Multi-View Learning

Angela Serra and Roberto Tagliaferri

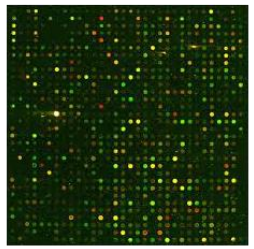
UNIVERSITÀ DEGLI STUDI DI SALERNO

Introduction

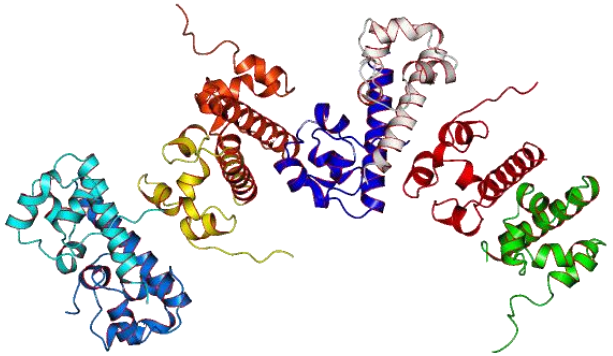
- Multi-view learning is concerned with the problem of machine learning from data represented by multiple distinct feature sets.
- The recent emergence of this learning mechanism is largely motivated by the property of data from real applications where examples are described by different feature sets or different views.
  - Bioinformatics: microarray gene expression, RNASeq, PPI, gene ontology, etc.;
  - Neuroinformatics: Functional magnetic resonance imaging (fMRI), diffusion tensor imaging (DTI)

Introduction

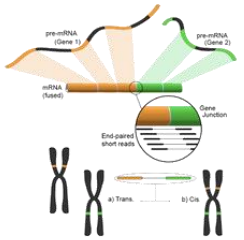
- How to put things together?



mRNA expression data



Protein-Protein Interaction

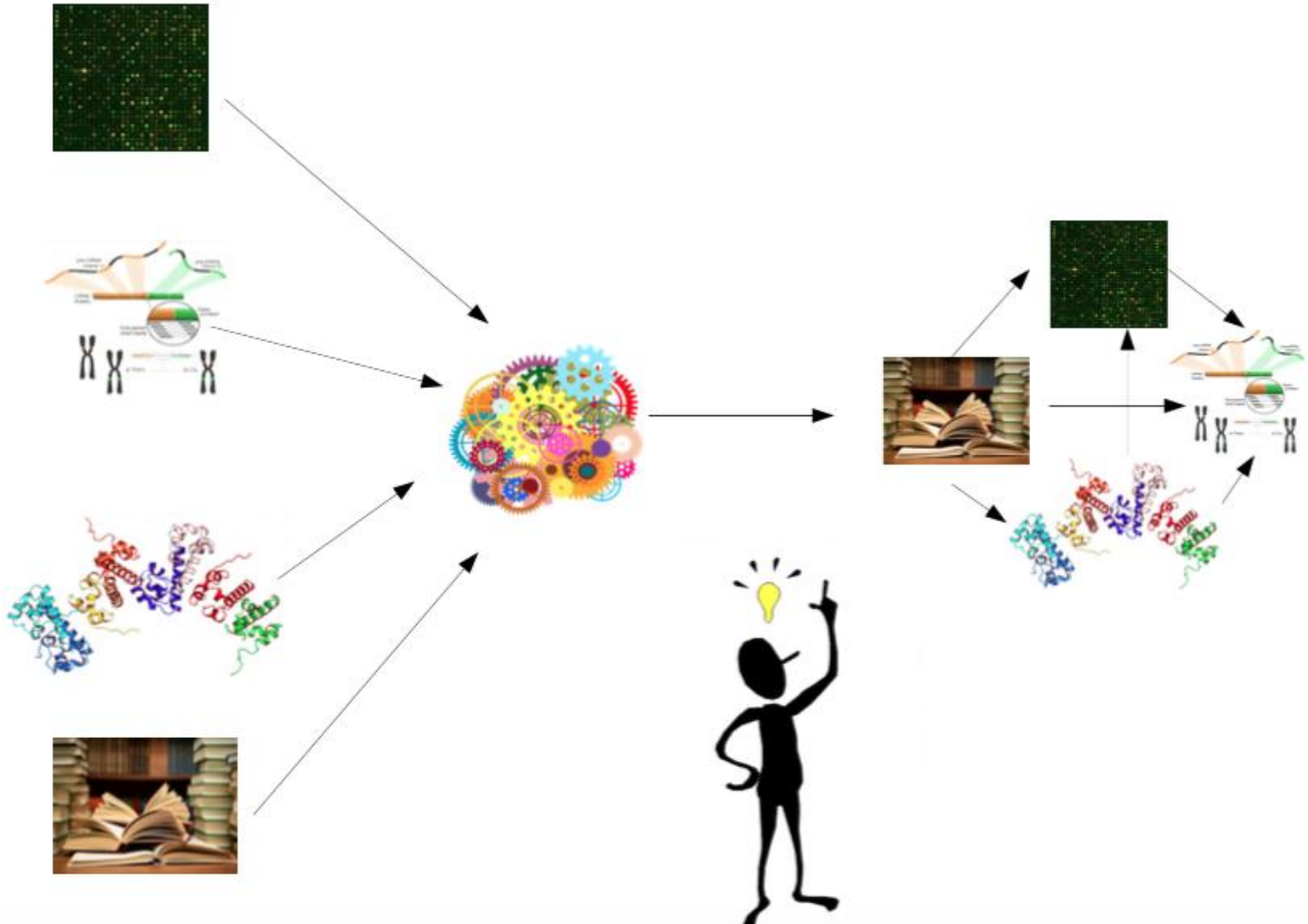


RNAseq

Medical Literature

Introduction

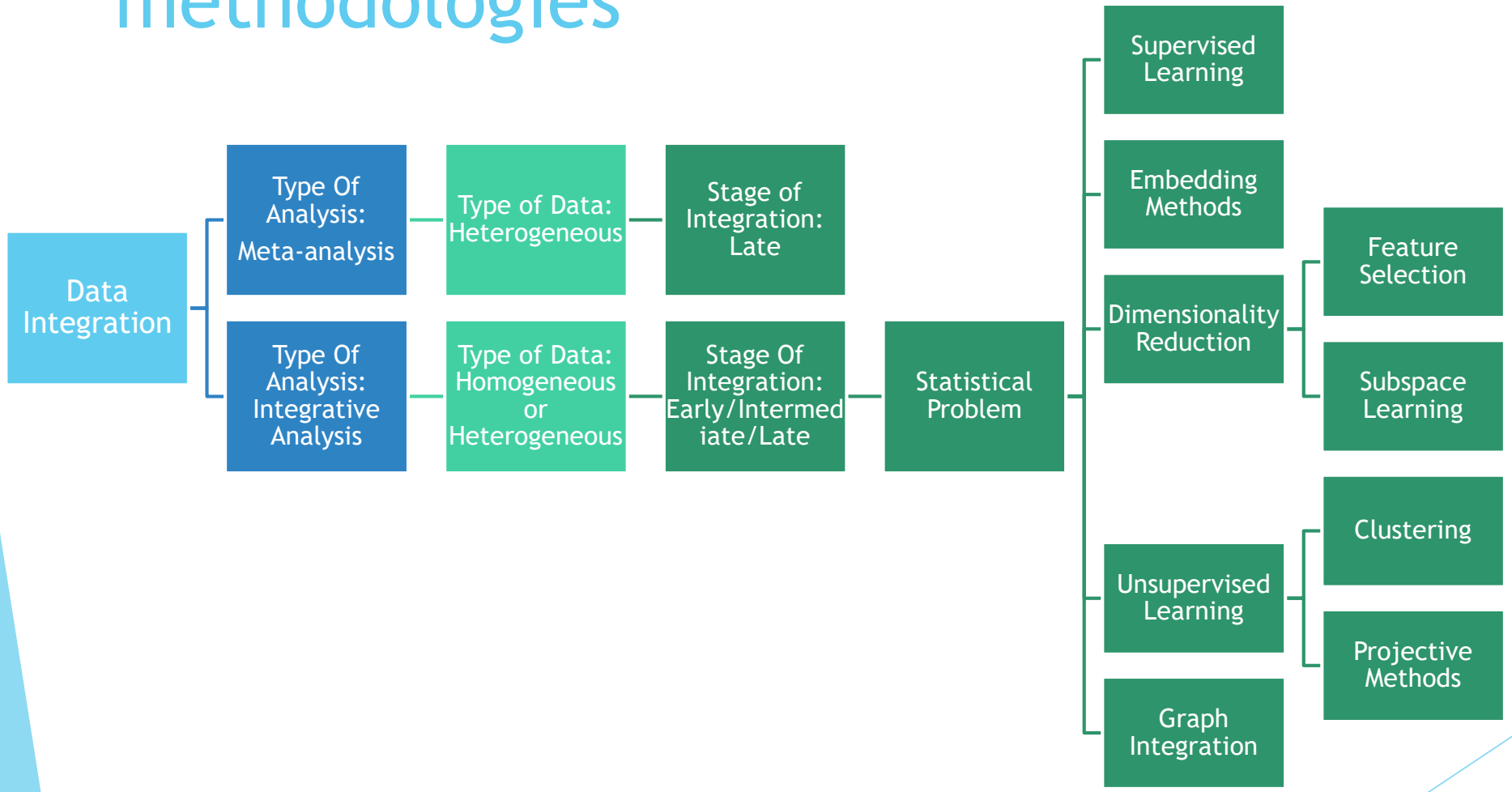
- Thanks to these multiple views, the learning task can be conducted with multi-view information.



Introduction

- In Bioinformatics multi-view approaches are useful since heterogeneous genome-wide data sources capture information on different aspects of complex biological systems.
- Each source provides a distinct “view” of the same domain, but potentially encodes different biologically-relevant patterns.
- Effective integration of such views can provide a richer model of an organism’s functional module than that produced by a single view alone

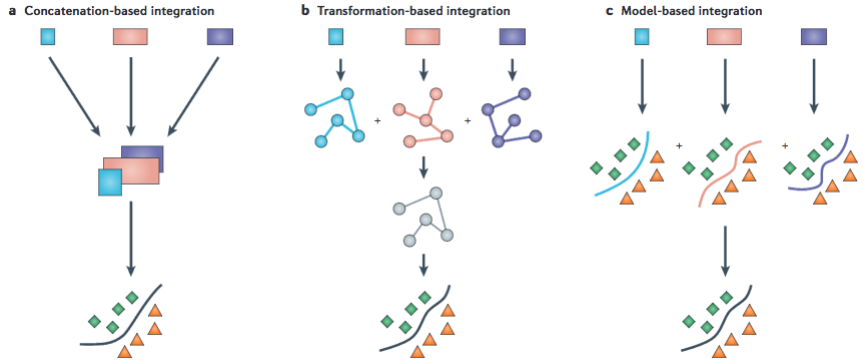
Classification of Data Integration methodologies



Classification of Data Integration methodologies

Table 1 | **Categorization of data analysis methods**

Approach	Methods	Software and/or tools	Refs
Multi-staged analysis			
Genomic variation analysis	eQTL, mQTL and causal analysis	Matrix eQTL[94] and QTDT[95]	43,94,95
	Allele-specific expression	AlleleSeq[96] and ChIP–SNP[48]	48,96
Domain knowledge-guided analysis	Correlation and mapping variation to pathway	ANNOVAR[97], HaploReg[98] and RegulomeDB[99]	97–100
Meta-dimensional analysis			
Concatenation-based integration	Grammatical evolution neural network	ATHENA	18,56,57
	Bayesian network	WinBUGS	54
	Multivariate Cox LASSO model	Glmpath	55
Transformation-based integration	Kernel-based integration	SKMsmo	59,60
	Graph-based semi-supervised learning	Graph-based semi-supervised learning	53,61,62
Model-based integration	Majority voting	ipred	64,65
	Ensemble classifier	Weka 3	66



- Meta-dimensional analysis can be divided into three categories.
  a) Concatenation-based integration involves combining data sets from different data types at the raw or processed data level before modelling and analysis.
  b) Transformation-based integration involves performing mapping or data transformation of the underlying data sets before analysis, and the modelling approach is applied at the level of transformed matrices.
  c) Model-based integration is the process of performing analysis on each data type independently, followed by integration of the resultant models to generate knowledge about the trait of interest.

Ritchie, Marylyn D., et al. "Methods of integrating data to uncover genotype-phenotype interactions." Nature Reviews Genetics 16.2 (2015): 85-97.

Type of Analysis

- The analysis to be performed is somehow limited by the type of data involved in the experiment and by the desired level of integration. Analyses can be divided in two categories:
  - Meta-analysis can be thought as an integrative study of previous results, usually performed aggregating the summary statistics from different studies. Due to its nature, meta-analysis can only be performed as a step of late integration involving homogeneous data.
  - Integrative analysis considers the fusion of different data sources in order to get more stable and reliable estimates. Based on the type of data and the stage of integration, new methodologies have been developed spanning a landscape of techniques comprising graph theory, machine learning and statistics.

Type Of Data

- Data integration methodologies in systems biology can be divided into two categories based on the type of data: integration of homogeneous or heterogeneous data types.
  - Usually biological data are thought to be homogeneous if they assay the same molecular level, for gene or protein expression, copy number variation, and so on.
  - On the other hand if data is derived from two or more different molecular levels they are considered to be heterogeneous. Integration of this kind of data poses some issues: first, the data can have different structure, for example they can be sequences, graphs, continuous or discrete numerical values.

Integration Stage

- Depending on the nature of the data and on the statistical problem to address, the integration of heterogeneous data can be performed at different levels:
  - Early integration
  - Intermediate Integration
  - Late Integration

Early Integration

- Early integration consists in concatenating data from different views in a single feature space, without changing the general format and nature of data.
- Early integration is usually performed in order to create a bigger pool of features by multiple experiments.
- The main disadvantage of early integration methodologies is given by the need to search for a suitable distance function. In fact, by concatenating views, the data dimensionality considerably increases, consequently decreasing the performance of the classical similarity measures .

Intermediate Integration

- Intermediate integration consists in transforming all the data sources in a common feature space before combining them.
- In classification problems, every view can be transformed in a similarity matrix that will be combined in order to obtain better results.

Late Integration

- In the late integration methodologies each view is analysed separately and the results are then combined.
- Late integration methodologies have some advantages over early integration techniques:
  - the user can choose the best algorithm to apply to each view based on the data;
  - the analysis on each view can be executed in parallel.

Supervised Learning

- In machine learning, supervised learning consists in inferring a function from labelled data.
- The input is a collection of samples defined as vectors on a set of features and a collection of labels, one for each sample.

Supervised Learning

Data Type	Aim	Stage of Integration	Testing Data	Comment
Heterogeneous	Classification	Early – Intermediate - Late	Real Dataset from Stanford University	Gene functional classification from heterogeneous data. Pavlidis et al.
Heterogeneous	Classification	Early – Intermediate - Late	Genomic Cancer Datasets	Information content and analysis methods for Multi-Modal High-Throughput Biomedical Data. Bisakha et al. .
Heterogeneous	Drugs classification and repositioning	Intermediate	CMAP Dataset	A multi layer drug repositioning approach. Napolitano et al.
Heterogeneous	Classification	Early	Webpage data and Advertisement data	Multi-view Fisher Discriminant Analysis (MFDA) which combines traditional FDA with multi-view learning. Chen et al.
Heterogeneous	Classification	Intermediate	PASCAL VOC (images)	Combines KCCA and SVM into a single optimisation termed SVM-2K. Larson et al.
Heterogeneous	Classification	Early - Late	CNN's audio and video	AVIS: a connectionist-based framework for integrated auditory and visual information processing. Kasabov et al.

Gene functional classification from heterogeneous data

- Brown et al. showed that SVM provides excellent classification performance on DNA microarray expression data.
- Pavlidis et al. extend the methodology of Brown et al. to learn gene functional classifications from a heterogeneous data set consisting of microarray expression data and phylogenetic profiles.
- SVMs are members of a larger class of algorithms, known as kernel methods, which can be non-linearly mapped to a higher-order feature space by replacing the dot product operation in the input space with a kernel function $K(\cdot, \cdot)$

Pavlidis, Paul, et al. "Gene functional classification from heterogeneous data." Proceedings of the fifth annual international conference on Computational biology. ACM, 2001.

Gene functional classification from heterogeneous data

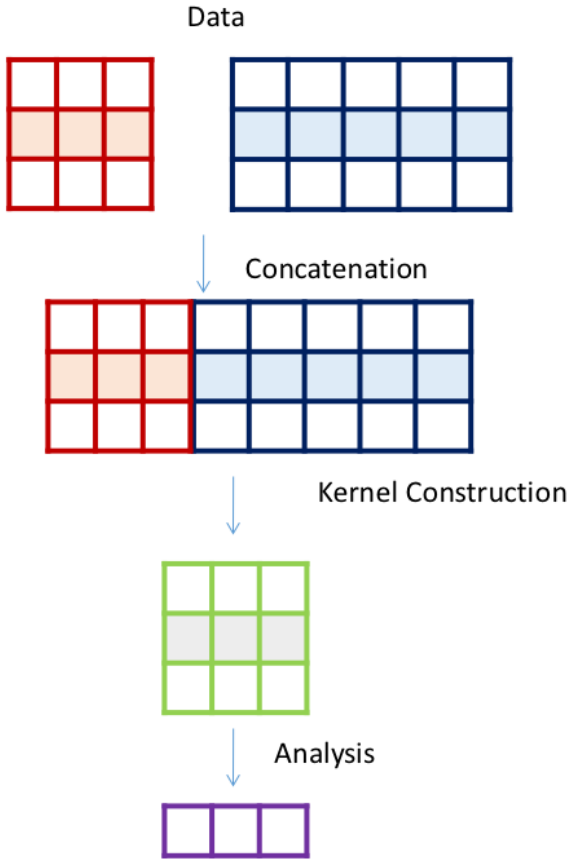
- The characteristics of the feature space are determined by a kernel function, which is selected a priori.
- The experiments employ this kernel function:

$$K(\vec{\mathbf{X}}, \vec{\mathbf{Y}}) = \left((\vec{\mathbf{X}} \cdot \vec{\mathbf{Y}} / \sqrt{\vec{\mathbf{X}} \cdot \vec{\mathbf{X}}} \sqrt{\vec{\mathbf{Y}} \cdot \vec{\mathbf{Y}}}) + 1\right)^3$$

Pavlidis, Paul, et al. "Gene functional classification from heterogeneous data." Proceedings of the fifth annual international conference on Computational biology. ACM, 2001.

