Efficient Krylov Approximation for Manifold Learning

Shinjae Yoo
Computational Science Initiative

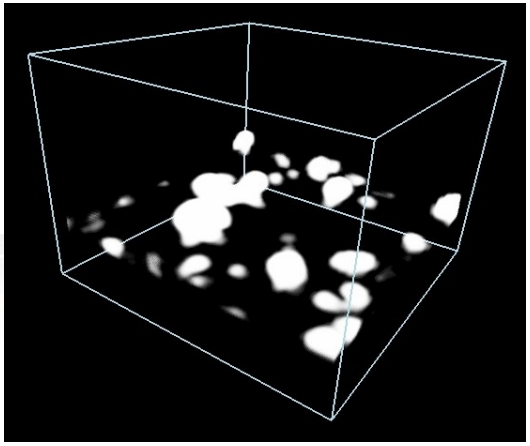
Outline

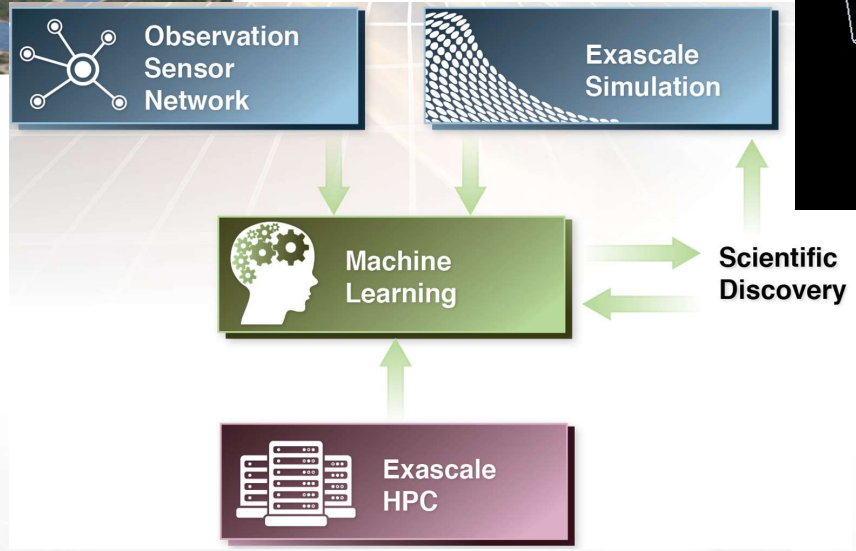
- Projects at BNL
- Big data and unsupervised learning
- Challenges of manifold learning in Big data
- Diverse Power Iteration Embedding
- Streaming version

Extreme Scale Spatio-Temporal Learning

- Fusing theory, simulation, experiments, and ML
  - Interplay of simulation, observation and ML

