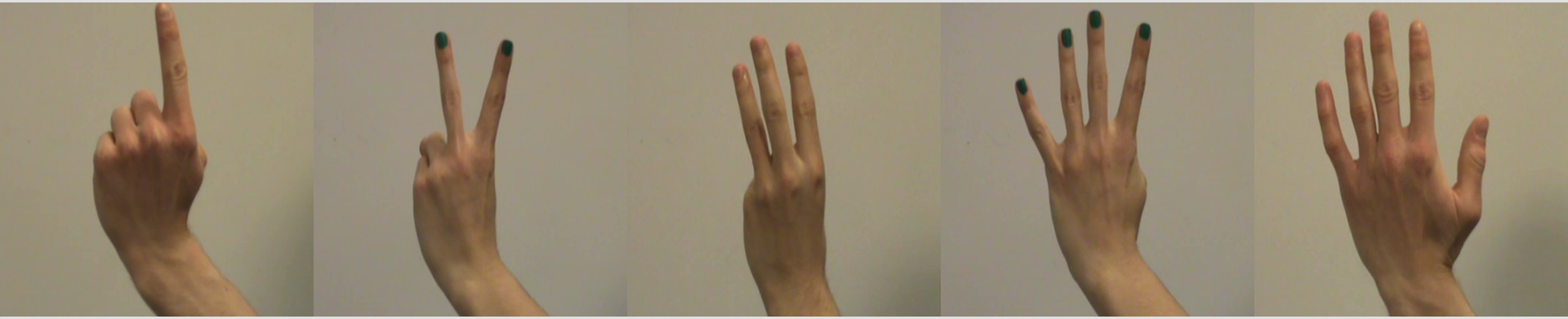
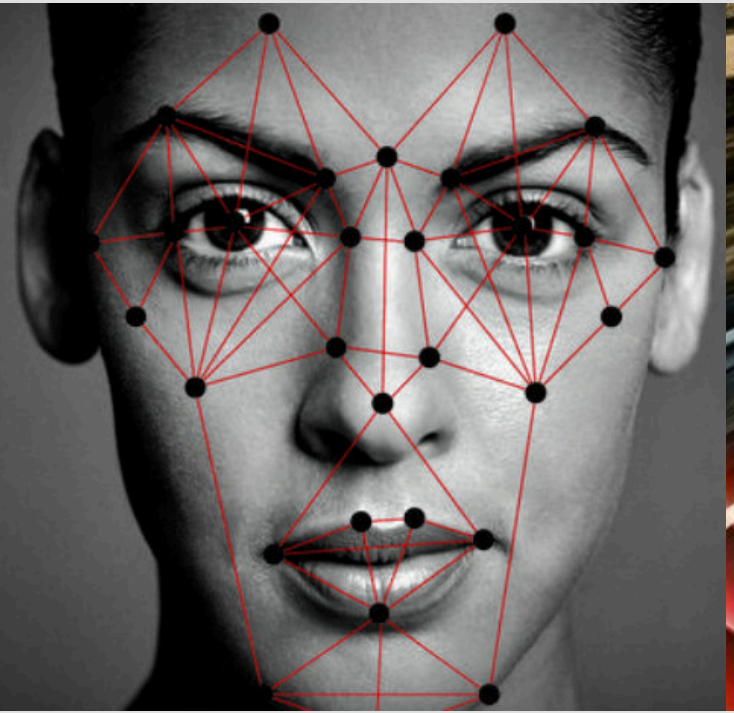


Parallelized Image Recognition in Spark & MPI

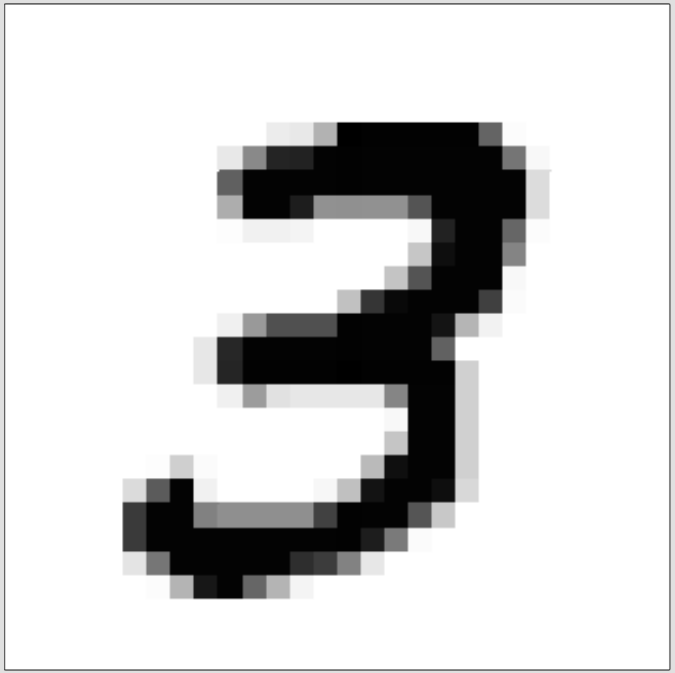
Tim Clements
Dan Cusworth
J.D. (Bram) Maasakkers

