Advancing the Bases: Tensor Methods and the Frontiers of Generality

Matthew D. Merris, Tensor500 Team, Department of Computer Science
Timothy Andersen, Tensor500 Team: P.I., Department of Computer Science
Richard Murphy, Tensor500 Team: Co-P.I., Micron Technology

BOISE STATE UNIVERSITY
COLLEGE OF ENGINEERING
Department of Computer Science

Key Research Questions

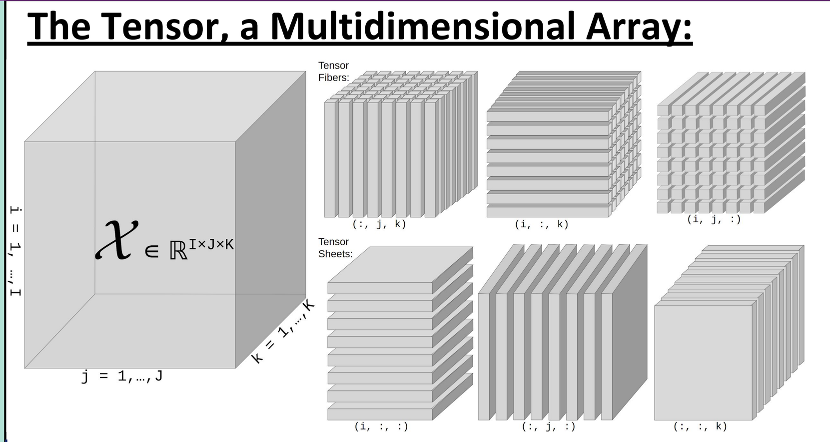
- Do workflow centric meta-kernels provide more actionable insight than traditional kernel based benchmarking?
- How to abstract workflows into a representative set?
- Can we build a means to synthesis sufficiently realistic tensors across the application domain of tensor methods?
- What does portability look like in the era of increased hardware heterogeneity?
- How fuzzy is too fuzzy as our technologies move increasingly towards favoring precision over accuracy?

Tensor500: a streaming analytics HPC benchmark

- A modular set of meta-kernels to enable composition of representative workflows tailored to a particular application space.
- Robust capacity to handle various streaming conditions from traditional to multi-aspect.
- Synthetic data generation of arbitrary size and dimension capable of mimicking real-world data specific to a particular application.
- Implementation that is easily scalable and portable beyond traditional architectures.

Overview

The Tensor, a Multidimensional Array:



Expanding Application Space:

- signal processing
- data mining
- network analysis and security
- machine learning
- neuroscience
- remote sensing
- environmental monitoring
- text based systems
- process control
- engineering
- medical
- forecasting
- streaming

Modern data problems:

- Massive Data Scale
- Sparsity
- Faulty or Missing Data
- Data Fusion
- Preprocessing/Format

Modern hardware problems:

- The walls are closing in
- Resolving the backend
- Smaller, faster, funnier or…
  - We need a hydroelectric power plant

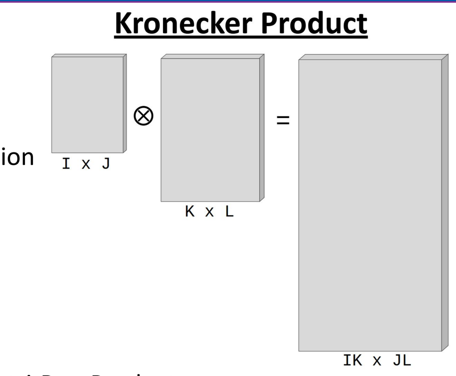
Hardware innovations:

- GPU
- TPU
- FPGAs
- EMU
- Wafer-scale Processor
- Quantum Computing
- Photonics

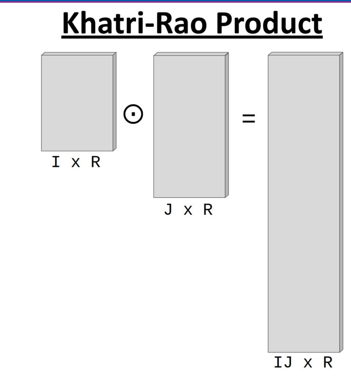
Tensor kernel examples:

