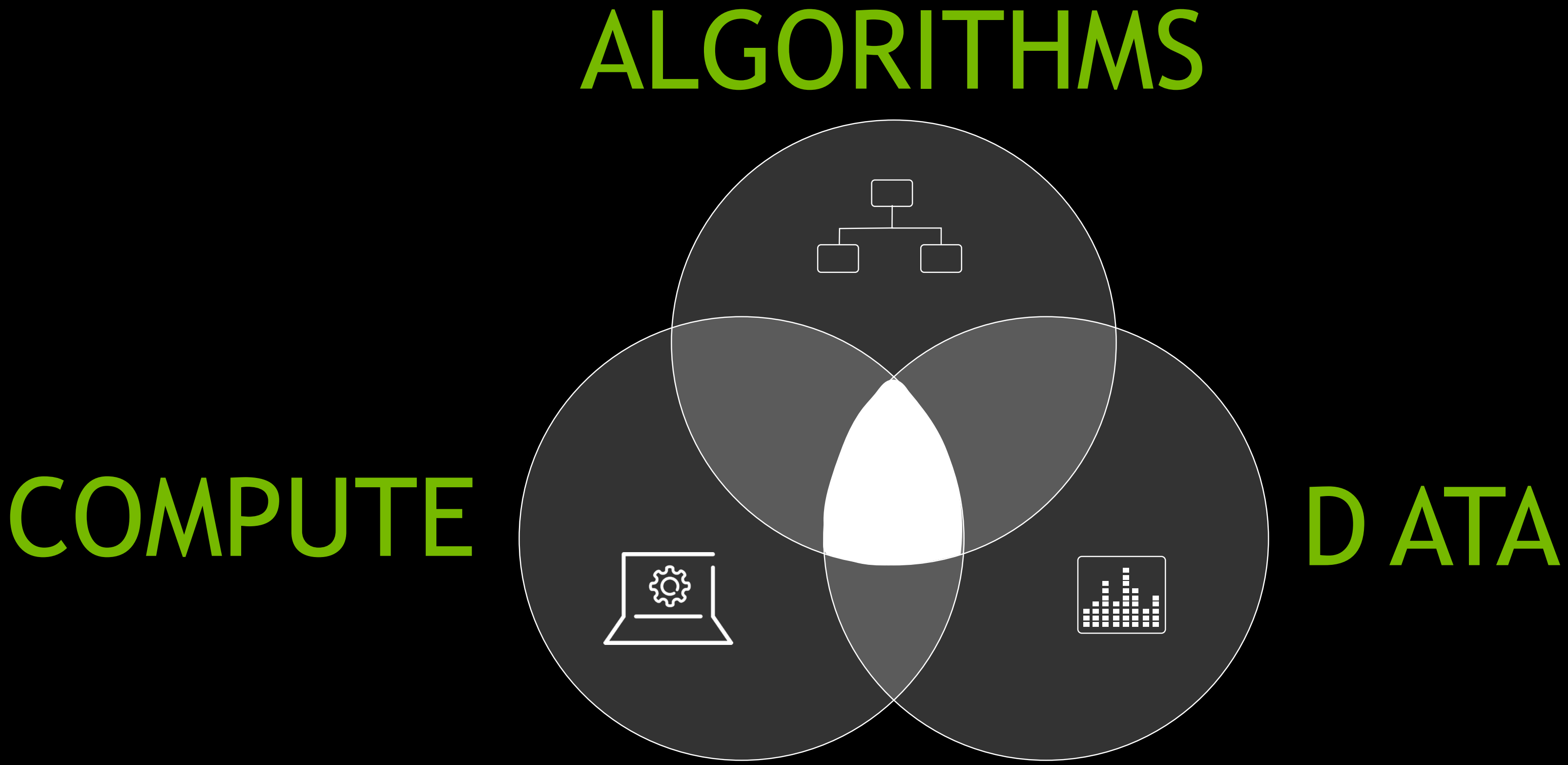
Anima Anandkumar

Caltech

NVIDIA

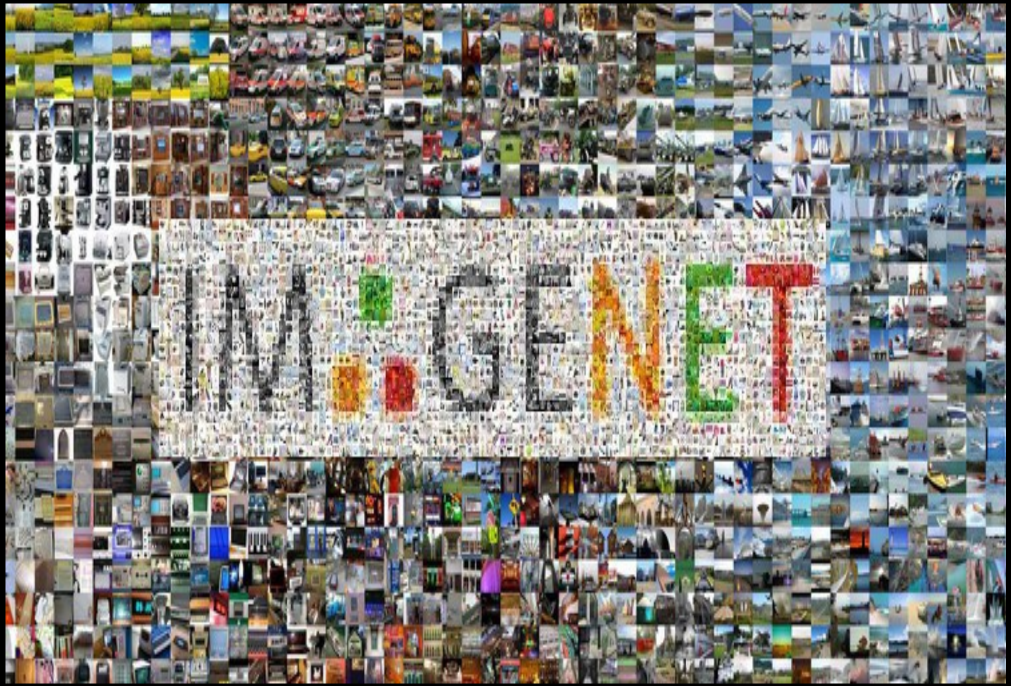
TRINITY OF AI: DATA + ALGORITHMS + COMPUTE

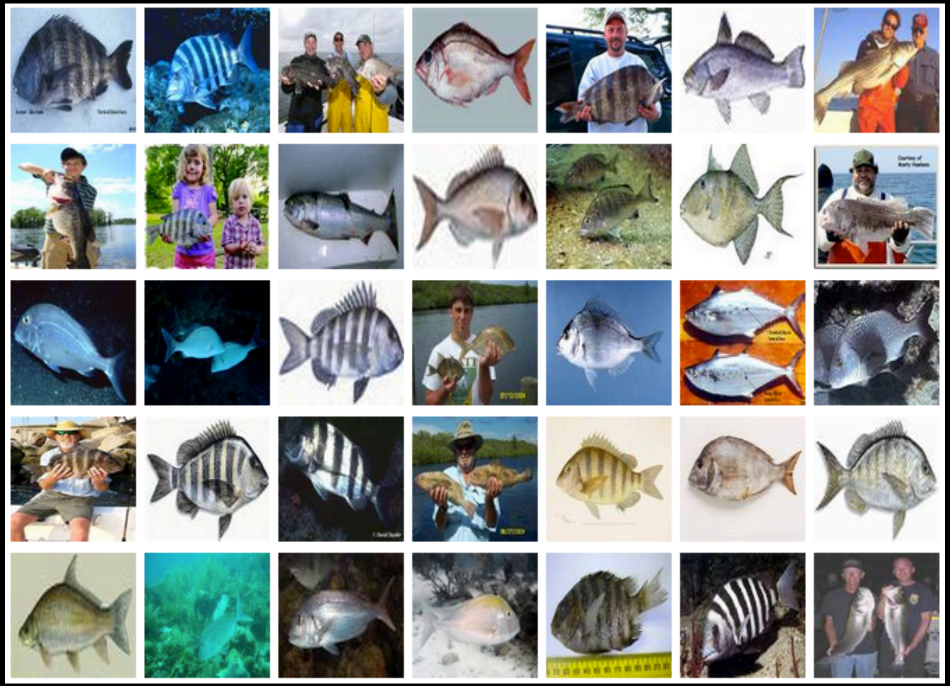
TRINITY OF AI/ML



EXAMPLE AI TASK: IMAGE CLASSIFICATION

DATA: LABELED IMAGES FOR TRAINING AI



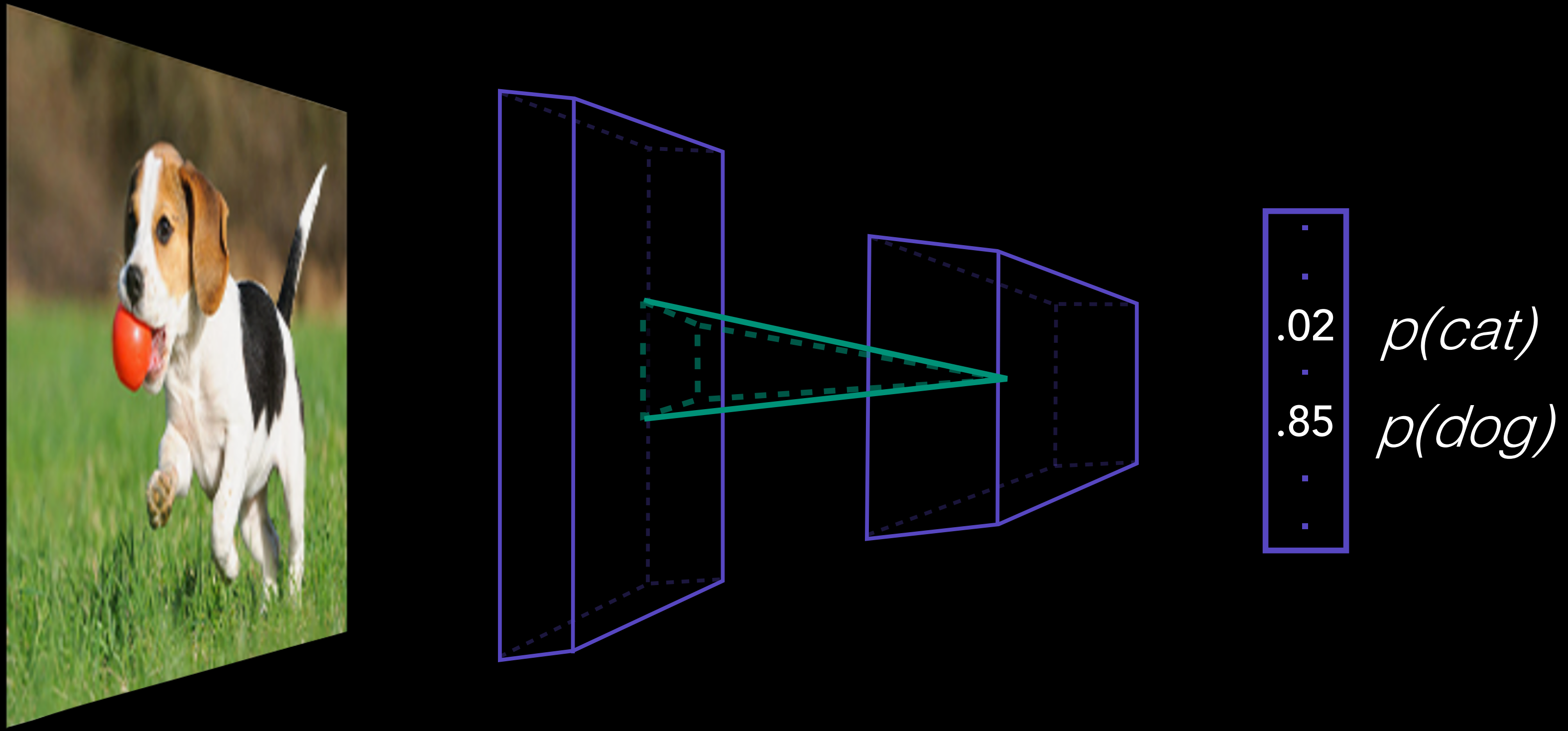


- 14 million images and 1000 categories.
- Largest database of labeled images.

