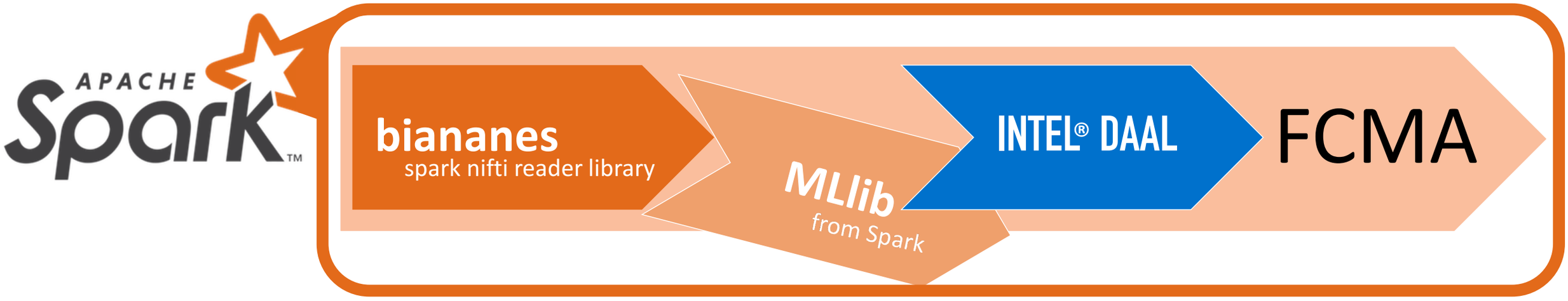
Big Data Approach to fMRI Data Analysis with Intel DAAL and Full Correlation Matrix Analysis

Haoran Shu (CUHK)
Yin Lok Wong (HKU)

Mentors:
Pragnesh Kumar
Kwai Wong
Junqi Yin

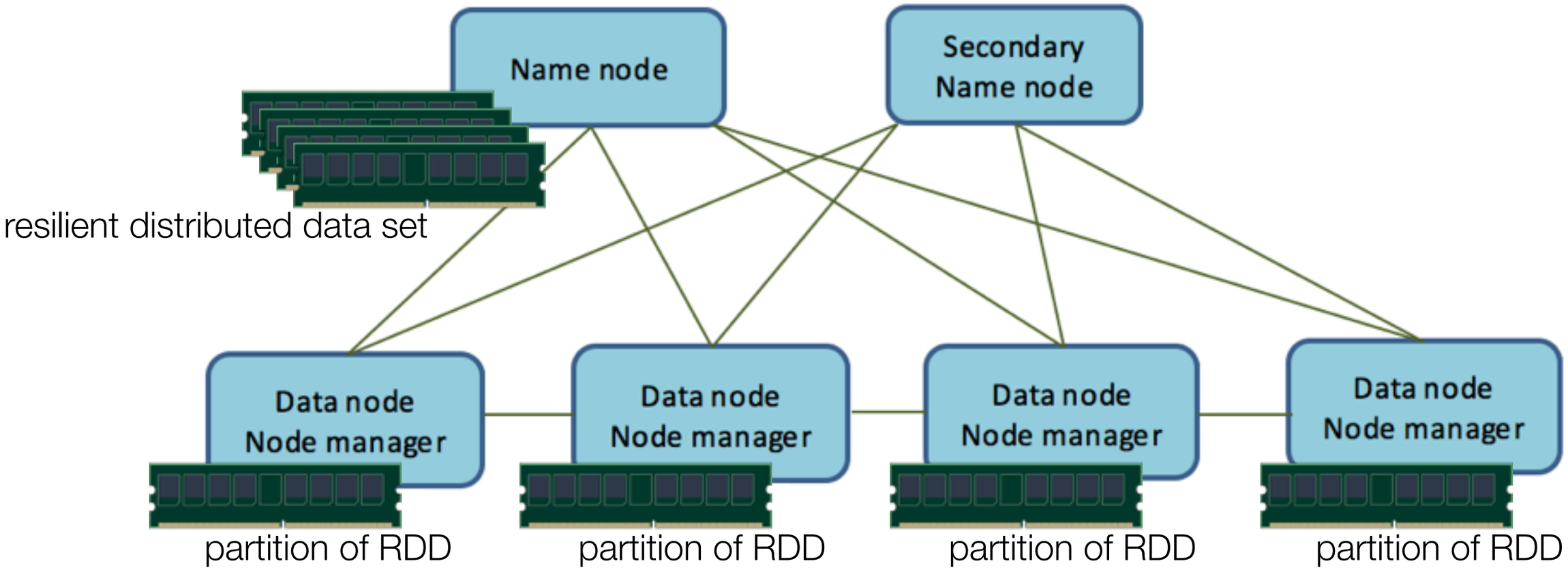
Big Data

fMRI data analysis deals with data in tera bytes and operations on large scale matrices, which breeds an ideal scenario for the use of Big Data framework such as Apache Spark.



