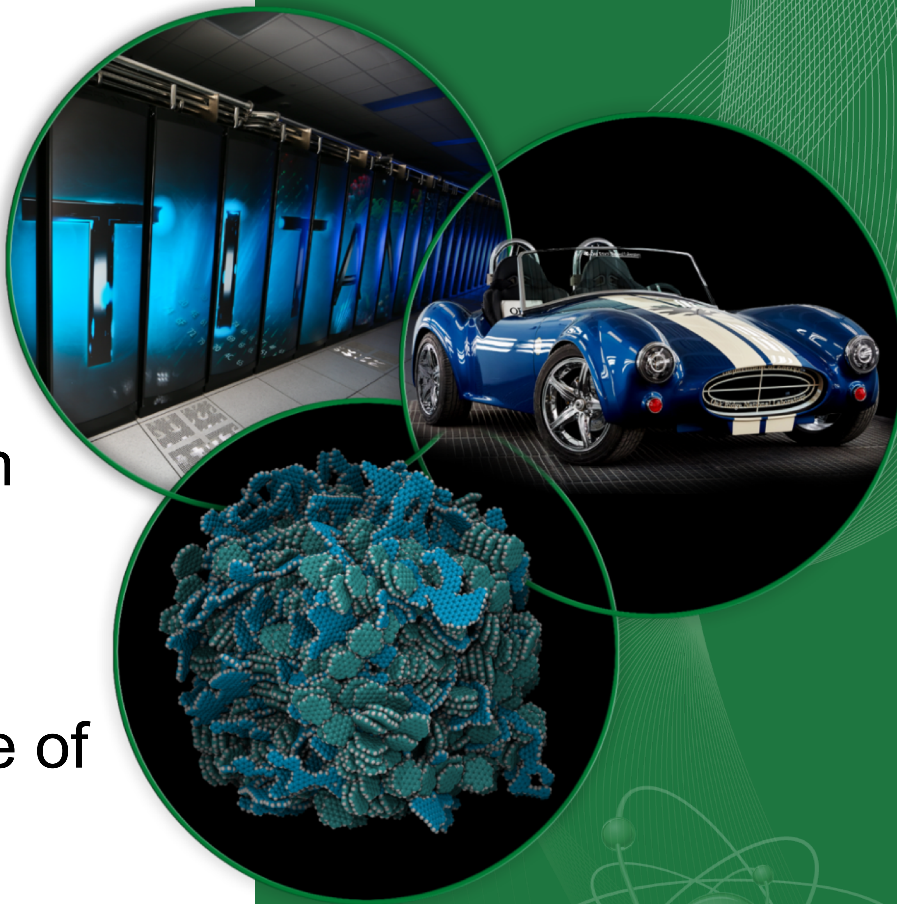
Distributed Nonnegative Tensor Low Rank Approximation for Large-Scale Clustering



Ramakrishnan(Ramki) Kannan

Grey Ballard (Wake Forest University)

Haesun Park (Georgia Institute of Technology)

Barry Drake (Georgia Tech Research Institue)

https://github.com/ramkikannan/

Acknowledgements

This manuscript has been co-authored by UT-Battelle, LLC under Contract No. DE-AC05-00OR22725 with the U.S. Department of Energy. This project was partially funded by the Laboratory Director's Research and Development fund. This research used resources of the Oak Ridge Leadership Computing Facility at the Oak Ridge National Laboratory, which is supported by the Office of Science of the U.S. Department of Energy.
This research used resources of the National Energy Research Scientific Computing Center, a DOE Office of Science User Facility supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.
Also, partial funding for this work was provided by AFOSR Grant FA9550-13-1-0100, National Science Foundation (NSF) grants IIS-1348152, ACI-1338745, ACI-1642410, and ACI-1642385, Defense Advanced Research Projects Agency (DARPA) XDATA program grant FA8750-12-2-0309. We also thank NSF for the travel grant to present this work in the conference through the grant CCF-1552229.

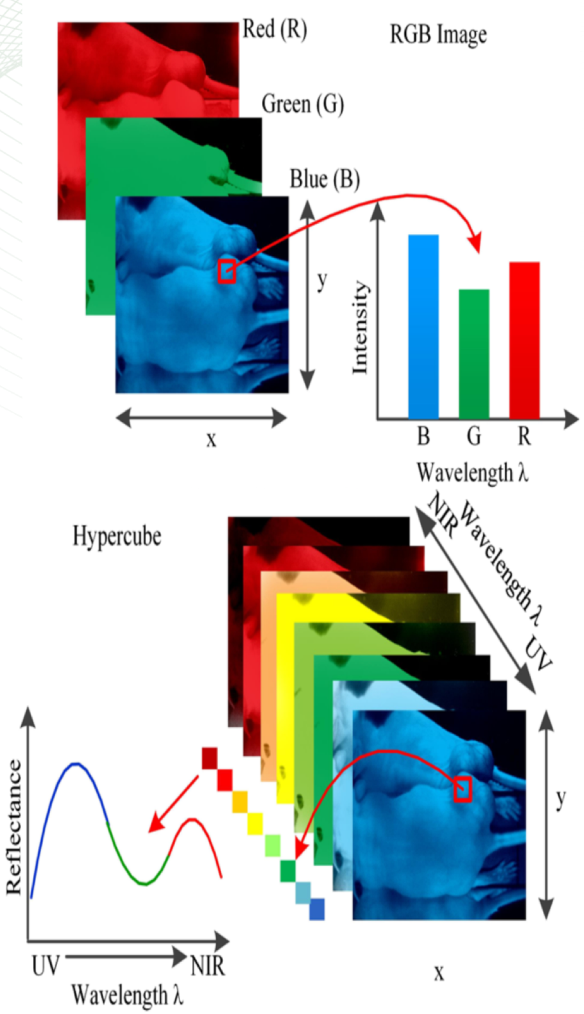
Agenda

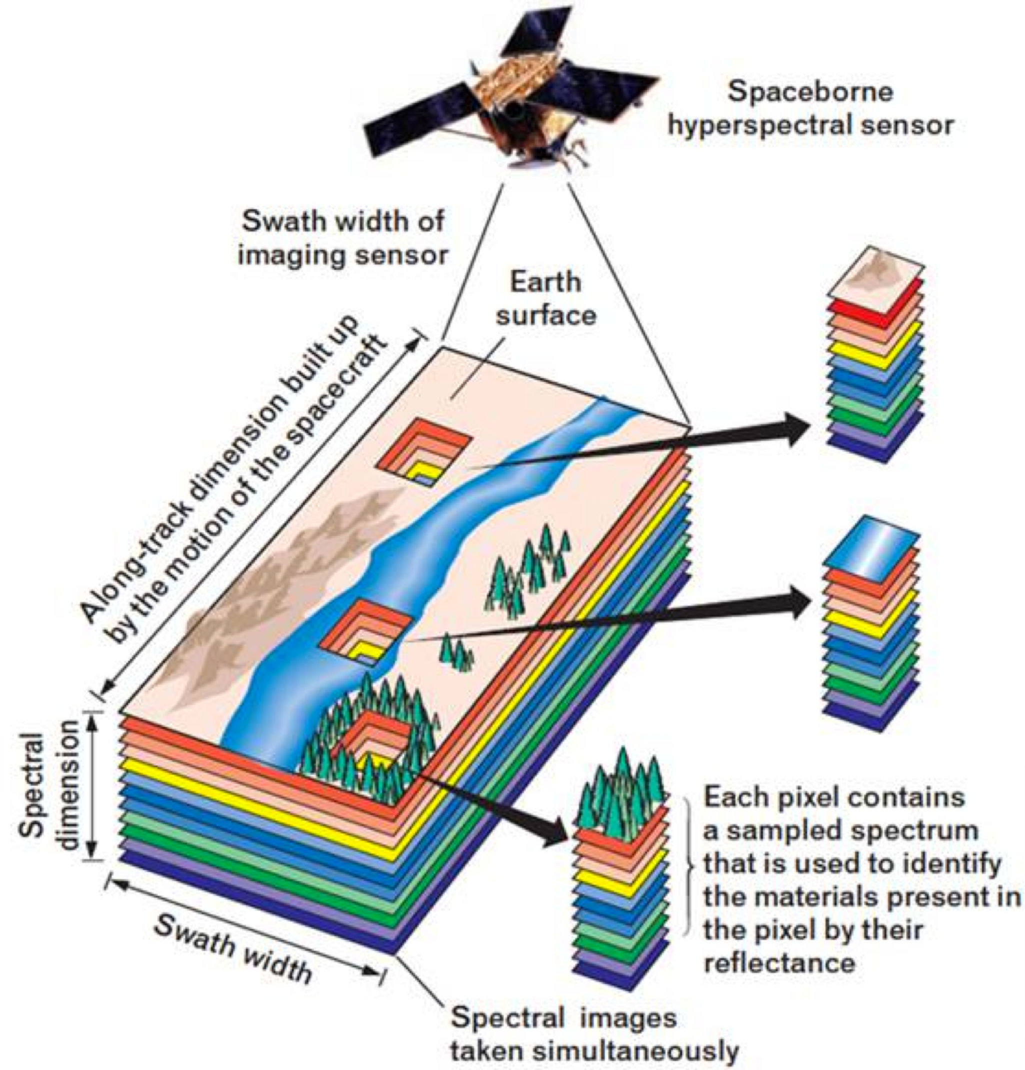
- Introduction and Motivation
- MPI-FAUN - Distributed NMF
  - Alternating-Updating NMF(AUNMF)
  - 1D Distribution
  - 2D Distribution
- NTF
  - Tensor Introduction and Operations
  - Distributed NTF

