Sparse and Low-Rank Modeling for High-Dimensional Data Analysis

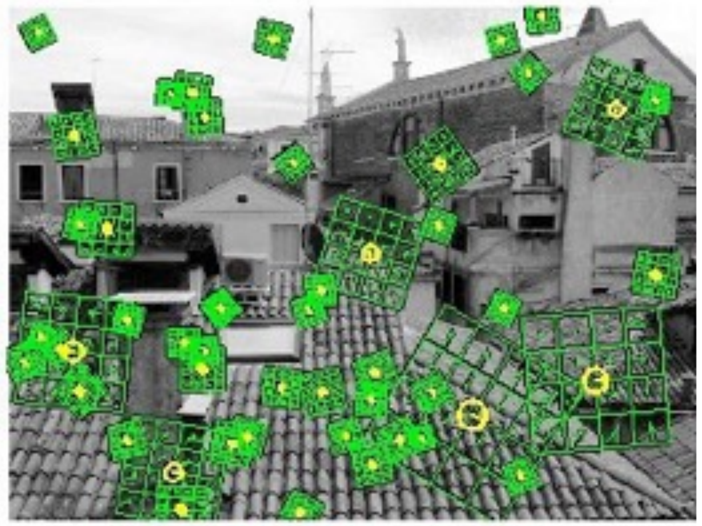
Ehsan Elhamifar, Rene Vidal,
John Wright, Guillermo Sapiro

