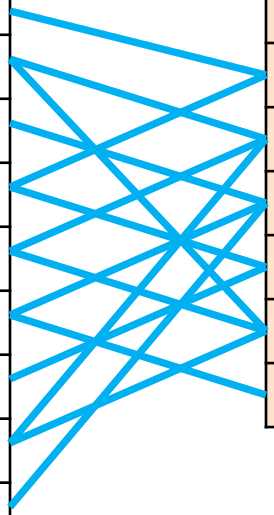
UNIFYING DEPENDENT CLUSTERING AND DISPARATE CLUSTERING FOR NON-HOMOGENEOUS DATA

M. Shahriar Hossain, Dept. of CS, Virginia Tech
Satish Tadepalli, Dept. of CS, Virginia Tech
Layne T. Watson, Dept. of CS, Virginia Tech
Ian Davidson, Dept. of CS, UC Davis
Richard F. Helm, Dept. of Biochemistry, Virginia Tech
Naren Ramakrishnan, Dept. of CS, Virginia Tech

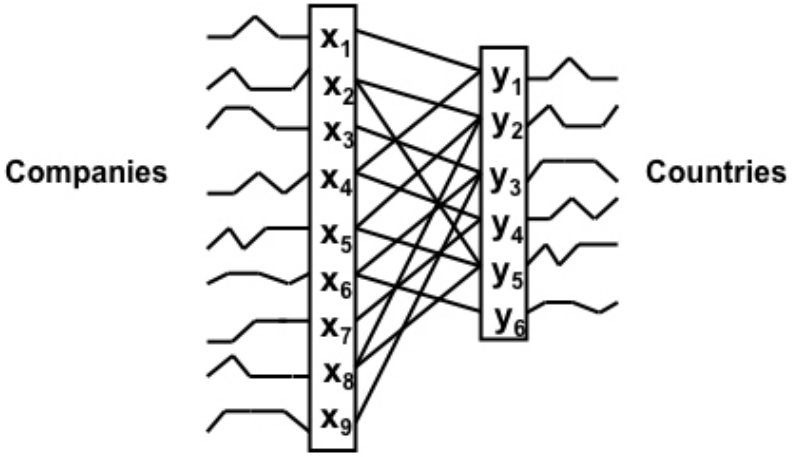
Problem Setting

Companies	Avg. salary of Employees	Stock values	Profit margins
x_1	1.0 K	25.11	11%
x_2	1.1 K	21.32	20%
x_3	1.2 K	28.81	12%
x_4	1.2 K	31.85	22%
x_5	1.1 K	85.32	5%
x_6	1.2 K	10.71	32%
x_7	0.9 K	11.61	18%
x_8	1.1 K	35.81	12%
x_9	1.2 K	20.81	4%



Countries	GDP	GNP	Inflation
y_1	$11832 B	$12970 B	-0.4%
y_2	$8219 B	$8153 B	2.0%
y_3	$6732 B	$7812 B	-0.3%
y_4	$1761 B	$2852 B	1.8%
y_5	$5022 B	$4391 B	0.0%
y_6	$7224 B	$8312 B	1.1%

Objective



Fortunes of individual companies are inter-twined with the fortunes of the countries.

Performances of companies may not necessarily be tied to the economies of the countries.

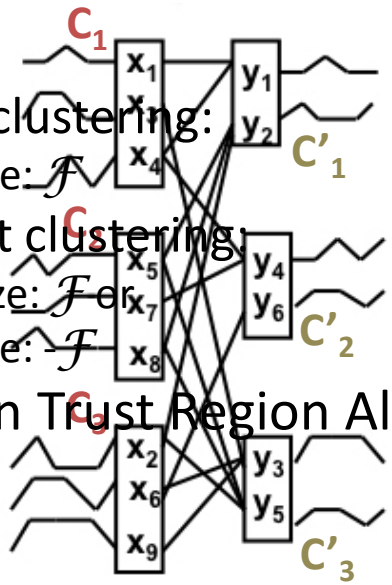
Companies \ Countries	C'_1	C'_2	C'_3
C_1	4	0	0
C_2	0	6	0
C_3	0	0	4

