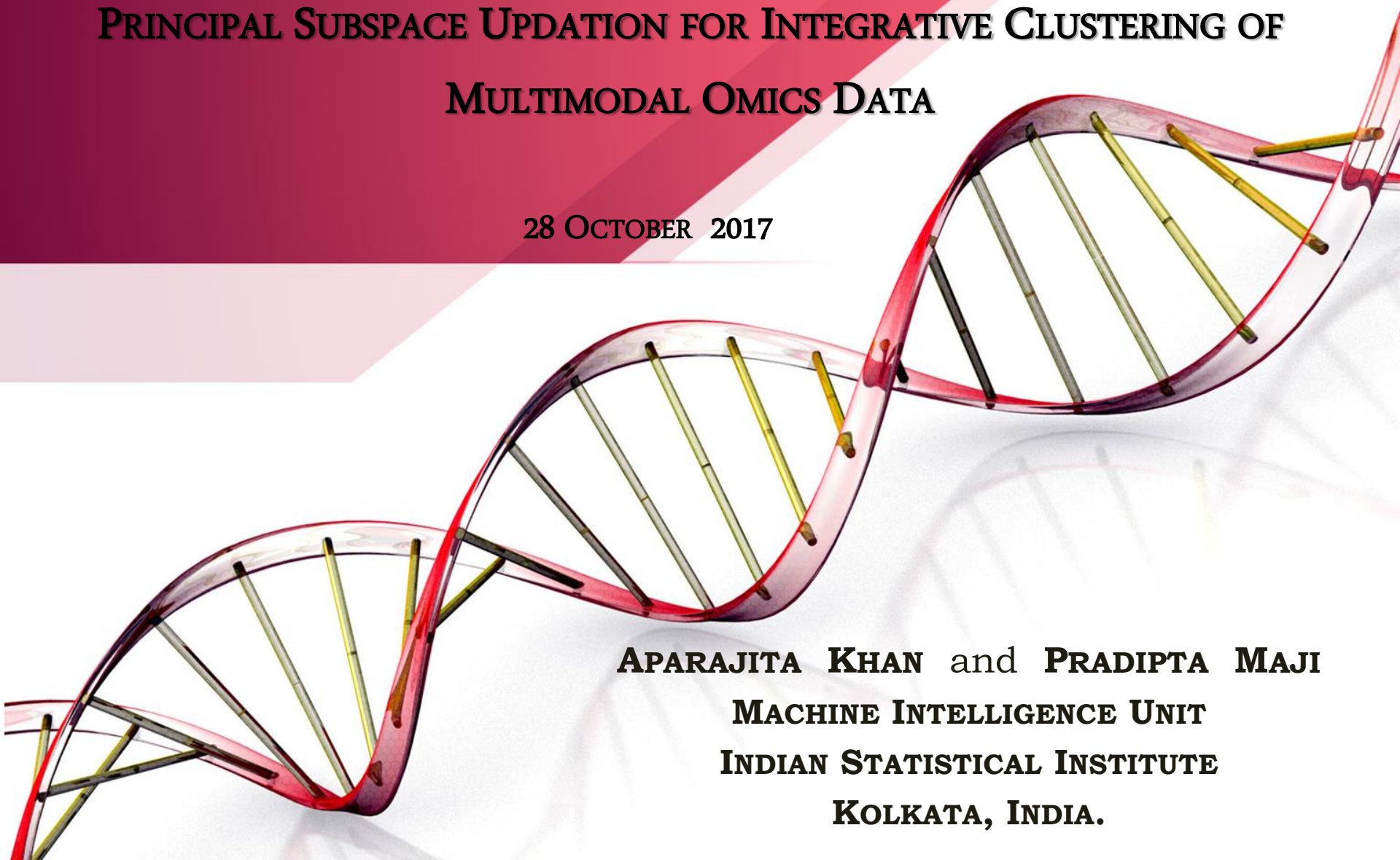
3rd International Conference On Computational Intelligence And Networks

Principal Subspace Updation for Integrative Clustering of Multimodal Omics Data

28 October 2017

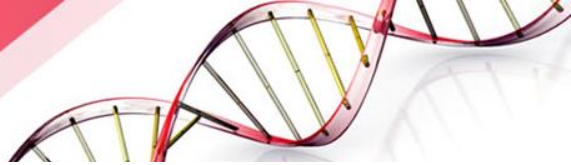


Aparajita Khan and **Pradipta Maji**
Machine Intelligence Unit
Indian Statistical Institute
Kolkata, India.

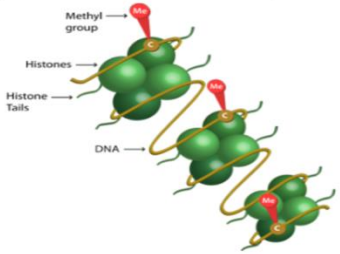
Outline

- Multimodal Data Integration
- Existing Approaches
- Principal Subspace
- Relevance and Concordance
- Proposed Algorithm
- Results and Survival Analysis
- Conclusion

