SAND2022-7515C

Sandia National Laboratories

Exceptional service in the national interest

Low-Rank Tensor Decompositions for Large Sparse Count Data

Danny Dunlavy, Rich Lehoucq, Oscar López

ESCO 2022 - Applied Statistics and Data Science

SAND2022-YYYYP

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

U.S. DEPARTMENT OF ENERGY

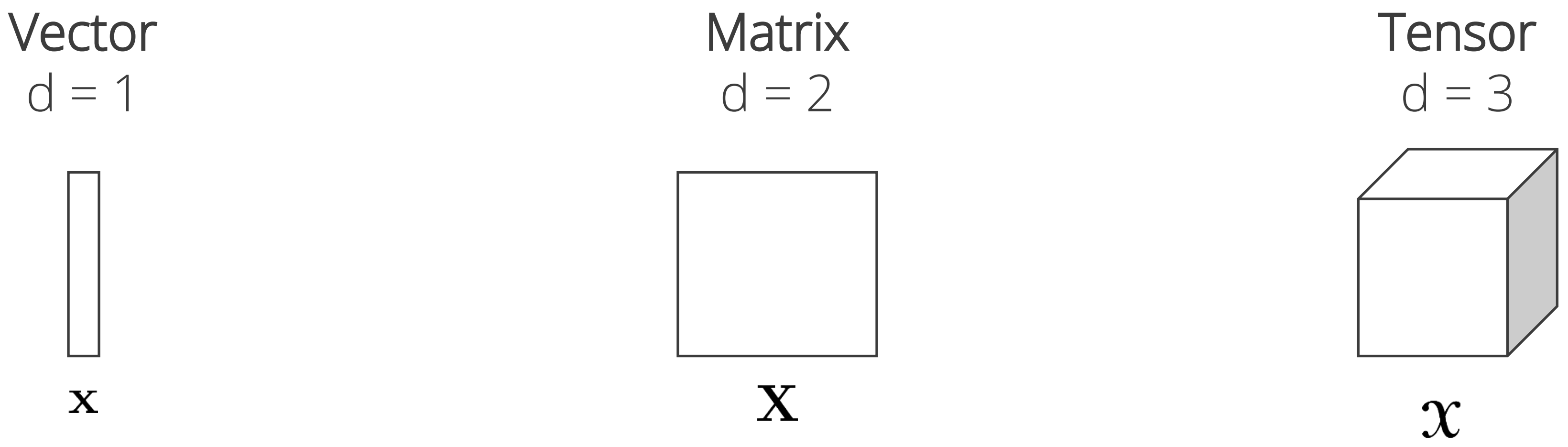
NNSA

Low-Rank Tensor Decompositions in Data Analysis

- What are they?
- Why are they useful?
- How much data is required to use them reliably?

Low-Rank Tensor Decompositions

Tensors: d-way Data Arrays

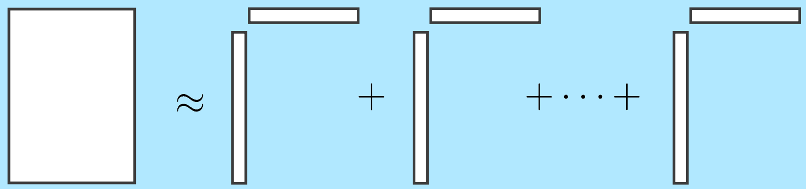


We refer to data arrays with 3 or more ways as *tensors.*

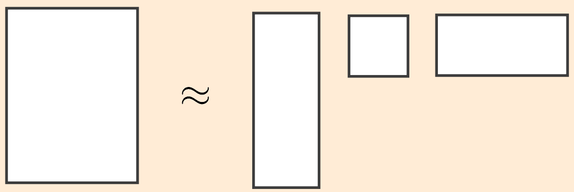
Low-Rank Decompositions: Two Points of View

