NeuroData SIMPLEX Report: Q1 2017

The following report documents the progress made by the labs of PI Joshua T. Vogelstein and Co-PIs Randal Burns and Carey Priebe at Johns Hopkins University towards goals set by the DARPA SIMPLEX grant.

Contents

1 Statistical Theory and Methods

1.1 LOL @jovo

This month we finalized the real data analysis using LOL. In particular, we considered four datasets, the Prostate and Colon datasets have extensively been studied in the sparse literature. LOL yields better perforance, and for Colon, with lower dimensionality. MNIST is an even more prominent dataset, LOL achieves the best performance for all dimensions. MRN is a new dataset that we generated; it has over 500 million features, and 112 samples. We subsampled to 100 samples for cross-validation purposes. To our knowledge, no other machine learning tool is capable of even operating on 500 million features. Moreover, we demonstrate that our implementation outperforms first doing PCA on the data, for any number of dimensions that we embed into. We then also investigate the amount of time it takes to run LOL on very wide datasets. For a 128 million dimensional dataset, with 2000 samples, requiring nearly half a terabyte of space just to store, we have an approximate implementation that runs on a single machine and only takes about 3 minutes.

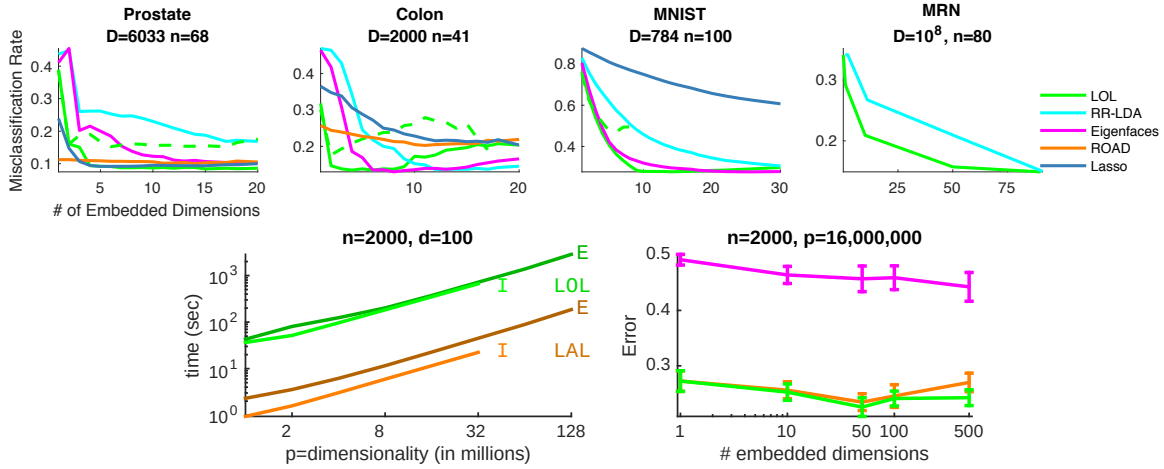

Figure 1: Top: LOL outperforms essentially all other methods for essentially all problems and number of dimensions to embed into. Bottom: LOL (green) and LAL (brown, using random projections to approximate) for both fully in memory (IE) and semi-external memory (E) implementations. Magenta compares performance to our scalable implementation of eigenfaces on the same problem.

1.2 meda @JesseLP

1.3 Multiscale Generalized Correlation (MGC)

We developed the Multiscale Generalized Correlation method to better detect associations between two datasets X and Y. We demonstrate that Oracle MGC is a consistent test statistic (power converge to 1 as sample size increases) under standard regularity conditions, is equivalently to the global correlation under linear dependency (i.e., each observation X_i is a linear transformation of Y_i), and can be strictly better than the global correlation under common nonlinear dependencies. Thus Oracle MGC dominates the global correlation, and the sample MGC (i.e., choose the optimal scale by p-value map approximation, as the testing power are not available in the absence of the true model and training data) also empirically dominates the global correlation. A flowchart to illustrate the advantage of MGC is shown in Figure 2.
Upon final draft polishing and addressing feedback from statisticians and biologists, the newest draft is updated to arXiv and submitted for publication.