Companies \ Countries	C'_1	C'_2	C'_3
C_1	2	1	1
C_2	2	1	2
C_3	1	1	3

Objective Function

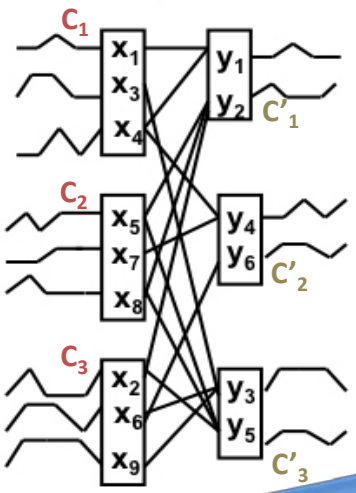
$v_i^{x_s}$ $v_j^{y_s}$

- Optimize $\mathcal{F}$
  - Disparate clustering:
    - minimize: $\mathcal{F}$
  - Dependent clustering:
    - maximize: $\mathcal{F}$ or
    - minimize: $-\mathcal{F}$
- Quasi Newton Trust Region Algorithm



	C'1	C'2	C'3
C1	2	1	1
C2	2	1	2
C3	1	1	3

Formulations



	C'1	C'2	C'3
C1	2	1	1
C2	2	1	2
C3	1	1	3

w

	C'1	C'2	C'3
C1	0.33	0.33	0.33
C2	0.33	0.33	0.33
C3	0.33	0.33	0.33

U

α

	C'1	C'2	C'3
C1	0.5	0.25	0.25
C2	0.4	0.2	0.4
C3	0.2	0.2	0.6

Row distributions

β

	C'1	C'2	C'3
C1	0.4	0.33	0.17
C2	0.4	0.33	0.33
C3	0.2	0.33	0.5

Col. distributions

Membership Probability

$$v_i^{(\mathbf{x}_s)} = \frac{\exp(-\frac{\rho}{D}||\mathbf{x}_s - \mathbf{m}_{i,\mathcal{X}}||^2)}{\sum_{i'=1}^{k_x}\exp(-\frac{\rho}{D}||\mathbf{x}_s - \mathbf{m}_{i',\mathcal{X}}||^2)}$$

Contingency Table Counts

$$w_{ij} = \sum_{s=1}^{n_x}\sum_{t=1}^{n_y} B(s,t)v_i^{(\mathbf{x}_s)}v_j^{(\mathbf{y}_t)}$$

Probability Distributions (row and column)

$$p(\alpha_i = j) = p(C_{(y)} = j|C_{(x)} = i) = \frac{w_{ij}}{w_{i.}}$$

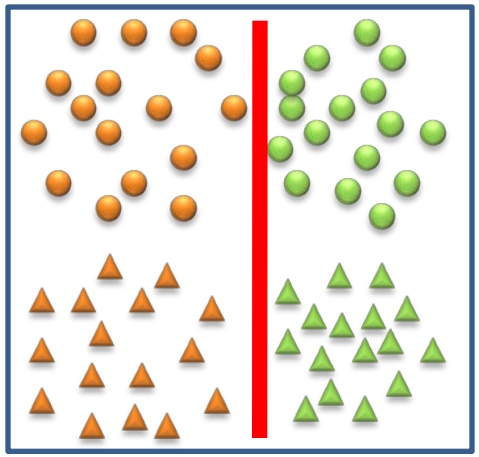
$$p(\beta_j = i) = p(C_{(x)} = i|C_{(y)} = j) = \frac{w_{ij}}{w_{.j}}$$