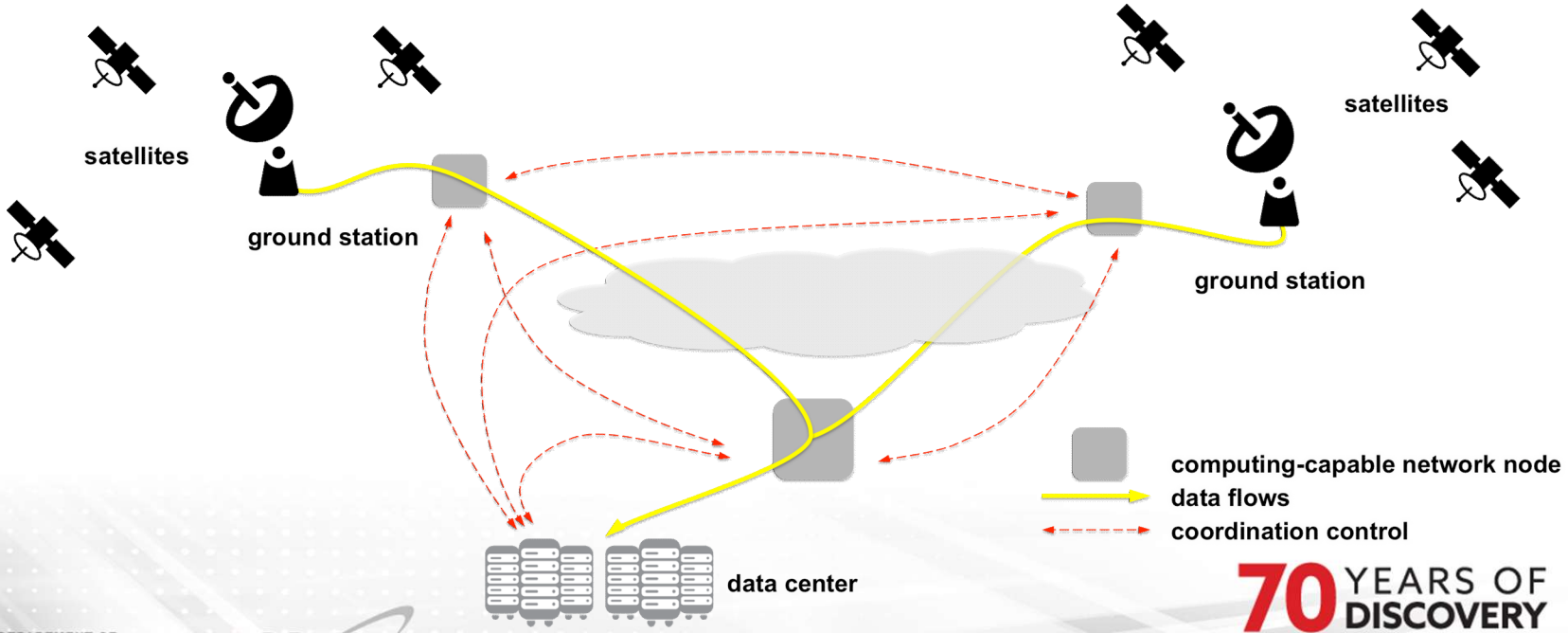
Long Island Solar Farm



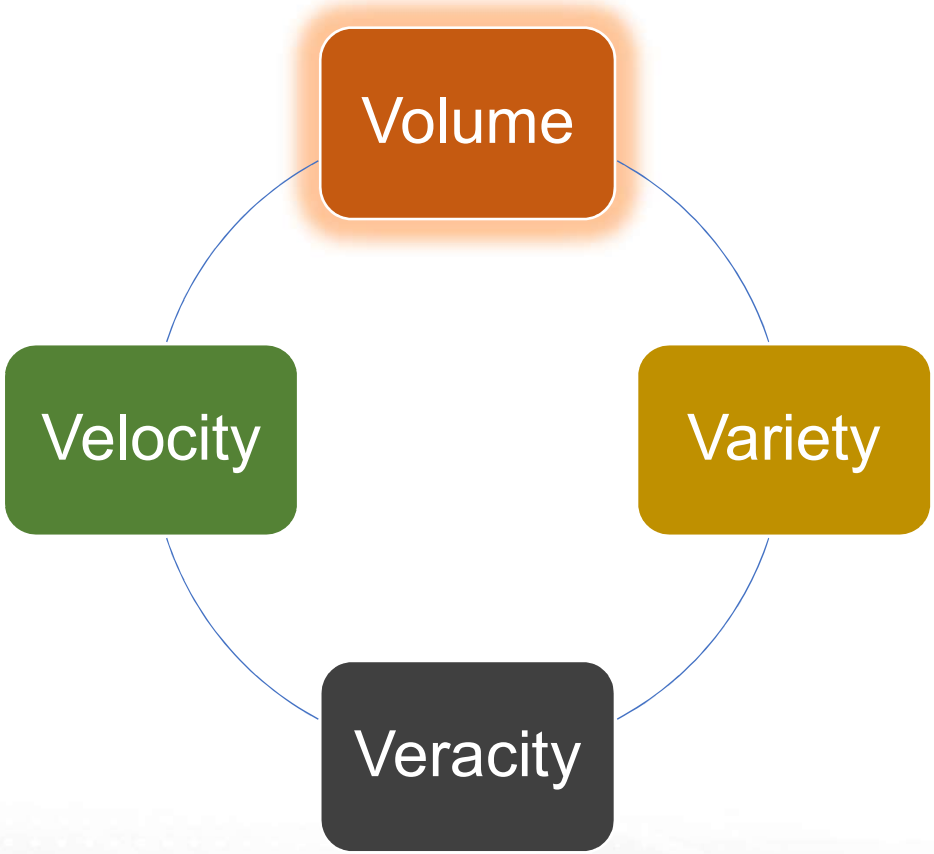


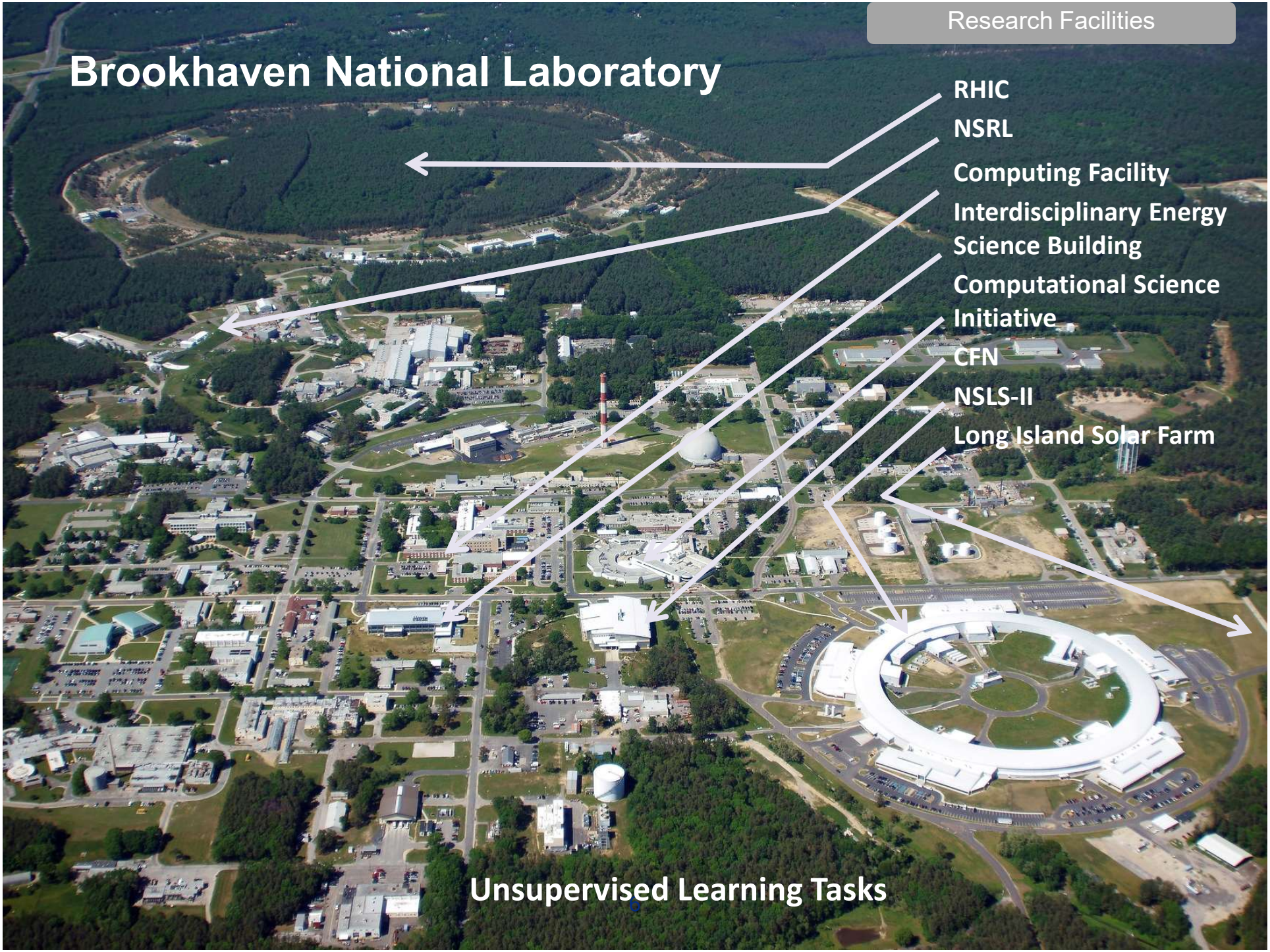
Analysis on the Wire

- **Selectively and transparently perform generic computations on data while in transit in the network fabric.**
  - Process streaming data (e.g., imagery) for early decision-making and reduced downstream bandwidth requirements
