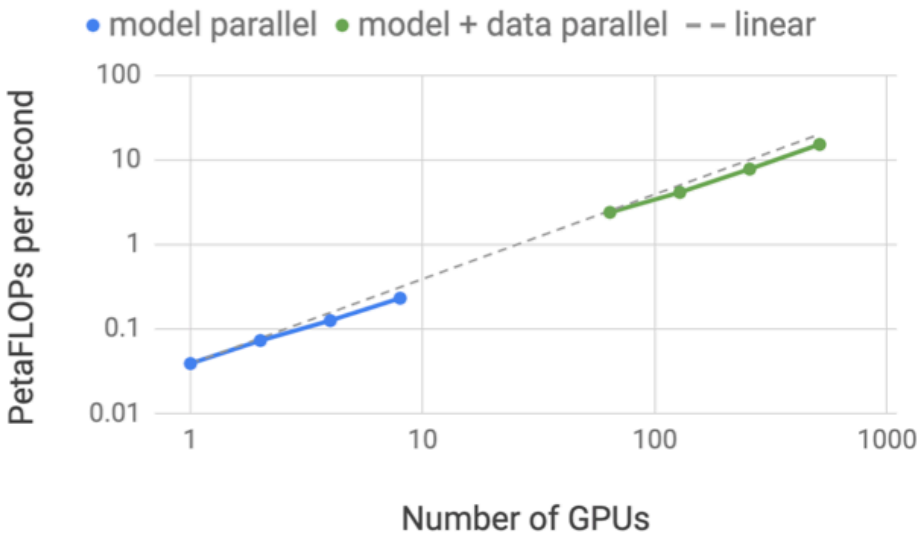
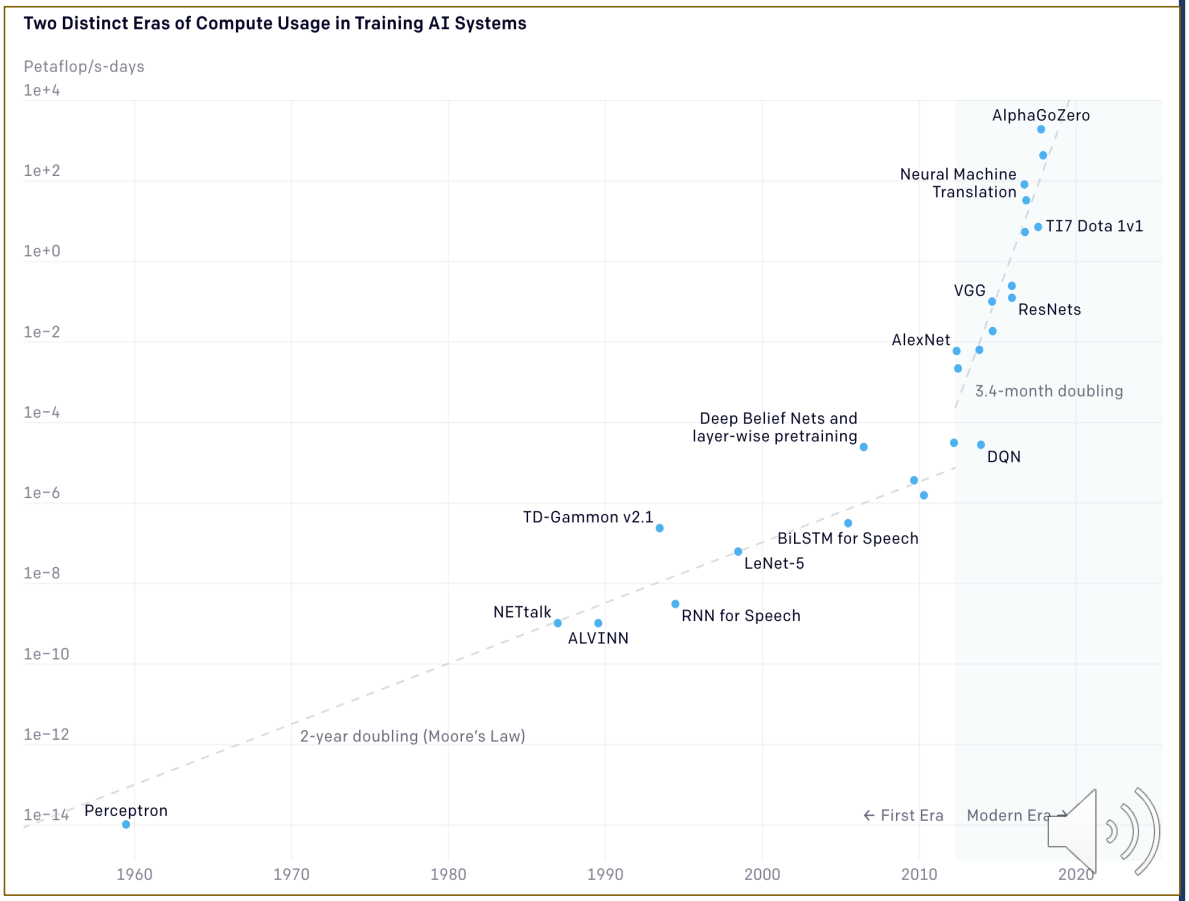