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

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

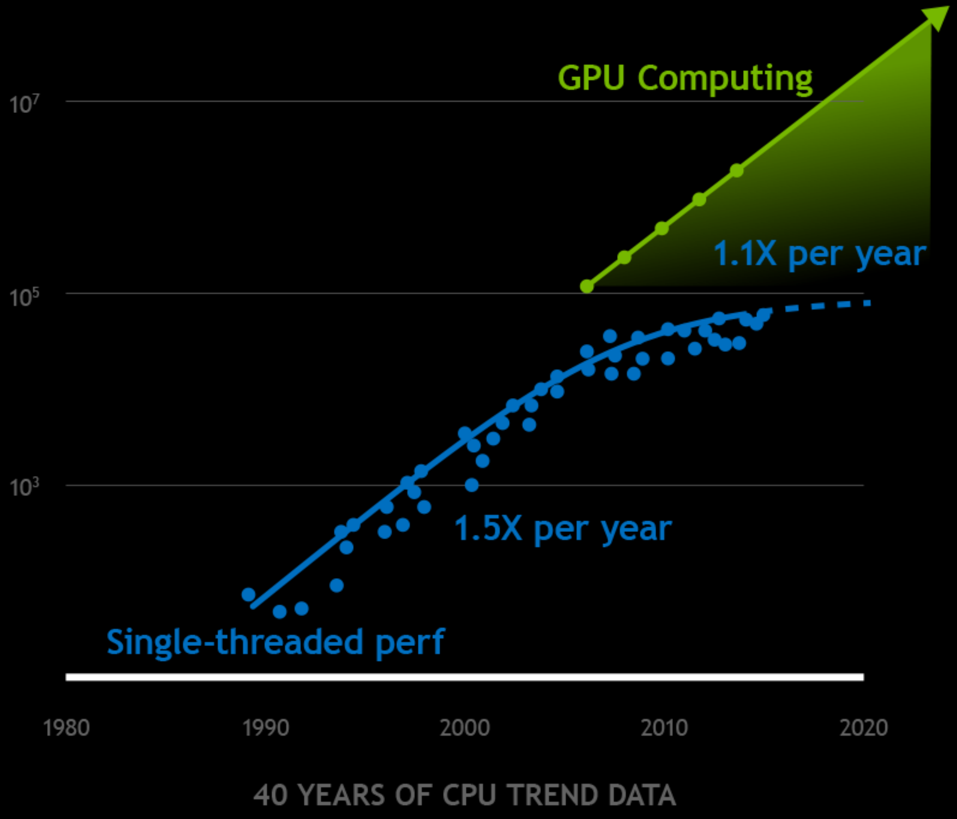
MODEL: CONVOLUTIONAL NEURAL NETWORK



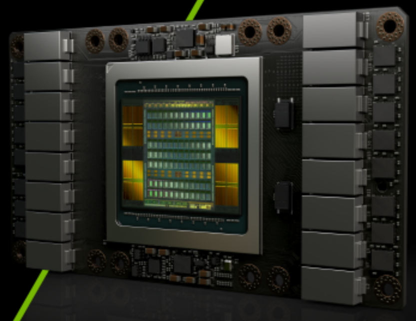
- **Deep learning:** Many layers give large capacity for model to learn from data
- **Inductive bias:** Prior knowledge about natural images.

COMPUTE INFRASTRUCTURE FOR AI: GPU

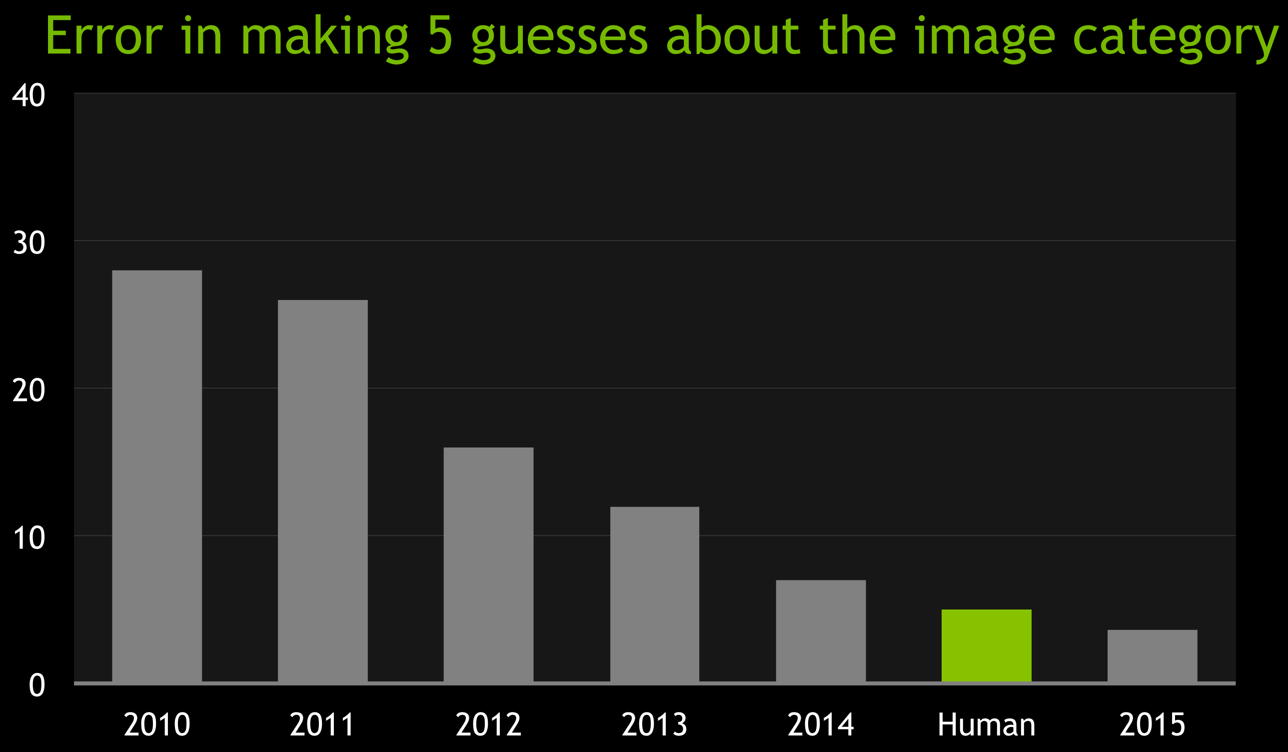
- More than a billion operations per image.
- NVIDIA GPUs enable parallel operations.
- Enables Large-Scale AI.



MOORE’S LAW: A SUPERCHARGED LAW



PROGRESS IN TRAINING IMAGENET



Need Trinity of AI : Data + Algorithms + Compute

Statista: Statistics Portal

Dealing with Data Scarcity

DATA IS EVERYTHING

Credits: Liza Donnelly at Women in Data Science (WIDS), Stanford 2019

LACK OF LABELED DATA IN MANY DOMAINS

Strategies to cope with it

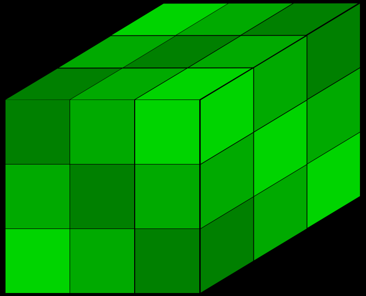
- Semi-supervised learning
- Active learning
- Crowdsourcing
- Domain adaptation/transfer learning
- Domain knowledge and structure

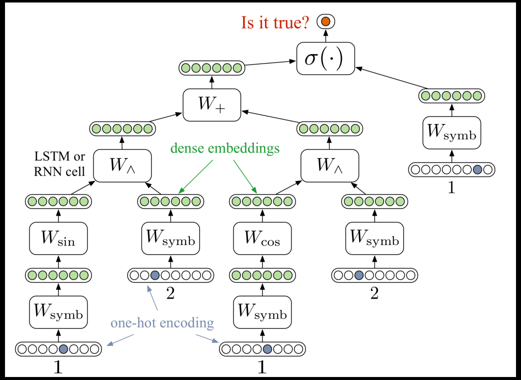
USE OF PRIORS FOR DATA EFFICIENCY

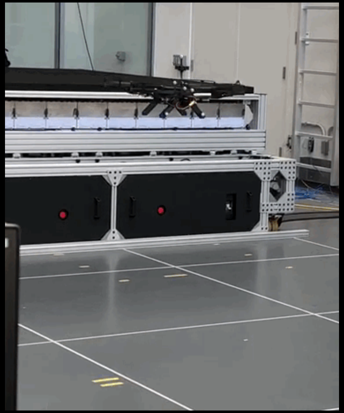
Learning = Data + Priors

Examples of Priors

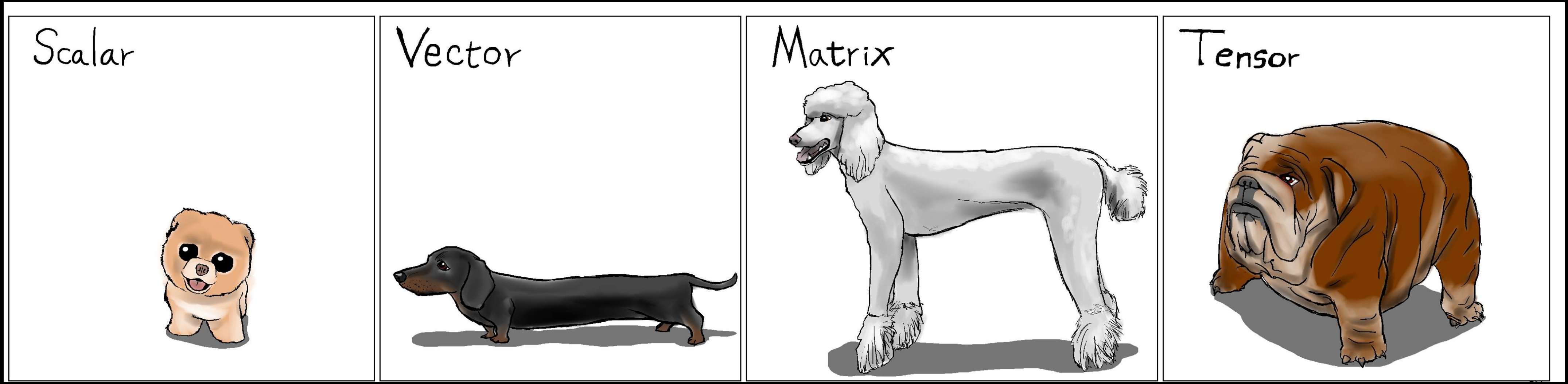
- Tensors and graphs
- Symbolic rules
- Physical laws
- Simulations

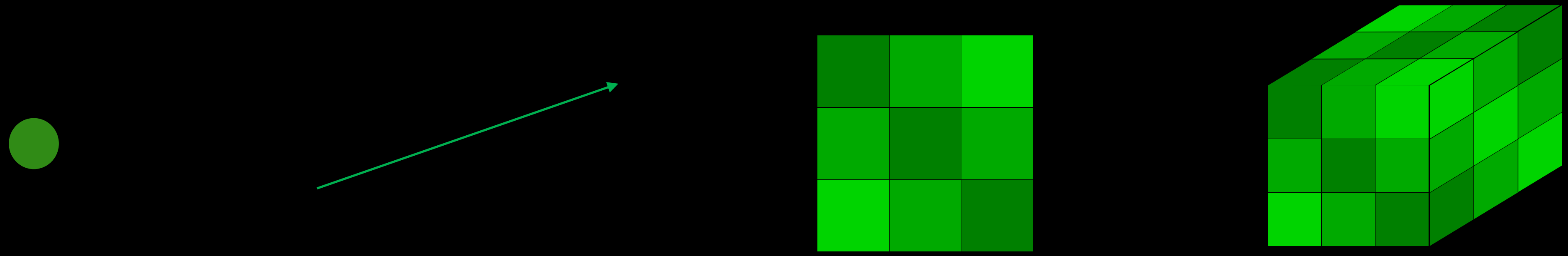






TENSOR : EXTENSION OF MATRIX





TENSORS FOR DATA
ENCODE MULTI-DIMENSIONALITY

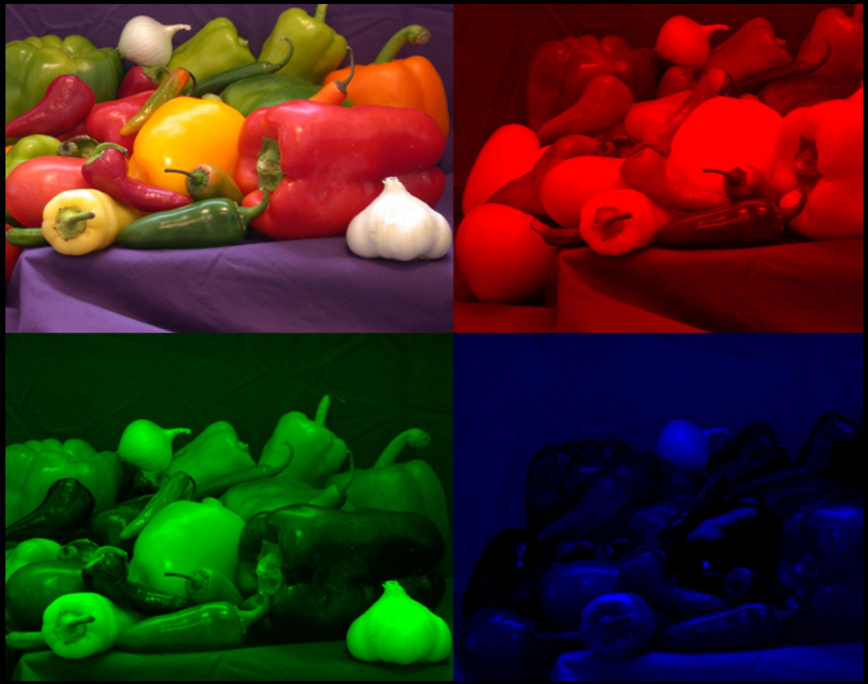


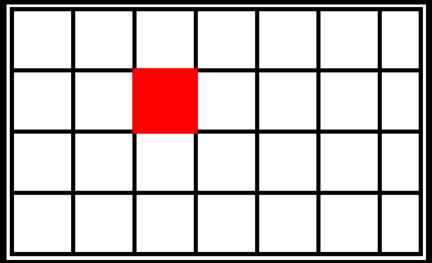
Image: 3 dimensions
Width * Height * Channels