  - Extract data analytics, perform generic computations, use distributed computing capabilities
  - Examples: Forecasting, deep learning, pattern recognition (e.g., cyber security, automation)

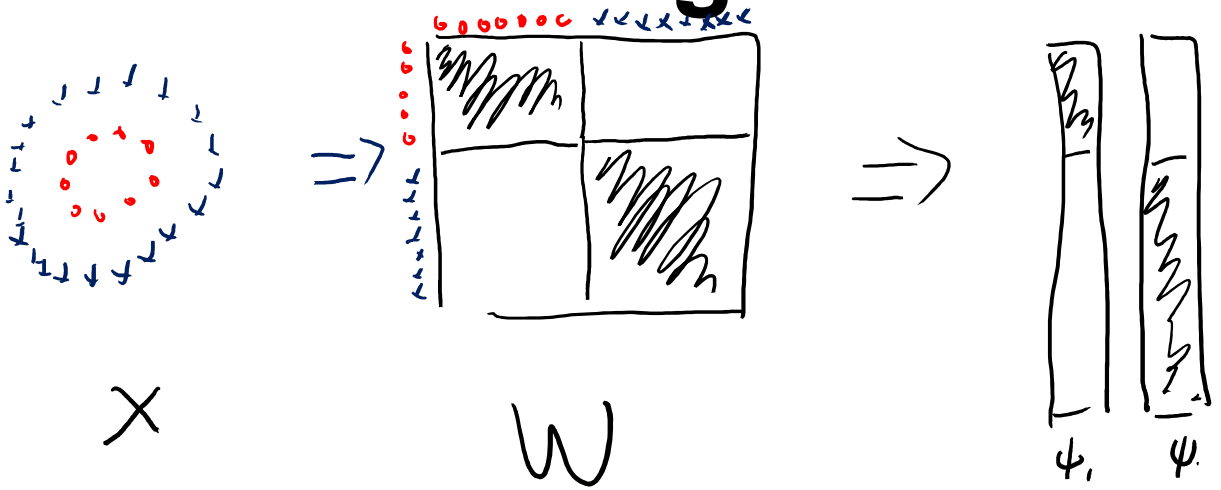


Big Data




Research Facilities
Brookhaven National Laboratory
RHIC
NSRL
Computing Facility
Interdisciplinary Energy Science Building
Computational Science Initiative
CFN
NSLS-II
Long Island Solar Farm
Unsupervised Learning Tasks