Motivation

- Observed features/collected metrics/independent variable/predictor cannot explain the dependent variable/response/outcome variable
- Eg., temperature, humidity, precipitation, etc. are insufficient to explain the probability to rain
- It is impossible to collect all the features that explain an outcome
- Sometimes, statistically significant latent features contained in the factors offer explanation

NHOT Illustration: Hyper Spectral Image



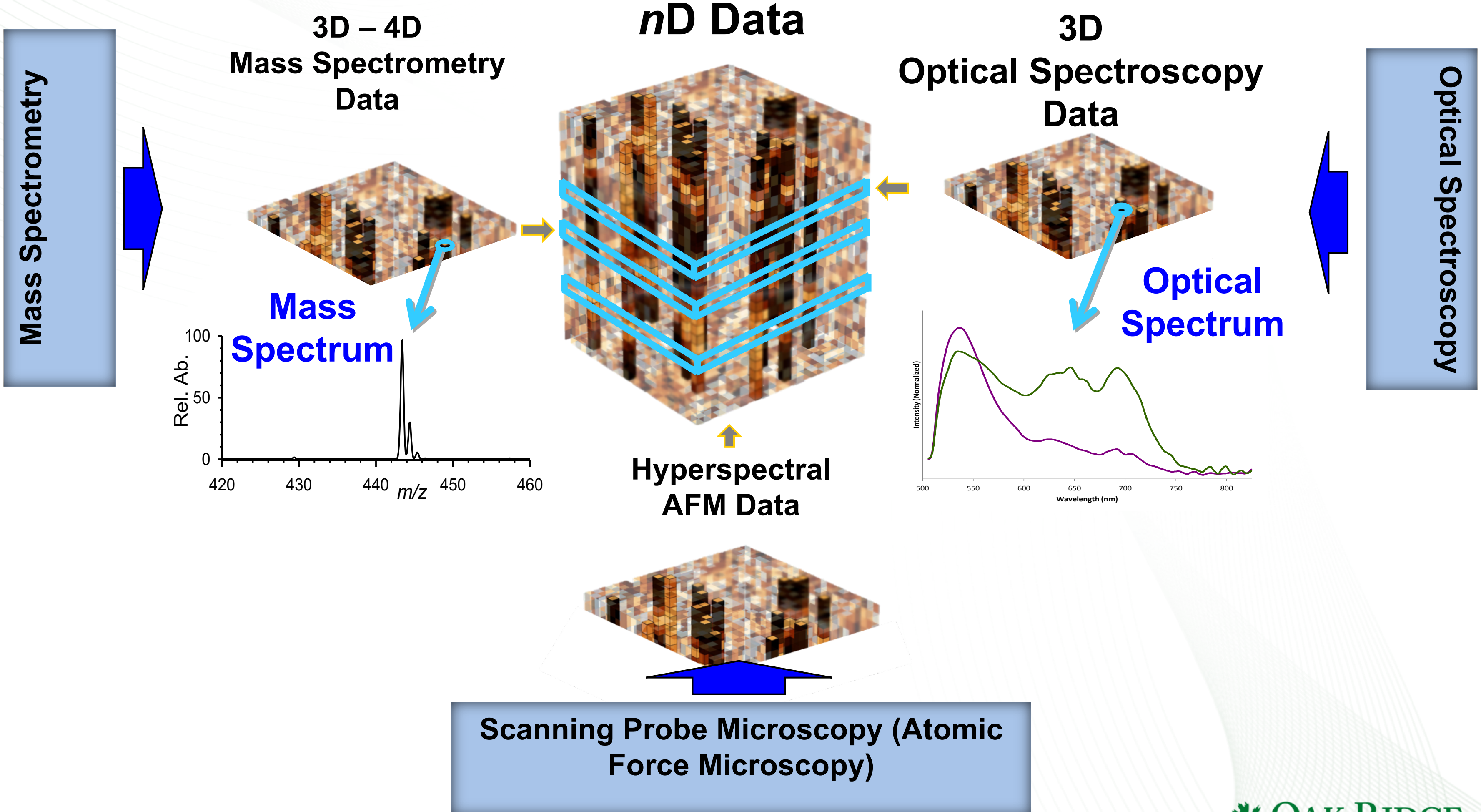


http://www.harrisgeospatial.com/Portals/0/blogs/imageryspeaks/USGS%20PRISM/BlogPost_Figure1.jpg

Lu G, Fei B; Medical hyperspectral imaging: a review. J. Biomed. Opt. 0001;19(1):010901. doi:10.1117/1.JBO.19.1.010901.

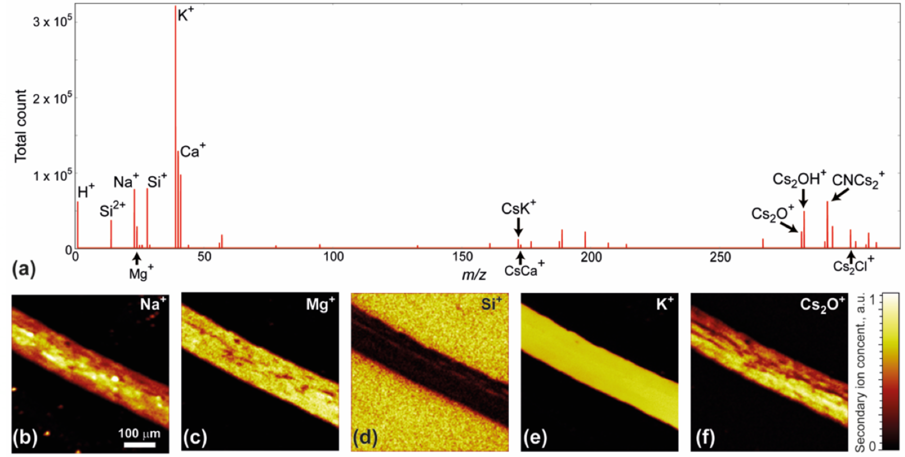
Dimensionality Reduction in Scientific Data

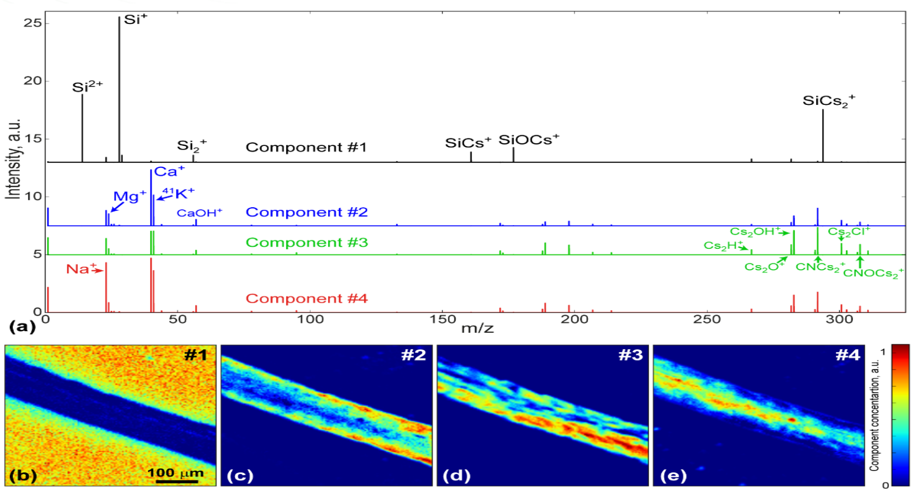
- Multimodal characterization of materials – *comprehensive characterization from chemical composition to functional properties on the nanoscale*



