High Performance Computing with Sparse Data

Graphs, Matrices and Tensors

Kamer Kaya, Sabancı University

ncc@ulakbim.gov.tr

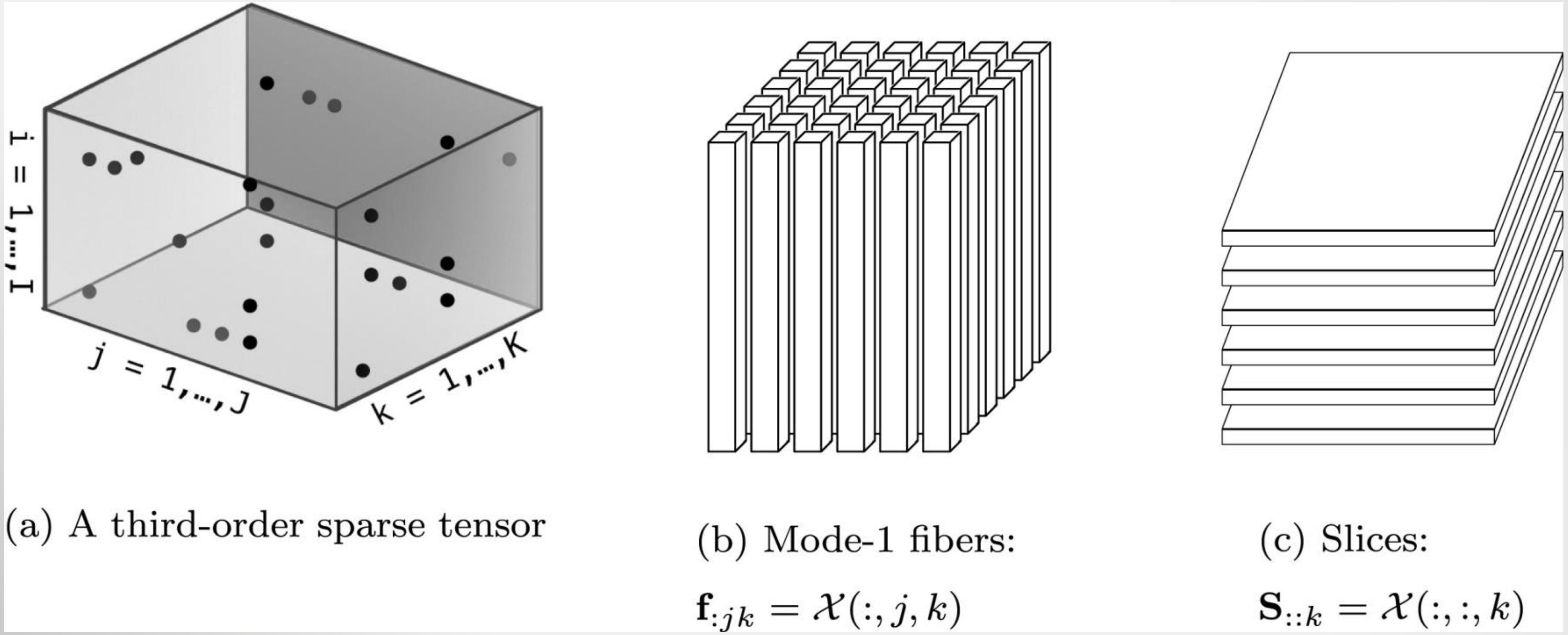
Sparse Tensors

A **sparse tensor** is a generalization of sparse matrices to higher dimensions.

Like sparse matrices, sparse tensors are used to represent multi-dimensional arrays (or tensors) where most of the elements are zero.

Instead of explicitly storing every element, sparse tensors store only the non-zero values and their corresponding indices, which saves memory and computation time.

Sparse Tensors



Yuchen Ma, Jiajia Li, Xiaolong Wu, Chenggang Yan, Jimeng Sun, Richard Vuduc, Optimizing sparse tensor times matrix on GPUs, Journal of Parallel and Distributed Computing, Volume 129, 2019

Why Sparse Tensors?