Image recognition is everywhere



Two datastreams

MNIST



28 x 28

Our own hand images

40 x 60

Multi-class linear classifier

Recognition uses a linear classification with L-2 penalty:

$$\arg\min_{w \in \mathbb{R}} \left\{ ||Xw - Y||_2^2 + \lambda ||w||_2^2 \right\}$$

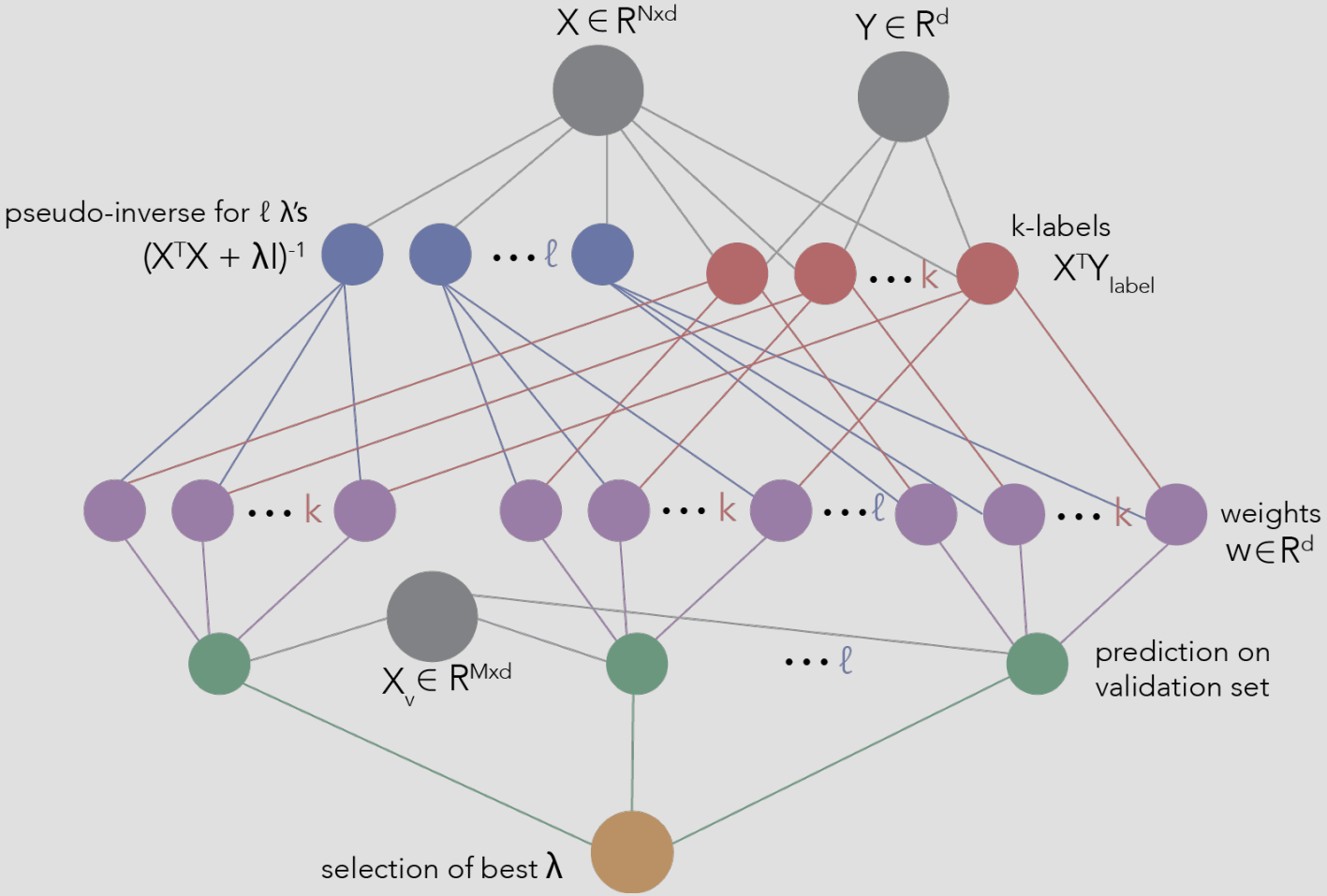
$$X \in \mathbb{R}^{N \times d}; w \in \mathbb{R}^d, \lambda \in \mathbb{R}, y_i \in \{-1, 1\}$$