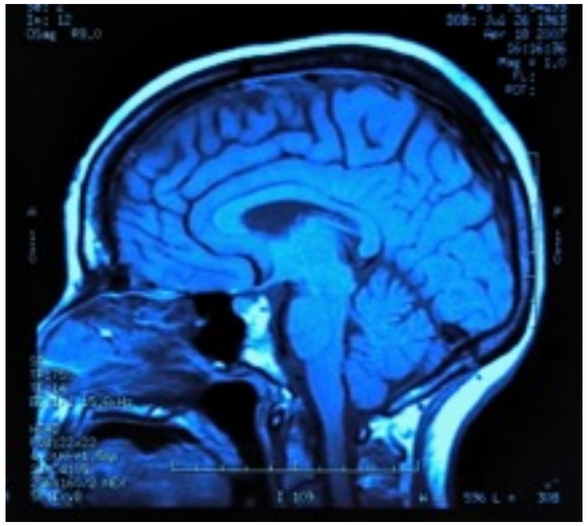
CVPR 2015 Tutorial
Boston, MA

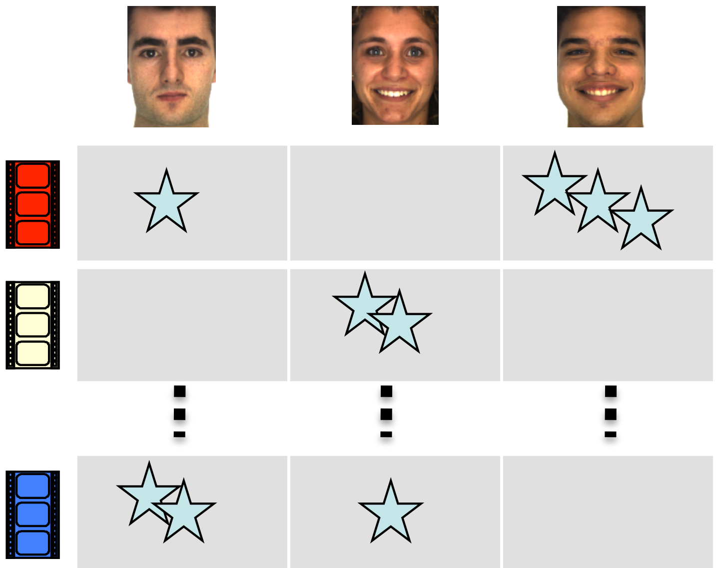
High-dimensional data deluge

72 hrs new **videos** / minute

300M new **photos** / day

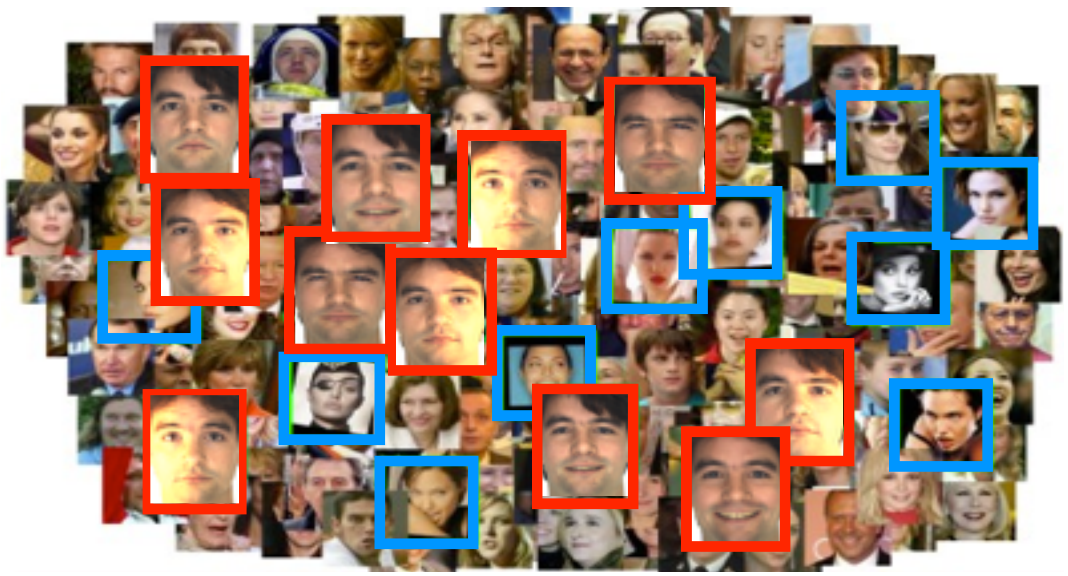


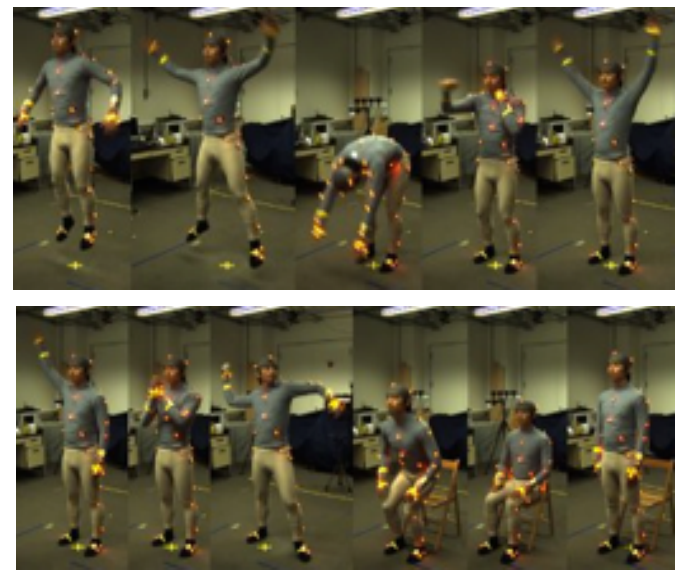


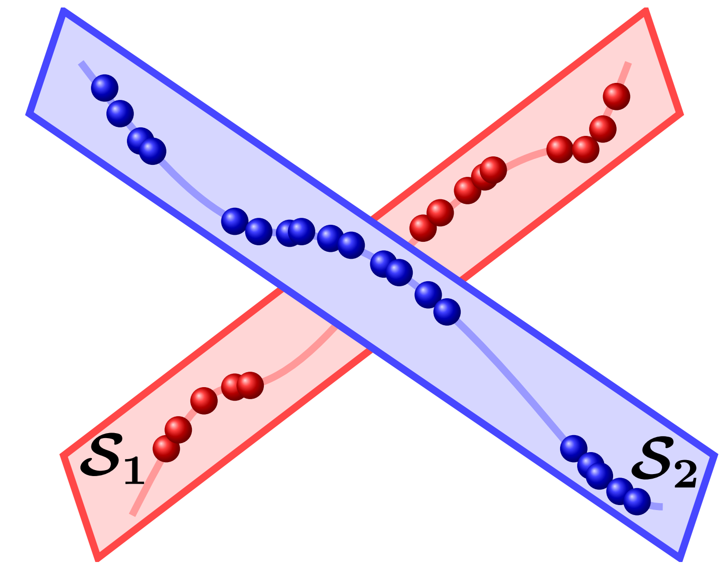


Low-dimensional structures

- Intrinsic structures are low-dimensional

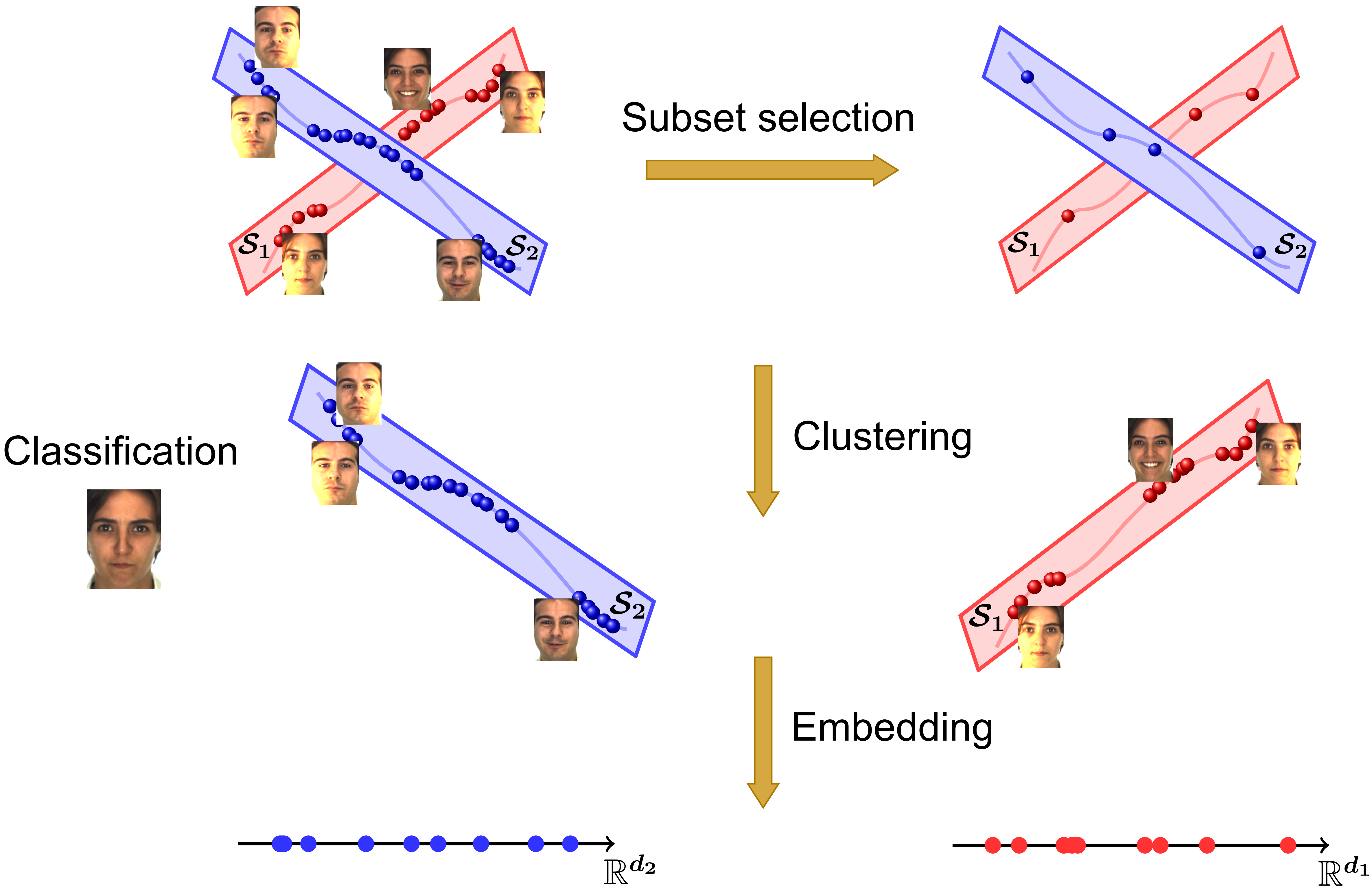






Basri-Jacobs'03, Tomasi-Kanade'92, Deerwester et al '90, Goldberg-Nichols-Oki-Terry '92

