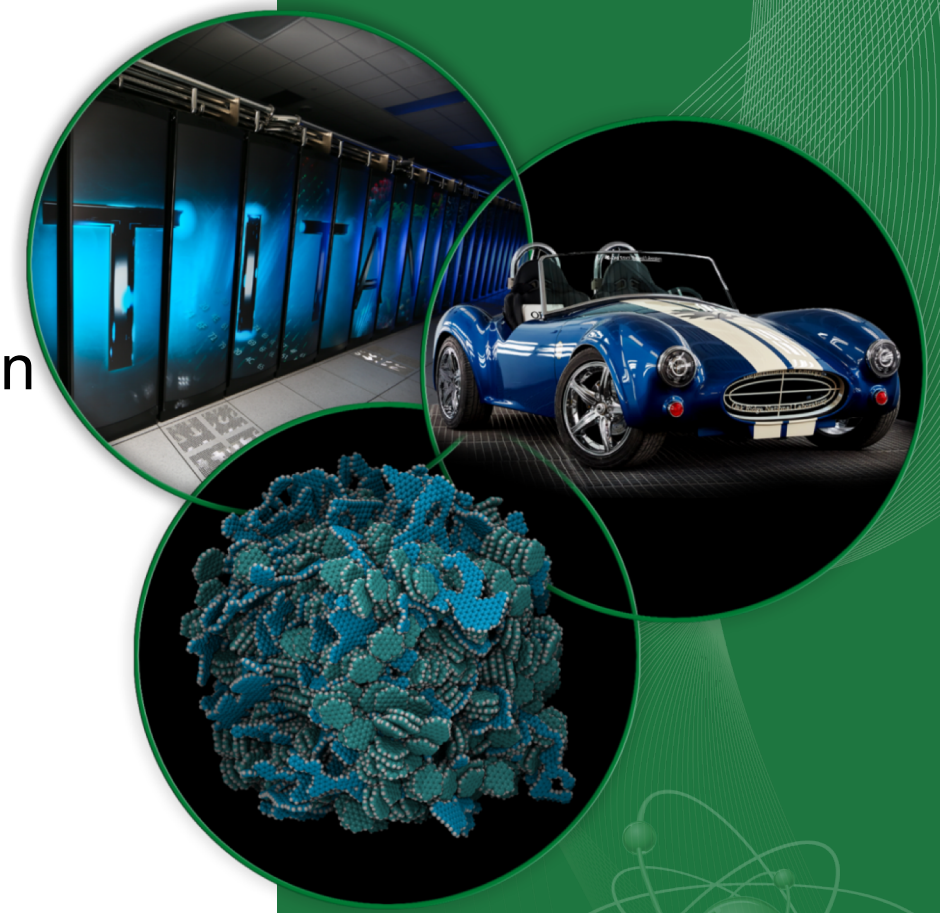
Non-negative Matrix and Tensor Factorization

Ramakrishnan(Ramki) Kannan



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.

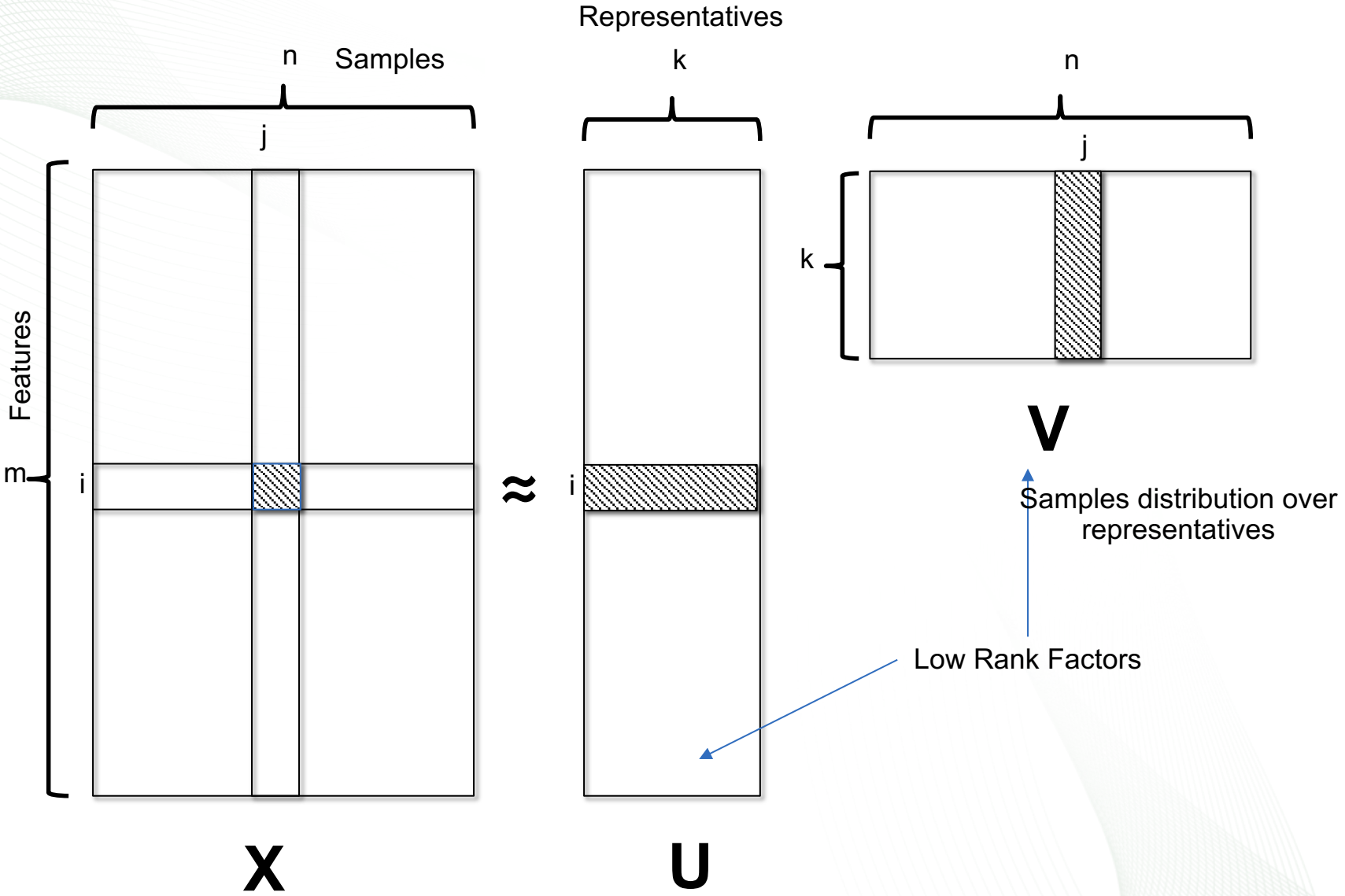
Agenda

- Research
  - Introduction to Matrix Factorization
  - Scientific and Internet Applications
  - NMF on HPC Platforms
  - Higher order NTF
  - Deep learning and NTF

Acknowledgements

- Collaborators
  - Internal - Seung-Hwan, Arvind, Maxim, Rama Vasudevan, Anton Ievlev, John Harney, Dale Stansberry, David Hughes, Hoony Park, Srikanth, Dmitry, Maxim Ziatdinov
  - External – Prof. Haesun Park (GATech), Prof. Grey Ballard(Wake Forest), Sheikh Ghafoor (TNTech)

Matrix Factorization (MF)



