Anima Anandkumar

Caltech

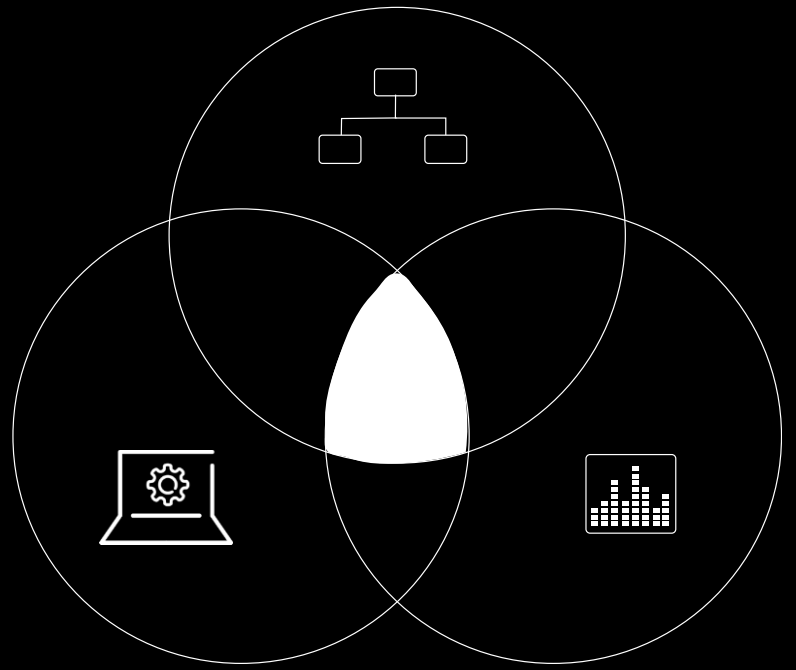
NVIDIA

ROLE OF TENSORS IN MACHINE LEARNING

TRINITY OF AI/ML

ALGORITHMS

COMPUTE

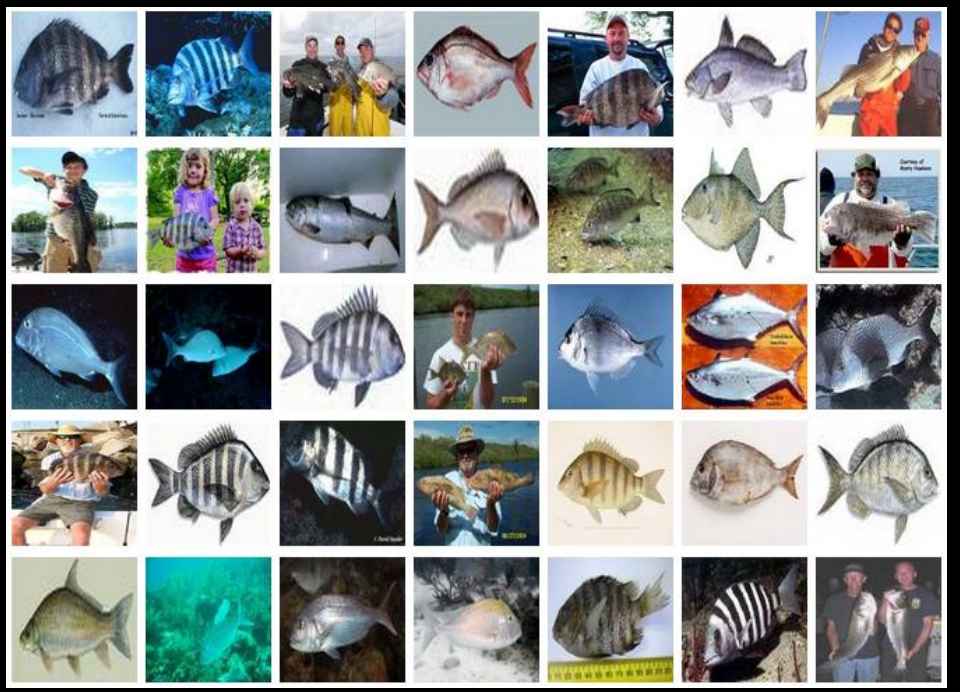


DATA

EXAMPLE AI TASK: IMAGE CLASSIFICATION

DATA: LABELED IMAGES FOR TRAINING AI

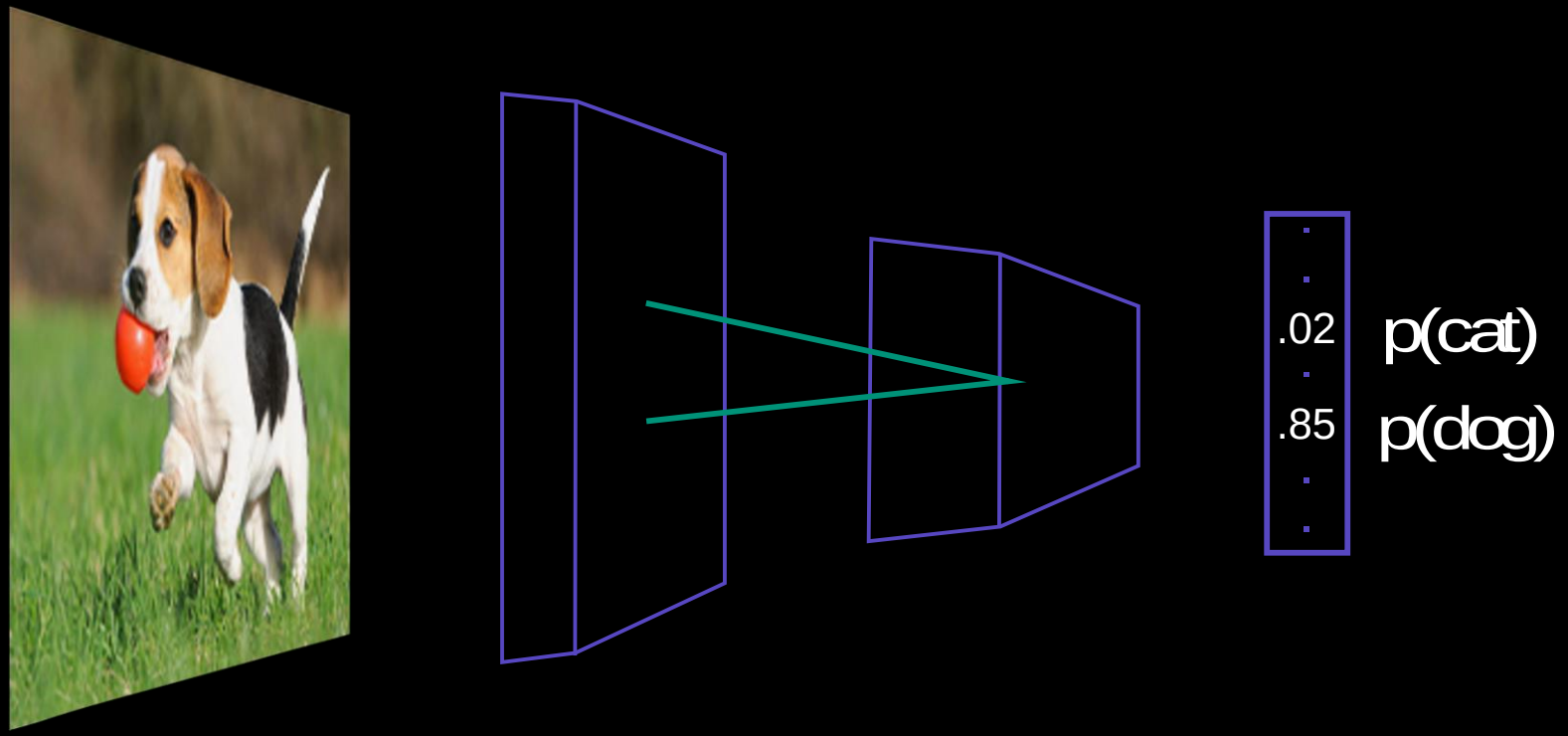
- 14 million images and 1000 categories.
- Largest database of labeled images.



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

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

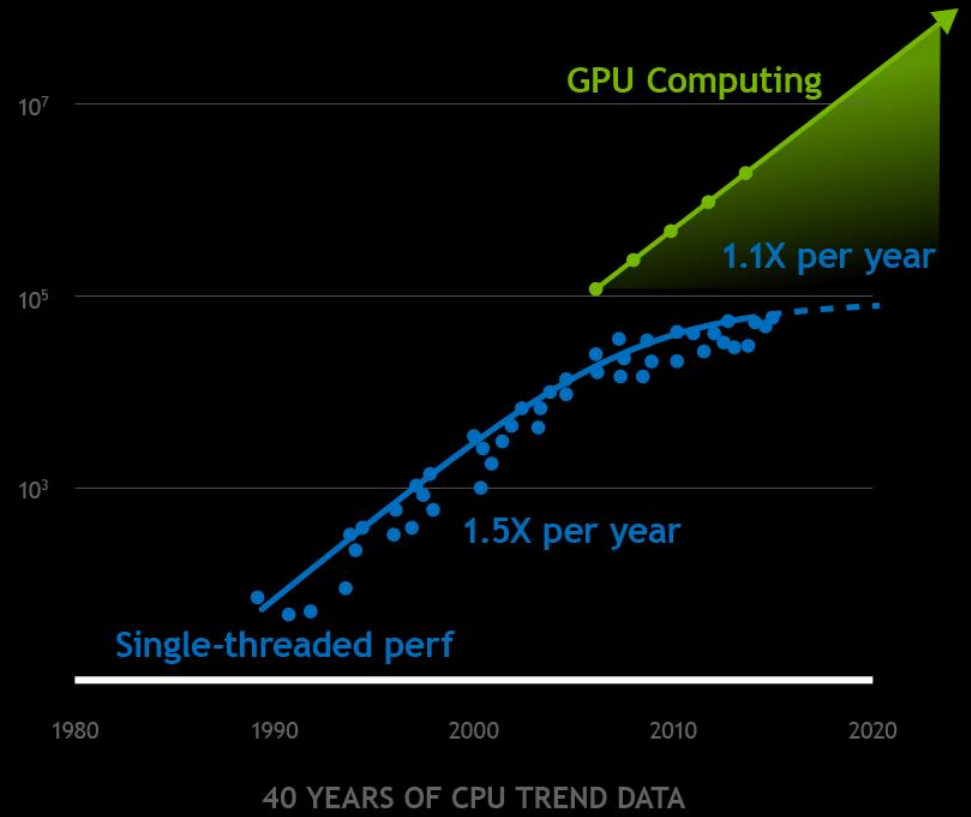
MODEL: CONVOLUTIONAL NEURAL NETWORK



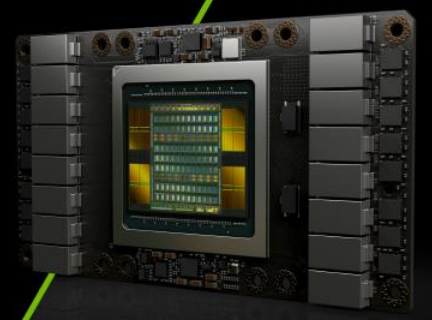
- **Deep learning:** Many layers give large capacity for model to learn from data
- **Inductive bias:** Prior knowledge about natural images.