Where $\boldsymbol{\lambda}$ is a tuning parameter, to be determined separately, and weights $\boldsymbol{w}$:

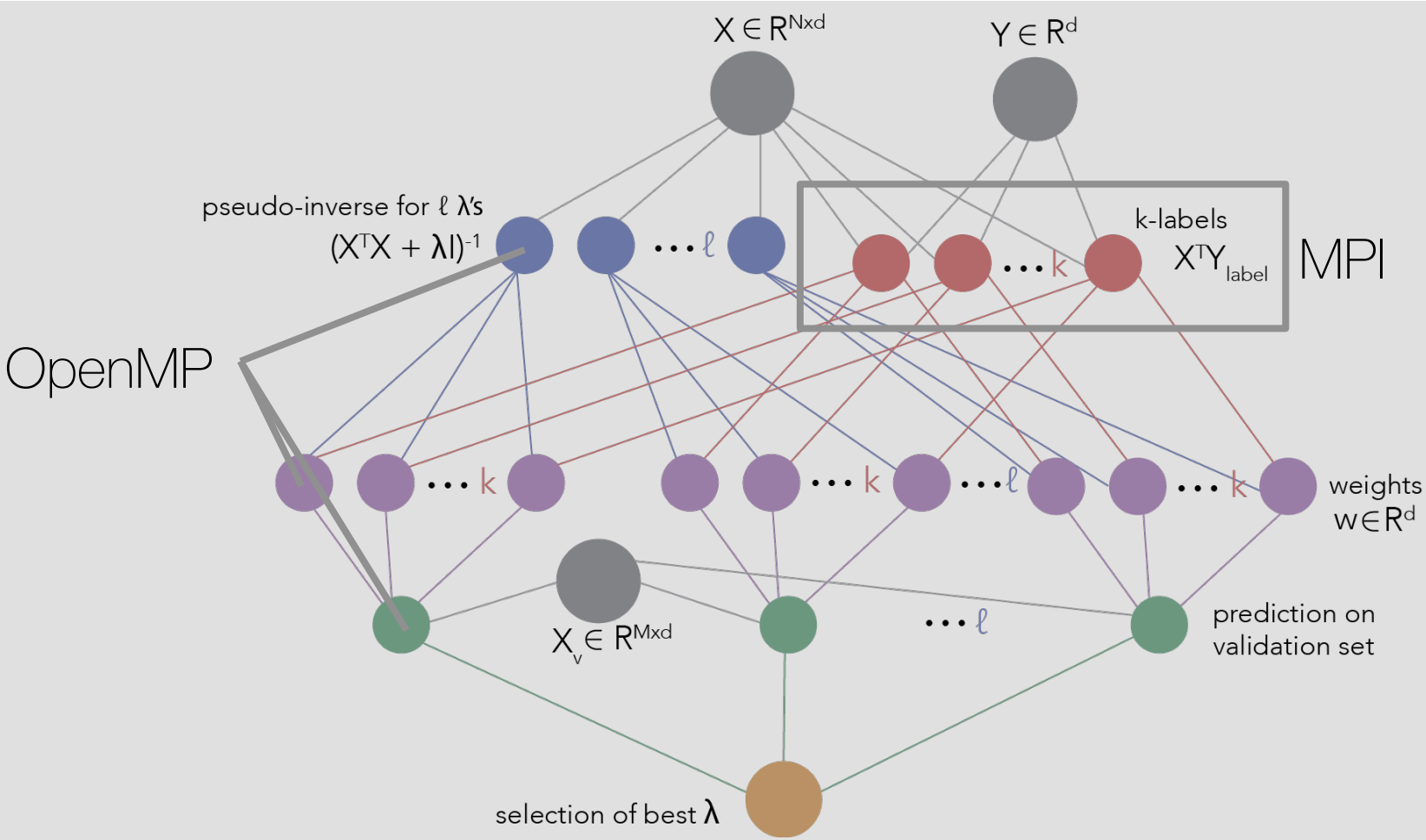
$$w^* = \left(X^T X + \lambda I_d \right)^{-1} X^T Y$$

Weights need to be fitted for all $\boldsymbol{k}$ classes.