Low-Rank Matrix Decompositions

Viewpoint 1: Sum of vector outer products, useful for interpretation



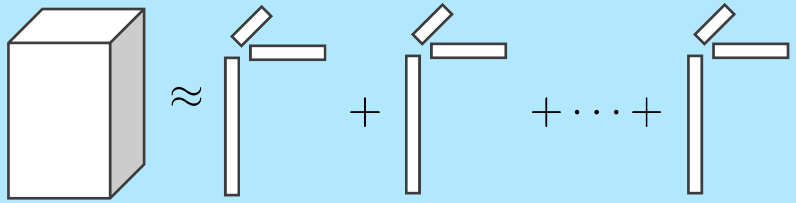
Viewpoint 2: High-variance subspaces, useful for compression



Singular value decomposition (SVD), eigendecomposition (EVD), nonnegative matrix factorization (NMF), etc.

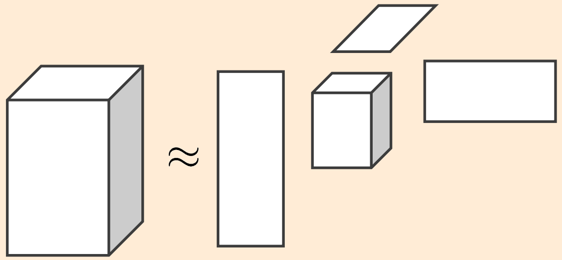
Low-Rank Tensor Decompositions

CP Model: Sum of d-way vector outer products, useful for interpretation



Canonical Polyadic, CANDECOMP, PARAFAC, CP

Tucker Model: Project onto high-variance subspaces to reduce dimensionality



HO-SVD, Best Rank-$(R_1, R_2, \ldots, R_d)$ decomposition

Other models for compression include hierarchical Tucker, tensor train, tensor ring, tensor network, etc.

Kolda and Bader (2009), Tensor Decompositions and Applications, https://doi.org/10.1137/07070111X

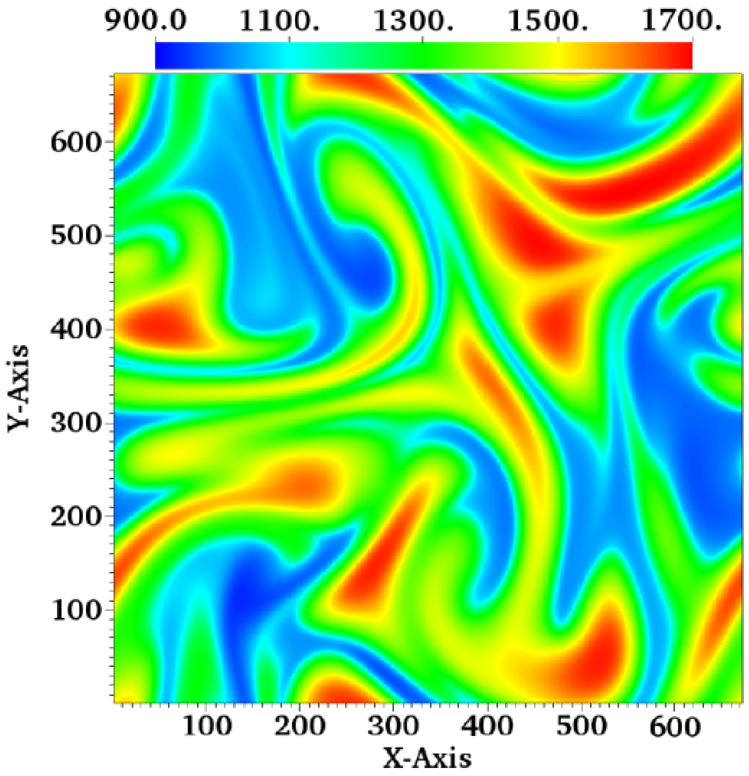
Low-Rank Decompositions: Benefits

- Unsupervised data models
- Reduced memory usage
- Noise reduction
- Identification of most important patterns and/or strongest signals in data
- Interpretability of complex data relationships