COMPUTE INFRASTRUCTURE FOR AI: GPU

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

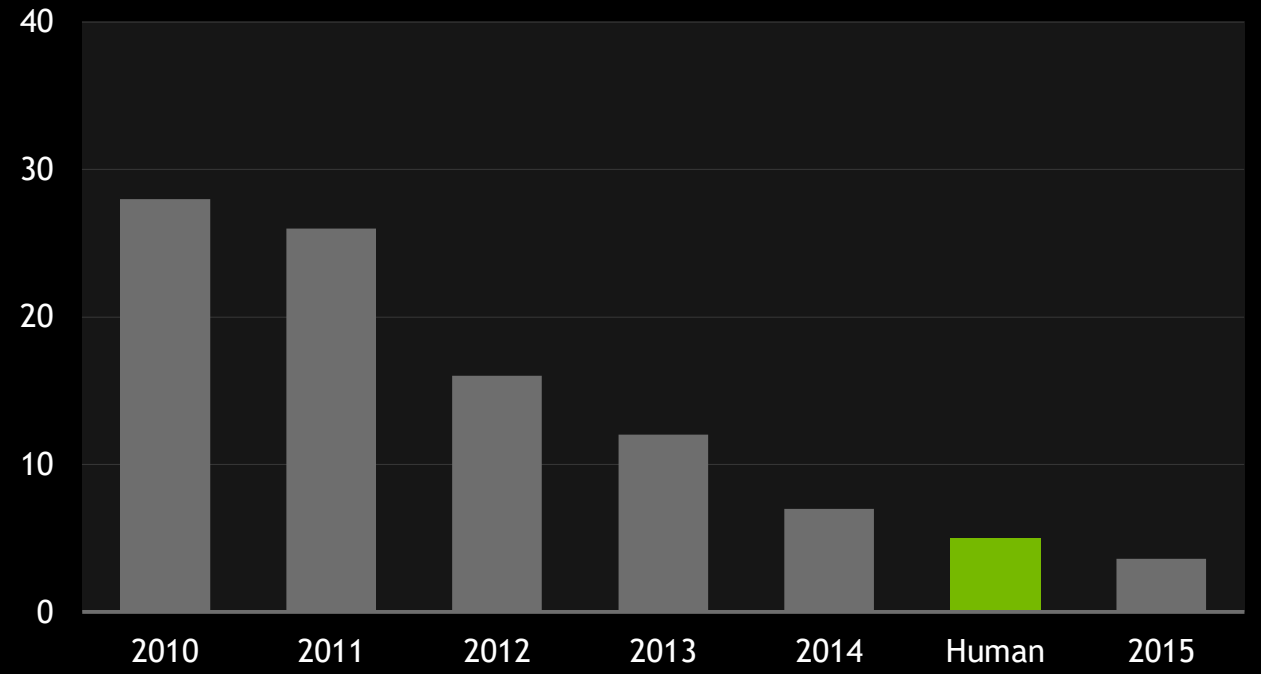


MOORE'S LAW: A SUPERCHARGED LAW



PROGRESS IN TRAINING IMAGENET

Error in making 5 guesses about the image category



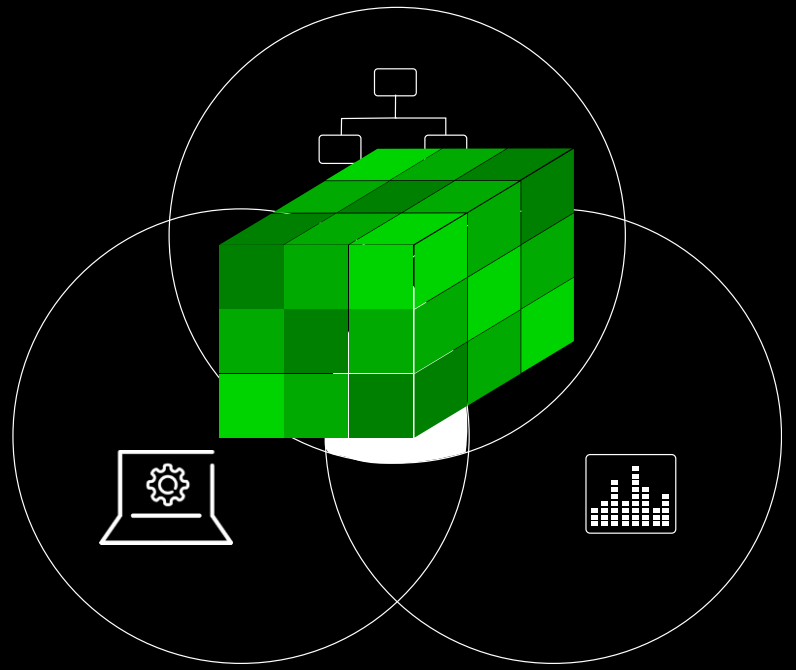
Need Trinity of AI : Data + Algorithms + Compute

Statista: Statistics Portal

TENSORS PLAY A CENTRAL ROLE

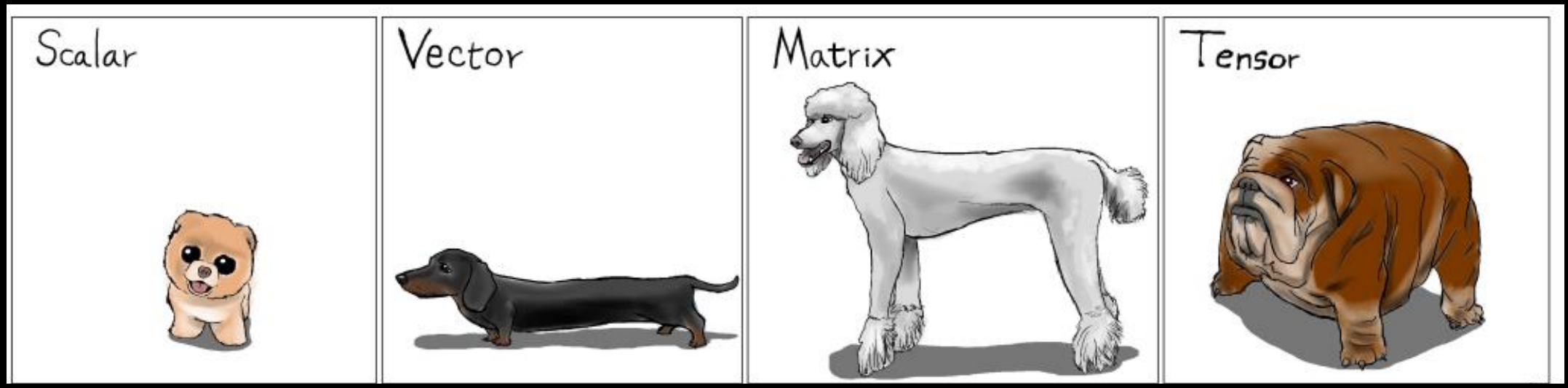
ALGORITHMS

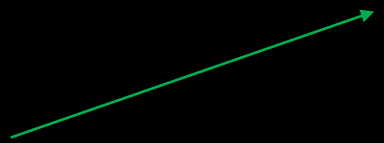
COMPUTE

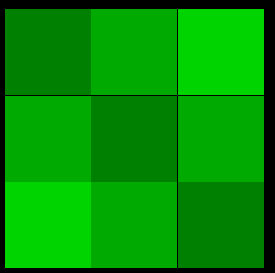


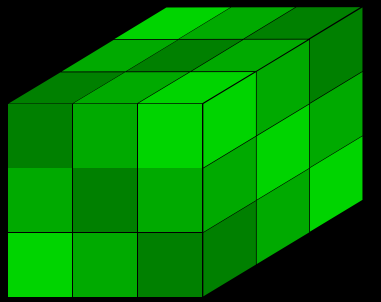
DATA

TENSOR : EXTENSION OF MATRIX









WHY TENSORS?

TENSORS FOR DATA
ENCODE MULTI-DIMENSIONALITY

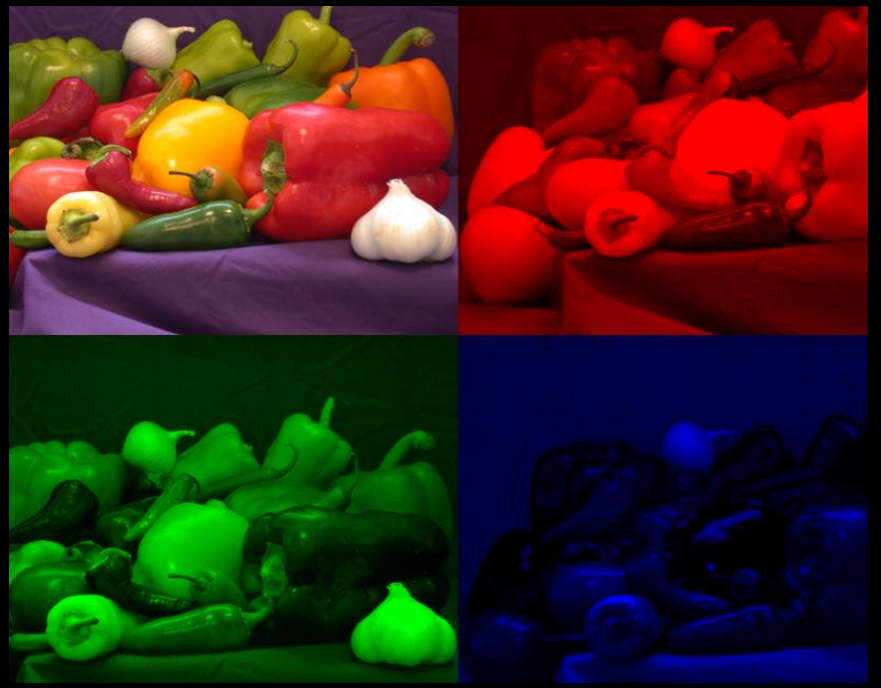


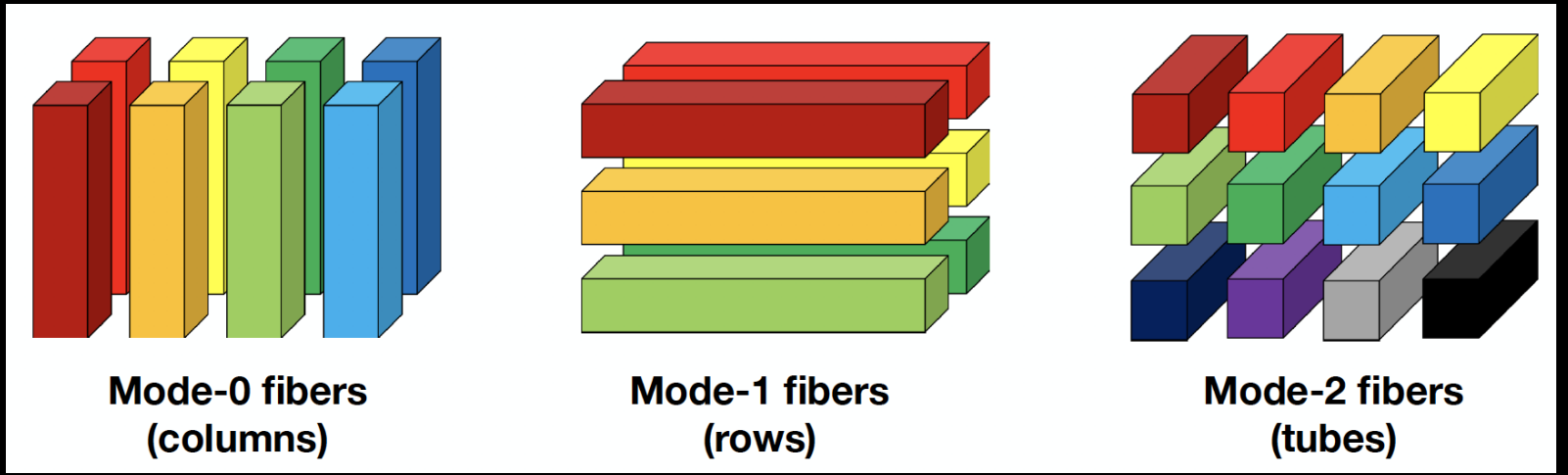
Image: 3 dimensions
Width * Height * Channels

Video: 4 dimensions
Width * Height * Channels * Time

INDEXING A TENSOR

Notion of a fiber

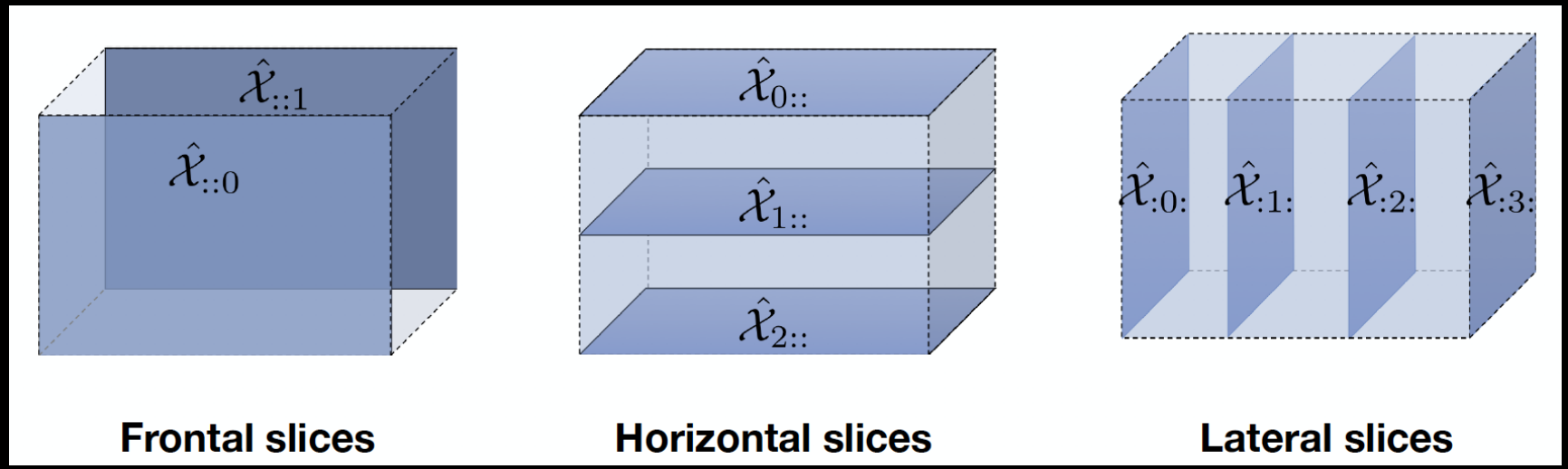
- Fibers = generalization of the concept of rows and columns for matrices
- Obtained by fixing all indices but one