- TEW: Tensor Element-Wise
- TS: Tensor-Scalar
- TTV: Tensor-Times-Vector
  - Result: Reduced by 1 dimension
- TTM: Tensor-Times-Matrix
  - Result: 1 dimension rescaled
- Matrix products
  - Kronecker Product
  - Khatri-Rao Product
  - MTTKRP: Matricized-Tensor-Times-Khatri-Rao-Product

Kronecker Product



Khatri-Rao Product



References:

H. Fanaee-T and J. Gama, "Tensor-based anomaly detection: An interdisciplinary survey," Knowledge-Based Systems, vol. 98, pp. 130–147, Apr. 2016.

T. G. Kolda and B. W. Bader, "Tensor Decompositions and Applications," SIAM Review, vol. 51, no. 3, pp. 455–500, Aug. 2009.

J. Li, Y. Ma, X. Wu, A. Li, and K. Barker, "PASTA: A Parallel Sparse Tensor Algorithm Benchmark Suite," arXiv:1902.03317 [cs], Feb. 2019.

E. E. Papalexakis, C. Faloutsos, and N. D. Sidiropoulos, "Tensors for Data Mining and Data Fusion: Models, Applications, and Scalable Algorithms," ACM Trans. Intell. Syst. Technol., vol. 8, no. 2, pp. 1–44, Oct. 2016.

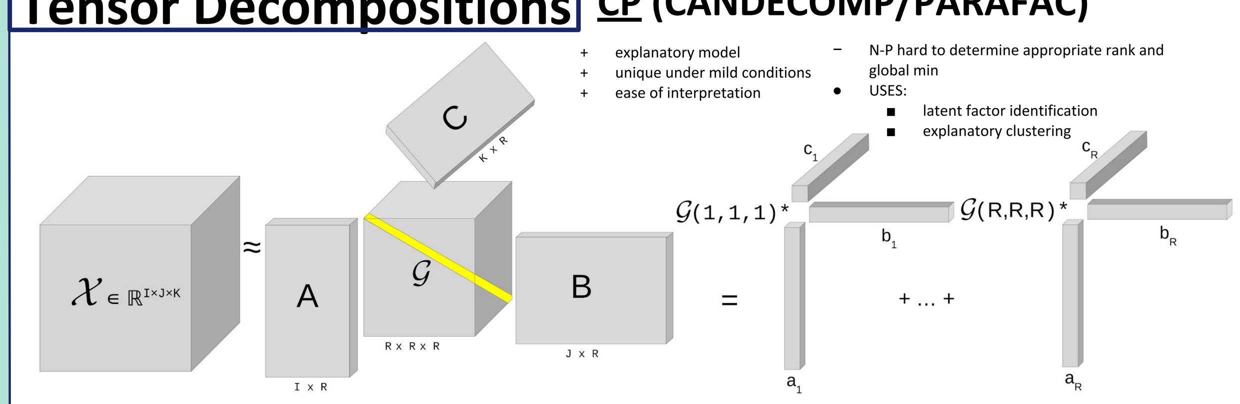
Q. Song, X. Huang, H. Ge, J. Caverlee, and X. Hu, "Multi-Aspect Streaming Tensor Completion," in Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '17, Halifax, NS, Canada, 2017, pp. 435–443.

J. Sun, D. Tao, and C. Faloutsos, "Beyond Streams and Graphs: Dynamic Tensor Analysis," in Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, NY, USA, 2006, pp. 374–383.

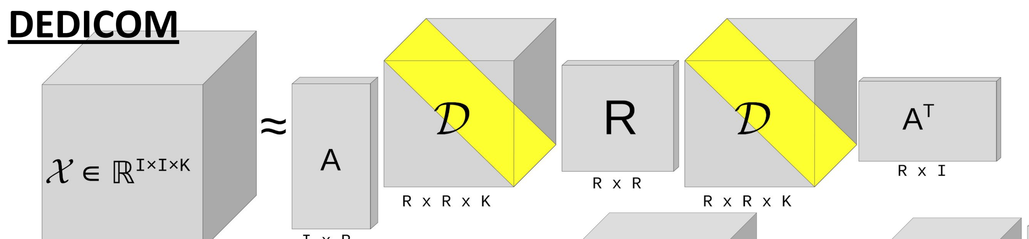
Tensor Decompositions

CP (CANDECOMP/PARAFAC)

