SIAM Conference on
Computational Science
and Engineering
February 25-March 1, 2019
Spokane Convention Center
Spokane, Washington, USA

Welcome to Friday!

1 Mar 2019

Thank you!

- **Executing a successful conference takes a small army**
- **Let's thank again**
  - Organizing Committee
  - SIAM Staff
  - Spokane Convention Center Staff
  - A/V Staff
  - Our Corporate, Government, and Academic Sponsors

Thank you!

- **SIAM CSE is also successful because of community involvement**
- **Thank you to YOU!**
  - Plenary Speakers
  - Panelists
  - Minisymposium and Minisymposterium Organizers
  - Special Event Organizers
  - Poster Judges
  - Presenters
  - Attendees

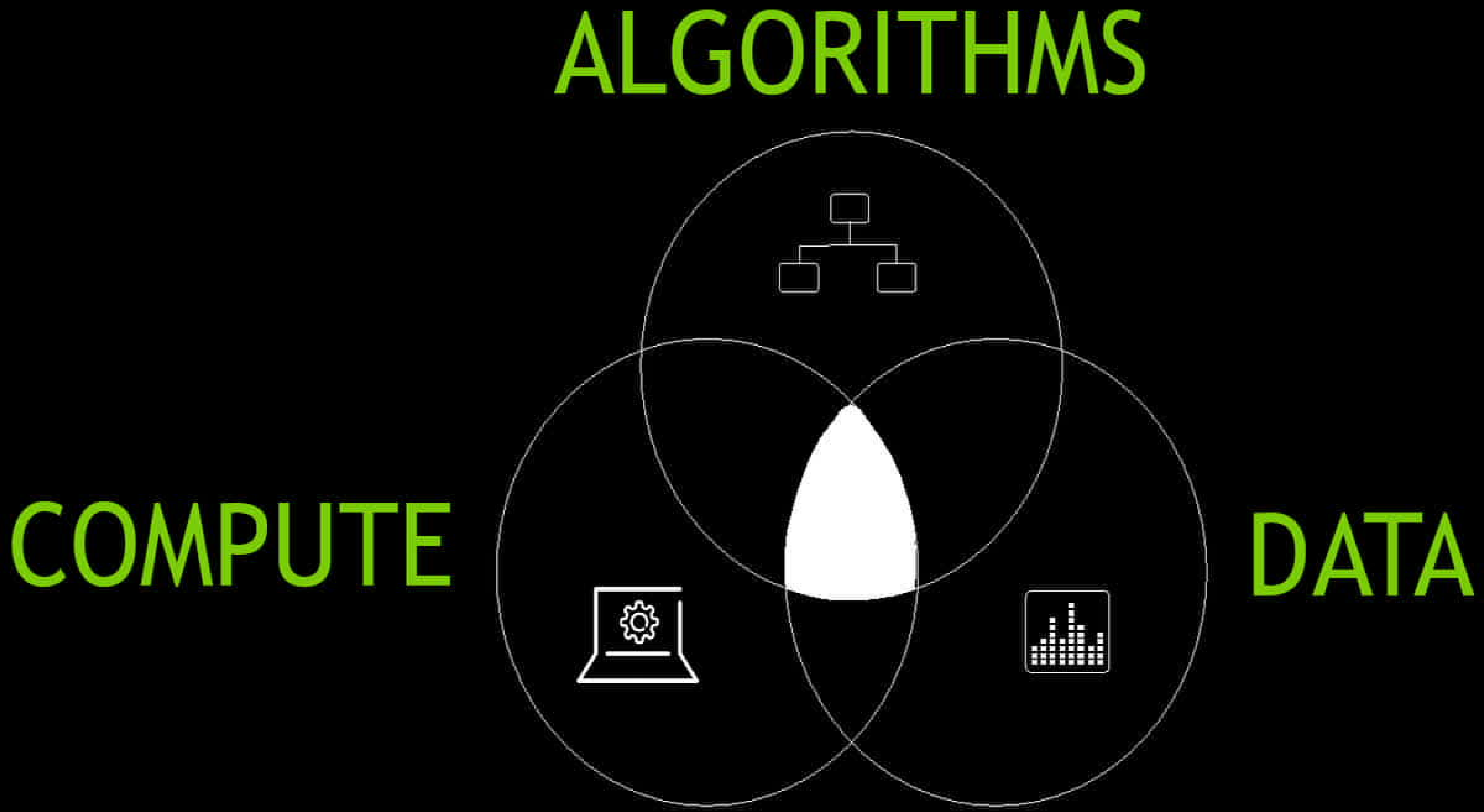
Today's Main Events

Time	Room	Events
8:30-9:15 am	100BC	*Role of Tensors in Machine Learning* **Anima Anandkumar**, Amazon and California Institute of Technology

Odds and Ends

- **Your opinion counts!**
  - Look for an email with a link to an online evaluation form
  - Please give us feedback so that we can improve future conferences

TRINITY OF AI/ML



EXAMPLE AI TASK: IMAGE CLASSIFICATION

DATA: LABELED IMAGES FOR TRAINING AI

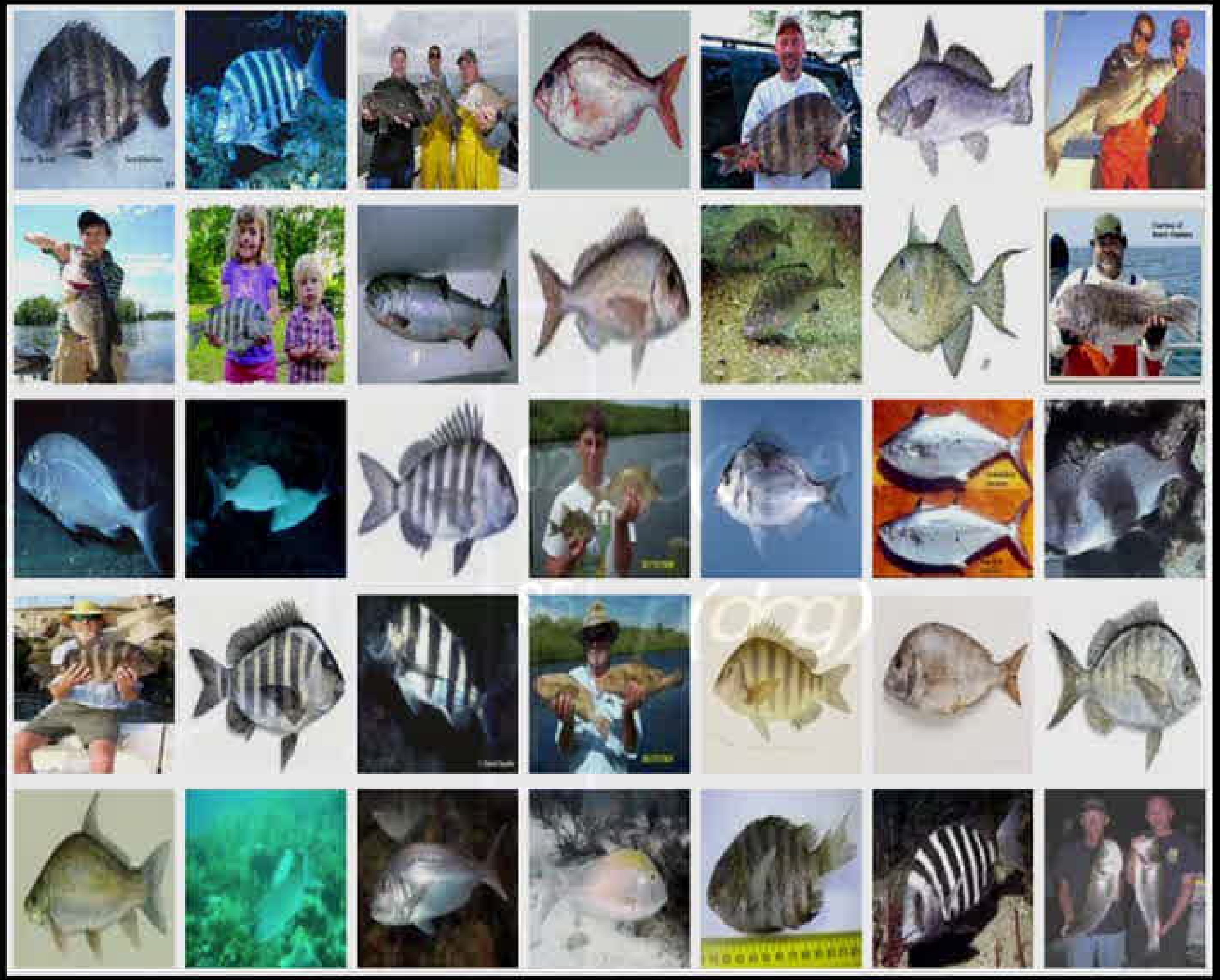
Picture credits: Image-net.org, ZDnet.com