INDEXING A TENSOR

Notion of a slice

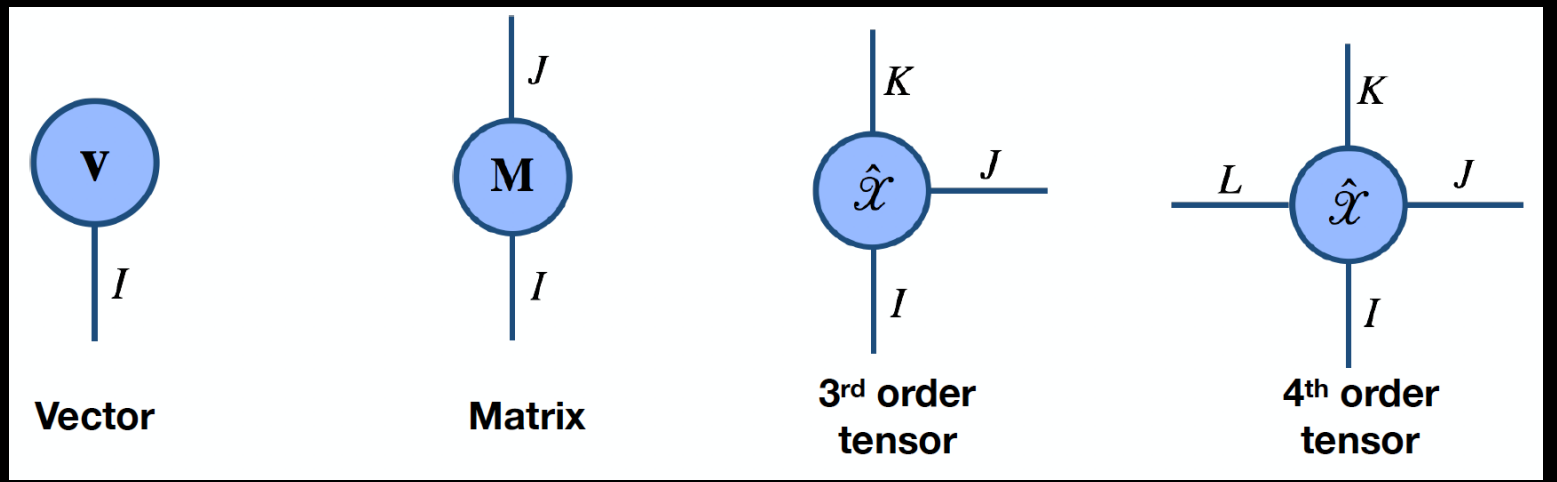
- Slices are obtained by fixing all indices but 2
- Useful to make examples by stacking matrices



TENSOR DIAGRAMS

Succinct notation

- Represent only variables and indices (dimensions)
- Tensors = vertices, mode = edge, order = degree

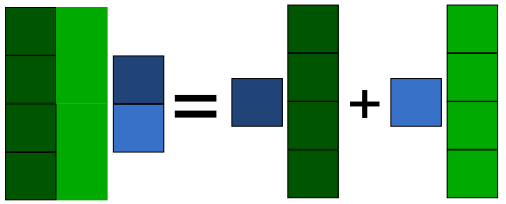


TENSORS OPERATIONS
TENSOR CONTRACTION PRIMITIVE

Extends the notion of matrix product

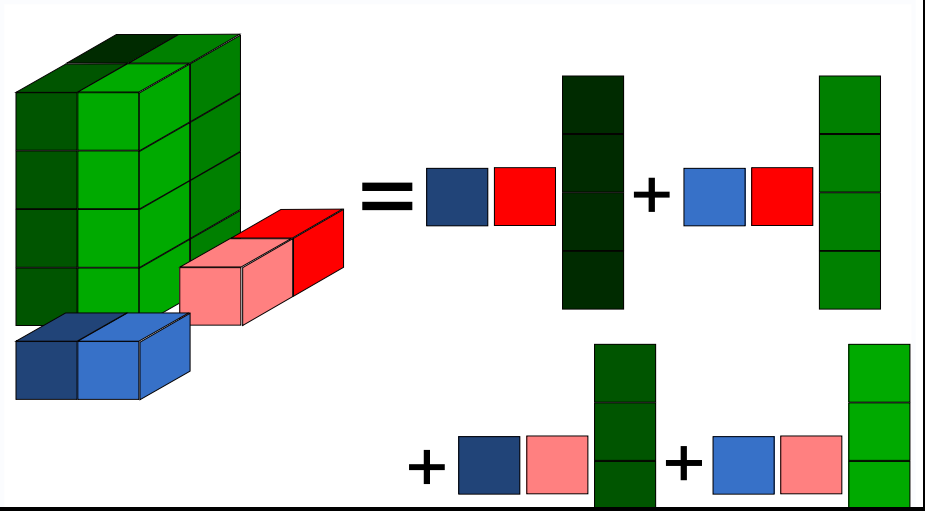
Matrix product

$$Mv = \sum_j v_j M_j$$



Tensor Contraction

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

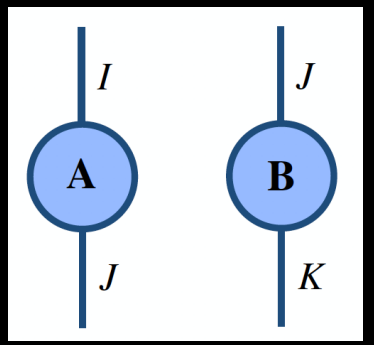


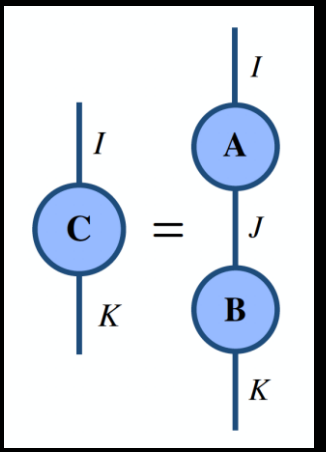
TENSOR DIAGRAMS

Succinct notation

- Contraction on a given dimension: simply link the indices over which to contract together!

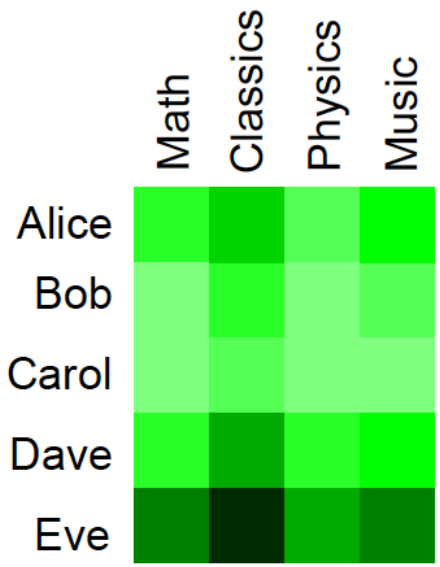
$$\mathbf{C} = \mathbf{AB} = \sum_{j=1}^{J} \mathbf{a}_{:,j}\mathbf{b}_{j,:}^{\top}$$





EXAMPLE: DISCOVERING HIDDEN FACTORS

A Matrix of Measurements

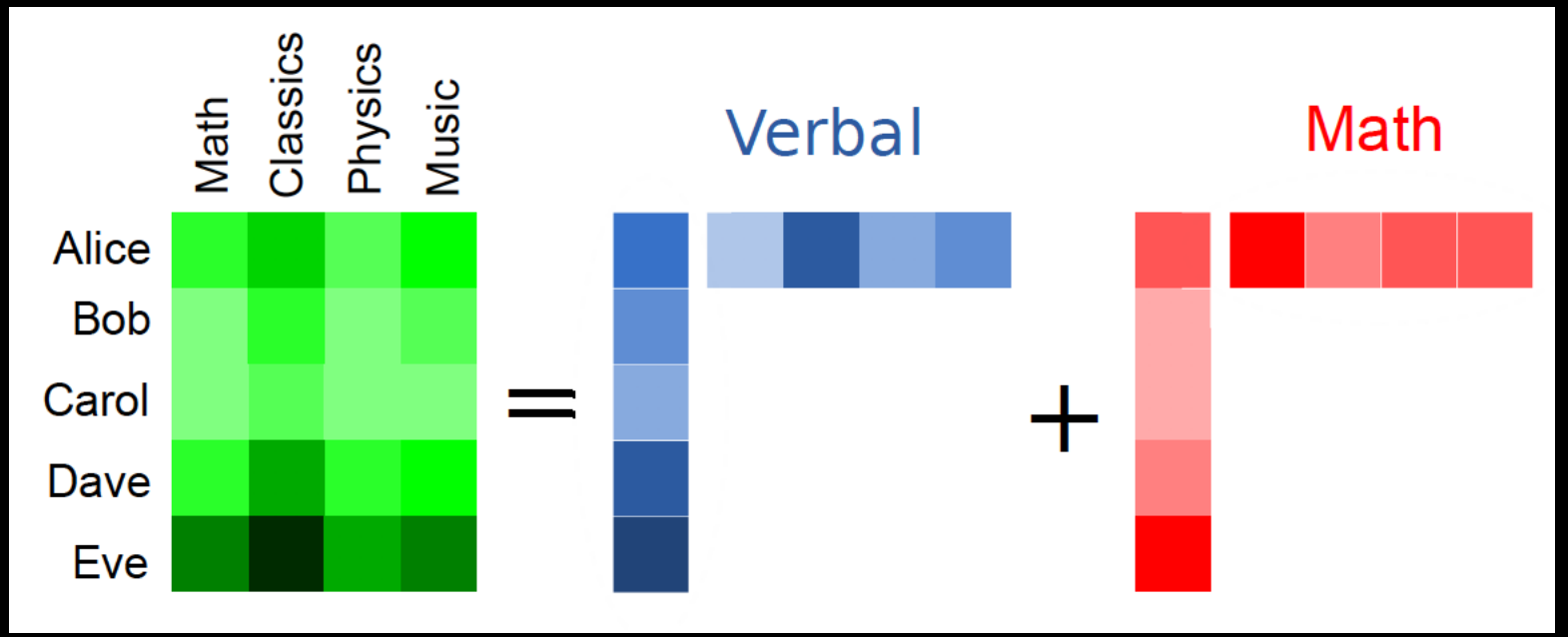


- List of scores for students in different tests
- Learn hidden factors for Verbal and Mathematical Intelligence [C. Spearman 1904]

$$\text{Score}\ (\text{student},\text{test}) = \text{student}_{\text{verbal-intlg}} \times \text{test}_{\text{verbal}} + \text{student}_{\text{math-intlg}} \times \text{test}_{\text{math}}$$