- \+ explanatory model
- \+ unique under mild conditions
- \+ ease of interpretation
- − N-P hard to determine appropriate rank and global min
- USES:
  - latent factor identification
  - explanatory clustering

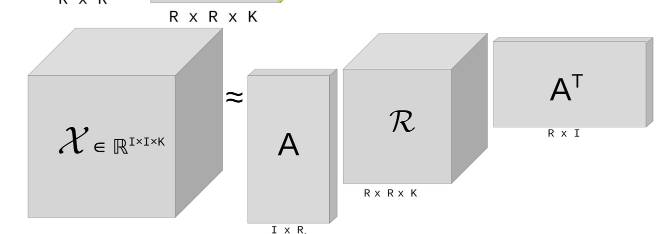


DEDICOM



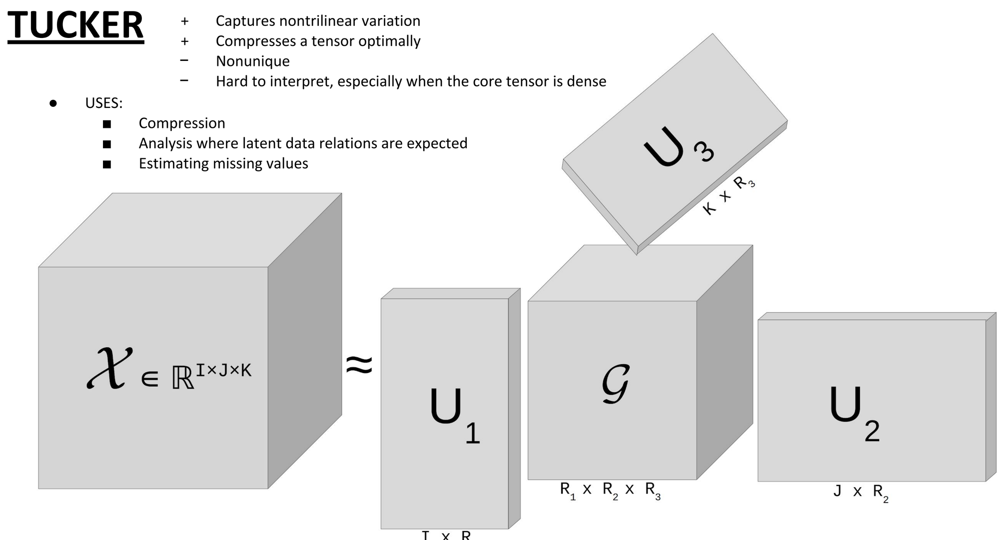
- \+ Model multirelational data that exhibits asymmetries
- \+ Unique
- − Restricted to multirelationsional data
- USES:
  - Modeling asymmetric relationships in social network data
  - Knowledge base relations
  - Linked data

RESCAL



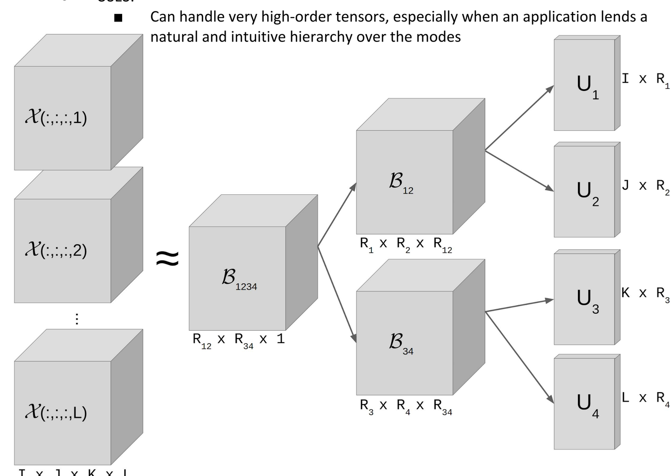
TUCKER

- \+ Captures nontrilinear variation
- \+ Compresses a tensor optimally
- − Nonunique
- − Hard to interpret, especially when the core tensor is dense
- USES:
  - Compression
  - Analysis where latent data relations are expected
  - Estimating missing values