biananes - Spark Nifti Reader Library

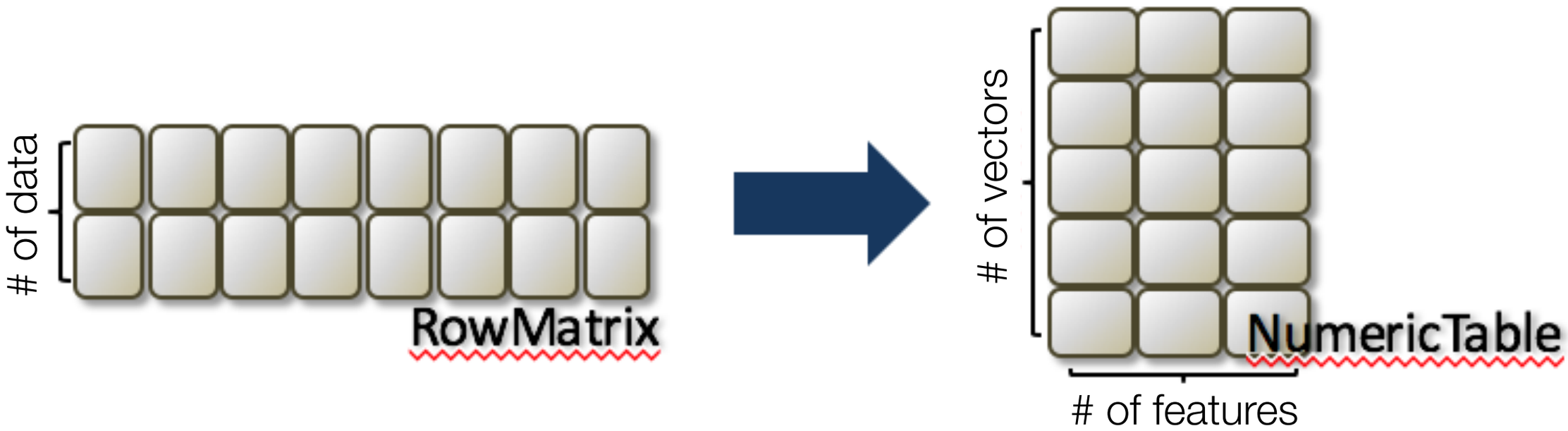
Read Nifti(.nii) to RowMatrix RDD



Incorporate with Intel DAAL

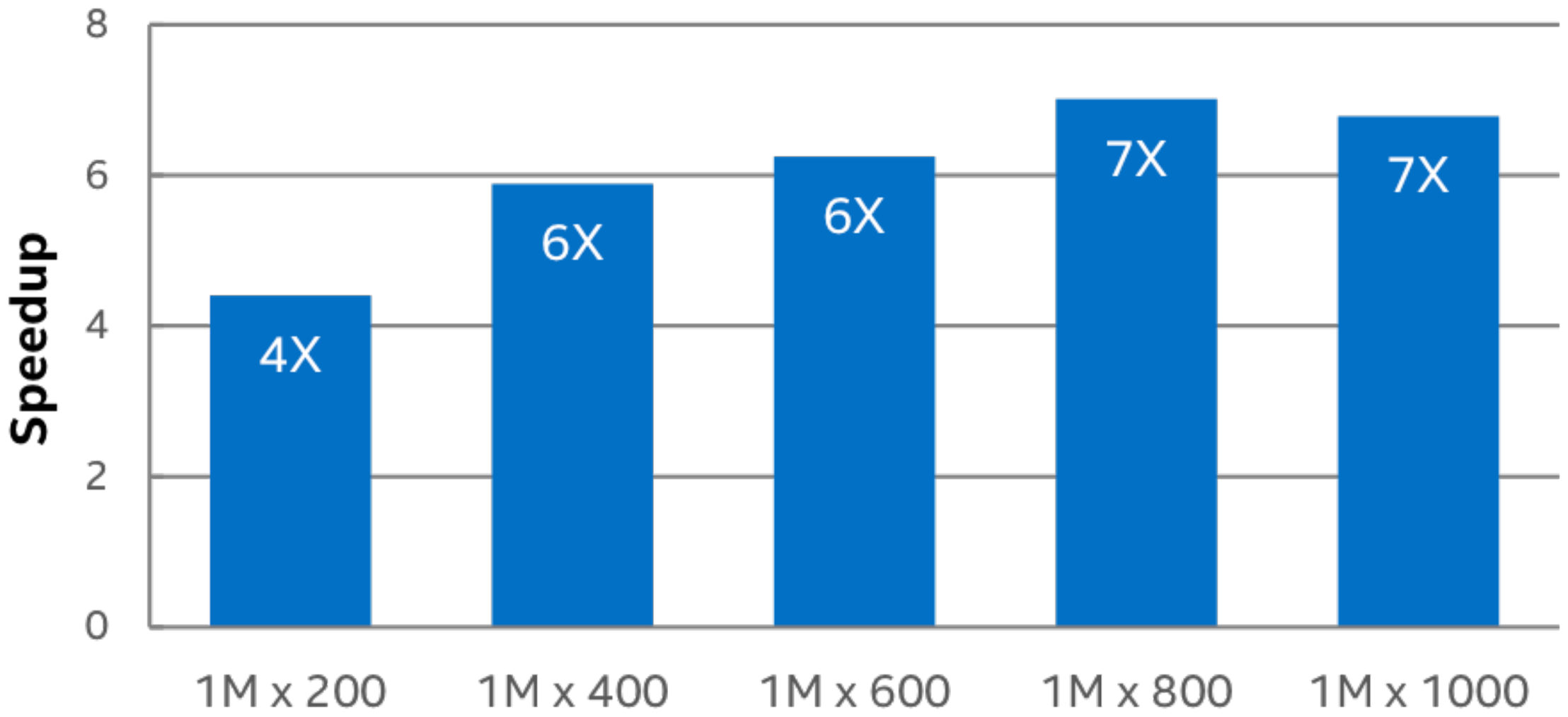
APACHE Spark MLlib

INTEL® DAAL



Supposedly, rows in RowMatrix will be represented as vectors(rows) in NumericTable.

PCA Performance Boost
Using Intel® DAAL vs. Spark* MLLib



Configuration Info - Versions: Intel® Data Analytics Acceleration Library 2016, CDH v5.3.1, Apache Spark* v1.2.0; Hardware: Intel® Xeon® Processor E5-2699 v3, 2 Eighteen-core CPUs (45MB LLC, 2.3GHz), 128GB of RAM per node; Operating System: CentOS 6.6 x86_64. PCA normalized input.