Gene functional classification from heterogeneous data

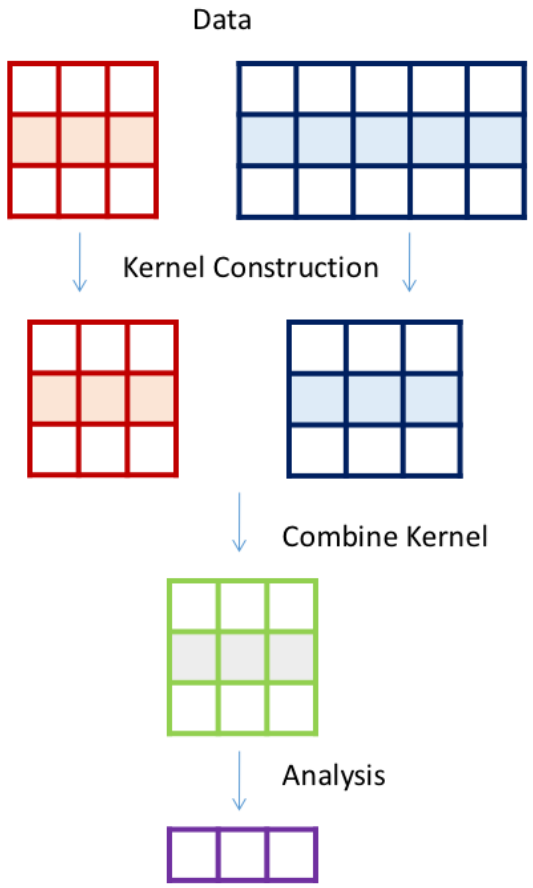
- The two types of data — gene expression and phylogenetic profiles — are combined in three different fashions, which we refer to as early, intermediate and late integration.
  - In early integration, the two types of vectors are concatenated to form a single vector which serve as input for the SVM.



Pavlidis, Paul, et al. "Gene functional classification from heterogeneous data." Proceedings of the fifth annual international conference on Computational biology. ACM, 2001.

Gene functional classification from heterogeneous data

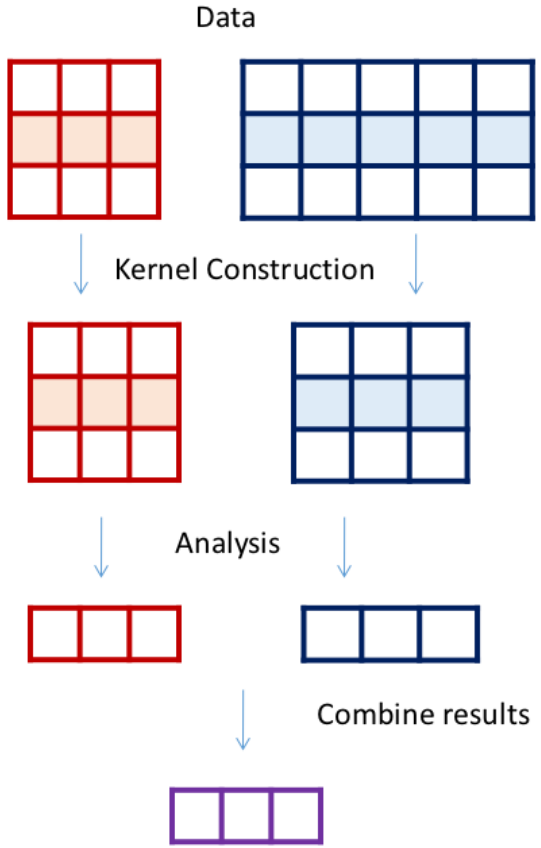
- The two types of data — gene expression and phylogenetic profiles — are combined in three different fashions, which we refer to as early, intermediate and late integration.
  - In intermediate integration, the kernel values for each type of data are pre-computed separately, and the resulting values are added together. These summed kernel values are used in the training of the SVM.



Pavlidis, Paul, et al. "Gene functional classification from heterogeneous data." Proceedings of the fifth annual international conference on Computational biology. ACM, 2001.

Gene functional classification from heterogeneous data

- The two types of data — gene expression and phylogenetic profiles — are combined in three different fashions, which we refer to as early, intermediate and late integration.
  - In late integration, one SVM is trained from each data type, and the resulting discriminant values are added together to produce a final discriminant for each gene.



Pavlidis, Paul, et al. "Gene functional classification from heterogeneous data." Proceedings of the fifth annual international conference on Computational biology. ACM, 2001.

Gene functional classification from heterogeneous data

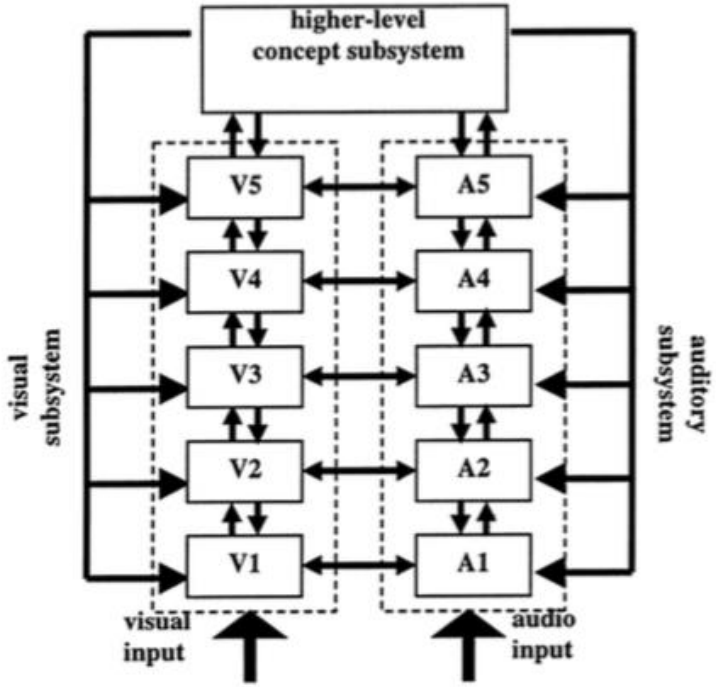
Class	Exp	Phylo	Early	Intermediate	Late
amino acid transporters	0.05 ± 0.04	$\mathbf{0.77 \pm 0.10}$	0.50 ± 0.04	$\mathbf{0.71 \pm 0.08}$	0.49 ± 0.07
ribosomal proteins	0.71 ± 0.02	0.09 ± 0.03	$\mathbf{0.76 \pm 0.01}$	0.71 ± 0.01	0.69 ± 0.01
sugar and carbohydrate transporters	0.33 ± 0.07	$\mathbf{0.67 \pm 0.02}$	$\mathbf{0.68 \pm 0.06}$	$\mathbf{0.70 \pm 0.01}$	0.63 ± 0.03
glycolysis and gluconeogenesis	0.21 ± 0.03	$\mathbf{0.43 \pm 0.05}$	0.28 ± 0.02	$\mathbf{0.39 \pm 0.05}$	$\mathbf{0.39 \pm 0.04}$
mitochondrial organization	$\mathbf{0.40 \pm 0.03}$	0.15 ± 0.01	$\mathbf{0.43 \pm 0.03}$	$\mathbf{0.42 \pm 0.02}$	0.35 ± 0.02
tricarboxylic acid pathway	0.21 ± 0.11	0.15 ± 0.07	$\mathbf{0.32 \pm 0.08}$	$\mathbf{0.42 \pm 0.07}$	$\mathbf{0.25 \pm 0.13}$
deoxyribonucleotide metabolism	0.07 ± 0.05	$\mathbf{0.31 \pm 0.11}$	$\mathbf{0.24 \pm 0.15}$	$\mathbf{0.39 \pm 0.11}$	$\mathbf{0.31 \pm 0.12}$
organization of cytoplasm	0.35 ± 0.01	0.18 ± 0.01	$\mathbf{0.38 \pm 0.01}$	0.34 ± 0.02	$\mathbf{0.35 \pm 0.02}$
transport ATPases	0.13 ± 0.04	$\mathbf{0.37 \pm 0.05}$	0.23 ± 0.05	$\mathbf{0.32 \pm 0.04}$	0.22 ± 0.03
amino acid biosynthesis	0.18 ± 0.02	0.28 ± 0.02	0.29 ± 0.03	$\mathbf{0.36 \pm 0.04}$	0.27 ± 0.02
purine ribonucleotide metabolism	0.17 ± 0.03	$\mathbf{0.26 \pm 0.05}$	0.20 ± 0.04	$\mathbf{0.33 \pm 0.04}$	0.19 ± 0.03
⋮					
organization of endoplasmatic reticulum	$\mathbf{0.20 \pm 0.02}$		$\mathbf{0.22 \pm 0.03}$	$\mathbf{0.19 \pm 0.05}$	0.13 ± 0.03
organization of cell wall	$\mathbf{0.12 \pm 0.04}$	$\mathbf{0.19 \pm 0.06}$	$\mathbf{0.14 \pm 0.08}$	$\mathbf{0.16 \pm 0.07}$	$\mathbf{0.21 \pm 0.08}$
anion transporters		$\mathbf{0.21 \pm 0.02}$			
Mean cost savings	0.19 ± 0.02	0.21 ± 0.04	0.27 ± 0.03	0.31 ± 0.03	0.24 ± 0.03
Number of best-performing	10	12	17	21	8
Number of non-learnable	4	6	3	2	3

- Each row in the table contains the cost savings for one MYGD classification. Each cost savings is computed via three-fold cross-validation, with standard deviation calculated across five repetitions. Decide to integrate or to not integrate and the type of integration to perform strongly depend on the data.

Pavlidis, Paul, et al. "Gene functional classification from heterogeneous data." Proceedings of the fifth annual international conference on Computational biology. ACM, 2001.

AVIS: a connectionist-based framework for integrated auditory and visual information processing

- Kasabov et al proposed the AVIS framework for studying the integrated processing of auditory and visual information in order to recognize people.
- They proposed a hierarchical architecture consists of three subsystems
  - an auditory subsystem
  - a visual subsystem
  - A higher-level conceptual subsystem



Kasabov, Nikola, Eric Postma, and Jaap Van Den Herik. "AVIS: a connectionist-based framework for integrated auditory and visual information processing." Information Sciences 123.1 (2000): 127-148.

AVIS: a connectionist-based framework for integrated auditory and visual information processing

- They proposed four modes of operation:
  a) The unimodal visual mode takes visual information as input (e.g., a face), and classifies it. The classification result is passed to the conceptual subsystem for identification.
  b) The unimodal auditory mode deals with the task of voice recognition. The classification result is passed to the conceptual subsystem for identification.
  c) The bimodal (or early-integration) mode combines the bimodal and cross- modal modes of AVIS by merging auditory and visual information into a single (multimodal) subsystem for person identification.
  d) The combined mode synthesises the results of all three modes (a), (b) and (c). The three classification results are fed into the conceptual subsystem for person identification.

Kasabov, Nikola, Eric Postma, and Jaap Van Den Herik. "AVIS: a connectionist-based framework for integrated auditory and visual information processing." Information Sciences 123.1 (2000): 127-148.

AVIS: a connectionist-based framework for integrated auditory and visual information processing

- They performed experiment on digital video downloaded from the CNN's website:
  - They downloaded a digital video containing small fragments of four American talk-show
  - They recorded CNN broadcasts of eight fully-visible and audibly-speaking presenters of sport and news programs
- The experimental results support the hypothesis that the recognition rate is considerably enhanced by combining visual and auditory dynamic features.

Kasabov, Nikola, Eric Postma, and Jaap Van Den Herik. "AVIS: a connectionist-based framework for integrated auditory and visual information processing." Information Sciences 123.1 (2000): 127-148.

Embedding Methods

- Dimensionality reduction of high dimensional multi-view data can be a non-trivial task because of the underlying connections between the features in the different views.
- A solution is to embed the multi-view patterns simultaneously into a low-dimensional space shared by all features.

Embedding Methods

- An example of embedding methods is Stochastic Neighbour Embedding (SNE) that constructs a low-dimensional manifold such that the density of low-dimensional data approximates the original density in the original high-dimensional space.
- Density is estimated as pairwise distances in the original feature space and the resulting embedding is obtained minimising the Kullback-Leibler divergence among the high- and low- dimensional densities.
- Multi-view SNE is an extension of the original method that replaces the original estimated density with a combination of pairwise densities, each constructed from a different view. The corresponding objective includes 2-norm regularization among the combination weights, plus a trade-off to balance the objective and the regularise.

Hinton, Geoffrey E., and Sam T. Roweis. "Stochastic neighbor embedding." Advances in neural information processing systems. 2002.

Xie, Bo, et al. "m-SNE: Multiview stochastic neighbor embedding." Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on 41.4 (2011): 1088-1096.

Dimensionality Reduction: Feature Selection

- The goal of feature selection is to express high-dimensional data with a low number of features to reveal significant underlined information. It is mainly used as a pre-processing step for other computational methodologies.
- Three different approaches are proposed in literature:
  - The univariate filter methods
  - The multivariate wrapper
  - The multivariate embedded methods.
- They have the common goal of finding the smallest set of features useful to correctly classify objects. Accuracy and stability are the two main requirements for feature selection methodologies.

Dimensionality Reduction: Feature Selection

Data Type	Aim	Stage of Integration	Testing Data	Comment
Heterogeneous	Feature Selection	-	Gene Expression	A Robust and Accurate Method for Feature Selection and Prioritization from Multi- Class OMICs Data. Fortino et al. [25]
Heterogeneous	Feature Selection	Late	Gene Expression Multiple Tissues	A sparse multi-view matrix factorization method for gene prioritization in gene expression datasets for multiple tissues. Larson et al. [31]

Dimensionality Reduction: Feature Selection

- Fortino et al. proposed a wrapper feature selection method based on fuzzy logic and random forests that is able to guarantee good performance and high stability.
- The first step of their algorithm consists in a discretization step where the gene expression data are transformed into Fuzzy Patterns (FP) that give information about the most relevant features of each category.
- Then a random forest is used to classify data using priori knowledge about the fuzzy patterns.
- As last step, they ranked the selected features with a permutation variable importance measure.

Fortino, Vittorio, et al. "A Robust and Accurate Method for Feature Selection and Prioritization from Multi-Class OMICs Data." PloS one 9.9 (2014): e107801.

Dimensionality Reduction: Feature Selection

- They tested their method on different gene expression multi-class datasets and compared their results with other two random forest based feature selection methods: varSelRF and Borda.
- Accuracy was estimated with F -score and G-score, two measures particularly appropriate for multi-class unbalanced problems.
- Stability was evaluated by executing the method for 30 bootstrap iterations. During the iterations, the significantly consistent features were selected.
- The final stability metric was defined as the ratio between the number of consistent features and the total number of selected features.
- Results show that their system has similar or better results compared to the other methods proposed in literature.

Fortino, Vittorio, et al. "A Robust and Accurate Method for Feature Selection and Prioritization from Multi-Class OMICs Data." PloS one 9.9 (2014): e107801.

Dimensionality Reduction: Subspace Learning

- The aim of subspace learning approaches is to find a latent subspace shared by multiple views.
- One of the most cited approaches used to model the relationships between two (or more) views is Canonical Correlation Analysis (CCA).
- Consider two sets of variables X and Y
- How to find the connection between the two sets of variables?
  - CCA: find a projection direction w_x in the space of X and w_y in the space of Y, so that projected data onto w_x and w_y has max correlation.
  - Note: CCA simultaneously makes dimensional reduction for both the two feature spaces
- It was defined for datasets with two views but it was later generalized to data with more than two representations in several ways (Kettenring, 1971 - Batch, 2002)

Dimensionality Reduction: Subspace Learning

- The problem with CCA is that most of the connections between objects in real datasets cannot be expressed with linear relations.
- A solution is given by kernel methods that map data into a higher dimensional space and then apply linear methods in that space.
- Kernel Canonical Correlation Analysis (KCCA) is the kernelized non linear version of CCA.

Dimensionality Reduction: Subspace Learning

- KCCA is widely used in genomics, in particular for the analysis of data from Genome-Wide Association Studies (GWAS).
- GWAS is used for the detection of genetic variants of complex diseases. So far, studies focused on the association of a Single Nucleotide Polymorphism (SNP) with a specific trait.
- Applying more sophisticated methods like KCCA, researchers can focus on more complex interactions between genes and specific traits of interest
  - For example, Larson et al. developed a KCCA method able to identify associations between genes for complex phenotypes from a case-control study in genome-wide SNP data.
  - They applied the approach to find interaction between genes in an ovarian cancer dataset with 3869 cases and 3276 controls.
  - They were able to identify 13 gene pairs highly predictive of ovarian cancer risk.