EXAMPLE: DISCOVERING HIDDEN FACTORS

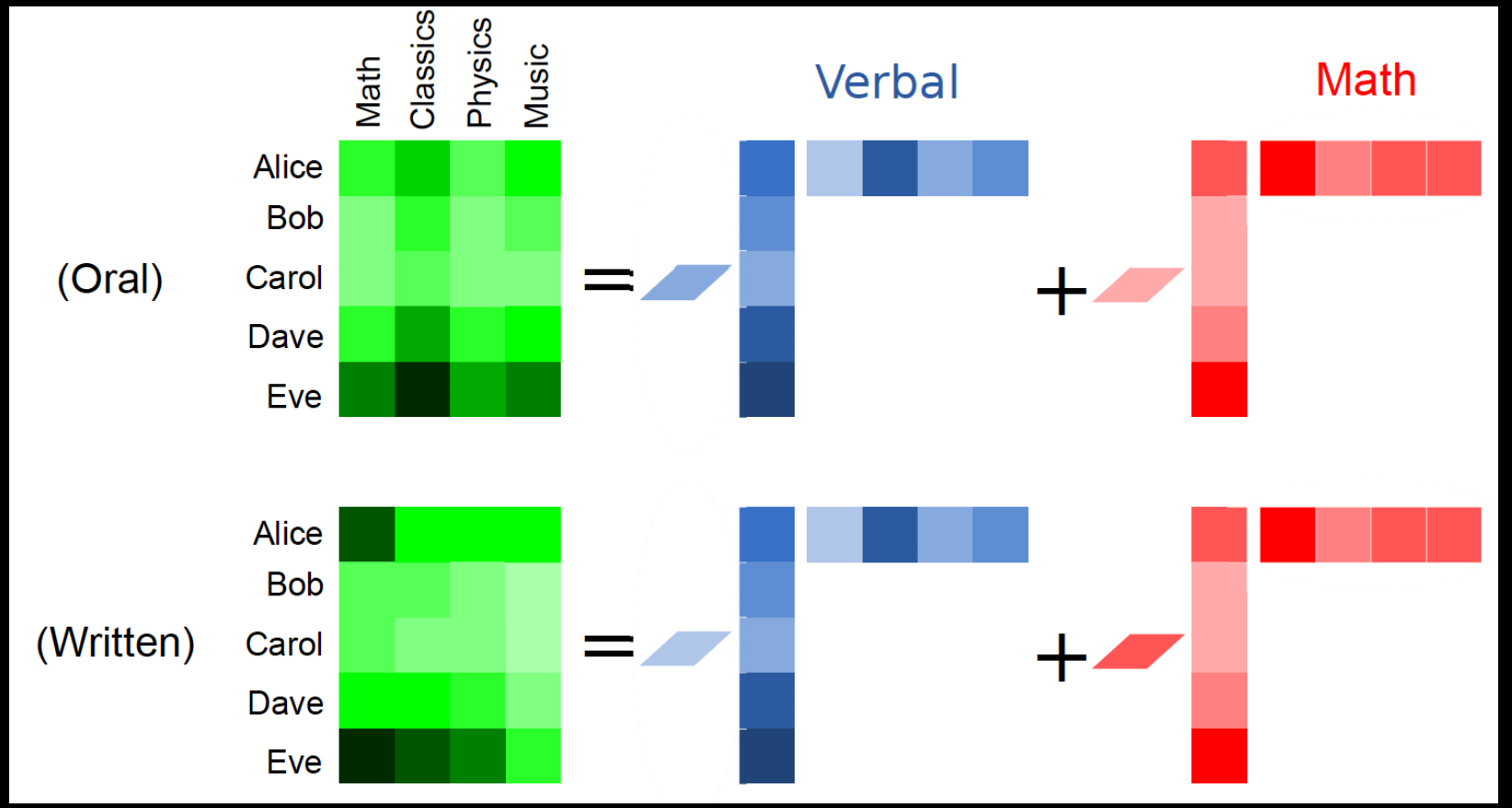
Matrix Decomposition Methods



- Find low rank Approx. of matrix.
- Each component is a latent factor

EXAMPLE: DISCOVERING HIDDEN FACTORS

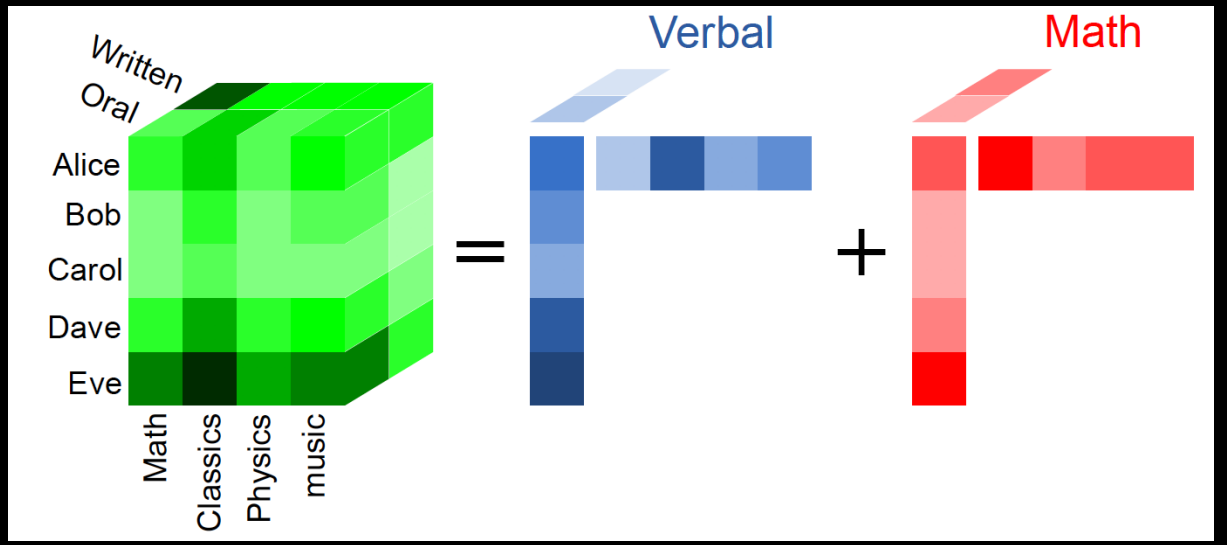
Adding more dimensions to data through tensors



- Collect more data in another dimension.
- Represent it as a tensor.
- How do we exploit this additional dimension?

EXAMPLE: DISCOVERING HIDDEN FACTORS

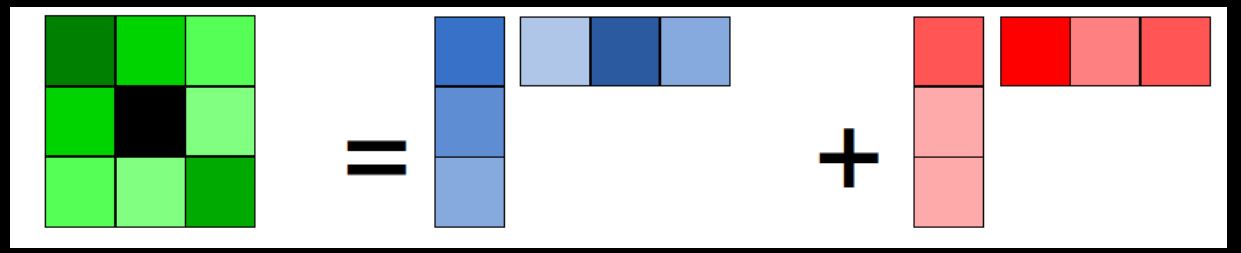
Low rank approximations of a tensor

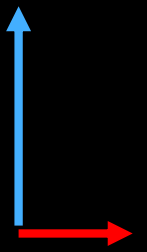


- Decompose tensor into rank-1 components.
- Declare each component as a hidden factor
- Why is this more powerful than a matrix decomposition?

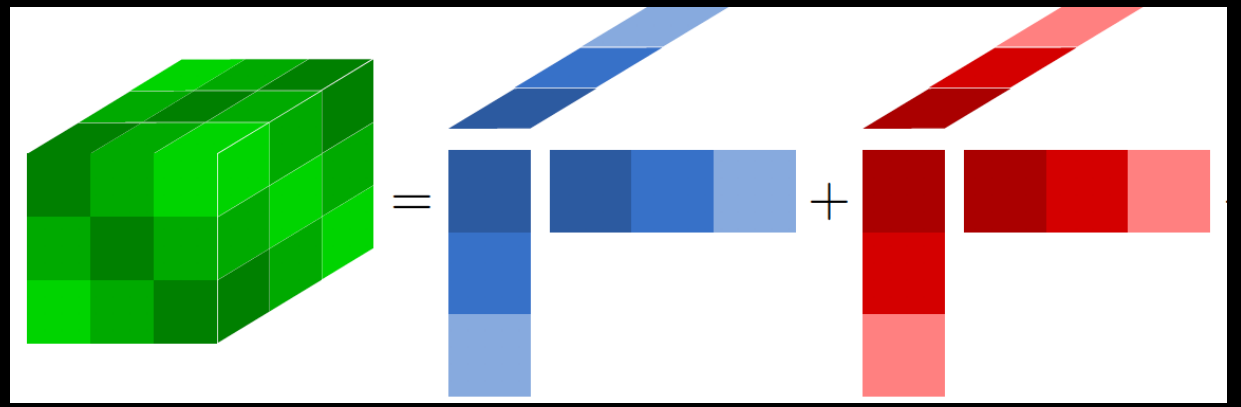
MATRIX VS TENSOR DECOMPOSITION

Conditions for unique decomposition?





Unique only when components are orthogonal



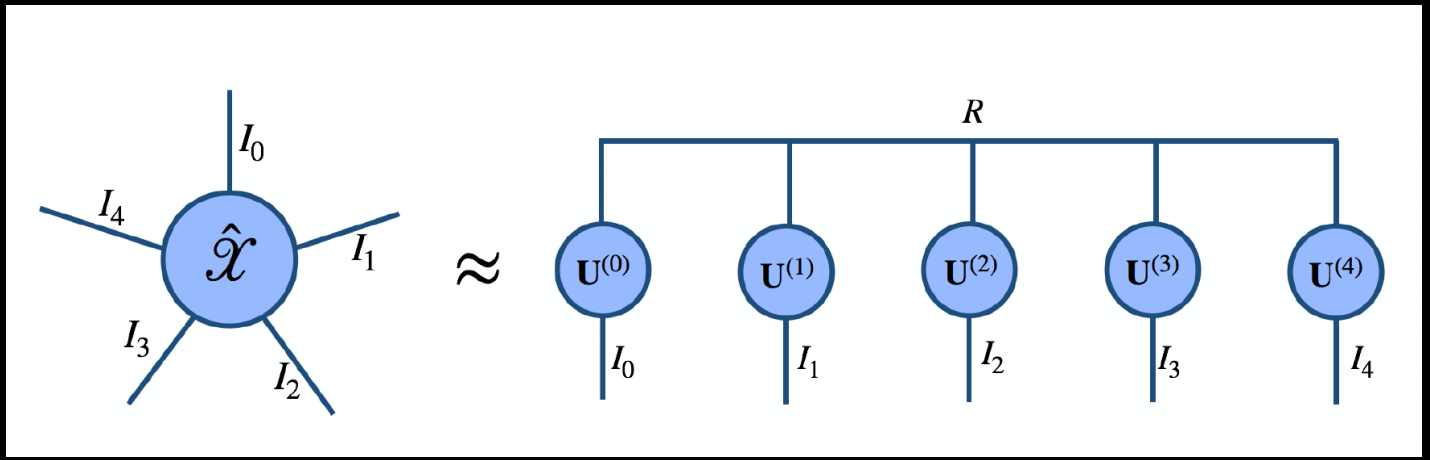
Unique when components are linearly independent

TENSOR DIAGRAMS

Notation for Tensor CP decomposition

- Contraction on a given dimension: simply link the indices over which to contract together!

$$\hat{\mathcal{X}} = \underbrace{\sum_{k=0}^{R-1} \underbrace{\mathbf{u}_k^{(0)} \circ \mathbf{u}_k^{(1)} \circ \mathbf{u}_k^{(2)}}_{\text{rank-1 components}}}_{\text{rank-R tensor}}$$



$$\mathbf{U}^{(0)} = \left[\mathbf{u}_0^{(0)}, \quad \mathbf{u}_1^{(0)}, \quad , \cdots, \mathbf{u}_{R-1}^{(0)}\right] \in \mathbb{R}^{I,R}$$

$$\mathbf{U}^{(1)} = \left[\mathbf{u}_0^{(1)}, \quad \mathbf{u}_1^{(1)}, \quad , \cdots, \mathbf{u}_{R-1}^{(1)}\right] \in \mathbb{R}^{J,R}$$

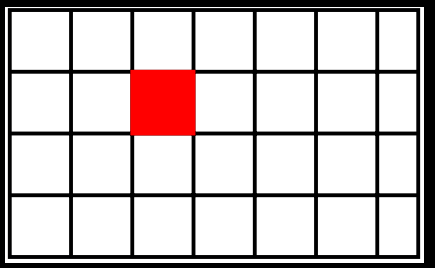
$$\mathbf{U}^{(2)} = \left[\mathbf{u}_0^{(2)}, \quad \mathbf{u}_1^{(2)}, \quad , \cdots, \mathbf{u}_{R-1}^{(2)}\right] \in \mathbb{R}^{K,R}$$

TENSORS FOR HIGHER ORDER MOMENTS

WHY IS IT MORE POWERFUL?