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

Supervised Learning

Regression

Bedrooms	Area	Baths	…	School rating	Price

Features

Labels

Source	Destn	Language	…	Num words	Spam/Not spam

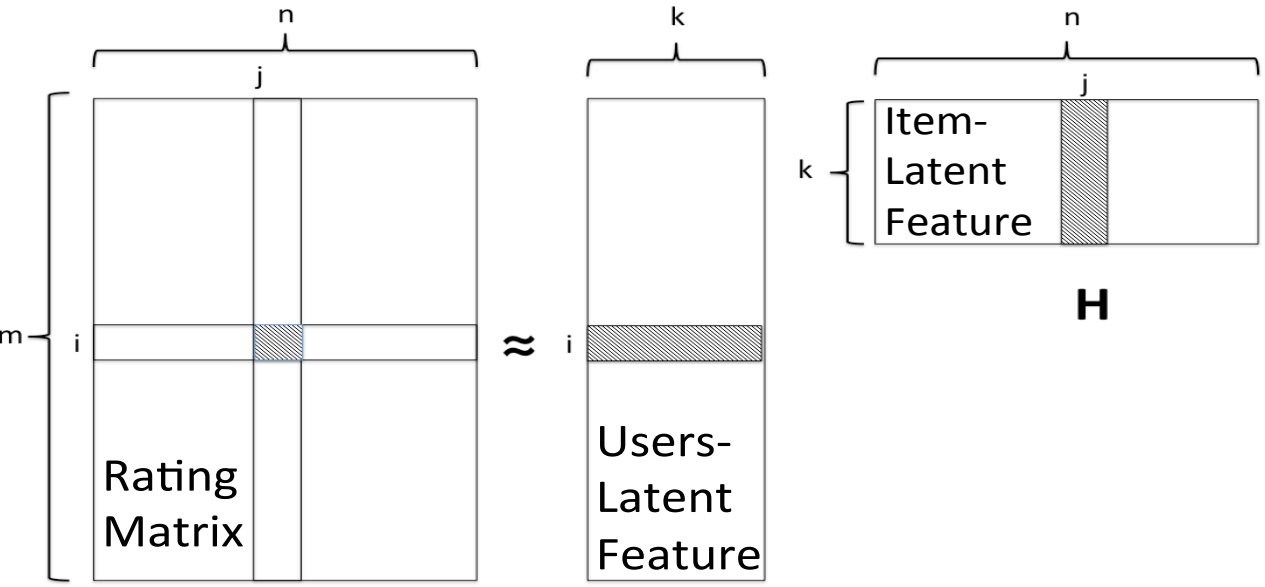
Classification

Matrix Factorization

Rating Matrix

	Item 1	Item 2	…	Item n
User 1				
User 2				
User m				

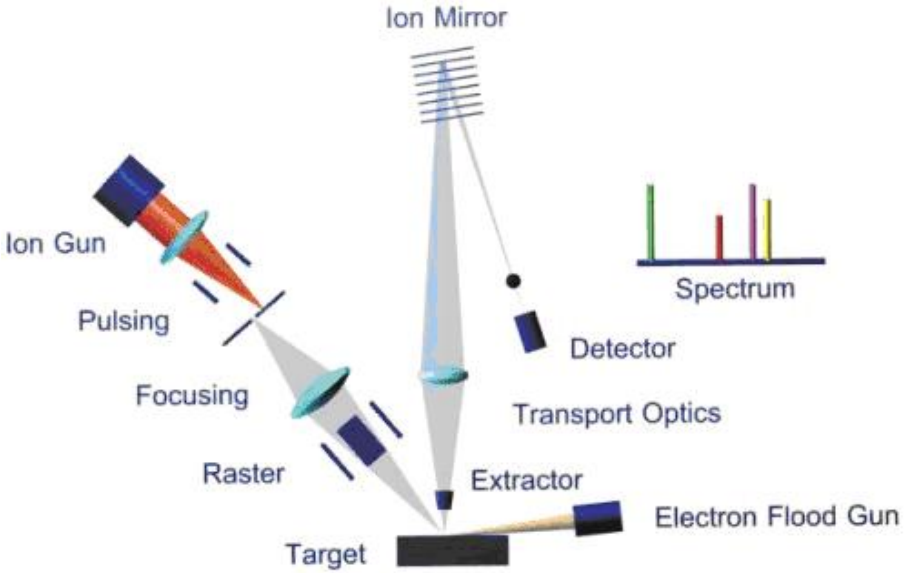
Labels 1 Labels 2 Labels n



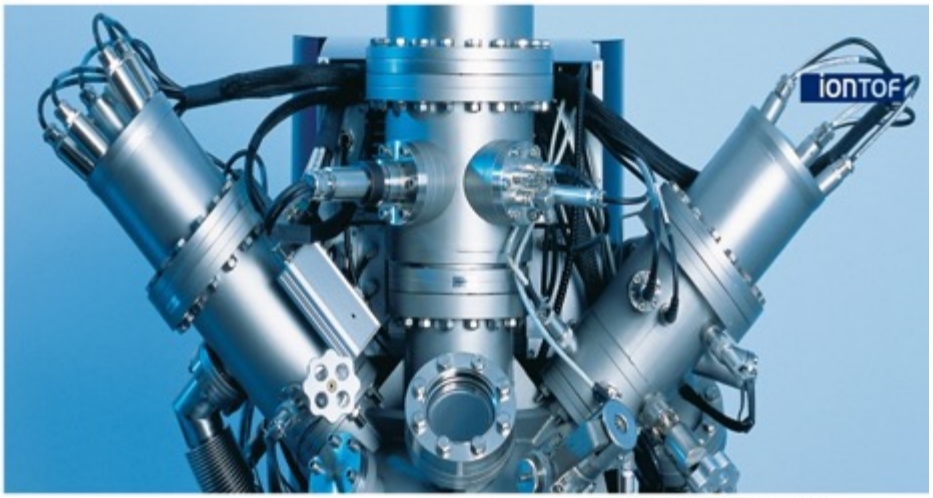
Example 1 : ToF SIMS

- Time of Flight Secondary Ion Mass Spectrometry
  - Local investigations of the sample chemical composition
  - Ionization by Bi^+ ions
  - Time of flight of secondary ions is proportional to *m/z*
  - Sputtering by Cs^+ ions for investigations in the bulk

ToF SIMS scheme



IONTOF TOF.SIMS[5] (4100 C151)

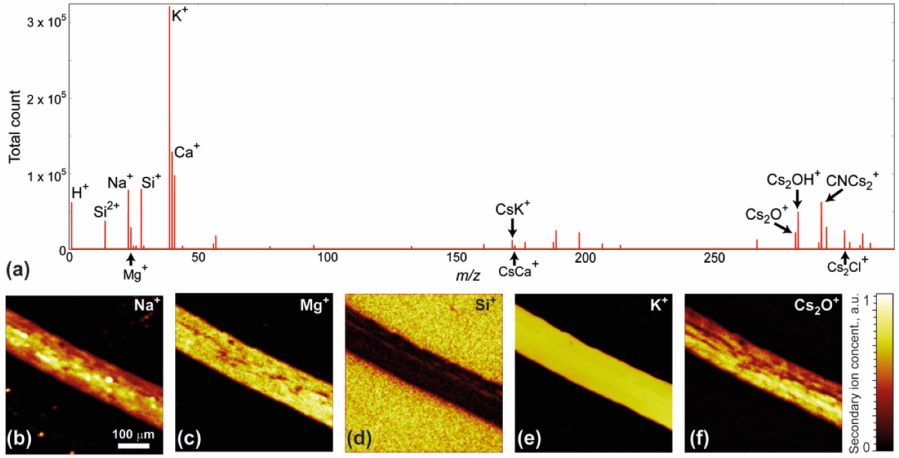


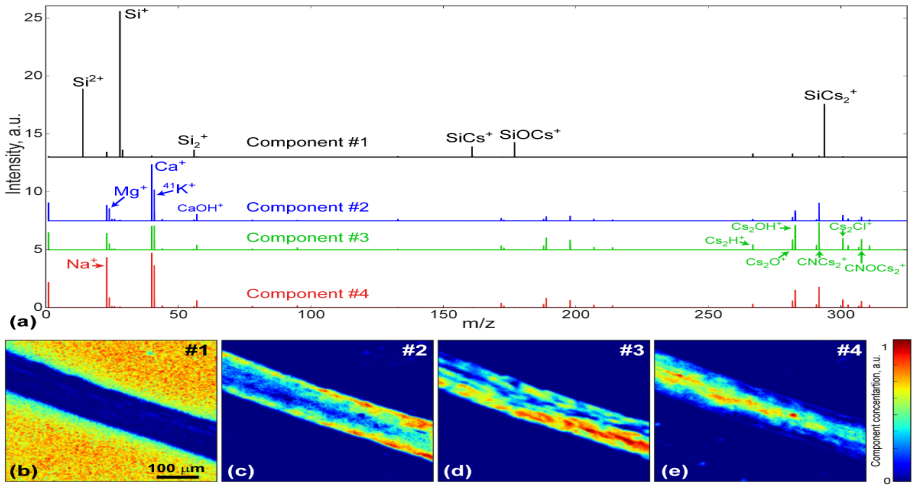
Thanks Anton

Example 1

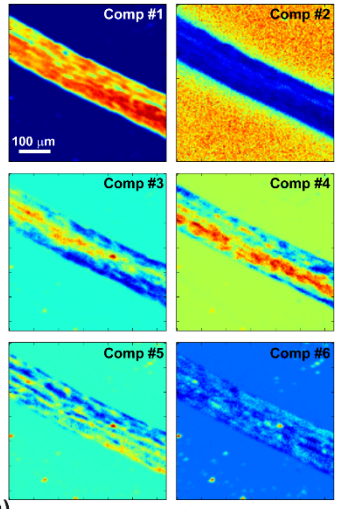
- TOF SIMS Data comes as a dense matrix
- Direct unmixing of this matrix is difficult
- We will have to find a U matrix that represents the end members and a V matrix for the abundance map (mixture among these end members)
- For this example, m = 128*128(image size), n = 1200 (signal length) and k is between 2-6.

Example 1 : NMF vs. PCA

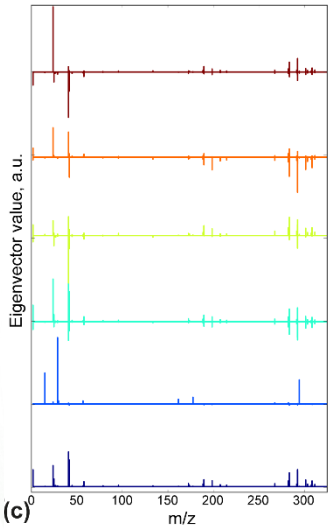




PCA abundance maps



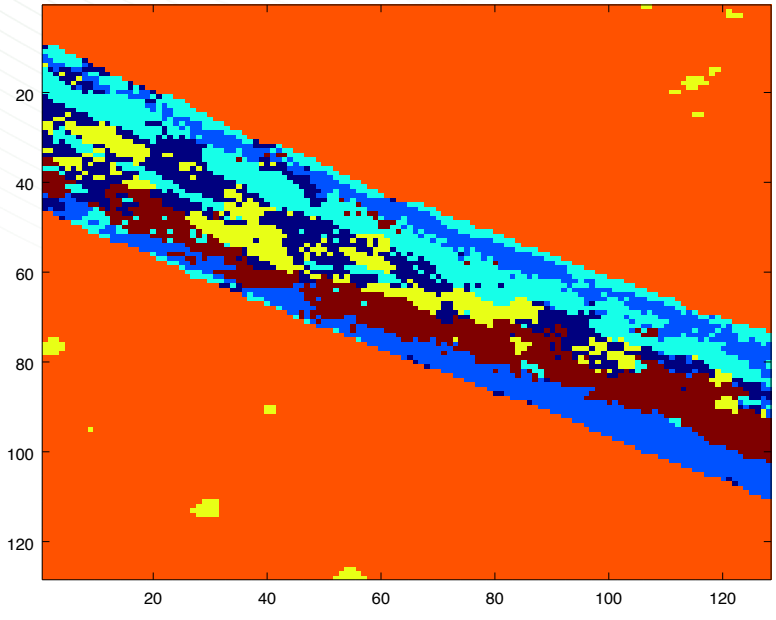
PCA Eigen vectors

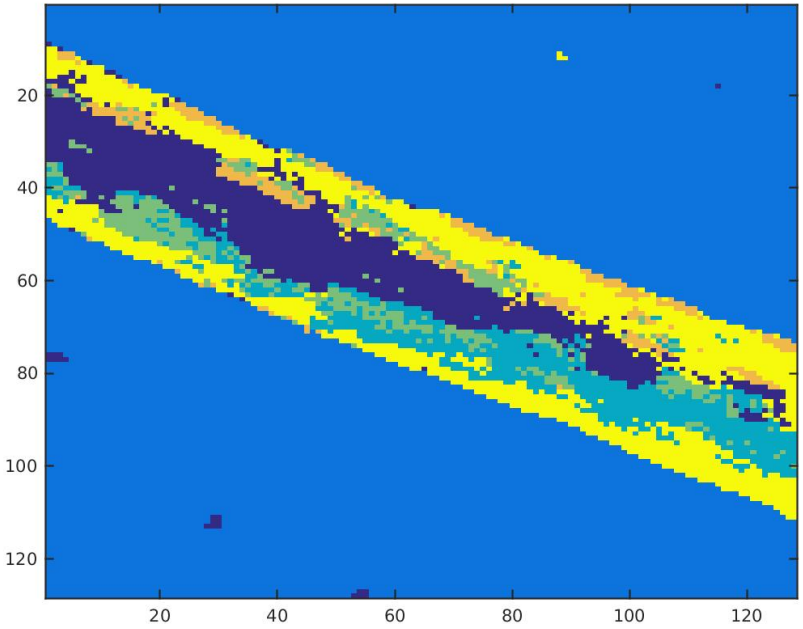


Both PCA and NMF are insufficient
They do not consider the neighbourhood information
To consider this information, we use regularization