Unsupervised Learning

- In machine learning, the unsupervised learning is defined as the problem of identifying hidden structures in unlabelled data.
- This means that the learner tries to group data by comparing the patterns based on their similarities.
- Here we focus in particular on multi-view clustering techniques.

Unsupervised Learning: Clustering

- Clustering is used when we want to extract information from data without any previous knowledge
- What does clustering mean?
- Given a set of objects X = {x1,...,xn}, clustering is a partition P = {P1, . . . , Pk } of these objects such that

$$\bigcup_{i=1}^{k} P_i = X \text{ e } P_i \cap P_j = \emptyset \; \forall i \neq j$$

- Each cluster contains similar objects and different objects are in different clusters.
- The result depends on the (dis)similarity function.

Unsupervised Learning: Clustering Differences between traditional and Multi View Clustering

- Traditional clustering methods take multiple views as a flat set of variables and ignore the differences among different views,
- Multiview clustering exploits the information from multiple views and take the differences among different views into consideration in order to produce a more accurate and robust data partitioning.

Unsupervised Learning: Clustering

Data Type	Aim	Stage of Integration	Testing Data	Comment
Heterogeneous	Clustering	Early	Swissprot protein database and Image Dataset	Multi-view DBSCAN. Kailing et al
Heterogeneous	Clustering	Early	UCI Machine Learning Repository:	Multi-View weighted version of K-means. Chen et al.
Heterogeneous	Clustering	Late	Synthetic Dataset	A General Model for Multiple View Unsupervised Learning. Long et al.
Heterogeneous	Clustering	Late	Synthetic Dataset	Matrix Factorization. Greene, Derek.
Heterogeneous	Clustering	Late	Genomic Cancer datasets	A multi-view clustering integration methodology for cancer subtype. Serra et al.
Heterogeneous	Biclustering	Late	Ovaria Cancer	A non-negative matrix factorization method for detecting modules in heterogeneous omics multi-modal data Yang et al.
Heterogeneous	Clustering	Late	TCGA Dataset	Multi-omic integration approach that supports visual exploration of the data, and inspection of the contribution of the different genome-wide data-types. Taskesen et al.

Unsupervised Learning: Clustering DBSCAN Multi-View

- The method proposed by Kailing et al. is based on the DBSCAN algorithm.
- The method works with as many views as you want.
- It finds a multi-view clustering by combining core objects found in each view with two approach:
  - Union method: for sparse data
  - Intersection method: well suited for data containing
  - unreliable representations

Kailing, Karin, et al. "Clustering multi-represented objects with noise." Advances in Knowledge Discovery and Data Mining. Springer Berlin Heidelberg, 2004. 394-403.

Unsupervised Learning: Clustering DBSCAN Multi-View

- DBSCAN MV – Example Of Application:

- Clustering image data is a good example for the usefulness of the intersection-method.
- A lot of different similarity models exists for image data, each having its own advantages and disadvantages.
- Using for example text descriptions of images, one is able to cluster all images related to a certain topic, but these images must not look alike.
- Using color histograms instead, the images are clustered according to the distribution of color in the image.

Kailing, Karin, et al. "Clustering multi-represented objects with noise." Advances in Knowledge Discovery and Data Mining. Springer Berlin Heidelberg, 2004. 394-403.

Unsupervised Learning: Clustering DBSCAN Multi-View

- DBSCAN MV – Example Of Application:

- The first representation was a 64-dimensional colour histogram. In this case, we used the weighted distance between those colour histograms.
- The second representation were segmentation trees. An image was first divided into segments of similar colour by a segmentation algorithm. In a second step, a tree was created from those segments by iteratively applying a region-growing algorithm which merges neighbouring segments, if their colours are alike. The similarity between two such trees is computed using filters for the complex edit-distance measure.

Kailing, Karin, et al. "Clustering multi-represented objects with noise." Advances in Knowledge Discovery and Data Mining. Springer Berlin Heidelberg, 2004. 394-403.

Unsupervised Learning: Clustering DBSCAN Multi-View . .

Cluster IC5 found by intersection-method

Images that are in the corresponding cluster built on histograms but not added to IC5

Images that are in the corresponding cluster built on segmentation trees but not added to IC 5

Fig. 3. Example of an image cluster. The left rectangle contains images clustered by the intersection-method. The right rectangles display additional images that were grouped with the corresponding cluster when clustering the images with respect to a single representation.

Kailing, Karin, et al. "Clustering multi-represented objects with noise." Advances in Knowledge Discovery and Data Mining. Springer Berlin Heidelberg, 2004. 394-403.

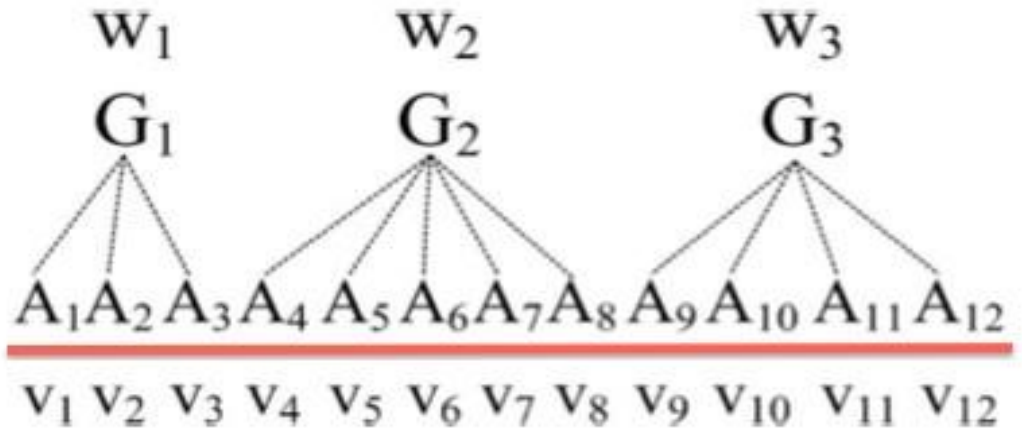
Unsupervised Learning: Clustering TW-Kmeans

- It is a two level variable weighting k-means clustering algorithm for multi-view data.
- The weights of views and individual variables are included into the distance function.
- It is an extension of the k-means algorithm with two more steps that should not require intensive computation so it should have the same computation complexity as k-means.

Chen X, Xu X, Huang JZ, Ye Y. Tw- (k)-means: Automated two-level variable weighting clustering algorithm for multiview data. Knowl Data Eng IEEE Trans. 2013; 25(4):932–44.

Unsupervised Learning: Clustering TW-Kmeans

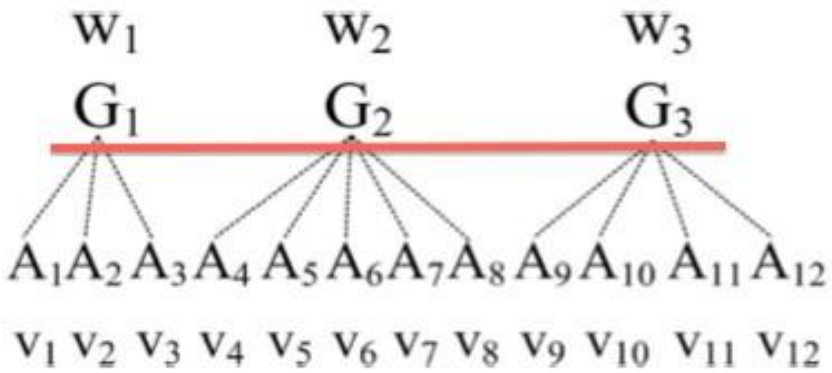
- Let $X = \{X_1, X_2, \ldots, X_n\}$ be a set of n objects represented by a set A of m variables.



Chen X, Xu X, Huang JZ, Ye Y. Tw- (k)-means: Automated two-level variable weighting clustering algorithm for multiview data. Knowl Data Eng IEEE Trans. 2013; 25(4):932–44.

Unsupervised Learning: Clustering TW-Kmeans

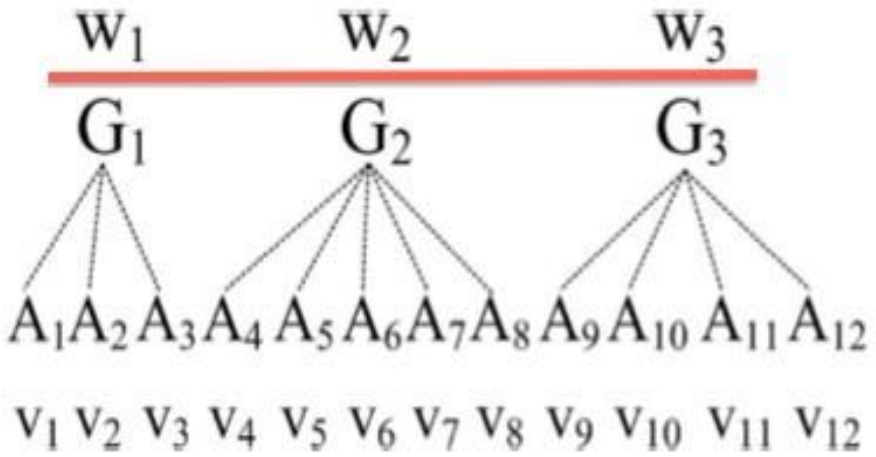
- Assume A is divided into T views $\{G_t\}_{t=1}^{T}$ where $G_t \cap G_s = \emptyset$ for $s \neq t$ and $\cup_{t=1}^{T} G_t = A$.



Chen X, Xu X, Huang JZ, Ye Y. Tw- (k)-means: Automated two-level variable weighting clustering algorithm for multiview data. Knowl Data Eng IEEE Trans. 2013; 25(4):932–44.

Unsupervised Learning: Clustering TW-Kmeans

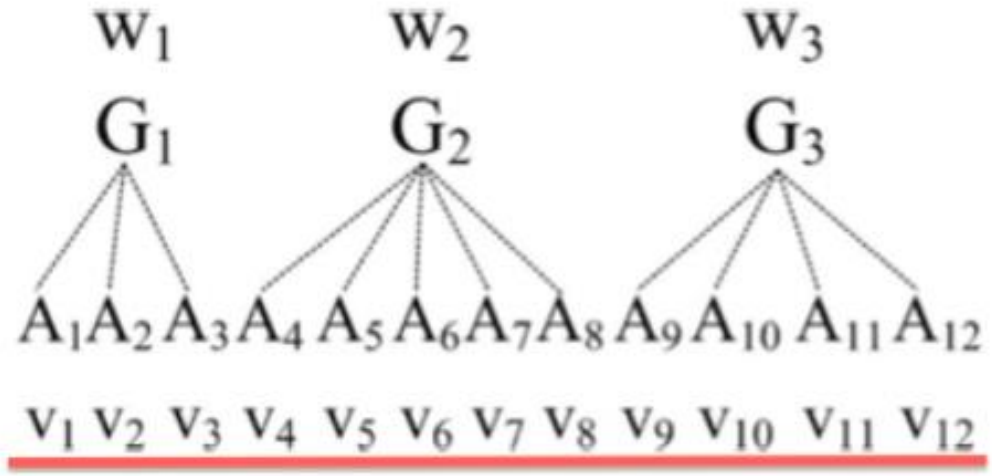
- Let $W = \{w_1, w_2, \ldots, w_T\}$ be a set of T weights, where w_t indicates the weight that is assigned to the tth view and $\sum_{t=1}^{T} w_t = 1$.



Chen X, Xu X, Huang JZ, Ye Y. Tw- (k)-means: Automated two-level variable weighting clustering algorithm for multiview data. Knowl Data Eng IEEE Trans. 2013; 25(4):932–44.

Unsupervised Learning: Clustering TW-Kmeans

- Let $V = \{V_j\}$ be a set of m variable weights, where v_j indicates the weight that is assigned to the j th variable and $\sum_{j \in G_t} v_j = 1\ (1 \leq t \leq T)$, $\sum_{j=1}^{m} v_j = T$



Chen X, Xu X, Huang JZ, Ye Y. Tw- (k)-means: Automated two-level variable weighting clustering algorithm for multiview data. Knowl Data Eng IEEE Trans. 2013; 25(4):932–44.

Unsupervised Learning: Clustering TW-Kmeans

- Assume that X contains k clusters. We want to identify:
  - the set of k clusters from G.
  - the relevant views from the view weight matrix W = $[wt]_T$
  - the relevant variables from the variable weight matrix V =$[vj]_m$

Chen X, Xu X, Huang JZ, Ye Y. Tw- (k)-means: Automated two-level variable weighting clustering algorithm for multiview data. Knowl Data Eng IEEE Trans. 2013; 25(4):932–44.

Unsupervised Learning: Clustering TW-Kmeans

- The Optimization Model
- The clustering process to partition X into k clusters with weights for both views and individual variables is modeled as the minimization problem:

$$P(U,Z,V,W) = \sum_{l=1}^{k}\sum_{i=1}^{n}\sum_{t=1}^{T}\sum_{j\in G_t} u_{i,l} w_t v_j d(x_{i,j}, z_{l,j}) + \eta \sum_{j=1}^{m} v_j \log(v_j) + \lambda \sum_{t=1}^{T} w_t \log(w_t) \quad (1)$$

subject to

$$\begin{cases} \sum_{l=1}^{k} u_{i,l} = 1, & u_{i,l} \in \{0,1\}, \quad 1 \le i \le n, \\ \sum_{t=1}^{T} w_t = 1, & 0 \le w_t \le 1, \\ \sum_{j\in G_t} v_j = 1, & 0 \le v_j \le 1, \quad 1 \le t \le T, \end{cases} \quad (2)$$

Chen X, Xu X, Huang JZ, Ye Y. Tw- (k)-means: Automated two-level variable weighting clustering algorithm for multiview data. Knowl Data Eng IEEE Trans. 2013; 25(4):932–44.

Unsupervised Learning: Clustering TW-Kmeans

- The Optimization Model
- Where:

$$P(U,Z,V,W) = \sum_{l=1}^{k}\sum_{i=1}^{n}\sum_{t=1}^{T}\sum_{j\in G_t} u_{i,l} w_t v_j d(x_{i,j}, z_{l,j}) + \eta \sum_{j=1}^{m} v_j \log(v_j) + \lambda \sum_{t=1}^{T} w_t \log(w_t) \quad (1)$$

U is a nxk partition matrix whose elements $u_{i,l}$ are binary where $u_{i,l} = 1$ indicates that object i is allocated to cluster l;

Chen X, Xu X, Huang JZ, Ye Y. Tw- (k)-means: Automated two-level variable weighting clustering algorithm for multiview data. Knowl Data Eng IEEE Trans. 2013; 25(4):932–44.

Unsupervised Learning: Clustering TW-Kmeans

- The Optimization Model
- Where:

$$P(U,Z,V,W) = \sum_{l=1}^{k}\sum_{i=1}^{n}\sum_{t=1}^{T}\sum_{j\in G_t} u_{i,l} w_t v_j d(x_{i,j}, z_{l,j}) + \eta \sum_{j=1}^{m} v_j \log(v_j) + \lambda \sum_{t=1}^{T} w_t \log(w_t) \quad (1)$$

$Z = \{Z_1, Z_2, \ldots, Z_k\}$ is a set of k vectors representing the centers of the k clusters;

Chen X, Xu X, Huang JZ, Ye Y. Tw- (k)-means: Automated two-level variable weighting clustering algorithm for multiview data. Knowl Data Eng IEEE Trans. 2013; 25(4):932–44.

Unsupervised Learning: Clustering TW-Kmeans

- The Optimization Model
- Where:

$$P(U, Z, V, W) = \sum_{l=1}^{k}\sum_{i=1}^{n}\sum_{t=1}^{T}\sum_{j\in G_t} u_{i,l} w_t v_j d(x_{i,j}, z_{l,j}) + \eta \sum_{j=1}^{m} v_j \log(v_j) + \lambda \sum_{t=1}^{T} w_t \log(w_t) \quad (1)$$

$V = \{v_1, v_2, \ldots, v_m\}$ are m weights for m variables;

Chen X, Xu X, Huang JZ, Ye Y. Tw- (k)-means: Automated two-level variable weighting clustering algorithm for multiview data. Knowl Data Eng IEEE Trans. 2013; 25(4):932–44.

Unsupervised Learning: Clustering TW-Kmeans

- The Optimization Model
- Where:

$$P(U,Z,V,W) = \sum_{l=1}^{k}\sum_{i=1}^{n}\sum_{t=1}^{T}\sum_{j\in G_t} u_{i,l} w_t v_j d(x_{i,j}, z_{l,j}) + \eta \sum_{j=1}^{m} v_j \log(v_j) + \lambda \sum_{t=1}^{T} w_t \log(w_t) \quad (1)$$

$W = \{w_1, w_2, \ldots, w_T\}$ are T weights for T views;

Chen X, Xu X, Huang JZ, Ye Y. Tw- (k)-means: Automated two-level variable weighting clustering algorithm for multiview data. Knowl Data Eng IEEE Trans. 2013; 25(4):932–44.