Manifold Learning



$O(n^2)$

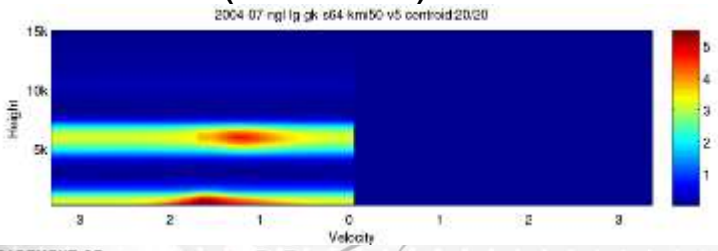
$= EVD(W)$

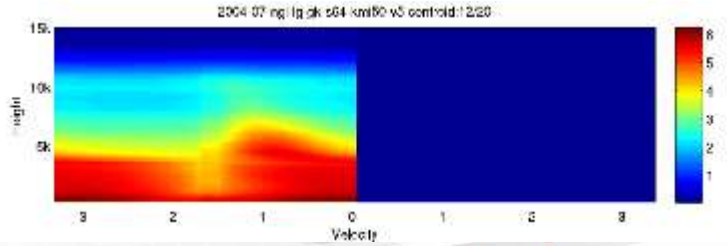
$G(n^3)$

$W\psi = \lambda\psi$

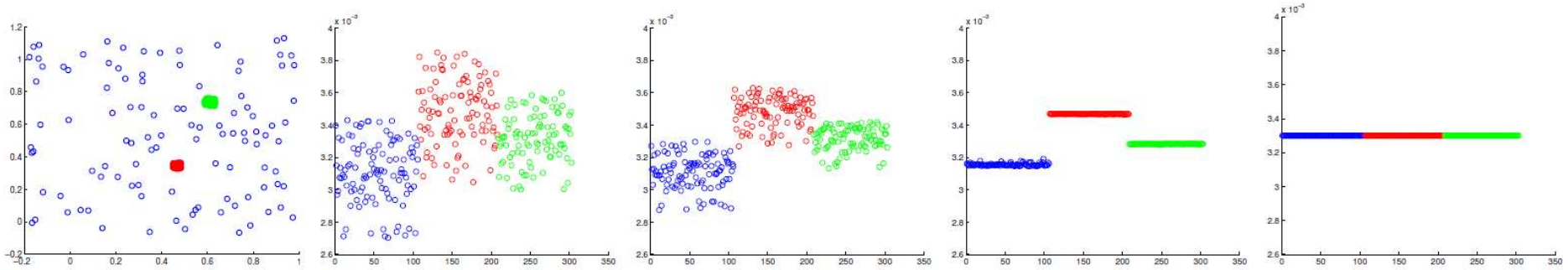
MapReduce: Not Complete Solution in 2010

- **Task**: Find cluster patterns in Doppler Radar Spectra
- **Data**: 1hr≈130MB, 1yr ≈1TB, 2004~2008 ≈ 5TB
- MapReduce (K-Means)
  - Map: Find closest centroids
  - Reduce: Update centroids
- MapReduce (Spectral Clustering)
  - Distributed Affinity Matrix Computation : $O(n^2)$
  - Distributed Lanczos Methods to compute EVD
- Scalability Analysis
  - 12 cores (1 node) Spectral clustering took 1 week for one month data
  - 616 cores (77 nodes) Spectral Clustering took less than 2 hours for three months (~300GB)



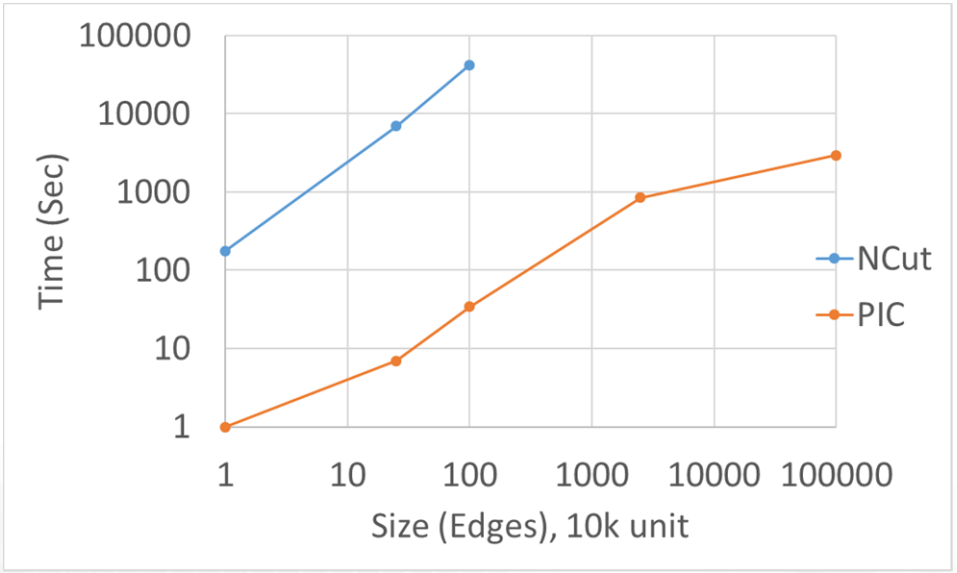


Power-iteration-based Method



(a) 2D dataset (b) $t = 1$ (c) $t = 10$ (d) $t = 20$ (e) $t = 200$

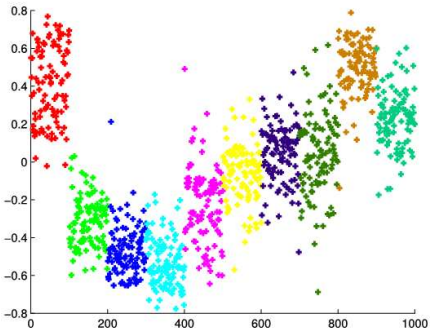
$$v_t = W_t v_0 = a_1 \lambda_1^t \psi_1 + a_2 \lambda_2^t \psi_2 + \ldots + a_n \lambda_n^t \psi_n = a_1 \psi_1 + \lambda_2^t \left(\sum_{i=2}^{n} a_i (\frac{\lambda_i}{\lambda_2})^t \psi_i \right)$$



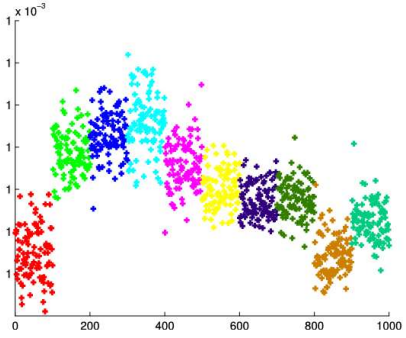
F. Lin, W. Cohen, "Power Iteration Clustering", (ICML 2010)