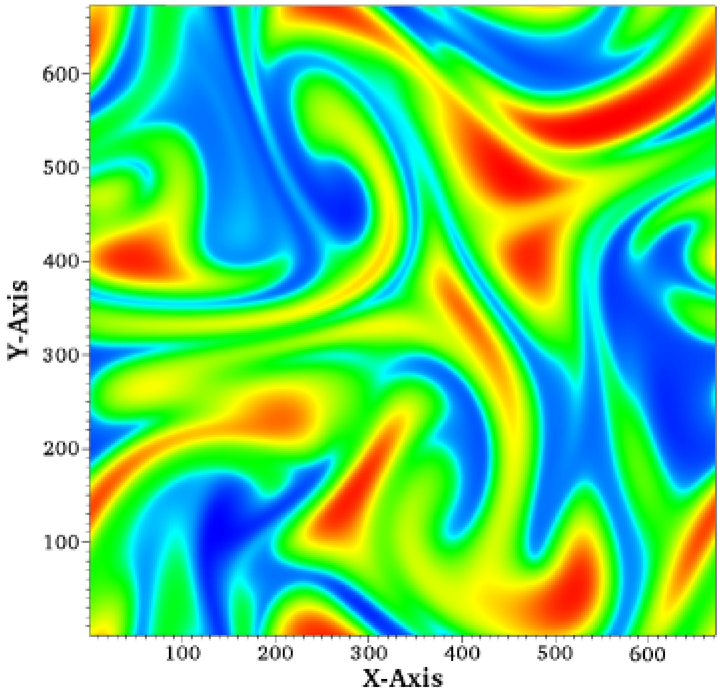
Examples of Tensor Decompositions

Tucker Decompositions: Scientific Computing Data Compression

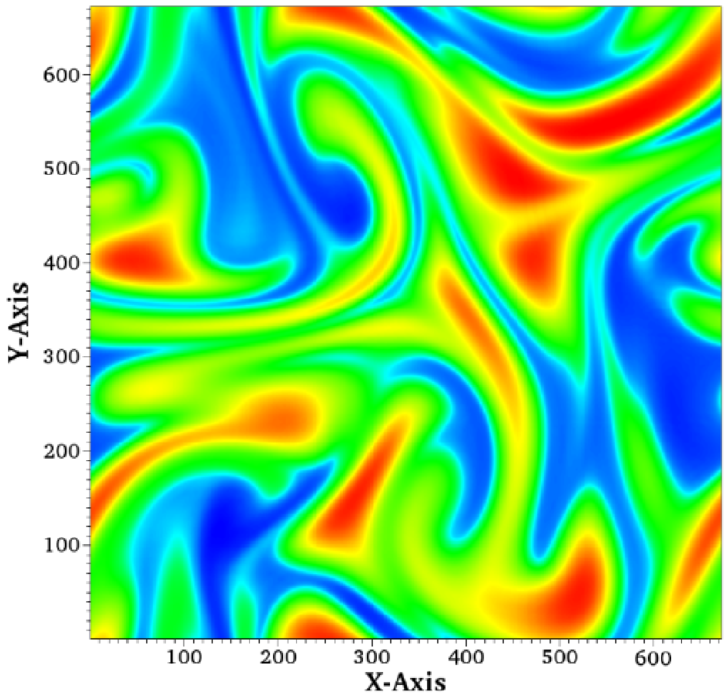
Contour plots of temperature (Kelvin) for combustion processes using simulation data



Original Data
(at one time instance)



Low-Rank Tensor Data
(~10X compression)