Unsupervised Learning: Clustering TW-Kmeans

- The Optimization Model
- Where:

$$P(U,Z,V,W) = \sum_{l=1}^{k}\sum_{i=1}^{n}\sum_{t=1}^{T}\sum_{j\in G_t} u_{i,l} w_t^{\beta} v_j^{\alpha} d(x_{i,j}, z_{l,j}) + \eta \sum_{j=1}^{m} v_j \log(v_j) + \lambda \sum_{t=1}^{T} w_t \log(w_t) \quad (1)$$

$d(x_{i,j}, z_{l,j})$ is a distance or dissimilarity measure on the jth variable between the ith object and the center of the lth cluster. If the variable is numerical, then

$$d(x_{i,j}, z_{l,j}) = (x_{i,j} - z_{l,j})^2.$$

If the variable is categorical, then

$$d(x_{i,j}, z_{l,j}) = \begin{cases} 0, & (x_{i,j} = z_{l,j}), \\ 1, & (x_{i,j} \neq z_{l,j}). \end{cases}$$

Chen X, Xu X, Huang JZ, Ye Y. Tw- (k)-means: Automated two-level variable weighting clustering algorithm for multiview data. Knowl Data Eng IEEE Trans. 2013; 25(4):932–44.

Unsupervised Learning: Clustering TW-Kmeans

- The Optimization Model
- Where:

$$P(U,Z,V,W) = \sum_{l=1}^{k}\sum_{i=1}^{n}\sum_{t=1}^{T}\sum_{j\in G_t} u_{i,l} w_t v_j d(x_{i,j}, z_{l,j}) + \eta \sum_{j=1}^{m} v_j \log(v_j) + \lambda \sum_{t=1}^{T} w_t \log(w_t) \quad (1)$$

λ and η are two given parameters

Chen X, Xu X, Huang JZ, Ye Y. Tw- (k)-means: Automated two-level variable weighting clustering algorithm for multiview data. Knowl Data Eng IEEE Trans. 2013; 25(4):932–44.

Unsupervised Learning: Clustering TW-Kmeans

- The Optimization Model

$$P(U,Z,V,W) = \underbrace{\sum_{l=1}^{k}\sum_{i=1}^{n}\sum_{t=1}^{T}\sum_{j\in G_t} u_{i,l} w_t v_j d(x_{i,j}, z_{l,j})}_{} + \underbrace{\eta\sum_{j=1}^{m} v_j \log(v_j) + \lambda\sum_{t=1}^{T} w_t \log(w_t)}_{} \quad (1)$$

- The first term in (1) is the sum of the within cluster dispersion
- The second and the third terms are two negative weight entropies
- Two positive parameters λ and η control the strengths of the incentive for clustering on more views and variables

Chen X, Xu X, Huang JZ, Ye Y. Tw- (k)-means: Automated two-level variable weighting clustering algorithm for multiview data. Knowl Data Eng IEEE Trans. 2013; 25(4):932–44.

Unsupervised Learning: Clustering TW-Kmeans

- The Optimization Model

subject to

$$\begin{cases} \sum_{l=1}^{k} u_{i,l} = 1, & u_{i,l} \in \{0,1\}, \quad 1 \le i \le n, \\ \sum_{t=1}^{T} w_t = 1, & 0 \le w_t \le 1, \\ \sum_{j \in G_t} v_j = 1, & 0 \le v_j \le 1, \quad 1 \le t \le T, \end{cases} \tag{2}$$

- An object i can be part of only one cluster l
- The sum of the view weights must be one
- The sum of the variable weights inside a view must be one

Chen X, Xu X, Huang JZ, Ye Y. Tw- (k)-means: Automated two-level variable weighting clustering algorithm for multiview data. Knowl Data Eng IEEE Trans. 2013; 25(4):932–44.

Unsupervised Learning: Clustering TW-Kmeans

- The Optimization Model
- We can minimize (1) by iteratively solving the following four minimization problems:

1. Problem P_1: Fix $Z = \hat{Z}$, $V = \hat{V}$, and $W = \hat{W}$, and solve the reduced problem $P(U, \hat{Z}, \hat{V}, \hat{W})$;
2. Problem P_2: Fix $U = \hat{U}$, $V = \hat{V}$, and $W = \hat{W}$, and solve the reduced problem $P(\hat{U}, Z, \hat{V}, \hat{W},)$;
3. Problem P_3: Fix $U = \hat{U}$, $Z = \hat{Z}$ and $W = \hat{W}$, and solve the reduced problem $P(\hat{U}, \hat{Z}, V, \hat{W})$;
4. Problem P_4: Fix $U = \hat{U}$, $Z = \hat{Z}$, and $V = \hat{V}$, and solve the reduced problem $P(\hat{U}, \hat{Z}, \hat{V}, W)$.

Chen X, Xu X, Huang JZ, Ye Y. Tw- (k)-means: Automated two-level variable weighting clustering algorithm for multiview data. Knowl Data Eng IEEE Trans. 2013; 25(4):932–44.

Unsupervised Learning: Clustering TW-Kmeans

- To investigate the performance of the TW-k-means algorithm in classifying real-life data, the authors selected three data sets from the UCI Machine Learning Repository:
  - the Multiple Features data set,
  - the Internet Advertisement data set
  - the Image Segmentation data set
- With these data sets, they compared TW-k-means with four individual variable weighting clustering algorithms, W-k-means, EW-k-means, LAC, EWKM and a weighted multi-view clustering algorithm WCMM

Chen X, Xu X, Huang JZ, Ye Y. Tw- (k)-means: Automated two-level variable weighting clustering algorithm for multiview data. Knowl Data Eng IEEE Trans. 2013; 25(4):932–44.

Unsupervised Learning: Clustering TW-Kmeans

- The next table summarizes the total 1,503 clustering results. From these results, we can see that TW-k-means significantly outperformed the other five algorithms in almost all results

TABLE 2
Summary of Clustering Results on Three Real-Life Data Sets by Six Clustering Algorithms

Data	Evaluation indices	W-k-means	EW-k-means	LAC	EWKM	WCMM	TW-k-means
MF	Precision	−0.06(.10)∗	−0.07(.10)∗	−0.07(.09)∗	−0.24(.08)∗	−0.59(.00)∗	0.79(.09)
	Recall	−0.09(.09)∗	−0.09(.09)∗	−0.09(.08)∗	−0.36(.12)∗	−0.56(.00)∗	0.82(.08)
	F-measure	−0.08(.10)∗	−0.08(.10)∗	−0.08(.08)∗	−0.41(.12)∗	−0.59(.00)∗	0.80(.09)
	Accuracy	−0.09(.09)∗	−0.09(.09)∗	−0.09(.08)∗	−0.36(.12)∗	−0.56(.00)∗	0.82(.08)
IA	Precision	−0.16(.19)∗	−0.16(.20)∗	−0.14(.20)∗	−0.22(.19)∗	−0.56(.00)∗	0.72(.12)
	Recall	−0.14(.04)∗	−0.10(.07)∗	−0.10(.08)∗	−0.13(.06)∗	−0.33(.00)∗	0.72(.07)
	F-measure	−0.23(.04)∗	−0.17(.12)∗	−0.17(.12)∗	−0.21(.09)∗	−0.47(.00)∗	0.69(.11)
	Accuracy	−0.14(.04)∗	−0.10(.07)∗	−0.10(.08)∗	−0.13(.06)∗	−0.33(.00)∗	0.72(.07)
IS	Precision	−0.03(.07)∗	−0.04(.08)∗	−0.03(.07)∗	−0.03(.09)∗	−0.37(.00)∗	0.62(.09)
	Recall	−0.03(.05)∗	−0.03(.03)∗	−0.03(.05)∗	−0.03(.05)∗	−0.41(.00)∗	0.64(.05)
	F-measure	−0.01(.07)∗	−0.02(.05)∗	−0.01(.07)∗	−0.02(.07)∗	−0.40(.00)∗	0.60(.07)
	Accuracy	−0.03(.05)∗	−0.03(.03)∗	−0.03(.05)∗	−0.03(.05)	−0.41(.00)∗	0.64(.05)

The value of the TW-k-means algorithm is the mean value of 100 results and the other values are the differences of the mean values between the corresponding algorithms and the TW-k-means algorithm. The value in brackets is the standard deviation of 100 results. "" indicates that the difference is significant.*

Chen X, Xu X, Huang JZ, Ye Y. Tw- (k)-means: Automated two-level variable weighting clustering algorithm for multiview data. Knowl Data Eng IEEE Trans. 2013; 25(4):932–44.

Unsupervised Learning: Clustering

Late Integration

- Unification of patterns can also be seen as the next step of a data mining pipeline in which the preceding step is the clustering of objects on each single view.
  This distributed approach (as opposed to the centralized one) has some benefits as:
  - Clustering algorithms can be chosen with respect to the application domain.
  - Natural parallelization possibility.
  - Representation issues are avoided since clustering results are the inputs.
  - Suitable in privacy-preserving use cases.

Greene D. A Matrix Factorization Approach for Integrating Multiple Data Views. Mach Learn Knowl Discov Databases. 2009; 5781:423–38.

Unsupervised Learning: Clustering

Notation and Formulation

- Given a set of views $\{V_1, \ldots, V_v\}$ denoting n objects $x_1, \ldots, x_n$, the goal is a consistent clustering between the views.
- The input is a set of clusterings C = $\{C_1, \ldots, C_v\}$ where each C_h represents a clustering of the view V_h. Clustering can be obtained by
  - Partitive clustering algorithms (k-means)
  - Probabilistic models (EM clustering)
  - Threshold based hierarchical clustering
  - Any other reasonable clustering method

Greene D. A Matrix Factorization Approach for Integrating Multiple Data Views. Mach Learn Knowl Discov Databases. 2009; 5781:423–38.

Unsupervised Learning: Clustering

Notation and Formulation

- Each clustering is represented as a membership matrix
- $\mathbf{M}_h \in \mathbb{R}^{n \times k_h}$ where k_h is the number of clusters on view V_h. If an object is not present in V_h then the corresponding row is filled with zeros.

Example

$$\mathcal{C}_1 = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix} \quad \mathcal{C}_2 = \begin{bmatrix} 0.5 & 0.5 \\ 0.5 & 0.5 \\ 0.7 & 0.3 \\ 0.6 & 0.4 \\ 0.2 & 0.8 \\ 0.1 & 0.9 \end{bmatrix}$$

Greene D. A Matrix Factorization Approach for Integrating Multiple Data Views. Mach Learn Knowl Discov Databases. 2009; 5781:423–38.

Unsupervised Learning: Clustering

Matrix Factorization for Multi-View Clustering

- This algorithm combines information by factorizing the “matrix of clusters”.
- This factorization produces a projection of the original clusters into a new set of meta-clusters.
- These meta-clusters represent the additive combinations of clusters generated on one or more different views.

Greene D. A Matrix Factorization Approach for Integrating Multiple Data Views. Mach Learn Knowl Discov Databases. 2009; 5781:423–38.

Unsupervised Learning: Clustering

Matrix Factorization for Multi-View Clustering

- We start by transposing all the membership matrices and stacking them vertically obtaining the matrix of clusters $\mathbf{X} \in \mathbb{R}^{l \times n}$ where l is the total number of clusters in C. We want to find the best approximation of $\mathbf{X}$ such that

$$\mathbf{X} \approx \mathbf{PH} \text{ and } \mathbf{P} \geq 0, \mathbf{H} \geq 0$$

Greene D. A Matrix Factorization Approach for Integrating Multiple Data Views. Mach Learn Knowl Discov Databases. 2009; 5781:423–38.

Unsupervised Learning: Clustering

Matrix Factorization for Multi-View Clustering

- The rows of $\mathbf{P} \in \mathrm{R}^{l \times k^r}$ project the clusters in a new set of k^t meta-clusters.
- The columns of $\mathbf{H} \in \mathrm{R}^{k^r \times n}$ can be viewed as the membership of the original objects in the new set of meta-clusters.

Greene D. A Matrix Factorization Approach for Integrating Multiple Data Views. Mach Learn Knowl Discov Databases. 2009; 5781:423–38.

Unsupervised Learning: Clustering

Matrix Factorization for Multi-View Clustering

- The approximation error is measured by the Frobenius norm

$$\|\mathbf{X} - \mathbf{PH}\|_F^2 = \sum_{i=1}^{l}\sum_{j=1}^{n}\left[\mathbf{X}_{ij} - (\mathbf{PH})_{ij}\right]^2$$

- to minimize the approximation error these multiplicative update rules are iteratively applied until a termination criteria is reached

$$P_{ic} \leftarrow P_{ic}\frac{(\mathbf{XH}^T)_{ic}}{(\mathbf{PHH}^T)_{ic}} \qquad H_{cj} \leftarrow H_{cj}\frac{(\mathbf{P}^T\mathbf{X})_{cj}}{(\mathbf{P}^T\mathbf{PH})_{cj}}$$

- each iteration has a computational cost of O(lnk') when multiplying dense matrices.

Greene D. A Matrix Factorization Approach for Integrating Multiple Data Views. Mach Learn Knowl Discov Databases. 2009; 5781:423–38.

Unsupervised Learning: Clustering

Matrix Factorization for Multi-View Clustering

- Based on the values in the projection matrix $\mathbf{P}$, we can calculate a matrix $\mathbf{T} \in \mathrm{R}^{v \times k^r}$.
- T_{hf} indicates the contribution of the view V_h to the f-th meta-cluster, calculated as

$$T_{hf} = \frac{\sum_{c_h^j \in C_h} P_{jf}}{\sum_{g=1}^{l} P_{gf}}$$

- If T_{hf} is close to 0, the contribute of view V_h to the f-th meta-cluster is poor
- If T_{hf} is close to 1, the contribute of view V_h to the f-th meta-cluster is strong

Greene D. A Matrix Factorization Approach for Integrating Multiple Data Views. Mach Learn Knowl Discov Databases. 2009; 5781:423–38.

Unsupervised Learning: Clustering

Matrix Factorization for Multi-View Clustering: Initialization

- Since IMF is based on an iterative algorithm the choice of a good starting point is important.
  It can be used a stochastic initialization, but the resulting clustering will probably vary with different starting factors. A good method is the deterministic NNDSVD (non-negative double SVD) that produces a pair of matrices suitable as a starting point.

Greene D. A Matrix Factorization Approach for Integrating Multiple Data Views. Mach Learn Knowl Discov Databases. 2009; 5781:423–38.

Unsupervised Learning: Clustering

Matrix Factorization for Multi-View Clustering: Model Selection

- We need to find a suitable value for k^t. If it is too low the integration process will merge unrelated clusters, if it is too high it will fail merge related clusters.
- To identify an appropriate value for k^t we will search into some range $[k_{min}, k_{max}]$ determined by the knowledge of the domain.
- For each candidate k^t we consider the uncertainty of the mapping between clusters based on the uncertainty of the values of matrix $\mathbf{P}$.

Greene D. A Matrix Factorization Approach for Integrating Multiple Data Views. Mach Learn Knowl Discov Databases. 2009; 5781:423–38.

Unsupervised Learning: Clustering

Matrix Factorization for Multi-View Clustering: Model Selection

- We start by normalizing the rows of the projection matrix yelding $\hat{\mathbf{P}}$.
  Ideally $\hat{\mathbf{P}}$ would contain a single value 1 and $(k' - 1)$ zeros, i.e. each cluster is perfectly matched with a meta-cluster.
  Formally we will evaluate the normalized entropy of the rows of $\hat{\mathbf{P}}$ for a given value k'.

Greene D. A Matrix Factorization Approach for Integrating Multiple Data Views. Mach Learn Knowl Discov Databases. 2009; 5781:423–38.

Unsupervised Learning: Clustering

Matrix Factorization for Multi-View Clustering: Model Selection

- For each row j in $\hat{\mathbf{P}}$ the normalized entropy of the projection values is

$$e\left(\hat{\mathbf{P}}_j\right) = -\frac{1}{\log k'} \sum_{h=1}^{k'} P_{jh} \log\left(P_{jh}\right)$$

The suitability criterion for a particular value of k' is given by

$$s\left(k'\right) = 1 - \frac{1}{l} \sum_{j=1}^{l} e\left(\hat{\mathbf{P}}_j\right)$$

and we are looking for the value k' that maximize the criterion.

Greene D. A Matrix Factorization Approach for Integrating Multiple Data Views. Mach Learn Knowl Discov Databases. 2009; 5781:423–38.

Unsupervised Learning: Clustering

Matrix Factorization for Multi-View Clustering: Evaluation

- IMF has been evaluated on both synthetic and real-world datasets.
  In both settings IMF produced more informative clusterings with respect to the single view clustering counterparts.
- It turned out that IMF can effectively take advantage of cases when a variety of different clusterings are available for each view and in fact out-performed popular ensemble clustering algorithms.

Greene D. A Matrix Factorization Approach for Integrating Multiple Data Views. Mach Learn Knowl Discov Databases. 2009; 5781:423–38.

Unsupervised Learning: Projective Methods

- Projective methods are based on the concept of embedding the patterns into a new feature space learned by optimizing a criteria such as minimum reconstruction error from principal component analysis.
- Recently, this methodology has been applied in the context of multi-view data.
- For example Tyagi et al. proposed an intermediate integration approach for soft-hard clustering.

G.Tyagi,N.Patel,andI.Sethi,"Soft-hardclusteringformultiviewdata," in Information Reuse and Integration (IRI), 2015 IEEE International Conference on. IEEE,

Unsupervised Learning: Projective Methods

- The method consists in mapping all the objects from the different views into a unit hypercube.
- The projected views were concatenated and then clustered with k-means.
- They tested the method on three different benchmark data sets: the first contains acoustic and seismic sensors for different type of vehicles, the second is the Handwritten Numeral dataset and the third is a multi-view image dataset.
- The results were evaluated by using three performance measures: Clustering accuracy, Normalized Mutual Information (NMI) and Clustering purity.
- They demonstrated that their methods have good performances and are not too sensitive to input parameters.