1.4 Network Dependence Test via Diffusion Maps and MGC

Deciphering the association between network structures and corresponding nodal attributes of interest is a core problem in network science. We propose a new nonparametric procedure for testing dependence between network topology and nodal attributes, via diffusion maps and `MGC`. Specifically, under an exchangeable graph, we verify that the diffusion maps provide a set of conditionally independent multivariate coordinates for the nodes, which can be combined with `MGC` (or in general, any distance-based correlation measures) to yield consistent statistic for network dependence testing. Moreover, our method is computationally inexpensive and robust against parameter mis-specifications, very efficient in capturing a wide variety of nonlinear and high-dimensional relationships, and readily extend-able to testing independence between two graphs.
Figure 3 illustrates the advantage of the proposed method on testing dependency between two graphs. The graphs are simulated by the random dot product graph, with the underlying latent variables being related by a quadratic function. By repeatedly generating dependent sample graphs, the testing power equals the percentage of rejection of the independence hypothesis. Although all methods are consistent (having power 1 as number of vertices increases), the proposed approach using `MGC` is able to achieve perfect testing power at a very small size, which is significantly better than other benchmarks.
An early draft is recently awarded the Best Student Paper Awards by the American Statistical Association Nonparametric Statistics Section, which will be presented in a special section in the Joint Statistical Meeting this year. We collected and addressed feedback from experts in graph inference, and submitted the complete manuscript this month.

1.5 Randomer Forest (RerF)

Most recently, we have begun investigating the theoretical behavior of RerF. Mathematical analysis of the RF and RerF procedures is difficult. Therefore, we have started with simplified procedures. The main simplifications we have made are that trees only have a depth of one

(also known as decision stumps), and that each tree is trained on the full training set rather than a bootstrap sample. In the figure, we surprisingly see that RerF outperforms RF across a variety of settings (see caption for more details).
The R version of RerF is now functional and is an order of magnitude faster than the Matlab implementation. This version of RerF allows the user to specify the minimum size of a node and a parameter to tweak the rotation matrix. Additional basic functionality is being added to this tool including bagging, out-of-bag error reporting, max tree depth, and pruning.

1.6 Non-Parametric Shape Clustering

Energy statistics provides a nonparametric test for equality of distributions. It is rotational invariant which is a highly desirable quality for clustering. For a two-class problem, $X, X' \sim \mu$ and $Y, Y' \sim \nu$, where μ, ν are CDFs, it reads

$$\mathcal{E}(X, Y) = 2\mathbb{E}\|X - Y\| - \mathbb{E}\|X - X'\| - \mathbb{E}\|Y - Y'\|. \tag{1}$$

We are developing a clustering framework based on (1). Our criteria is that $\mathcal{E}$ should be a maximum when data points are correctly classified. It is possible to show that there is a map from the data space of X, Y to the probability space of μ, ν which is a Hilbert space whose inner product can be obtained from a kernel function related to (1), $\langle \mu, \nu \rangle = k(x, y)$. This enables us to formulate our clustering problem as follows:

$$\max \left\{ \mathrm{Tr} L^{1/2} Z^T K Z L^{1/2} \right\} \quad \text{s.t.} \quad Z_{ij} \in \{0, 1\}, \sum_i Z_{ij} = N_j, \sum_j Z_{ij} = 1, Z^T Z = L^{-1} \tag{2}$$

where N_j is the number of elements in the jth cluster, $L^{-1} = \mathrm{diag}(N_1, \ldots, N_k)$, and K is the Gram matrix obtained from the kernel. Let $\mathcal{X}$ be the pooled data matrix. If we replace $K \to \mathcal{X}^T \mathcal{X}$ we recover the well-known k-means problem, which in this formulation is related to spectral clustering and normalized cuts. Problem (2) is NP-hard and a numerical implementation is prohibitive even for small data sets. We are investigating how to solve (2) in a feasible way. As an evidence that (2) is the correct optimization problem, and more importantly, it illustrates the power behind our proposal, in Fig. 5 we generate data and plot the objective in (2) versus n, where n is the number of points randomly shuffled from one class to the other. Therefore, for $n = 0$ the function must be a maximum. We do this for the kernel related to (1) (blue dots) and compare with the kernel related to k-means (red dots). Clearly, (2) based on (1) is able to distinguish between different cluster even for complex data sets that are not linearly separable. Moreover, in our formulation there are no free-parameters in the kernel.

1.7 Discriminability

We develop a measure of discriminability (or reliability). It is intuitive to understand and easy to implement. Discriminability is defined to be the probability of within subject distances being smaller than the cross subject distances. If we let $x_{i,t}$ denote the t$^{\text{th}}$ trial of subject i and $\Delta(,)$ be the metric, the (population) discriminability D is:

$$D := P(\Delta(x_{i,t}, x_{i,t'}) \leq \Delta(x_{i,t}, x_{i',t''})).$$

Previously, we search for the optimal processing pipeline which has the maximal discriminability.