Objective Function

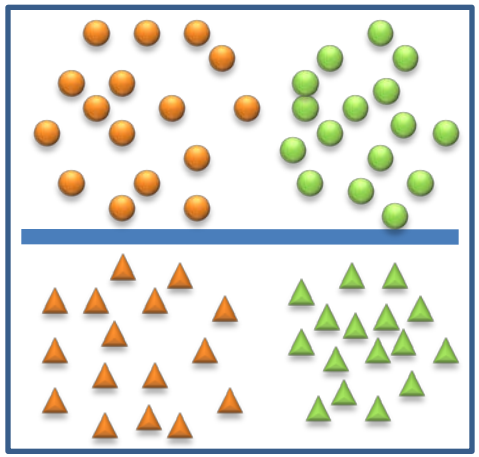
$$\mathcal{F} = \frac{1}{k_x}\sum_{i=1}^{k_x} D_{KL}\left(\alpha_i||U(\frac{1}{k_y})\right) + \frac{1}{k_y}\sum_{j=1}^{k_y} D_{KL}\left(\beta_j||U(\frac{1}{k_x})\right)$$

Single Dataset Scenarios

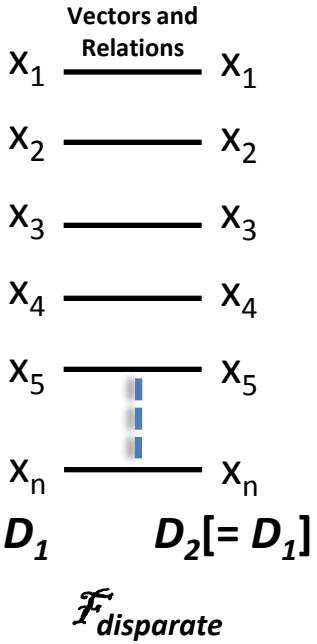
- **ALTERNATIVE CLUSTERING**



D_1



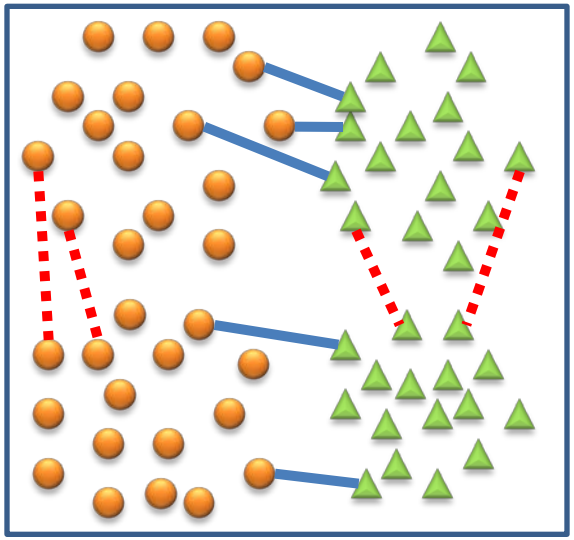
D_2



$\mathcal{F}_{disparate}$

Single Dataset Scenarios

- **CONSTRAINED CLUSTERING**
  - Instance-level constraints

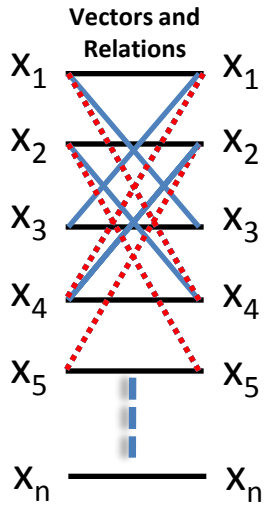


D_1

Must-Link (ML)

Must-Not-Link (MNL)

- **Clustering of D_1 is given.**
- **The desired constrained clustering is obtained in D_2.**