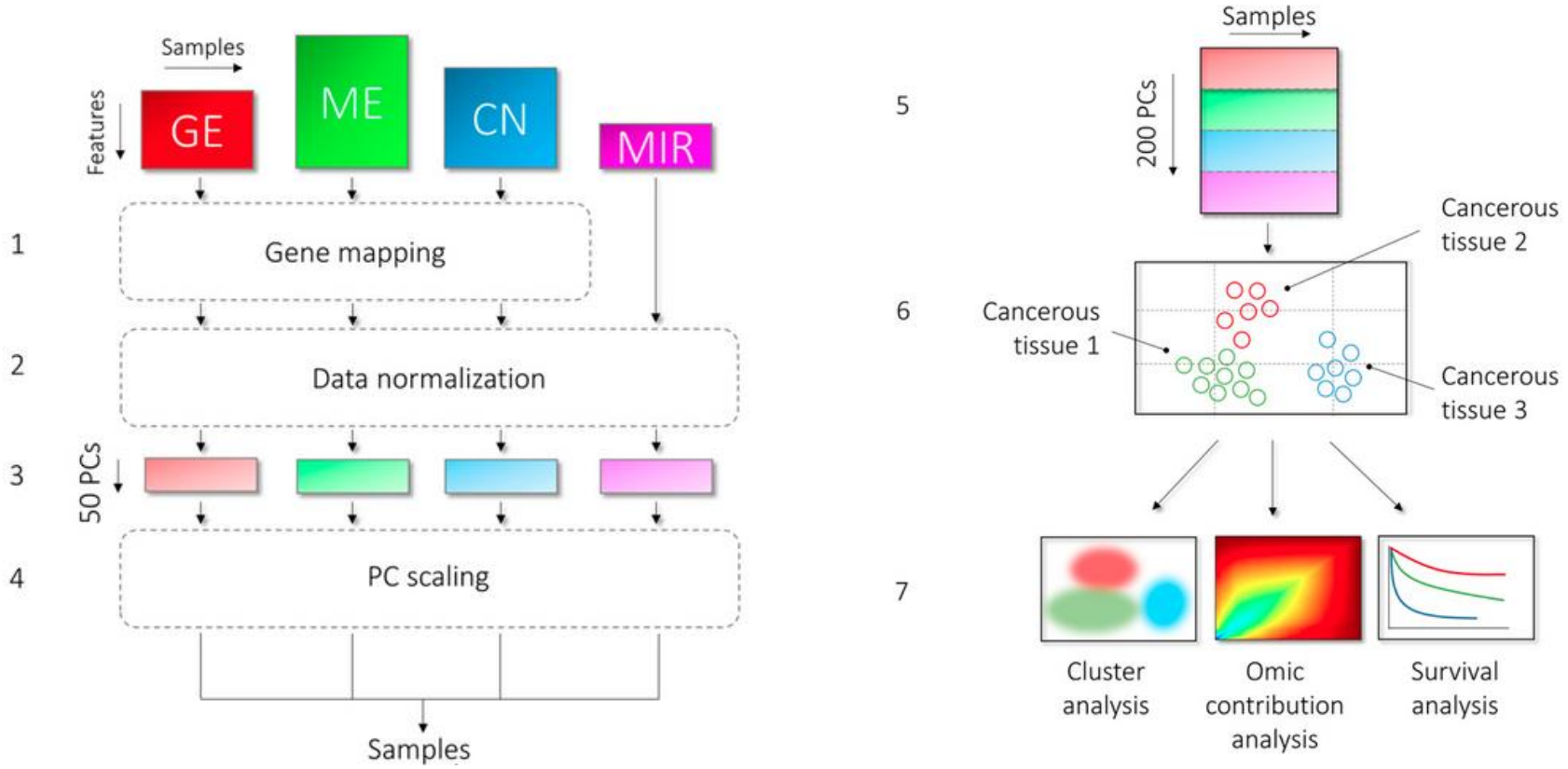
G.Tyagi,N.Patel,andI.Sethi,"Soft-hardclusteringformultiviewdata," in Information Reuse and Integration (IRI), 2015 IEEE International Conference on. IEEE,

Multi-View Clustering on TCGA Dataset

- Taskesen et al. proposed a multi-omic integration approach (MEREDITH) that exploits the joint behaviour of the different molecular characteristics
- It supports visual exploration of the data by a two-dimensional landscape
- It is useful for inspect of the contribution of the different genome-wide data-types.
- Experiments were performed among 4,434 patients taken from The Cancer Genome Atlas (TCGA) across 19 cancer-types based on genome-wide measurements of four different molecular characteristics:
  - gene expression (GE; 18,882 features),
  - DNA-methylation (ME; 11,429 features),
  - copy-number variation (CN; 23,638 features)
  - microRNA expression (MIR; 467 features).

Taskesen, Erdogan, et al. "Pan-cancer subtyping in a 2D-map shows substructures that are driven by specific combinations of molecular characteristics." Scientific Reports 6 (2016).

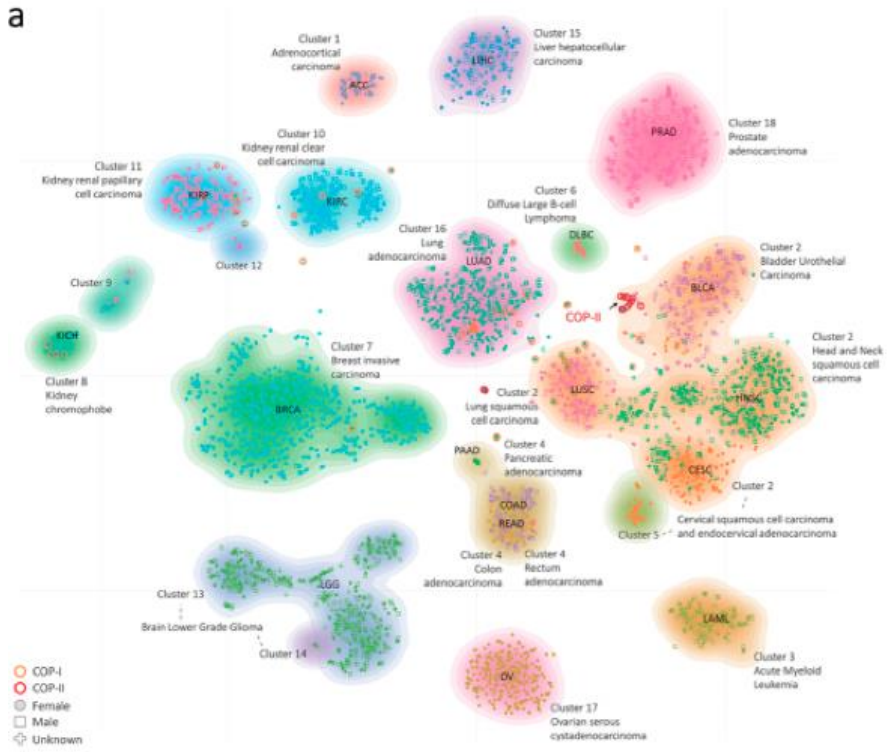
Multi-View Clustering on TCGA Dataset



Taskesen, Erdogan, et al. "Pan-cancer subtyping in a 2D-map shows substructures that are driven by specific combinations of molecular characteristics." Scientific Reports 6 (2016).

Multi-View Clustering on TCGA Dataset

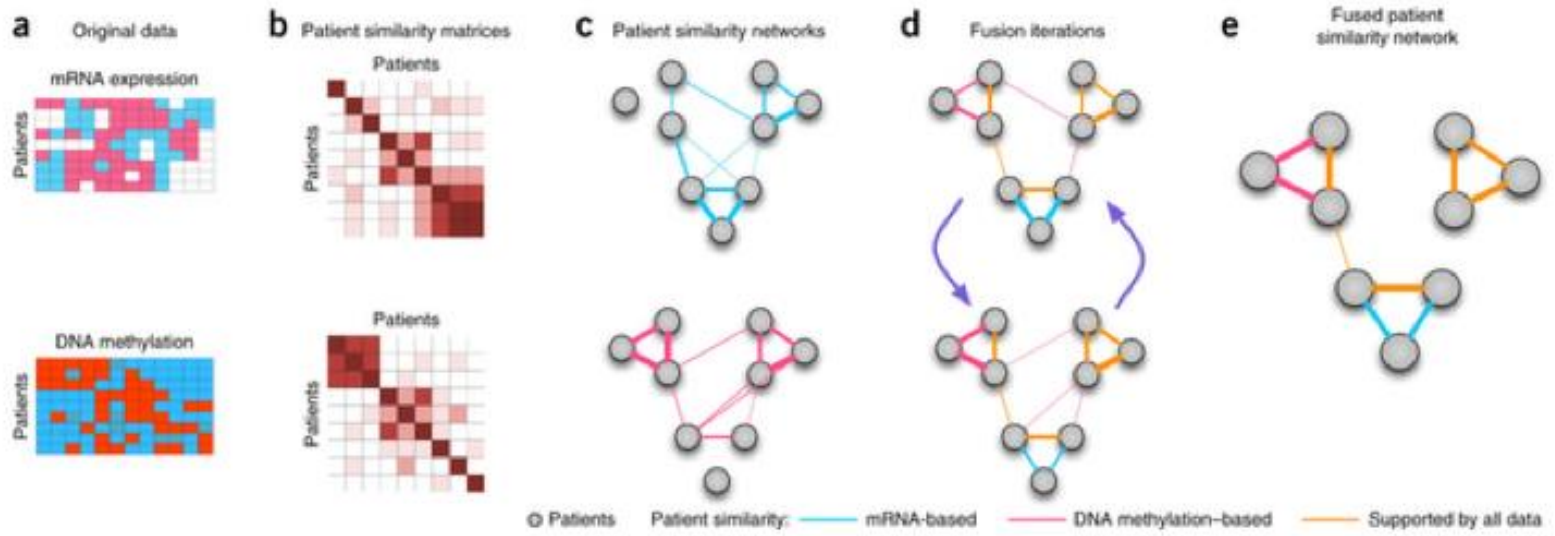
- Patient-sample projection in a two-dimensional map illustrating the cancer-landscape.
- The clustering is based on DBSCAN with the Davies-Bouldin index score for selecting the number of clusters



Taskesen, Erdogan, et al. "Pan-cancer subtyping in a 2D-map shows substructures that are driven by specific combinations of molecular characteristics." Scientific Reports 6 (2016).

Graph Integration: Similarity Network Fusion

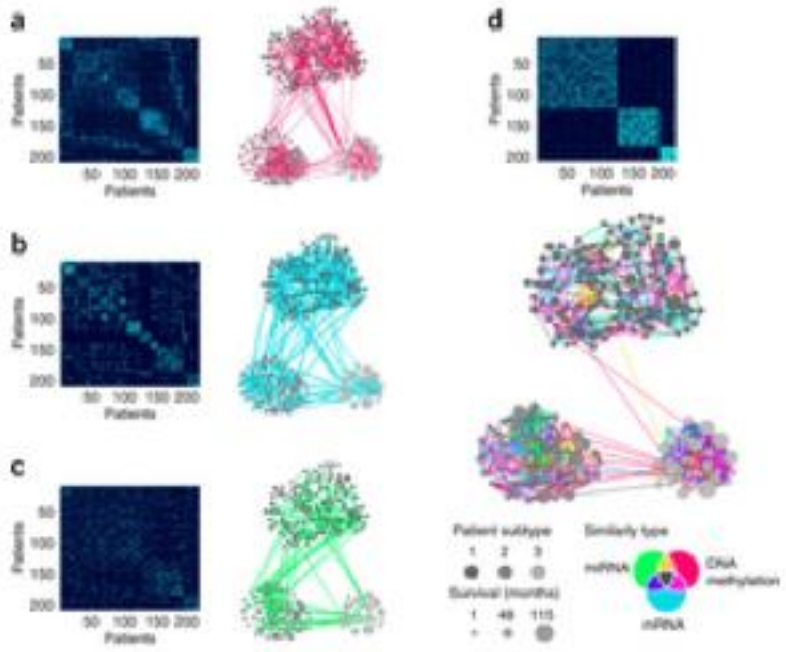
- Wang et al. proposed an **intermediate** integration **network fusion** methodology in order to integrate multiple genomic data and clustering patients.



Wang, Bo, et al. "Similarity network fusion for aggregating data types on a genomic scale." Nature methods 11.3 (2014): 333-337.

Graph Integration: Similarity Network Fusion

- They first constructed a patients similarity network for each view.
- Then, they iteratively updated the network with the information coming from other networks in order to make them more similar at each step.
- At the end, this iterative process converged to a final fused network.



Wang, Bo, et al. "Similarity network fusion for aggregating data types on a genomic scale." Nature methods 11.3 (2014): 333-337.

Graph Integration: Similarity Network Fusion

- The authors tested the method to combine mRNA expression, microRNA expression and DNA methylation from five cancer data sets.
- They showed that the similarity networks of each view have different characteristics related to patients similarity while the fused network gives a more clear picture of the patients clusters.
- They compared the proposed methodology with iClust and the clustering on concatenated views.
- Results were evaluated with the silhouette score for clustering coherence, Cox log-rank test p-value for survival analysis for each subtype and the running time of the algorithms.
- SNF outperformed single view data analysis and they were able to identify cancer subtypes validated by survival analysis.

Wang, Bo, et al. "Similarity network fusion for aggregating data types on a genomic scale." Nature methods 11.3 (2014): 333-337.

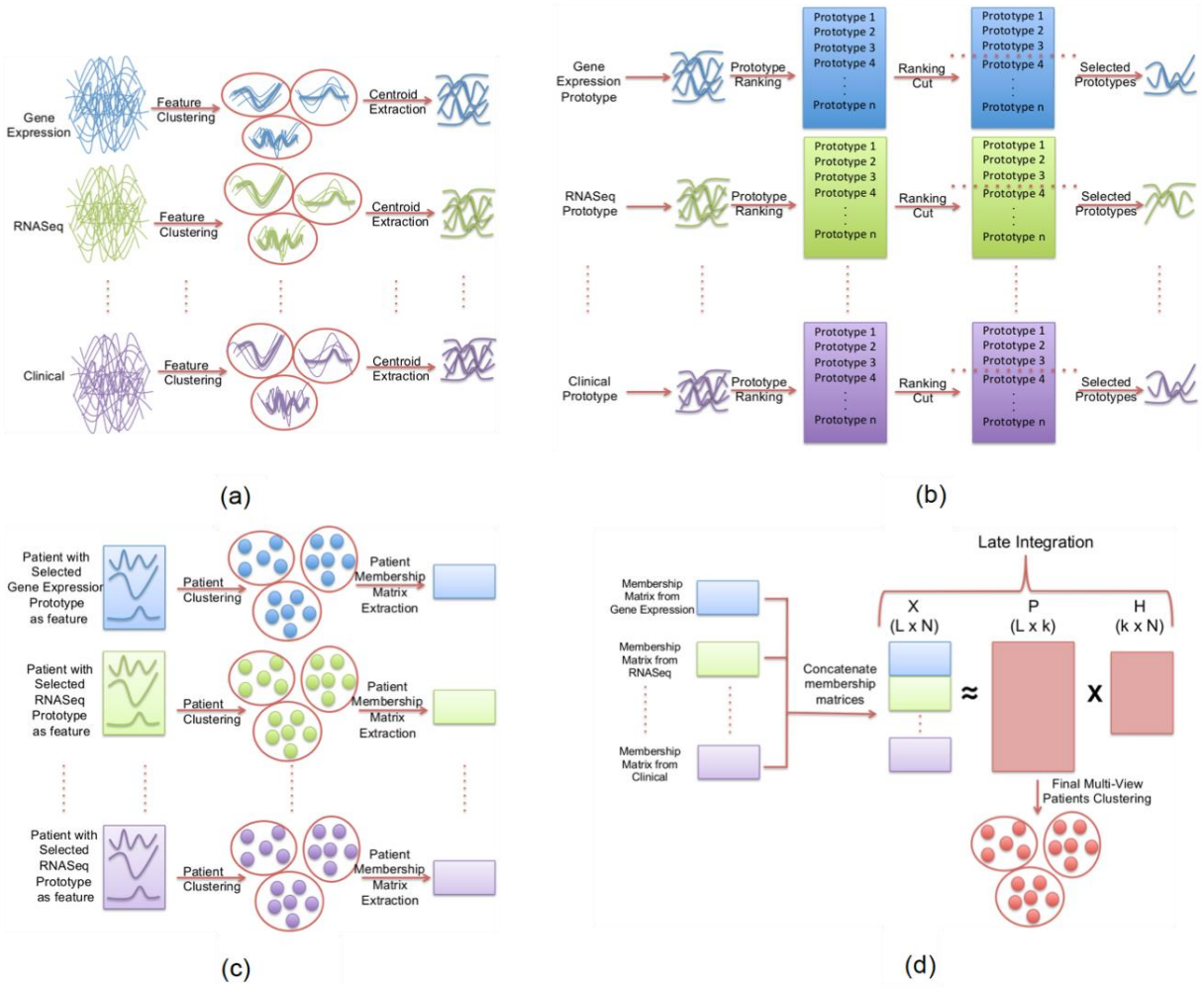
MVDA: A Multi-view genomic data integration methodology

- We propose a multi-view approach in which the information from different data layers is integrated at the levels of the results of each single view clustering iterations by means of a matrix factorization approach.
- We performed experiment on six genomic datasets spanning on seven different views.