Multimodal Data Integration



Epigenomic

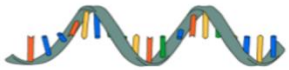


DNA Methylation

Genomic

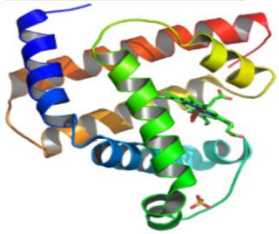
Copy Number Variation

Transriptomic



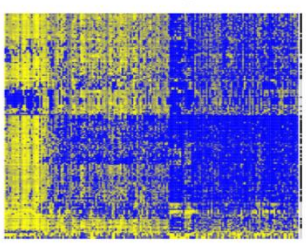
Gene Expression

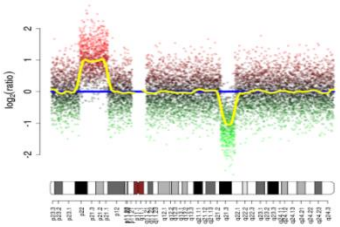
Proteomic

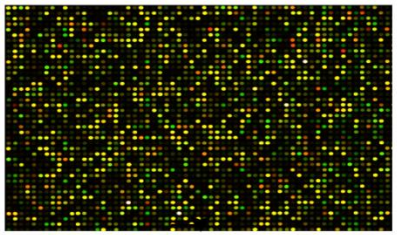


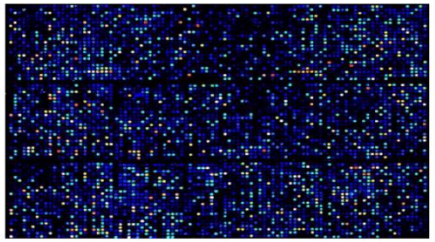
Protein Expression

N Samples

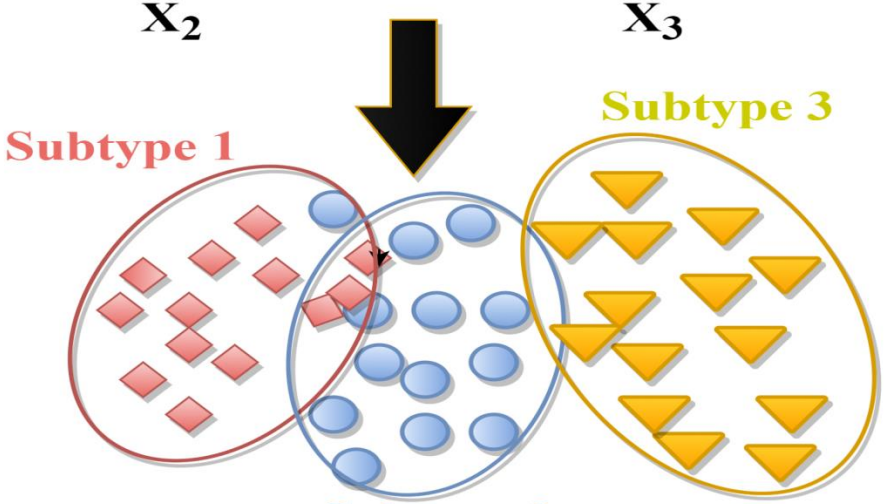






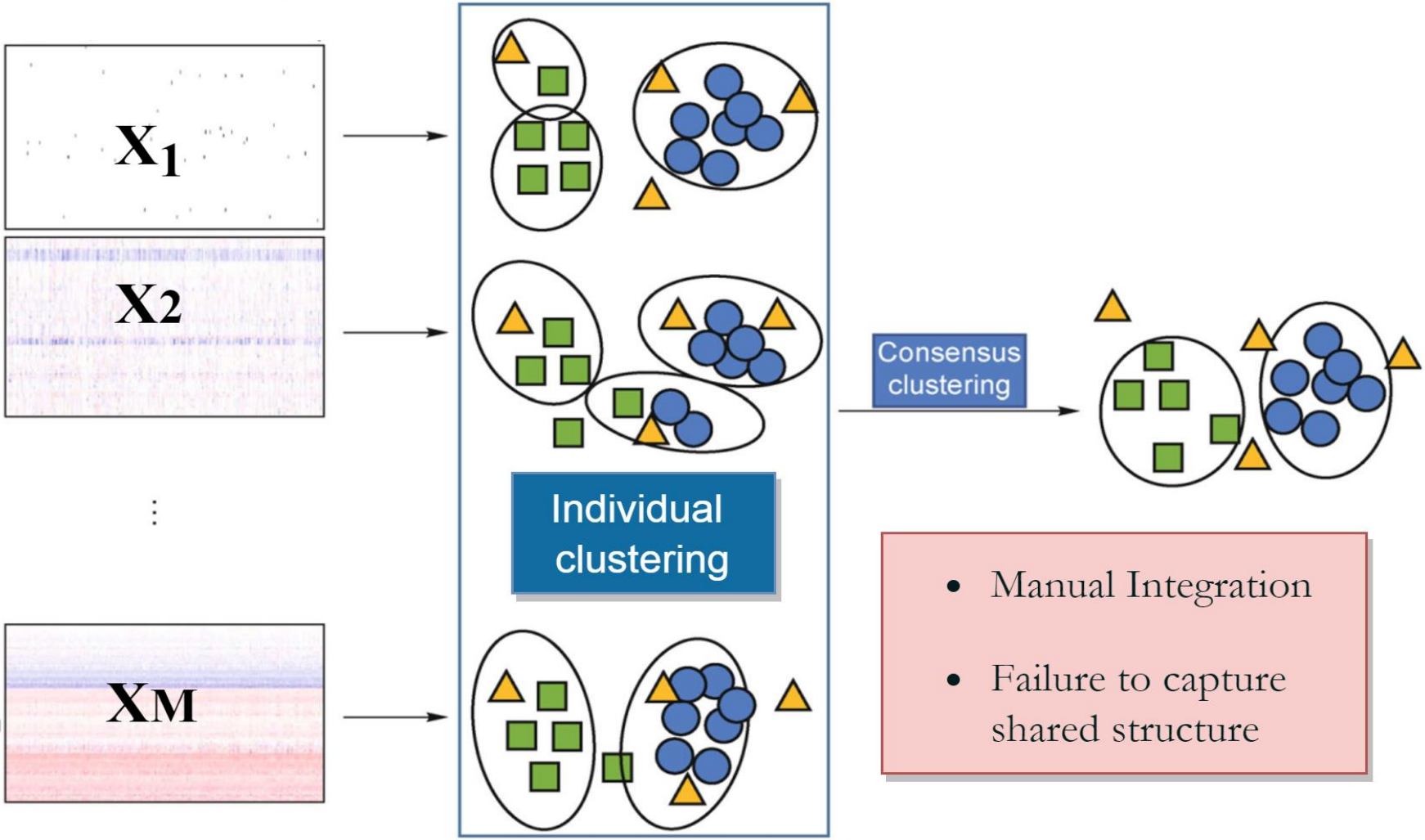


X_1 X_2 X_3 . . . X_M



Existing Approaches

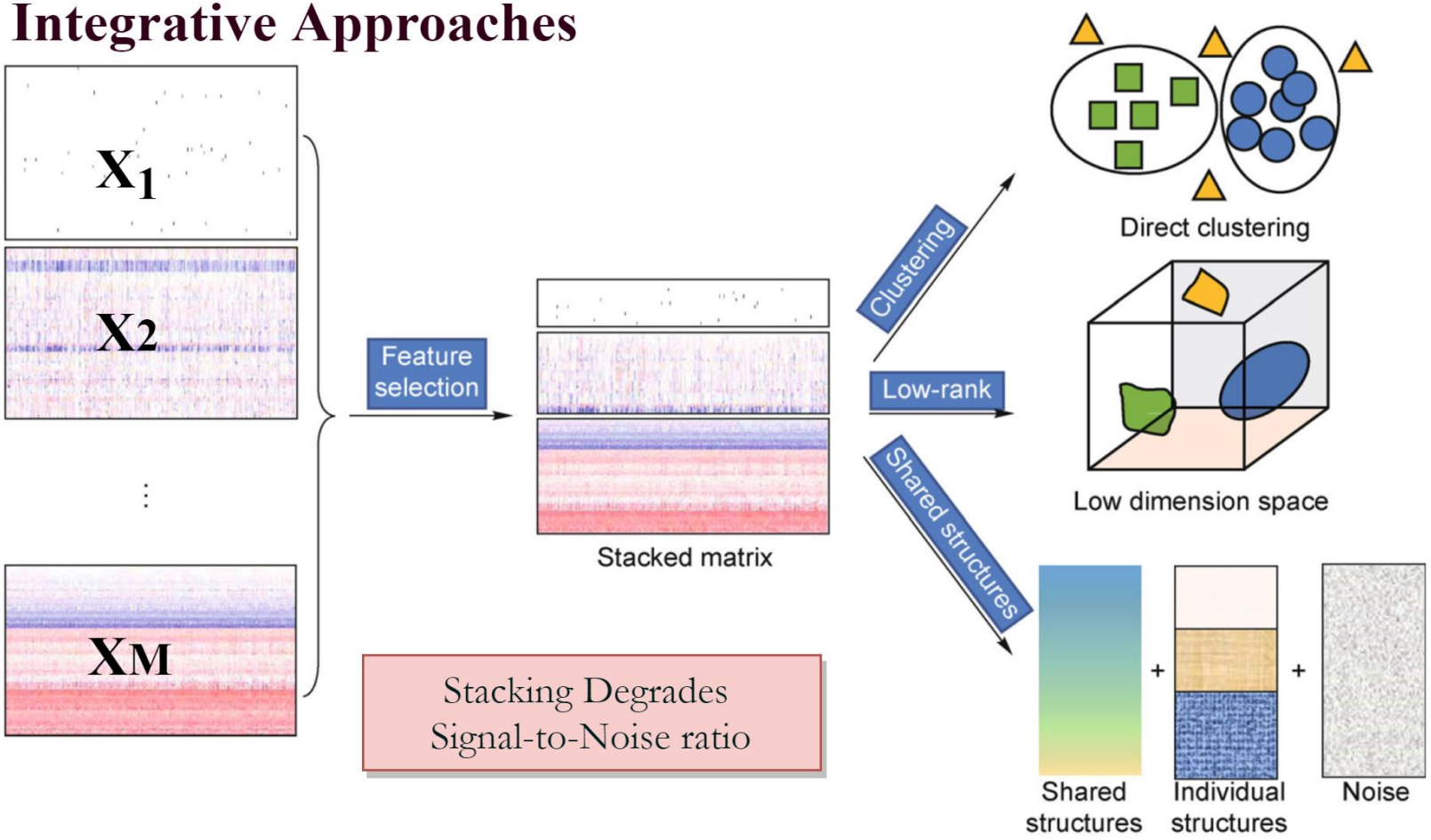
2 Stage Approaches:



- Manual Integration
- Failure to capture shared structure

Existing Approaches

Direct Integrative Approaches

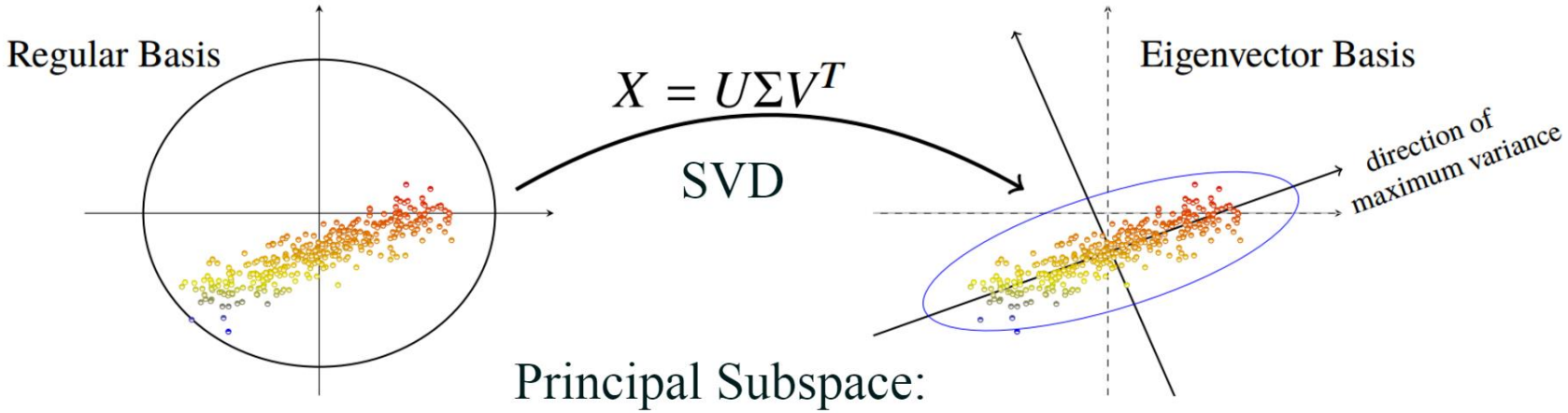


- **Relevance of individual modalities**
- **Concordance between modalities**

IDEA

Adaptively construct the joint subspace from individual subspaces

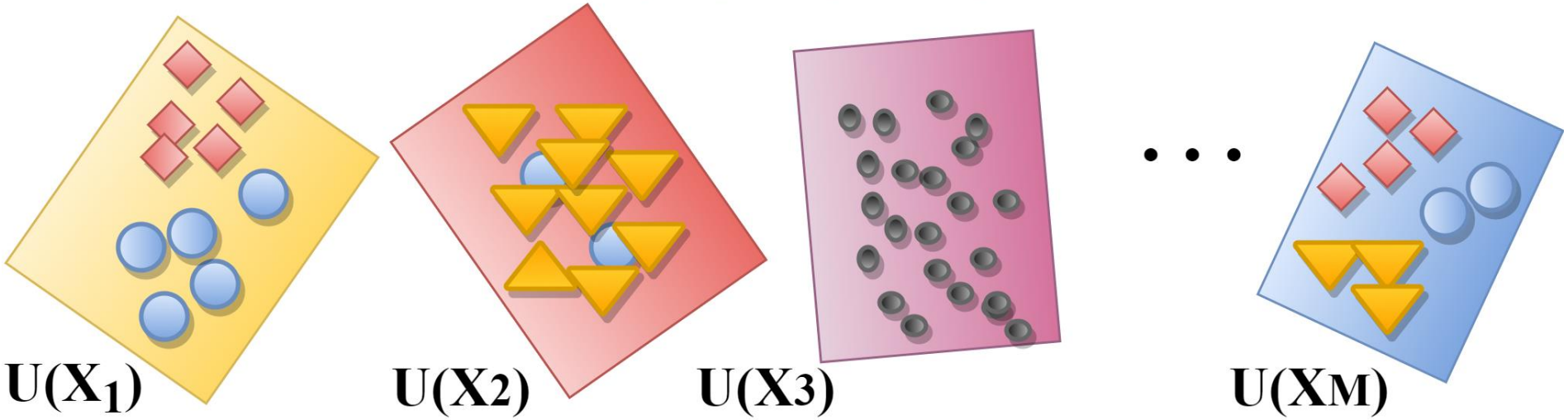
Principal Subspace



Principal Subspace:

$$\Psi(X_m) = \langle U(X_m), \Sigma(X_m) \rangle$$

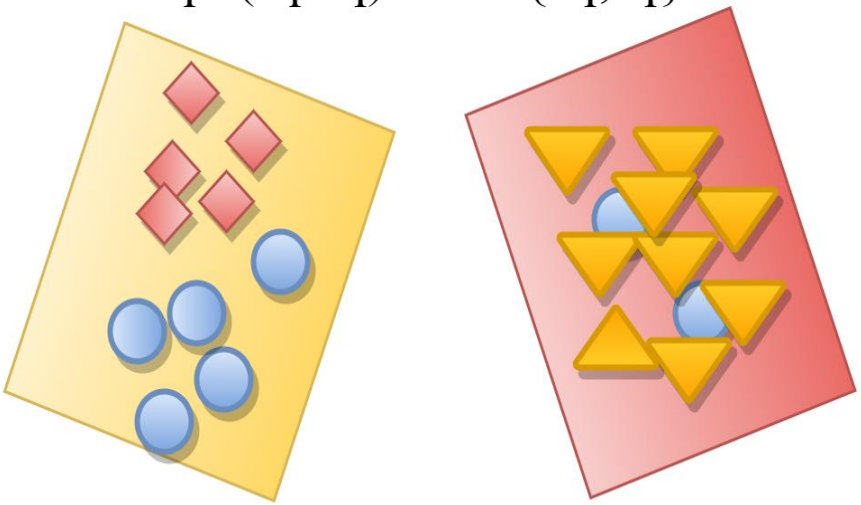
- K-cluster solution = top K principal components [1]