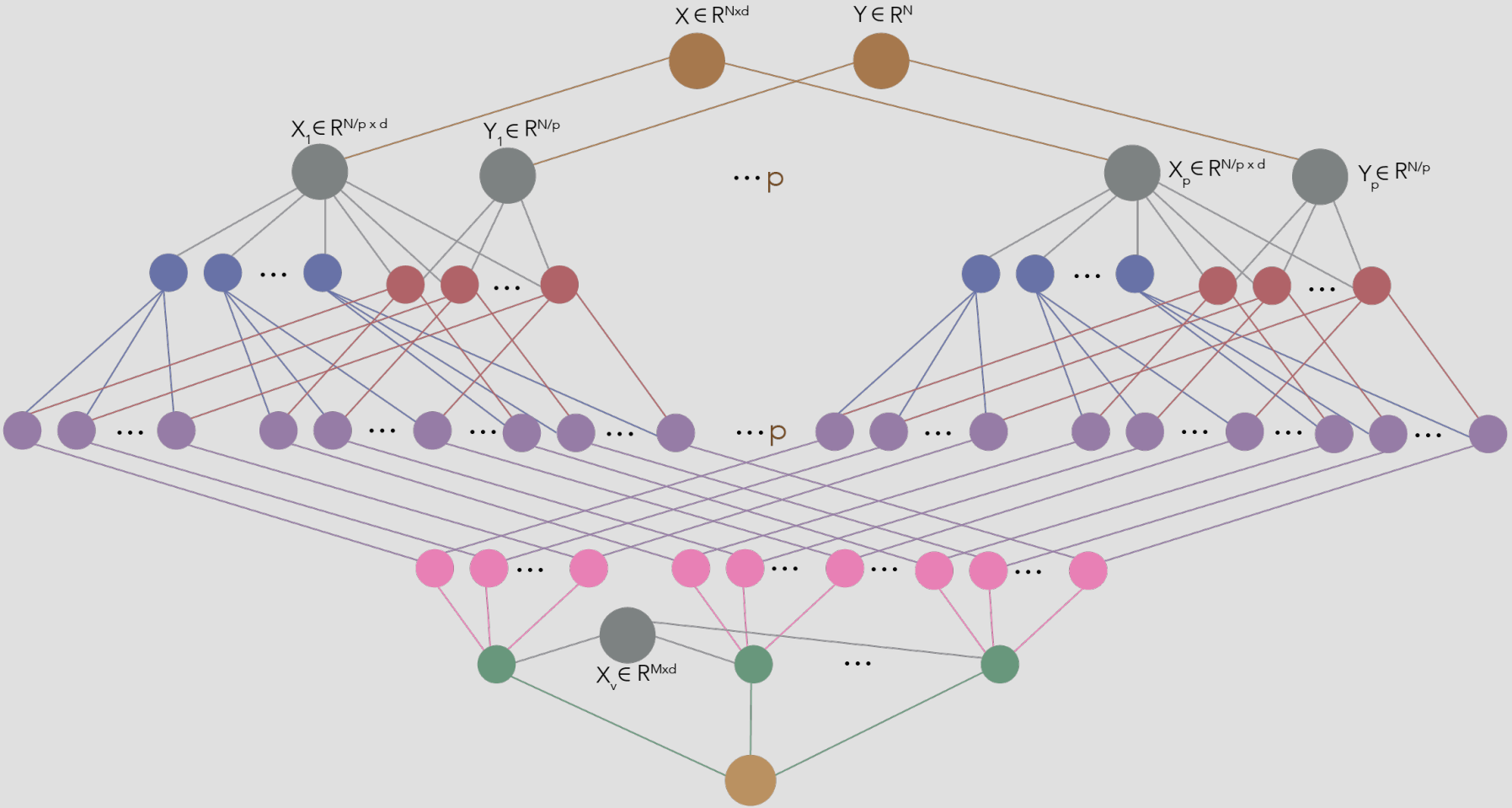
Computational Graph



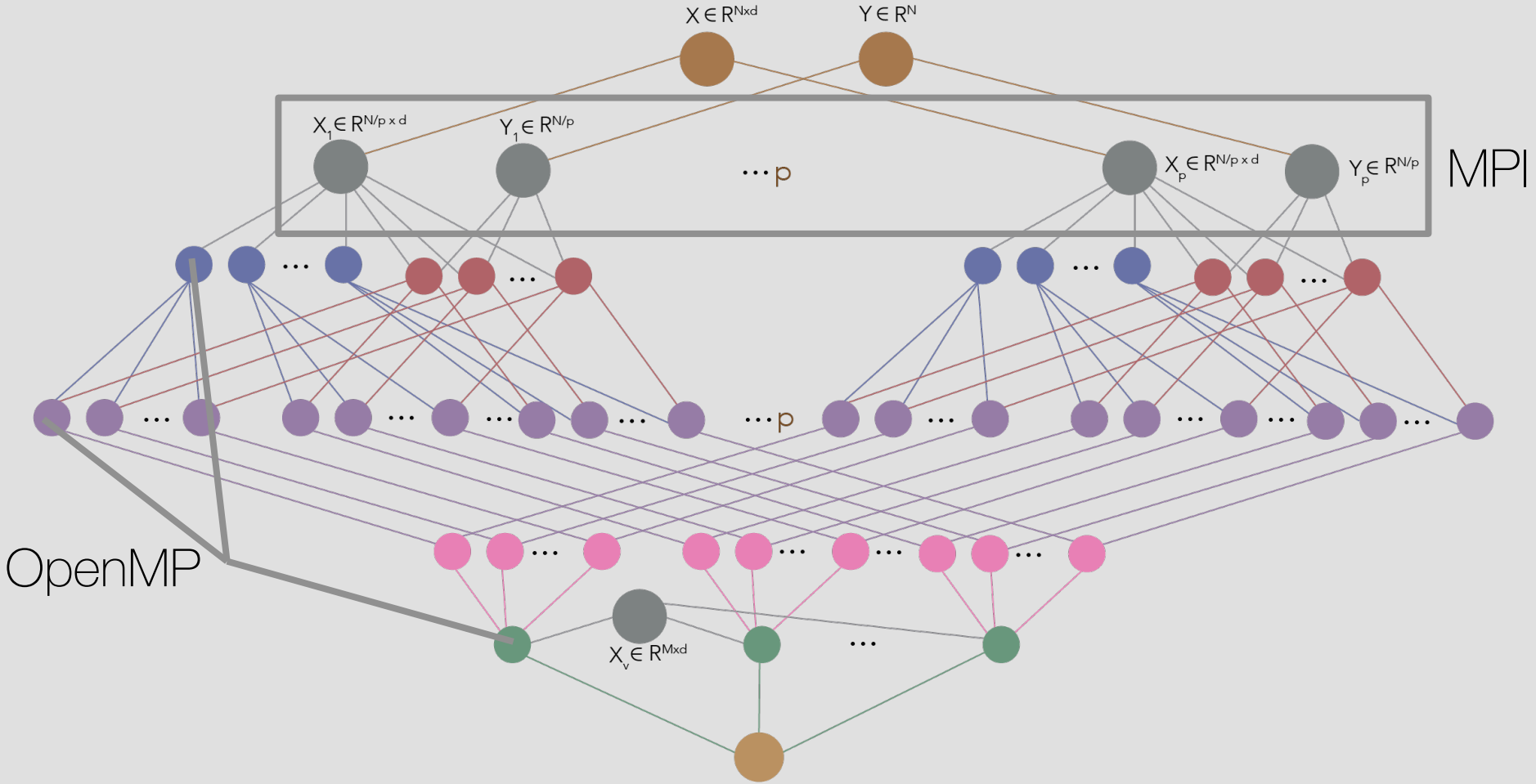
Computational Graph



Computational Graph – Data Parallel



Computational Graph – Data Parallel

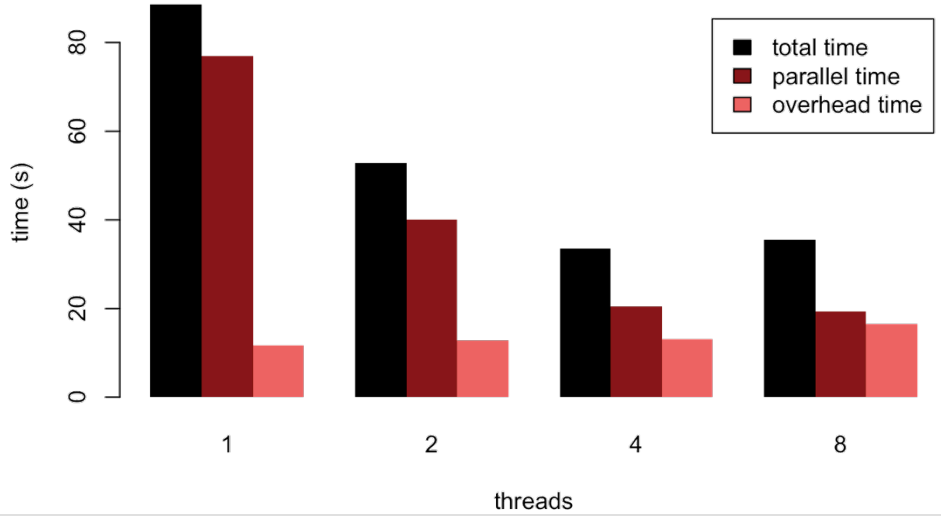