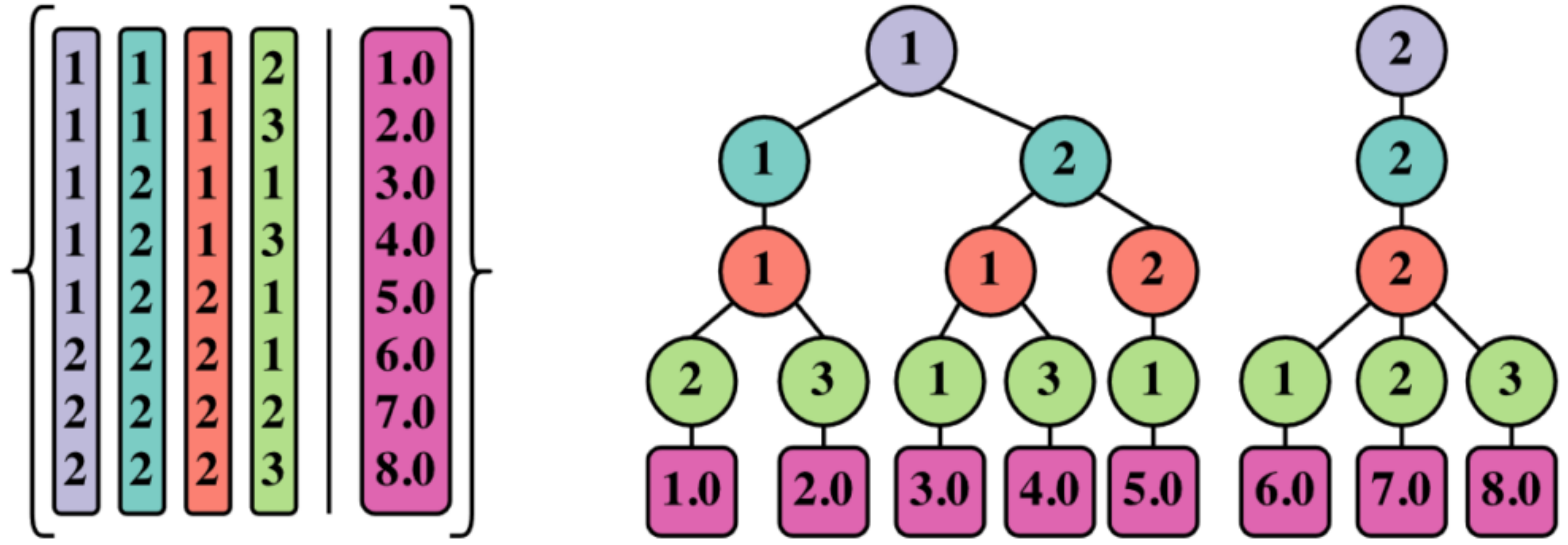
Memory Efficiency: Sparse tensors significantly reduce memory usage by not storing zeros.

Computational Efficiency: Operations on sparse tensors focus only on non-zero elements, avoiding unnecessary computations.

Applications:

- Machine learning (e.g., recommendation systems, NLP).
- Scientific computing (e.g., simulations with sparse systems).
- Graph and network analysis (e.g., adjacency tensors)