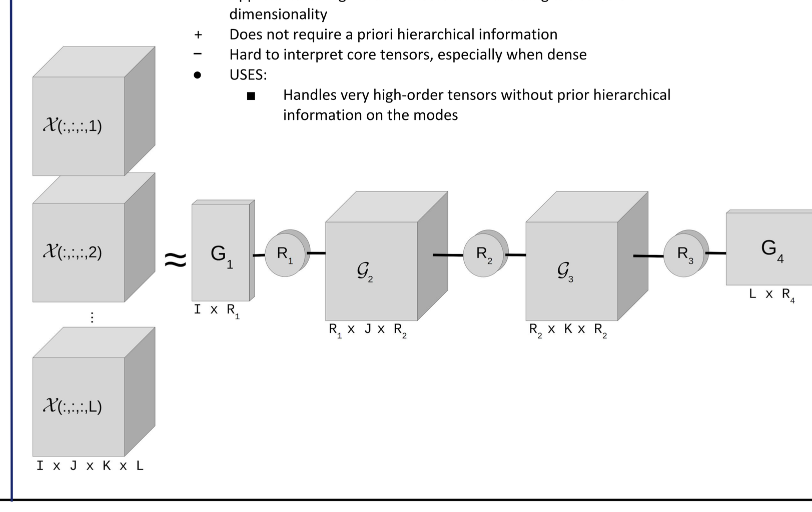
HIERARCHICAL TUCKER

- \+ Approximates high order tensors while avoiding the curse of dimensionality
- − Requires a priori knowledge of a binary tree of matricazation
- USES:
  - Can handle very high-order tensors, especially when an application lends a natural and intuitive hierarchy over the modes



TENSOR-TRAIN

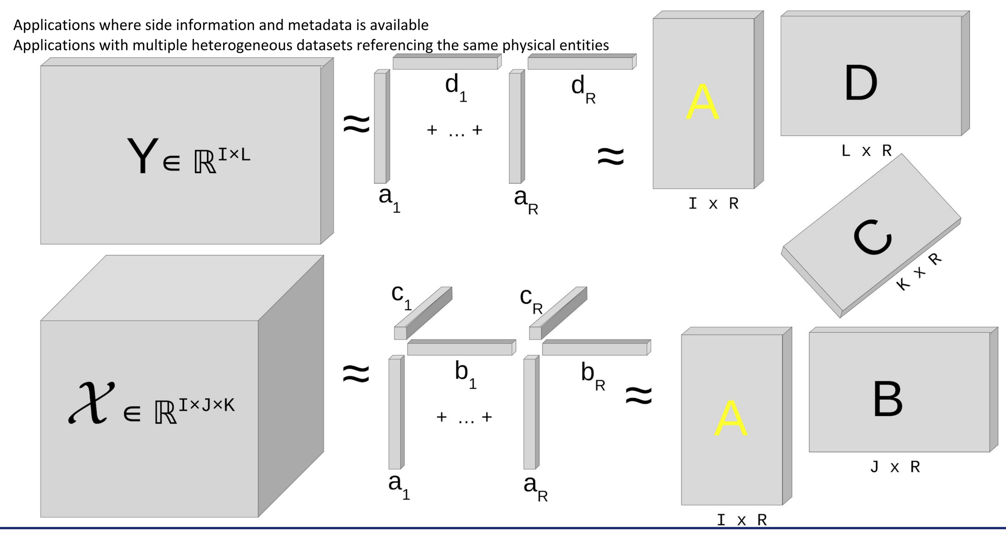
- \+ Approximates high order tensors while avoiding the curse of dimensionality
- \+ Does not require a priori hierarchical information
- − Hard to interpret core tensors, especially when dense
- USES:
  - Handles very high-order tensors without prior hierarchical information on the modes