Relevance and Concordance

$$\mathbf{R}_I(X_m)=Silhouette(U(X_m))$$

$$s_i=(b_i-a_i)/\max\{b_i,a_i\}$$

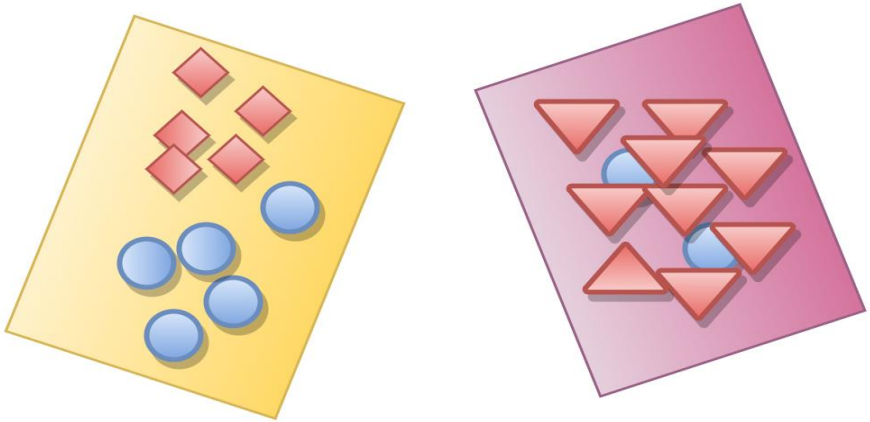


High Relevance

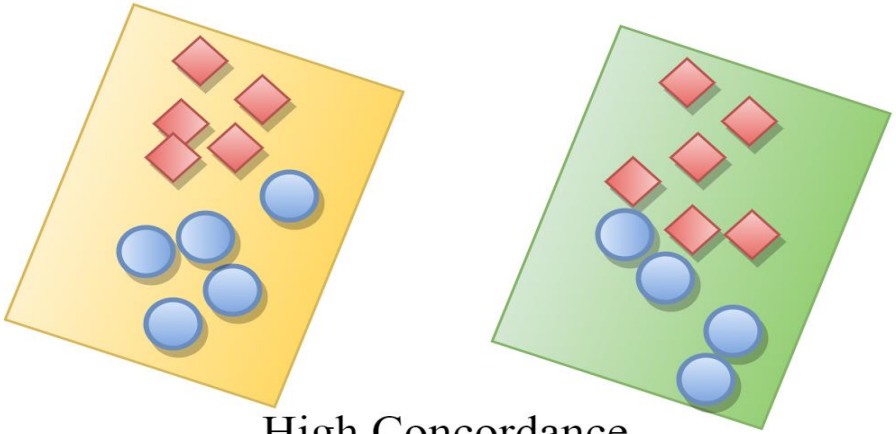
Low Relevance

$$C_d(X_i,X_j)=NMI(C_i,C_j)$$

$$NMI(C_i,C_j)=2I(C_i,C_j)/[H(C_i)+H(C_j)]$$

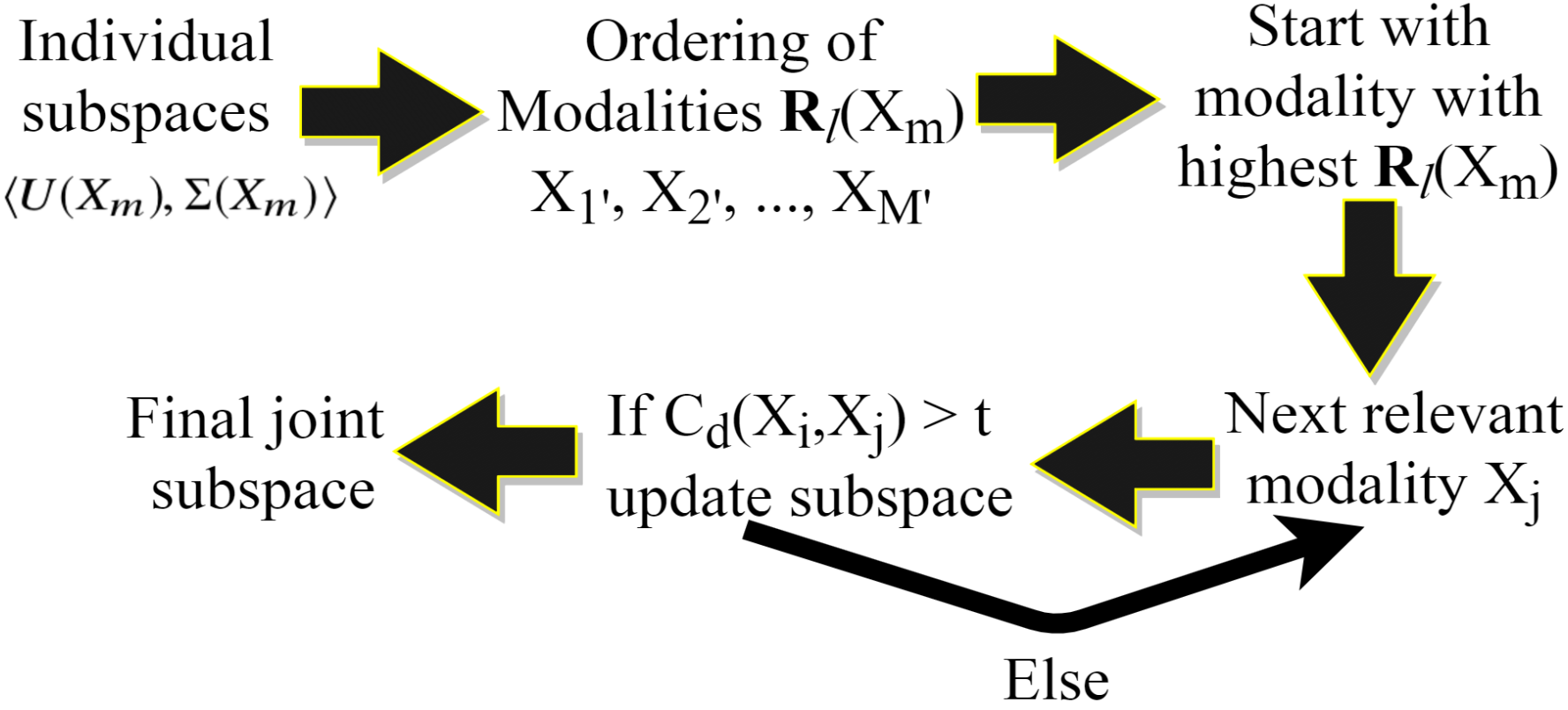


Low Concordance



High Concordance

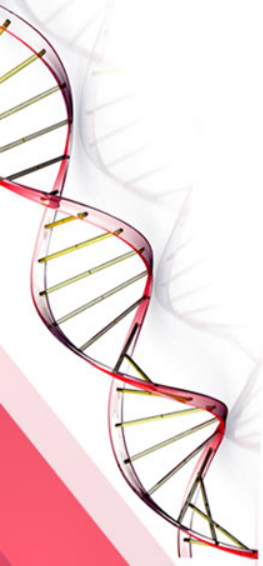
Proposed Algorithm



Joint Principal Components: $\mathbf{Y} = U(\widetilde{\mathbf{X}}_M)\Sigma(\widetilde{\mathbf{X}}_M)$

Identify subtypes: Cluster the rows of $\mathbf{Y}$ using standard k-means

Computational complexity is $O(Mn^2 d_{max})$.