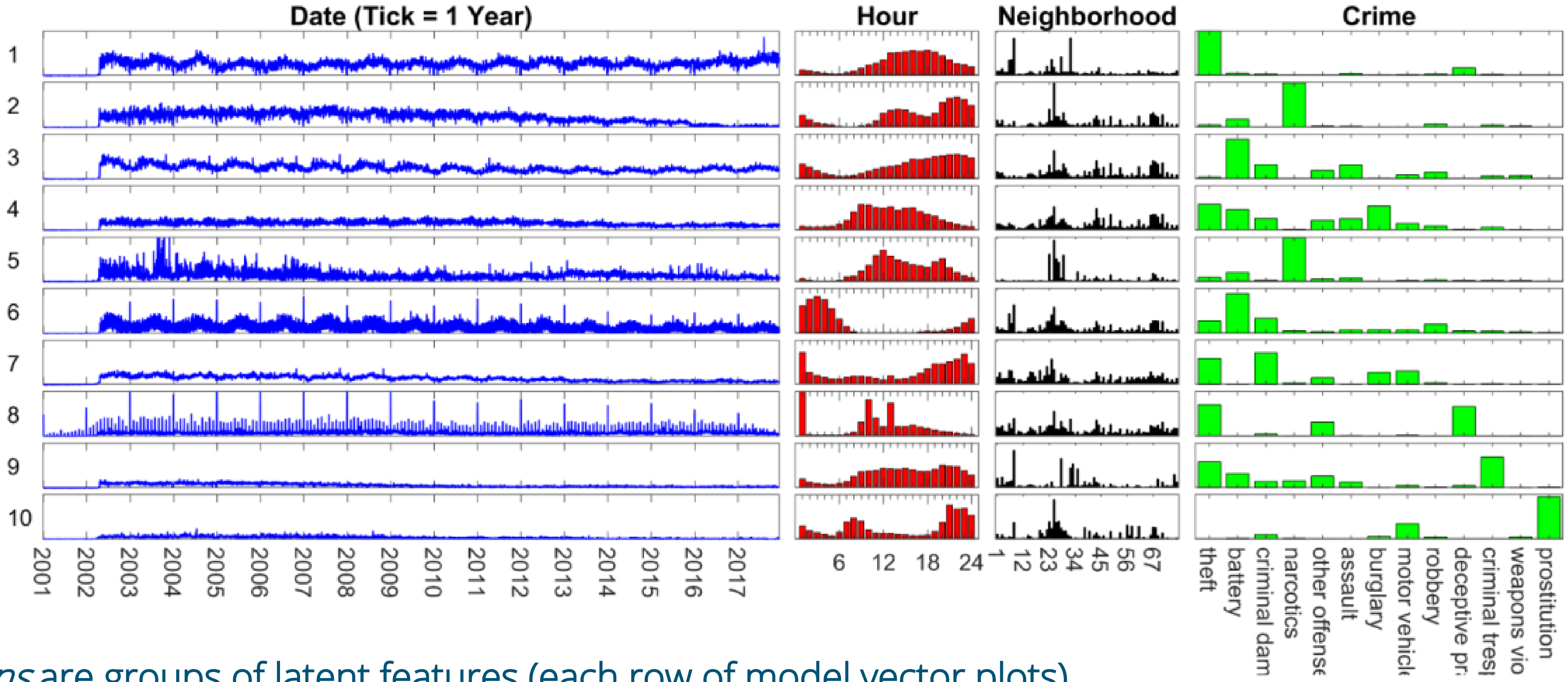


Low-Rank Tensor Data
(~700X compression)

Kolla, Aditya, and Chen (2020), Higher Order Tensors for DNS Data Analysis and Compression, https://doi.org/10.1007/978-3-030-44718-2_6

CP Decompositions: Extracting Patterns from Count Data

Crime reports in the city of Chicago, 2001-2017



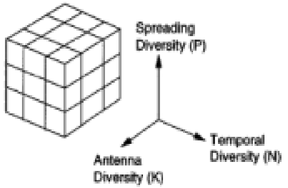
Patterns are groups of latent features (each row of model vector plots).