Power-iteration-based Method

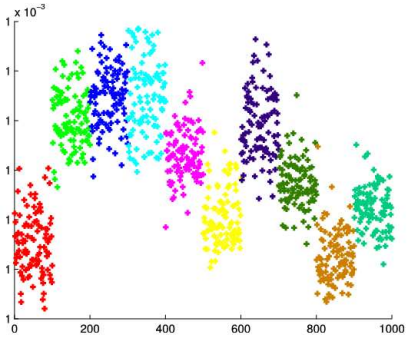
- Limitations



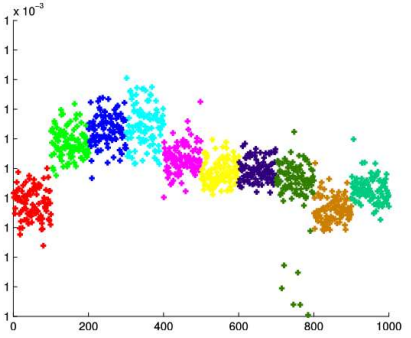
1st Eigenvector



PIE-1



PIE-2

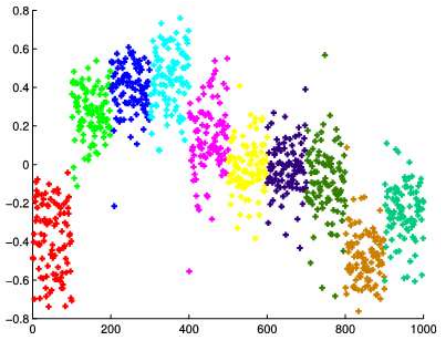


PIE-3

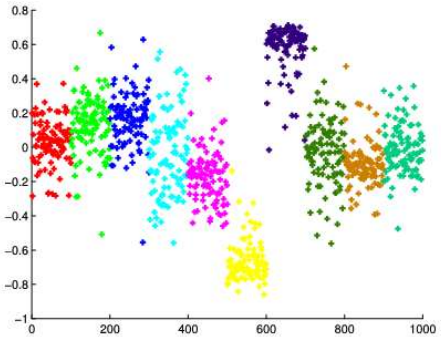
- Large number of cluster application
- Limited use of manifolds
  - Anomaly detection, feature selection, dimensionality reduction

Huang et al. ICDM ‘14 and TKDE ‘16

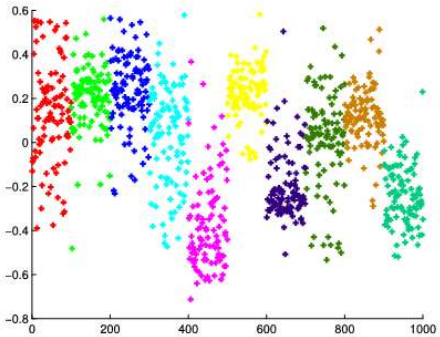
Diverse Power Iteration Embedding (DPIE)



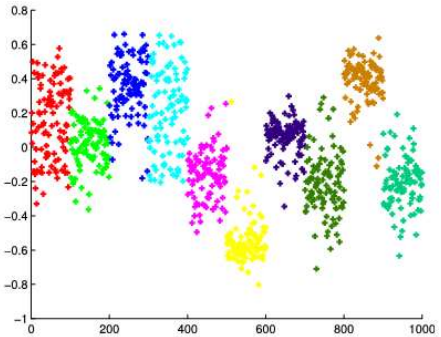
$DPIE_1$



$DPIE_2$



$DPIE_3$



$DPIE_4$

$$\arg\min_{f} = \left\| v_i^t - \psi'_{1:k-1} f \right\| \qquad \psi'_k = \frac{v_i^t - \psi'_{1:k-1} f}{\left\| v_i^t - \psi'_{1:k-1} f \right\|_1}$$

$$O(n^3) \Rightarrow O(nmT + ne\sqrt{\kappa})$$

Huang et al. ICDM '14 and TKDE '16

DPIE: Efficient Space Learning

Space Efficiency:

Cosine Similarity

$$W_{(COS)}(i,j) = \frac{X(i)\cdot X(j)}{\| X(i) \|_2 \cdot \| X(j) \|_2}.$$

$$N_{ii} = 1/\sqrt{X(i)X(i)^T}$$

Affinity matrix W and degree matrix D can be calculated with:

$$W = N \times X \times X^T \times N,$$

$$D = N \times X \times X^T \times N \times \mathbf{1},$$

where **1** is a constant vector of all 1's, and X^T denotes the transpose of X.

$$Wv^t = D^{-1} \times (N \times (X \times (X^T \times (N \times v^t))) - v^t).$$

Gaussian Similarity Approximation

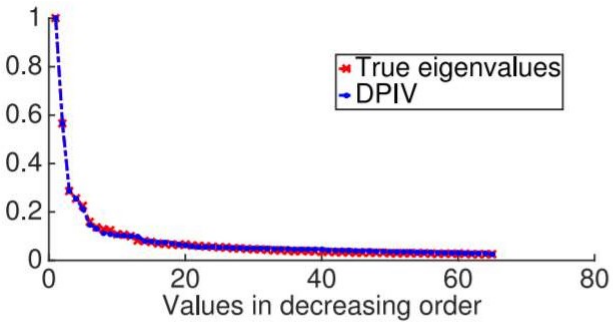
$$W_{(GAU)}(i,j) = exp(\frac{-\| X(i) - X(j) \|^2}{2\sigma^2}),$$