DATA: LABELED IMAGES FOR TRAINING AI

- 14 million images and 1000 categories.
- Largest database of labeled images.

Picture credits: Image-net.org, ZDnet.com

DATA: LABELED IMAGES FOR TRAINING AI

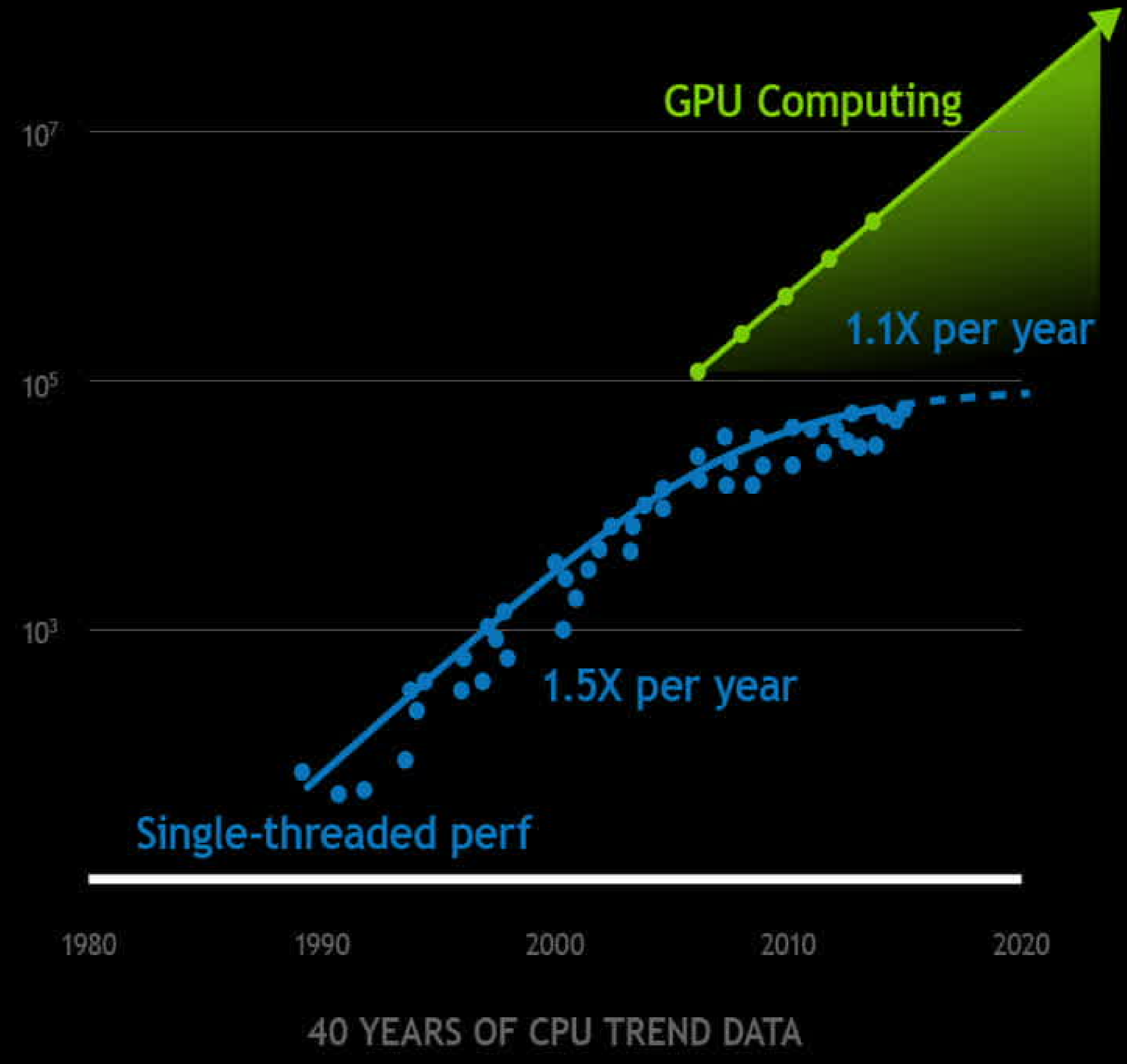


- 14 million images and 1000 categories.
- Largest database of labeled images.

- Images in Fish category.
- Captures variations of fish.