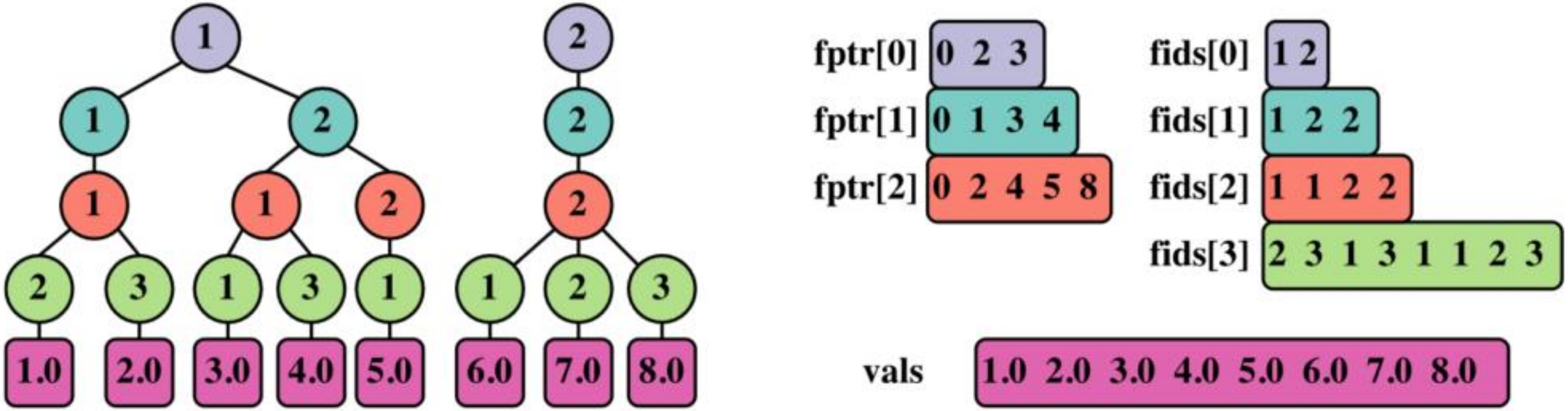
Sparse Tensor Data Structures - CSF



Coordinate format. CSF: conceptual.

https://www.boristhebrave.com/2021/01/01/compressed-sparse-fibers-explained/

Sparse Tensor Data Structures - CSF

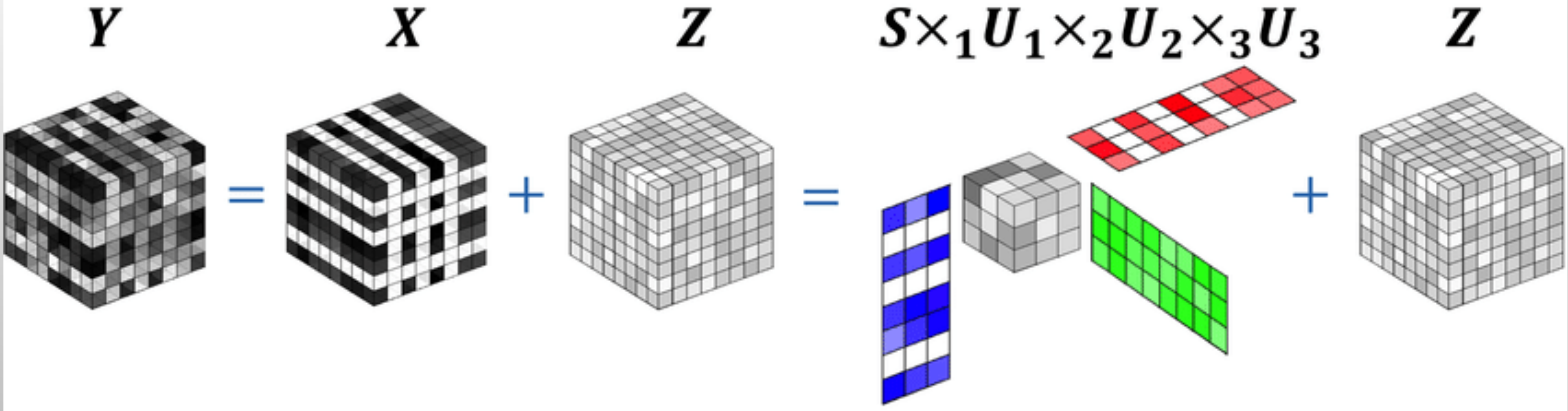


CSF: conceptual.

CSF: implementation.

https://www.boristhebrave.com/2021/01/01/compressed-sparse-fibers-explained/

Sparse Tensor Operations



Optimal Sparse Singular Value Decomposition for High-dimensional High-order Data Anru Zhang, Rungang Han, arXiv:1809.01796

```
// Perform MTTKRP
for (const auto &[i, j, k, value] : tensor) {
    for (int r = 0; r < R; ++r) {
        result[i][r] += value * A[j][r] * B[k][r];
    }
}
```