Dataset	Response	N(0)	N(1)	N(2)	N(3)	Gene Expression	RNASeq	microRNA Expression	miRNASeq	Protein Expression	Copy Number	Clinical Data
Breast Cancer patient from The Cancer genome Atlas (TCGA), N = 151												
TCGA.BRC	Pam50 (Her2,Basal,LumA,LumB)	24	13	55	59		x		x			
Breast Cancer patient from The Gene Expression Omnibus (GEO), N = 201												
OXF.BRC.1	Pam50 (Her2,Basal,LumA,LumB)	26	6	117	52	x		x				
OXF.BRC.2	Clinical (Level1, Level2, Level3, Level4)	73	54	42	32	x		x				
Prostate cancer patient from Memorial Sloan-Kettering Cancer Center (MSKCC), N=88												
MSKCC.PRCA	Tumor stages T1 vs. T2, T3, T4	53	35			x		x			x	x
Ovarian cancer patient from The Cancer Genome Atlas (TCGA), N=93												
TCGA.OVG	Venus invasion present vs. absent	40	53			x		x		x		
Glioblastoma Multiforme patient from The Cancer genome Atlas (TCGA), N = 167												
TCGA.GBM	(Classical, Mesechymal, Neural, Proneural)	37	54	24	52	x		x				

Serra, Angela, et al. "MVDA: a multi-view genomic data integration methodology." BMC bioinformatics 16.1 (2015): 1.

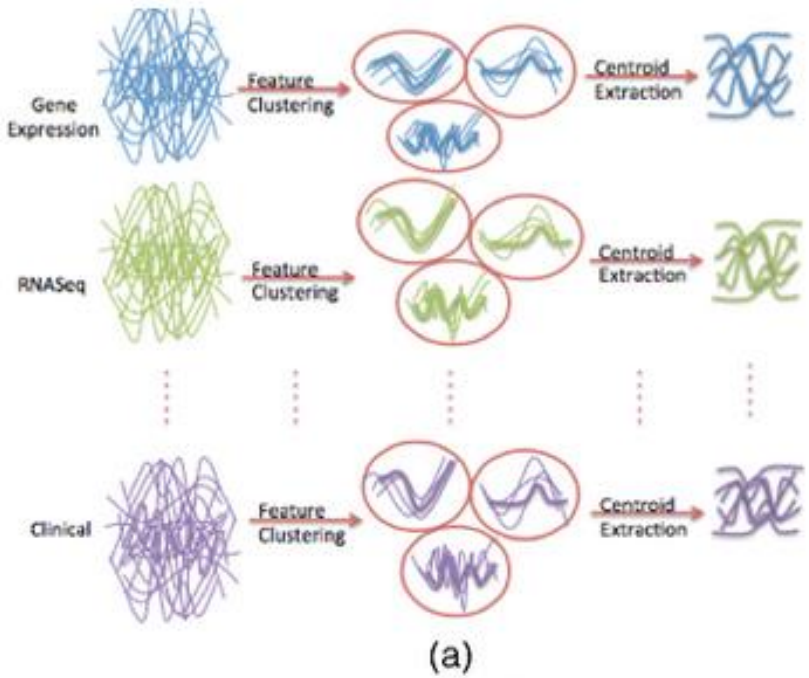
MVDA: A Multi-view genomic data integration methodology



Serra, Angela, et al. "MVDA: a multi-view genomic data integration methodology." BMC bioinformatics 16.1 (2015): 1.

MVDA: A Multi-view genomic data integration methodology

- Goal: input dimension reduction and relevant patterns discover.
- We tried different kinds of clustering algorithms using the Pearson coefficient as metric.
  - Pvclust
  - SOM
  - Hierarchical (Ward)
  - Pam
  - Kmeans



Serra, Angela, et al. "MVDA: a multi-view genomic data integration methodology." BMC bioinformatics 16.1 (2015): 1.

MVDA: A Multi-view genomic data integration methodology

- Clustering of genes

	Number of clusters	Min	Max	Mean	Number of Singleton	Correlation	Connectivity	Dunn	Entropy
Pvclust	388	1	118	8.13	6	0.54	5804.46	0.02	5.67871368848465
SOM	300	2	560	10.52	0	0.48	7759.67	0	3.84247601565554
Pam	240	1	125	13.15	8	0.53	6361.09	0.03	5.12678431184979
Kmeans	300	1	325	10.52	92	0.72	7062.1	0.03	4.52143446800201
Hierarchical (Ward)	300	3	165	10.52	0	0.46	5746.12	0.11	5.40406208883471

- Normalized Mutual Information Value

	Pvclust	SOM	Pam	Kmeans	Hierarchical (Ward)
Pvclust	1.00	0.45	0.71	0.60	0.63
SOM	0.45	1.00	0.41	0.47	0.38
Pam	0.71	0.41	1.00	0.59	0.61
Kmeans	0.60	0.47	0.59	1.00	0.59
Hierarchical (Ward)	0.63	0.38	0.61	0.59	1.00

Serra, Angela, et al. "MVDA: a multi-view genomic data integration methodology." BMC bioinformatics 16.1 (2015): 1.

MVDA: A Multi-view genomic data integration methodology

- Clustering of miRNA

	Number of cluster	Min	Max	Mean	Number of Singleton	Correlation	Connectivity	Dunn	Entropy
Pvclust	31	1	20	5.9	2	0.69	186.4	0.05	3.2247841438082
SOM	28	2	70	6.54	0	0.56	340.05	0.01	2.64735154070016
Pam	21	3	28	8.71	1	0.7	144.97	0.07	2.77097117056627
Kmeans	28	1	22	6.54	9	0.81	204.45	0.03	2.85281018957383
Hierarchical (Ward)	24	3	27	7.62	0	0.67	161.7	0.1	2.95147237386223

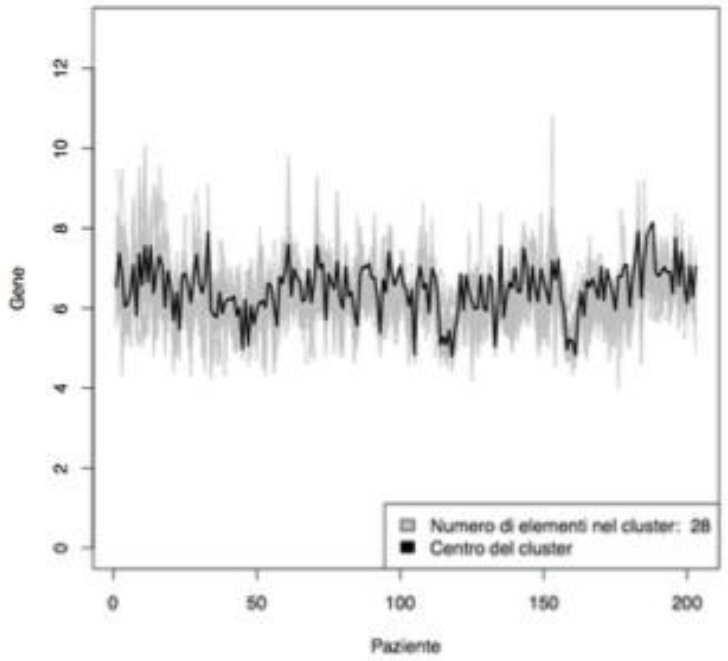
- Normalized Mutual Information Value

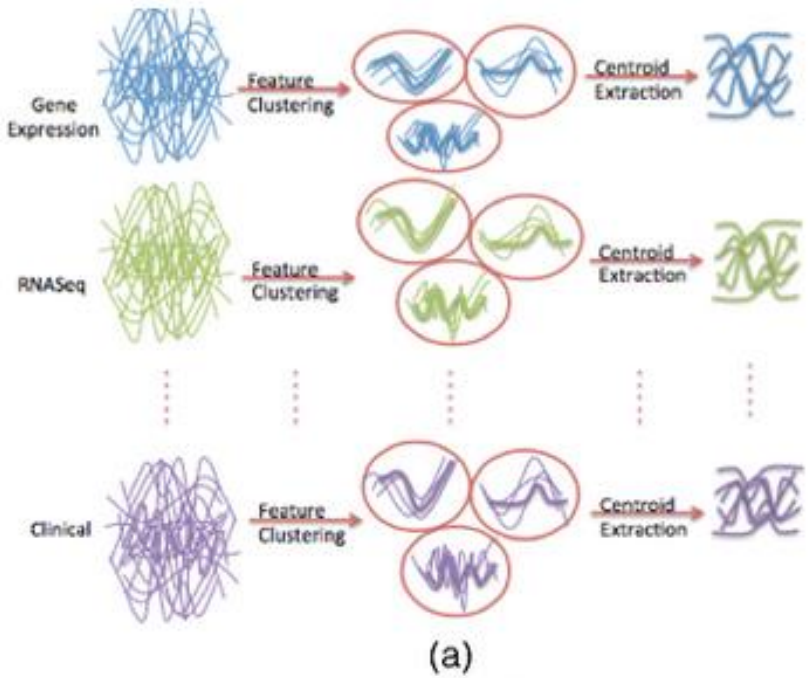
	Pvclust	SOM	Pam	Kmeans	Hierarchical (Ward)
Pvclust	1.00	0.42	0.73	0.51	0.60
SOM	0.42	1.00	0.36	0.40	0.42
Pam	0.73	0.36	1.00	0.47	0.59
Kmeans	0.51	0.40	0.47	1.00	0.61
Hierarchical (Ward)	0.60	0.42	0.59	0.61	1.00

Serra, Angela, et al. "MVDA: a multi-view genomic data integration methodology." BMC bioinformatics 16.1 (2015): 1.

MVDA: A Multi-view genomic data integration methodology

- For each cluster a prototype element has been extracted

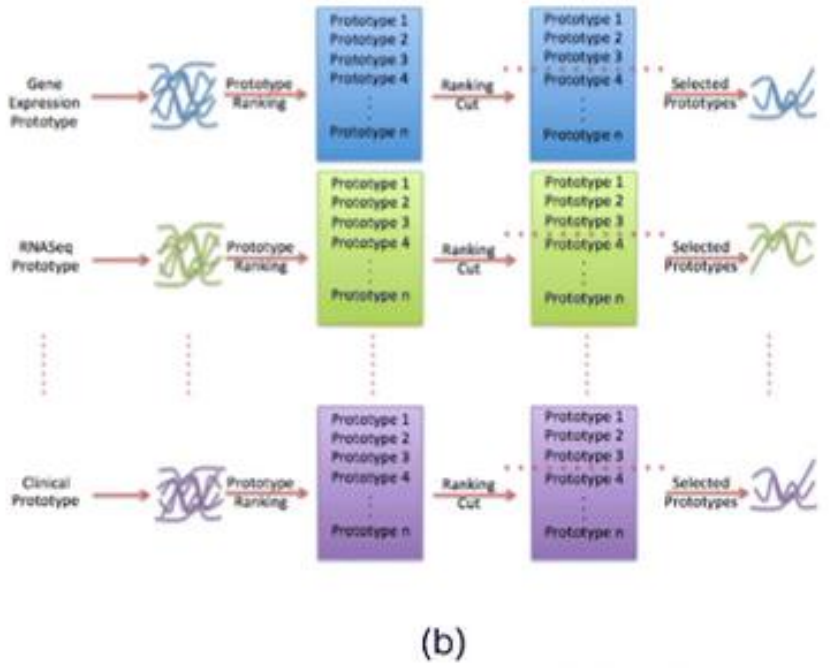




Serra, Angela, et al. "MVDA: a multi-view genomic data integration methodology." BMC bioinformatics 16.1 (2015): 1.

MVDA: A Multi-view genomic data integration methodology

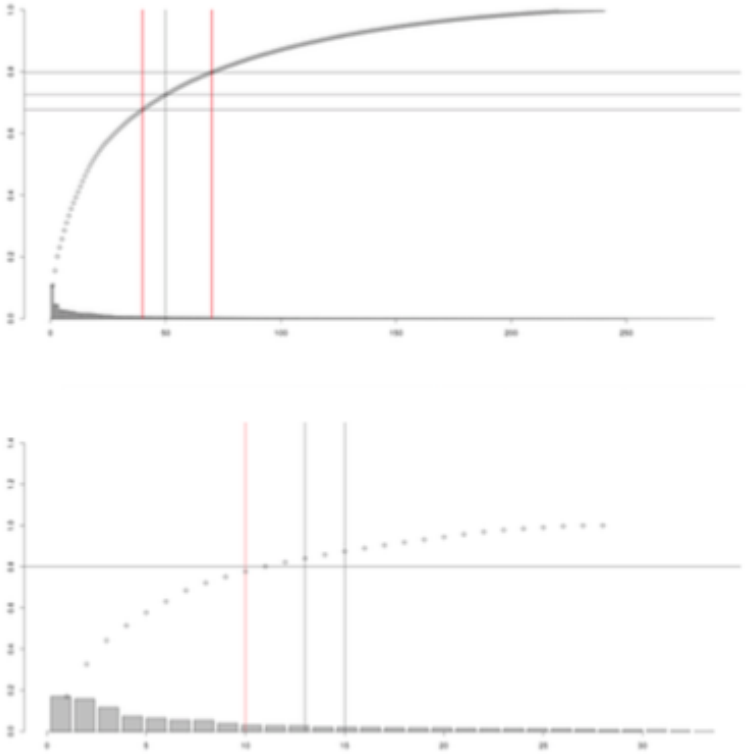
- By selecting prototypes obtained at the previous step we find the most relevant features when working in the patients' space.
- Feature selection is performed:
  - By computing the CAT t score.
    - The correlation-adjusted t-score (cat score) is a modification of the Student t-statistic to account for dependencies among variables
    - Zuber and Strimmer have shown that the cat score improves ranking of genes to detect differential expression in the presence of correlation.
  - By computing the mean decrease accuracy index of the random forest classifier

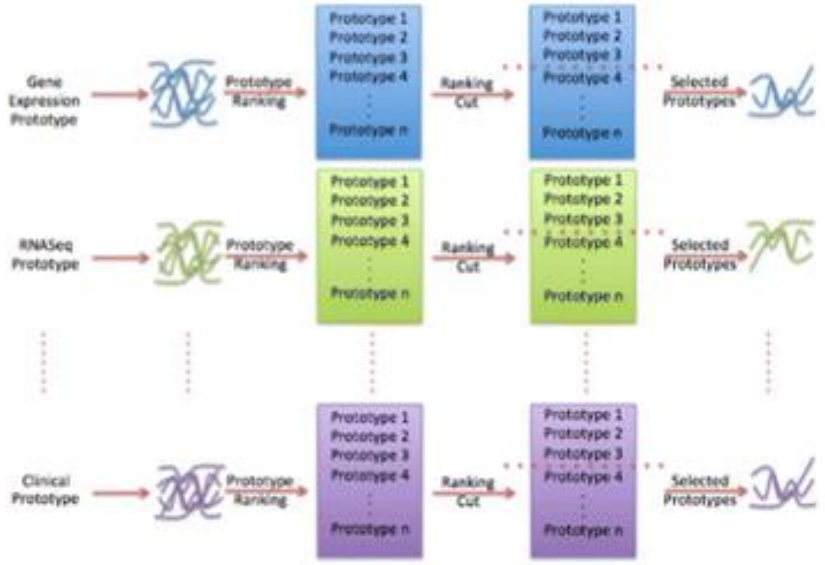


Serra, Angela, et al. "MVDA: a multi-view genomic data integration methodology." BMC bioinformatics 16.1 (2015): 1.

MVDA: A Multi-view genomic data integration methodology

- We select the top % relevant element for each view



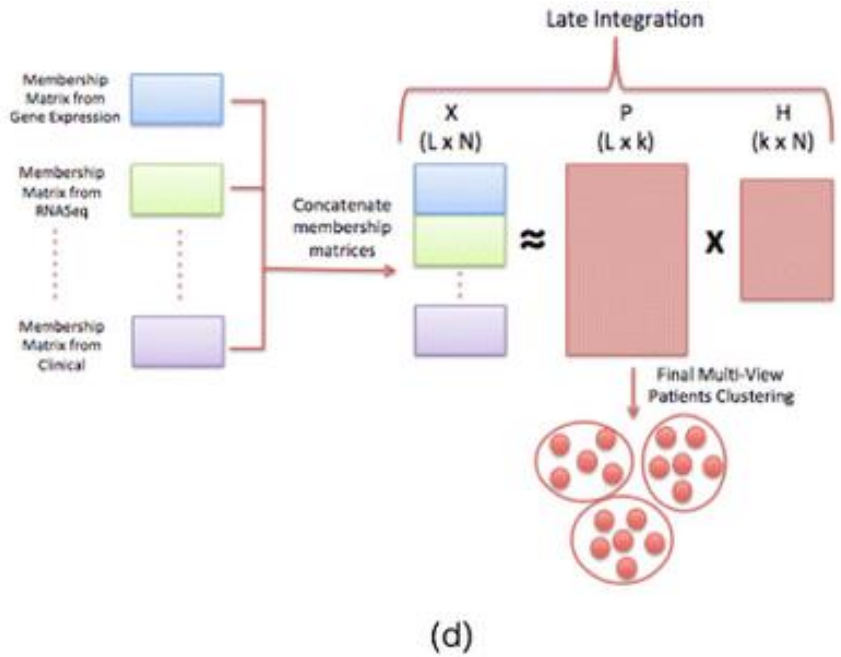


(b)

Serra, Angela, et al. "MVDA: a multi-view genomic data integration methodology." BMC bioinformatics 16.1 (2015): 1.