Parallel implementation

OpenMP + MPI	SPARK
Hybrid parallelism using MPI and Cython	Functional parallelism
Odyssey (8 nodes with 8 threads)	AWS EMR (m2xlarge, 4 workers)

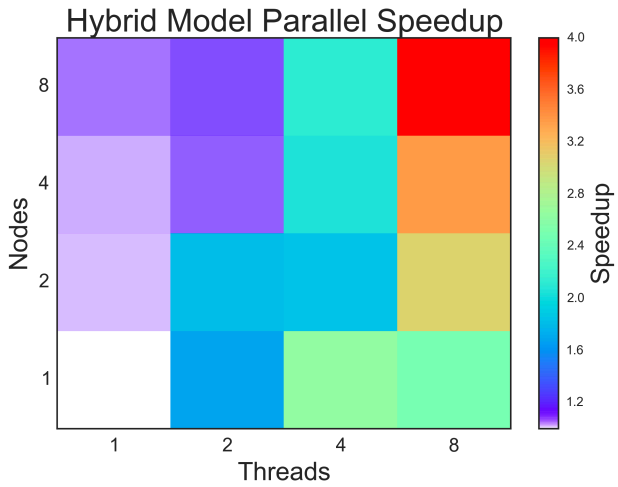
OpenMP parallelism analysis

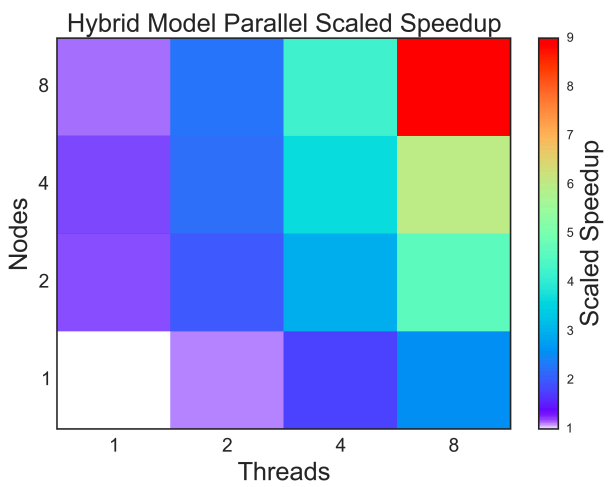
Speedup is approaching a constant after 8 nodes, so we are reaching maximum speedup

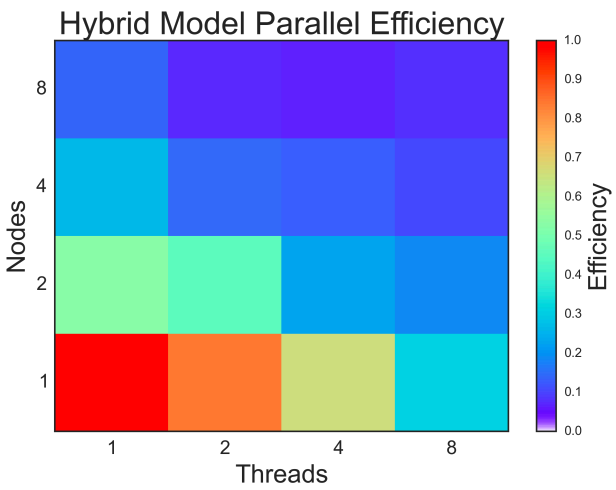


Hybrid Model Parallelism (OpenMP + MPI)

Optimal speedup occurs for 8 nodes/threads, but efficiency drops.