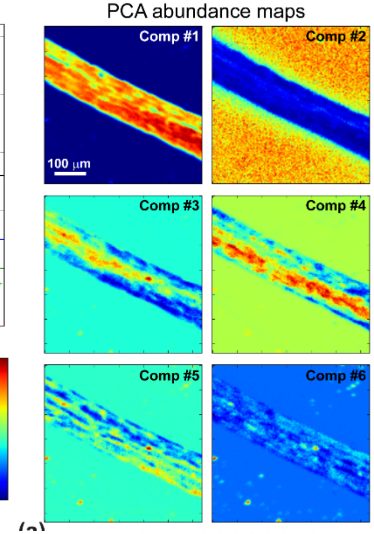
Thanks Ielev Anton and Sergei Kalinin

Example 1 : NMF vs. PCA

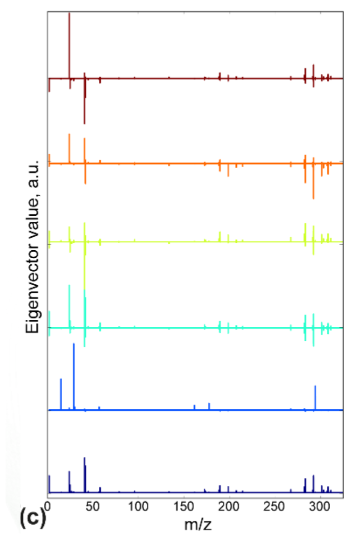




PCA abundance maps



PCA Eigen vectors

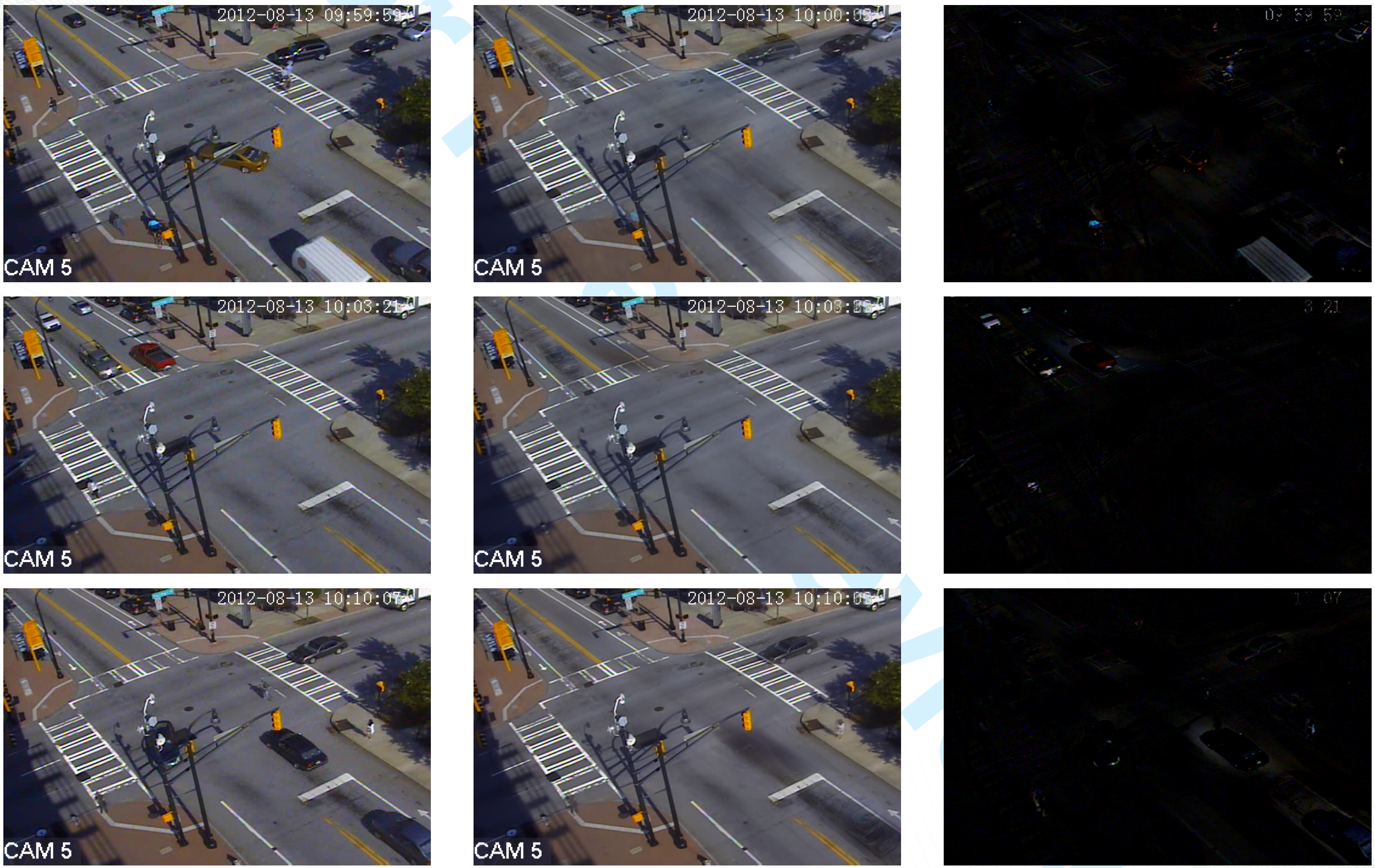


Both PCA and NMF are insufficient
They do not consider the neighbourhood information
To consider this information, we use regularization

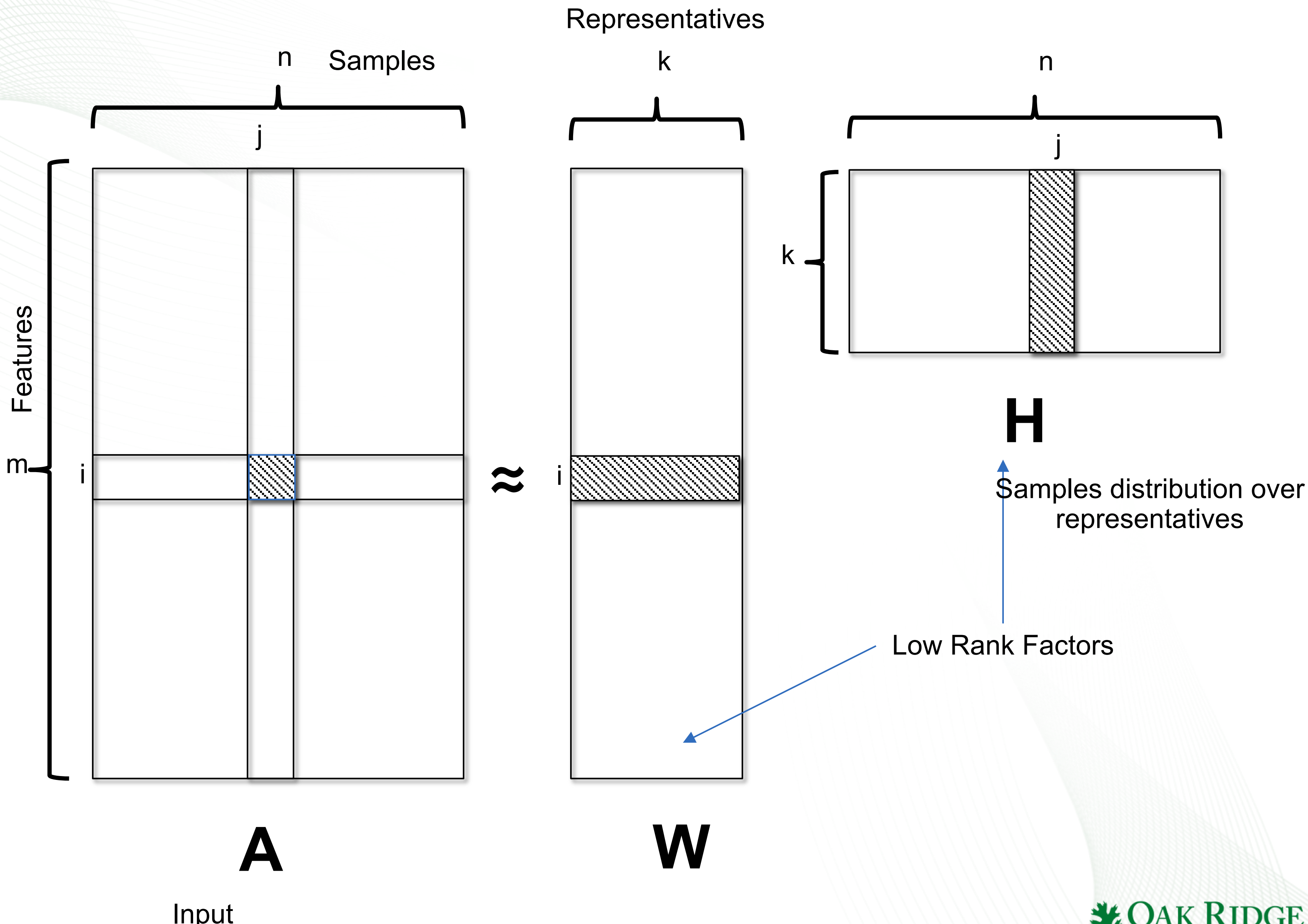
TOF SIMS Data – Collaboration w/ Ielev Anton