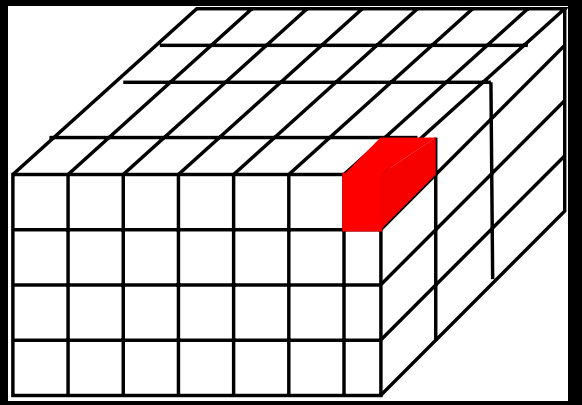
Pairwise correlations

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



Third order correlations

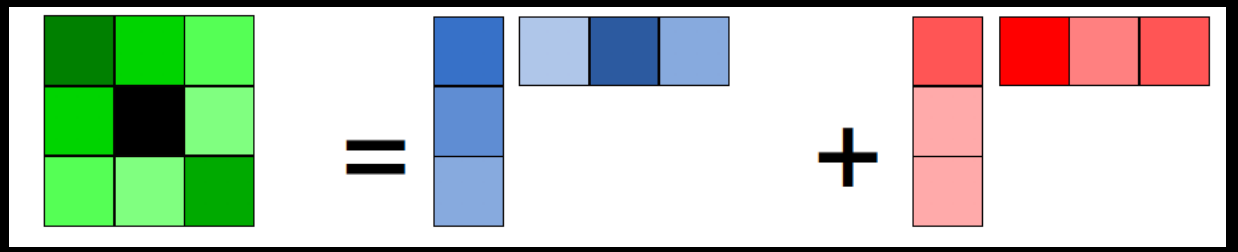
$$E(x \otimes x \otimes x)_{i,j,k} = E(x_i x_j x_k)$$



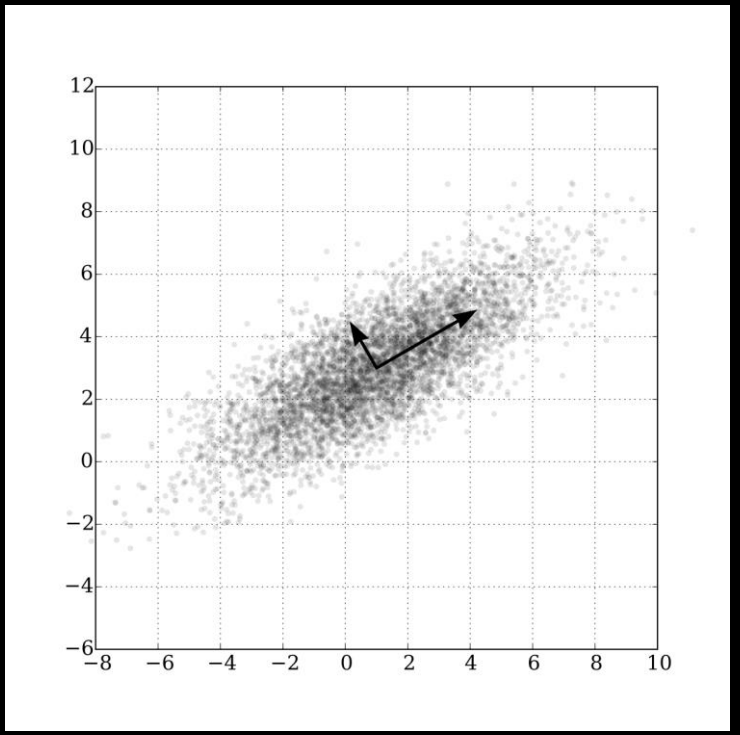
PRINCIPAL COMPONENT ANALYSIS (PCA)

Low-rank approximation of Covariance Matrix

- Problem: Find best rank-k projection of (centered) data
- Solution: Top Eigen components of Covariance matrix



- Limitation: Uses first two moments. Gaussian approx.
- But data tends to be far from Gaussian.



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

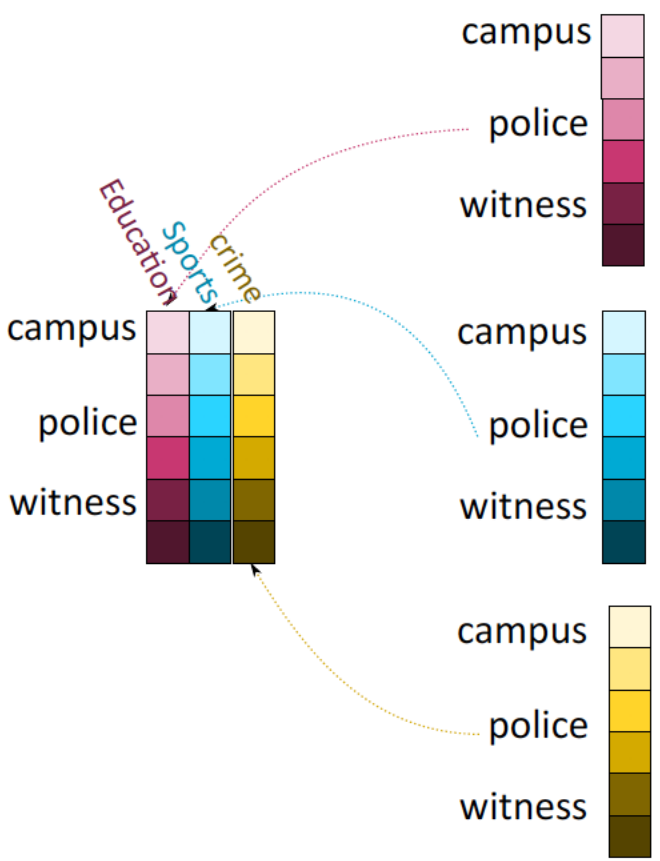
As The Times reported last April, the Tallahassee police also failed to aggressively investigate the rape accusation. It did not become public until November, when a Tampa reporter, Matt Baker, acting on a tip, sought records of the police investigation.

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."

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.

UNSUPERVISED LEARNING TOPIC MODELS THROUGH TENSORS



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.

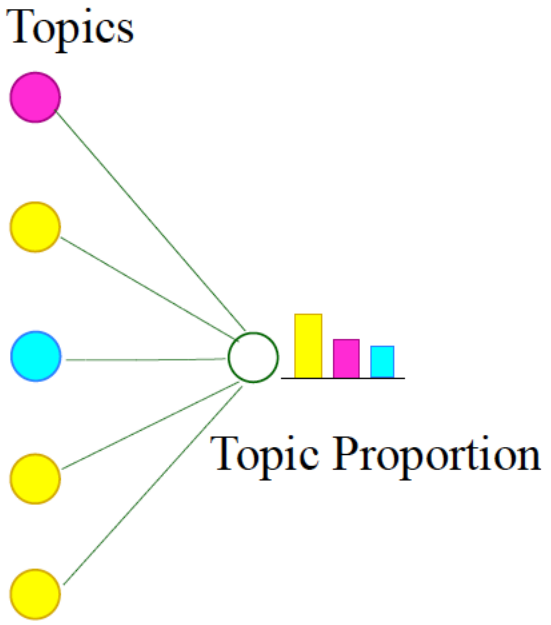
In the 21 months since, Florida State officials have said little about how they handled the case, which is no investigated by the federal Depart

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.

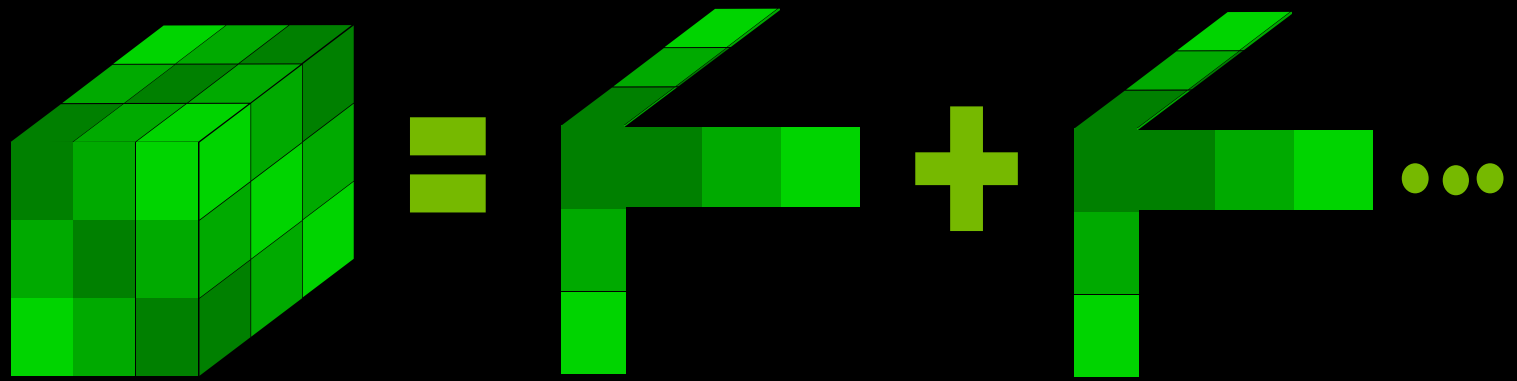
Most recently, university officials suspended Mr. Winston for one game after he stood in a public place on campus and, playing off a running Internet gag, shouted a crude reference to a sex act. In a news conference afterward, his coach, Jimbo Fisher, said, "Our hope and belief is Jameis will learn from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

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



Co-occurrence of word triplets

Topic 1

Topic 2

WHY TENSORS?

Statistical reasons:

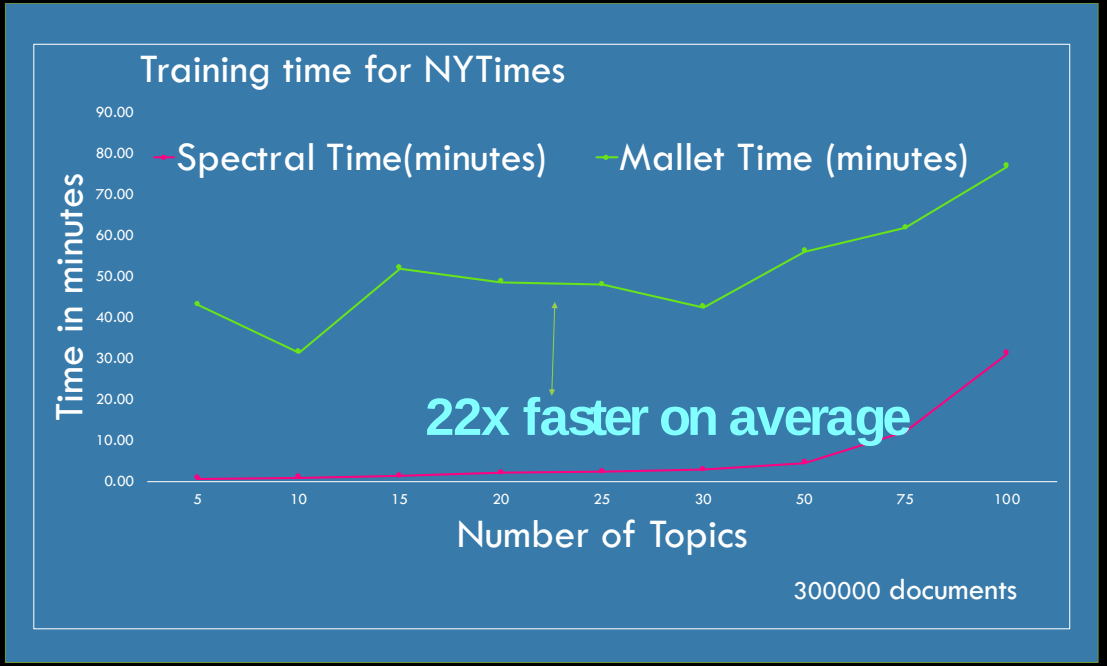
- Incorporate **higher order** relationships in data
- Discover hidden topics (not possible with matrix methods)