Currently, we are considering the discriminability from a different perspective. Specifically, we want to use the discriminability as an internal measure of the consistency in clustering. If we let $x_{i,t}$ denote the t$^{\text{th}}$ sample of cluster i and $\Delta(,)$ be the metric, the discriminability D is: $D := P(\Delta(x_{i,t}, x_{i,t'}) \leq \Delta(x_{i,t}, x_{i',t''}))$. Large discriminability implies the clusters are more consistent, that is we can better differentiate samples from different clusters.

We are also developing a clustering algorithm 1 which maximizes the discriminability. The algorithm will assign each sample i a cluster identity k_i such that the discriminability is maximized. The algorithm is similar to K-means, but we believe discriminability provides a more robust measure than within cluster distances which is maximized in K-means.

Algorithm 1 Cluster samples through maximizing discriminability.

Require: Samples $\{\boldsymbol{x}_i\}$ and number of clusters K.
Ensure: Cluster identity $\{k_i\}$.

```
function DiscriminabilityClustering
    Initialize {k_i} randomly
    while not convergent do
        for i do
            for j in 1 : K do
                Set k_i = j
                ComputeDiscriminability({x_i},{k_i})
            Set k_i to the cluster with maximum discriminability
    Output {k_i}
```

1.8 Law of Large Graphs

1.9 Robust Law of Large Graphs

Although we only present the results under exponential distributions, the results can be generalized to a broader class of distribution families, and even a different entry-wise robust estimator other than MLqE with the following conditions:

1. Let $A_{ij} \overset{ind}{\sim} (1-\epsilon) f_{P_{ij}} + \epsilon f_{C_{ij}}$, then $E[(A_{ij} - E[\hat{P}_{ij}^{(1)}])^k] \leq \text{const}^k \cdot k!$, where $\hat{P}^{(1)}$ is the entry-wise MLE as defined before;
   This is to ensure that observations will not deviate from the expectation too far away, so that the concentration inequalities hold.
2. There exists $C_0(P_{ij}, \epsilon) > 0$ such that under the contaminated model with $C > C_0(P_{ij}, \epsilon)$,

$$\lim_{m\to\infty} \left| E[\hat{P}_{ij}] - P_{ij} \right| < \lim_{m\to\infty} \left| E[\hat{P}_{ij}^{(1)}] - P_{ij} \right|;$$

   It requires the contamination of the model to be large enough (a restriction on the distribution) and $\hat{P}$ to be robust enough with respect to the contamination (a condition on the estimator).

3. $\hat{P}_{ij} \leq \text{const} \cdot \hat{P}_{ij}^{(1)}$;
   Since we use the results of $\hat{P}^{(1)}$ to bound $\hat{P}^{(q)}$, the proof can apply directly with this condition for an arbitrary $\hat{P}$.
4. $\text{Var}(\hat{P}_{ij}) = O(m^{-1})$, where m is the number of observations.
   We will get exactly the same results based on this order. However, even if the variance of the new estimator is not of order $O(m^{-1})$, we will get similar results with a different term related to m.

Previously we consider the model to be based on exponential distribution, which is continuous and monotone. Now we consider Poisson distribution instead. Poisson distribution is a commonly used distribution for nonnegative graphs with integer values. And we will prove that it satisfies the conditions for generalization and as a result all theories apply directly.
Let $A_{ij} \overset{ind}{\sim} (1-\epsilon)f_{P_{ij}} + \epsilon f_{C_{ij}}$ with f to be Poisson, then we proved that $E[(A_{ij} - E[\hat{P}_{ij}^{(1)}])^k] \leq \text{const}^k \cdot k!$, where $\hat{P}^{(1)}$ is the entry-wise MLE as defined before. So Condition 1 is satisfied. Intuitively, since exponential distribution has a fatter tail compare to Poisson, we should have the bound for central moment of Poisson directly from the results for exponential distribution. Condition 2 is satisfied as long as the contamination is large enough while keep using the robust MLqE. For Condition 3, the extreme case happens when there are m data $x_1, \cdots, x_m$ with $0 \leq x_1 = \cdots = x_k \leq \bar{x} \leq x_{k+1} = \cdots = x_m \leq m\bar{x}/(m-k)$. In order to have ML$q$E larger than MLE $\bar{x}$, we need the weights of the first m data to be smaller than the weights of the rest $m-k$ data. So $e^{-\bar{x}} < \bar{x}^{x_m} e^{-\bar{x}}/x_m!$. Then $x_m! < \bar{x}^{x_m}$. By the lower bound in Stirling's formula, we have $x_m < e\bar{x}$ when $x_m > 0$. Note that if $x_m = 0$ then MLE equals MLqE since all data equals zero. Thus MLqE is bounded by $e\bar{x}$. As a result, $\hat{P}_{ij} \leq e\hat{P}_{ij}^{(1)}$ and Condition 3 is satisfied. At last, Condition 4 follows directly from theory of minimum contrast estimators.
So for all the theorems proved before, we can replace the exponential distribution by Poisson distribution and all the results still hold.