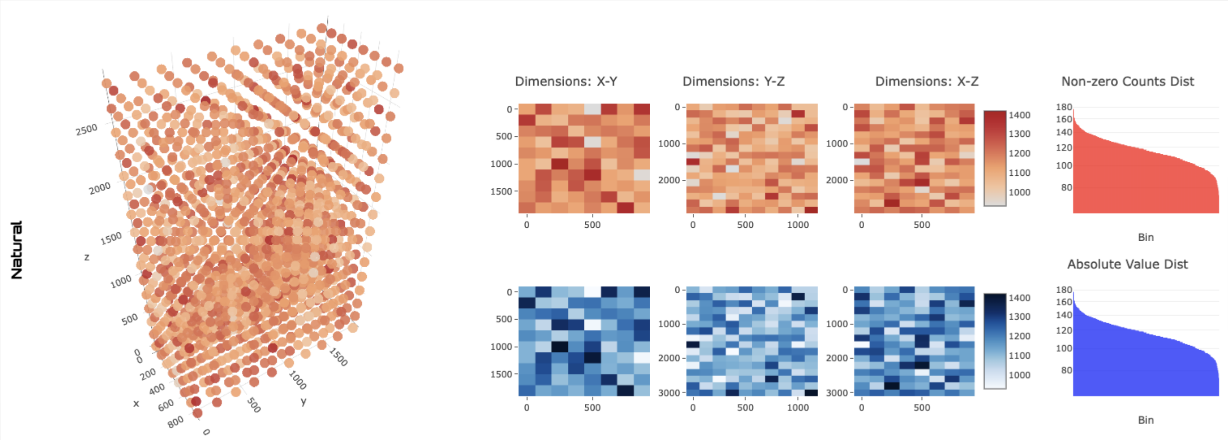
The main access pattern for MTTKRP and many tensor kernels.

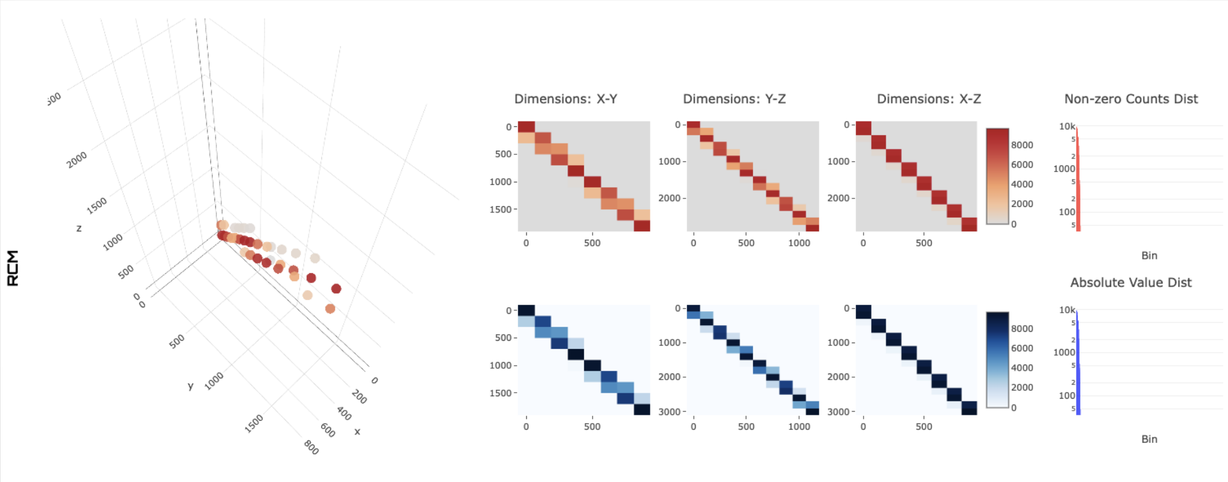
Sparse Tensor Operations

```
result[i][r] += tensor_value * A[j][r] * B[k][r];
```