Example 2 : Video Data

Input Frame(**A**) | Background (**WH**) | Moving Object **A** – **WH**



Matrix Factorization (MF)



Oak Ridge National Laboratory

Alternating Updating NMF (AUNMF)

Given A, find W, H such that $\min_{W\geq 0, H\geq 0} ||A - WH||_F$

ANLS-BPP (Alternating NLS – Block Principal Pivoting)

$$\mathbf{W} \leftarrow \operatorname*{argmin}_{\tilde{\mathbf{W}}\geqslant 0} \left\| \mathbf{A} - \tilde{\mathbf{W}}\mathbf{H} \right\|_F ,$$

$$\mathbf{H} \leftarrow \operatorname*{argmin}_{\tilde{\mathbf{H}}\geqslant 0} \left\| \mathbf{A} - \mathbf{W}\tilde{\mathbf{H}} \right\|_F .$$

HALS (Hierarchical Alternating Least Squares)

$$\mathbf{w}^i \leftarrow \left[\mathbf{w}^i + \frac{(\mathbf{A}\mathbf{H}^T)^i - \mathbf{W}(\mathbf{H}\mathbf{H}^T)^i}{(\mathbf{H}\mathbf{H}^T)_{ii}} \right]_+$$

$$\mathbf{h}_i \leftarrow \left[\mathbf{h}_i + \frac{(\mathbf{W}^T\mathbf{A})_i - (\mathbf{W}^T\mathbf{W})_i\mathbf{H}}{(\mathbf{W}^T\mathbf{W})_{ii}} \right]_+$$

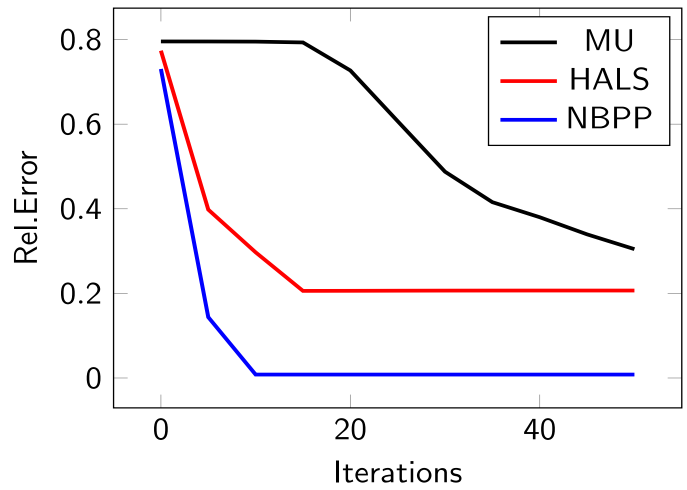
Multiplicative Update (MU)

$$w_{ij} \leftarrow w_{ij} \frac{(\mathbf{A}\mathbf{H}^T)_{ij}}{(\mathbf{W}\mathbf{H}\mathbf{H}^T)_{ij}} \quad h_{ij} \leftarrow h_{ij} \frac{(\mathbf{W}^T\mathbf{A})_{ij}}{(\mathbf{W}^T\mathbf{W}\mathbf{H})_{ij}}$$

AUNMF-Algorithm

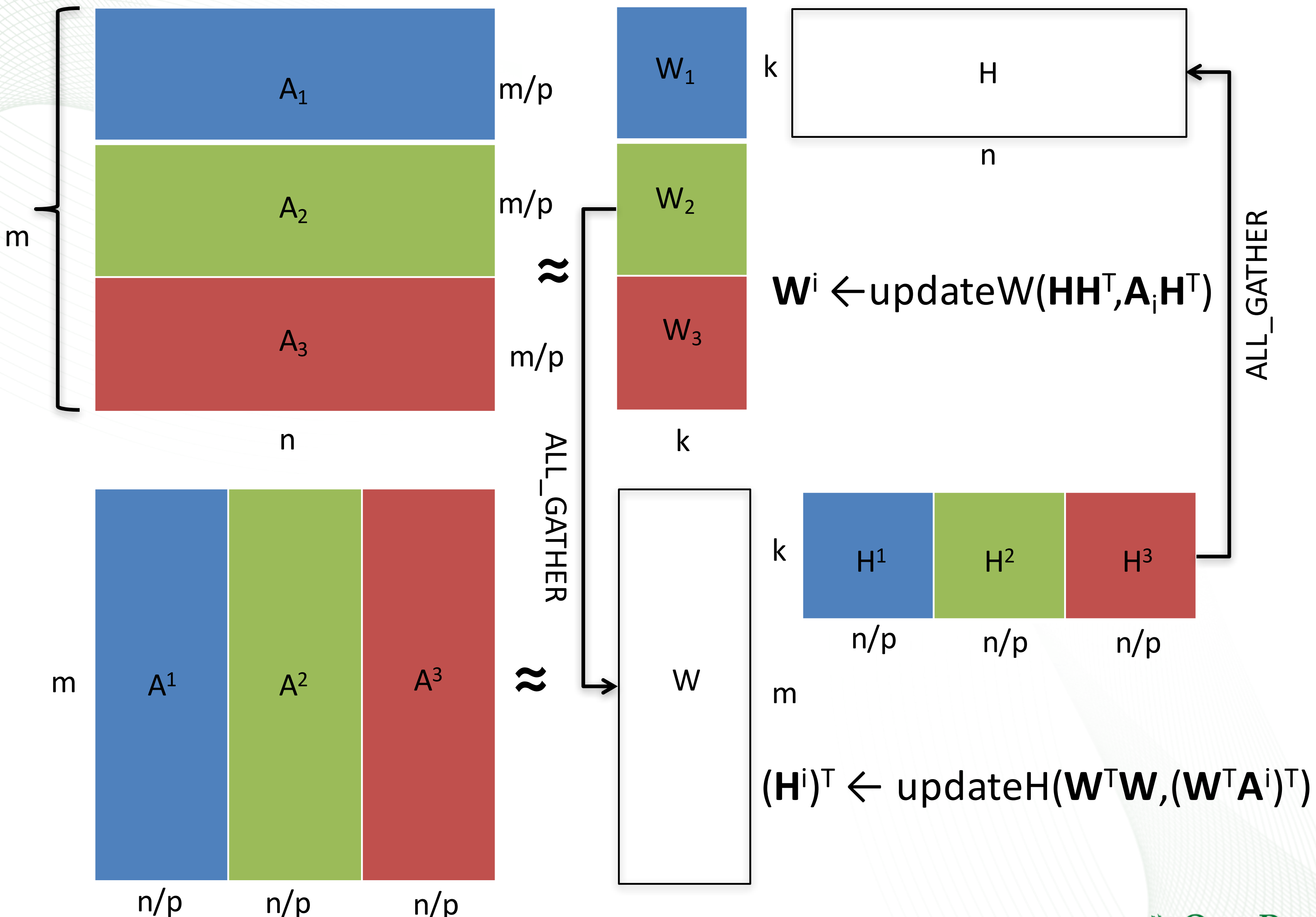
Require: $\mathbf{A}$ is an $m \times n$ matrix, k is rank of approximation

1: Initialize $\mathbf{H}$ with a non-negative matrix
2: **while** stopping criteria not satisfied **do**
3: Update $\mathbf{W}$ using $\mathbf{H}\mathbf{H}^T$ and $\mathbf{A}\mathbf{H}^T$
4: Update $\mathbf{H}$ using $\mathbf{W}^T\mathbf{W}$ and $\mathbf{W}^T\mathbf{A}$
5: **end while**