High-dimensional data analysis

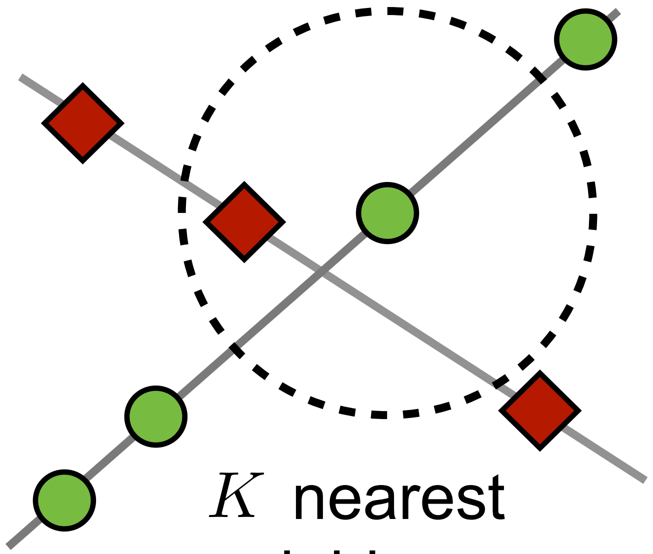


Challenges

- Clustering and subset selection: **Non-convex** and **NP-hard**
- Real data are often **corrupted**

- **Little prior knowledge** about low-dim structures
- Points in different groups can be **very close**
  - Ext YaleB dataset (38 subjects, 64 images)

$K=1$	$K=2$	$K=3$	$K=4$
6%	14%	23%	31%



This tutorial

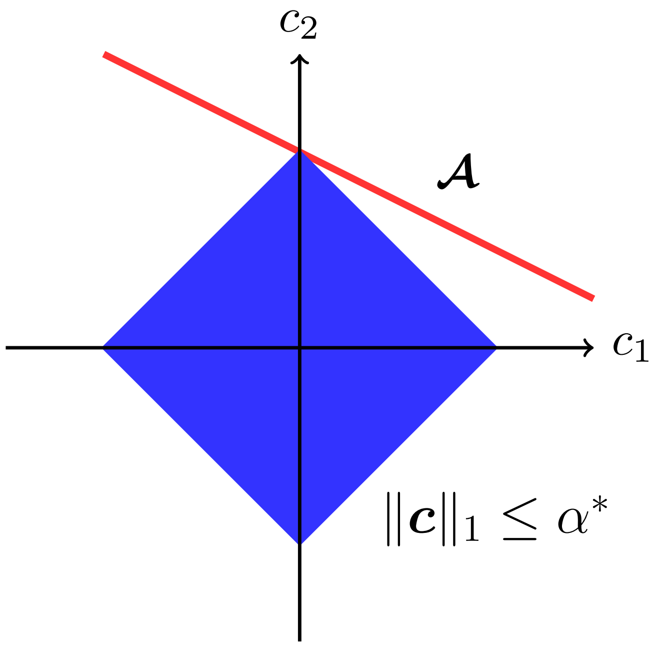
Efficient, **robust** and **provably correct** algorithms for

(1) **clustering, subset selection**

(2) **classification, dimension reduction**

Tools:

- Sparse & low-rank representation
- High-dimensional statistics & geometry
- Convex programming & analysis



This tutorial

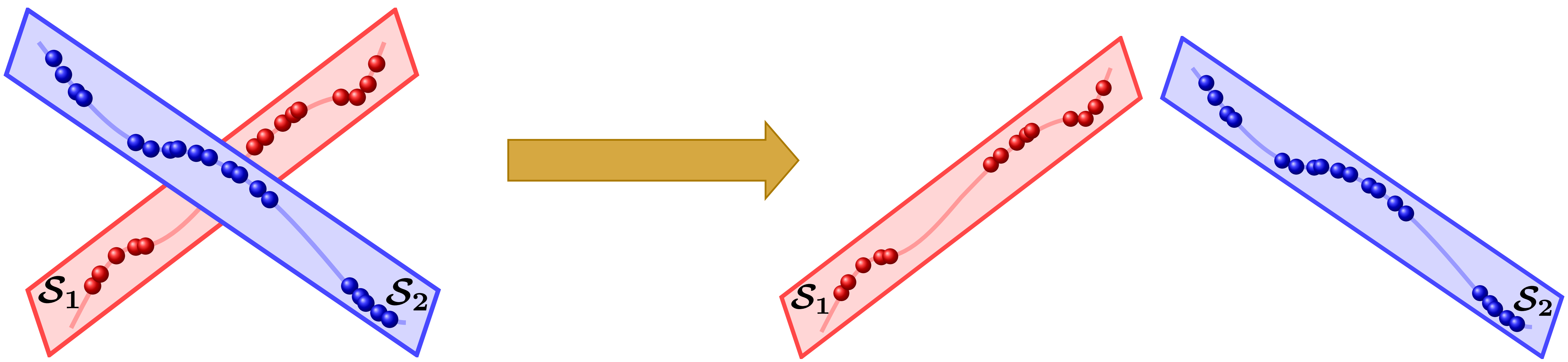
1) **Clustering, Subset selection: algorithm, theory, applications**
    - Ehsan Elhamifar
    - Rene Vidal

— Coffee Break 3:30pm — 4:15pm

2) **Robust PCA, Learning low-rank transformations: algorithm, theory, applications**
    - John Wright
    - Guillermo Sapiro