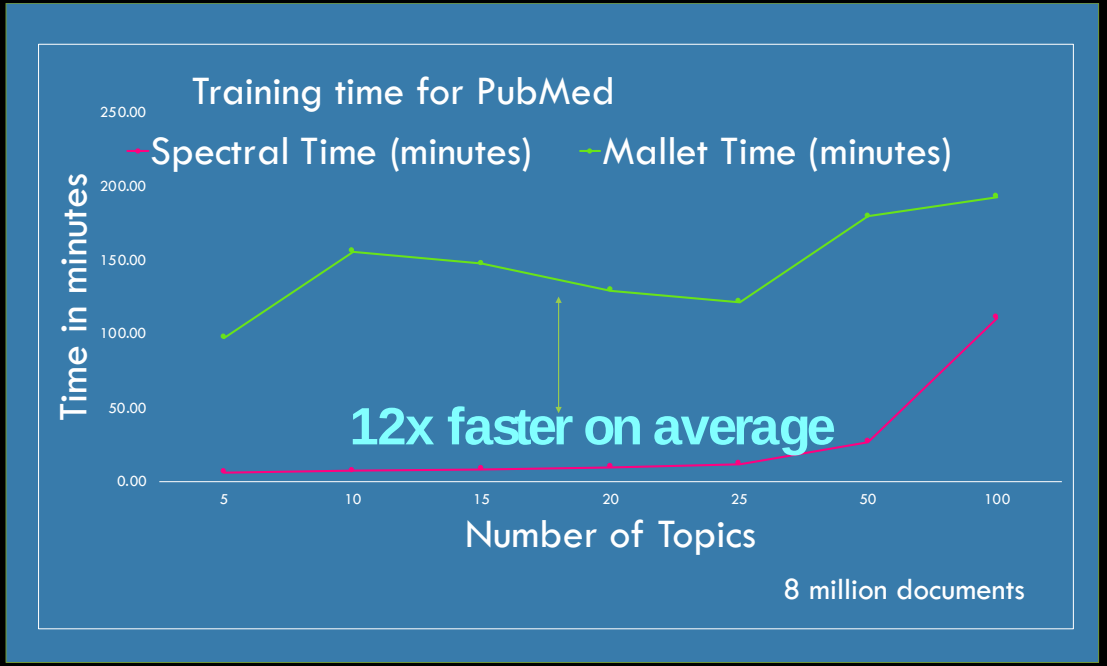
Computational reasons:

- Tensor algebra is **parallelizable** like linear algebra.
- **Faster** than other algorithms for LDA
- **Flexible:** Training and inference decoupled
- **Guaranteed** in theory to converge to global optimum

A. Anandkumar etal,Tensor Decompositions for Learning Latent Variable Models, JMLR 2014.

TENSOR-BASED TOPIC MODELING IS FASTER





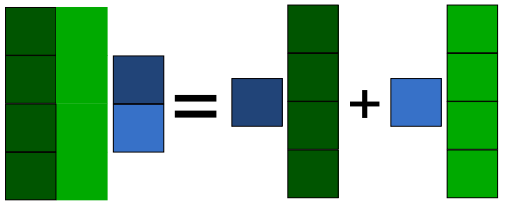
- Mallet is an open-source framework for topic modeling
- Benchmarks on AWS SageMaker Platform
- Bulit into AWS Comprehend NLP service.