OAK RIDGE National Laboratory

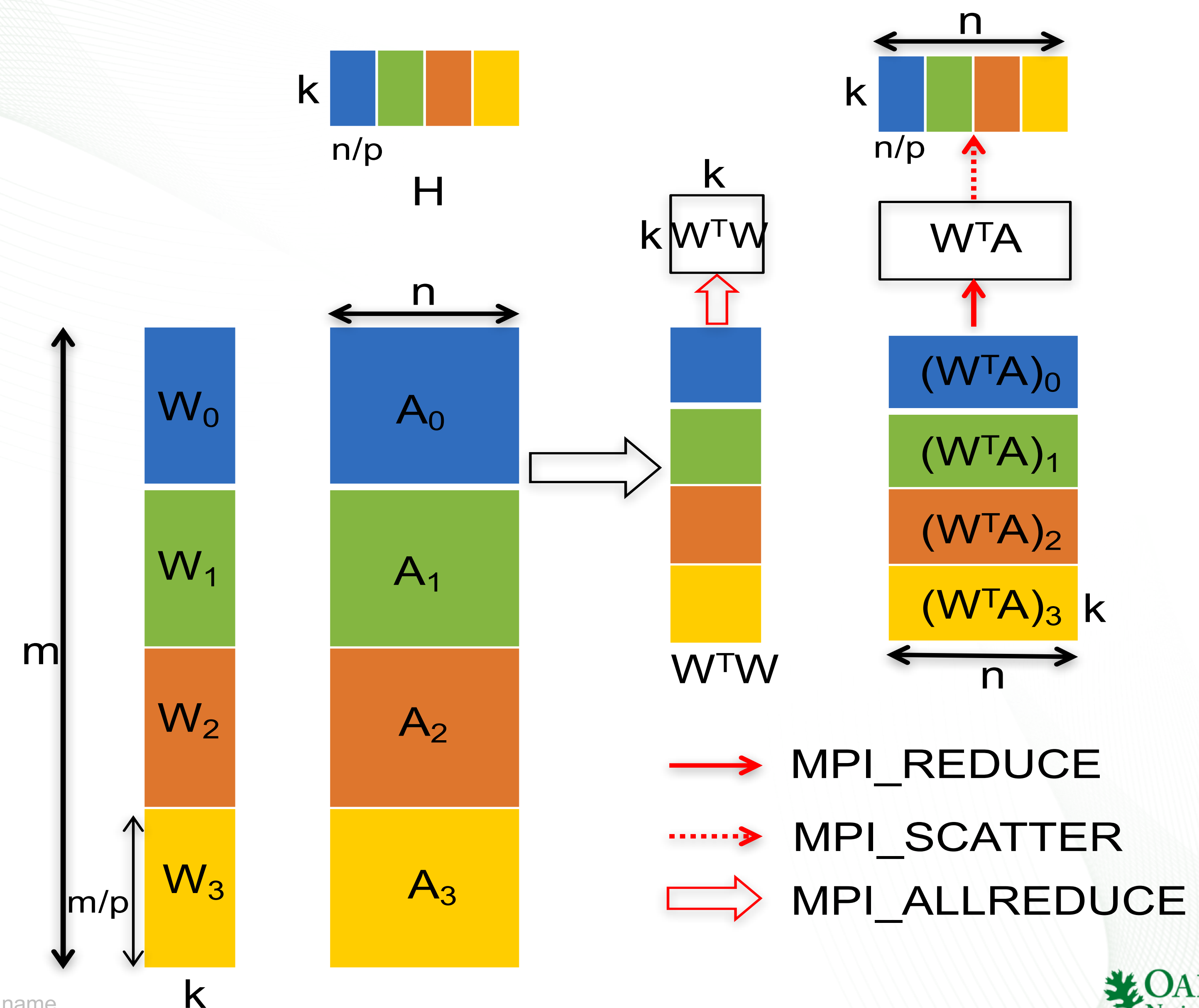
Naïve Parallel ANLS-BPP



MPI-FAUN

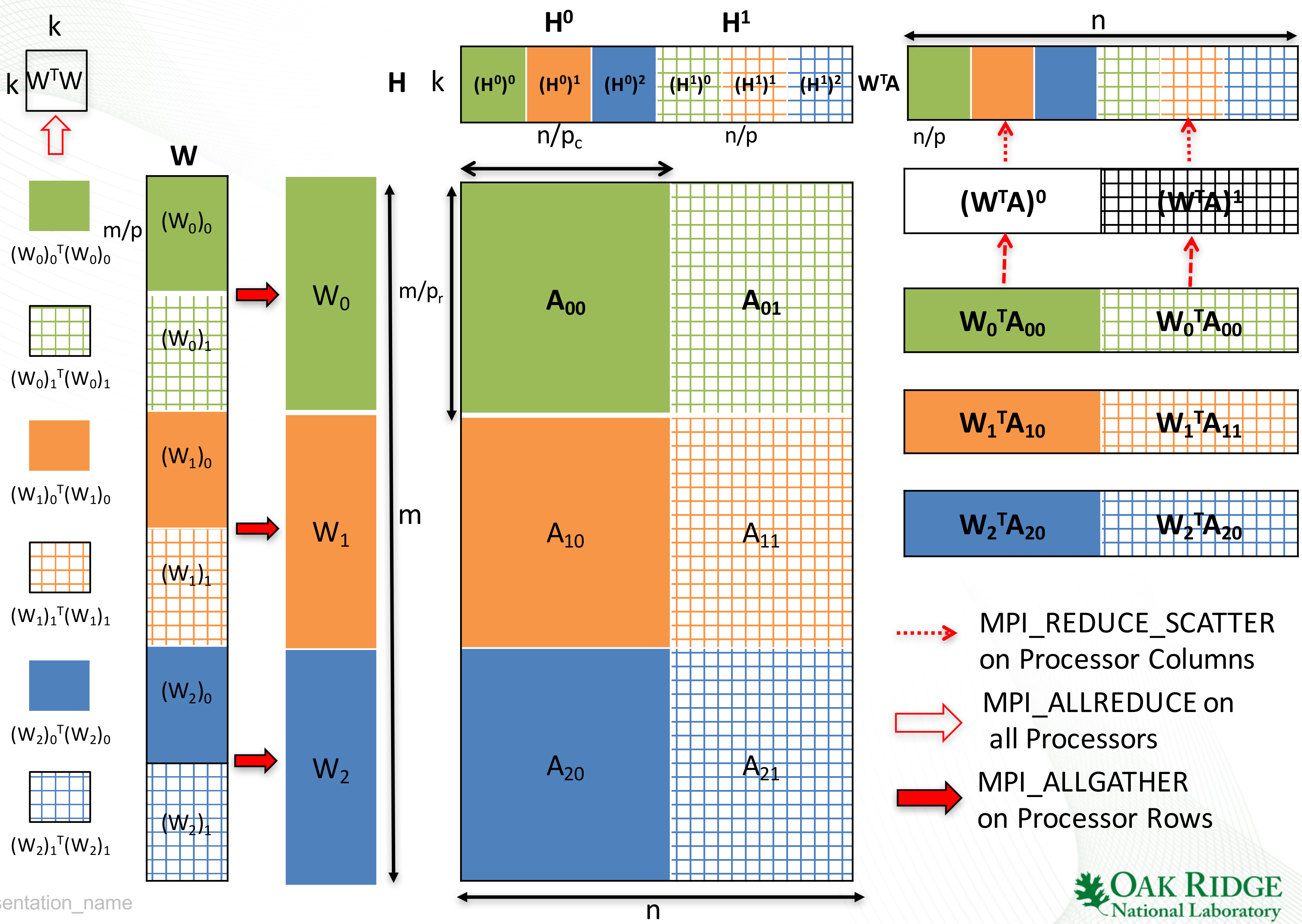
- Scalability is achieved by reducing the communication cost
- Intelligent tensor distribution so that entire computation happen in-situ
- Operations sequencing
- Collective MPI calls to reduce latency