Incorporate with Intel DAAL

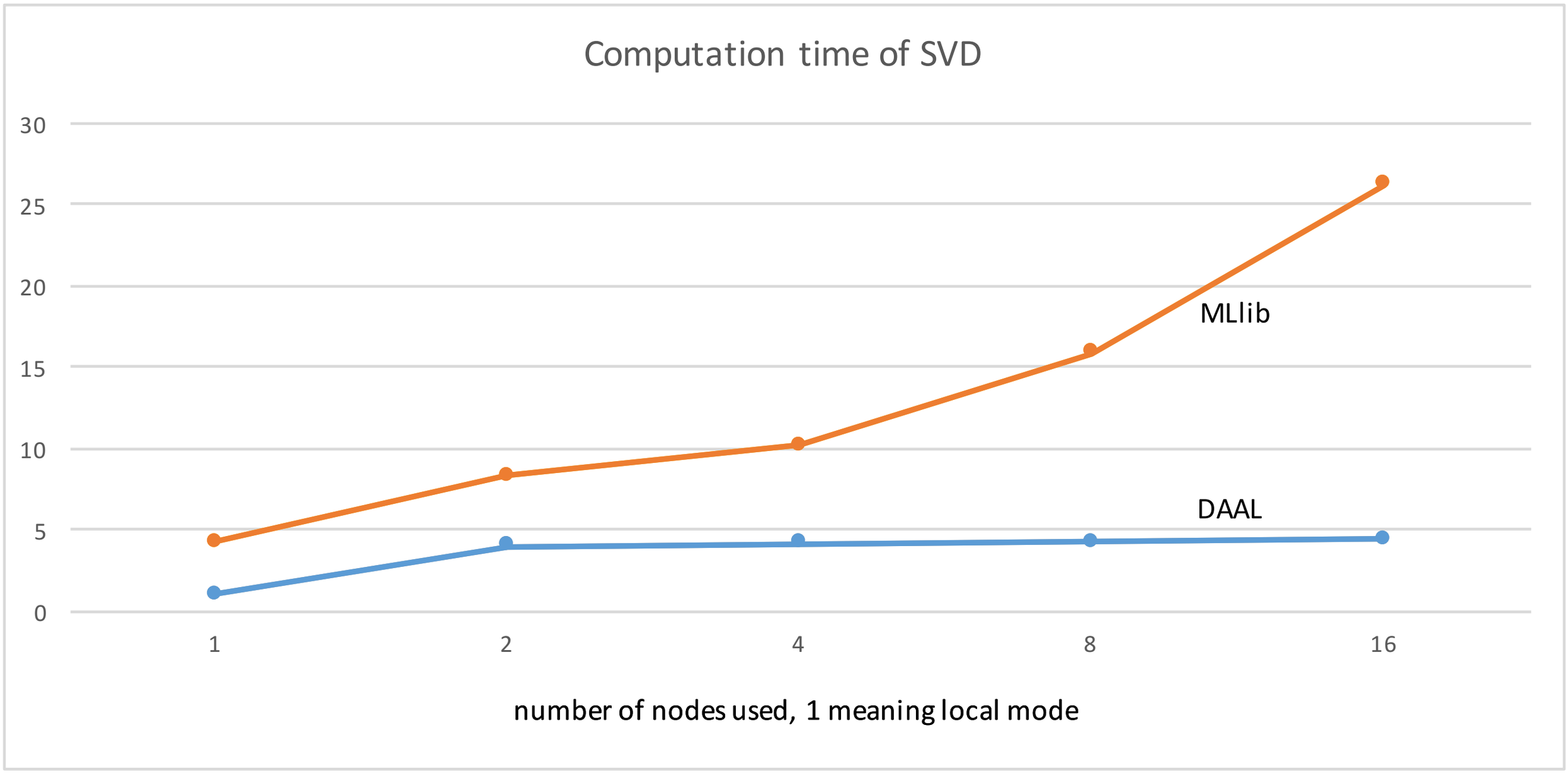
<

BLAS **MKL**

Theoretical performance boost