1) Draw d i.i.d. samples $\varpi(1), \ldots, \varpi(d)$ from $p(\varpi \sim \frac{1}{\sigma^2}\mathcal{N}(0,1))$ where $p(*)$ is fast Fourier transform;
2) Draw d i,i,d. samples (offsets) $b(1), \ldots, b(d)$ from uniform distribution on $[0, 2\pi]$;
3) Compute R where $R(i,j) = \sqrt{2/d}[cos(\varpi(j)^T x(i) + b)]$;

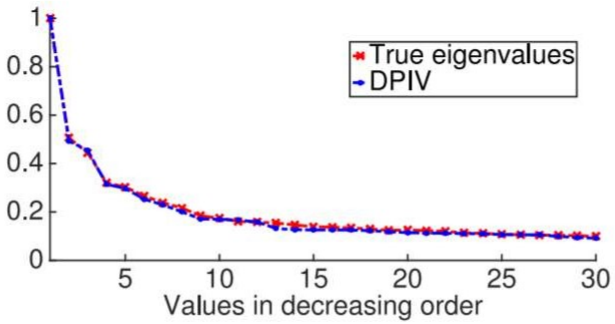
Using the equations listed in cosine similarity, by replacing X with R

Huang et al. ICDM '14 and TKDE '16

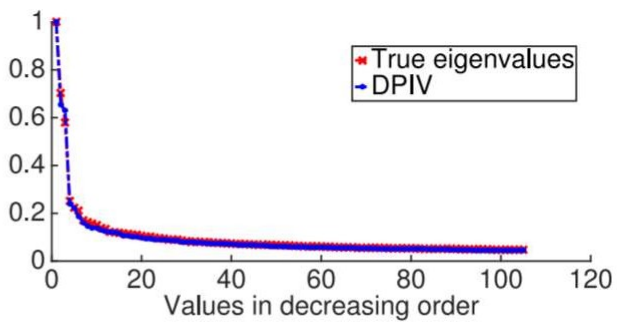
Diverse Power Iteration Value (DPIV)



(b) Reuters21578



(c) RVC1 (30 largest cluster subset)

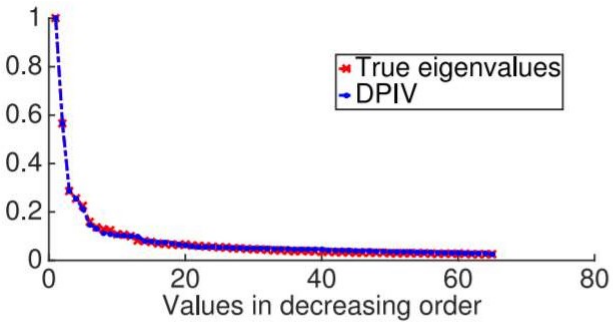


(d) Sector-scale

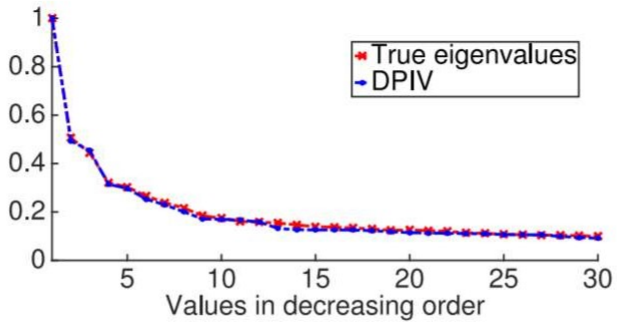
$$W\psi'_i = \lambda'_i \psi'_i \qquad \lambda'_i = (W\psi'_i)\psi'^{-1}_i$$

Huang et al. TKDE '16

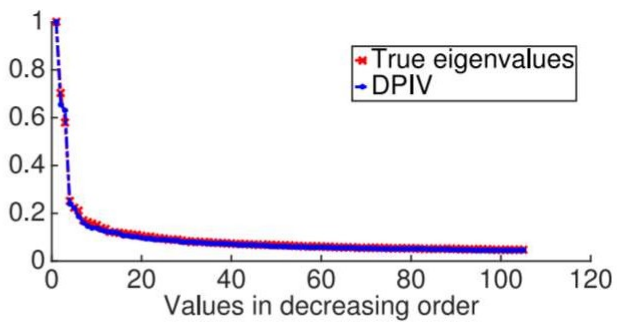
Diverse Power Iteration Value (DPIV)



(b) Reuters21578



(c) RVC1 (30 largest cluster subset)

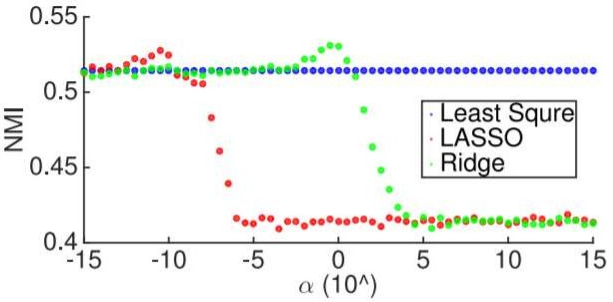


(d) Sector-scale

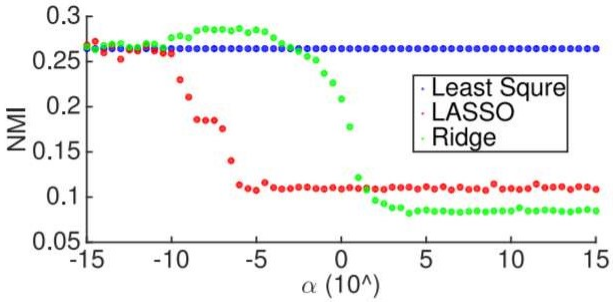
$$W\psi'_i = \lambda'_i \psi'_i \qquad \lambda'_i = (W\psi'_i)\psi'^{-1}_i$$