1D NMF – Long and Thin matrices



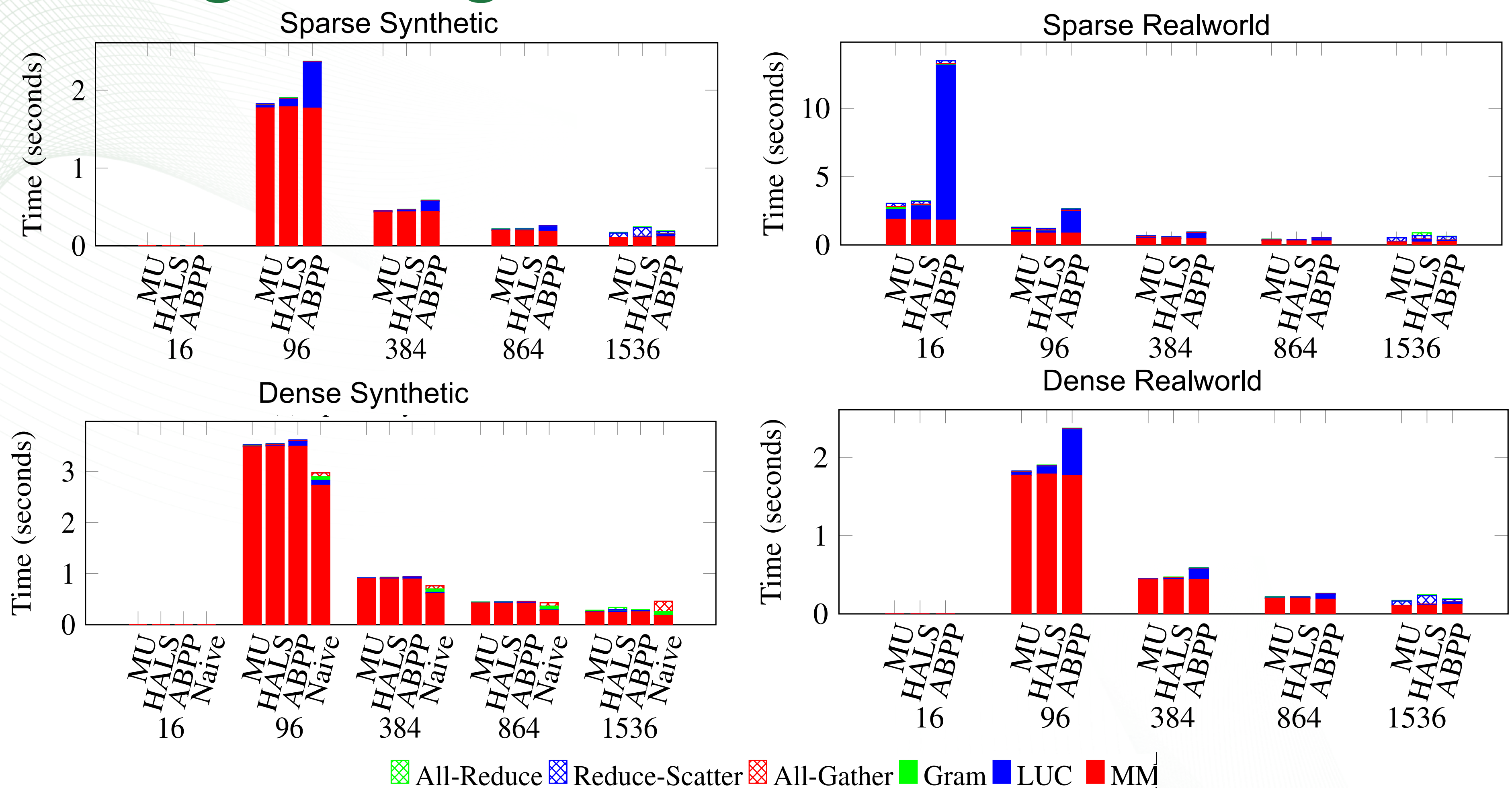
Oak Ridge National Laboratory

MPI-FAUN Framework



Oak Ridge National Laboratory

Strong Scaling



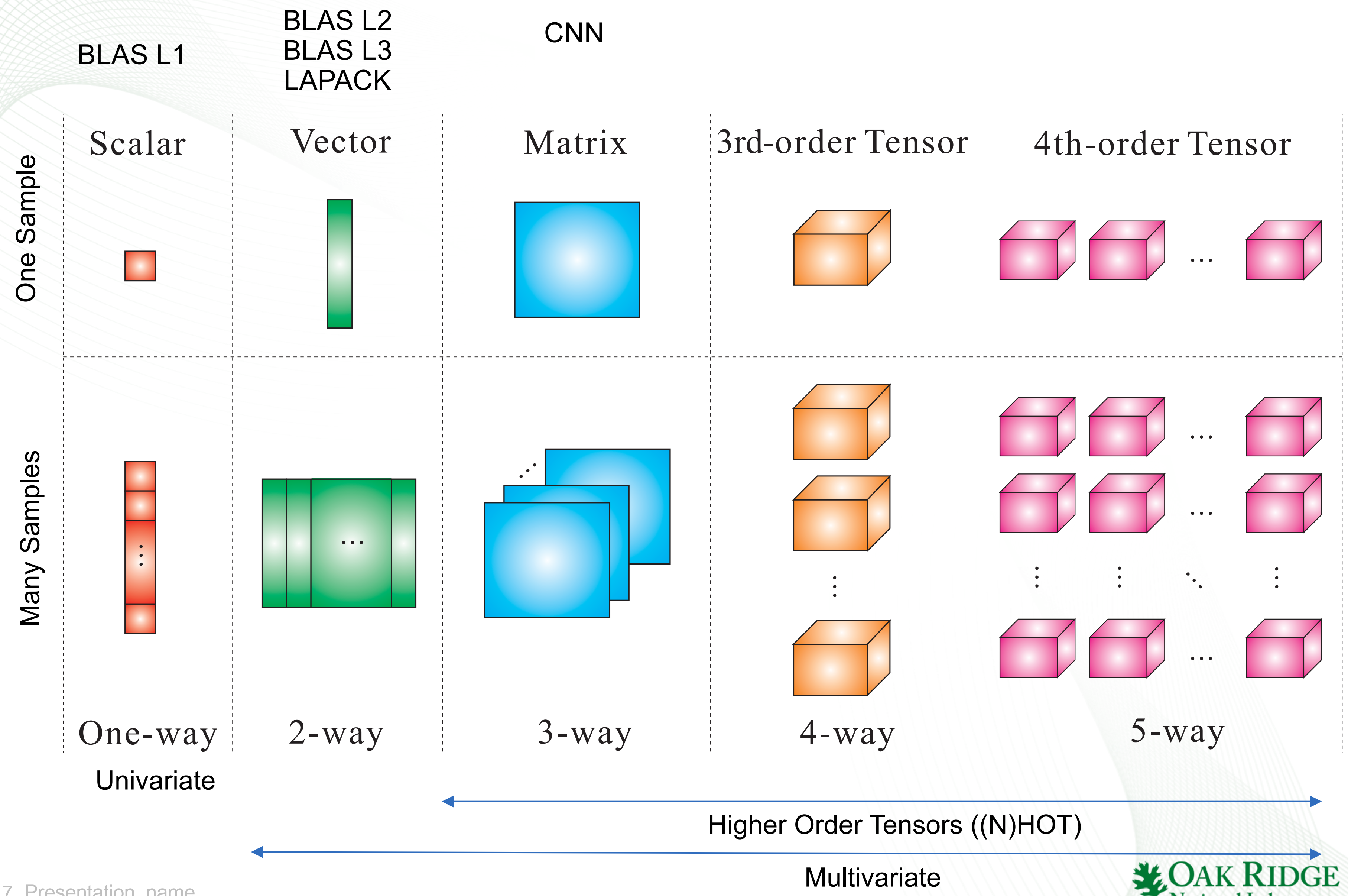
Dense/ Sparse Syn	207,360 × 138,240	Sparse Real world	1 million nodes, 3 million edges	Dense Real world	1,013,400 × 13,824 (12 min, 20 fps)

MPI-FAUN

- Distributed Communication avoiding NMF Algorithms
- https://github.com/ramkikannan/nmflibrary
- https://arxiv.org/abs/1609.09154
- Miniapp and benchmarked on OLCF Platforms