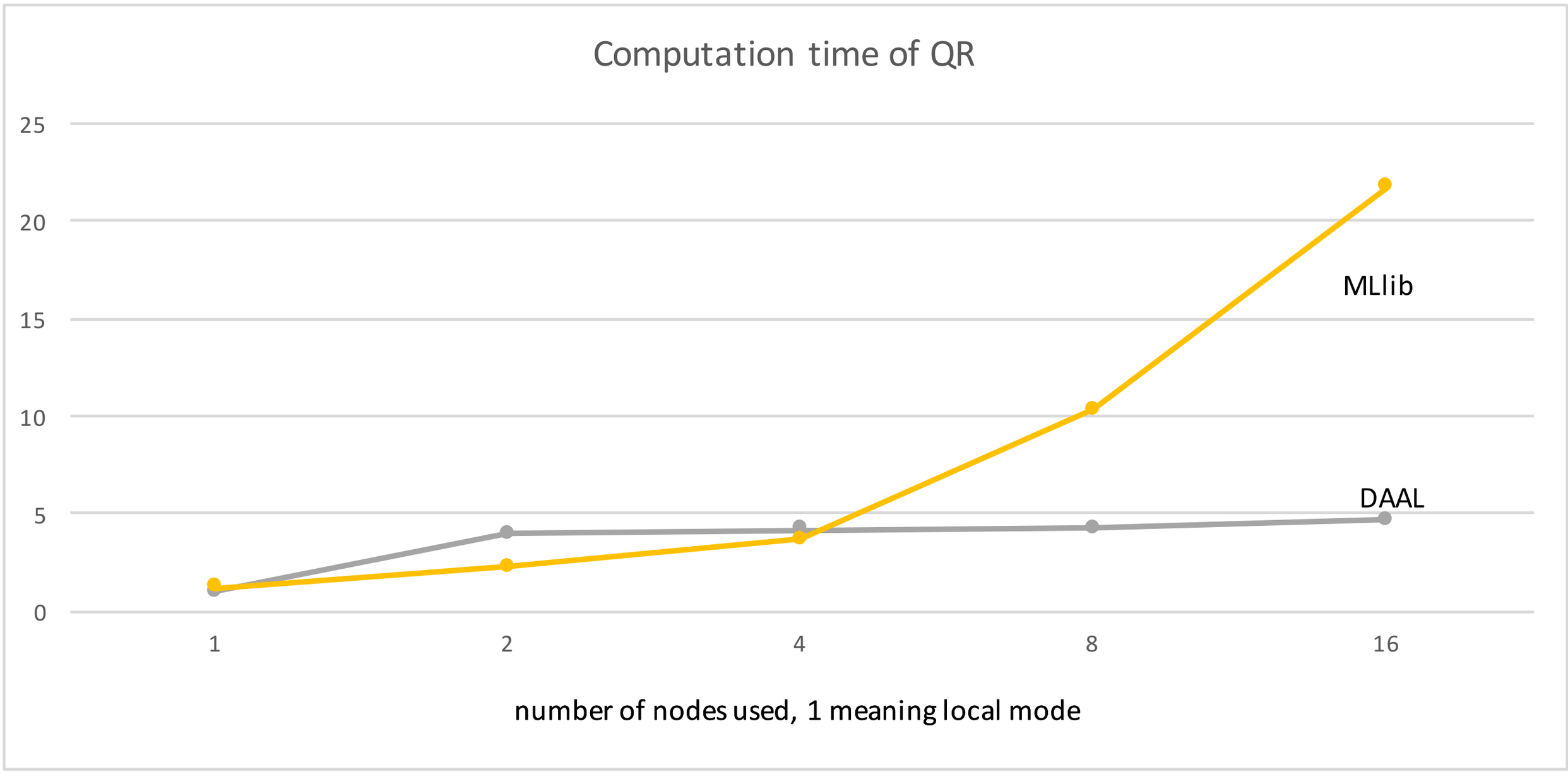
1. Intel Math Kernel Library is optimised for Intel processors
2. Set up overhead reduced in Apache Spark