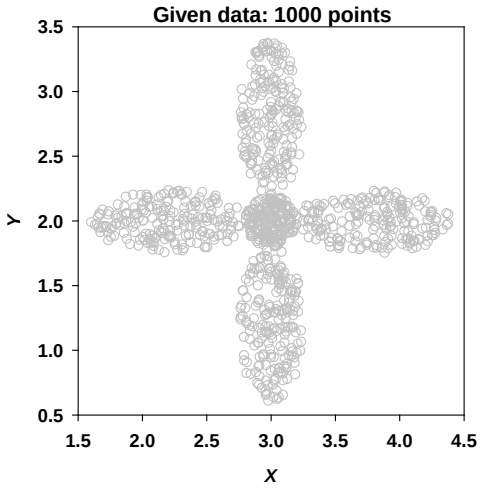


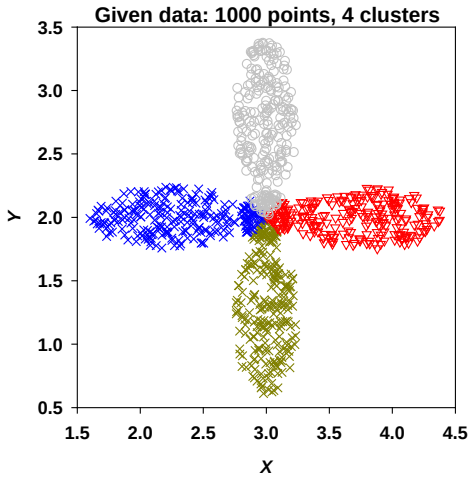
D_1 $D_2[= D_1]$

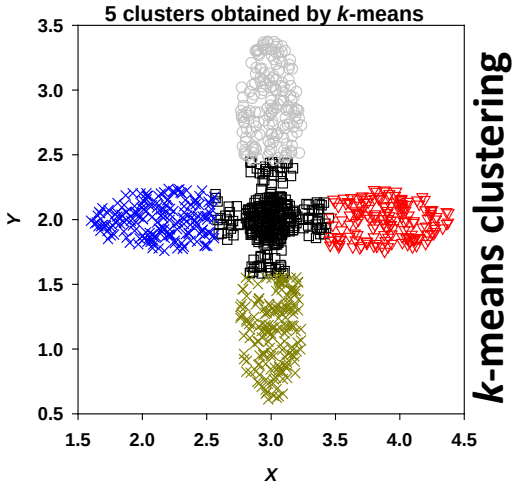
$$\mathcal{F}=\alpha\,\mathcal{F}_{dep}+(1-\alpha)\,\mathcal{F}_{disparate}$$