Huang et al. TKDE '16

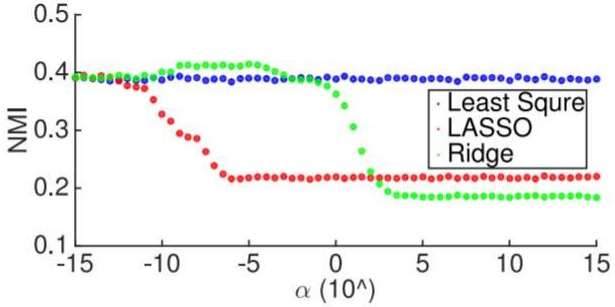
DPIE: Choice of regression types



(b) Reuters21578 ($\#clu. = 65$)



(c) RVC1 ($\#clu. = 103$)



(d) Sector-scale ($\#clu. = 105$)

$$f^* = argmin_f = \|v_i^t - \Psi'_{1:k-1} f\| + \alpha \sum_j |f_j|^p$$

Huang et al. TKDE '16

DPIE: Orthogonalization

Algorithm 4. DPIE-Orthogonalization(Ψ', Λ')

Input: $\Psi' \in R^{n \times k}$ where n is #instances and k is #DPIE, $\Lambda' \in R^{k \times k}$ is the diagonal matrix of DPIV.

output: Orthogonal DPIE $\hat{\Psi} \in R^{n \times k}$ and the corresponding DPIV $\hat{\Lambda} \in R^{k \times k}$.

1 $P \leftarrow \Psi'^T \Psi'$;

2 Perform eigen-decomposition: $P = V \Sigma V^T$;

3 $B \leftarrow \Sigma^{1/2} V^T \Lambda' V \Sigma^{1/2}$;

4 Perform eigen-decomposition: $B = V' \hat{\Lambda} V'^T$;

5 $\hat{\Psi} \leftarrow \Psi' V \Sigma^{-1/2} V'$;

Huang et al. TKDE '16

Experiment

- Evaluation Metrics
  - Clustering and Feature Selection: NMI (Normalized Mutual Information)
  - Anomaly Detection: AUC

Clustering, Feature Selection

	Dataset	# ins.	# fea.	# clu.
1	20Newsgroups	18846	26214	20
2	Reuters21578	8293	18933	65
3	RCV1	193844	47236	103
4	USPS	9298	256	10
5	MNIST	70000	784	10

Anomaly Detection

	Dataset	# ins.	# fea.	# ano.
6	20NG-10-11	4991	26214	100
7	Reuters21578AD	6261	18933	493
8	RCV1AD	7803	29992	200
9	magic04	19020	10	6688
10	satellite	6435	36	2036

Experiment: Clustering

NMI	NJW	PIE	PIE-k	MatSket	DeflationPIC	DPIE
20Newsgroups	0.5326	0.2519	0.3266	0.4877	0.4847	**0.5061** (1)
Reuters21578	0.5048	0.2557	0.2718	**0.5322**	0.5014	0.5143 (2)
RCV1	[23]*0.2875*	0.1022	0.1237	0.1521	0.1941	**0.2644** (1)
USPS	0.6207	0.2026	0.2401	0.4667	**0.5871**	0.5786 (2)
MNIST	0.4433	0.0022	0.0028	0.3523	0.3788	**0.4032** (1)
Average	0.4778	0.1629	0.1930	0.3982	0.4292	**0.4533** (1)
Time(s)	NJW	PIE	PIE-k	MatSket	DeflationPIC	DPIE
20Newsgroups	5653.0193	0.1461	5.0816	4131.7741	35.4688	5.0834
Reuters21578	1958.5777	0.0671	2.3548	830.7118	13.7681	1.6388
RCV1	——	5.1961	110.5477	108998.2234	923.6324	127.6903
USPS	1665.3840	0.0675	1.9807	395.9329	7.2451	0.6584
MNIST	201581.2017	4.0707	38.8645	46072.8311	196.3723	43.6582
Average	——	1.9095	31.7659	32085.8947	235.2973	35.7458

Experiment: Anomaly Detection

AUC	HKS-SE	HKS-PIE	HKS-PIEK	HKS-MatSket	HKS-DFL	IForest	HKS-DPIE
20NG-10-11	0.9042	0.3294	0.4858	0.6331	0.2318	0.6176	**0.8844** (1)
Reuters21578AD	0.7845	0.3034	0.5131	0.4824	0.7863	0.6048	**0.9271** (1)
RCV1AD	0.5428	0.4403	0.5049	0.4619	**0.5925**	0.4879	0.5547 (2)
magic04	0.7286	0.5757	0.5757	0.5799	0.4205	**0.7506**	0.7179 (3)
satellite	0.7078	0.3378	0.3378	0.5062	0.5416	0.7173	**0.7193** (1)
Average	0.7336	0.3973	0.4835	0.5327	0.5145	0.6356	**0.7607** (1)

Time(s)	HKS-SE	HKS-PIE	HKS-PIEK	HKS-MatSket	HKS-DFL	IForest	HKS-DPIE
20NG-10-11	876.9247	0.0297	0.8683	181.7283	5.7138	7.6199	0.8193
Reuters21578AD	4141.9718	0.0528	1.1995	170.0181	7.3392	8.2016	1.0608
RCV1AD	4199.1405	0.0476	1.3253	475.9983	10.6519	5.5944	1.1128
magic04	14732.0387	0.1252	0.3402	3241.6766	20.3112	53.8751	2.2759
satellite	779.7334	0.0145	0.1121	152.7320	8.9713	49.3959	0.5889
Average	4945.9618	0.0540	0.7691	844.4307	10.5975	24.9374	1.1715