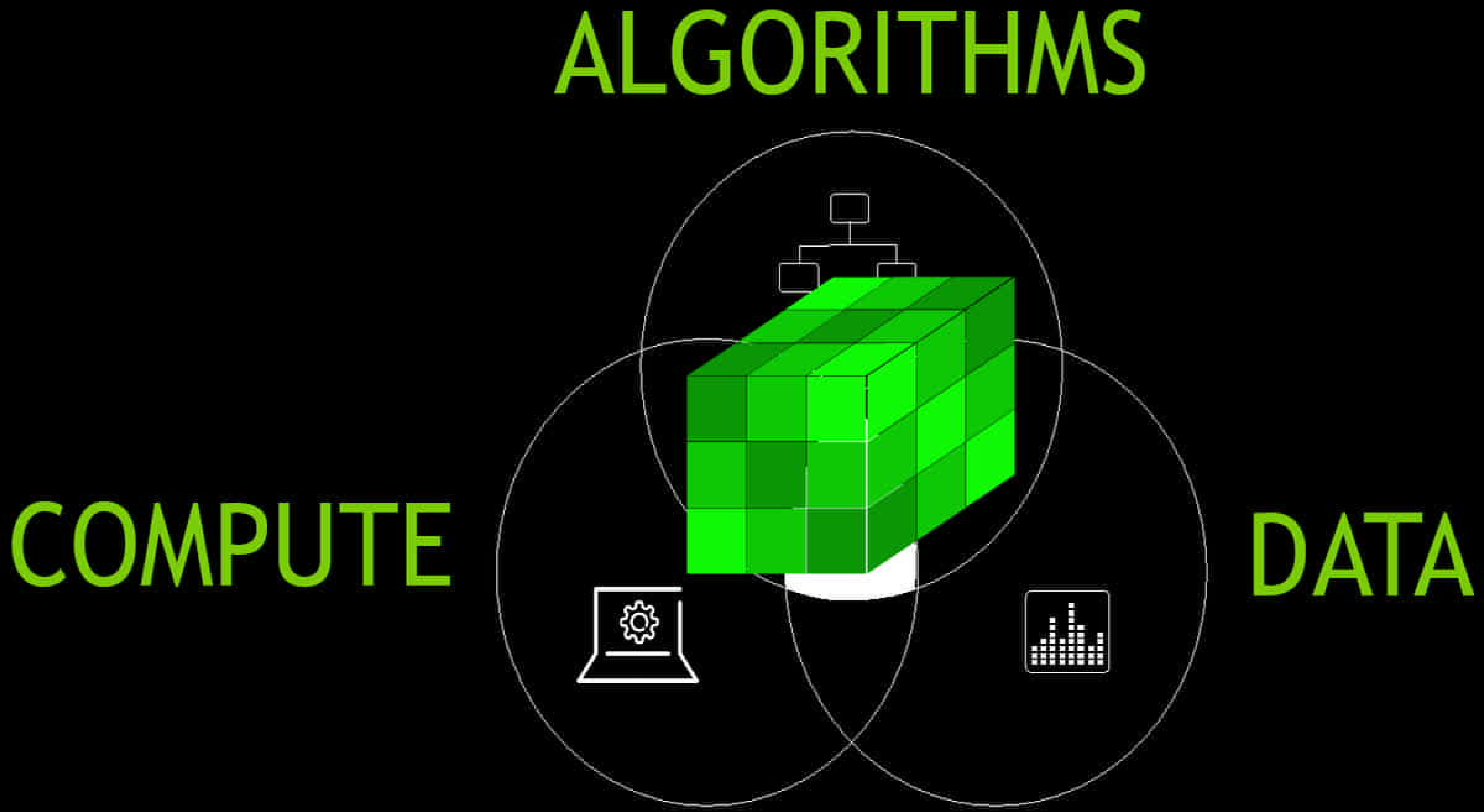
Picture credits: Image-net.org, ZDnet.com