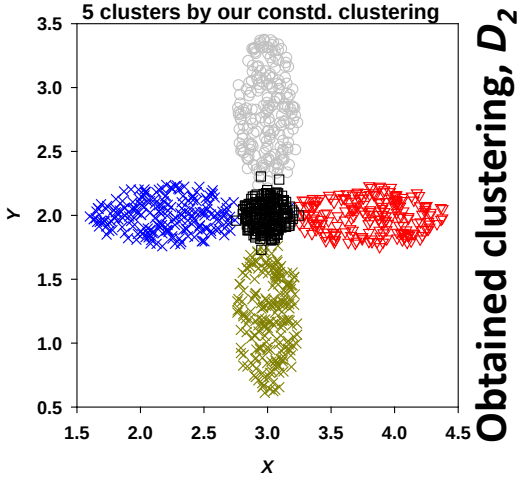
Single Dataset Scenarios

- **CONSTRAINED CLUSTERING**
  - Cluster-level constraints

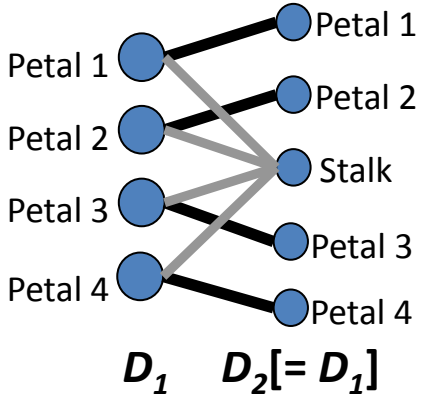






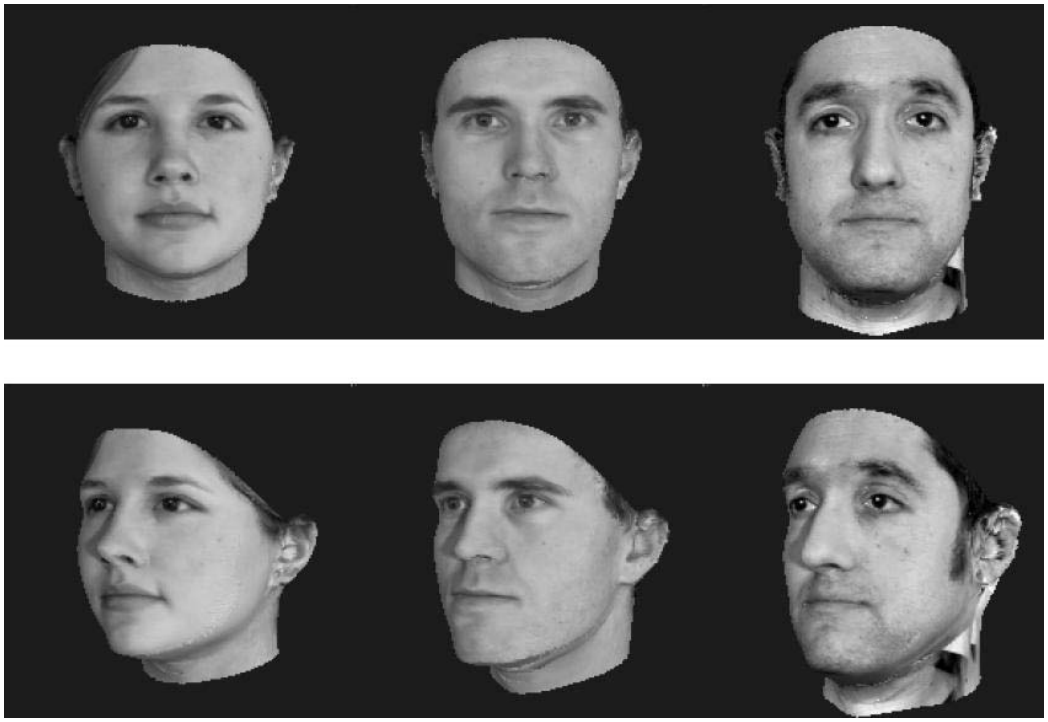


- **Clustering of D_1 is given.**
- **The desired constrained clustering is obtained in D_2.**