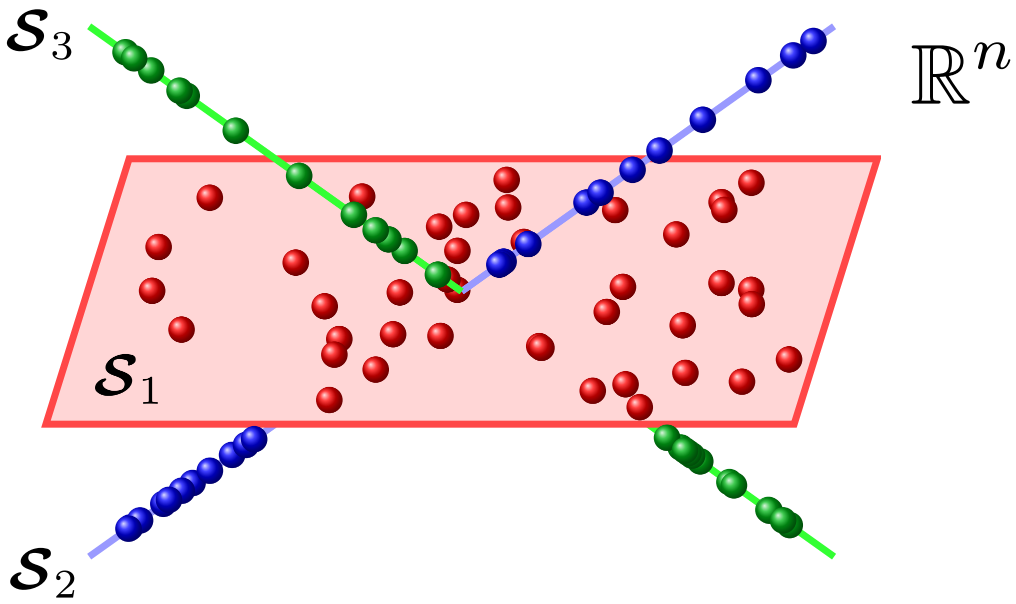
Sparse Subspace Clustering

Ehsan Elhamifar



Subspace clustering problem

- Given points $\{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_N\}$ in $\mathbb{R}^n$ lying in $\mathcal{S}_1 \cup \ldots \cup \mathcal{S}_L$, find
  - **Basis** for each subspace
  - **Clustering** of the data



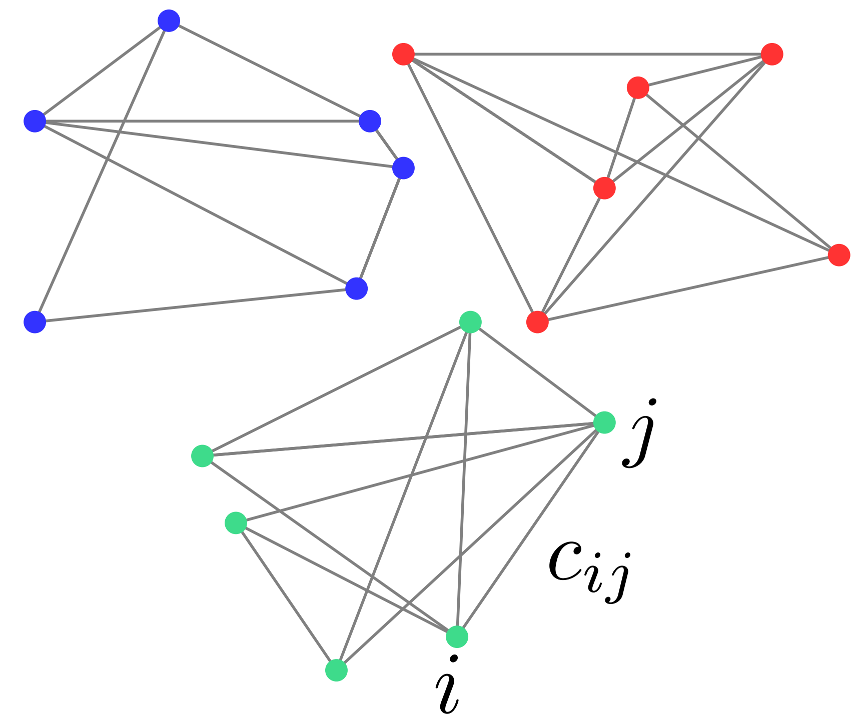
- **Challenging** for multiple subspaces
  - Do not know subspace **bases**
  - Do not know **memberships** of points
  - Corruption by **noise**, **missing entries**, **outliers**, ...

→ Possible approach: **Expectation Maximization**, Issue: **Local minima**

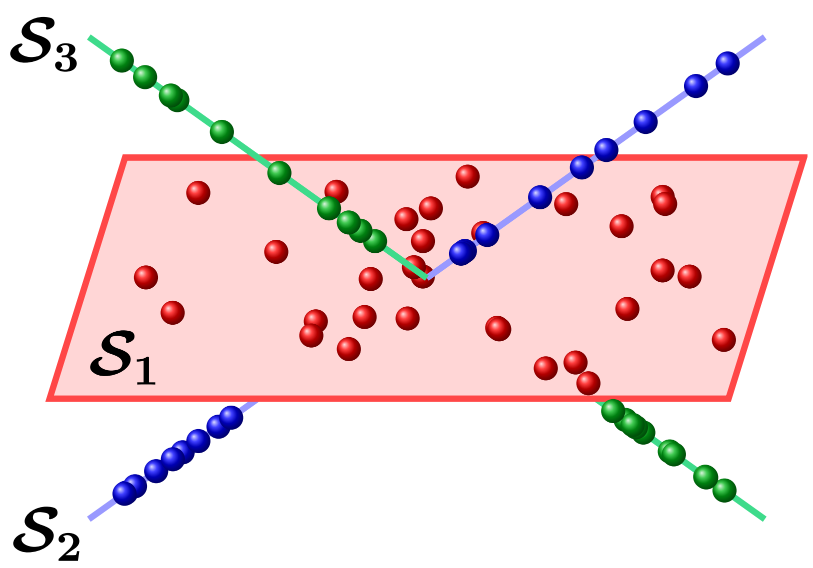
Tomasi-Kanade'92, Tipping-Bishop'99, Tseng'00, Kanatani'01, Vidal-Ma-Sastry'05, Yan-Pollefeys'06, Chen-Lerman'09