TOF SIMS Data – Collaboration w/ Anton

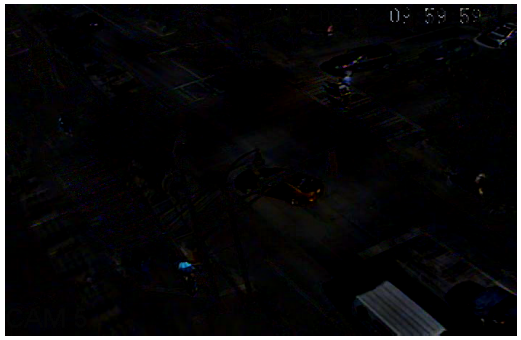
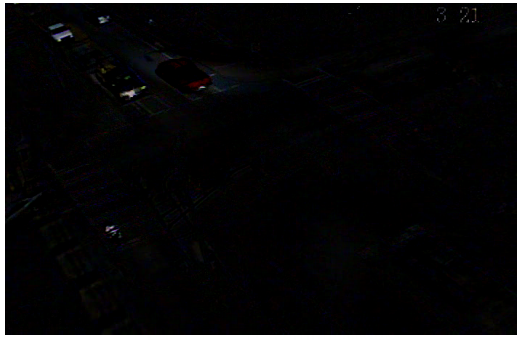
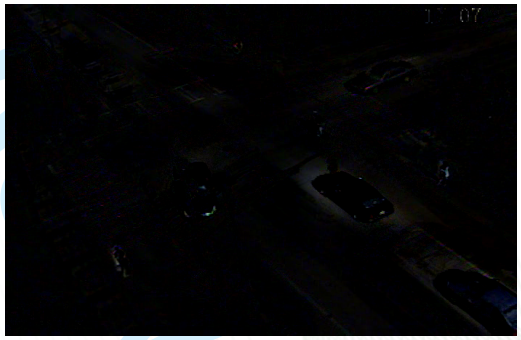
NMF with Spatial Regularization





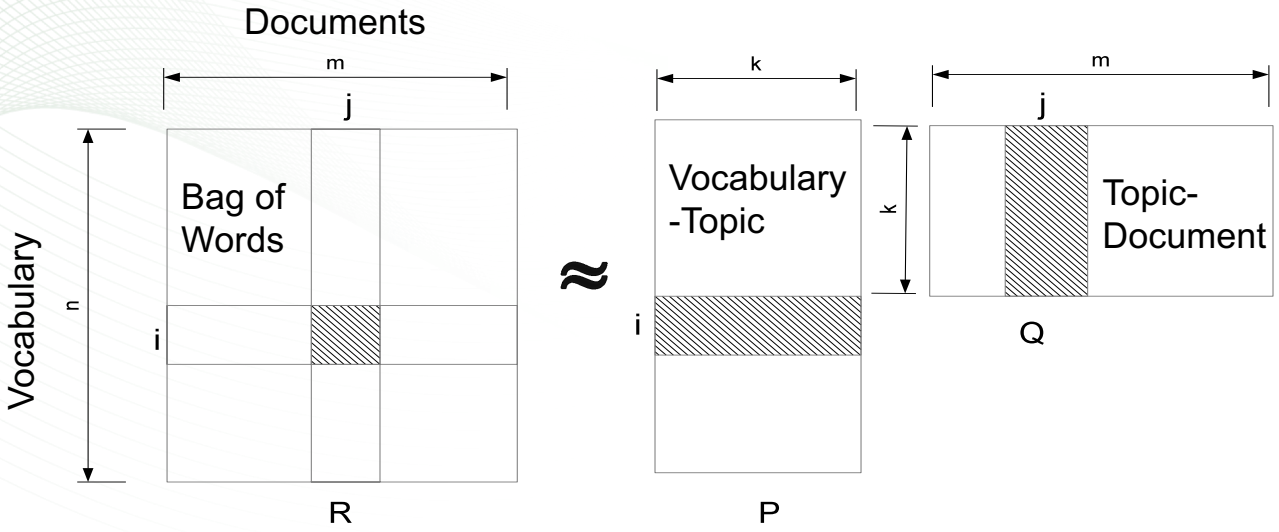
TOF SIMS Data – Collaboration w/ Anton

Example 2 : Video Data

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

Collaboration w/ Prof. Park and Prof. Ballard

Example 3 : Topic Modeling



Top Keywords from Topics 1-25					Top Keywords from Topics 26-50				
word1	word2	word3	word4	word5	word1	word2	word3	word4	word5
refer	undefin	const	key	compil	echo	type=text	php	form	result
text	field	box	word	static	test	perform	fail	unit	result
imag	src	descript	alt=ent	size	tabl	key	queri	databas	insert
button	click	event	form	add	user	email	usernam	login	log
creat	bean	add	databas	except	data	json	store	read	databas
string	static	final	catch	url	page	load	content	url	link
width	height	color	left	display	privat	static	final	import	float
app	applic	servic	thread	work	row	column	date	cell	valu
ipsum	lorem	dolor	sit	amet	line	import	command	print	recent
node	list	root	err	element	var	map	marker	match	url
0x00	0xff	byte	0x01	0xc0	server	connect	client	messag	request
file	directori	read	open	upload	number	byte	size	print	input

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

Collaboration w/ Prof. Park and Prof. Ballard

Alternating Updating NMF (AUNMF)

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

AUNMF-Algorithm

ANLS-BPP (Alternating NLS – Block Principal Pivoting)

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

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

HALS (Hierarchical Alternating Least Squares)

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

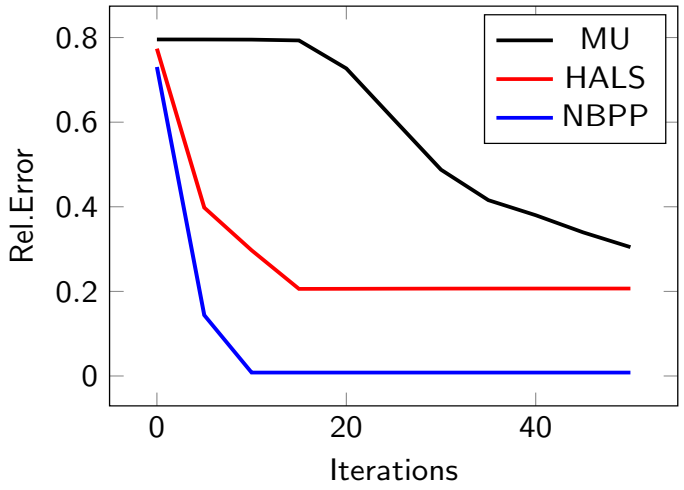
$$\mathbf{h}_i \leftarrow \left[\mathbf{h}_i + \frac{(\mathbf{W}^T\mathbf{A})_i - (\mathbf{W}^T\mathbf{W})_i\mathbf{H}}{(\mathbf{W}^T\mathbf{W})_{ii}} \right]_+$$

Multiplicative Update (MU)

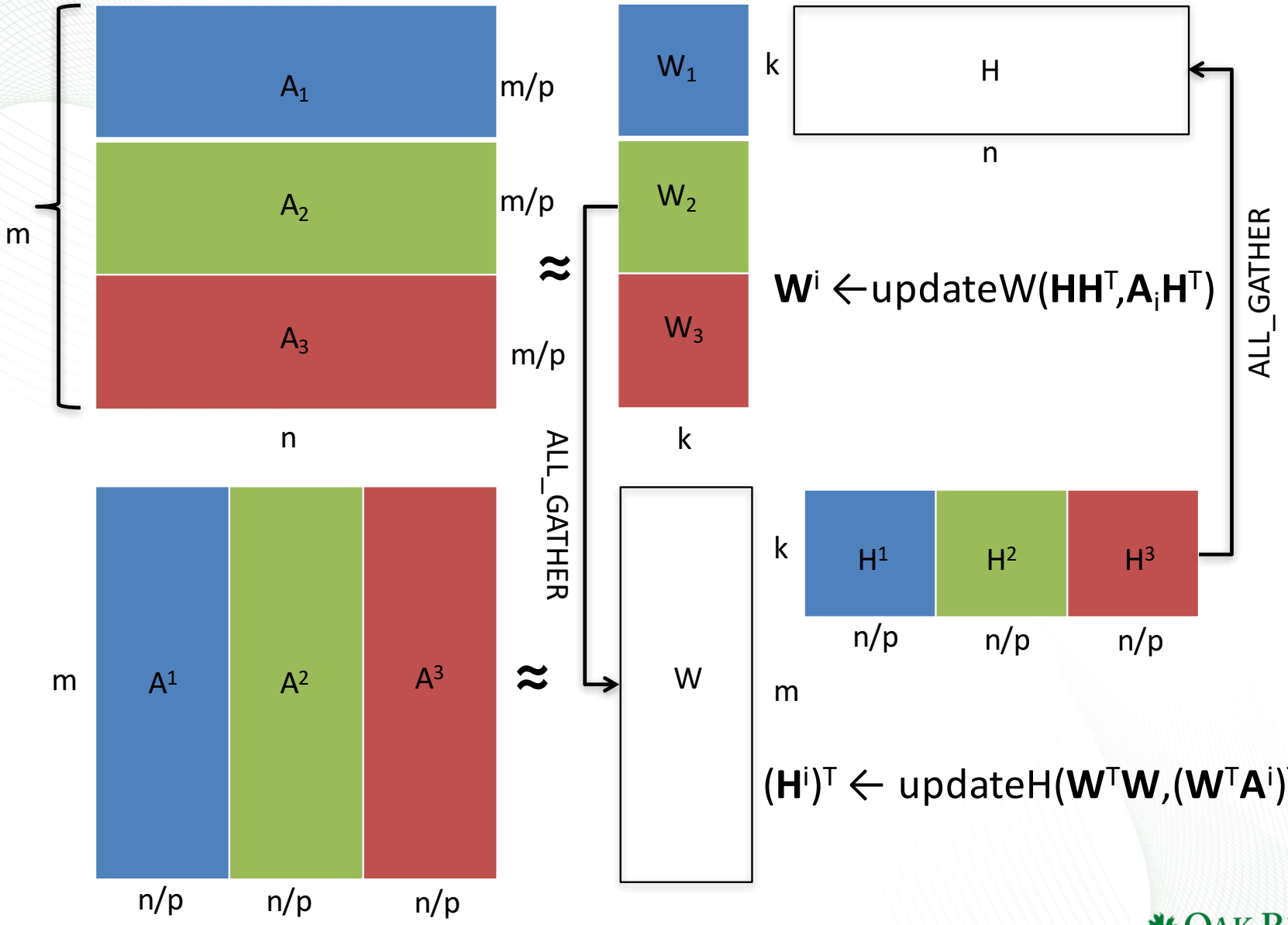
$$w_{ij} \leftarrow w_{ij}\frac{(\mathbf{A}\mathbf{H}^T)_{ij}}{(\mathbf{W}\mathbf{H}\mathbf{H}^T)_{ij}} \quad h_{ij} \leftarrow h_{ij}\frac{(\mathbf{W}^T\mathbf{A})_{ij}}{(\mathbf{W}^T\mathbf{W}\mathbf{H})_{ij}}$$