Experimental Results

- **ALTERNATIVE CLUSTERING**
- Portrait Dataset

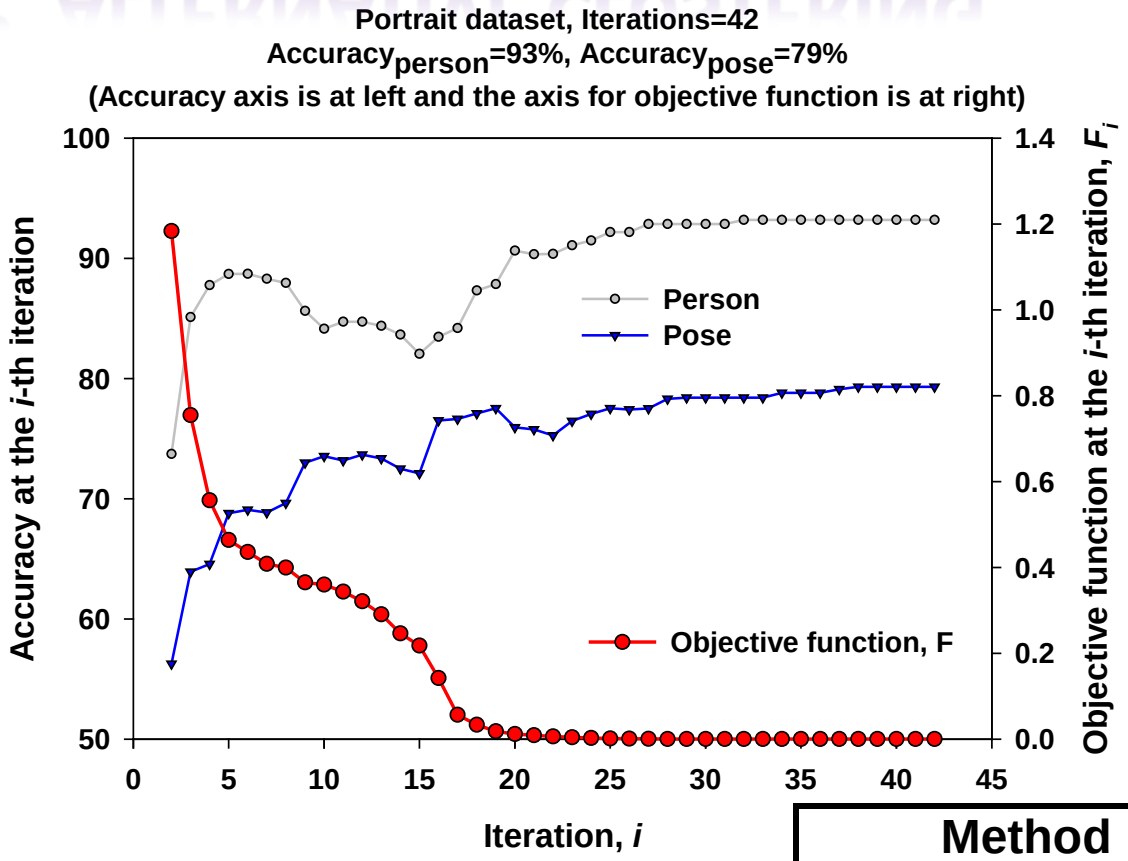


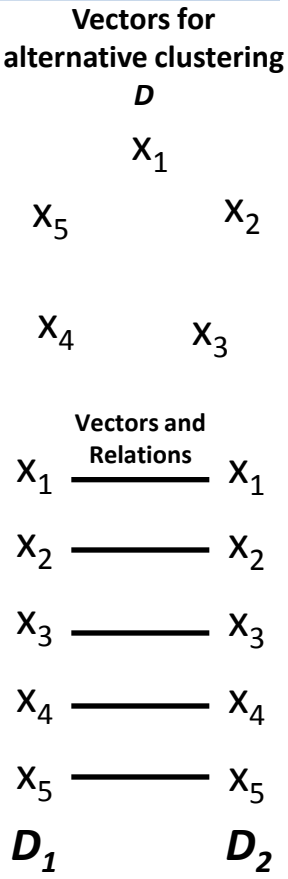
Prateek Jain et al. 2008

- **3** people each in **3** poses and **36** illuminations (i.e., **324** images.)
- **300** features