COMPUTE INFRASTRUCTURE FOR AI: GPU



MOORE'S LAW: A SUPERCHARGED LAW

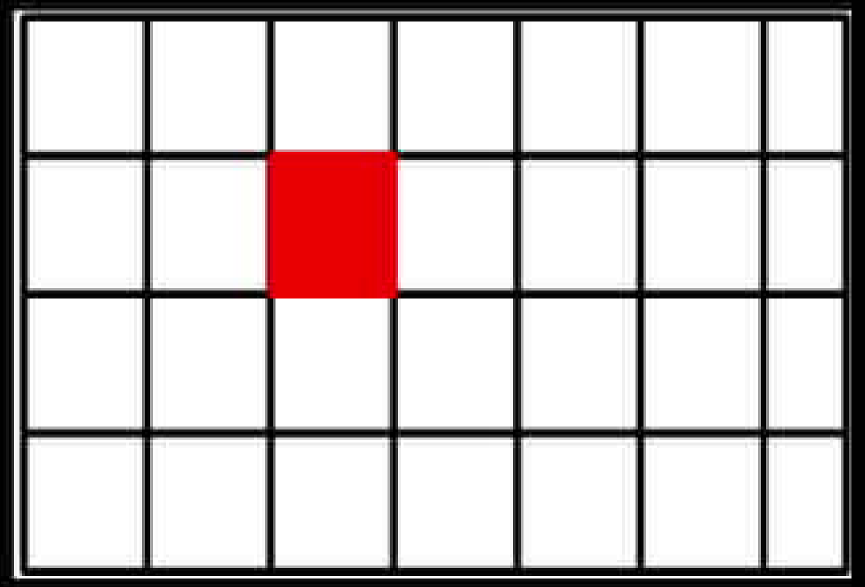
TENSORS PLAY A CENTRAL ROLE



TENSORS FOR ML ALGORITHMS
ENCODE HIGHER ORDER MOMENTS

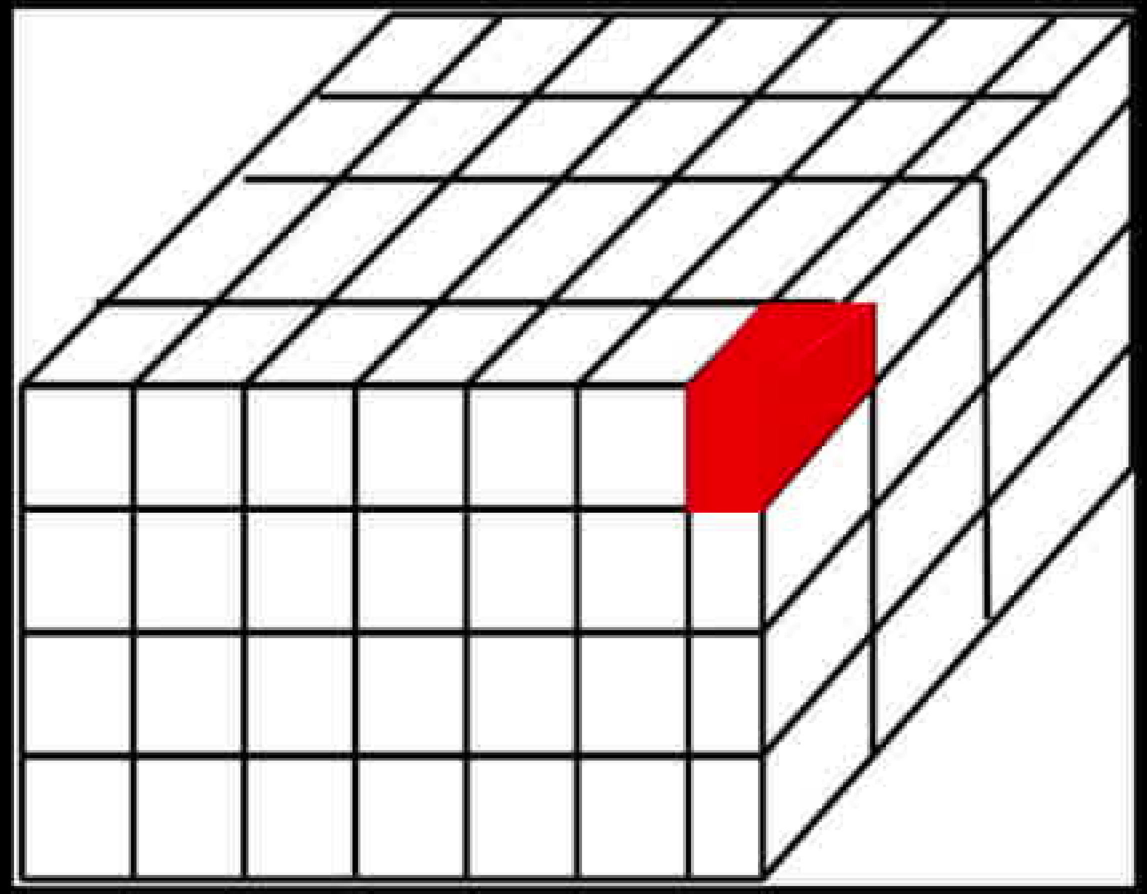