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

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



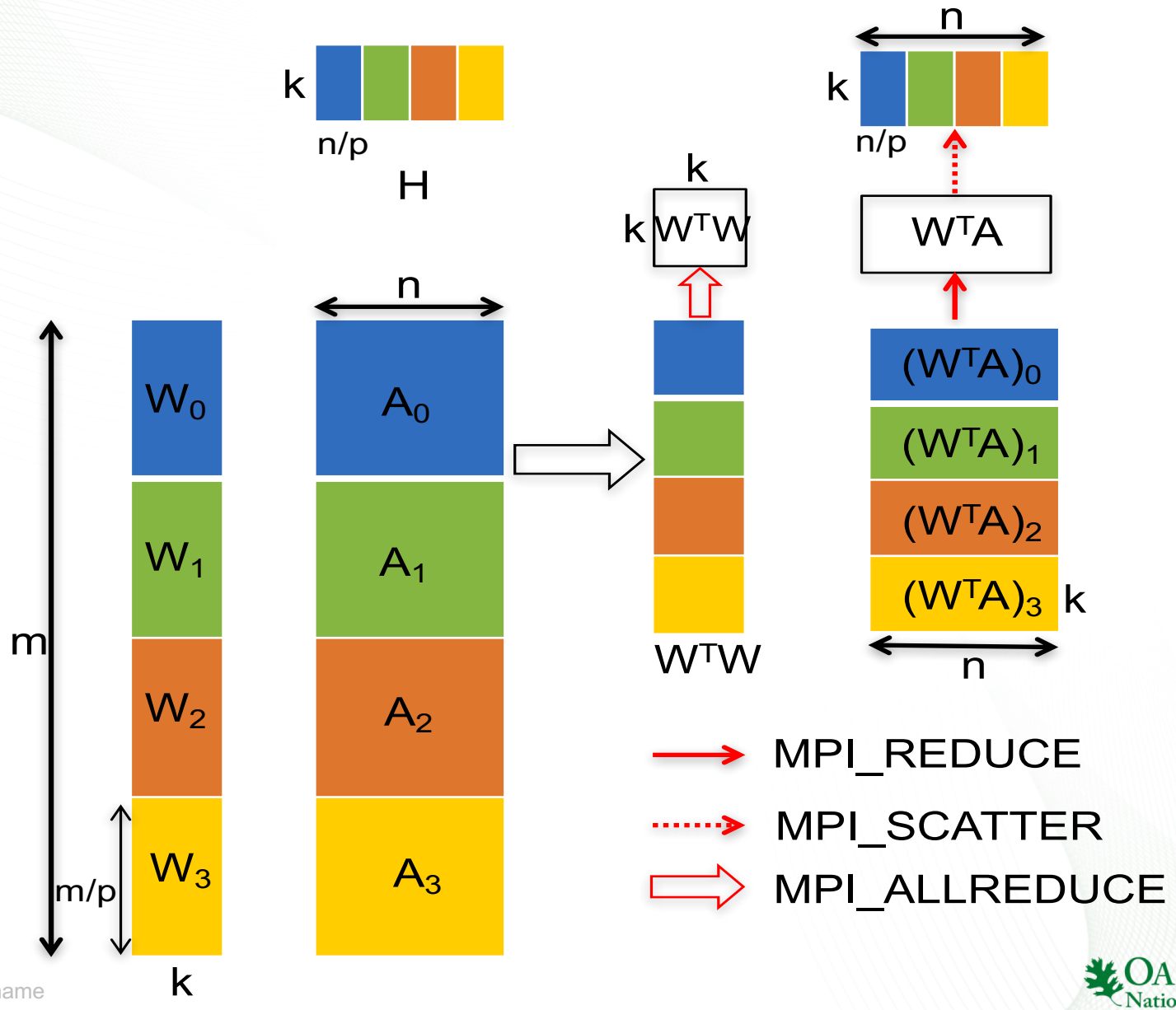
Naïve Parallel ANLS-BPP



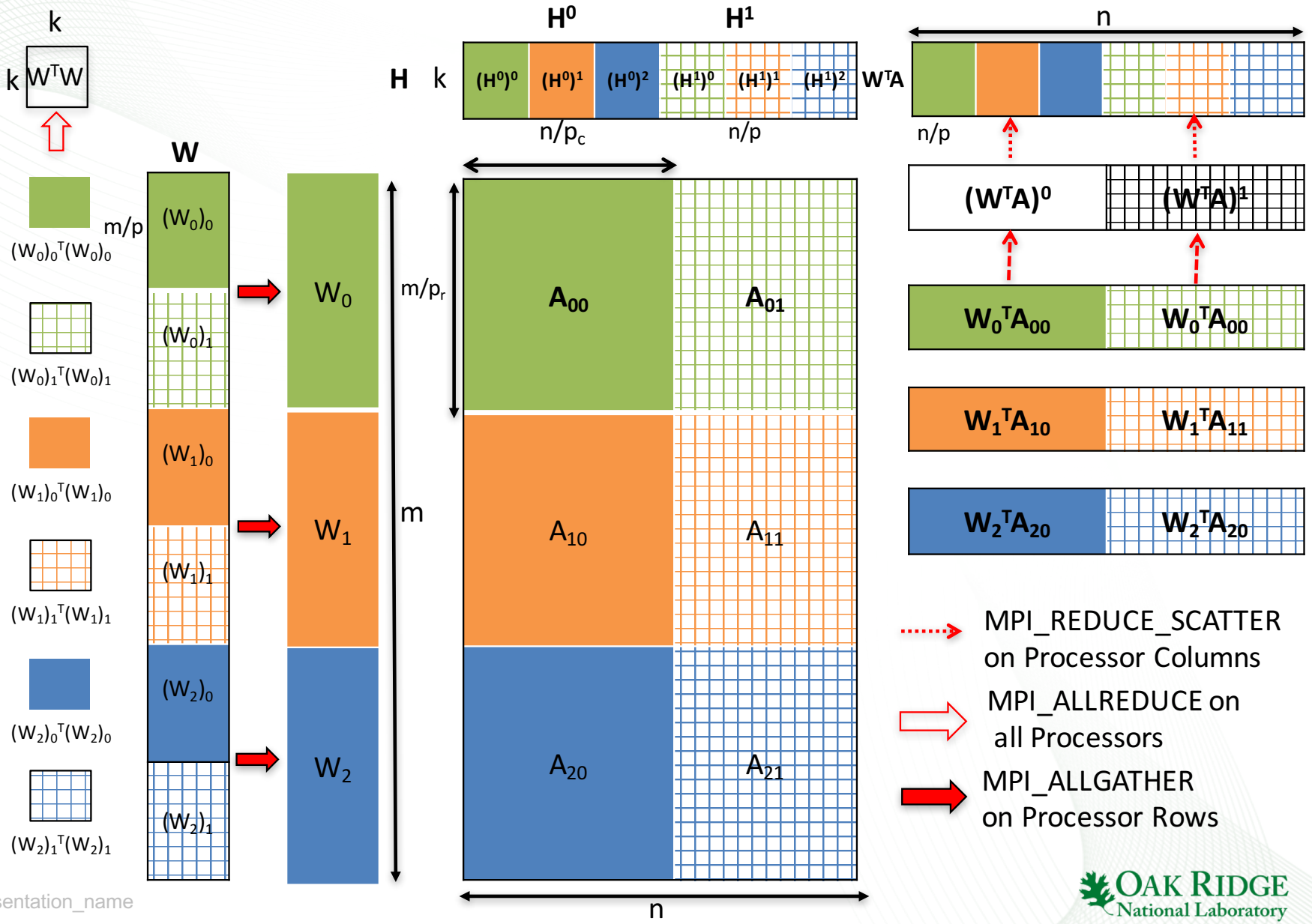
MPI-FAUN

- 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

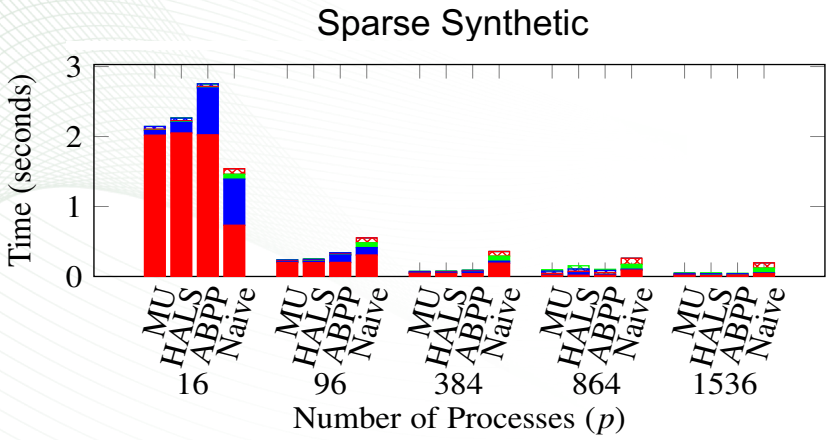
1D NMF – Long and Thin matrices

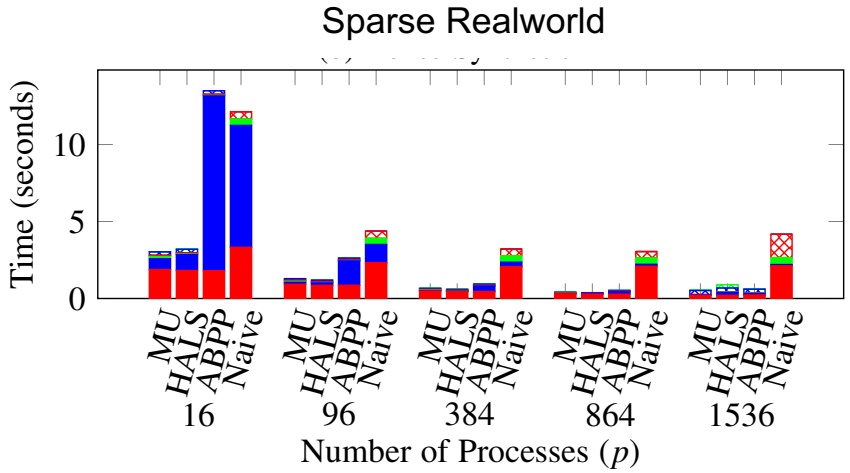