MVDA: A Multi-view genomic data integration methodology

- The goal was to integrate the single view results in order to find patient clusters.
- We used a late integration approach.
- On each view we executed the same clustering algorithms of the first step to cluster patients
- The algorithm used for multi-view data integration performed an iterative matrix factorization method



Serra, Angela, et al. "MVDA: a multi-view genomic data integration methodology." BMC bioinformatics 16.1 (2015): 1.

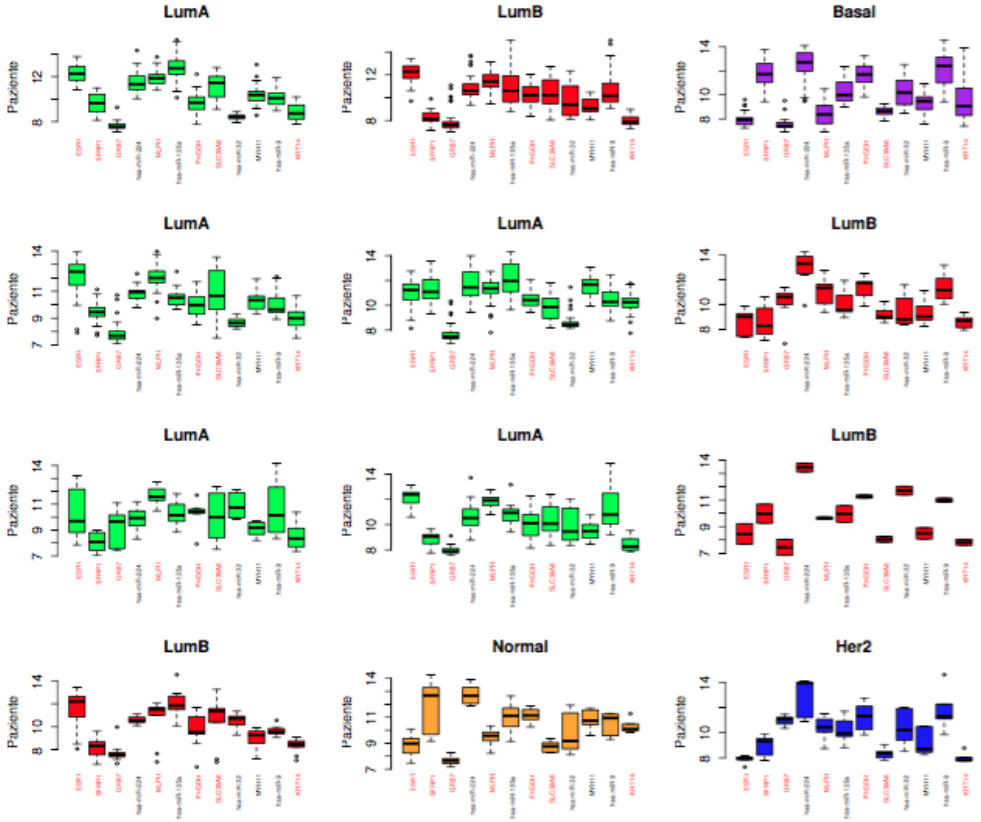
MVDA: A Multi-view genomic data integration methodology

	Feature	Integration	Algorithm	Error	NMI	Stability
Single View	All Feature	-	Ward	30,08%	26%	86%
		-	Kmeans	30,93%	25%	51%
		-	Pamk	30,75%	24%	94%
	Selected Prototype	-	Ward	30,72%	26%	89%
		-	Kmeans	30,36%	25%	52%
		-	Pamk	30,78%	24%	96%
Multi-View	All Feature	Early Integration	Tw-kmeans	37,10%	24%	69%
	All Feature	Intermediate Integration	SNF	30,83%	22%	83%
	All Feature in Cluster of Selected Prototype	Intermediate Integration	SNF	29,81%	20%	82%
	Selected Prototype	Late Integration unsupervised	MF/GLI	**26,83%**	**30%**	**96%**
	Selected Prototype	Late Integration semi-supervised	MF/GLI	**2,35%**	**60%**	**95%**

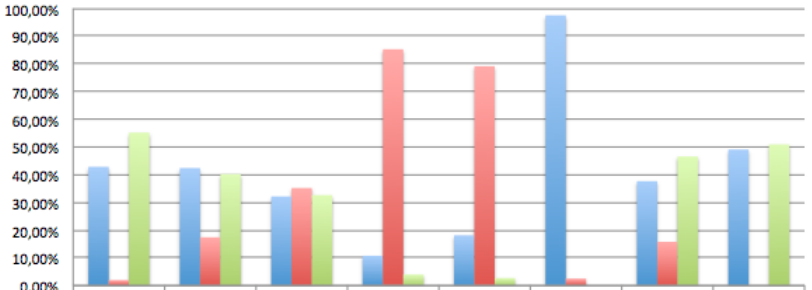
- We performed four kinds of experiments
  - One completely unsupervised with all the features.
  - One semi-supervised with all the features.
  - One completely unsupervised with the most relevant features.
  - One semi-supervised with the most significant features.
- The best result was obtained in the last case.

Serra, Angela, et al. "MVDA: a multi-view genomic data integration methodology." BMC bioinformatics 16.1 (2015): 1.

MVDA: A Multi-view genomic data integration methodology



OXF.BRC.1 Multi-View Clusters Statistics



	Cluster1 - LumA - Pvalue: 5,96E-07	Cluster2 - LumB - Pvalue: 4,47E-13	Cluster3 - Basal - Pvalue: 1,86E-17	Cluster5 - LumA - Pvalue: 8,37E-05	Cluster6 - LumA - Pvalue: 2,56E-05	Cluster7 - LumA - Pvalue: 2,65E-05	Cluster8 - Her2 - Pvalue: 9,09E-09	Cluster9 - LumB - Pvalue: 0,005
Gene	42,85%	42,39%	32,15%	10,76%	18,12%	97,47%	37,66%	49,07%
miRNA	1,97%	17,43%	35,20%	85,25%	79,13%	2,53%	15,78%	0,00%
Info	55,18%	40,19%	32,65%	4,00%	2,74%	0,00%	46,56%	50,93%

Serra, Angela, et al. "MVDA: a multi-view genomic data integration methodology." BMC bioinformatics 16.1 (2015): 1.

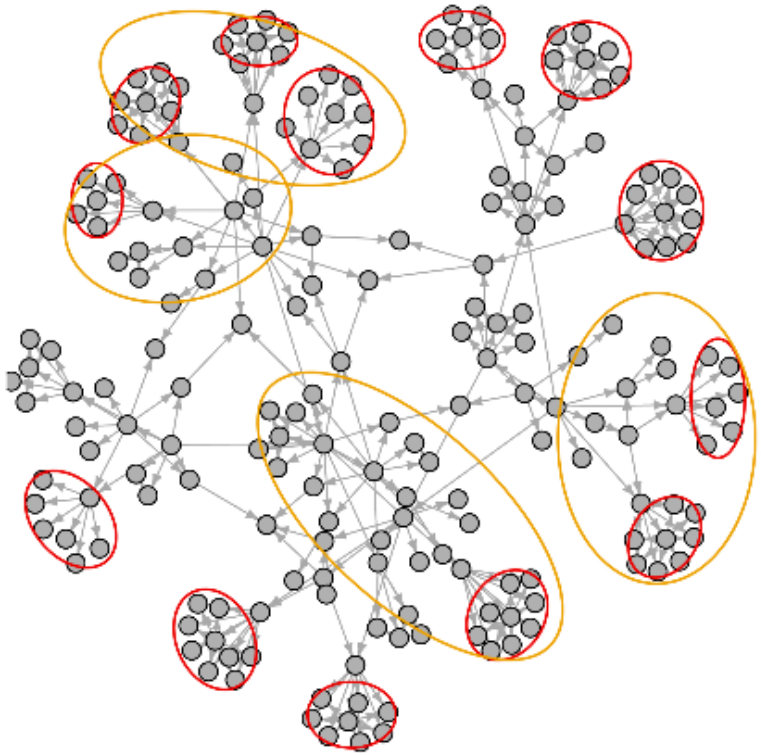
A multi-view genomic data simulator

- Integrative analysis has proven effective in terms of significance and stability
- New algorithms need to be benchmarked with annotated datasets which are expensive to produce and not under full control
- An alternative is to generate plausible synthetic datasets
- We propose a model for the simulation of multi-modal biological data modelled with regulatory networks and ordinary differential equations for the benchmark of data integration procedures

Fratello, Michele, et al. "A multi-view genomic data simulator." BMC bioinformatics 16.1 (2015): 1.

A multi-view genomic data simulator

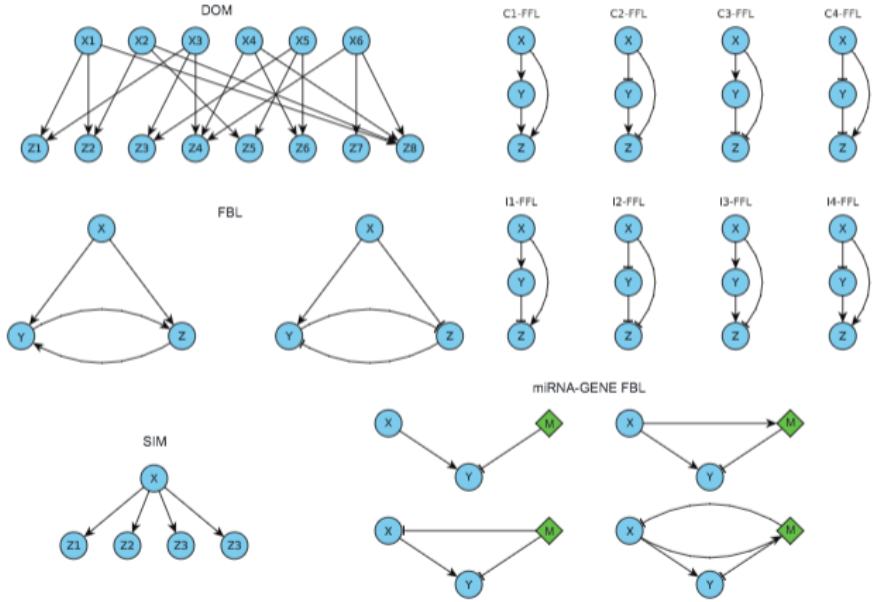
- Analysis performed on different organisms report common characteristics of TRNs:
  - Hierarchical architecture: A restricted set of genes can control whole biological processes. These genes have a higher-than-average number of connections
  - Modularity: At the local scale genes work in small modules tightly connected



Fratello, Michele, et al. "A multi-view genomic data simulator." BMC bioinformatics 16.1 (2015): 1.

A multi-view genomic data simulator

1. A pool of random motifs is constructed at each iteration
2. The utility of adding each motif to the network is estimated by a score
3. The motif to be added is sampled from a distribution proportional to the scores
4. A subset of nodes of the current network is sampled
5. The motif is used as a template for connecting them

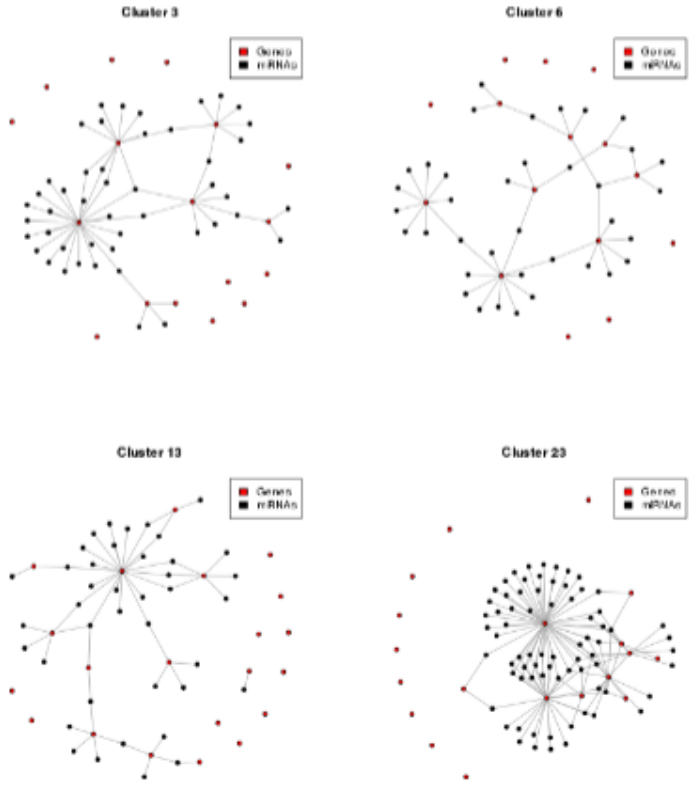


Fratello, Michele, et al. "A multi-view genomic data simulator." BMC bioinformatics 16.1 (2015): 1.

A multi-view genomic data simulator

We report three cases of analysis that can be performed on the generated datasets

- Reverse engineering of simulated networks
  - PANDA
  - ARACNE
- Gene Clustering
- Feature relevance determination
  - t-test
  - Random Forests



Fratello, Michele, et al. "A multi-view genomic data simulator." BMC bioinformatics 16.1 (2015): 1.

Semi-supervised Subgroup discovery in ALS

- Standard analysis aim at finding significant differences among groups defined *a priori* based on clinical and expert knowledge.

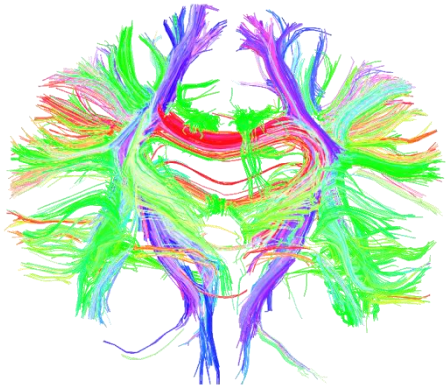
- We, instead, propose an approach in which we let the data group by themselves and then characterize *a posteriori* significant differences emerged by this grouping with clinical information.

Fratello, Michele, et al. Submitted for publication

Semi-supervised Subgroup discovery in ALS

- We consider each subject as an object represented in two different spaces, providing different kinds of information.

- The features (or characteristics) of these spaces are the voxels of the rsfMRI and DTI data respectively.

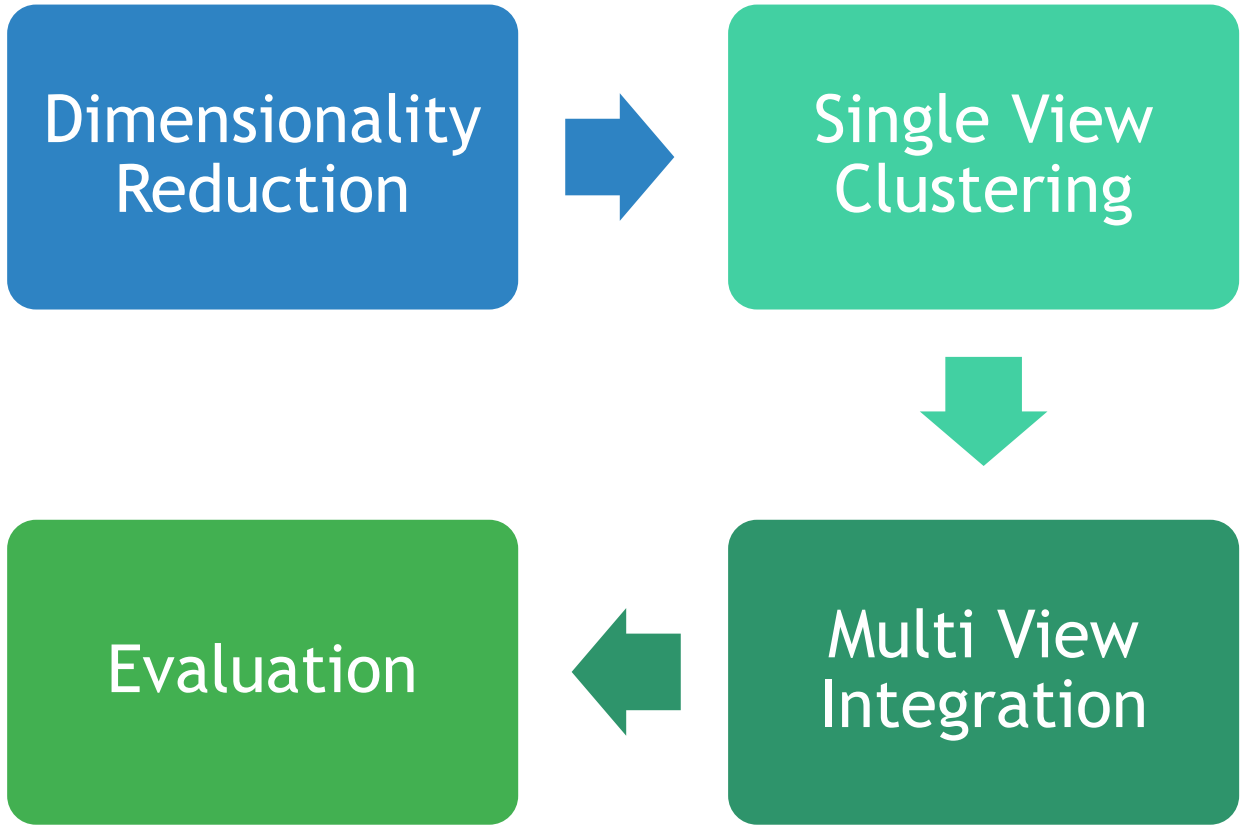


DTI

fMRI

Fratello, Michele, et al. Submitted for publication

Semi-supervised Subgroup discovery in ALS



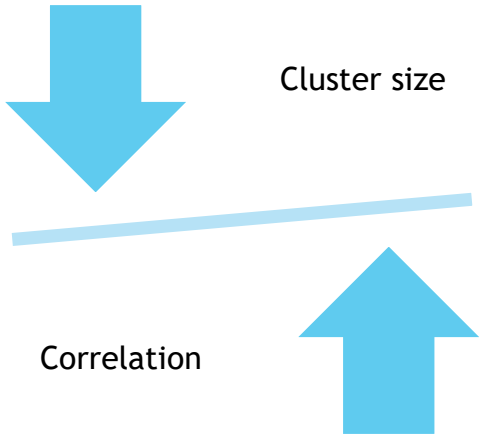
Fratello, Michele, et al. Submitted for publication