Experimental Results

- **ALTERNATIVE CLUSTERING**

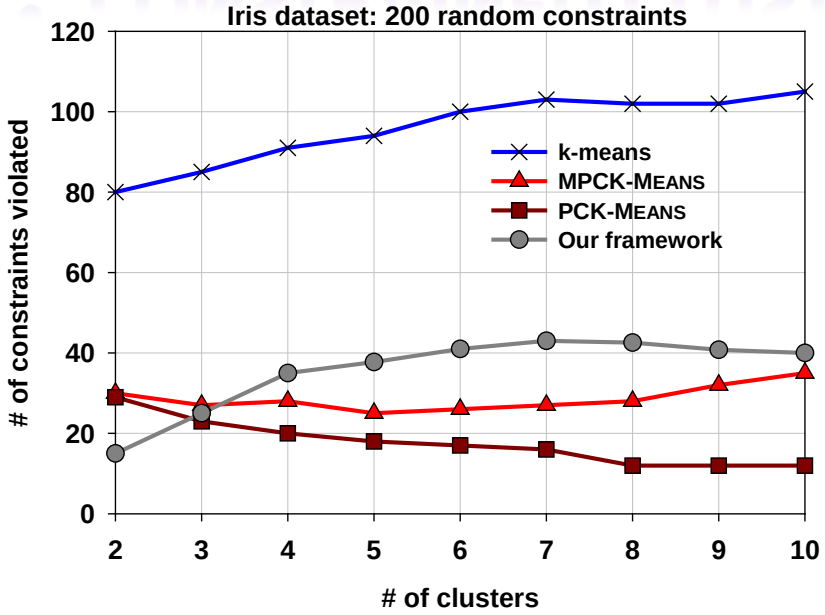


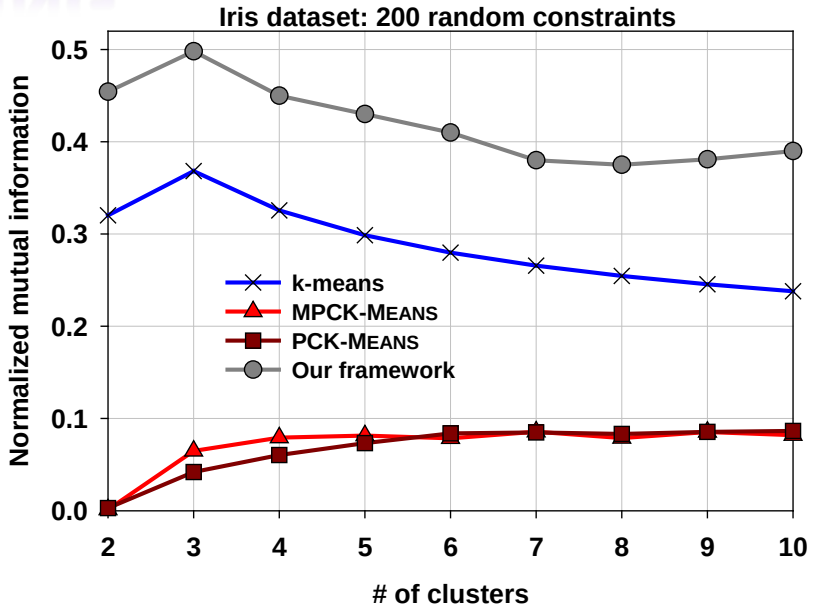


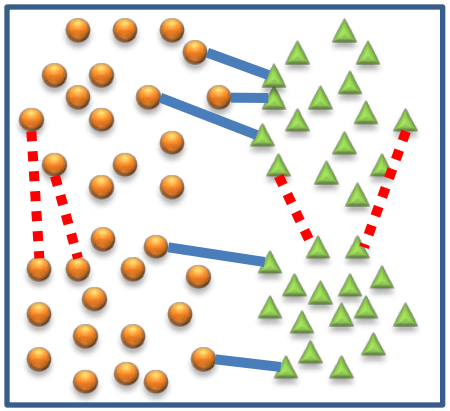
Method	Person	Pose
k-means	0.65	0.55
Conv-EM	0.69	0.72
Dec-kmeans	0.84	0.78
Our framework	0.93	0.79