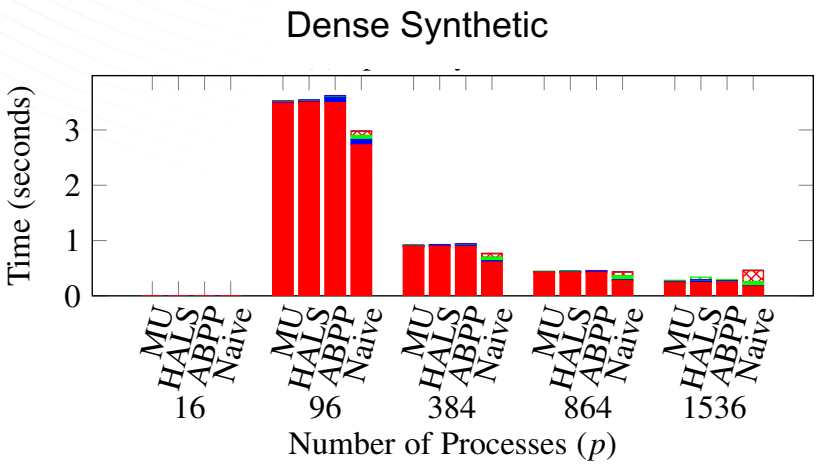
MPI-FAUN Framework

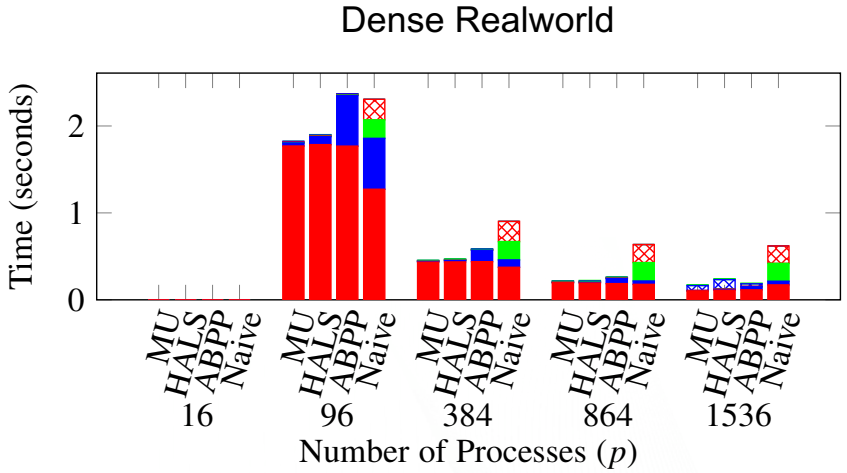


Strong Scaling





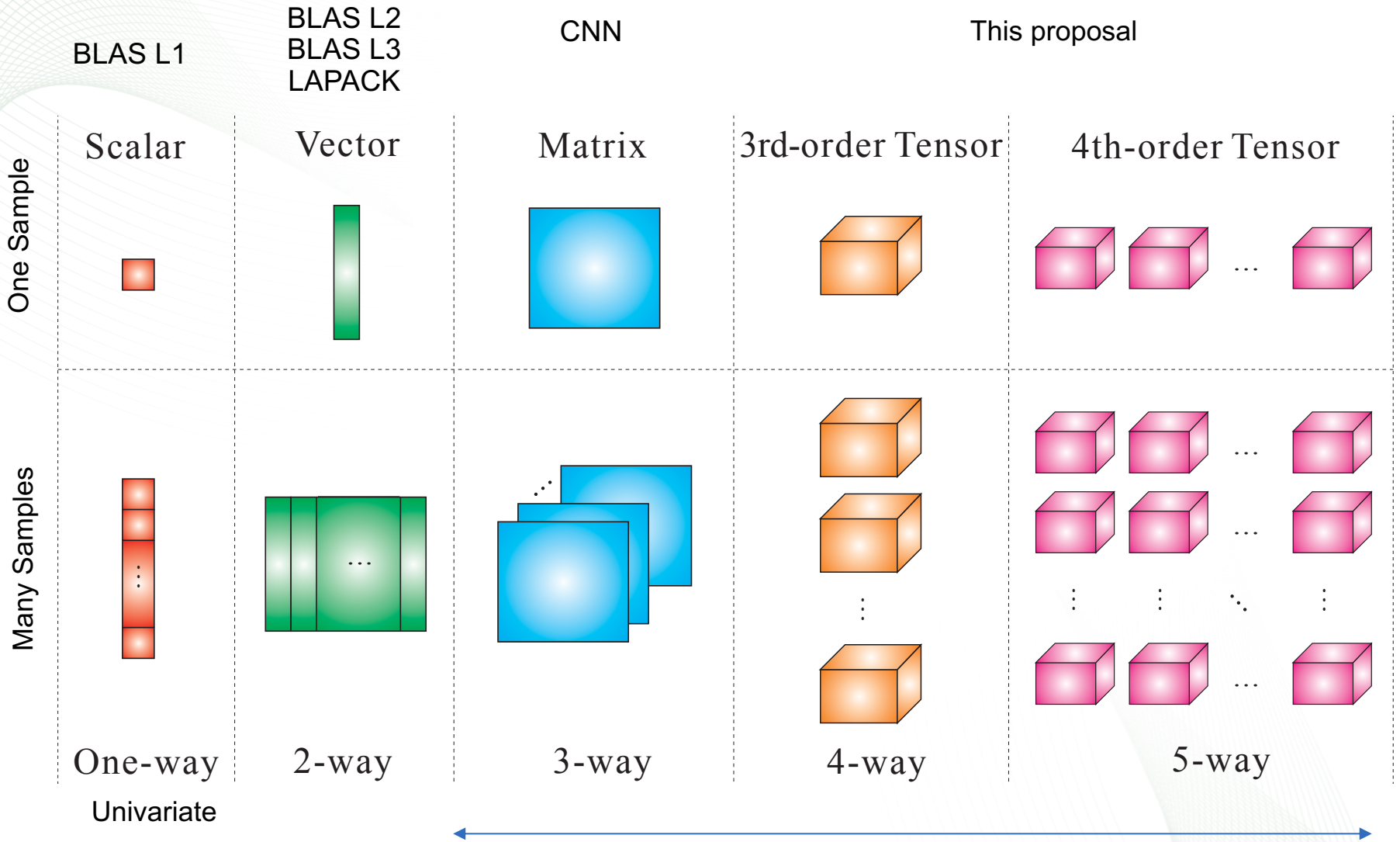




All-Reduce Reduce-Scatter All-Gather Gram LUC MM

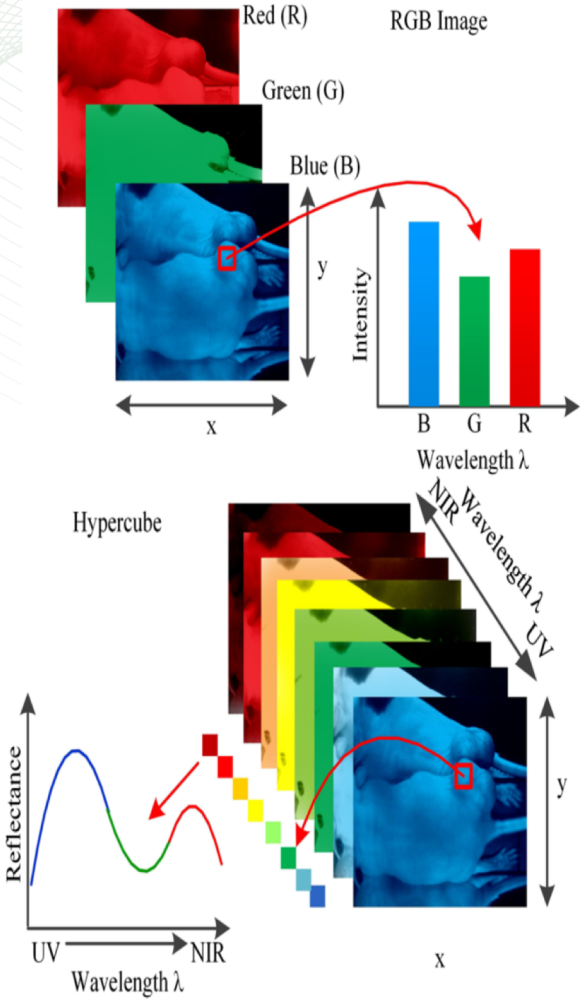
Dense/ Sparse Syn	207,360 × 138,240	Sparse Real world	1 million nodes, 3 million edges	Dense Real world	1,013,400 × 13,824 (12 min, 20 fps)