Pairwise correlations

$$E(x \otimes x)_{i,j} = E(x_i x_j)$$

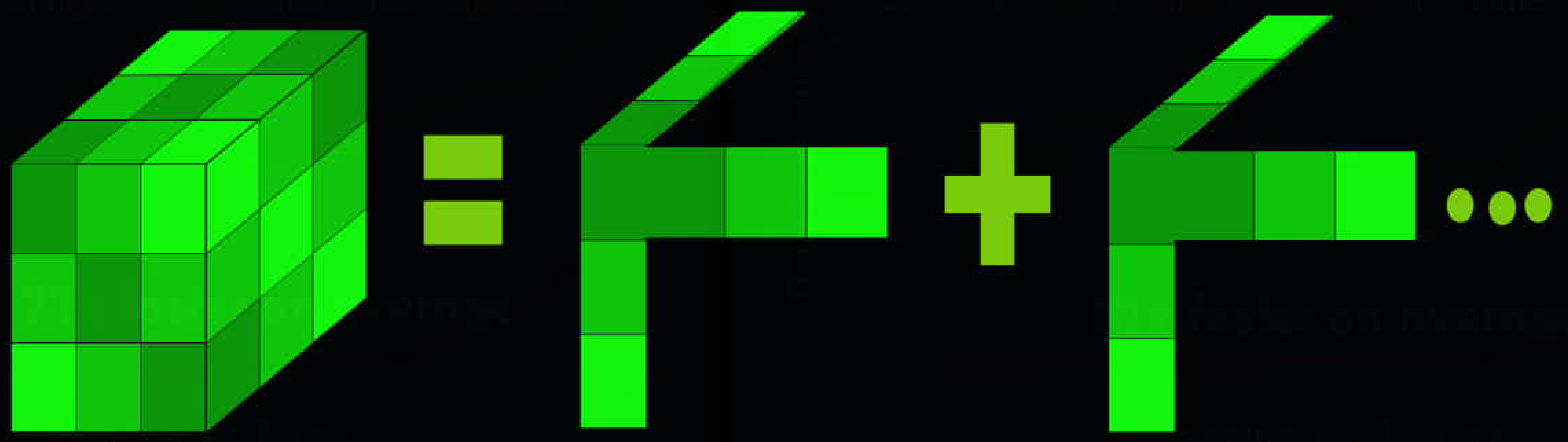


Third order correlations

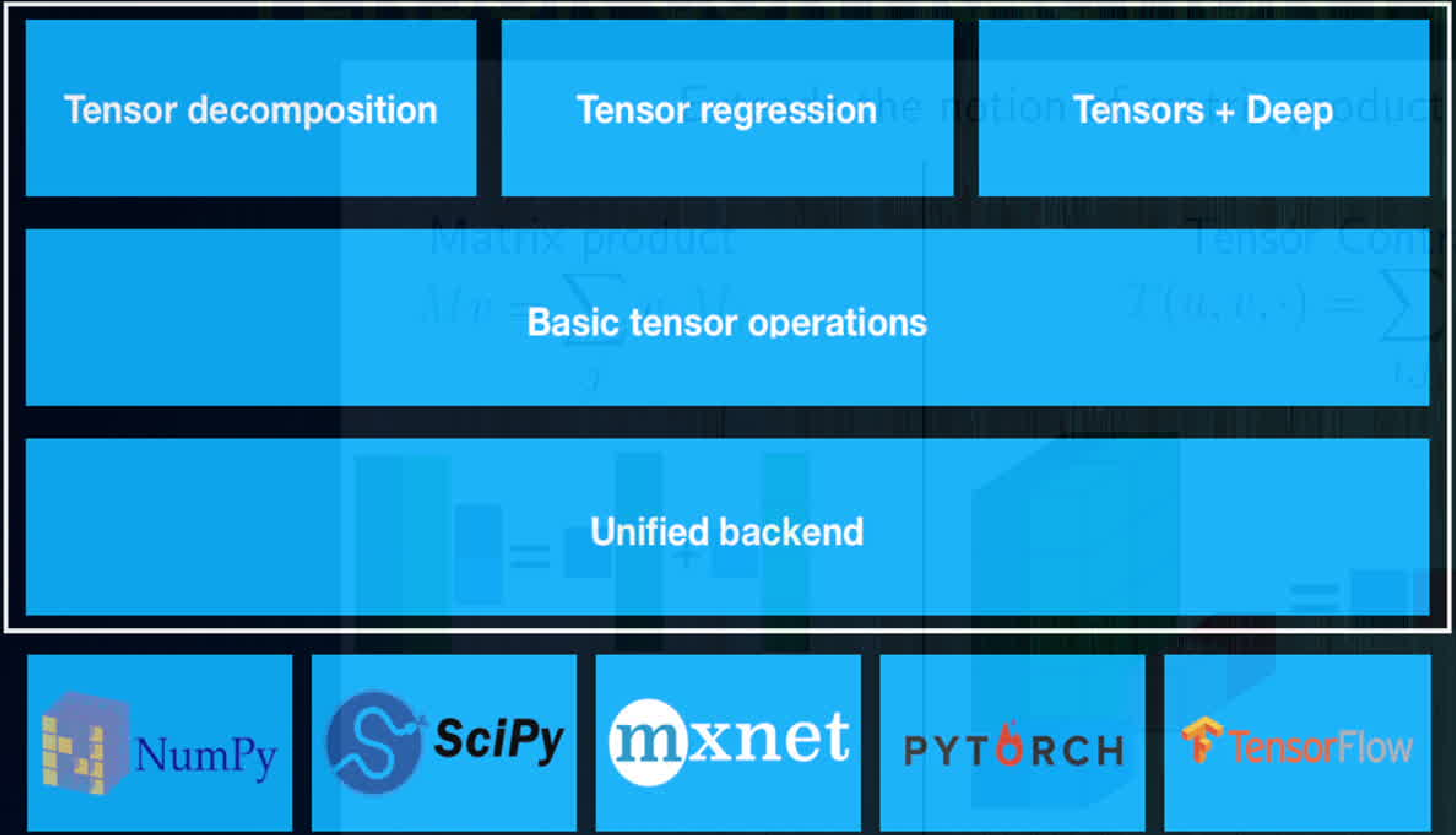
$$E(x \otimes x \otimes x)_{i,j,k} = E(x_i x_j x_k)$$