TENSORS OPERATIONS
TENSOR CONTRACTION PRIMITIVE

Extends the notion of matrix product

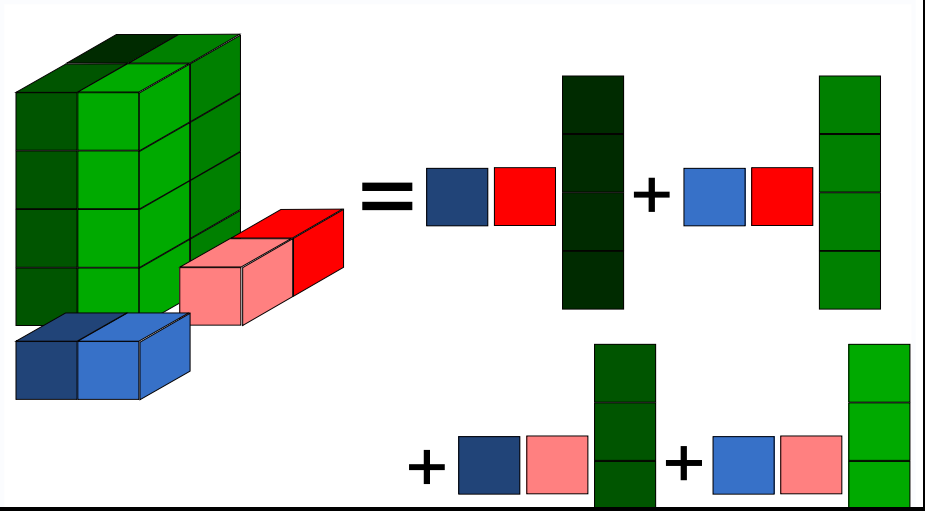
Matrix product

$$Mv = \sum_j v_j M_j$$



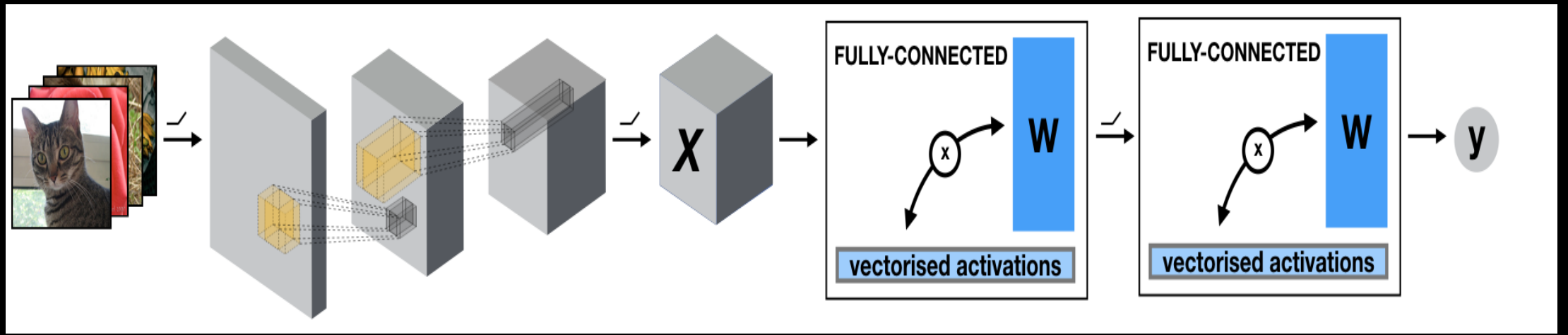
Tensor Contraction

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



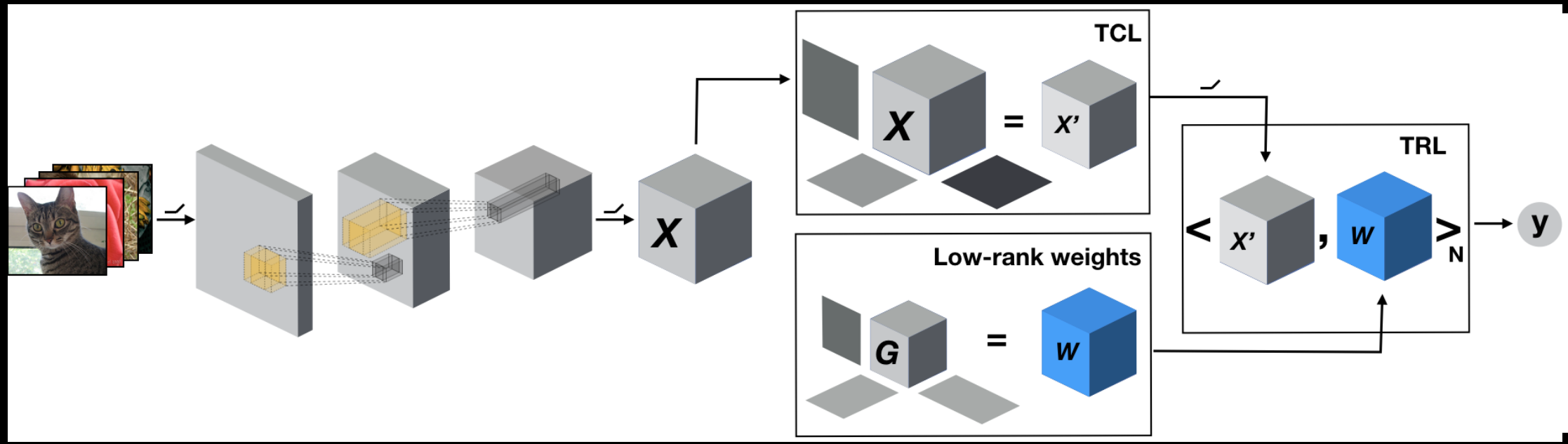
TENSORS FOR MODELS

STANDARD CNN USE LINEAR ALGEBRA



TENSORS FOR MODELS

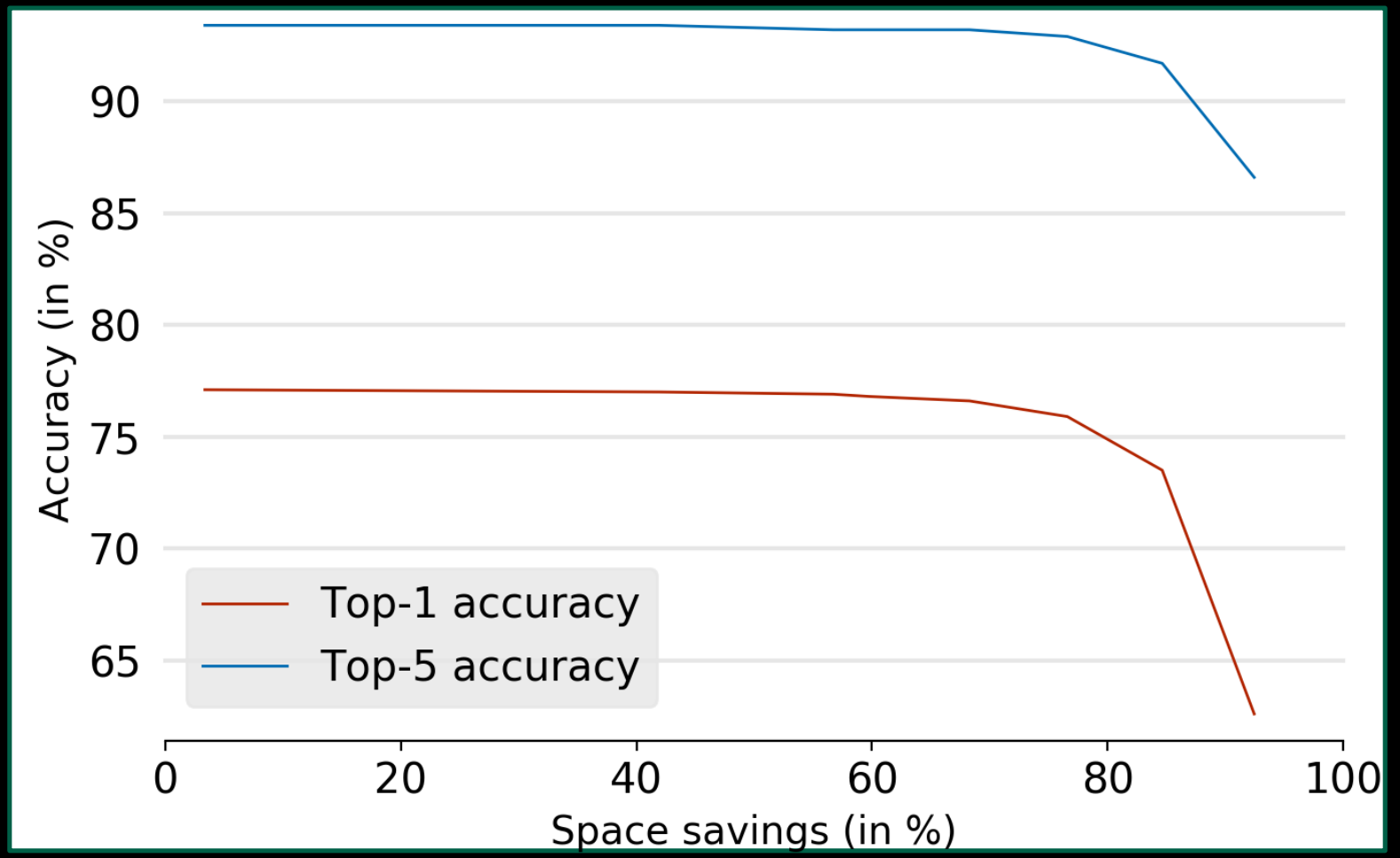
TENSORIZED NEURAL NETWORKS



Jean Kossaifi, Zack Chase Lipton, Aran Khanna, Tommaso Furlanello, **A**

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

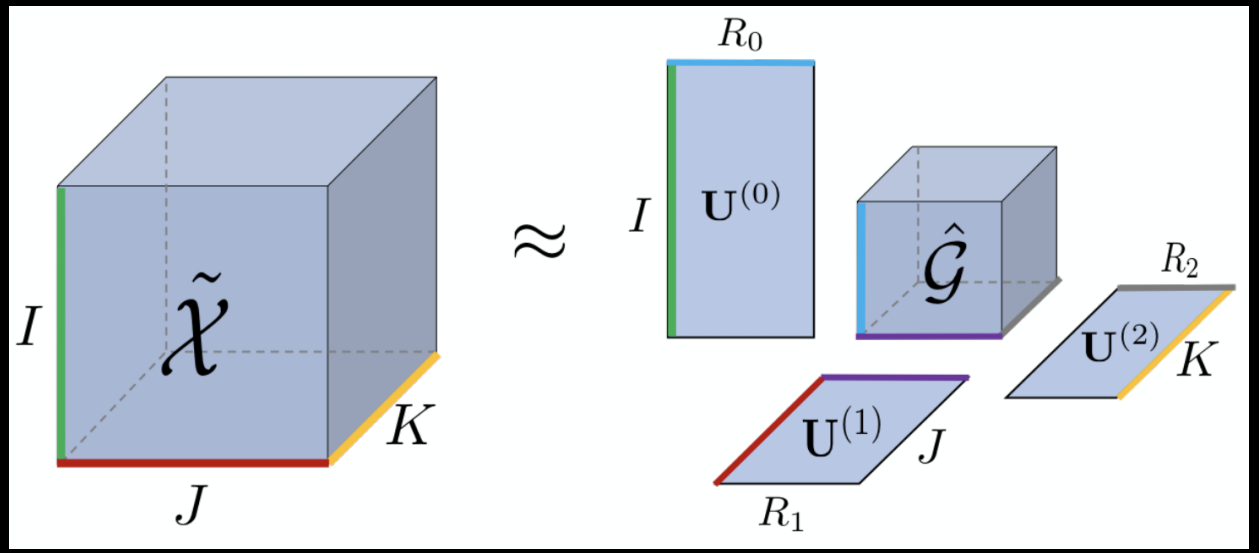
SPACE SAVING IN DEEP TENSORIZED NETWORKS



TUCKER DECOMPOSITION

Generalizing Tensor CP decomposition

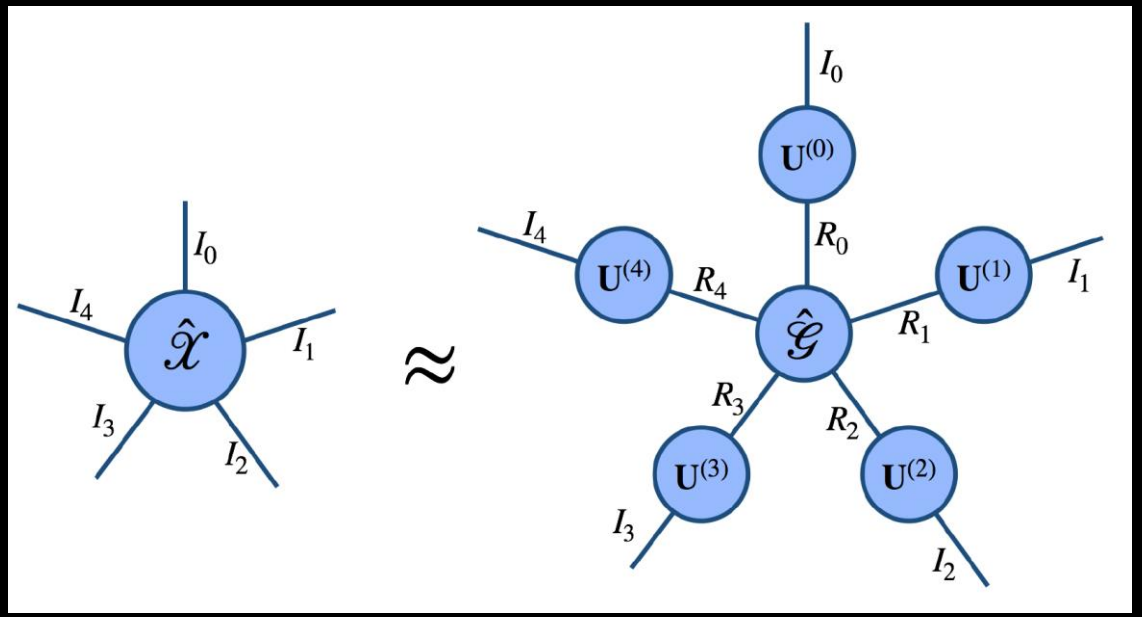
$$\hat{\mathcal{X}} = \hat{\mathcal{G}} \times_0 \mathbf{U}^{(0)} \times_1 \mathbf{U}^{(1)} \times_2 \mathbf{U}^{(2)}$$



TENSOR DIAGRAMS

Notation for Tucker Decomposition

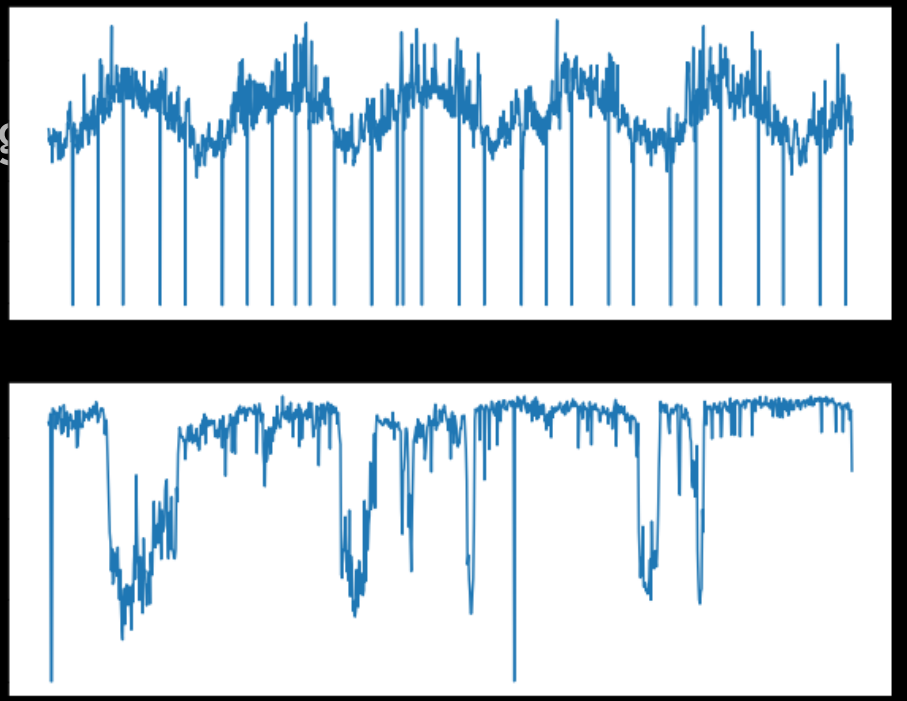
- Contraction on a given dimension: simply link the indices over which to contract together!



TENSORS FOR LONG-TERM FORECASTING

Difficulties in long term forecasting

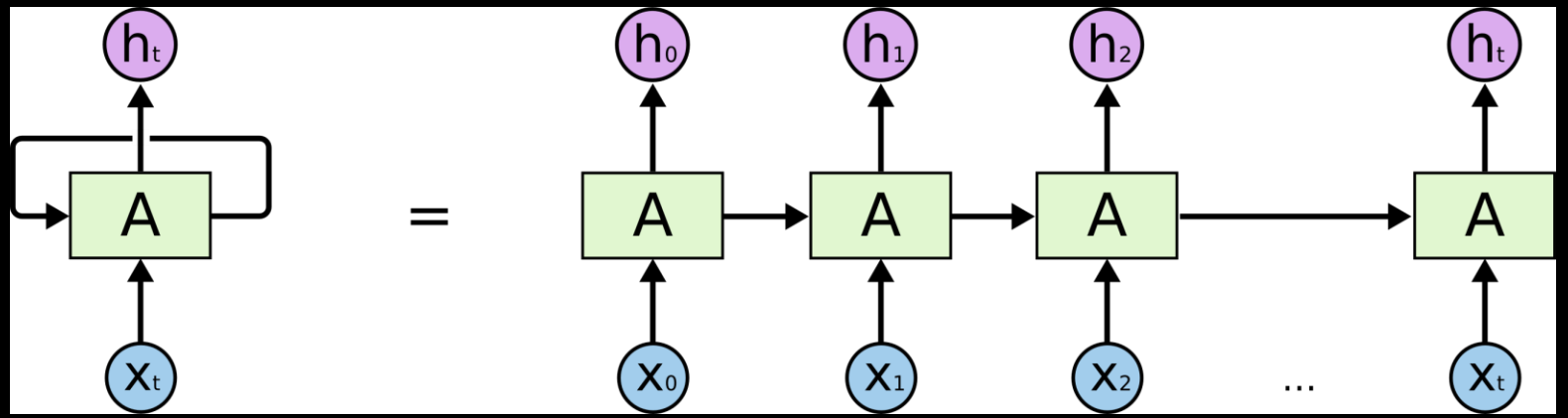
- Long-term dependencies
- High-order correlations
- Error propagation



39

RNNS: FIRST-ORDER MARKOV MODELS

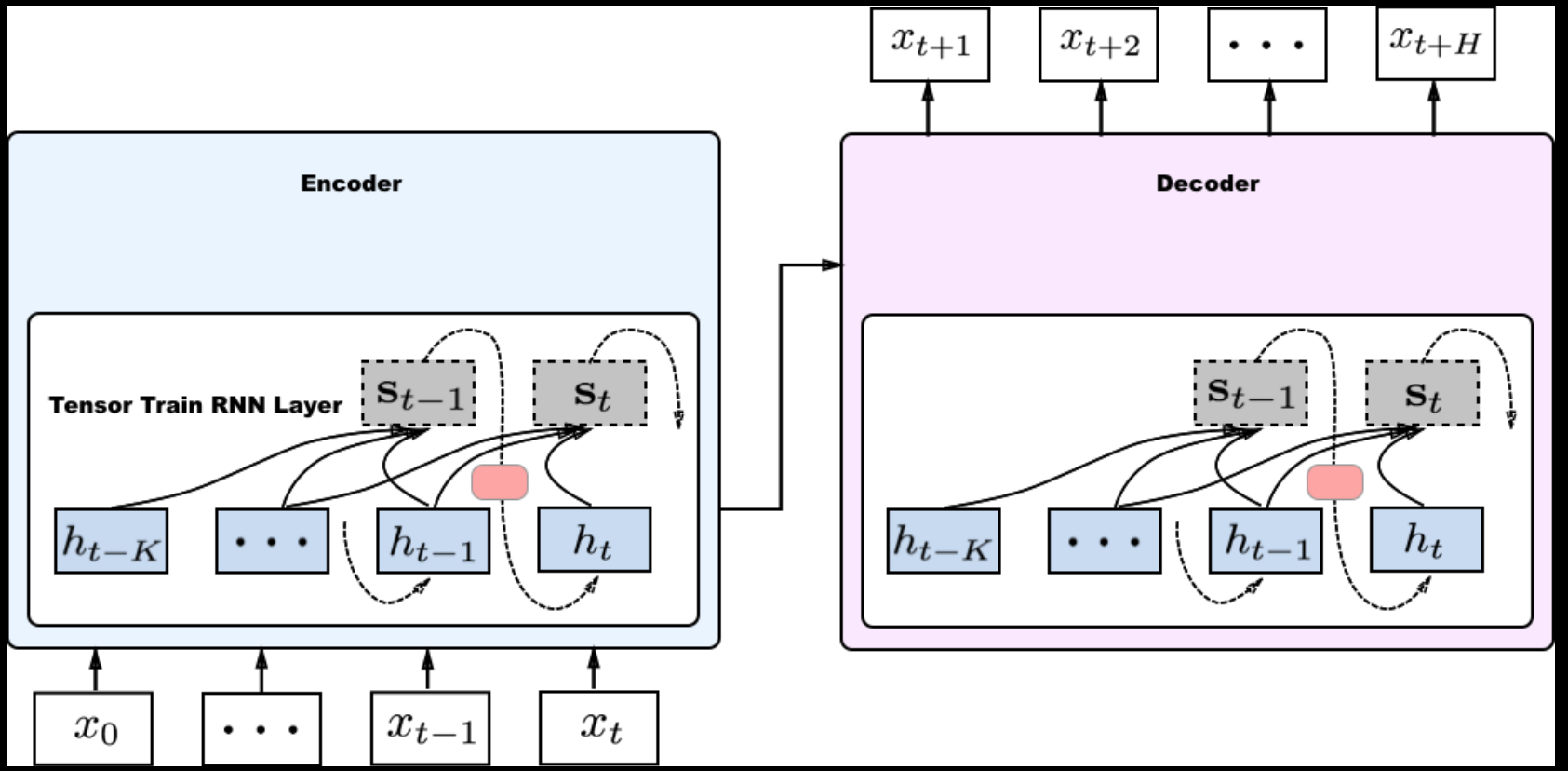
Input state x_t, hidden state h_t, output y_t,