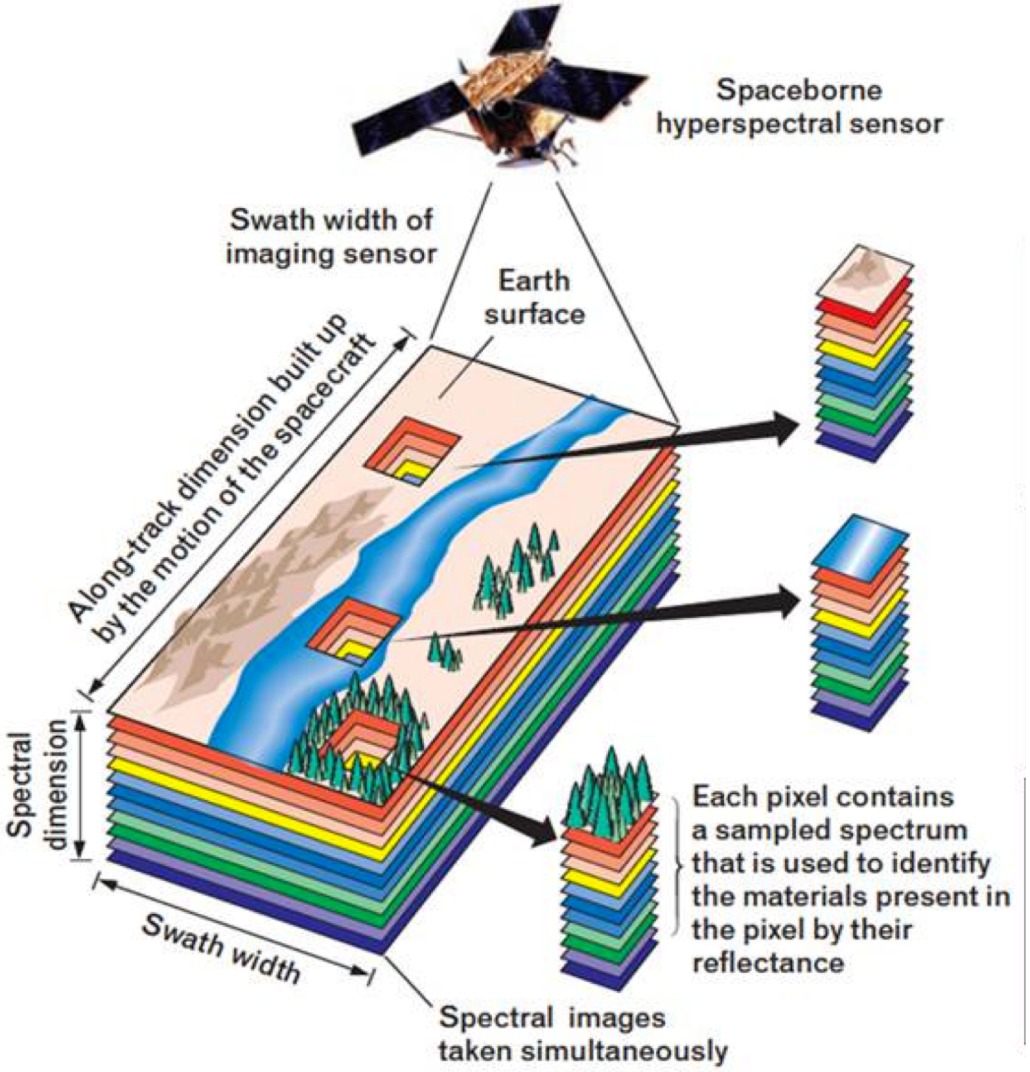
Higher Order Tensors



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

NHOT Illustration: Hyper Spectral Image



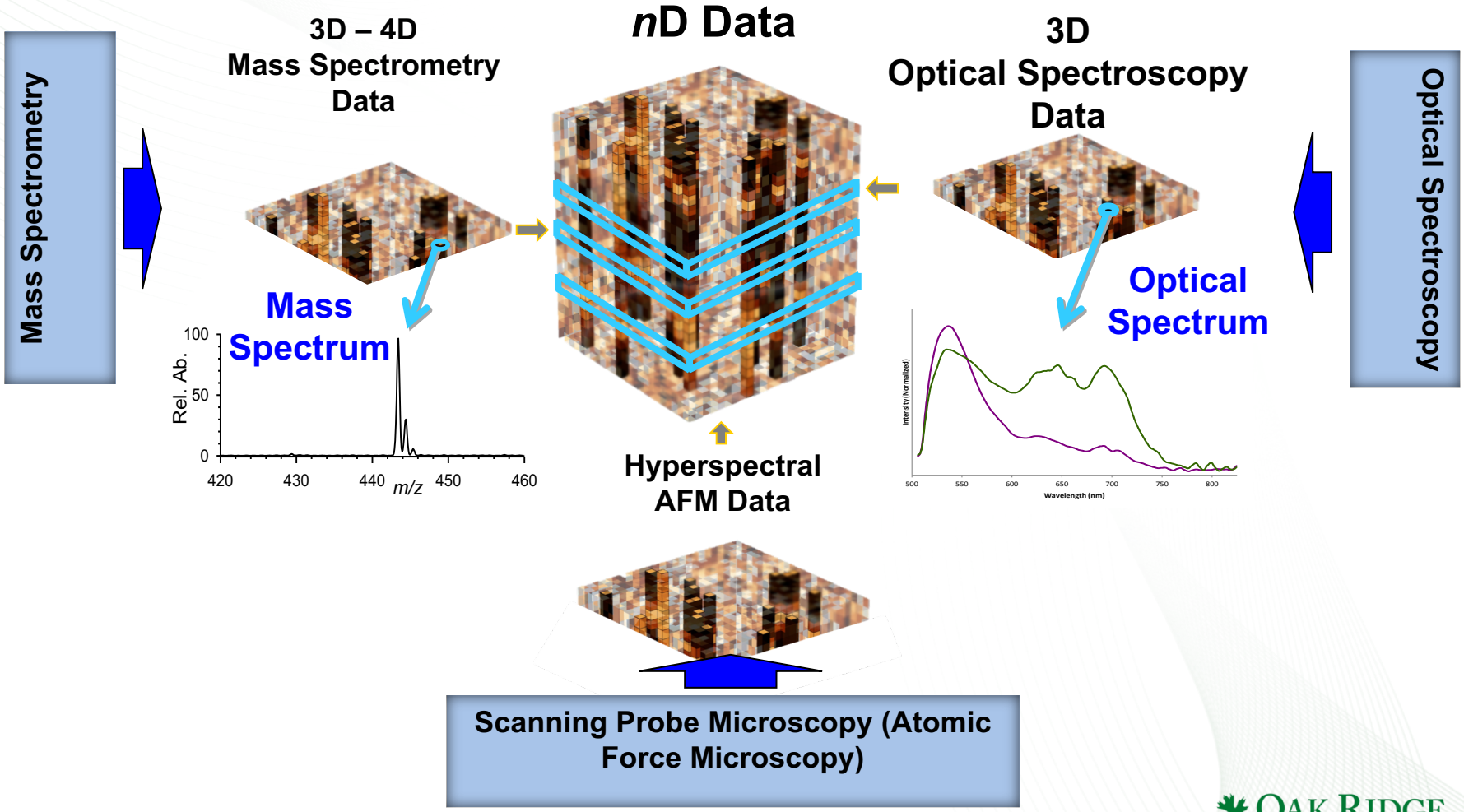


http://www.harrisgeospatial.com/Portals/0/blogs/imageryspeaks/USGS%20PRISM/BlogPost_Figure1.jpg

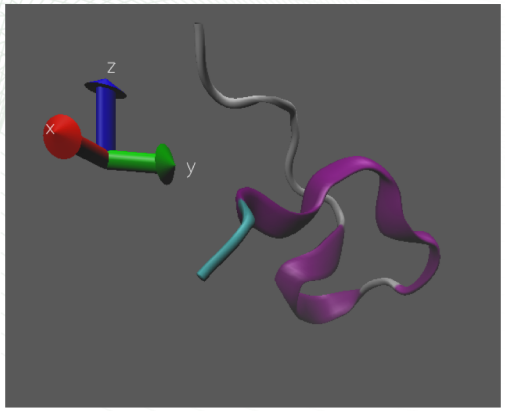
Lu G, Fei B; Medical hyperspectral imaging: a review. J. Biomed. Opt. 0001;19(1):010901. doi:10.1117/1.JBO.19.1.010901.

Dimensionality Reduction in Scientific Data

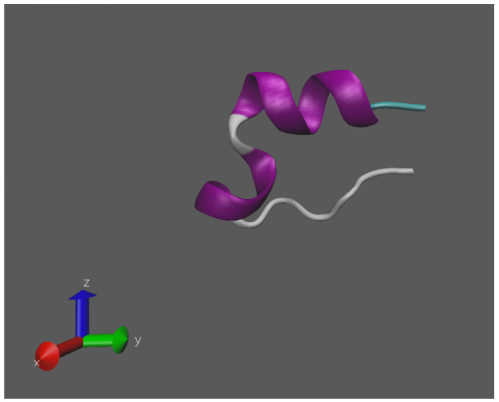
- Multimodal characterization of materials – *comprehensive characterization from chemical composition to functional properties on the nanoscale*