TENSORS FOR MODELING:
TOPIC DETECTION IN TEXT



TENSORLY: HIGH-LEVEL API FOR TENSOR ALGEBRA



- Python programming
- User-friendly API
- Multiple backends: flexible + scalable
- Example notebooks in repository

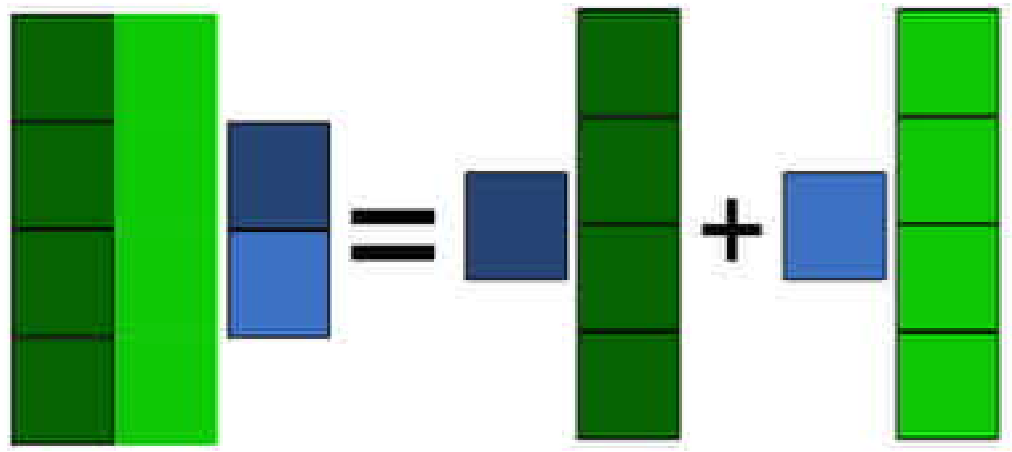
TENSORS FOR COMPUTE

TENSOR CONTRACTION PRIMITIVE

Extends the notion of matrix product

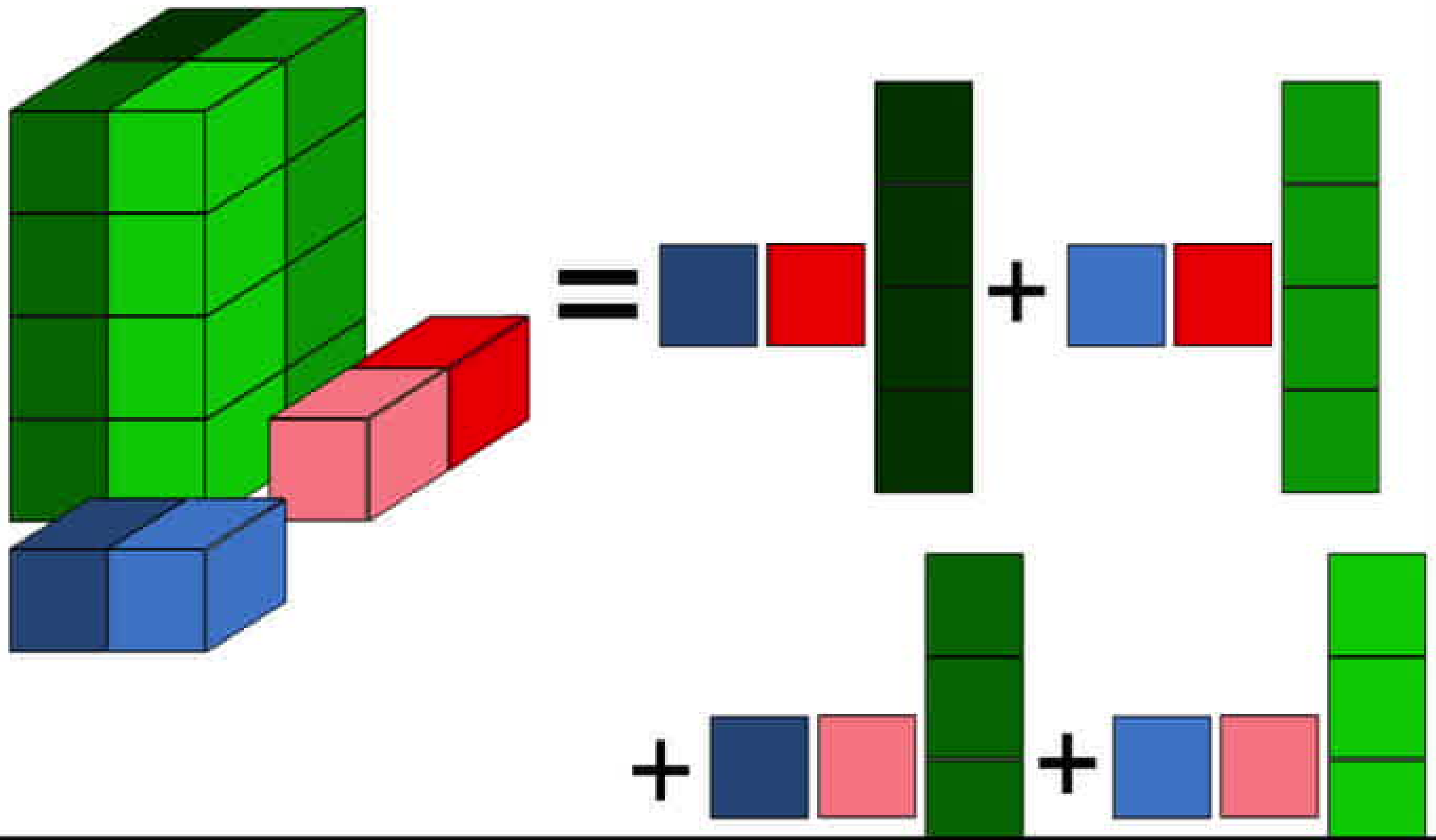
Matrix product

$$Mv = \sum_j v_j M_j$$



Tensor Contraction

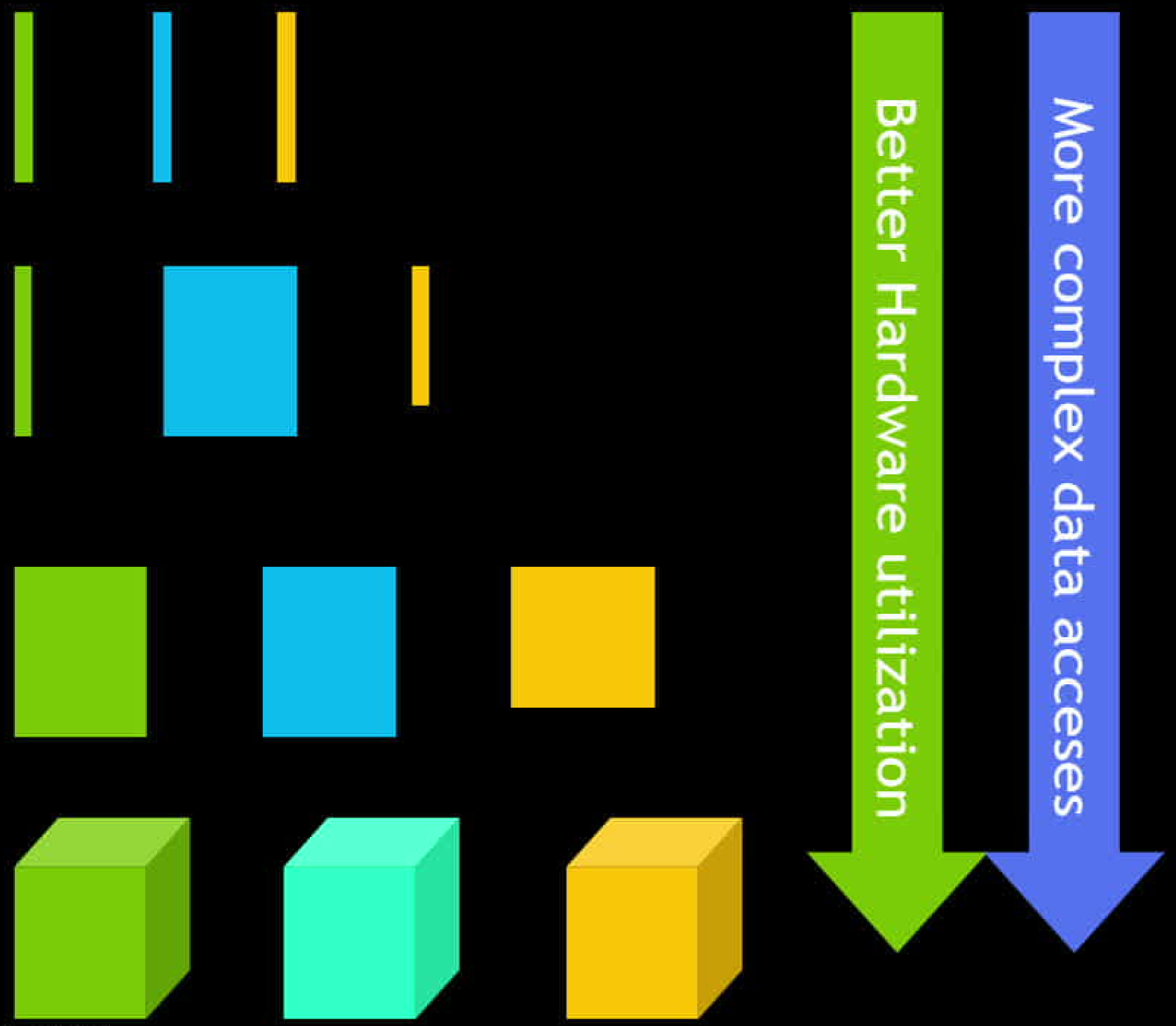
$$T(u, v, \cdot) = \sum_{i,j} u_i v_j T_{i,j,:}$$



TENSOR PRIMITIVES?

History & Future

- 1969 – BLAS Level 1: Vector-Vector
- 1972 – BLAS Level 2: Matrix-Vector
- 1980 – BLAS Level 3: Matrix-Matrix
- Now? – BLAS Level 4: Tensor-Tensor



Kim, Jinsung, et al. "Optimizing Tensor Contractions in CCSD (T) for Efficient Execution on GPUs." (2018).

NEW PRIMITIVE FOR TENSOR CONTRACTIONS

$$C[p] = \alpha\, \mathsf{op}(A[p])\, \mathsf{op}(B[p]) + \beta\, C[p]$$

- Pointer-to-Pointer BatchedGEMM requires memory allocation and pre-computation.
- Solution: StridedBatchedGEMM with fixed strides.
  - Special case of Pointer-to-pointer BatchedGEMM.
  - No Pointer-to-pointer data structure or overhead.

```
cublas<T>gemmStridedBatched(cublasHandle_t handle,
                            cublasOperation_t transA, cublasOperation_t transB,
                            int M, int N, int K,
                            const T* alpha,
                            const T* A, int ldA1, int strideA,
                            const T* B, int ldB1, int strideB,
                            const T* beta,
                            T* C, int ldC1, int strideC,
                            int batchCount)
```

A FEW OTHER RESEARCH THREADS..