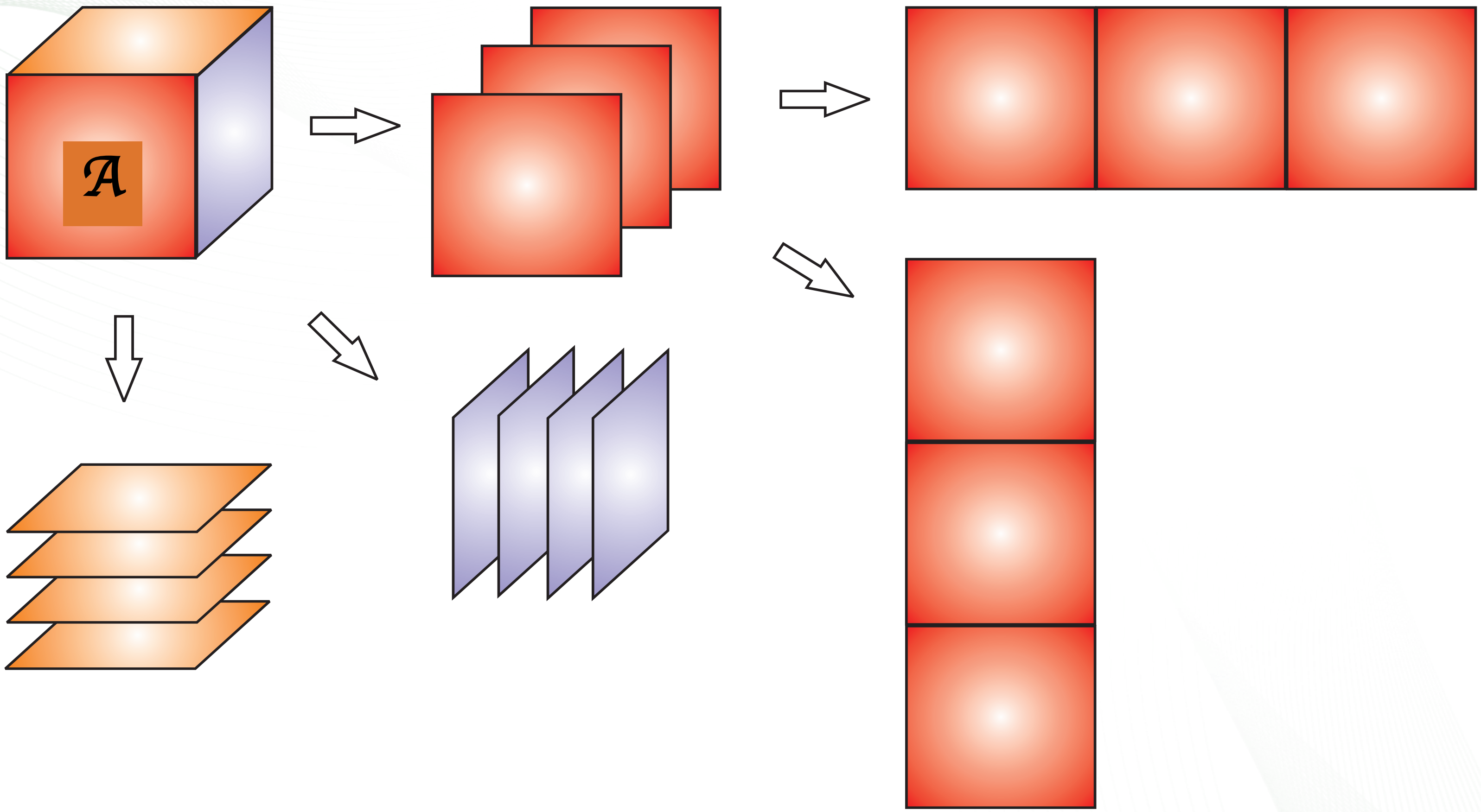
Dataset	Type	Matrix size	NMF Time
Video	Dense	1 Million x 13,824	5.73 seconds
Stack Exchange	Sparse	627,047 x 12 Million	67 seconds
Webbase-2001	Sparse	118 Million x 118 Million	25 minutes

Higher Order Tensors



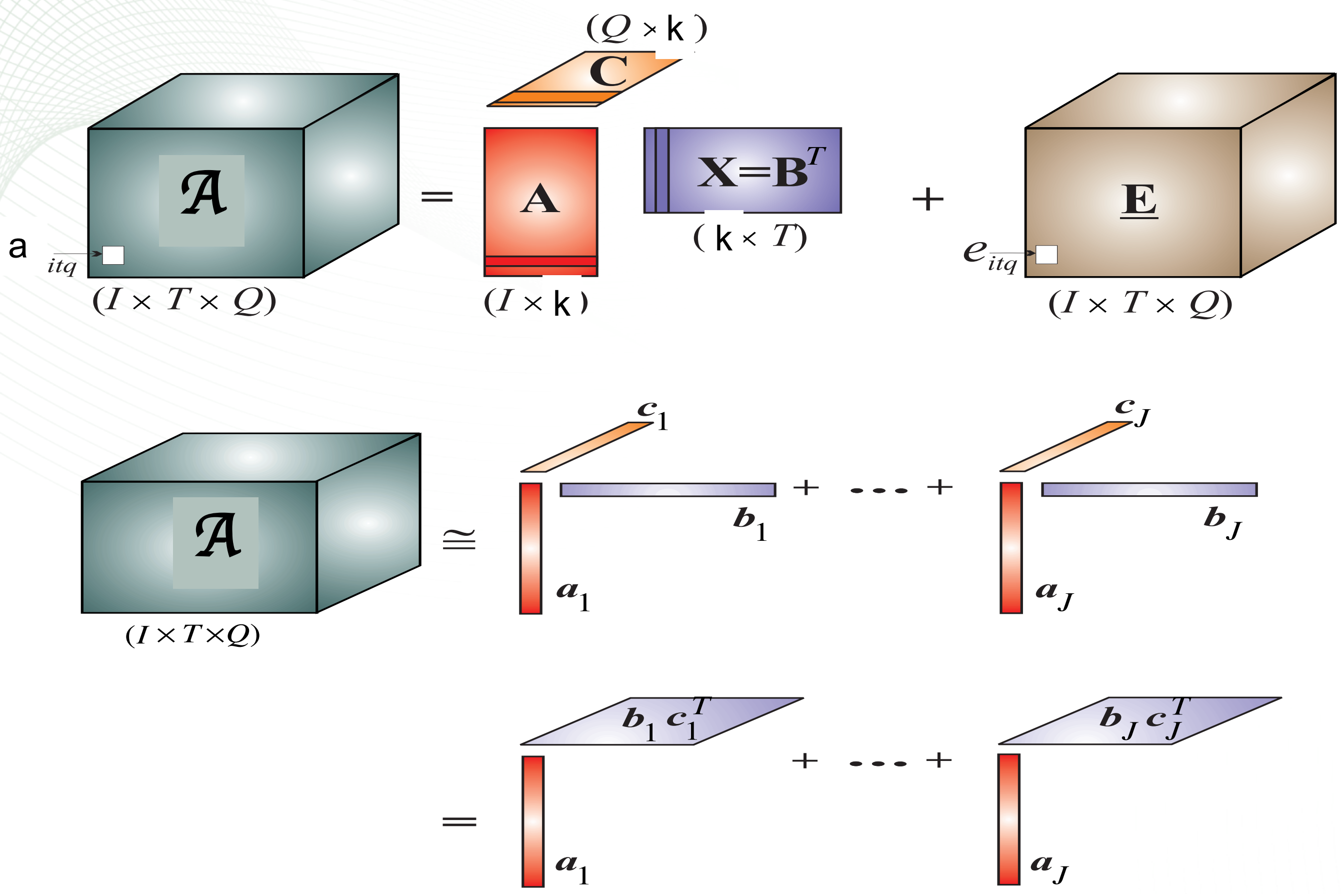
https://arxiv.org/abs/1609.00893v1

Existing DR for NHOT - Matricization



- Works only when some of the dimensions are independent
- Matricizing NHOT is non-trivial

Non-negative Tensor Factorization



Input

$\mathcal{A} \in \mathbb{R}^{M_1 \times \cdots \times M_N}$

Low Rank k

Output

A factor for every mode

$\mathbf{H}^{(1)}, \ldots, \mathbf{H}^{(N)}$

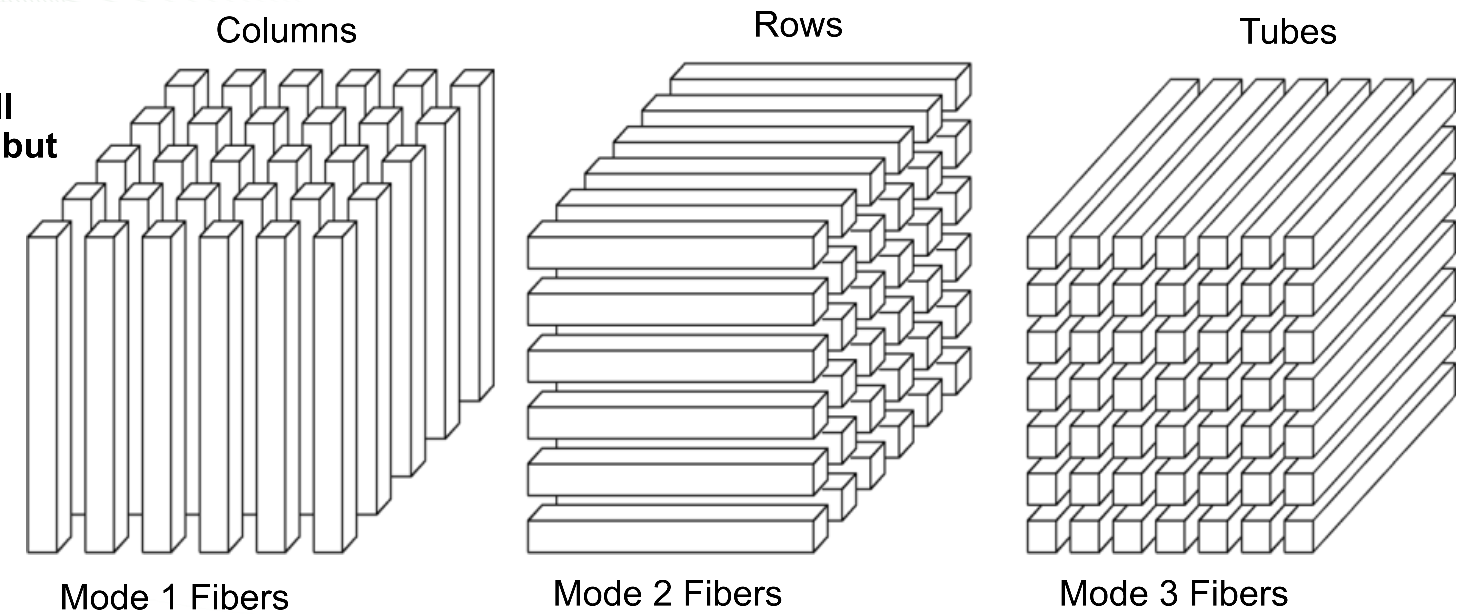
$\mathbf{H}^{(n)} \in \mathbb{R}^{M_n \times K}$

Novelty : Most of the tensor operations becomes infeasible on higher orders. Higher order tensors are going to be the defacto and we should be prepared with algorithms that can help us compute and interpret these higher order data.