Experimental Results

- **CONSTRAINED CLUSTERING**







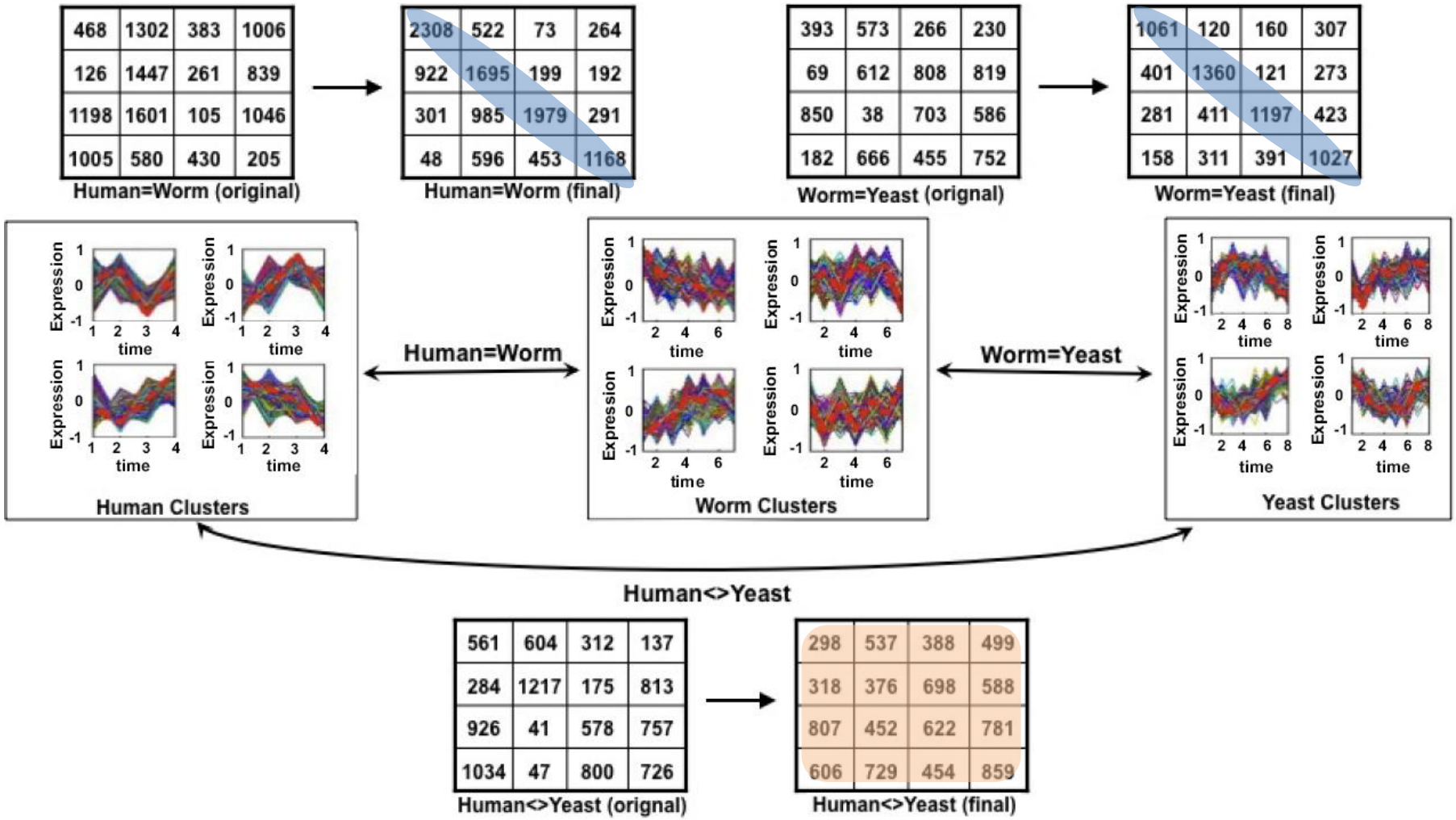
D_1

Must-Link (ML)

Must-Not-Link (MNL)

Experimental Results

- **COMPARING GENE EXPRESSION PROGRAMS**



Future Work & Conclusion

- Future directions
  - Capture more expressive relationships
    - Dependent and disparate clustering on same set of relationships
    - Different goal for different types of relationships (one-to-one, ML, MNL, etc.)
  - Clustering dependencies
- Conclusion
  - General, expressive framework for clustering non-homogenous datasets
  - The framework subsumes previously defined formulations
    - MDI (Kullback et al. ‘78), Disparate Clustering (Jain et al. ‘ 08), Clustering over Relation Graphs (Banerjee et al. ‘07), Multivariate Information Bottleneck (Friedman ‘01), etc.

Thank you