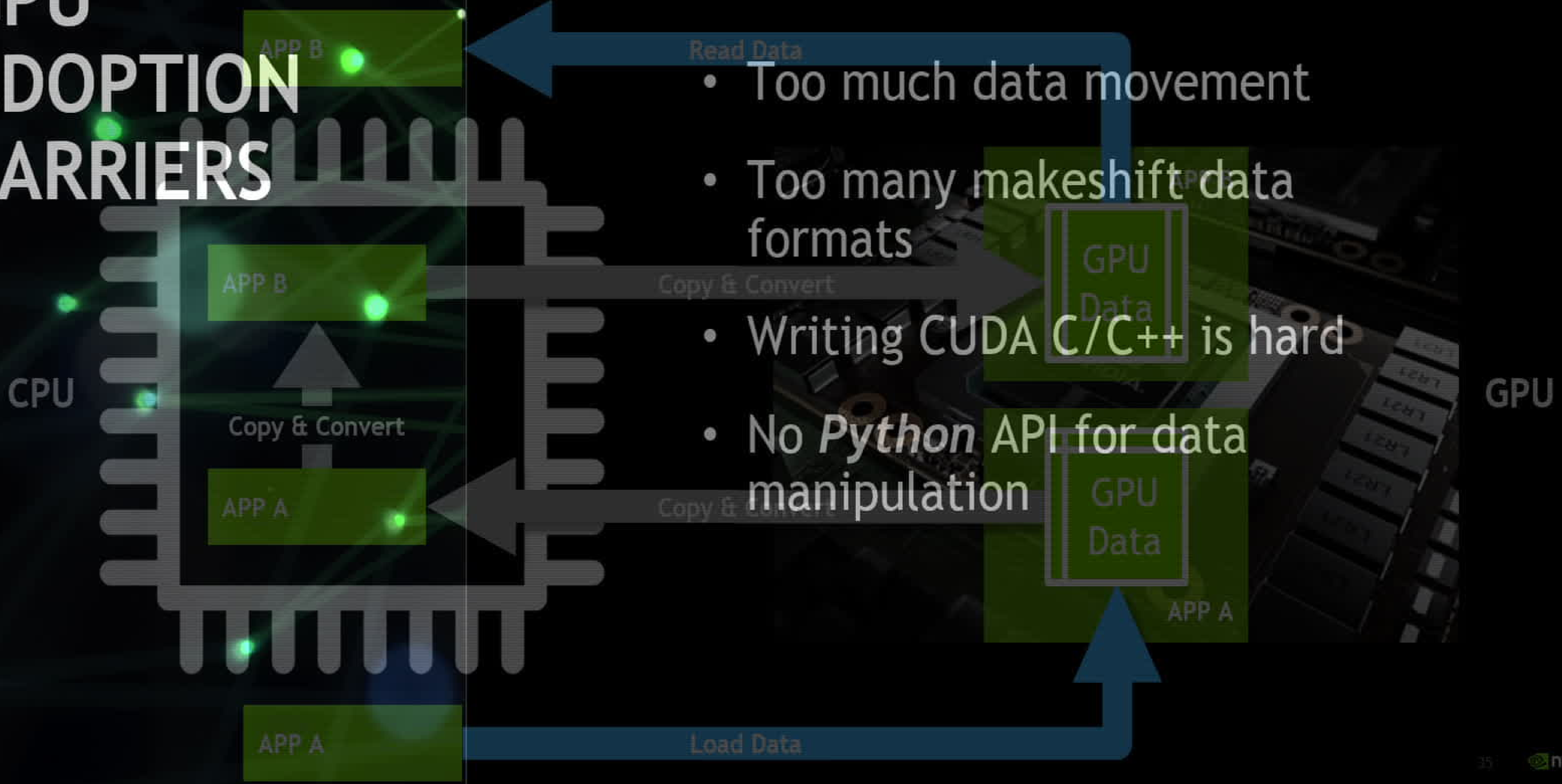
Yes GPUs are fast but …

DATA MOVEMENT AND TRANSFORMATION

The bane of productivity and performance

GPU ADOPTION BARRIERS

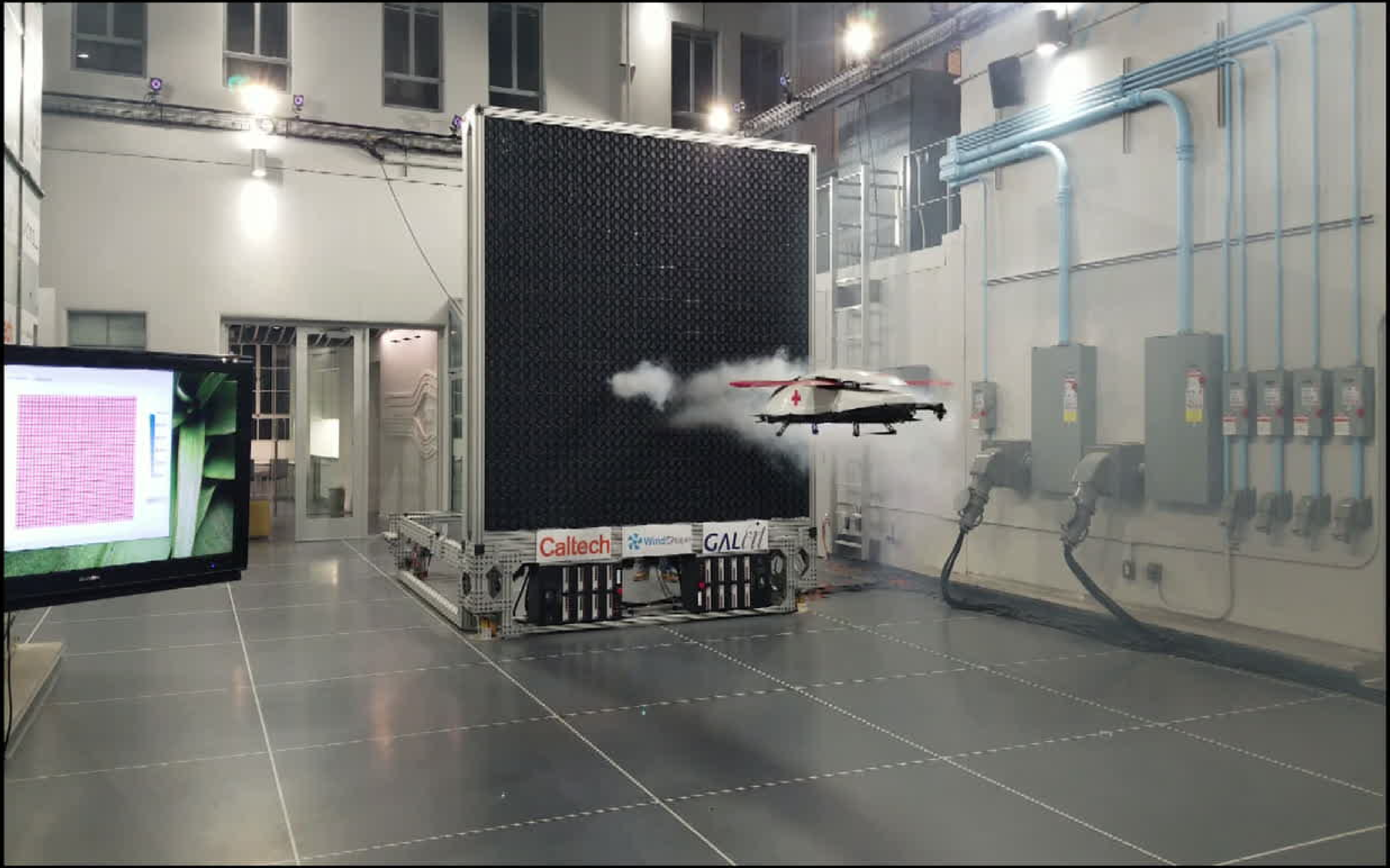
- Too much data movement
- Too many makeshift data formats
- Writing CUDA C/C++ is hard
- No *Python* API for data manipulation