Spectral clustering-based approach

- **Spectral Clustering**
  - Represent points as graph nodes
  - Connect i and j with weight c_{ij}
  - Infer clusters from graph Laplacian



- Good similarity for subspaces?
  - Points in the same subspace: $c_{ij} \neq 0$
  - Points in different subspaces: $c_{ij} = 0$
  - ~~Nearest neighbors~~

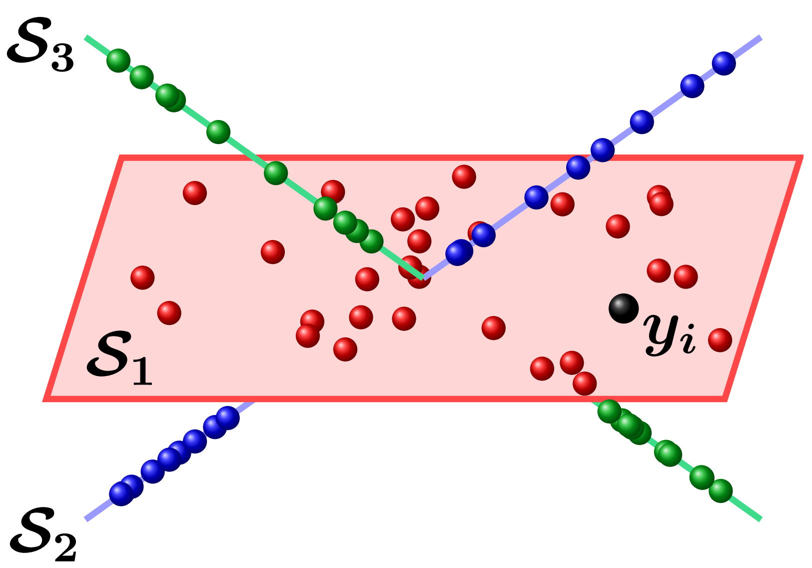


Fiedler'73, Shi-Malik'01, Belkin-Niyogi'01, Ng-Weiss-Jordan'01, Xing-Jordan'03, Von Luxburg'07

Subspace clustering: idea

- **Self-Expressiveness Property (SEP)**
  - $\boldsymbol{y}_i = \boldsymbol{Y}\boldsymbol{c}_i$ → many solutions
  - $\boldsymbol{Y} = [\boldsymbol{Y}_1 \quad \boldsymbol{Y}_2 \quad \cdots \quad \boldsymbol{Y}_L]\,\boldsymbol{\Gamma}$

    low column-rank



- In $\mathcal{S}_\ell$ of dim d_ℓ, a point can be reconstructed by d_ℓ other points

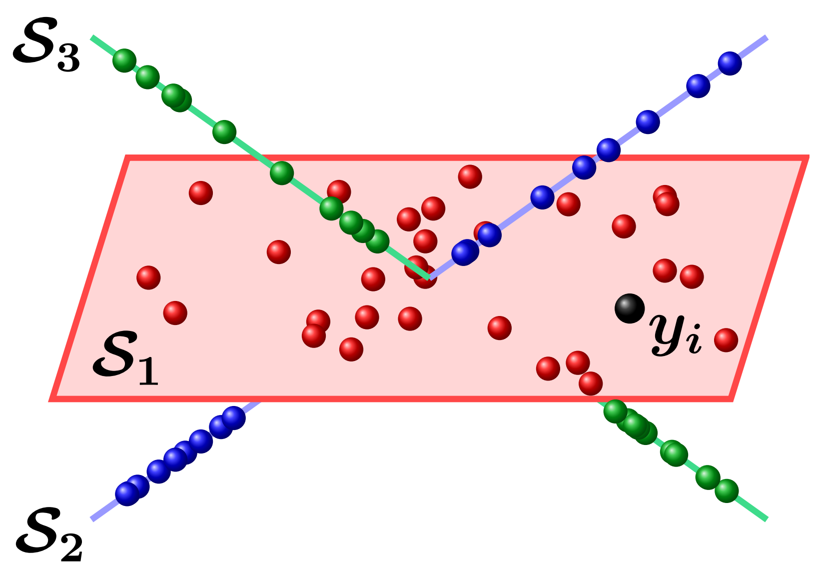
$$\min \|\boldsymbol{c}_i\|_0 \quad \text{s. t.} \quad \boldsymbol{y}_i = \boldsymbol{Y}\boldsymbol{c}_i, \;\; c_{ii} = 0 \qquad \text{NP-hard}$$

  - ℓ_0 : number of nonzero elements

Subspace clustering: idea

- **Self-Expressiveness Property (SEP)**
  - $\boldsymbol{y}_i = \boldsymbol{Y}\boldsymbol{c}_i$ ⟶ many solutions
  - $\boldsymbol{Y} = [\boldsymbol{Y}_1 \quad \boldsymbol{Y}_2 \quad \cdots \quad \boldsymbol{Y}_L]\,\boldsymbol{\Gamma}$

    low column-rank



- In $\mathcal{S}_\ell$ of dim d_ℓ, a point can be reconstructed by d_ℓ other points

$$\min \|\boldsymbol{c}_i\|_1 \quad \text{s.t.} \quad \boldsymbol{y}_i = \boldsymbol{Y}\boldsymbol{c}_i,\; c_{ii} = 0 \qquad \text{Convex}$$