Experiment: Feature Selection

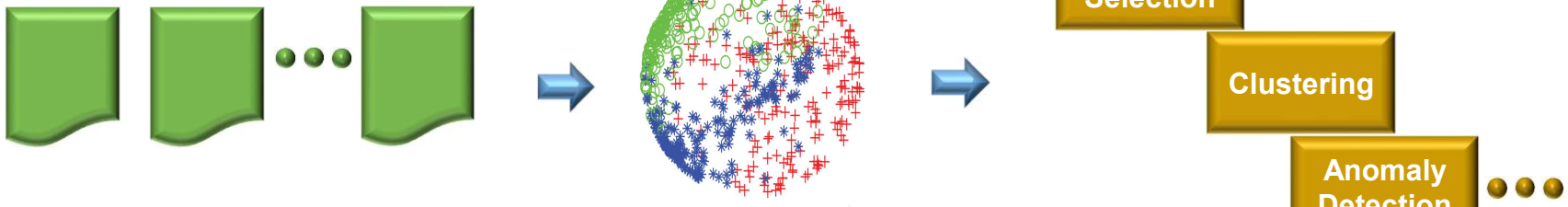
20Newsgroups	MCFS-SE	MCFS-PIE	MCFS-PIEK	MCFS-MatSket	MCFS-DFL	MCFS-DPIE
50	0.2971	0.1691	0.1590	0.2691	0.2552	**0.3446** (1)
200	0.3361	0.3089	0.3181	0.3603	0.3274	**0.3834** (1)
800	0.4118	0.3899	0.4115	0.4061	0.4256	**0.4372** (1)
1200	0.4256	0.4696	0.4498	0.4692	0.4335	**0.4819** (1)
1800	0.4865	0.4671	0.4587	0.4340	0.4748	**0.4993** (1)
Reuters21578	MCFS-SE	MCFS-PIE	MCFS-PIEK	MCFS-MatSket	MCFS-DFL	MCFS-DPIE
50	0.3957	0.3959	0.3889	**0.4399**	0.3973	0.4366 (2)
200	0.4607	0.4539	0.4598	0.4745	0.4677	**0.4814** (1)
800	0.5125	0.5021	**0.5183**	0.5113	0.4993	0.5176 (2)
1200	0.5125	0.4783	0.4882	0.4971	0.5122	**0.5297** (1)
1800	0.5081	0.5104	0.5078	0.4980	0.5200	**0.5308** (1)
Average	0.4347	0.4145	0.4160	0.4360	0.4313	**0.4646** (1)

Summary

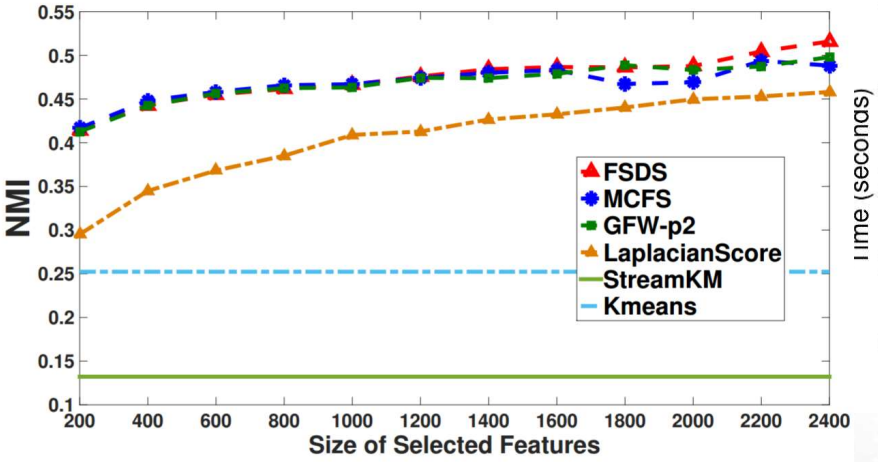
- Clustering: 4000 times faster and reach 95% of the best clustering performance.
- Anomaly Detection: 5000 times faster and reach 103% of the best performance.
- Feature Selection: 4000 times faster, and has similar performance of the best algorithms.
- Provides DPIV and Orthogonalization for various applications

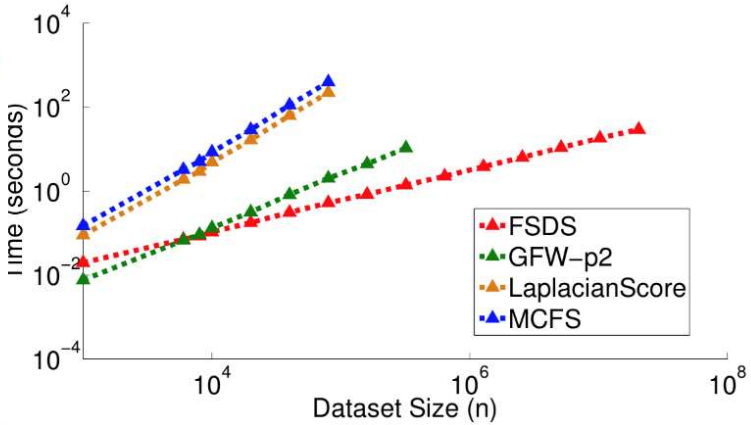
Streaming Approximations

High Dimensional Stream



Feature Selection





Questions?