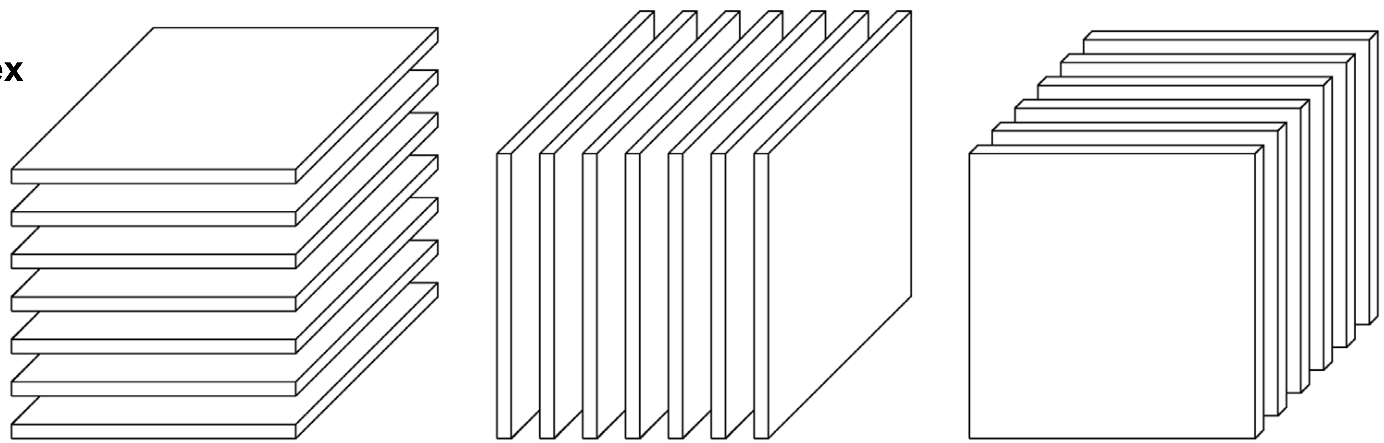
Cichocki, Andrzej, et al. *Nonnegative matrix and tensor factorizations: applications to exploratory multi-way data analysis and blind source separation*. John Wiley & Sons, 2009.

Fibers and Slices

Fix all indices but one



Fix one index



(a) Horizontal slices: $\mathbf{X}_{i::}$ (b) Lateral slices: $\mathbf{X}_{:j:}$ (c) Frontal slices: $\mathbf{X}_{::k}$ (or $\mathbf{X}_k$)

Thanks Bader and T. Kolda

Some tensor operations

Mode-n matricization: The mode-n matricization of $\mathcal{A} \in \mathbb{R}^{M_1 \times \cdots \times M_N}$, denoted by $\mathbf{A}^{<n>}$, is a matrix obtained by linearizing all the indices of tensor $\mathcal{A}$ except n. Specifically, $\mathbf{A}^{<n>}$ is a matrix of size $M_n \times (\prod_{\tilde{n}=1,\tilde{n}\neq n}^{N} M_{\tilde{n}})$, and the $(m_1, \ldots, m_N)$th element of $\mathcal{A}$ is mapped to the (m_n, J)th element of $\mathbf{A}^{<n>}$ where

$$J = 1 + \sum_{j=1}^{N} (m_j - 1) J_j \text{ and } J_j = \prod_{l=1, l\neq n}^{j-1} M_l .$$

Khatri-Rao product: The Khatri-Rao product of two matrices $\mathbf{A} \in \mathbb{R}^{J_1 \times L}$ and $\mathbf{B} \in \mathbb{R}^{J_2 \times L}$, denoted by $\mathbf{A} \odot \mathbf{B} \in \mathbb{R}^{(J_1 J_2) \times L}$, is defined as

$$\mathbf{A} \odot \mathbf{B} = \begin{bmatrix} a_{11}\mathbf{b}_1 & a_{12}\mathbf{b}_2 & \cdots & a_{1L}\mathbf{b}_L \\ a_{21}\mathbf{b}_1 & a_{22}\mathbf{b}_2 & \cdots & a_{2L}\mathbf{b}_L \\ \vdots & \vdots & \ddots & \vdots \\ a_{J_1 1}\mathbf{b}_1 & a_{J_1 2}\mathbf{b}_2 & \cdots & a_{J_1 L}\mathbf{b}_L \end{bmatrix} .$$

NMF vs NTF

NMF	NTF
$\min_{W\geq 0, H\geq 0} \|\|A - WH\|\|_F^2$	$\min_{H^{(i)}\geq 0} \|\|A - [\![H^{(1)}, \ldots, H^{(n)}]\!]\|\|_F^2$ $\forall i = 1, \ldots, n$
$\mathbf{H} \leftarrow \operatorname{argmin}_{\tilde{\mathbf{H}}\geqslant 0} \left\|\mathbf{A} - \mathbf{W}\tilde{\mathbf{H}}\right\|_F$	$\mathbf{H}^{(n)} \leftarrow \arg\min_{\mathbf{H}\geq 0} \left\|\mathbf{B}^{(n)}\mathbf{H}^T - (\mathbf{A}^{<n>})^T\right\|_F^2.$
	$\mathbf{B}^{(n)} = \mathbf{H}^{(N)} \odot \cdots \odot \mathbf{H}^{(n+1)} \odot \mathbf{H}^{(n-1)} \odot \cdots \odot \mathbf{H}^{(1)} \in \mathbb{R}^{\left(\prod_{\tilde{n}=1,\tilde{n}\neq n}^{N} M_{\tilde{n}}\right)\times K}.$ Khatri-Rao Prod
$(\mathbf{H}^i)^T \leftarrow$ updateH$(\mathbf{W}^T\mathbf{W},(\mathbf{W}^T\mathbf{A}^i)^T)$	$\left(\mathbf{B}^{(n)}\right)^T \mathbf{B}^{(n)} = \bigotimes_{\tilde{n}=1,\tilde{n}\neq n}^{N} \left(\mathbf{H}^{(\tilde{n})}\right)^T \mathbf{H}^{(\tilde{n})},$
	$\mathbf{B}^{(n)T} \left(\mathbf{A}^{<n>}\right)^T$ - MTTKRP

Oak Ridge National Laboratory

Distributed NCP Algorithm

- N-D Process Grid for N modes $P_1 \times \cdots \times P_N$
- Input Tensor is distributed as $\mathcal{A}_{p_1 \cdots p_N}$ is $(M_1/P_1) \times \cdots \times (M_N/P_N)$
- Factors are all_gathered as $\mathbf{H}^{(i)}_{p_i}$ is $(M_i/P_i) \times k$

that is redundant across $(\star, \ldots, \star, p_i, \star, \ldots, \star)$, for $1 \leqslant i \leqslant N$

- $\mathbf{U} = \text{Local-SYRK}(\mathbf{H}^{(i)}_{\mathbf{p}})$ where $\mathbf{H}^{(i)}_{\mathbf{p}}$ of dimensions $(M_i/P) \times k$
- $\mathbf{G}^{(i)} = \text{All-Reduce}(\mathbf{U}, (\star, \ldots, \star))$
- $\mathbf{S} = \circledast_{n \neq i} \mathbf{G}^{(i)}$
- $\mathbf{V} = \text{Local-MTTKRP}(\mathcal{A}_{p_1 \cdots p_N}, \{\mathbf{H}^{(n)}_{p_n}\}, i)$
- $\mathbf{W} = \text{Reduce-Scatter}(\mathbf{V}, (\star, \ldots, \star, p_i, \star, \ldots, \star))$
- Compute $\mathbf{H}^{(i)}_{\mathbf{p}}$ from **S** and **W** using local NLS

Conclusion and Future works

- Conclusion
  - MPI-FAUN
  - Distributed NTF
- Future work
  - Benchmarking on very large datasets
  - Optimal Communication
  - Interpretation for scientific datasets
  - Sparse Tensor with Hypergraph