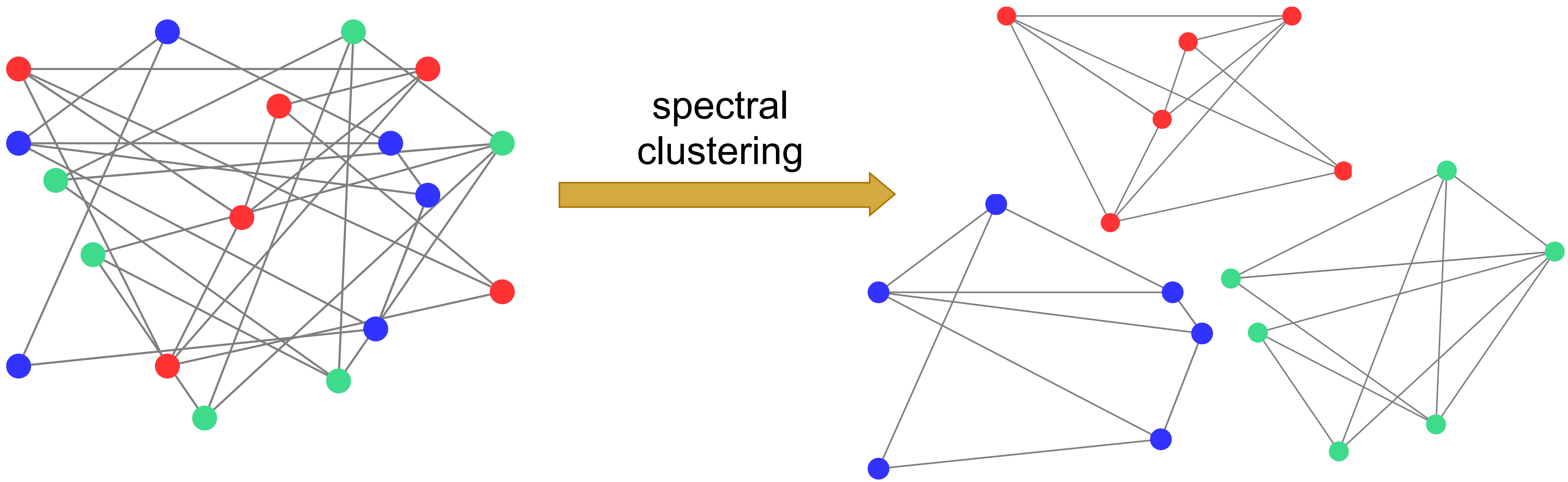
  - ℓ_1: sum of absolute values of elements

Sparse subspace clustering (SSC)

- 1: Solve the sparse optimization

$$\min \|\boldsymbol{c}_i\|_1 \quad \text{s.t.} \quad \boldsymbol{y}_i = \boldsymbol{Y}\boldsymbol{c}_i,\ c_{ii} = 0 \qquad \boldsymbol{c}_i^* = \begin{bmatrix} c_{i1}^* \\ \vdots \\ c_{iN}^* \end{bmatrix}$$

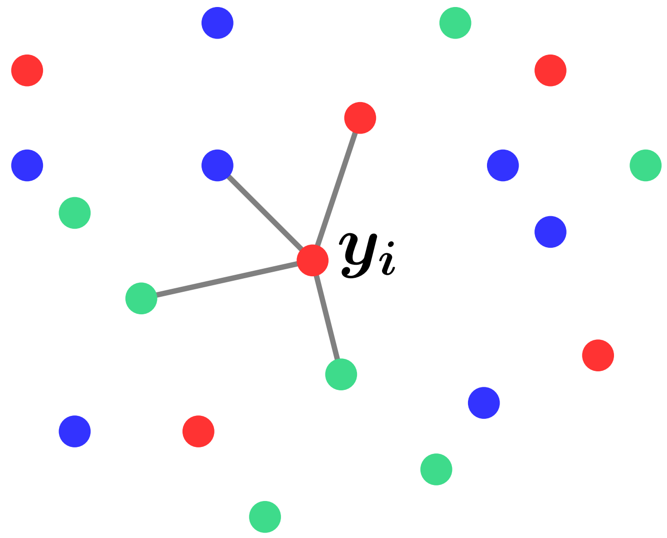
- 2: Infer clustering from similarity graph



E. Elhamifar and R. Vidal, CVPR 2009; E. Elhamifar and R. Vidal, IEEE Trans. PAMI, 2013

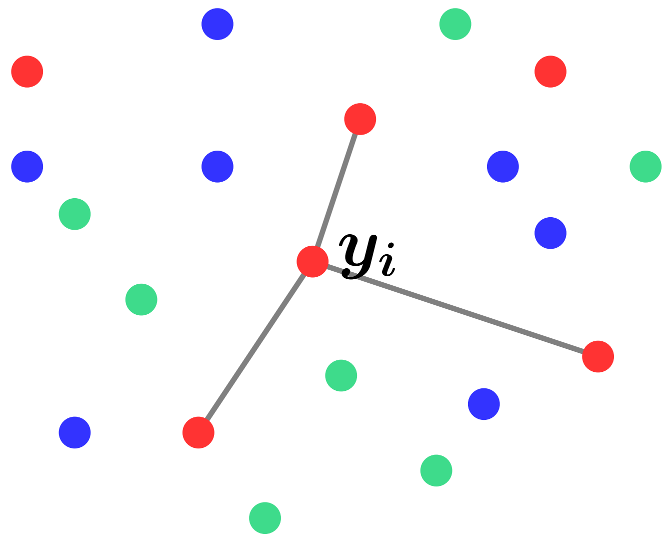
L1 graph vs k-NN graph

- Conventional graph clustering
  - 1) build a k-NN graph
  - 2) learn edge weights
  - 3) partition the graph