Kolda and Hong (2020), Stochastic Gradients for Large-Scale Tensor Decomposition, https://doi.org/10.1137/19M1266265

Low-Rank Tensor Decompositions: Numerous Other Applications

- Modeling fluorescence excitation-emission data (chemometrics)
- Signal processing
- Brain imaging (e.g., fMRI) data
- Network analysis and link prediction
- Image compression and classification; texture analysis
- Text analysis, e.g., multi-way LSI
- Approximating Newton potentials, stochastic PDEs, etc.
- Collaborative filtering
- Higher-order graph/image matching
- Neural network model compression

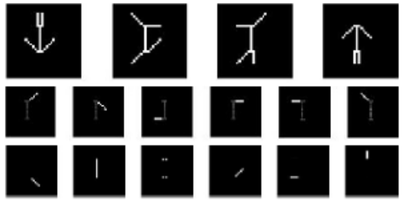
Furukawa, Kawasaki, Ikeuchi, and Sakauchi, *EGRW 2002*



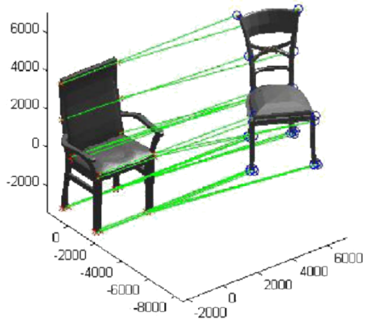
Sidiropoulos, Giannakis, Bro, *IEEE Trans. Signal Processing*, 2000

$$\mathcal{L}(x,t,\omega;u) = f(x,t,\omega) \quad (x,t)\in\mathcal{D}\times[0,T]$$
$$\mathcal{B}(x,t,\omega;u) = g(x,t) \quad (x,t)\in\partial\mathcal{D}\times[0,T]$$
$$\mathcal{I}(x,0,\omega;u) = h(x,\omega) \quad x\in\mathcal{D},$$

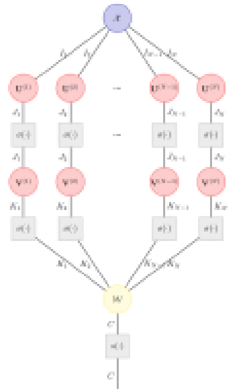
Doostan, Iaccarino, and Etemadi, *J. Computational Physics*, 2009



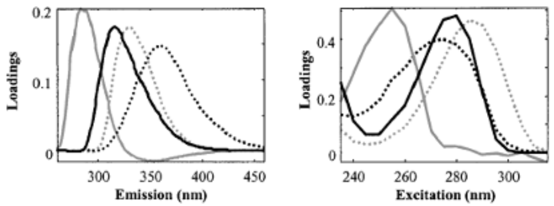
Hazan, Polak, and Shashua, *ICCV 2005*



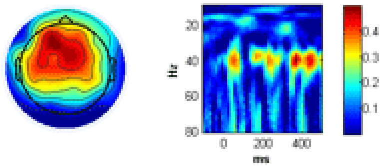
Duchenne, Bach, Kweon, Ponce, *TPAMI 2011*



