AWARE: Workload-aware, Redundancy-exploiting Linear Algebra
Sebastian Baunsonga & Matthias Boehm
Technische Universität Berlin

**MOTIVATION**
- The next step from Sparsity Exploitation → Redundancy Exploitation
- We changed the compression goal from the norm of Compression Ratio to optimizing for Execution Time in a workload-aware manner
- Guaranteed same results as uncompressed data via Lightweight Database Compression Techniques and Compressed Linear Algebra
- Improved performance of individual operations by up to 10,000x!
- Improved algorithmic performance including everything in end-to-end pipelines including Online Compression and algorithm!
- Grid search algorithms improving from 274.3 sec to 92.6 sec on same hardware due to reduced memory bandwidth requirements and faster direct compressed operations.

**WORKLOAD EXTRACTION**
User Script:
- cost(shift) = colMeans(x)
- cost(scale) = colSds(x)
- Output Matrix
- Built-in Functions:

**COMPRESSION WORKFLOW**
1. Classify Uncompressed Input Matrix
2. Grouping Extract column statistics Co-code Statistics
3. Transpose (optional) Compressed Output Matrix
4. Compress Extract & Compress

**COMPRESSION EXAMPLE**

**COMRESSED OPERATION EXAMPLE**

**OPERATIONS PERFORMANCE**

**LOCAL END-TO-END EXPERIMENTS**
Workload-awareness on Local End-to-End Algorithms (Data: US Census Enc)

**HYBRID END-TO-END EXPERIMENTS**

**TENSORFLOW COMPARISON**