Video: 4 dimensions
Width * Height * Channels * Time

TENSORS FOR ML ALGORITHMS
ENCODE HIGHER ORDER MOMENTS

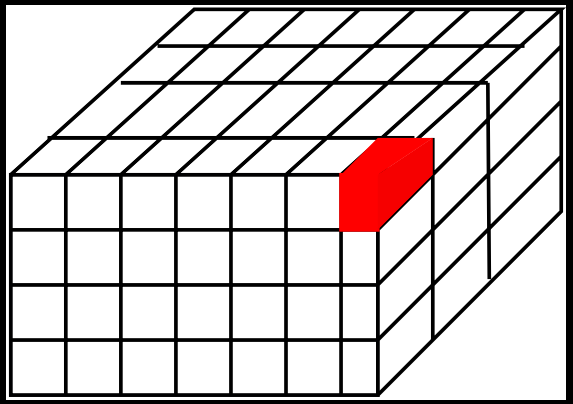
Pairwise correlations

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



Third order correlations

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



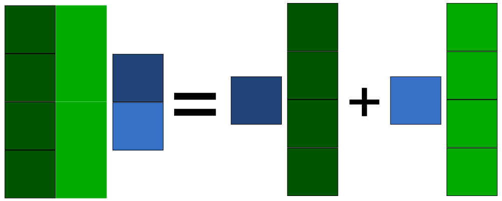
TENSORS FOR COMPUTE

TENSOR CONTRACTION PRIMITIVE

Extends the notion of matrix product

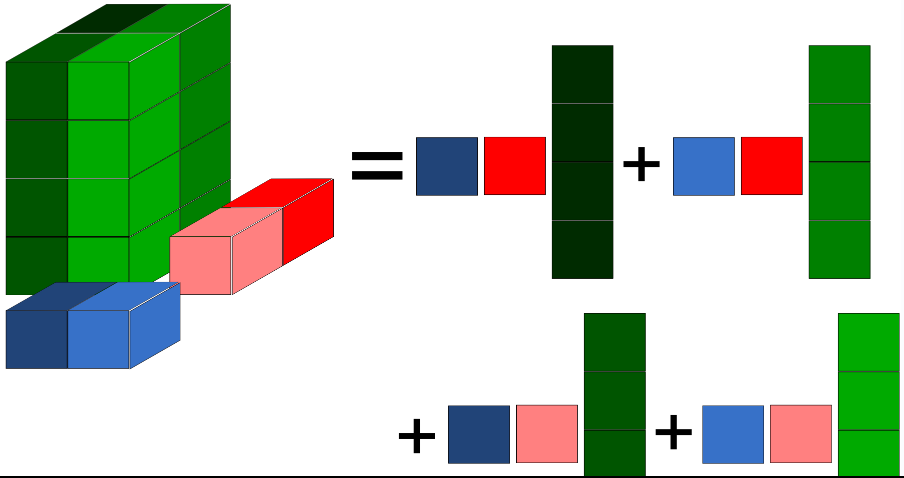
Matrix product

$$Mv = \sum_j v_j M_j$$

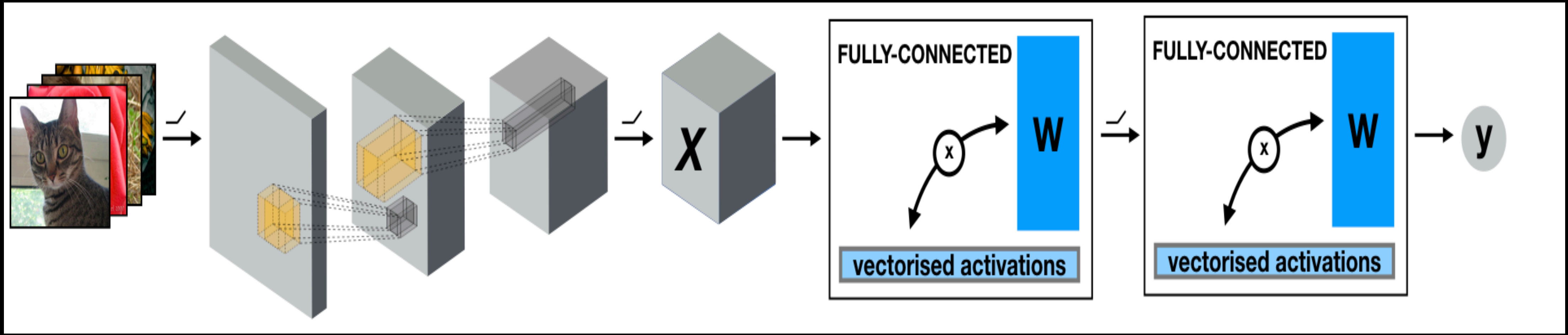


Tensor Contraction

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

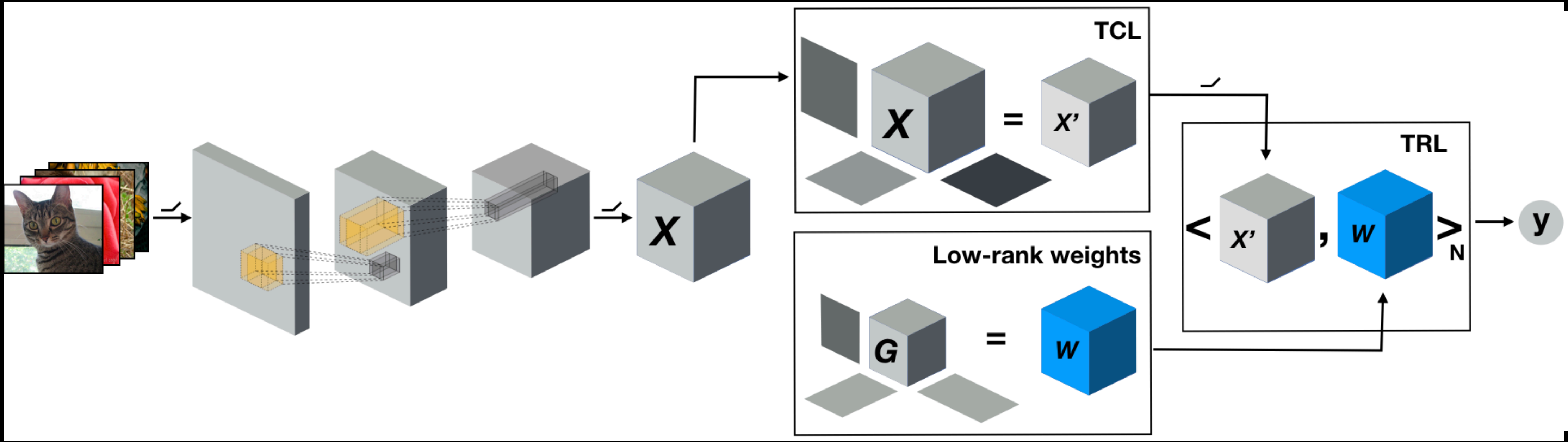


TENSORS FOR MODELS
STANDARD CNN USE LINEAR ALGEBRA



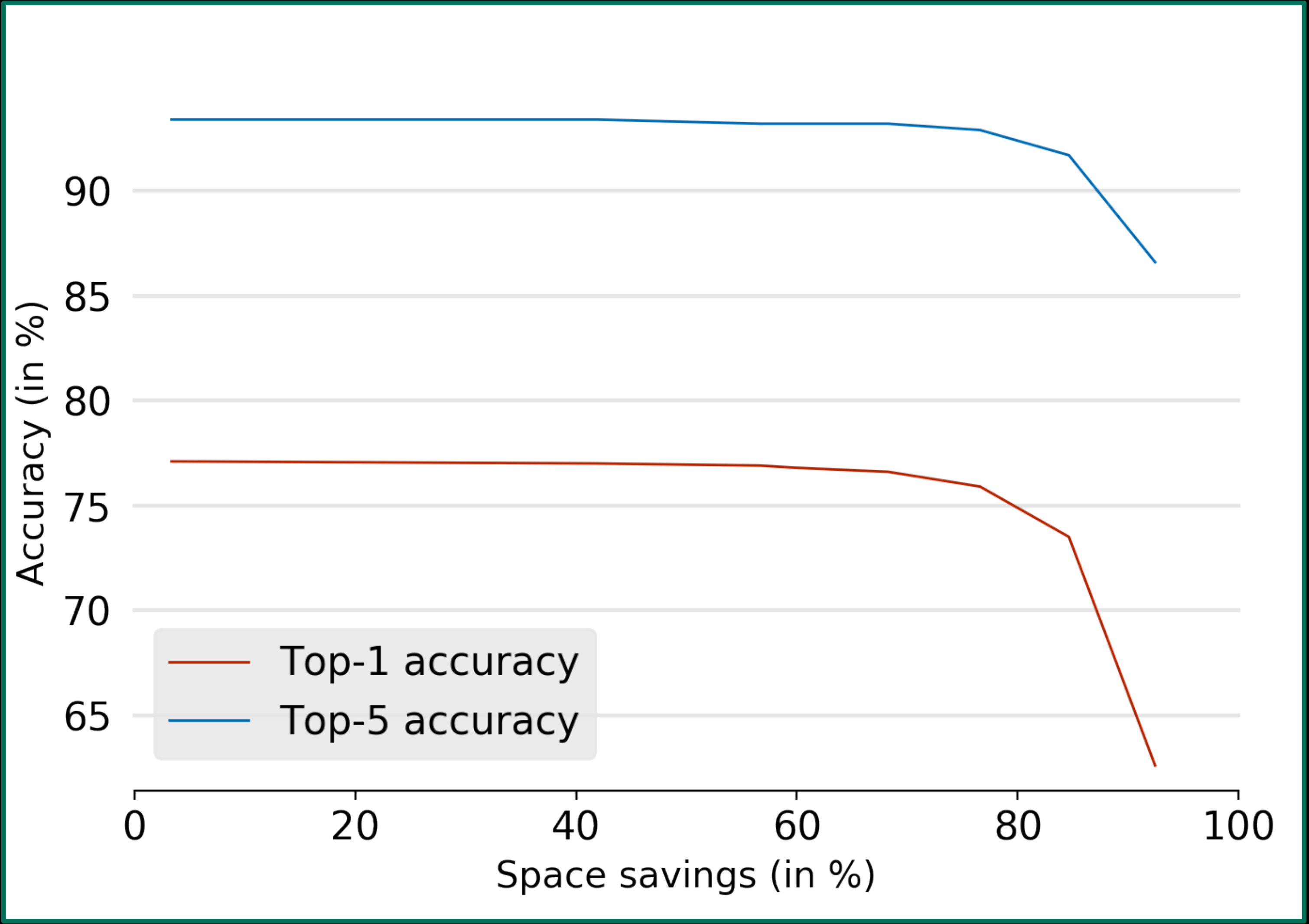
TENSORS FOR MODELS

TENSORIZED NEURAL NETWORKS



Jupyters notebook: https://github.com/JeanKossaifi/tensorly-notebooks

SPACE SAVING IN DEEP TENSORIZED NETWORKS



Jean Kossaifi

Zachary Lipton

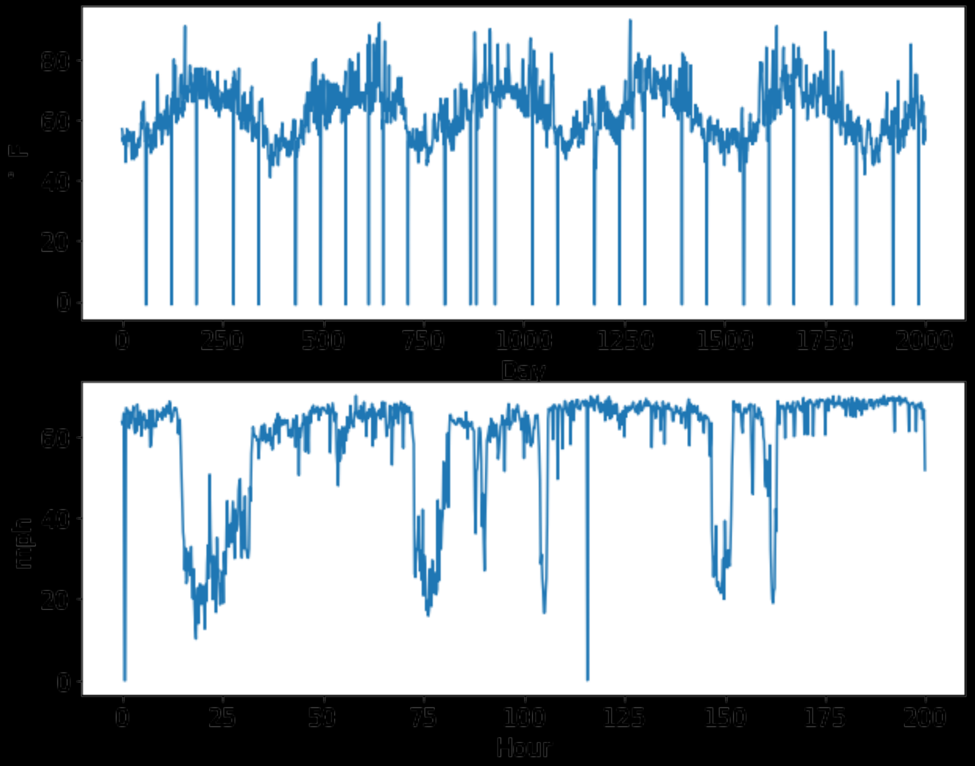
Aran Khanna

Tommaso Furlanello