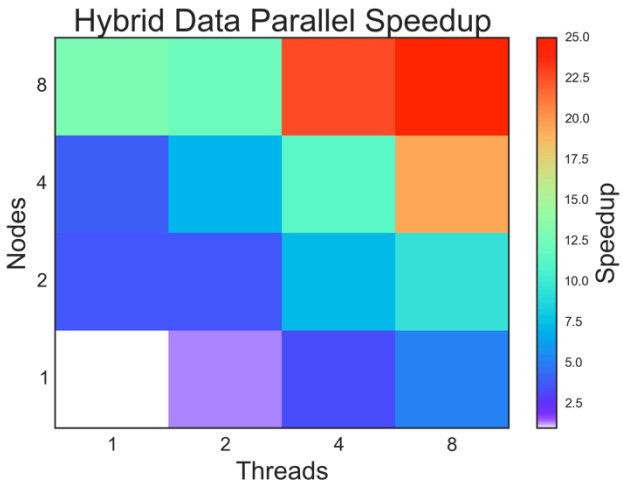


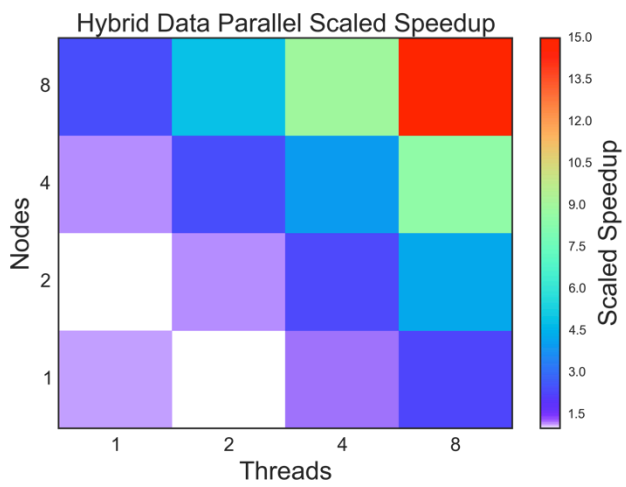


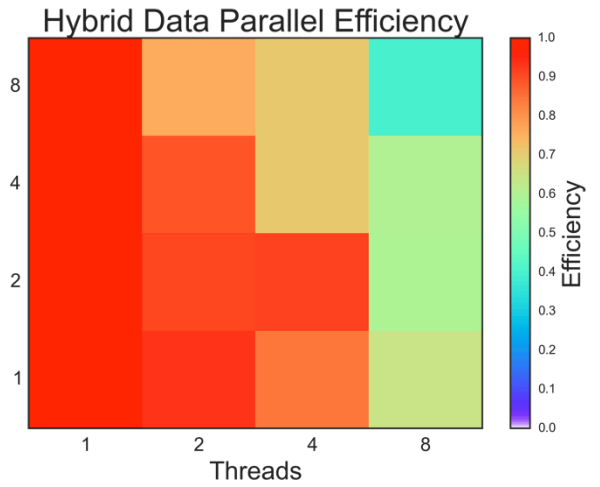


Hybrid Data Parallelism (OpenMP + MPI)

We see speedups of 25x, and efficiency is better sustained.