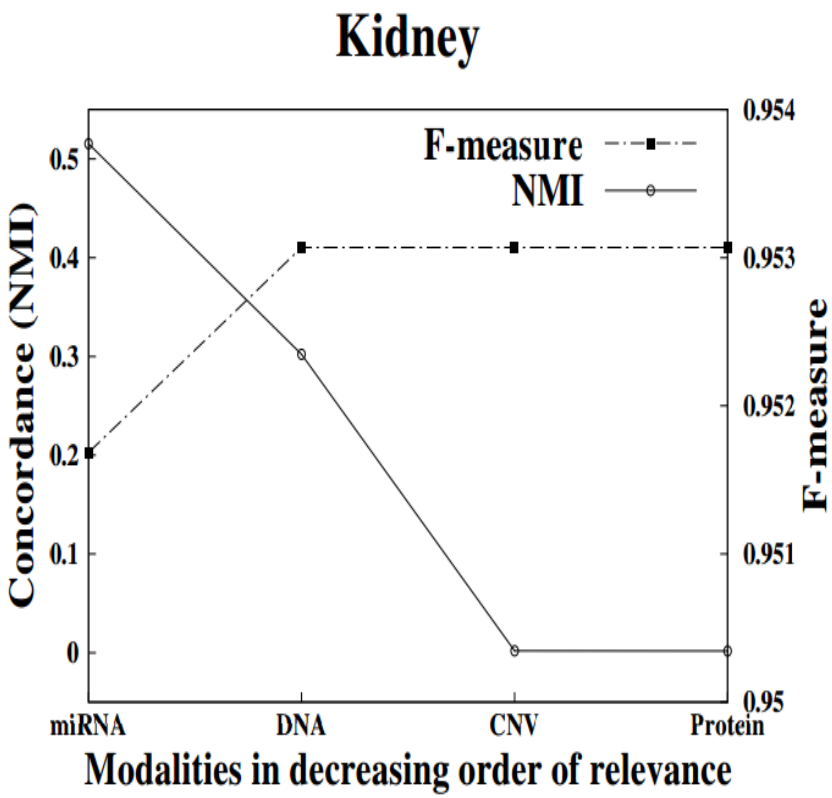
Datasets and Measures

- 2 real-life cancer data sets from TCGA
- Modalities:
  - Gene Expression, DNA methylation, miRNA, Protein, CNV
- Giloblastoma Multiforme (GBM):
  - 168 samples, 4534 features, 4 clusters
- Kidney cancer (KIDNEY):
  - 737 samples, 5979 features, 3 clusters
- Compared with:
  - BCC [2], COCA [3]
  - iCluster [4], LRAcluster [5], PCA-NI
- **External Indices**: F-measure, NMI
- **Survival Analysis**: p-value in log-rank test [7]

Results

Relevance of Each Modality and Selected Modalities

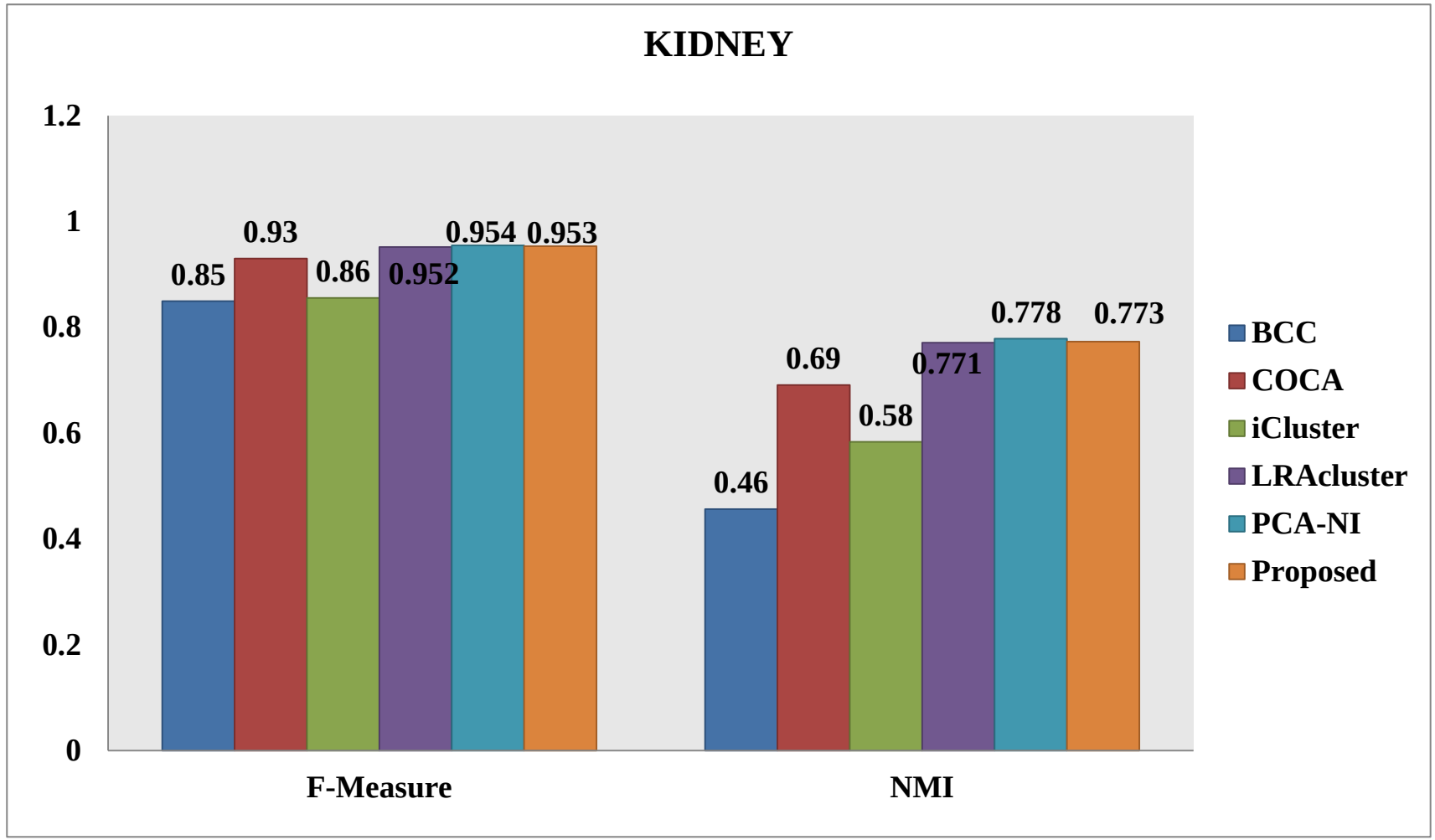
Different Modality	Kidney		GBM	
	Relevance $\mathcal{R}_l$	Selected	Relevance $\mathcal{R}_l$	Selected
CNV	0.2552738		0.2371261	
DNA	0.3840580	Gene	-	Gene
Gene	0.5172488	miRNA	0.2958170	CNV
miRNA	0.4014777	DNA	0.2196443	
Protein	0.2430872		-	