TENSORS FOR LONG-TERM FORECASTING

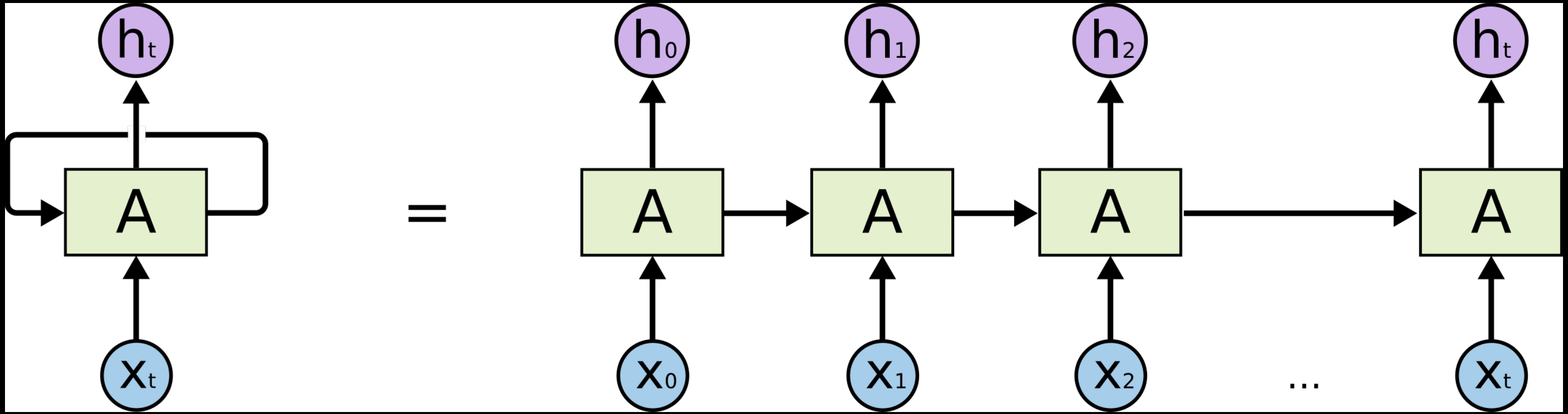
Difficulties in long term forecasting:

- Long-term dependencies
- High-order correlations
- Error propagation



RNN: FIRST-ORDER MARKOV MODELS

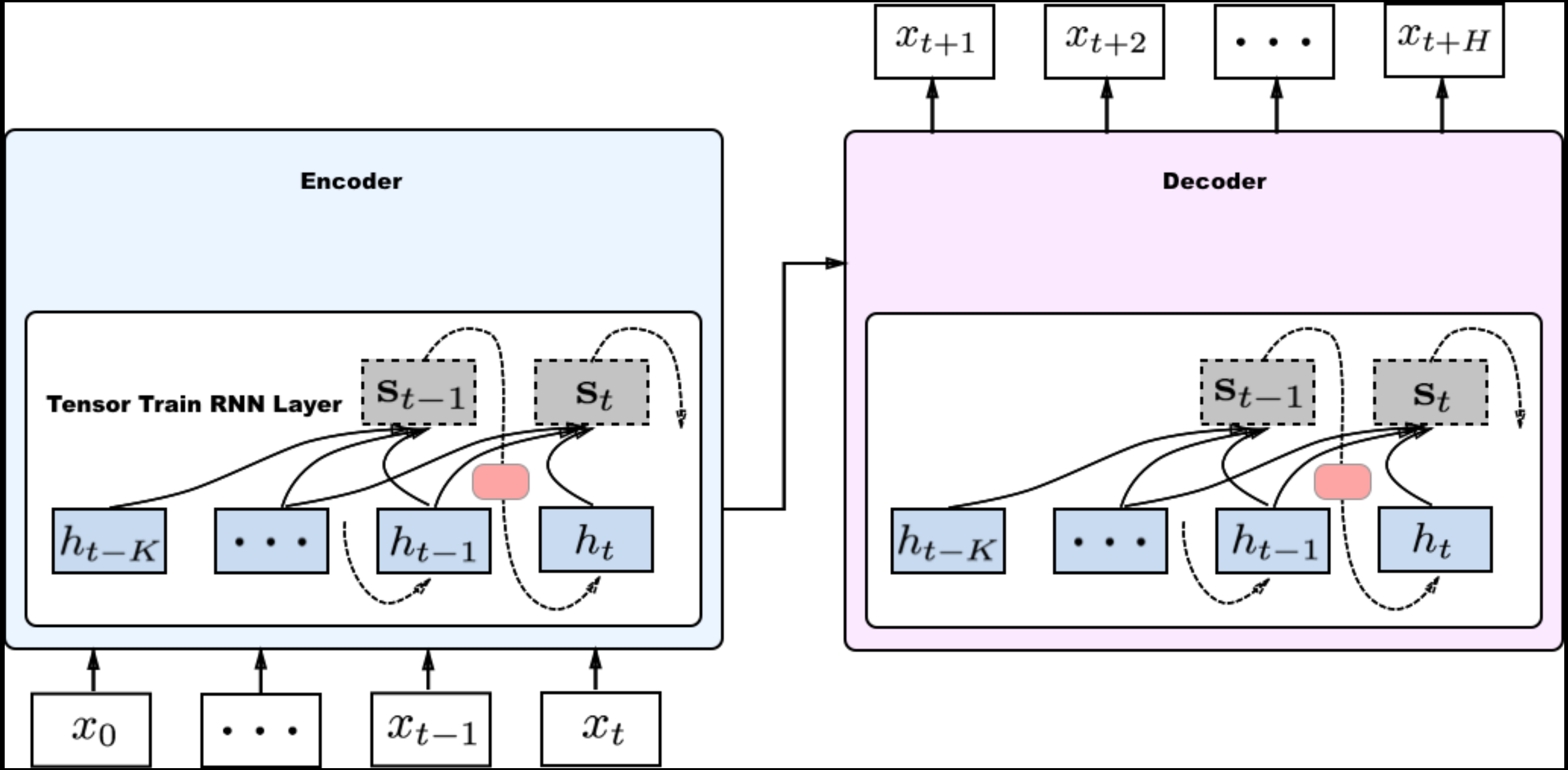
Input state x_t, hidden state h_t, output y_t,



TENSOR-TRAIN RNNS AND LSTMS

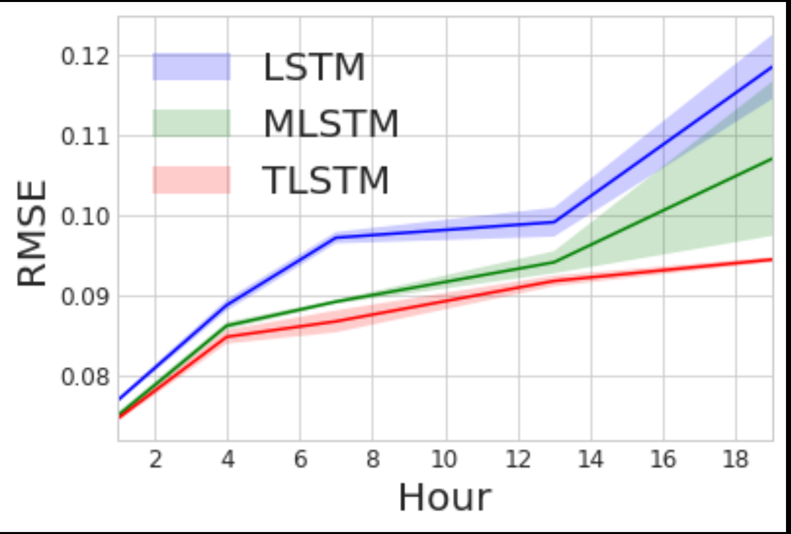
Seq2seq architecture

TT-LSTM cells

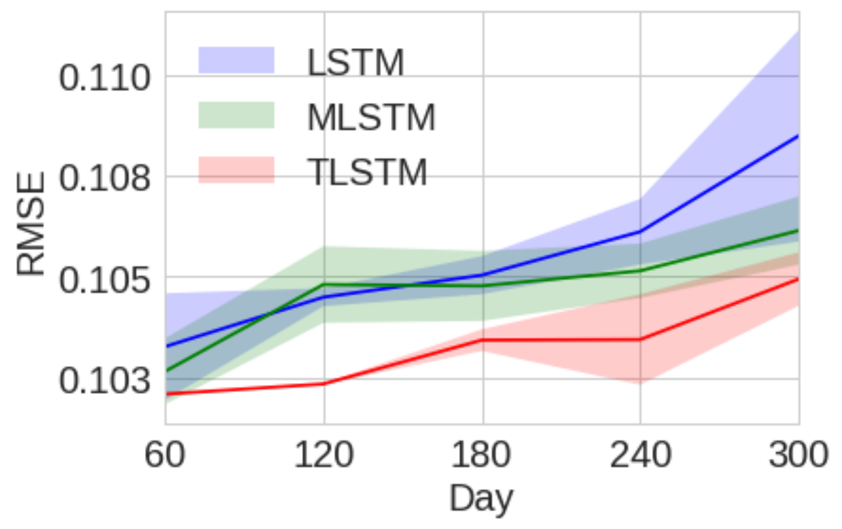


TENSOR LSTM FOR LONG-TERM FORECASTING

Traffic dataset



Climate dataset



Rose Yu

Stephan Zhang

Yisong Yue

LONG-TERM VIDEO PREDICTION WITH CONVOLUTIONAL TENSOR-TRAIN LSTM

Jiahao Su, Wonmin Byeon, Furong Huang, Jan Kautz, Anima Anandkumar





VIDEO PREDICTION

- Input: a sequence of frames

$$X = (X_1, X_2, \cdots, X_T)$$

- Output: a sequence of future frames

$$\hat{Y} = (\hat{X}_{T+1}, \hat{X}_{T+2}, \cdots, \hat{X}_{T+T'})$$

- Goal: The predicted future frames are close to their ground-truths

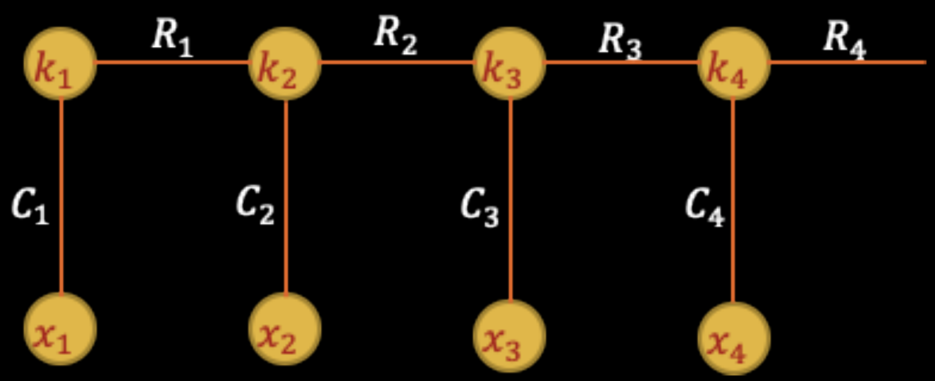
$$Y = (X_{T+1}, X_{T+2}, \cdots, X_{T+T'})$$