Benchmarking



1. Volumes in image data are scaled up according to the number of nodes.
2. 1 volume contains a column of 480115 elements.

Benchmarking



1. Volumes in image data are scaled up according to the number of nodes.
2. 1 volume contains a column of 480115 elements.

Benchmarking Remarks

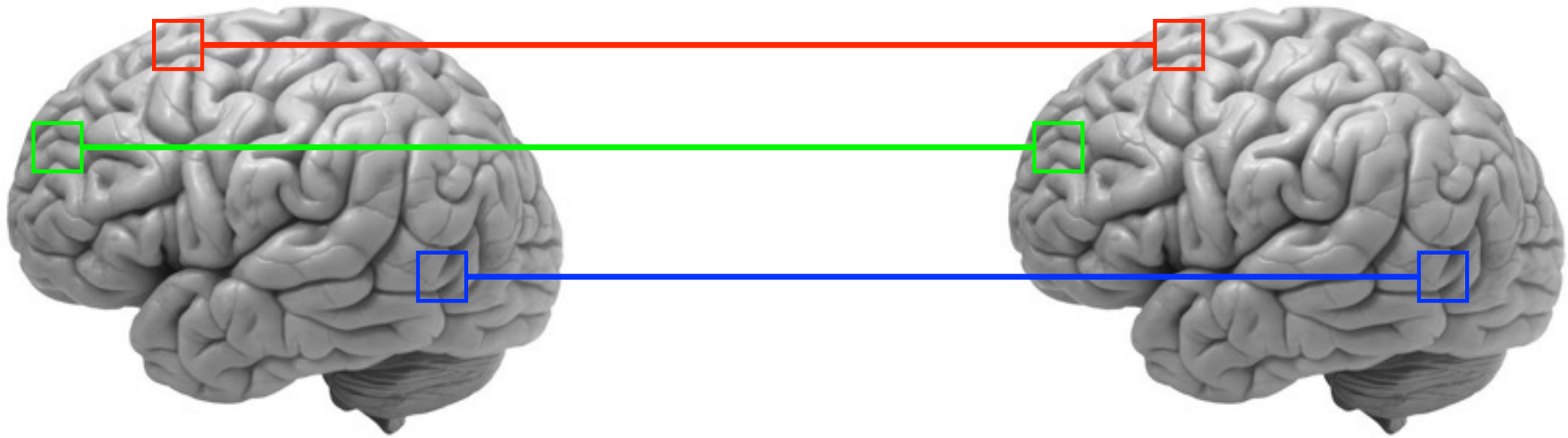
1. The above results do not give conclusive comparison between DAAL and MLlib. As the size of test data is small(~1MB), overhead for preliminary set up of the algorithms overwhelms the actual computation time.