Semi-supervised Subgroup discovery in ALS

Dimensionality Reduction

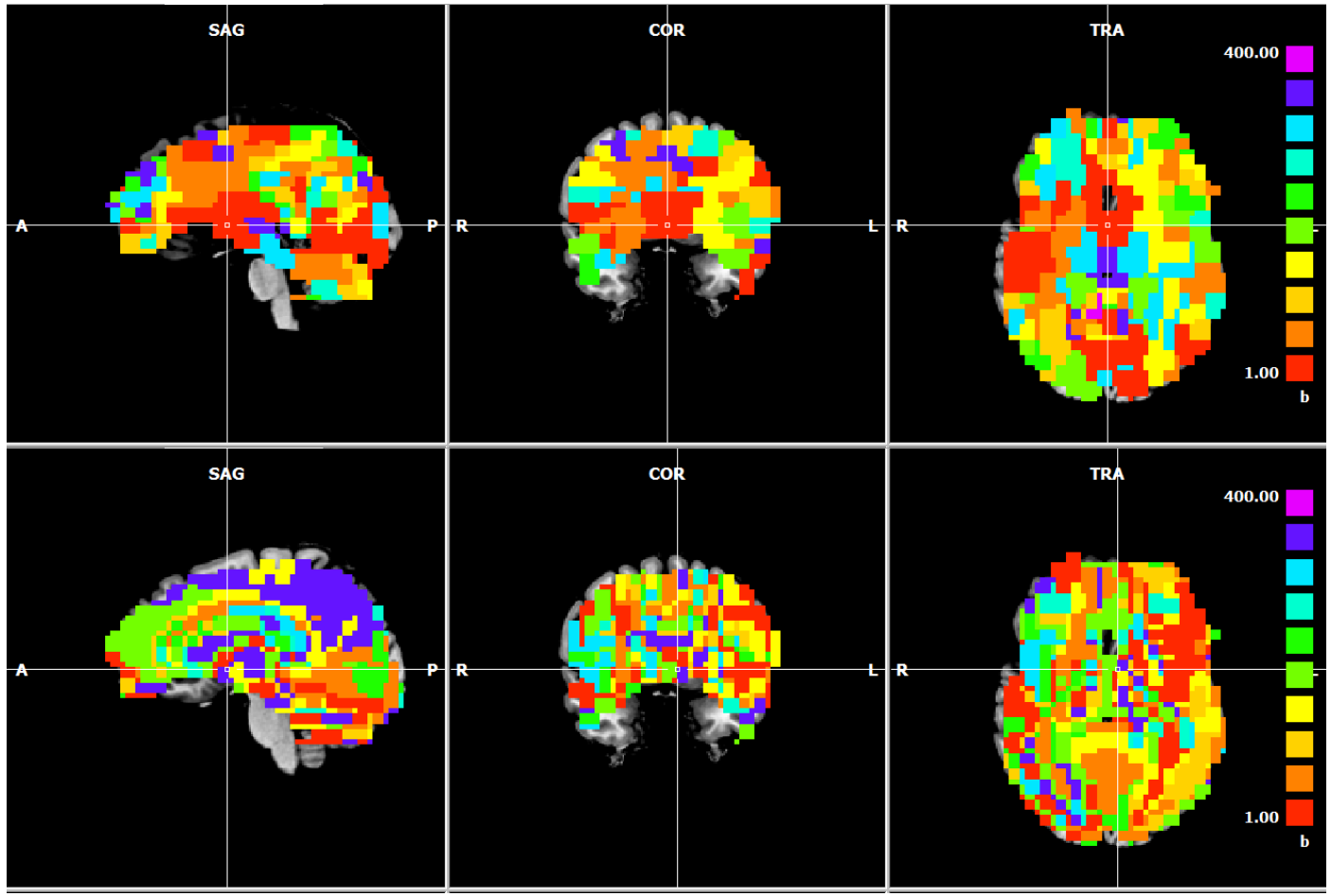
- To overcome the issues deriving from HDLSS data we reduced the size of each dataset.
- Adjacent voxels are then aggregated with clustering. Each resulting area is then represented by a single value, derived by the clustered voxels.
- Voxel clustering can be seen as a data-driven parcelation.

How many clusters?



Fratello, Michele, et al. Submitted for publication

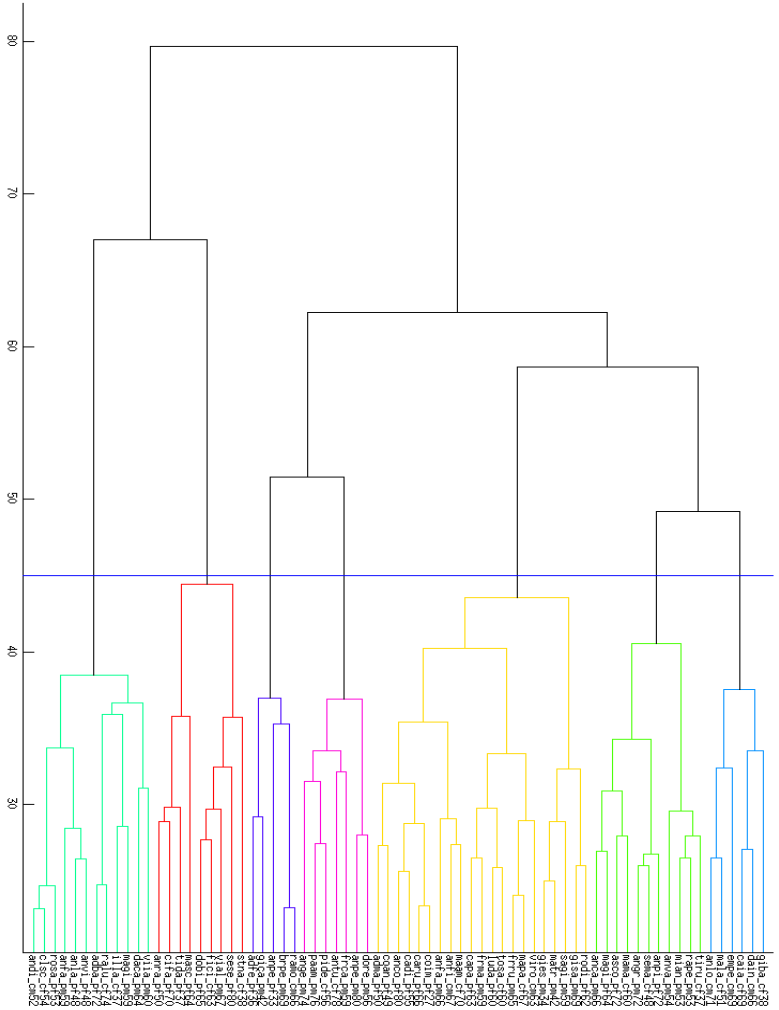
Semi-supervised Subgroup discovery in ALS



Clustered Voxels
Top: rsfMRI-
Bottom: DTI

Fratello, Michele, et al. Submitted for publication

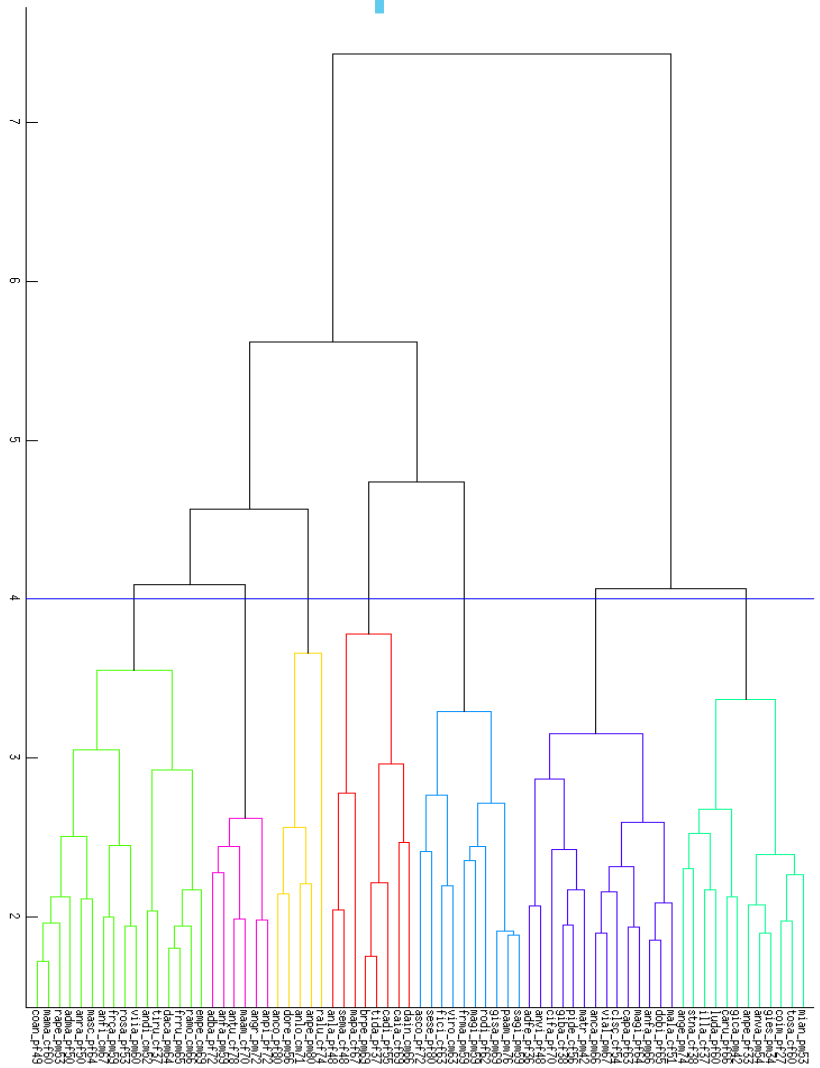
Semi-supervised Subgroup discovery in ALS



- We performed single view clustering of subjects on the reduced datasets
- The number of clusters was empirically fixed to 7

Fratello, Michele, et al. Submitted for publication

Semi-supervised Subgroup discovery in ALS



- We performed single view clustering of subjects on the reduced datasets
- The number of clusters was empirically fixed to 7

Fratello, Michele, et al. Submitted for publication

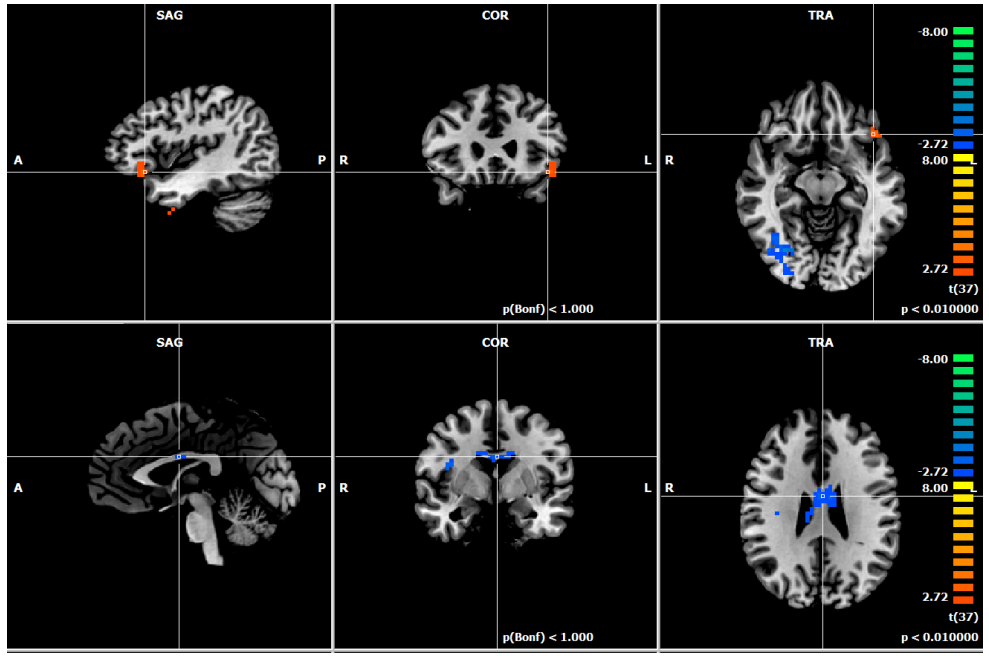
Semi-supervised Subgroup discovery in ALS

- Single View clusterings are integrated together with side information on patient class labels, into 6 clusters.

- With integration we can take into account simultaneously information from rsfMRI and DTI.

Fratello, Michele, et al. Submitted for publication

Semi-supervised Subgroup discovery in ALS



- We looked for relations with clinical information.
- We discovered that one of the clusters has an enriched group of subjects with lower limb onset and 2° clinical stage, with respect to the dataset
- The significance of the enriched group has been tested with a permutation test obtaining a p-value p=0.0033

Fratello, Michele, et al. Submitted for publication

Thank You! Questions?

UNIVERSITÀ DEGLI STUDI DI SALERNO

SECONDA UNIVERSITÀ DEGLI STUDI DI NAPOLI
unina2.it

People who participated to this work:

ANGELA SERRA
PhD Student

MICHELE FRATELLO
PhD Student

VITTORIO FORTINO
Ph.D., Researcher

GIANCARLO RAICONI
Full Professor

FABRIZIO ESPOSITO
Associate Professor

DARIO GRECO
Ph.D., Assoc. Prof. Genetics

ROBERTO TAGLIAFERRI
Full Professor

NANOSOLUTIONS

Part of this project has been realized under the FP7 European project Nanosolutions (grant agreement FP7-309329), WP11

References

- P. Pavlidis, J. Weston, J. Cai, and W. N. Grundy, “Gene functional classification from heterogeneous data,” in Proceedings of the fifth annual international conference on Computational biology. ACM, 2001, pp. 249–255.
- B. Ray, M. Henaff, S. Ma, E. Efstathiadis, E. R. Peskin, M. Picone, T. Poli, C. F. Aliferis, and A. Statnikov, “Information content and analysis methods for multi-modal high-throughput biomedical data,” Scientific reports, vol. 4, 2014.
- F. Napolitano, Y. Zhao, V. M. Moreira, R. Tagliaferri, J. Kere, M. D’Amato, and D. Greco, “Drug repositioning: a machine-learning approach through data integration.” J. Cheminformatics, vol. 5, p. 30, 2013.
- N. B. Larson, G. D. Jenkins, M. C. Larson, R. A. Vierkant, T. A. Sellers, C. M. Phelan, J. M. Schildkraut, R. Sutphen, P. P. Pharoah, S. A. Gayther et al., “Kernel canonical correlation analysis for

References

- Kasabov, Nikola, Eric Postma, and Jaap Van Den Herik. "AVIS: a connectionist-based framework for integrated auditory and visual information processing." Information Sciences 123.1 (2000): 127-148.
- Hinton, Geoffrey E., and Sam T. Roweis. "Stochastic neighbor embedding." Advances in neural information processing systems. 2002.
- Xie, Bo, et al. "m-SNE: Multiview stochastic neighbor embedding." Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on 41.4 (2011): 1088-1096.
- K. Kailing, H.-P. Kriegel, A. Pryakhin, and M. Schubert, “Clustering multi-represented objects with noise,” in Advances in Knowledge Discovery and Data Mining. Springer, 2004, pp. 394–403.
- X. Chen, X. Xu, J. Z. Huang, and Y. Ye, “Tw-(k)-means: Automated two-level variable weighting clustering algorithm for multiview data,” Knowledge and Data Engineering, IEEE Transactions on, vol. 25, no. 4, pp. 932–944, 2013.
- B. Long, S. Y. Philip, and Z. M. Zhang, “A general model for multiple view unsupervised learning.” in SDM. SIAM, 2008, pp. 822–833.

References

- D. Greene, “A Matrix Factorization Approach for Integrating Multiple Data Views,” Machine Learning and Knowledge Discovery in Databases, vol. 5781, pp. 423–438, 2009. [Online]. Available: http://www.springerlink.com/index/87g7r3p873w05m22.pdf
- A. Serra, M. Fratello, V. Fortino, G. Raiconi, R. Tagliaferri, and D. Greco, “Mvda: a multi-view genomic data integration methodology,” BMC bioinformatics, vol. 16, no. 1, p. 261, 2015.
- Z.Yang and G.Michailidis,“A non-negative matrix factorization method for detecting modules in heterogeneous omics multi-modal data,” Bioinformatics, vol. 32, no. 1, pp. 1–8, 2016.
- Taskesen, Erdogan, et al. "Pan-cancer subtyping in a 2D-map shows substructures that are driven by specific combinations of molecular characteristics." Scientific Reports 6 (2016).
- Wang, Bo, et al. "Similarity network fusion for aggregating data types on a genomic scale." Nature methods 11.3 (2014): 333-337.
- Fratello, Michele, et al. "A multi-view genomic data simulator." BMC