1.10 Batch effect removal in dimension reduction of multiway array data

1.11 Reduced Dimension Clustering

1.12 Graph-testing

2 Scalable Algorithm Implementations

2.1 FlashX

We advance FlashR for exploratory analysis on a billion-scale graph. This includes functions for filtering data points and plotting histograms and heatmaps on billions of data points. Figure 6 shows an example of plotting the in-degree histogram of vertices in the page graph and a heatmap for the distribution of the vertices of the page graph in a two-dimension space.

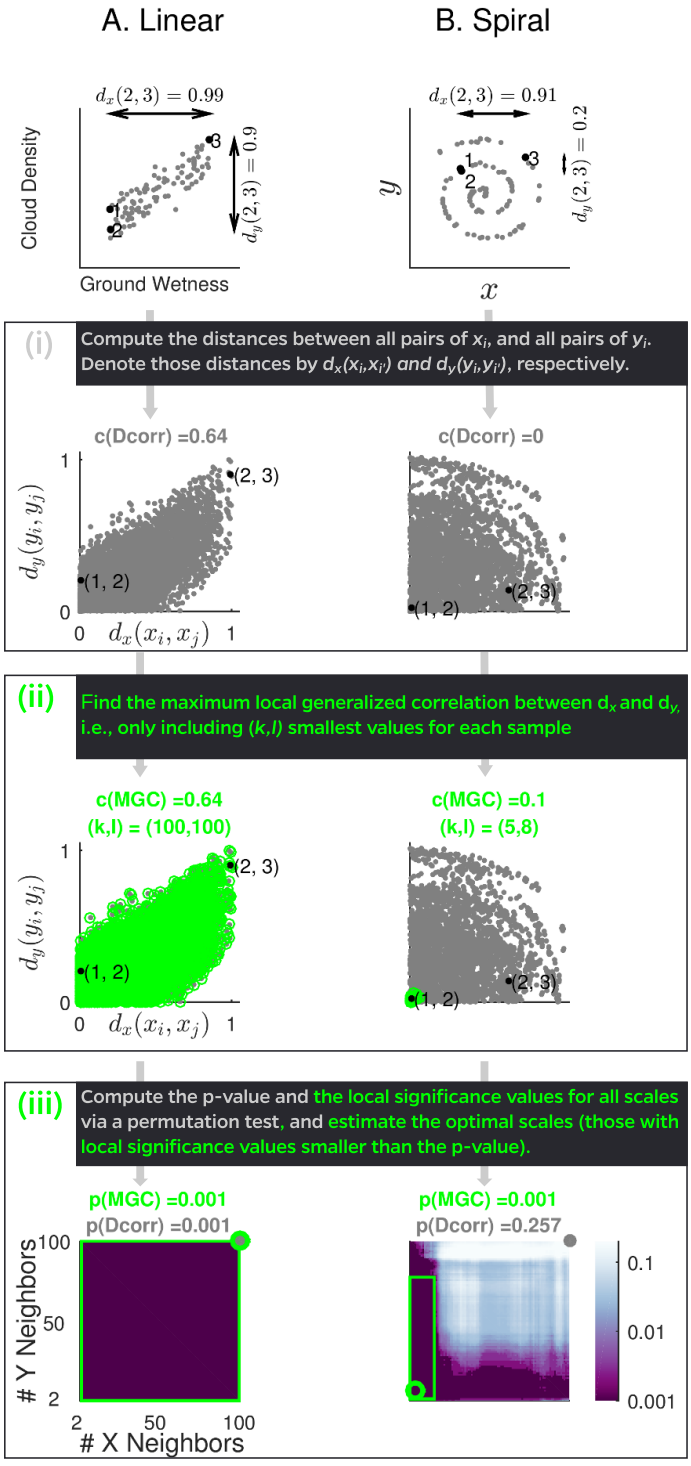

Figure 2: A flowchart to illustrate the advantages of MGC.