Data Fusion

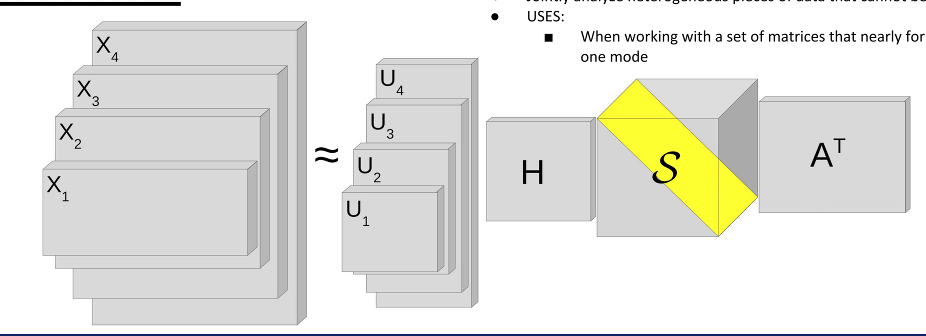
COUPLED MODEL

- \+ Jointly analyzes heterogeneous datasets with a common mode
- \+ Incorporate side information and metadata to the analysis
- − When datasets vary vastly in terms of size and volume, appropriate weighting must be applied to avoid drowning
- USES:
  - Applications where side information and metadata is available
  - Applications with multiple heterogeneous datasets referencing the same physical entities

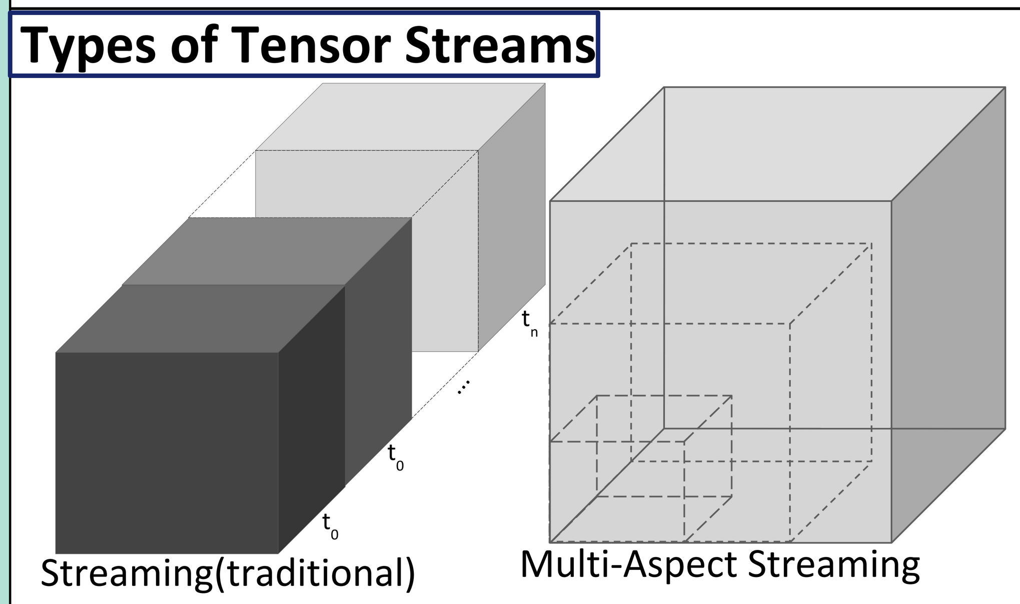


PARAFAC2

- \+ Jointly analyze heterogeneous pieces of data that cannot be expressed as a tensor
- USES:
  - When working with a set of matrices that nearly form a tensor, but vary along one mode