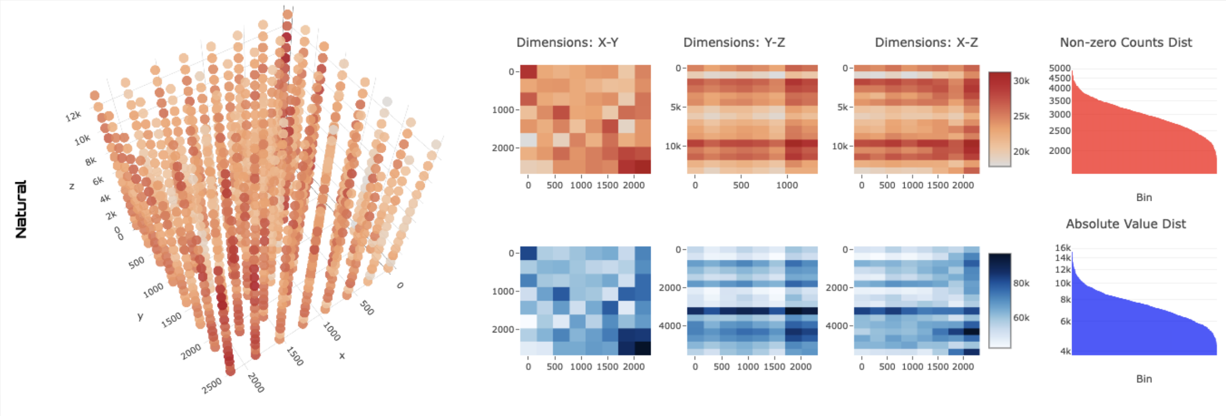
- Good Spatial Locality in result$[i][r]$
  - If the nonzeros are sorted w.r.t. their first dimension indices.
- Potentially Bad Spatial Locality in Factor Matrices:
  - Spatial locality depends on the indices j and k. If these indices correspond to widely separated rows in the matrices A, B, spatial locality may degrade.

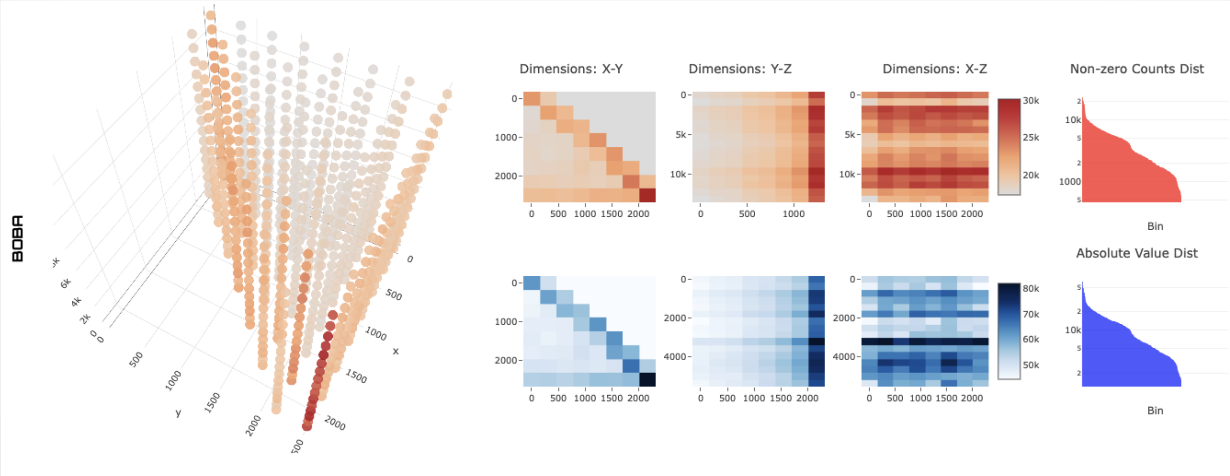
Sparse Tensor Operations





Sparse Tensor Operations





Thanks

This project has received funding from the European High-Performance Computing Joint Undertaking (JU) under grant agreement No 101101903. The JU receives support from the Digital Europe Programme and Germany, Bulgaria, Austria, Croatia, Cyprus, Czech Republic, Denmark, Estonia, Finland, Greece, Hungary, Ireland, Italy, Lithuania, Latvia, Poland, Portugal, Romania, Slovenia, Spain, Sweden, France, Netherlands, Belgium, Luxembourg, Slovakia, Norway, Türkiye, Republic of North Macedonia, Iceland, Montenegro, Serbia