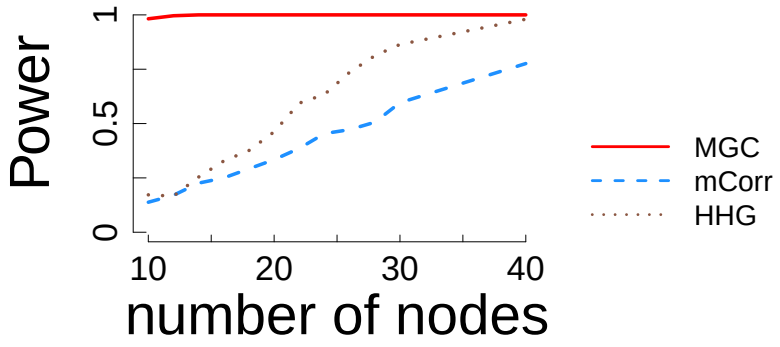

Figure 3: The power curve with respect to increasing number of vertices for the two-graph dependency testing simulation. The proposed approach quickly attains perfect power at a very small vertex size, while other benchmarks often require a much larger graph for perfect testing.

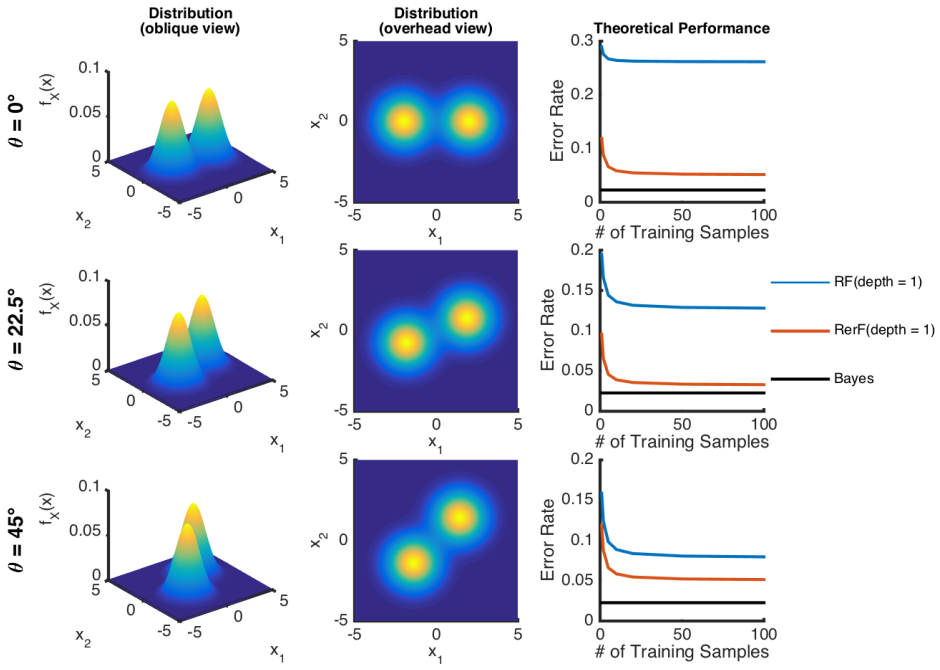

Figure 4: Theoretical performances of simplified RF and RerF models on a simple two-dimensional binary classification problem. In the top row, the two classes are distributed according to normal distributions having means that differ only in the first dimension and both having identity covariances. These distributions are shown in the left and middle panels. The right-most panel shows the theoretical error rate as a function of the number of training samples for RF and RerF. The Bayes optimal error rate is also shown for reference. The middle and bottom rows are the same as the top row, except the distributions have been rotated by 22.5° and 45° respectively. In all three cases, the RerF classifier outperforms RF for all training set sizes.