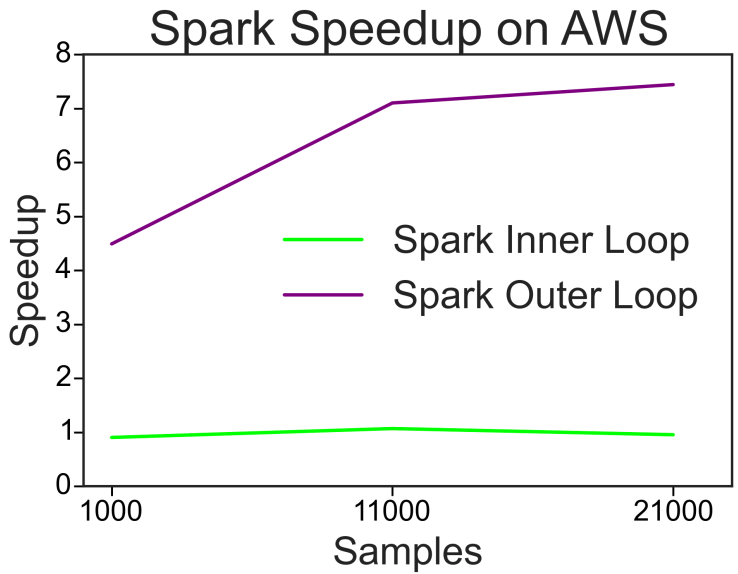






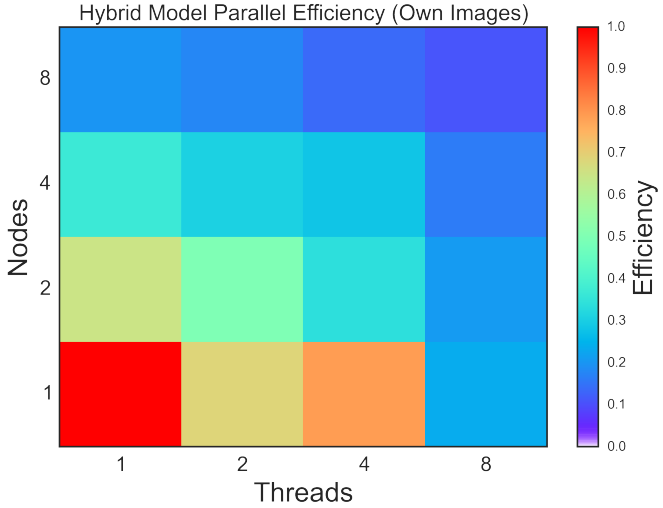
Advanced Feature – Spark parallelism

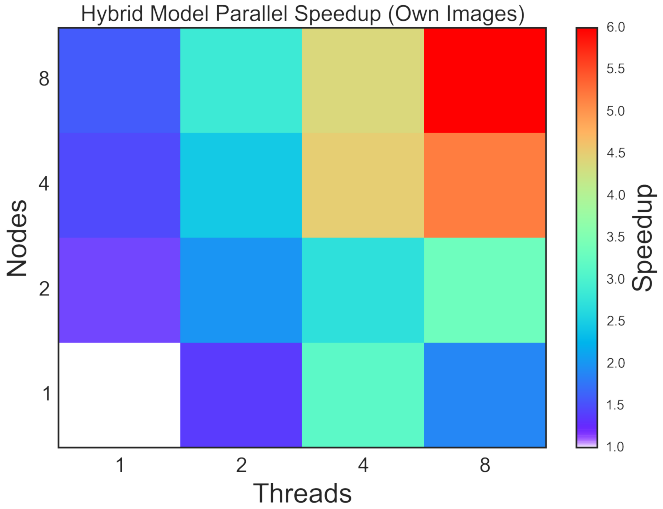
Spark speedup is between MNIST model and data parallelism.



Advanced Feature - Own images: 87% Accuracy

We see greater speedup and efficiency with our own images than with the MNIST.



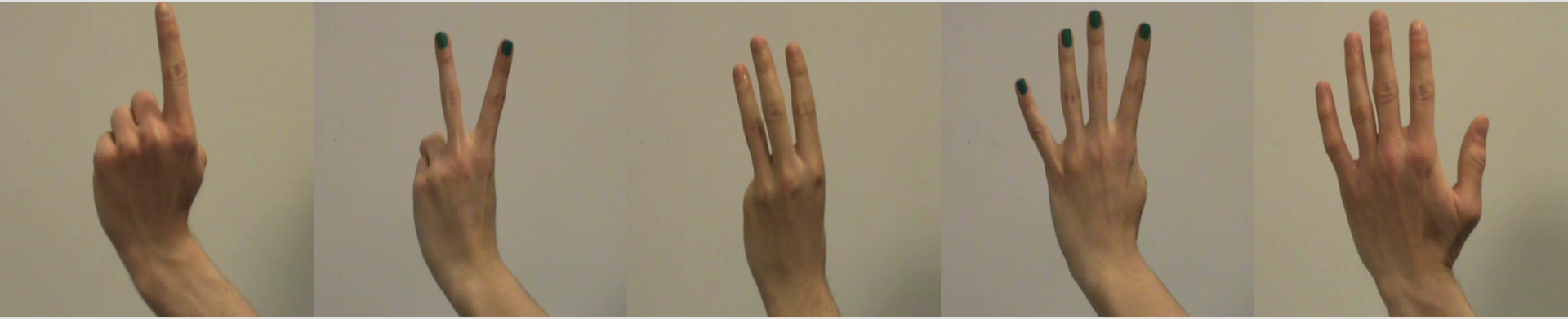


Conclusions

Using a simple regularized linear classifier, we obtain good predictive ability on the MNIST dataset and our own images.

The greatest computational bottleneck is in the computation of the pseudo-inverse.

Hybrid data parallel OpenMP + MPI gives the best performance. Spark performs between model and data parallel.



Parallelized Image Recognition in Spark & MPI

Architecture	Odyssey (8 Nodes with 8 Threads) & AWS EMR (m2xlarge, 4 workers)
Hybrid Parallelism	OpenMP (Cython) + MPI, both model and data parallel
Advanced Features	Spark on AWS & Own images
Showed	Parallelism analysis, weak/strong scaling, efficiencies, and computation graphs

Tim Clements, Dan Cusworth, and J.D. (Bram) Maasakkers