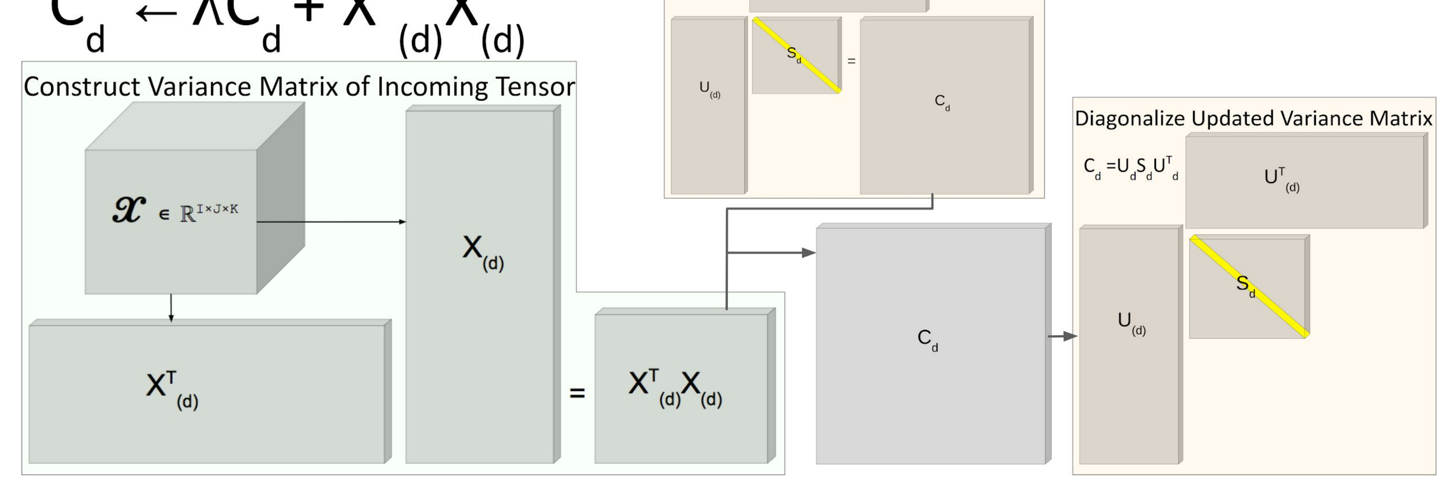


Types of Tensor Streams

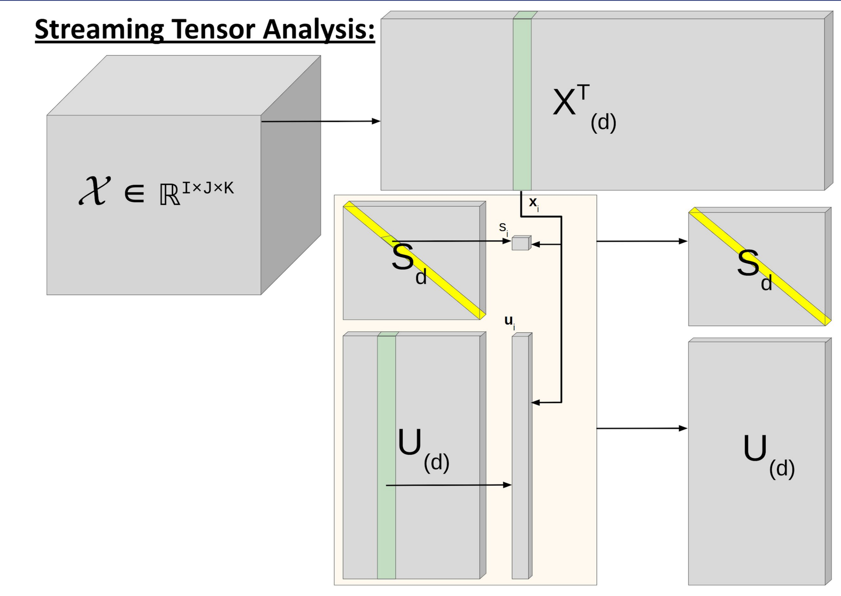


Dynamic Tensor Analysis:

$$C_d \leftarrow \lambda C_d + X^T_{(d)} X_{(d)}$$



Streaming Tensor Analysis:



Major Impediments

- Characterizing the statistical nature of multi-modal real-world data sets.
- Data explosion, all ways.
- Thinking outside the walls: memory, power, and ILP.
- Simple isn't easy: identifying what constitutes high quality, actionable metrics.
- Suitable characteristic workflow abstractions.

Where We Want To Go

- Generate statistically relevant synthetic datasets while preserving the data integrity of the training sets.
  - Methodologies to explore: GANs and multi-fractal distros
- Understand the broad domain of applicability and the landscape of data transport tensor methods encompass.
- Cultivate a series of meta-kernels to encompass complete workflows so as to better serve the expanding application space.

Acknowledgements

Work partially funded by the NSF award 1849463 & The Laboratory for Physical Sciences.
Special thank you to the Graph500 executive committee and its members.

NSF

The Laboratory for Physical Sciences