Example videos on Moving MNIST-2

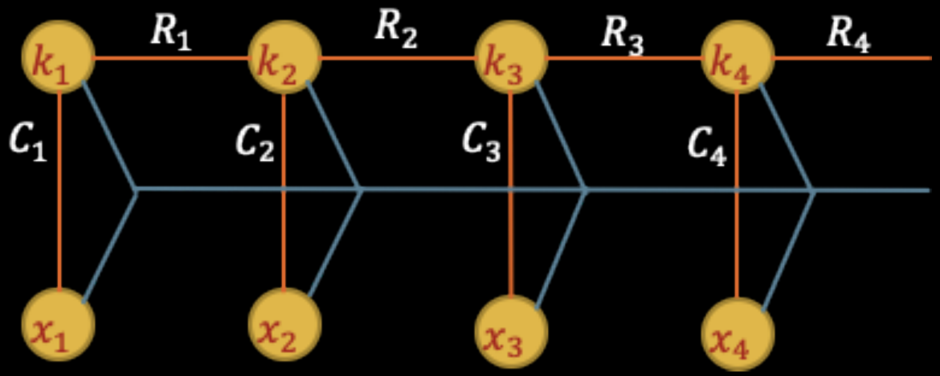
http://www.cs.toronto.edu/~nitish/unsupervised_video/

TENSOR-TRAIN VS CONVOLUTIONAL TENSOR-TRAIN

Standard Tensor-Train



Convolutional Tensor-Train (Conv-TT)



Blue edges: 2d-convolutional operations

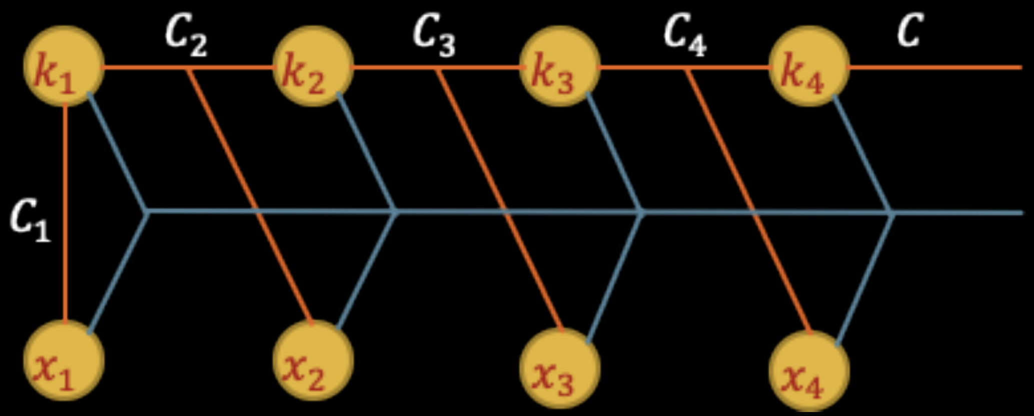
- Inputs {x_i}
  - Standard Tensor-Train: 1D-sequence
  - Conv-Tensor-Train (Conv-TT): 3D-sequence

- **Ranks {R_i}**: dimension of the kernels {k_i}
- **Order**: number of time steps (o=1, 2… t), t < T

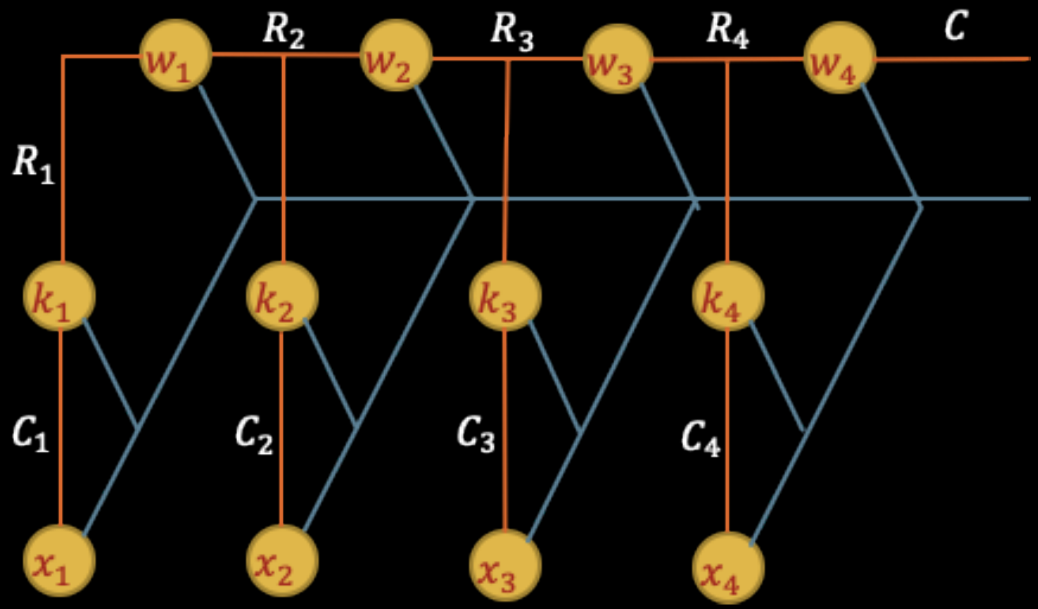
- **Conv-TT is expensive:** {k_i} are **5th-order kernels**

VARIANTS OF CONVOLUTIONAL TENSOR-TRAIN

Conv-TT Version 1



Conv-TT Version 2



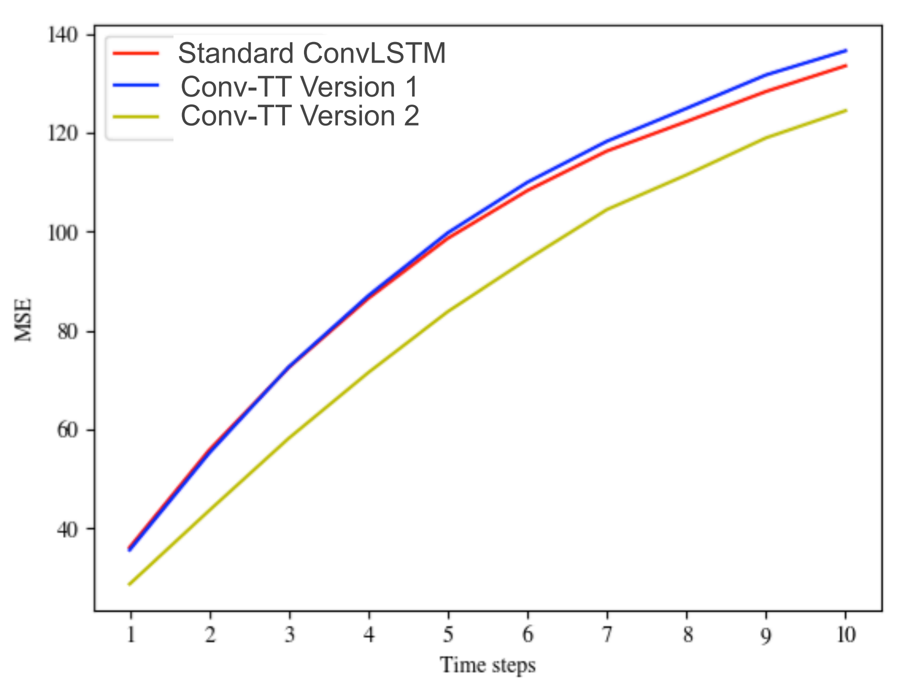
- Here, {k_i} are **4th-order kernels**
- Version 1
  - The numbers of input channels {C_i} = Tensor rank {R_i}
  - No low-rankness
- Version 2: Different tensor ranks {R_i} are allowed by {w_i}

PREDICTION RESULTS

- **Dataset:** Moving MNIST-2
- **Input:** 10 frames
- **Output:** 10 predictions
- **Base Architecture:** 12 Conv-LSTM Layers [Byeon, et al 2018]

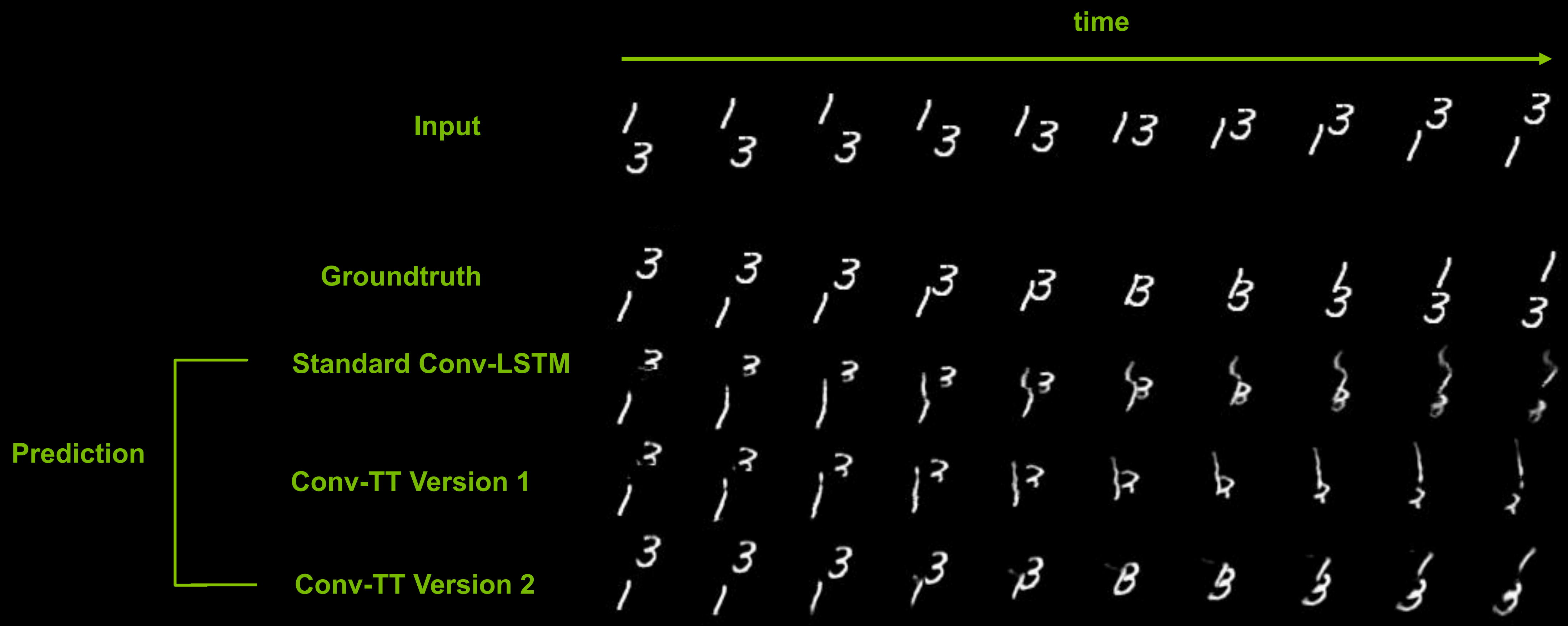
Models	MSE	SSIM
Standard ConvLSTM	96.7	0.831
Conv-TT Version 1 (order 3)	97.1	0.832
Conv-TT Version 2 (order 3, rank 4)	**83.97**	**0.853**

Per-frame MSE on test set



NVIDIA

PREDICTION RESULTS



UNSUPERVISED LEARNING TOPIC MODELS THROUGH TENSORS

Topics

Justice

Education

Sports

SECTIONS HOME SEARCH The New York Times

COLLEGE FOOTBALL

At Florida State, Football Clouds Justice

By MIKE McINTIRE and WALT BOGDANICH OCT. 10, 2014

Now, an examination by The New York Times of police and court records, along with interviews with crime witnesses, has found that, far from an aberration, the treatment of the Winston complaint was in keeping with the way the police on numerous occasions have soft-pedaled allegations of wrongdoing by Seminoles football players. From criminal mischief and motor-vehicle theft to domestic violence, arrests have been avoided, investigations have stalled and players have escaped serious consequences.