*Figure 1.* Model (blue) and model+data (green) parallel FLOPS as a function of number of GPUs. Model parallel (blue): up to 8-way model parallel weak scaling with approximately 1 billion parameters per GPU (e.g. 2 billion for 2 GPUs and 4 billion for 4 GPUs). Model+data parallel (green): similar configuration as model parallel combined with 64-way data parallel.

*Source: Megatron-LM: Training Multi-Billion Parameter Language Models Using Model Parallelism, 2020, arXiv:1909.08053v4, https://github.com/NVIDIA/Megatron-L...*

# Record Setting AI Doubles Compute Use Every 3.4 months

- Among AI researchers seeking to set new records, the accuracy benefits of large models is driving rapid increase in demand for computing
  - 2X compute every 3.4 months
- Vector emphasizes new methods, applications, optimization algorithms while also more compute than most universities

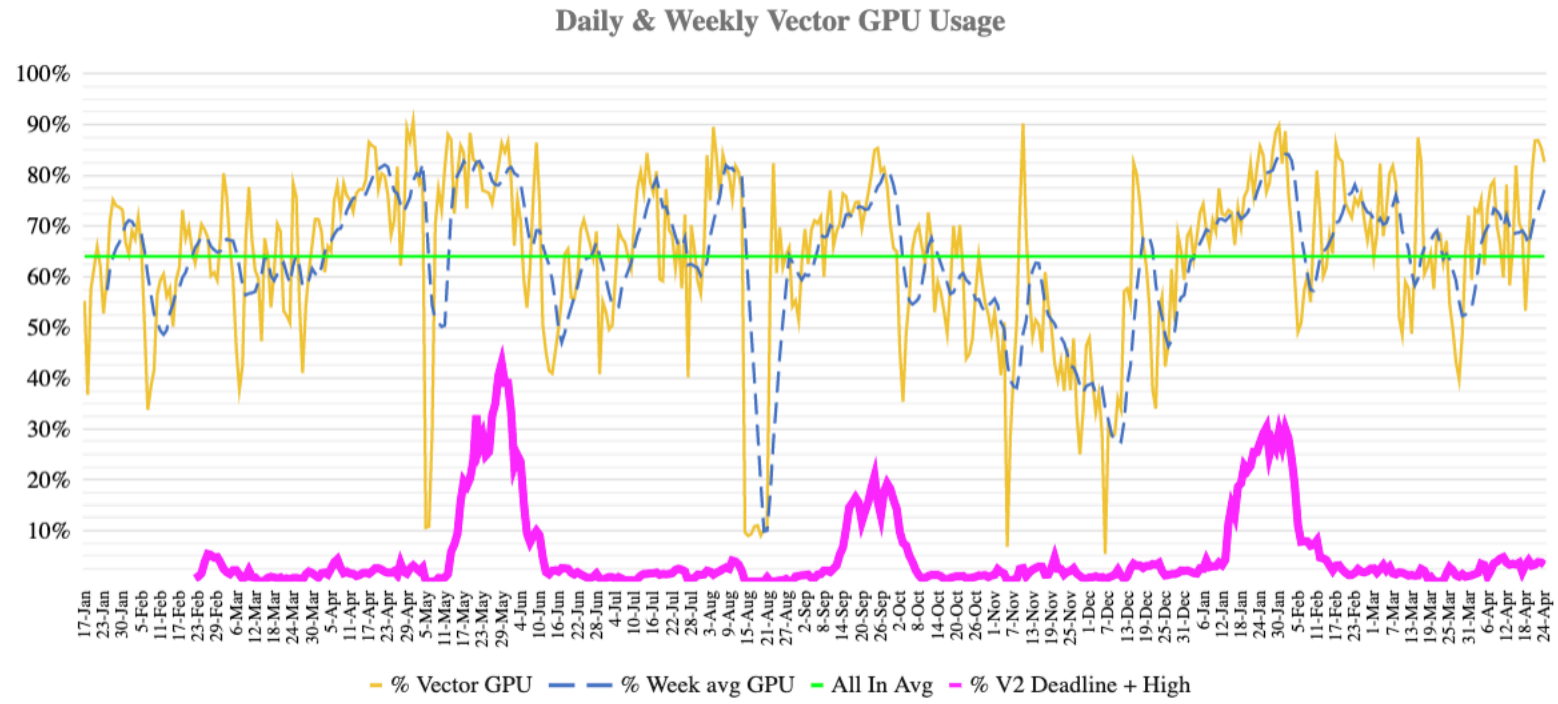


# Scaling ML/DL Expert Workforce Development

- Vector was created in 2017 to drive excellence in AI and foster economic growth using AI