$$c_{ij} = e^{-\frac{\|\boldsymbol{y}_i - \boldsymbol{y}_j\|^2}{2\sigma^2}}$$

- SSC algorithm
  - 1) learn graph & weights
  - 2) partition the graph



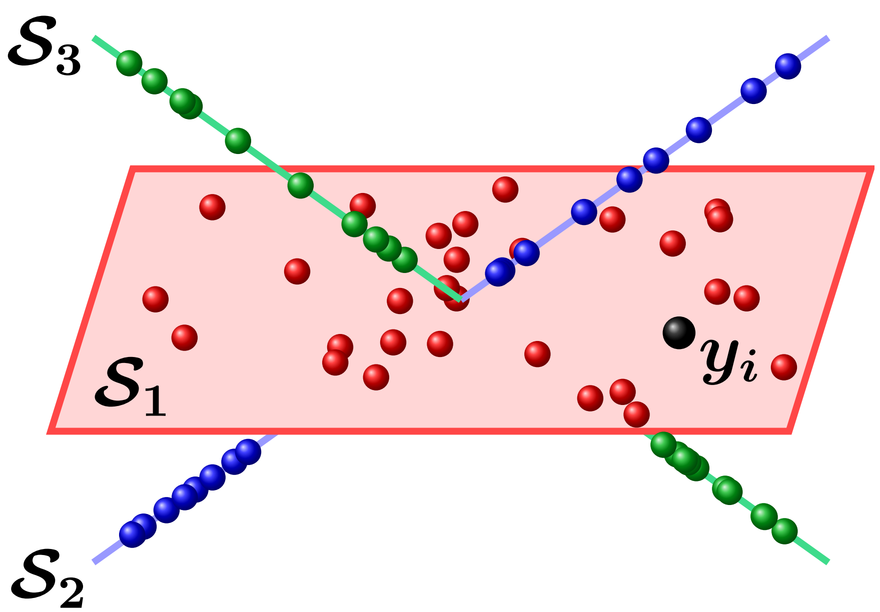
$$\boldsymbol{c}_i^* = \begin{bmatrix} 0 \\ 0.8 \\ 0 \\ \vdots \\ 0.3 \\ 0 \end{bmatrix}$$

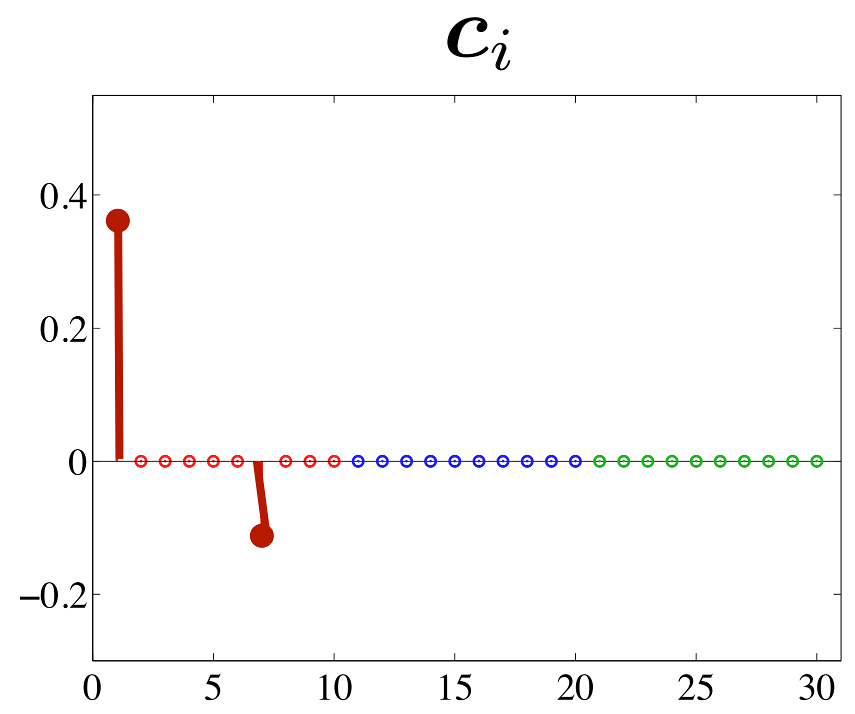
SSC *automatically* selects the right number of neighbors!
SSC can deal with subspaces of *different dimensions*!

E. Elhamifar and R. Vidal, CVPR 2009; E. Elhamifar and R. Vidal, IEEE Trans. PAMI, 2013

Theoretical analysis

- When does SSC succeed?
  - ℓ_1 selects points from the correct subspace: no false discovery





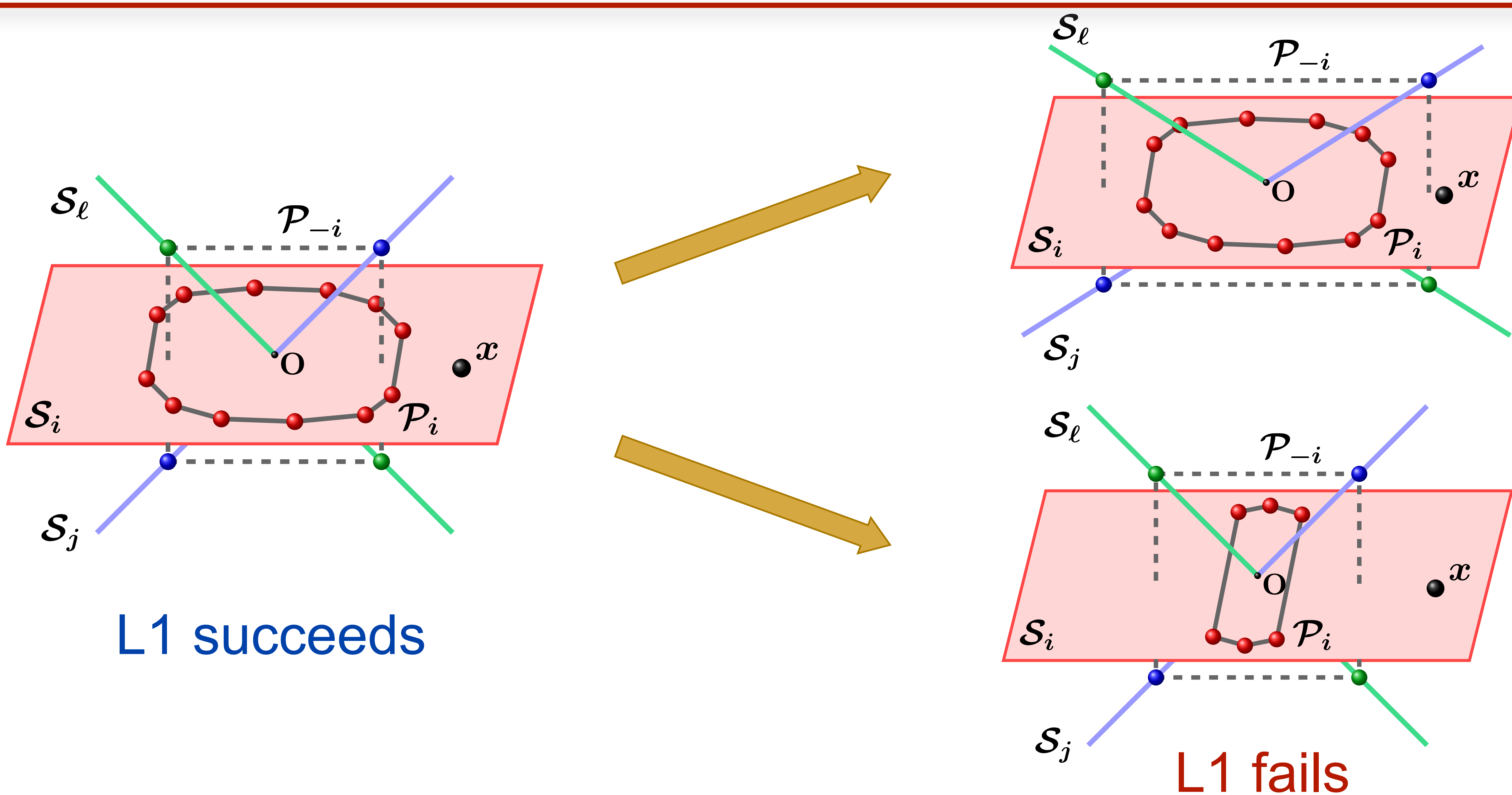
- **More challenging** than conventional sparse recovery
  - Sparse representation from the correct subspace
  - Sparse representation not unique