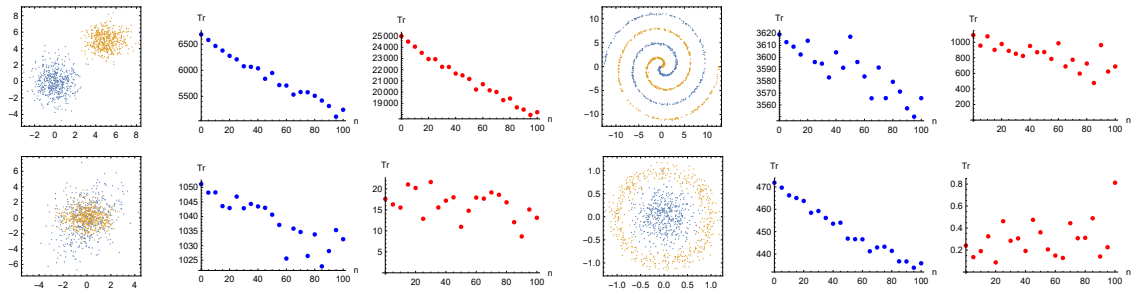

Figure 5: Two dimensional datasets and the objective function in (2) as a function of n, where n is the number of shuffled points from its correct class to the wrong class. Blue dots are for (1) and red dots for k-means. A good function must be monotonically decreasing. We can clearly see that (1) is way more powerful than k-means.

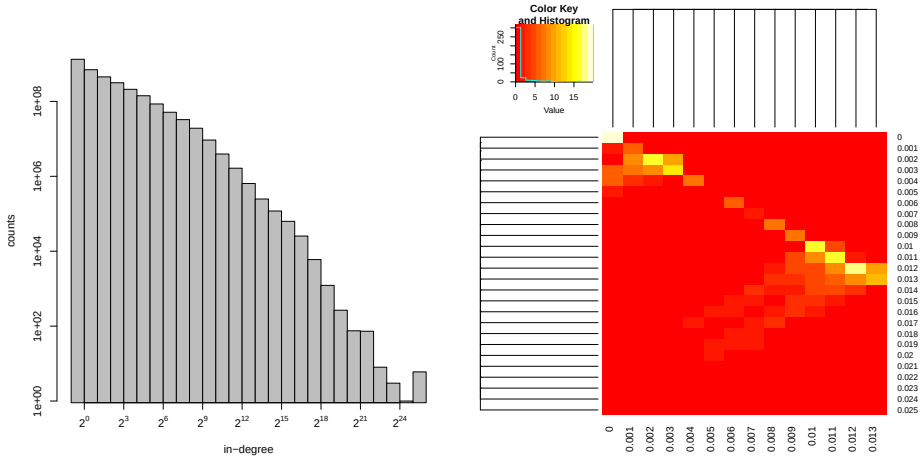

Figure 6: Left: the in-degree distribution of the Page graph. Right: a heatmap of the distribution of the vertices of the page graph in a two-dimension space; the coordinates of the vertices in the two-dimension space is determined by the first left and right singular vectors.

2.2 ndstore

Autoingest is a ndstore service for inserting image and annotation data into the data-store. The user posts information aboout the data inculding location, co-ordinates and datatype. The server uses this information and with a pull model ingests the raw image or annotation data from slices into cuboids and inserts them in the database. This service was earlier operational on local hardware deployed at Johns Hopkins. We modified this service so that it could now be run on Amazon Web Services and S3 object store. Raw data can now be staged on a publicly accessible web server or in a S3 bucket provided by us. The data can be now be inserted into MySQL as well as AWS S3 object store.
Propagate is a ndstore service for building scaling levels over base resolution of data. This is efficient for serving data at lower resolutions for processing and visualization. There also exist auto-zoom in and out capabilties to materialize the data at higher or lower resolutions on the fly if these scaling levels are not built or being built. This service was modified so that scaling levels could be built on the data inserted into AWS S3 object store.
In addition, we also added neariso scaling levels in addition to the exisitng scaling hierarchy. This reduces the data transfer size for 3D viewers and is the preferred interface for tools which use ndstore such as BigDataViewer and Neuroglancer. This will support easier insertion and visualization of data from collaborators.

2.3 ndviz

2.4 knor: K-means NUMA Optimized Routines

We previously described the **knor** library. As a set of tools enabling users to perform the popular k-means algorithm at scale at speeds of 10x-100x time of that of popular frameworks in use today like Spark's MLlib, H^2O and GraphLab(Turi, Dato).
We would like to report that our **knor** paper that was submitted to the ACM Symposium on High-Performance Parallel and Distributed Computing (HPDC 2017) and publicly released it to arxiv was **accepted** for publication to the 26th proceedings of HPDC.

2.5 ndreg

3 Scientific Pipelines: Infrastructure & Dataset Specific Progress

3.1 Science in the Cloud (SIC)

3.2 ndstore

3.3 ndmg

3.4 ndviz

3.5 MRI

3.6 CLARITY

3.7 Ophys