2. Transformation of rows in RowMatrix from MLlib to vectors in NumericTable from DAAL has failed. Instead, the above results are compared under transformation in which rows in RowMatrix are represented by columns in NumericTable.

3. Computation attempts on Principal Component Analysis and Covariance were taken but failed to draw comparison as with input matrix size larger than 65535, these computations in MLlib are not supported.

Full Correlation Matrix Analysis

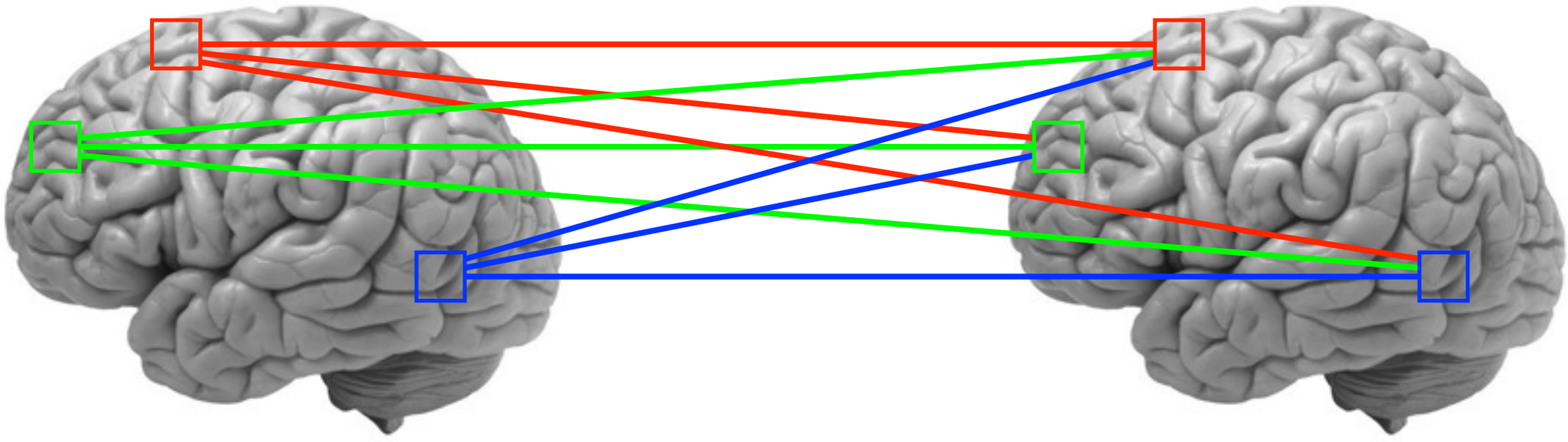
Traditional approach to fMRI analysis



activity 1 activity 2

pair-wise correlation, univariate analysis

Full Correlation Matrix Analysis



activity 1

activity 2

multivariate pattern analysis

Full Correlation Matrix Analysis

implementation pipeline with parallel computing speeds up computation for more than 50 times

1 **Correlation computation**
Pearson Correlation reduced to MM multiplication

2 **Normalisation of correlation results**
Fisher Transformation to normally distribute data

3 **Support Vector Machine**
Classifier to validate results via cross validation

Full Correlation Matrix Analysis

1. Correlation computation

Since the use of MKL is encapsulated in DAAL, direct call of Matrix-Matrix multiplication is not supported.

Resolution is to adapt to Covariance computation, which is currently an ongoing process with the need to review and refine the representation of data transformed in NumericTable.

Reference

1. rboubela, https://github.com/rboubela/biananes
2. Roland N Boubela et al., \Big Data approaches for the analysis of large-scale fMRI data using Apache Spark and GPU processing: A demonstration on resting-state fMRI data from the Human Connectome Project," Frontiers in neuroscience 9 (2015)
3. Intel, https://software.intel.com/en-us/blogs/daal
4. Yida Wang et al., Full correlation matrix analysis of fmri data, technical report
5. Wang, Yida, et al. "Optimizing Full Correlation Matrix Analysis of fMRI Data on IntelR Xeon PhiTM Coprocessors."

Acknowledgements

Support from the University of Hong Kong, the Chinese University of Hong Kong, Oak Ridge National Laboratory, Joint Institute for Computational Sciences and University of Tennessee is gratefully acknowledged. Our sincere thank to our mentors for their guidance and instructions.