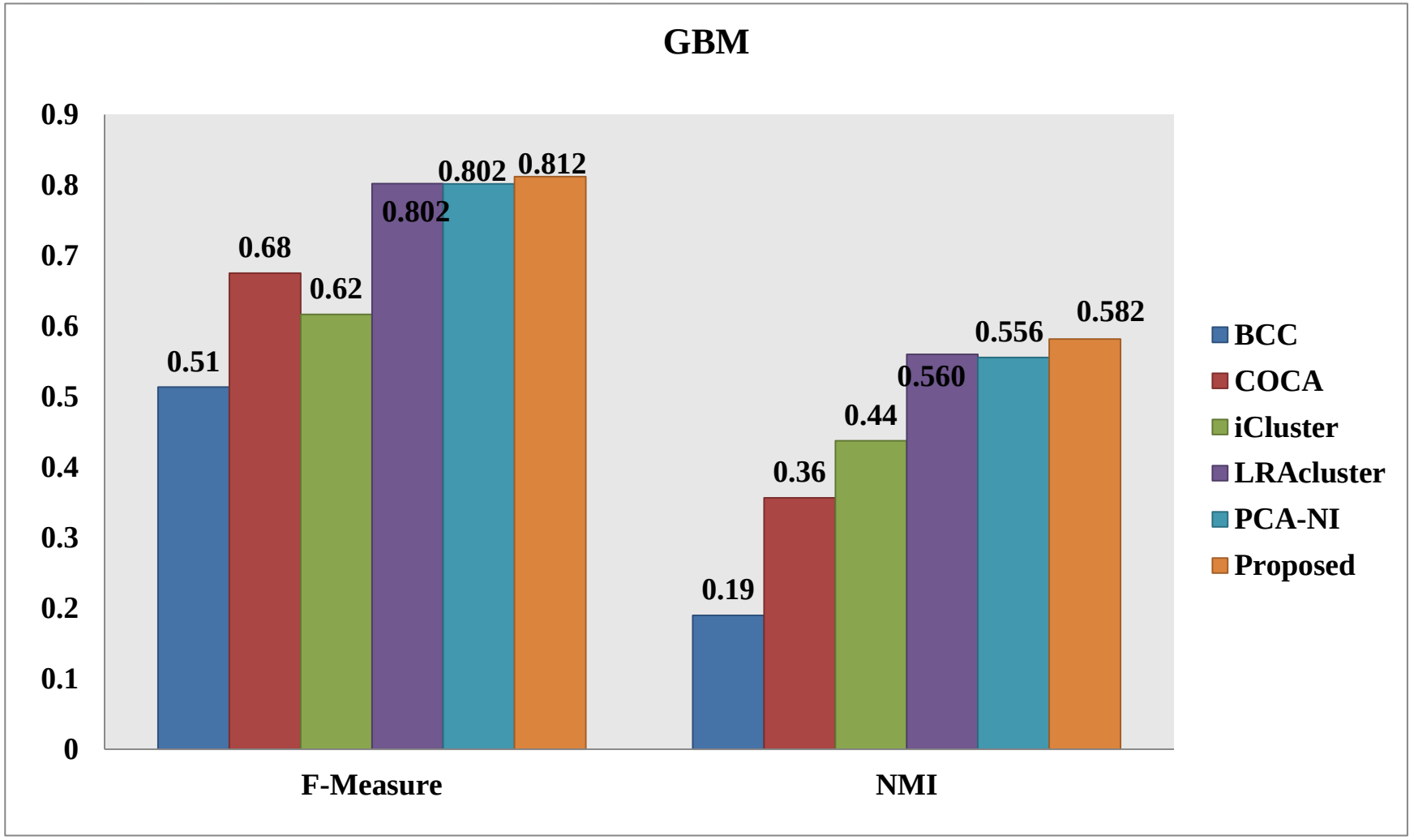
KIDNEY ordering:
Gene>miRNA>DNA>CNV>Protein

GBM ordering:
Gene>CNV>miRNA

RESULTS II


KIDNEY
1.2
1
0.8
0.6
0.4
0.2
0
0.85
0.93
0.86
0.952
0.954
0.953
0.46
0.69
0.58
0.771
0.778
0.773
F-Measure
NMI
BCC
COCA
iCluster
LRAcluster
PCA-NI
Proposed