CAST @ CALTECH

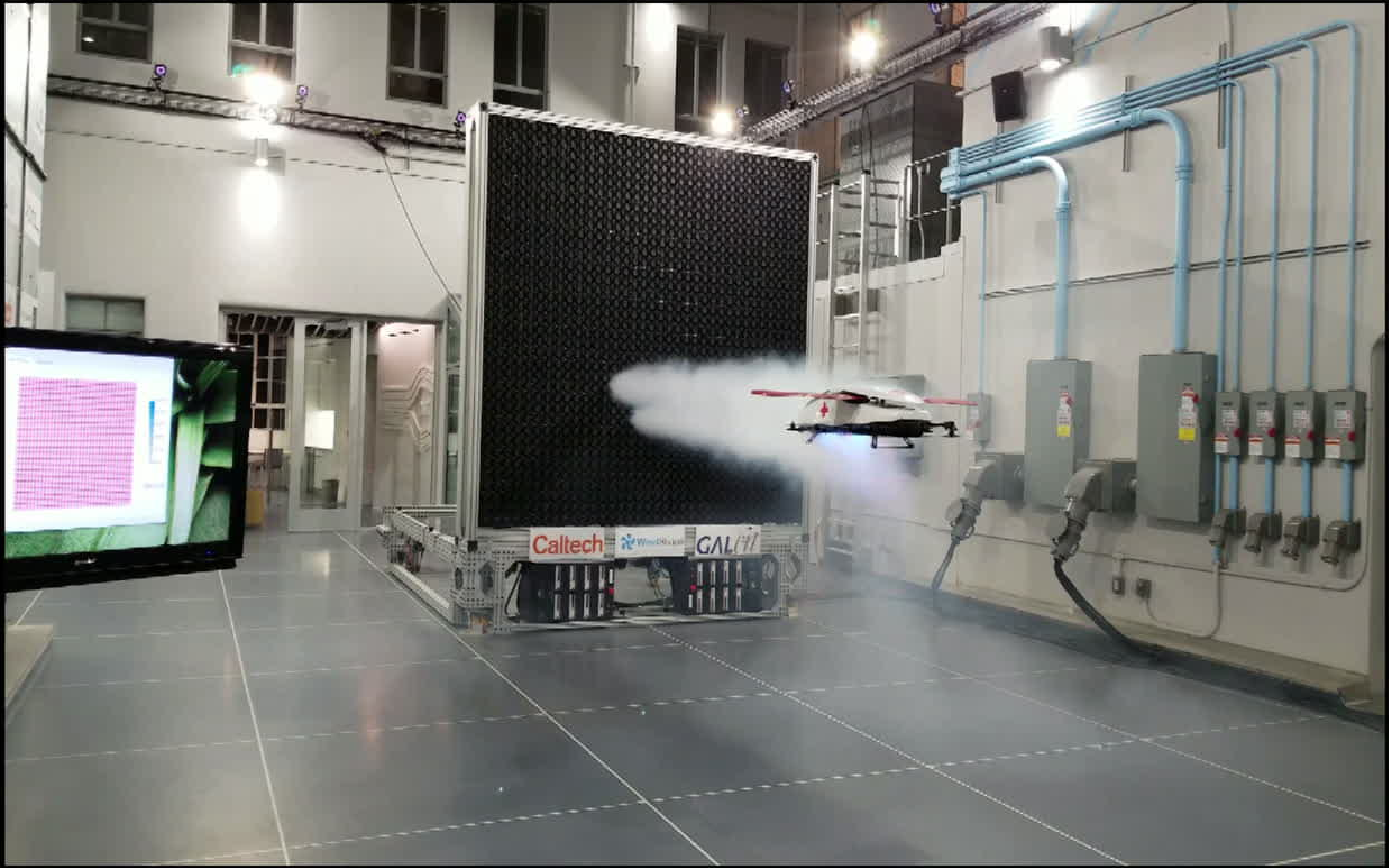
DRONE TESTING LAB

CAST @ CALTECH
DRONE TESTING LAB



CAST @ CALTECH

DRONE TESTING LAB

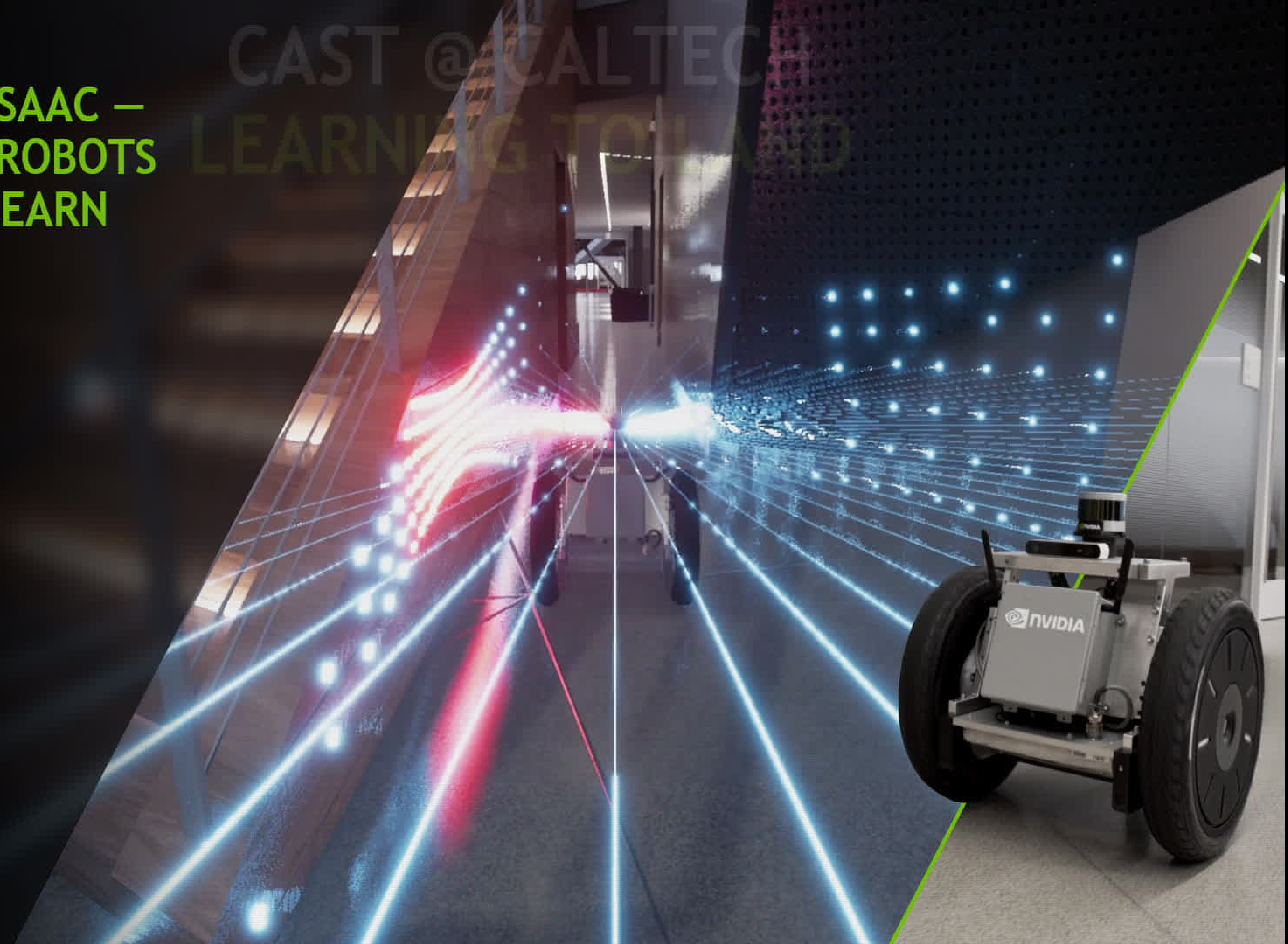


CAST @ CALTECH

LEARNING TO LAND

NVIDIA ISAAC — WHERE ROBOTS GO TO LEARN

NVIDIA ISAAC — WHERE ROBOTS GO TO LEARN



CAST @ CALTECH

LEARNING TO LAND

MIND & BODY
NEXT-GENERATION AI

Instinctive:

Fine-grained reactive Control

Behavioral:

Sense and react to human

Deliberative:

Making and adapting plans

Multi-Agent:

Acting for the greater good