In a community whose self-image and economic well-being are so tightly bound to the fortunes of the nation's top-ranked college football team, law enforcement officers are finely attuned to a suspect's football connections. Those ties are cited repeatedly in police reports examined by The Times. What's more, dozens of officers work second jobs directing traffic and providing security at home football games, and many express their devotion to the Seminoles on social media.

TMZ, the gossip website, also requested the police report and later asked the school's deputy police chief, Jim L. Russell, if the campus police had interviewed Mr. Winston about the rape report. Mr. Russell responded by saying his officers were not investigating the case, omitting any reference to the city police, even though the campus police knew of their involvement. "Thank you for contacting me regarding this rumor — I am glad I can dispel that one!" Mr. Russell told TMZ in an email. The university said Mr. Russell was unaware of any other police investigation at the time of the inquiry. Soon after, the Tallahassee police belatedly sent their files to the news media and to the prosecutor, William N. Meggs. By then critical evidence had been lost and Mr. Meggs, who criticized the police's handling of the case, declined to

lson after the Seminoles' first game; five

am's second-leading receiver.

On Jan. 10, 2013, a female student at Florida State spotted the man she believed had raped her the previous month. After learning his name, Jameis Winston, she reported him to the Tallahassee police.

In the 21 months since, Florida State officials have said little about how they handled the case, which is no

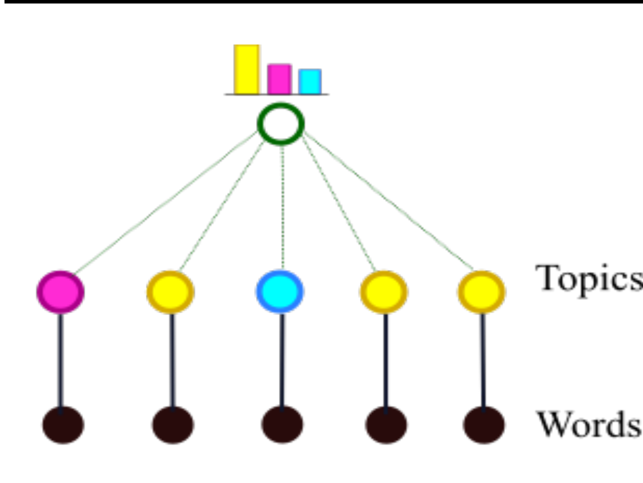
investigated by the federal Depart

Most recently, university officials suspended Mr. Winston for one game after he stood in a public place on campus and, playing off a running Internet gag, shouted a crude reference to a sex act. In a news conference afterward, his coach, Jimbo Fisher, said, "Our hope and belief is Jameis will learn from this and use better judgment and language and decision-making."

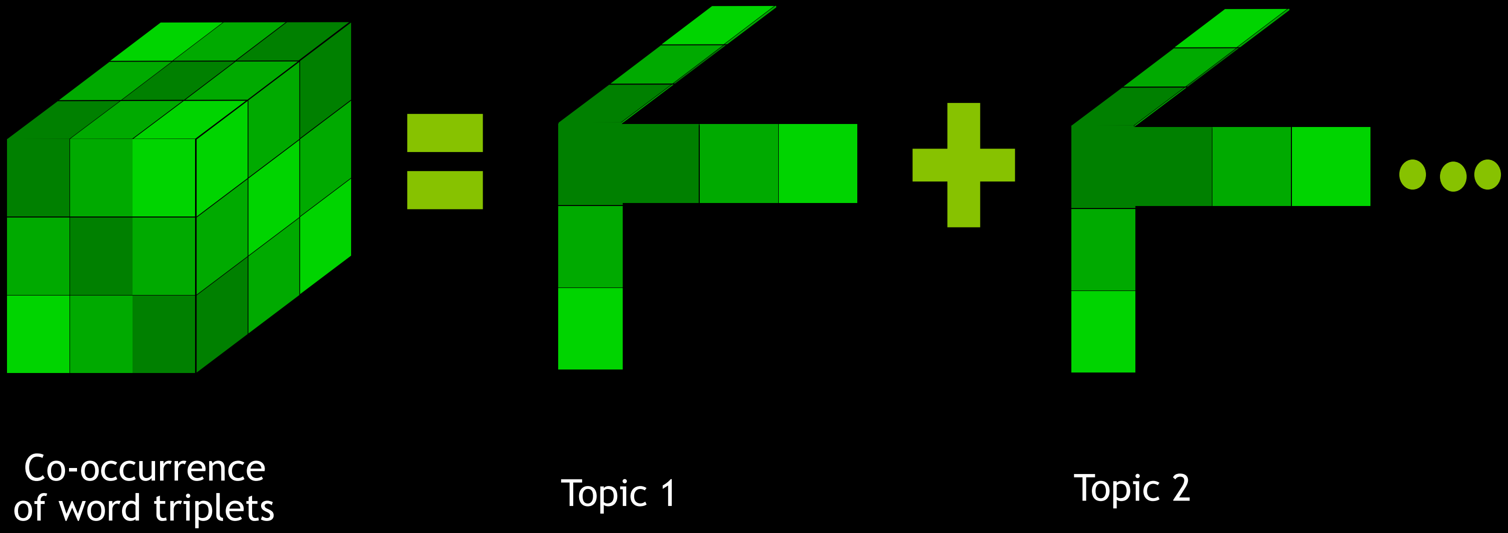
As The Times reported last April, the Tallahassee police also failed to aggressively investigate the rape accusation. It did not become public until November, when a Tampa reporter, Matt Baker, acting on a tip, sought records of the police investigation.

Upon learning of Mr. Baker's inquiry, Florida State, having shown little curiosity about the rape accusation, suddenly took a keen interest in the journalist seeking to report it, according to emails obtained by The Times.

"Can you share any details on the requesting source?" David Perry, the university's police chief, asked the Tallahassee police. Several hours later, Mr.



TENSORS FOR MODELING:
TOPIC DETECTION IN TEXT



TENSORLY: HIGH-LEVEL API FOR TENSOR ALGEBRA

Tensor decomposition | Tensor regression | Tensors + Deep

Basic tensor operations

Unified backend

NumPy | SciPy | mxnet | PYTORCH | TensorFlow

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

Jean Kossaifi

TENSORLY WITH PYTORCH BACKEND

```
import tensorly as tl
from tensorly.random import tucker_tensor

tl.set_backend('pytorch')
core, factors = tucker_tensor((5, 5, 5),
                              rank=(3, 3, 3))
core = Variable(core, requires_grad=True)
factors = [Variable(f, requires_grad=True) for f in factors]

optimiser = torch.optim.Adam([core]+factors, lr=lr)

for i in range(1, n_iter):
    optimiser.zero_grad()
    rec = tucker_to_tensor(core, factors)
    loss = (rec - tensor).pow(2).sum()
    for f in factors:
        loss = loss + 0.01*f.pow(2).sum()

    loss.backward()
    optimiser.step()
```

Set Pytorch backend

Tucker Tensor form

Attach gradients

Set optimizer

LACK OF LABELED DATA IN MANY DOMAINS

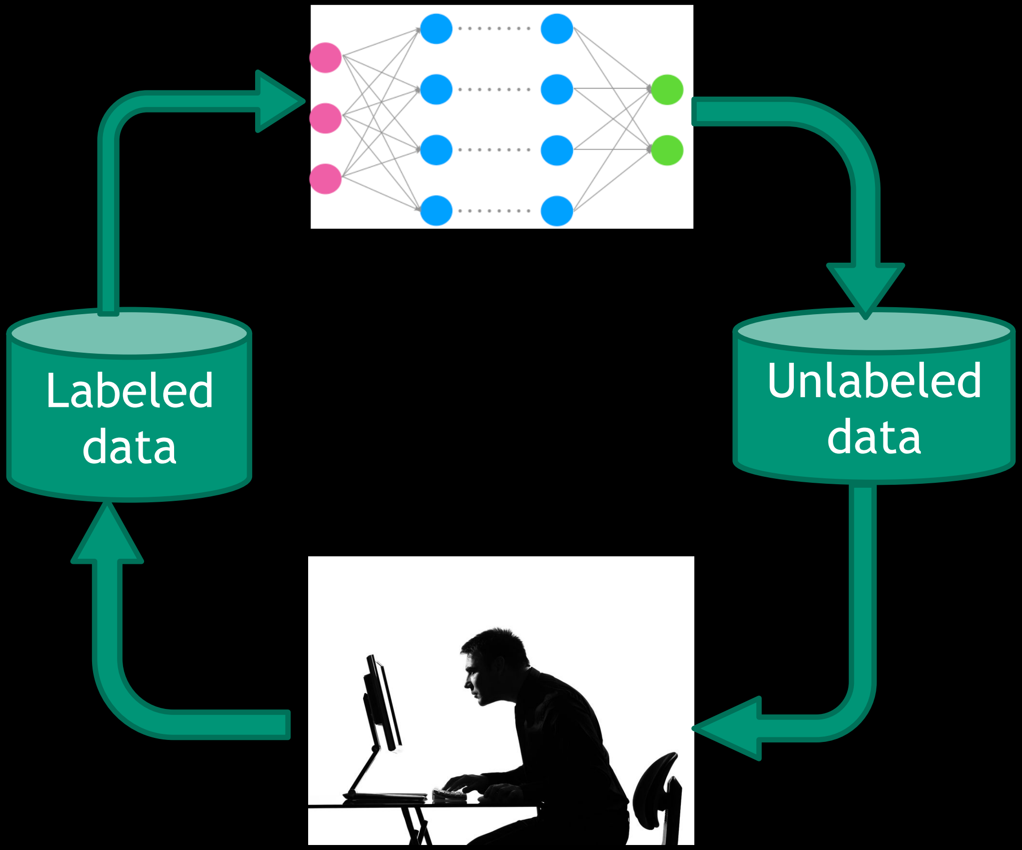
Strategies to cope with it

- Semi-supervised learning
- Active learning
- Crowdsourcing
- Domain adaptation/transfer learning
- Domain knowledge and structure

ACTIVE LEARNING

Can it work at scale with deep learning?

- Retraining from scratch not feasible: incremental training
- Minibatch size: balancing latency of labeling and training
- Acquisition function?



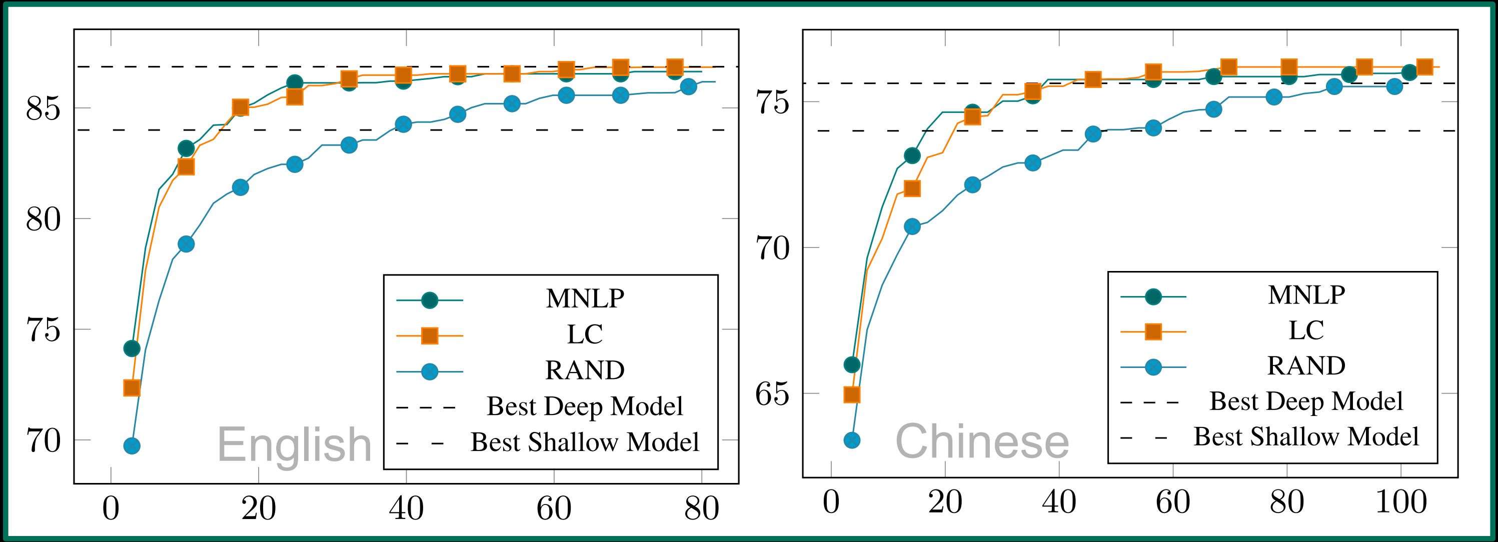
RESULTS

NER task on largest open benchmark (Onto-notes)

Acquisition functions for uncertainty sampling:

- Least confidence (LC)
- Max. normalized log probability (MNLP)

Test F1 score vs. % of labeled words



- Deep active learning matches :
  - SOTA with just 25% data on English, 30% on Chinese.
  - Best shallow model (on full data) with 12% data on English, 17% on Chinese.