  - To attract and retain excellence in academic researchers & educators
  - To build regional innovation, growth, productivity in large and small companies by exploiting AI
  - To foster growth of a large regional expert AI workforce at all levels
- $135M over its first 5 years
  - 1/3 federal, 1/3 Ontario, 1/3 private
  - ~$90M more recently committed

- In first four years Vector has grown to:
  - 500+ researchers, incl
    - 35 faculty, 100 faculty affiliates
  - 100 master scholarships per year
    - 1,100+ students in programs at 12 universities
  - 50 enterprise & startup/scaleup sponsors
  - 2 government sponsors (Federal, Ontario)
  - Independent but affiliated with 5 universities, 2 research hospitals
  - 60 professional staff

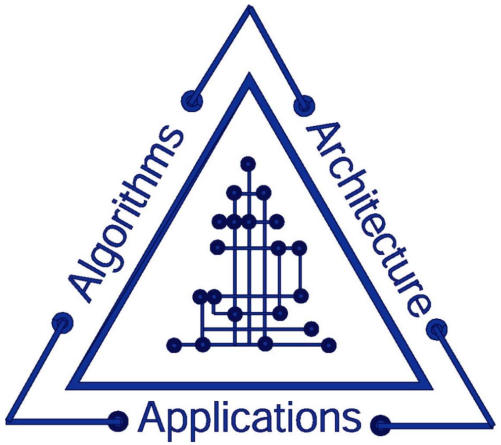
# Because this is a compute community …




~1150 NVidia GPUs providing ~10 PetaFLOP/s (FP32) utilized 64% including downtime – AI conference deadline in red

# Vectorinstitute.ai


VECTOR INSTITUTE




40TH ANNIVERSARY LECTURE SERIES