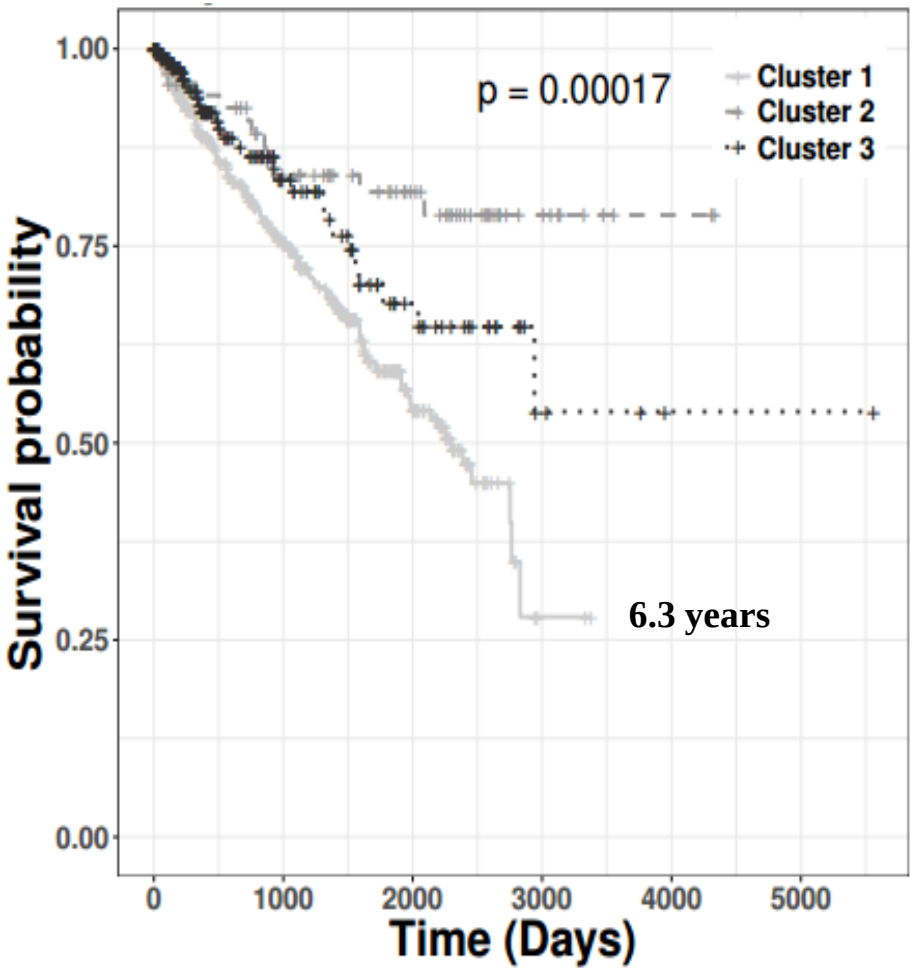
Results III


GBM
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0.51
0.68
0.62
0.802
0.802
0.812
0.19
0.36
0.44
0.560
0.556
0.582
F-Measure
NMI
BCC
COCA
iCluster
LRAcluster
PCA-NI
Proposed

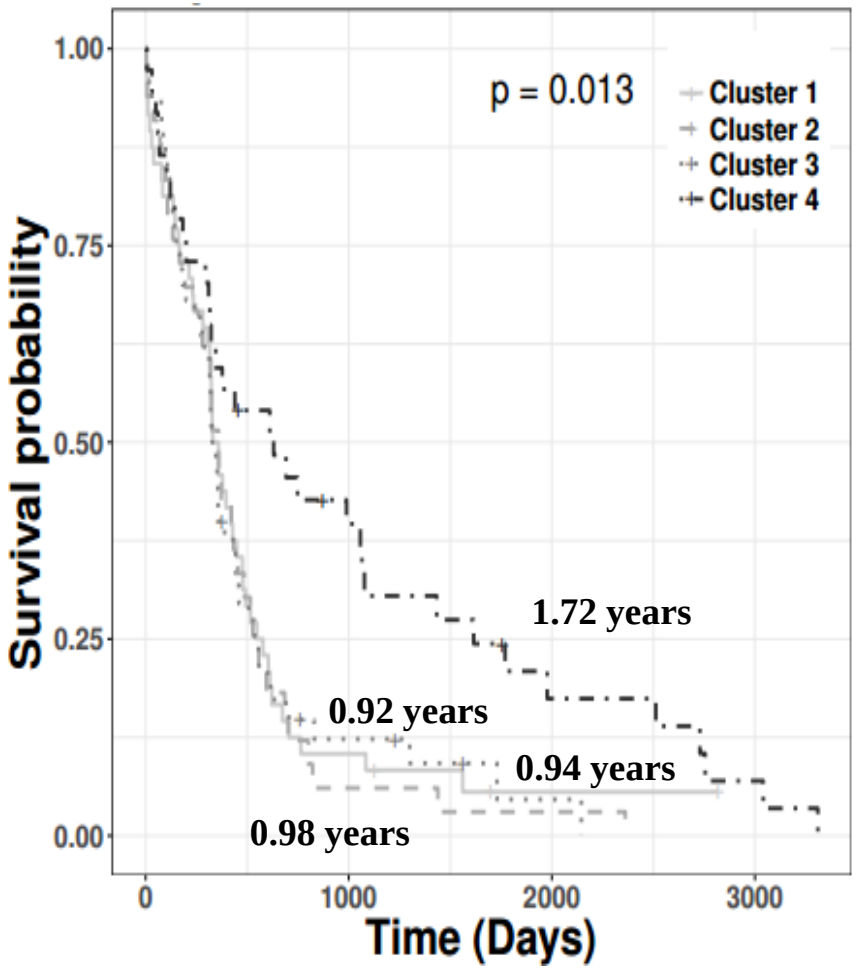
Survival Analysis

Kaplan-Meier survival plots with median survival time

KIDNEY



GBM



Conclusion

- Constructs low-rank joint subspace from individual subspaces
- Evaluates relevance and mutual information before naïve integration
- Filters out noisy and inconsistent modalities
- Computationally efficient

Future Work

- ❑ Dimensionality increases linearly with number of modalities
- ❑ Update subspace instead appending
- ❑ Effective rank estimation techniques

References

1. H. Zha, X. He, C. Ding, H. Simon, and M. Gu, "Spectral relaxation for k-means clustering," in Neural Information Processing Systems, vol. 14, (Vancouver, Canada), pp. 1057 - 1064, 2001.
2. E. F. Lock, et al., "Bayesian consensus clustering," Bioinformatics, no. 29(20), pp. 2610–2616, 2013.
3. K. A. Hoadley, C. Yau, et al., "Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin," Cell, vol. 158, pp. 929–944, 2014.
4. R. Shen, A. B. Olshen, and M. Ladanyi, "Integrative clustering of multiple genomic data types using joint latent variable model with application to breast and lung cancer subtype analysis," Bioinformatics, no. 25(22), pp. 2906–2912, 2009.
5. D. Wu, D. Wang, et al., , "Fast dimension reduction and integrative clustering of multi-omics data using low-rank approximation: Application to cancer molecular classification," BMC Genomics, 2015.
6. D. W. Hosmer, S. Lemeshow, and S. May, Applied Survival Analysis: Regression Modeling of Time to Event Data. New York, NY, USA: Wiley-Interscience, 2nd ed., 2008.
7. P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," Journal of Computational and Applied Mathematics, vol. 20, pp. 53 – 65, 1987.

Thank You!

Questions?