TAKE-AWAY

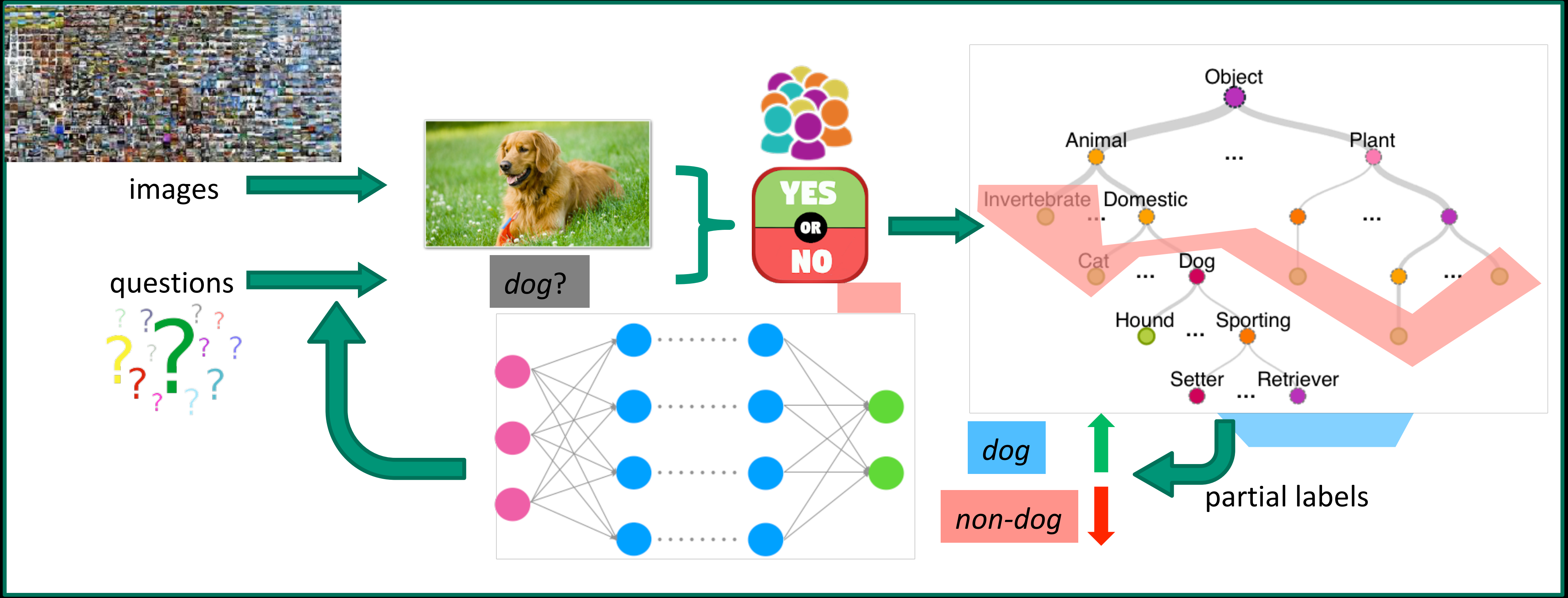
- Uncertainty sampling mostly works. Normalizing for length helps under low data. (Bayesian uncertainty more robust in subsequent work Siddhant & Lipton)
- With active learning, **deep beats shallow** even in low data regime.
- With active learning, **SOTA achieved** with far fewer samples.

Yanyao Shen

Hyokun Yun

Zachary Lipton

ACTIVE LEARNING WITH PARTIAL FEEDBACK

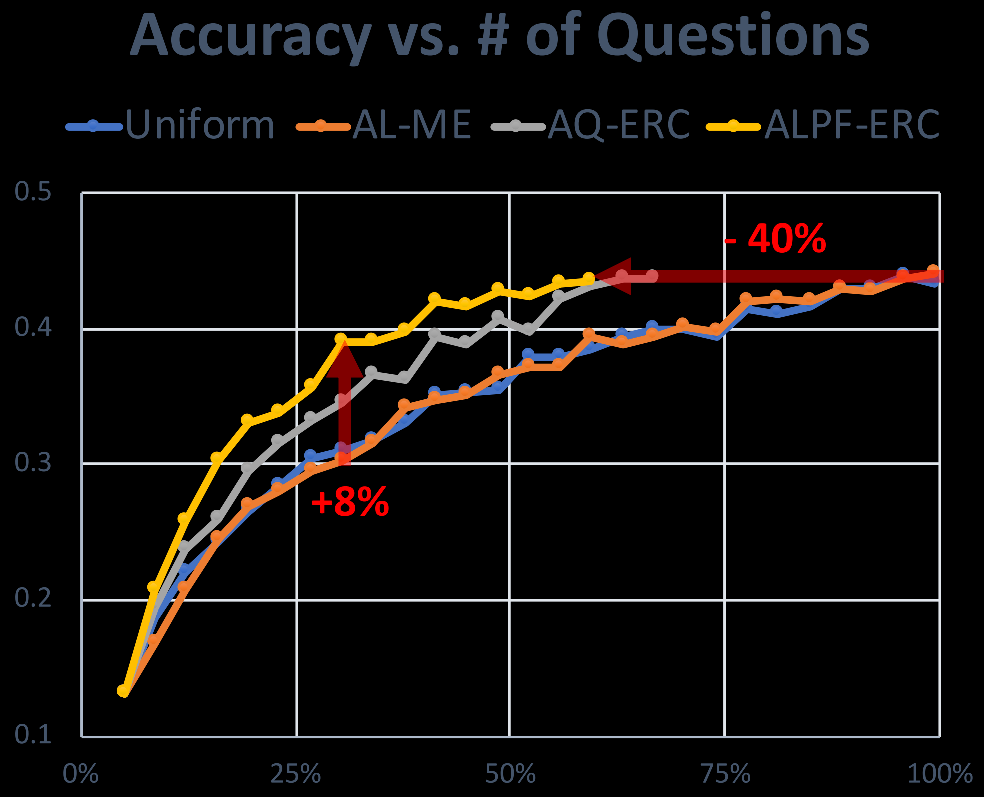


- Hierarchical class labeling: Labor proportional to # of binary questions asked
- **Actively pick informative questions ?**

RESULTS ON TINY IMAGENET (100K SAMPLES)

ALPF-ERC
active *data*
active *questions*

AQ-ERC
inactive *data*
active *questions*



Uniform
inactive *data*
inactive *questions*

AL-ME
active *data*
inactive *questions*

- Yield **8%** higher accuracy at **30%** questions (w.r.t. Uniform)
- Obtain full annotation with **40%** less binary questions

TAKE-AWAYS

Peiyun Hu

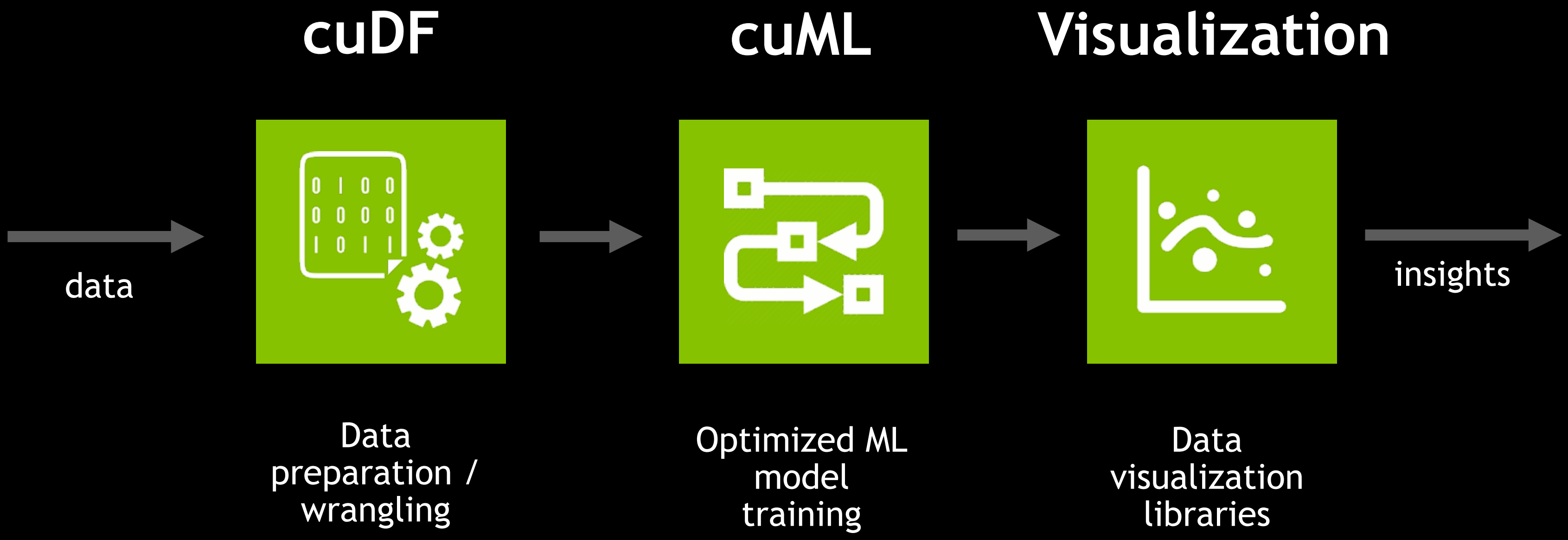
- Hierarchical structure in labels is very helpful
- Don't annotate from scratch
    - Select questions actively based on the learned model
- Don't sleep on partial labels
    - Re-train model from partial labels