TENSOR-TRAIN RNNS AND LSTMS

Seq2seq architecture

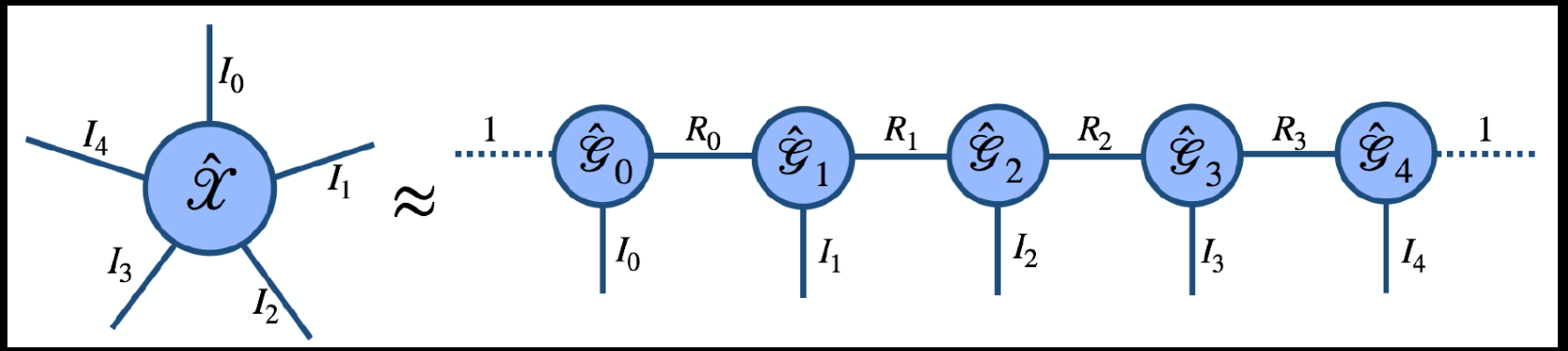
TT-LSTM cells



TENSOR DIAGRAMS

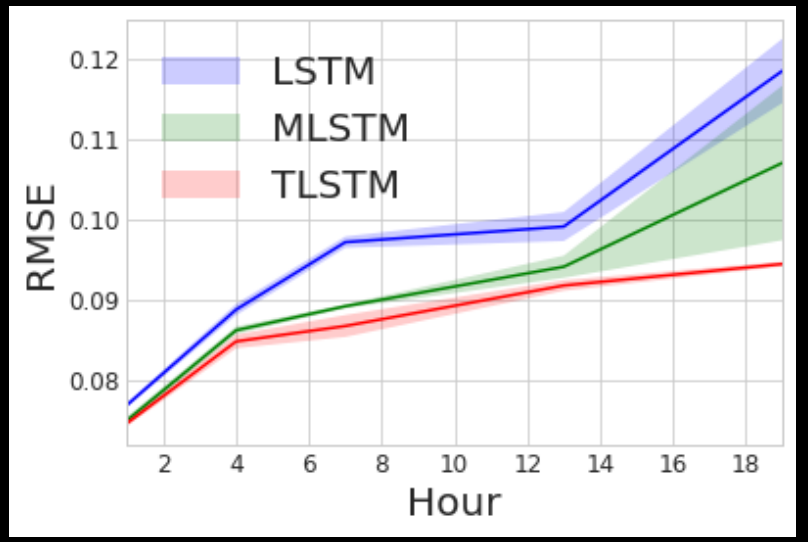
Notation for Tensor Train

- Contraction on a given dimension: simply link the indices over which to contract together!

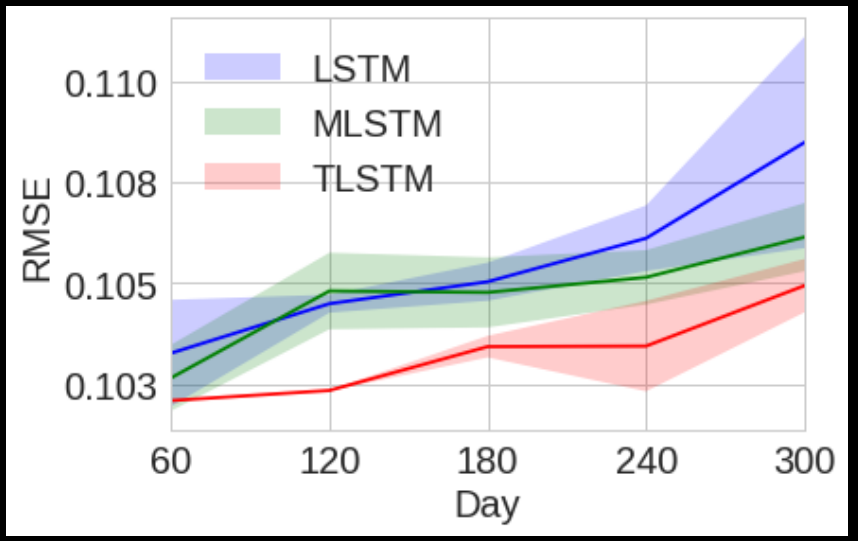


TENSOR LSTM FOR LONG-TERM FORECASTING

Traffic dataset



Climate dataset



Rose Yu

Stephan Zhang

Yisong Yue

APPROXIMATION GUARANTEES FOR TT-RNN

- Approximation error : bias of best model in function class.
- No such guarantees exist for RNNs.

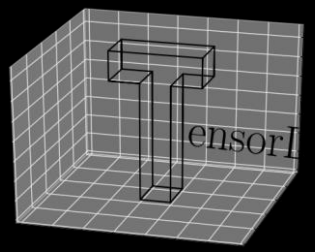
Theorem: TT-RNN with m units approx. with error ε

$$m \leq O\left(\frac{C^2}{\epsilon}(dr^{-k} + p^{-k})\right)$$

- Dimension d , tensor-train rank r. Window p.
- Bounded derivatives order k , smoothness C

- Easier to approximate if function is **smooth and analytic.**
- **Higher rank and bigger window** more efficient.

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

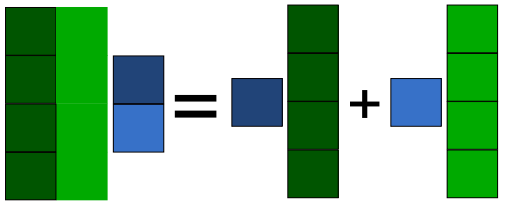
Set optimizer

TENSORS FOR COMPUTE
TENSOR CONTRACTION PRIMITIVE

Extends the notion of matrix product

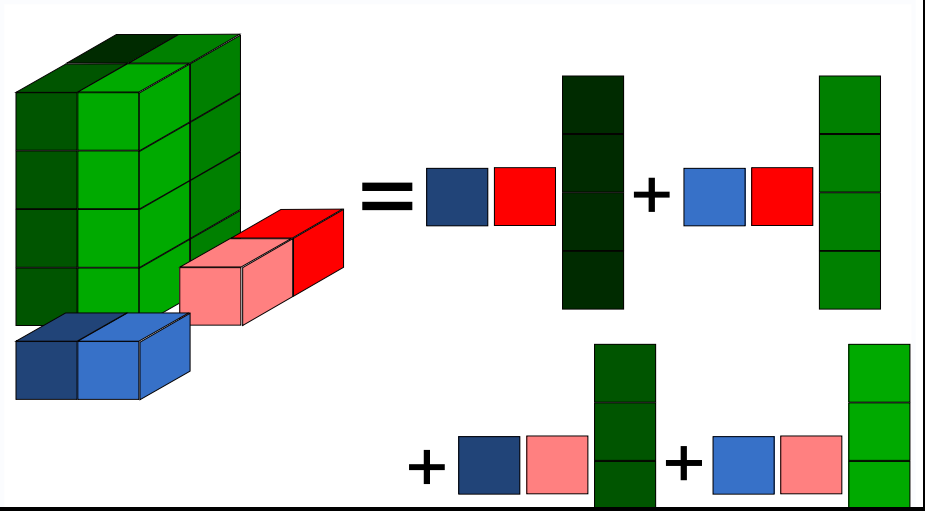
Matrix product

$$Mv = \sum_j v_j M_j$$



Tensor Contraction

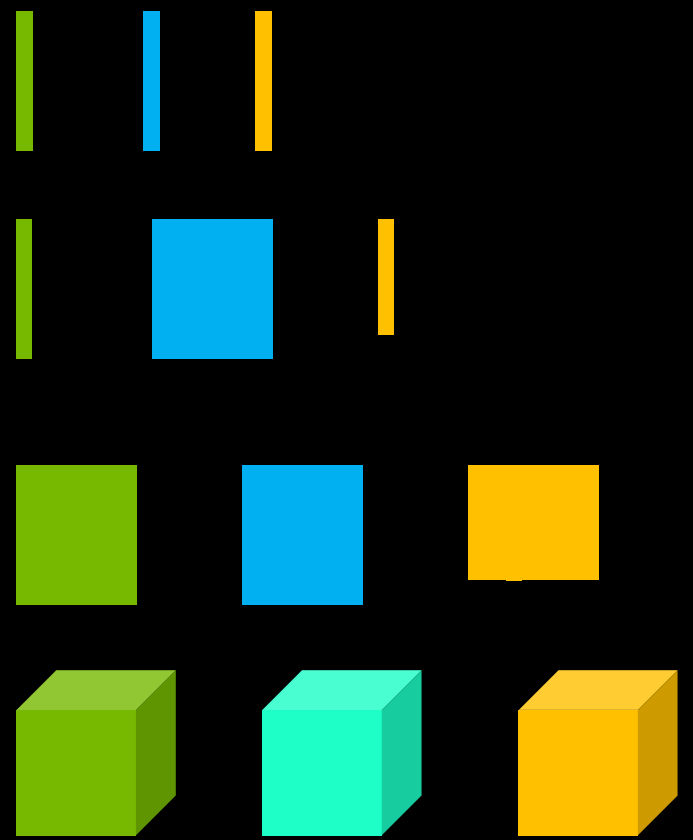
$$T(u, v, \cdot) = \sum_{i,j} u_i v_j T_{i,j,:}$$

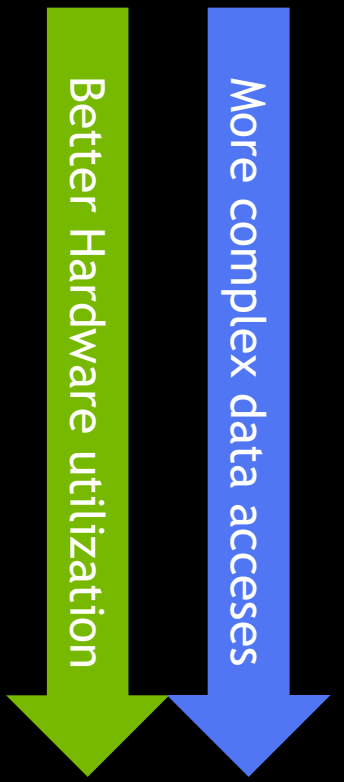


TENSOR PRIMITIVES?

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).

S9593 - cuTENSOR: High-performance Tensor Operations in CUDA

We'll discuss cuTENSOR, a high-performance CUDA library for tensor operations that efficiently handles the ubiquitous presence of high-dimensional arrays (i.e., tensors) in... View More

Talk

Paul Springer - Senior Software Engineer, NVIDIA
Chenhan Yu - Deep Learning Software Engineer, NVIDIA

Wednesday, Mar 20, 1:00 PM - 01:50 PM – SJCC Room 211B (Concourse Level)

Add to My Interests

Thank you