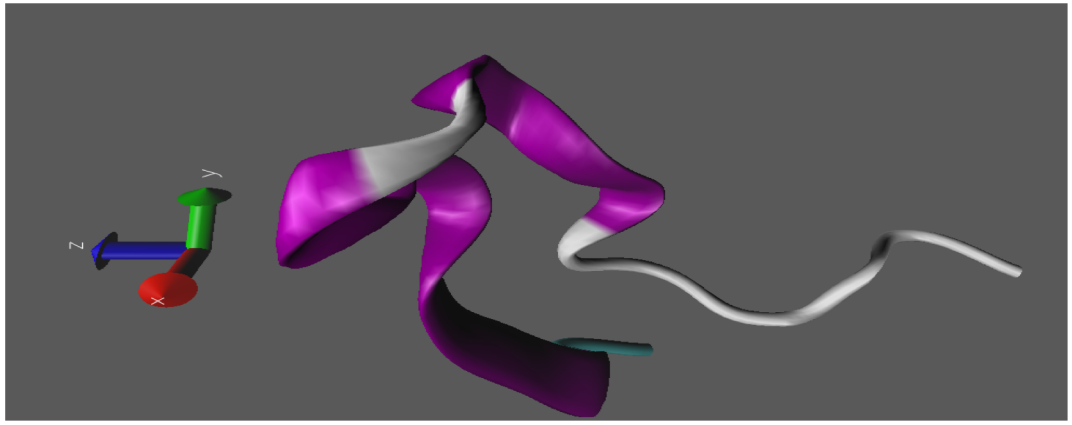
NHOT Illustration: Molecular Simulation



T=1



T=2



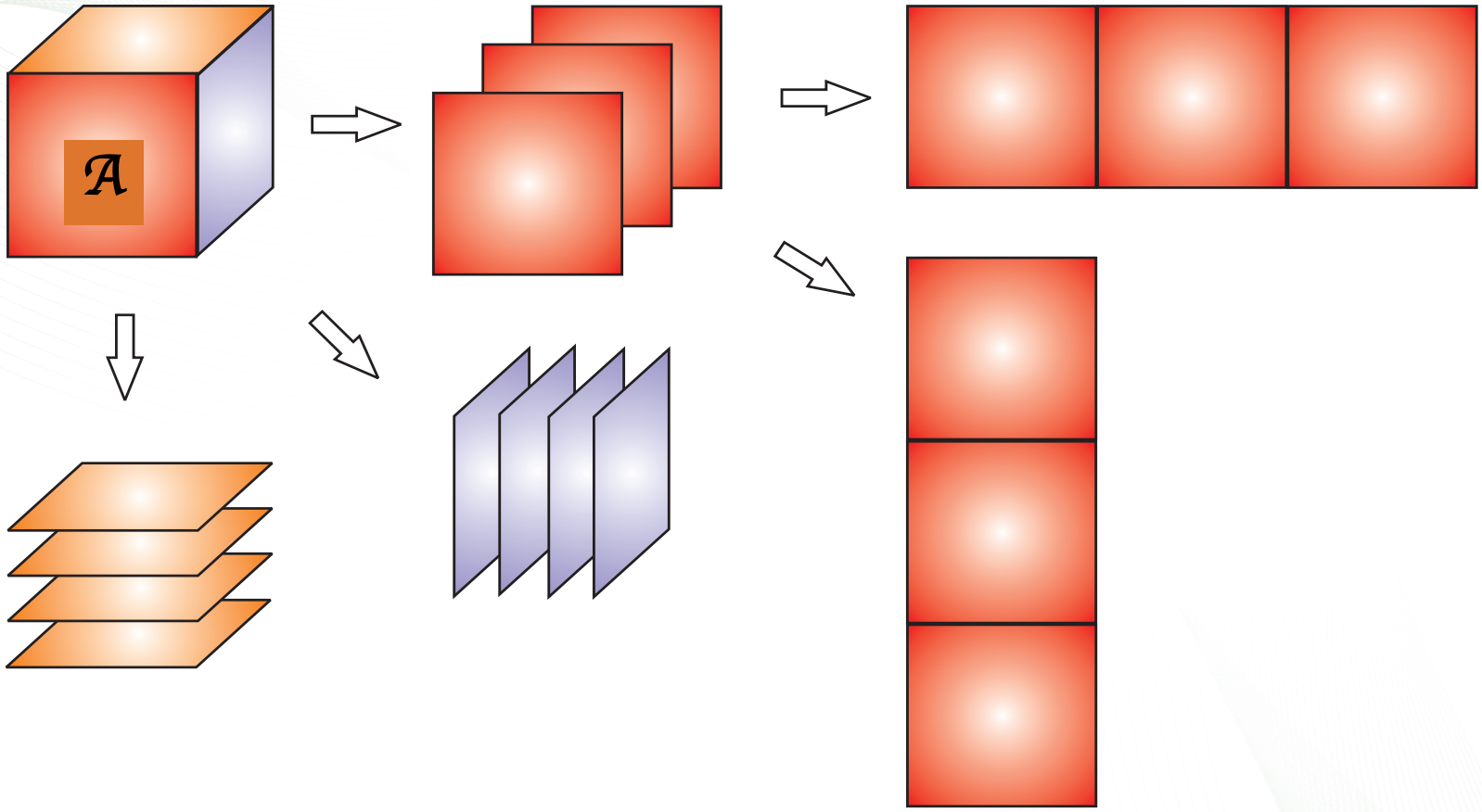
T=3

- Each molecule has an (x,y,z) coordinate that changes over time
- These changes are captured at very high resolution at pico second levels

The movements occuring in picosecond has over all impact on the molecular behaviour over duration at multiple scales

Thanks Arvind and Debsindhu

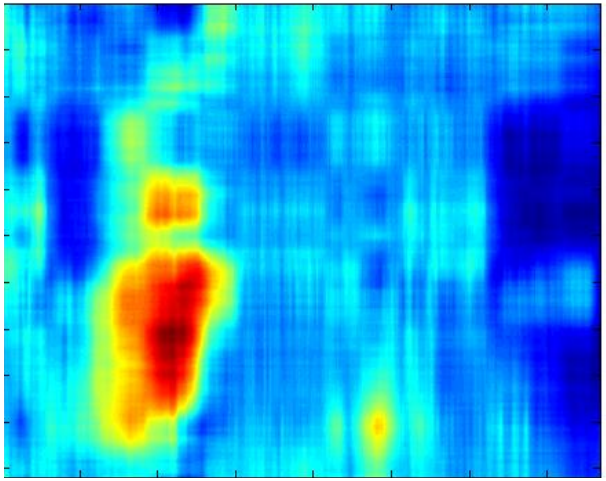
Existing DR for NHOT - Matricization



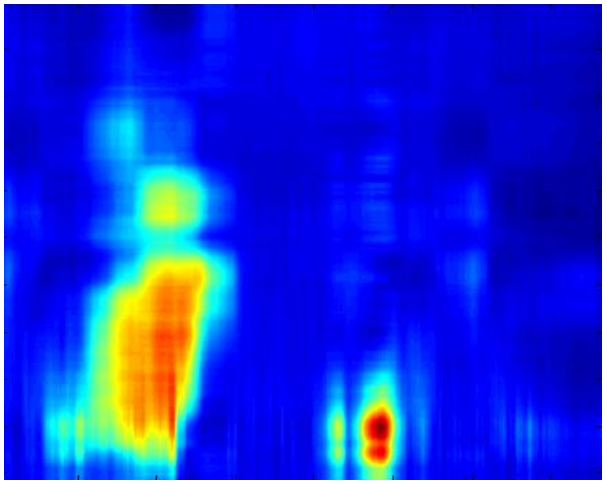
- Works only when some of the dimensions are independent
- Matricizing NHOT is non-trivial

Matricization vs. Proposed DR Results

Original



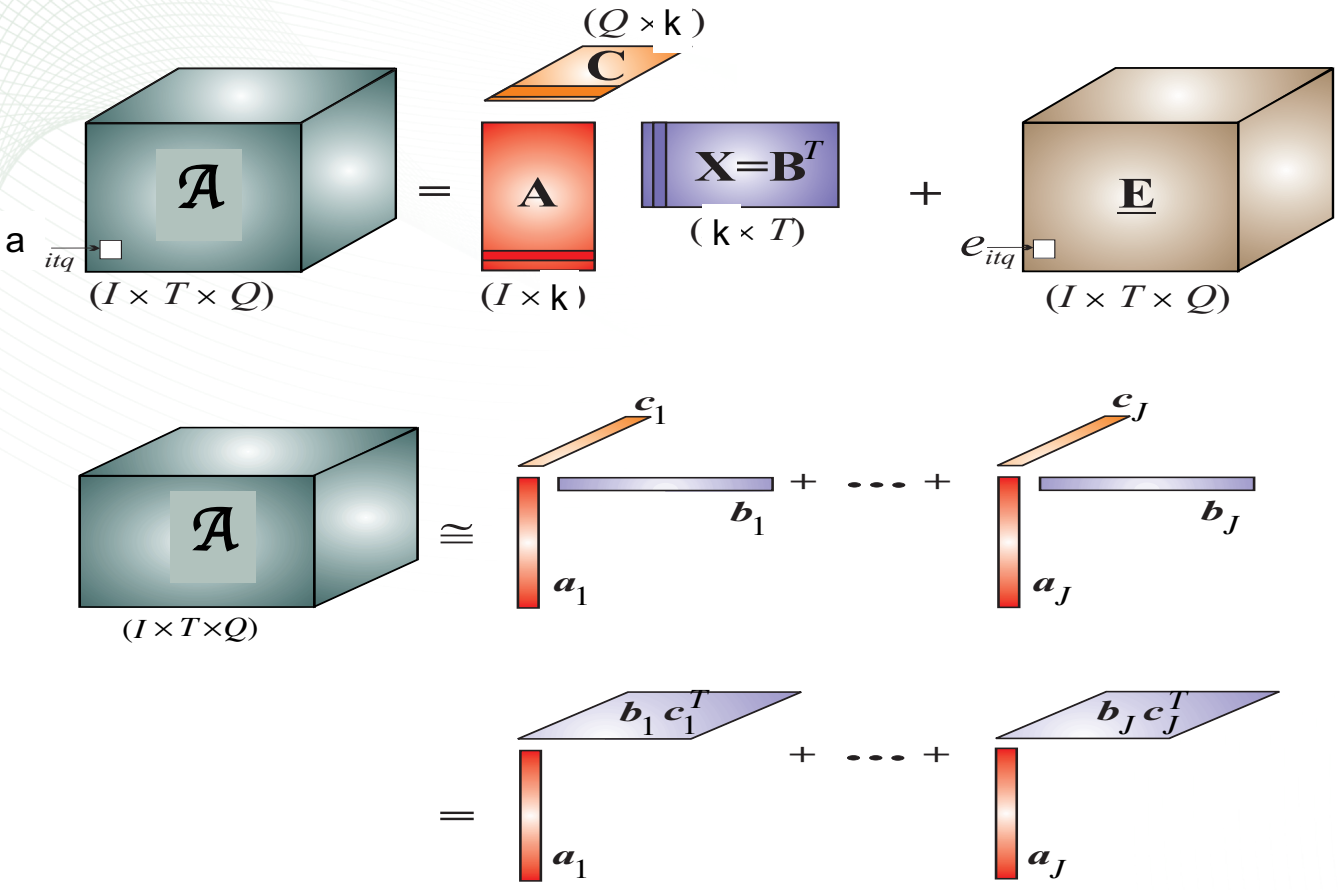
Non-negative Matrix Factorization



Non-negative Tensor Factorization

- Original Image is of size 2046x3406x4 – RGB+Near IR
- We matricized the four frames into rows yielding a matrix of size 8184x3406. NMF output is on this matrix
- NTF was performed directly on 2046x3406x4 tensor. NTF leveraged the spatial correlation.