Zachary Lipton

Deva Ramanan

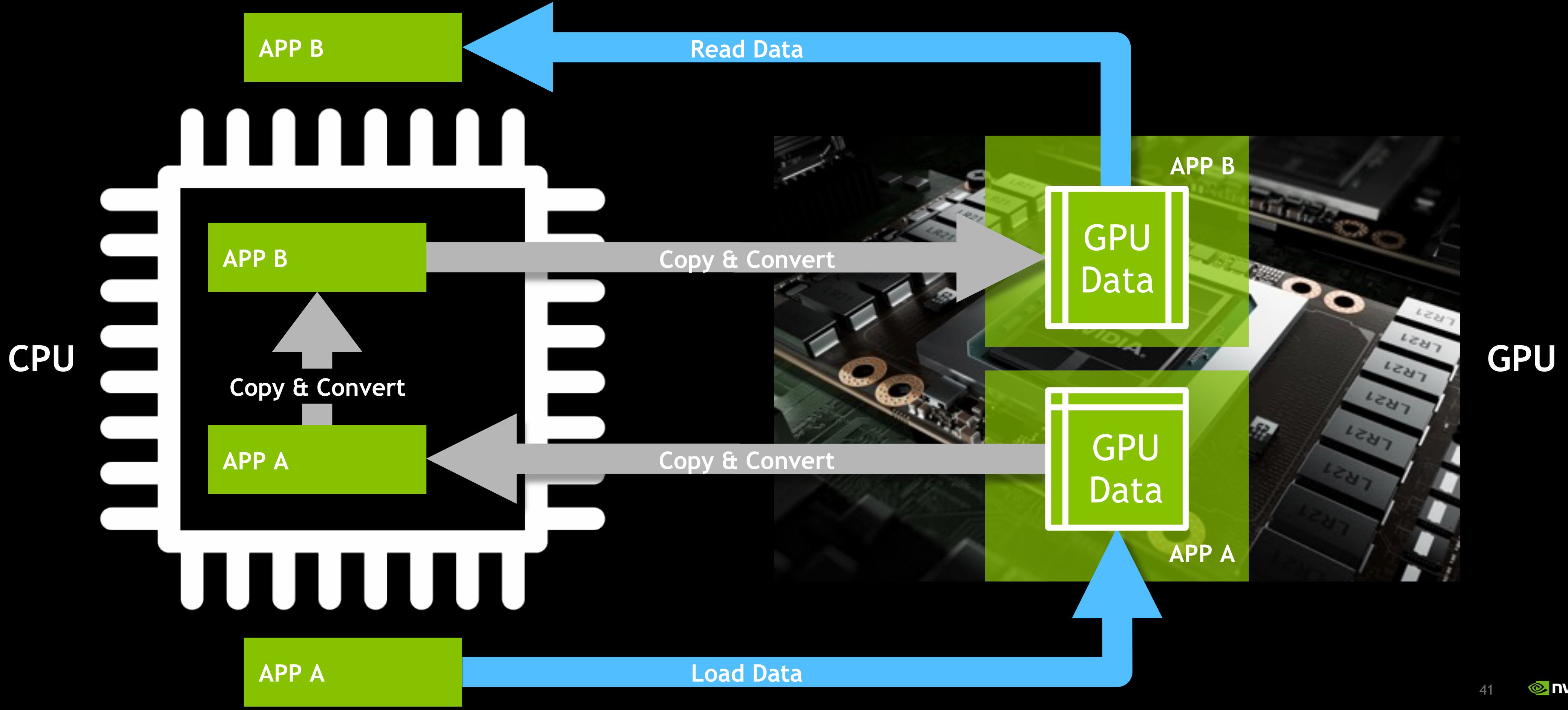
RE-IMAGINING DATA SCIENCE WORKFLOW

RAPDIS: Open Source, End-to-end GPU-accelerated Workflow



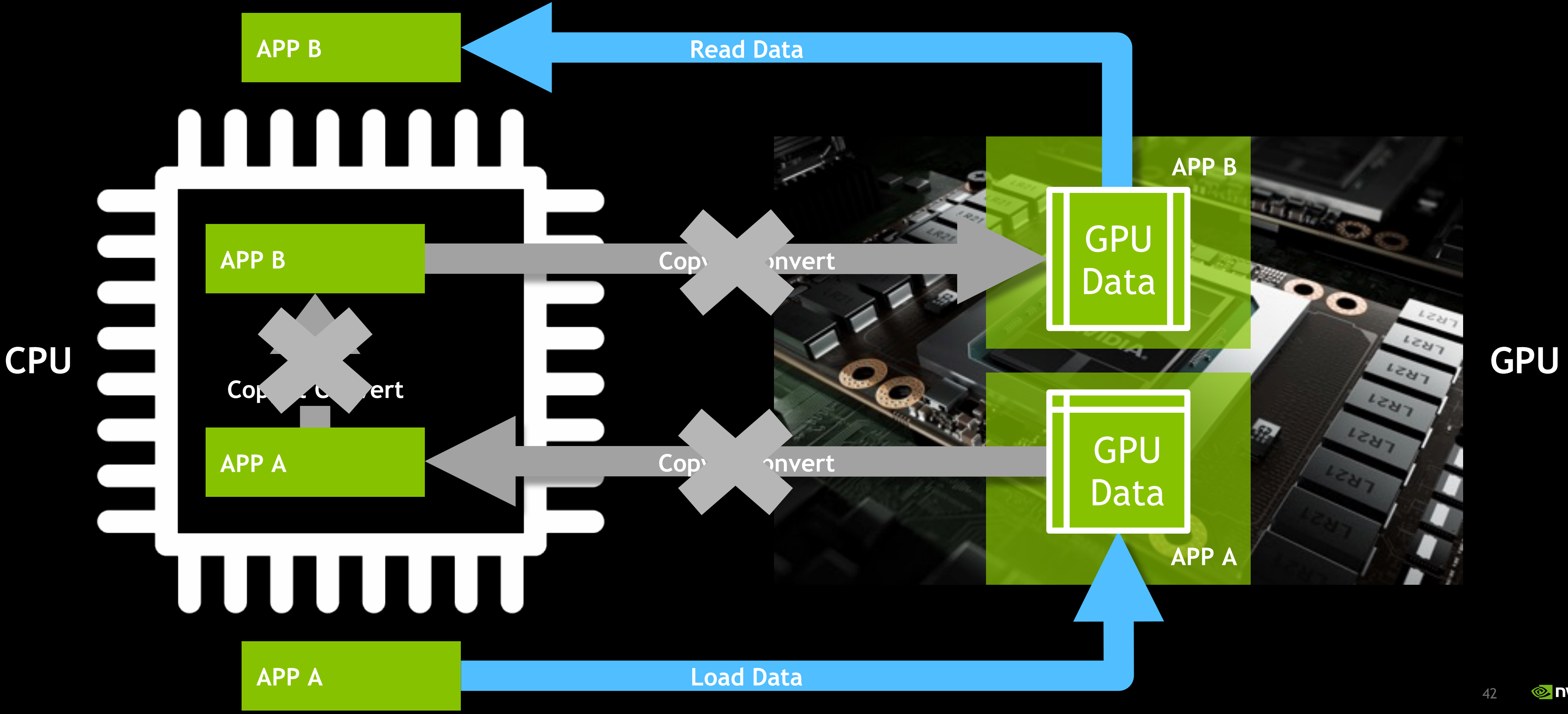
DATA MOVEMENT AND TRANSFORMATION

The bane of productivity and performance



DATA MOVEMENT AND TRANSFORMATION

What if we could keep data on the GPU?

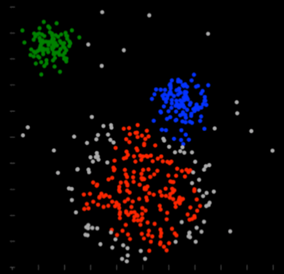


RAPID AI LIBRARIES

cuML & cuGraph

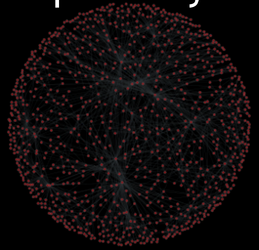
8x V100 20-90x faster than dual socket CPU

Machine Learning

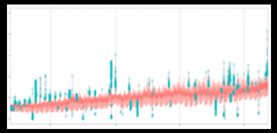


Decisions Trees
Random Forests
Linear Regressions
Logistics Regressions
K-Means
K-Nearest Neighbor
DBSCAN
Kalman Filtering
Principal Components
Single Value Decomposition

Graph Analytics



PageRank
BFS
Jaccard Similarity
Single Source Shortest Path
Triangle Counting
Louvain Modularity



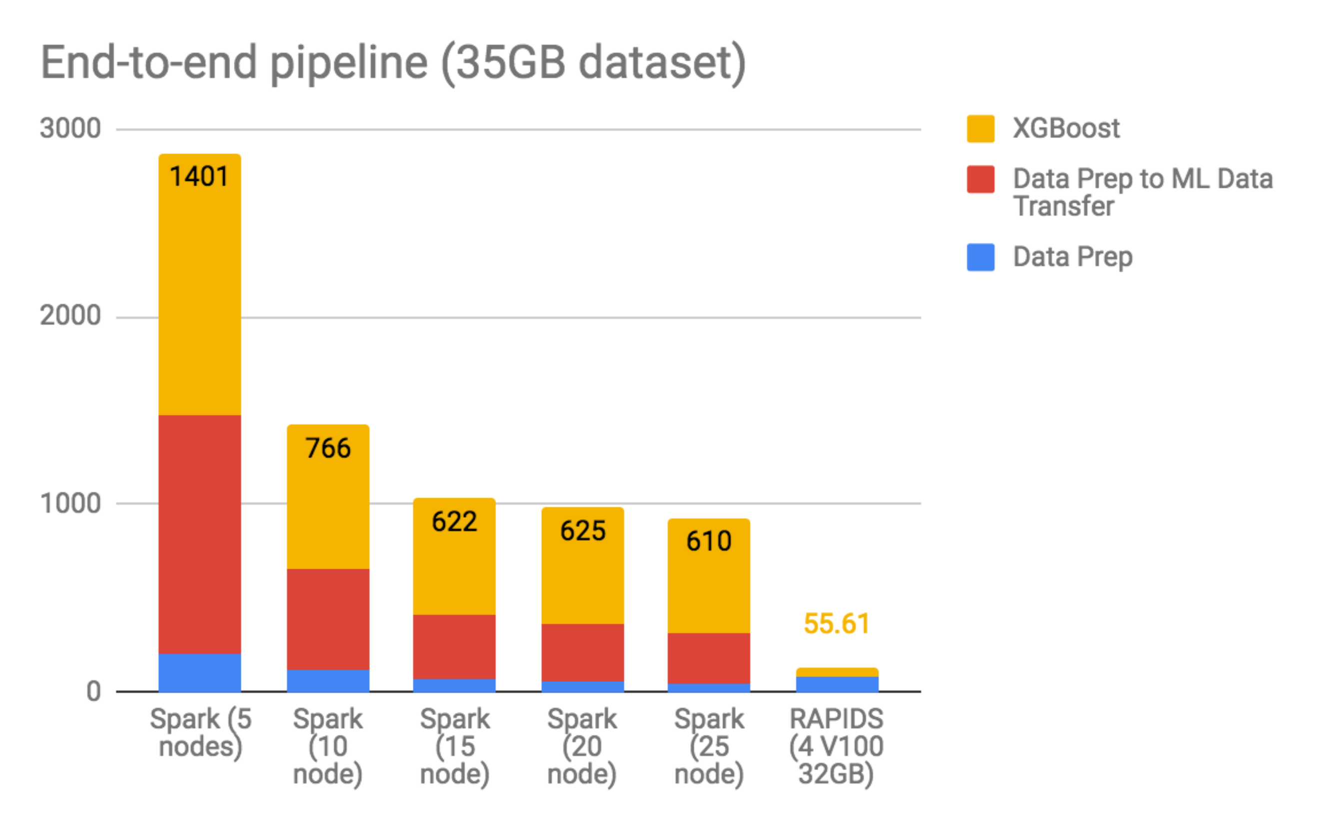
ARIMA
Holt-Winters

XGBoost,
Mortgage Dataset, 90x

3 Hours to 2 mins on DGX-1

CUDF + XGBOOST

Fully In- GPU Benchmarks



- Full end to end pipeline
- Leveraging DaskGDF
- No Data Prep time all in memory
- Arrow to Dmatrix (CSR) for XGBoost

RAPIDS

How do I get the software?

https://github.com/rapidsai

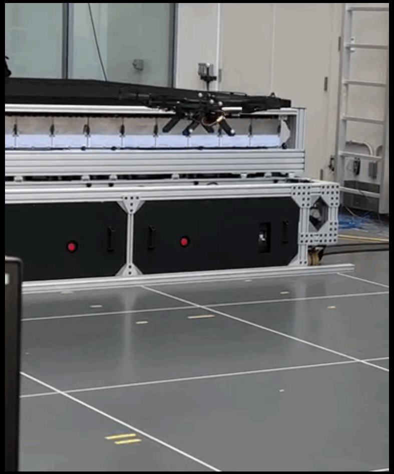
https://anaconda.org/rapidsai/

https://pypi.org/project/cudf
https://pypi.org/project/cuml

https://ngc.nvidia.com/registry/nvidia-rapidsai-rapidsai

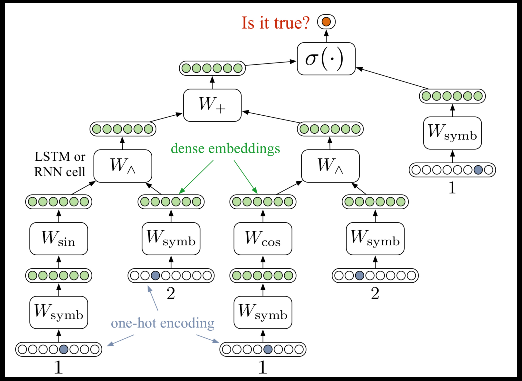
https://hub.docker.com/r/rapidsai/rapidsai/

TAKEAWAYS



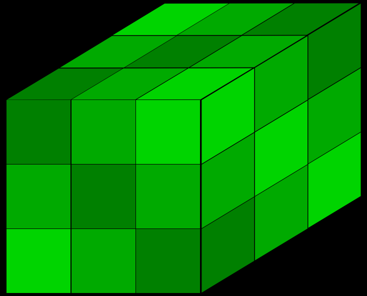
End-to-end learning from scratch is impossible in most settings

Blend DL w/ prior knowledge => improve data efficiency, generalization, model size



Outstanding challenge (application dependent):

what is right blend of prior knowledge vs data?



Thank you