Liu, Liu, Long, Zhu, in *Tensor Computation for Data Analysis*, 2022



Andersen and Bro, *J. Chemometrics*, 2003



ERPWAVELAB by Morten Mørup

Low-Rank Decompositions & Data Sampling

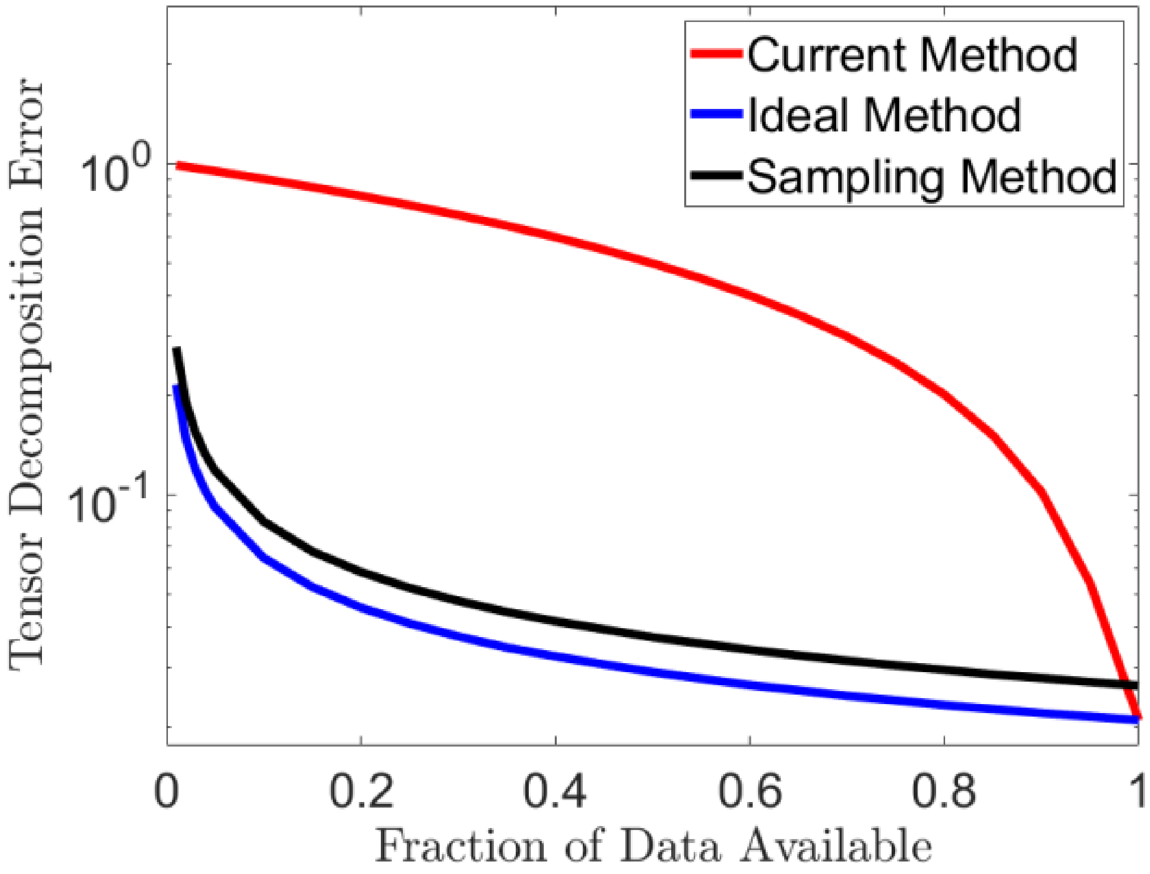
Large-Scale Data Analysis: Two Approaches

- Scale up computation

- Scale down data

CP Decompositions using Sampling for Sparse Count Data

- **Sparse data challenge:** determining which zeros are true values and which are placeholders in the data arrays
- **Our solution:** ignore zeros and fit tensor decompositions using only samples of non-zero values
- **Benefits:**
  - Better than assuming all zeros are true values (current methods)
  - No *a priori* knowledge of zeros needed (ideal method)
  - Can prove that only a small constant multiple of error will be incurred (our sampling method)



López, Dunlavy, and Lehoucq (2022), Zero-Truncated Poisson Regression for Zero-Inflated Multiway Count Data, [*https://doi.org/10.48550/arXiv.2201.10014*](https://doi.org/10.48550/arXiv.2201.10014)

Conclusions

- Many complex datasets can be modeled using low-rank tensor decompositions
- Low-rank decompositions can provide compression and interpretability of data
- Randomized tensor decompositions via data sampling can lead to great savings in terms of computation and memory usage at a modest cost in increased error
  - This is just the beginning of research in this area

Thank You

Low-Rank Tensor Decompositions for Large Sparse Count Data

Danny Dunlavy

dmdunla@sandia.gov