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

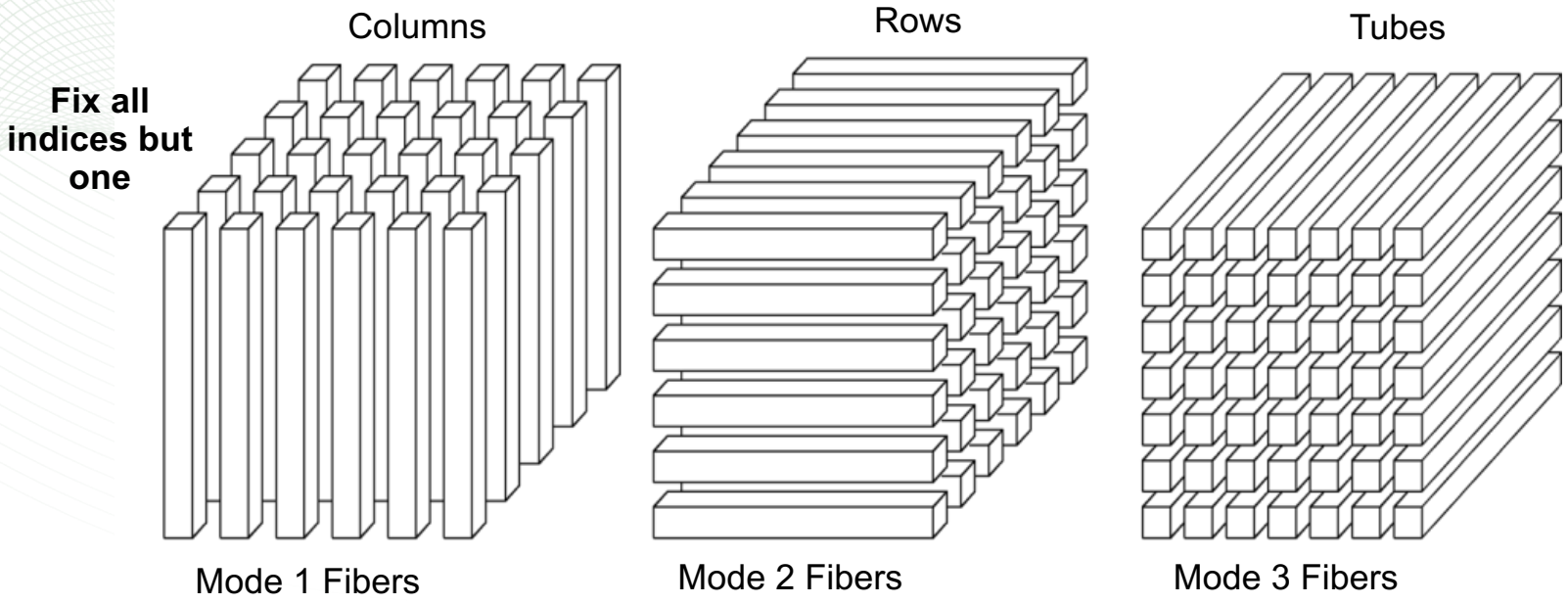
$\mathbf{H}^{(1)}, \ldots, \mathbf{H}^{(N)}$

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

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



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

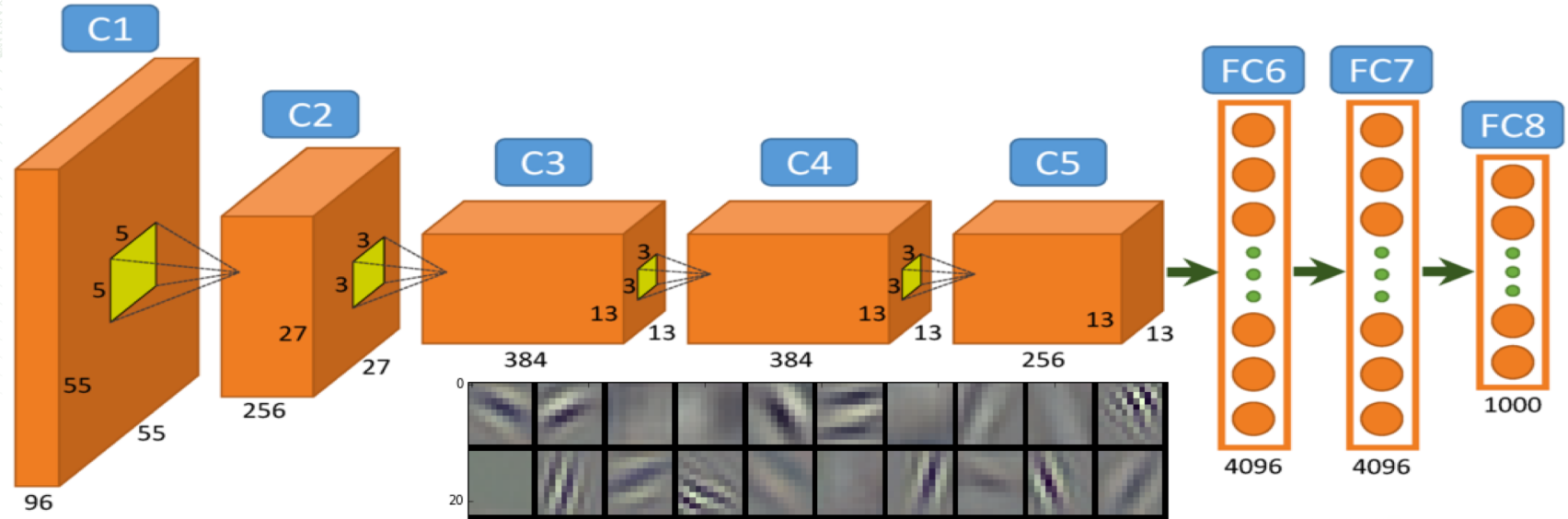
Distributed NCP Algorithm

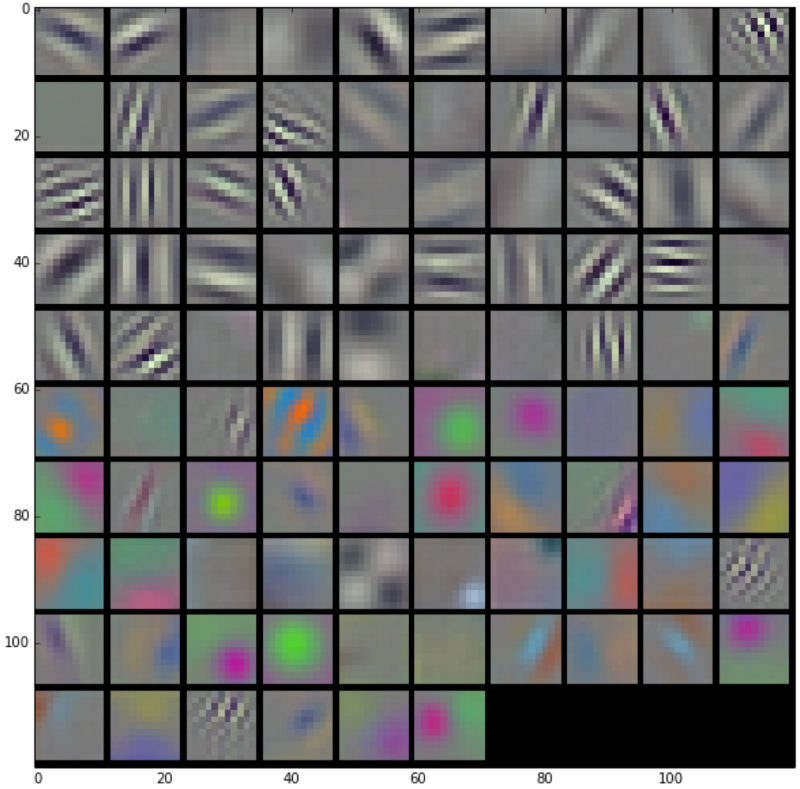
- N-D Process Grid for N modes $P_1 \times \cdots \times P_N$
- Input Tensor is distributed as $\mathcal{A}_{p_1 \cdots p_N}$ is $(M_1/P_1) \times \cdots \times (M_N/P_N)$
- Factors are all_gathered as $\mathbf{H}_{p_i}^{(i)}$ is $(M_i/P_i) \times k$

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

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

Deep Learning and NTF



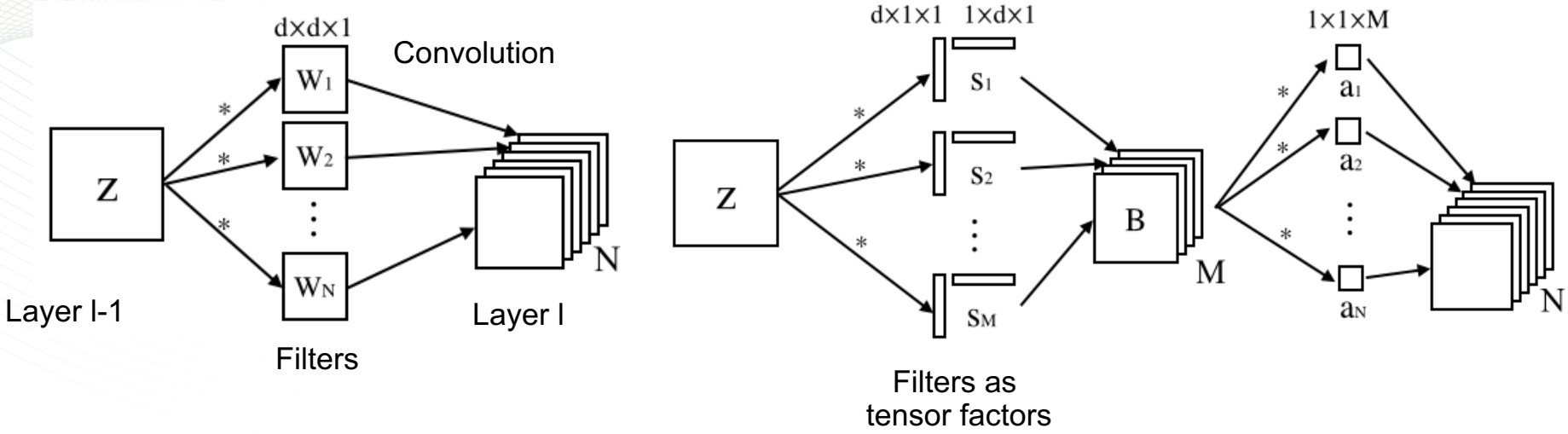


Deep Learning and NTF

- Denil et.al.,(2013) proposed that the weights within a layer can be accurately predicted from a small subset of them.
- Rigamonti et.al.,(2013) show that multiple image filters can be approximated by a shared set of separable filters
- Cohen et.al., (2016) Shallow network corresponds to CP (rank-1) decomposition and deep network correspond to Hierarchical Tucker Decomposition

Collaboration w/ Arvind, Srikanth and Dmitry

Deep-learning and NTF



- Total number of parameters in level C2-C3 in Alexnet is dxdx256x384. If d=11, the number of parameters is 11,894,784 approximately 12M
- In our proposed architecture, it needs only 10,592 parameters with low rank 16.

Collaboration w/ Arvind, Srikanth and Dmitry

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