Donoho-Elad'03, La-Do'05, Candes-Romberg-Tao'06, Tropp-Gilbert'07, Wright-Ma'10, Sankaranarayanan-Turaga-Baraniuk-Chellappa'10

Geometry-based theoretical guarantees

- **Theorem:** SSC has zero false discovery for any $\boldsymbol{y} \in \mathcal{S}_i$ if

$$\max_{j \neq i} \cos(\theta_{ij}) < \max_{\operatorname{rank}(\mathbf{Y}'_i)=d_i} \sigma_{d_i}(\mathbf{Y}'_i) \,/\, \sqrt{d_i}$$



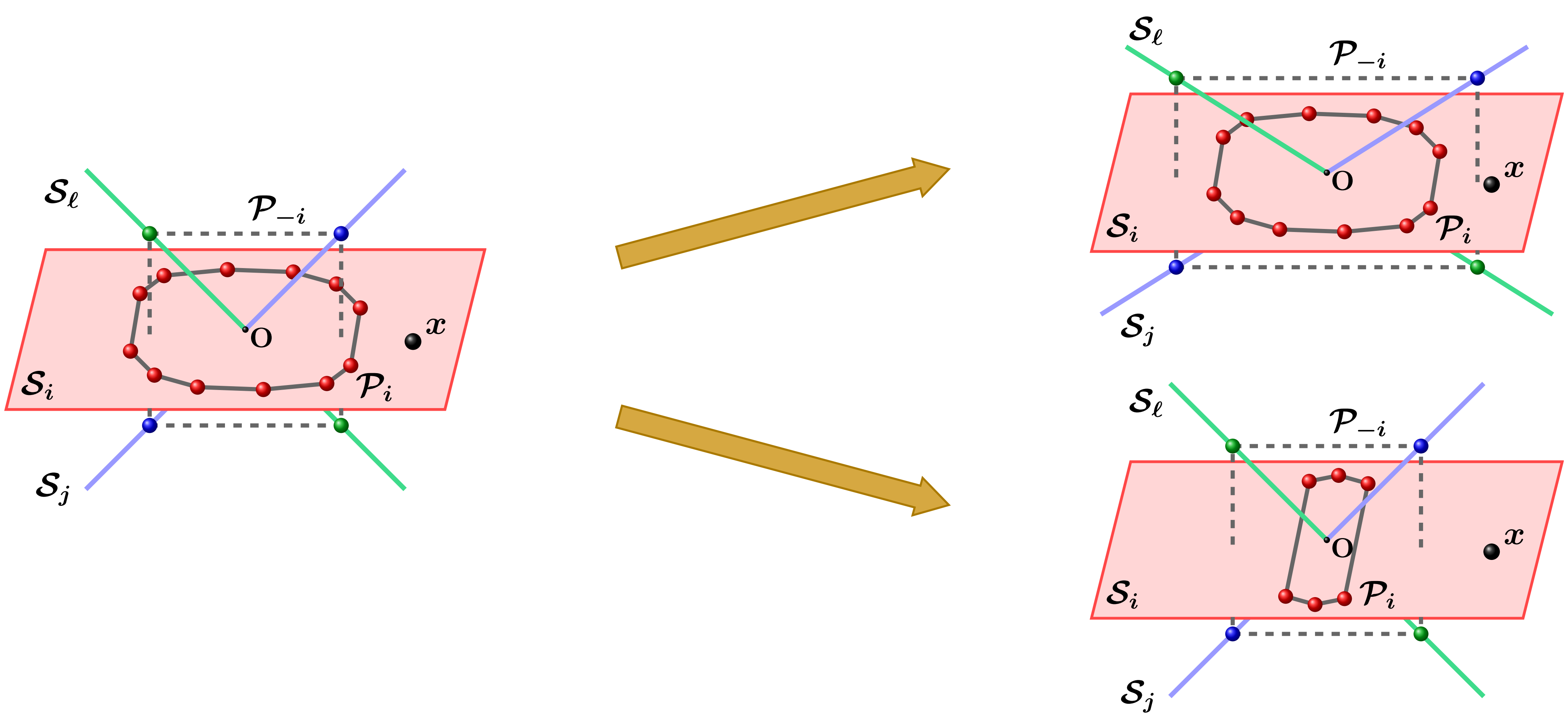
E. Elhamifar and R. Vidal, ICASSP 2010; E. Elhamifar and R. Vidal, IEEE Trans. PAMI, 2013

Geometry-based theoretical guarantees

- **Theorem:** SSC has zero false discovery for any $\boldsymbol{y} \in \mathcal{S}_i$ if

$$\max_{j \neq i} \cos(\theta_{ij}) < \max_{\operatorname{rank}(\mathbf{Y}_i^{'})=d_i} \sigma_{d_i}(\mathbf{Y}_i^{'}) \,/\, \sqrt{d_i}$$

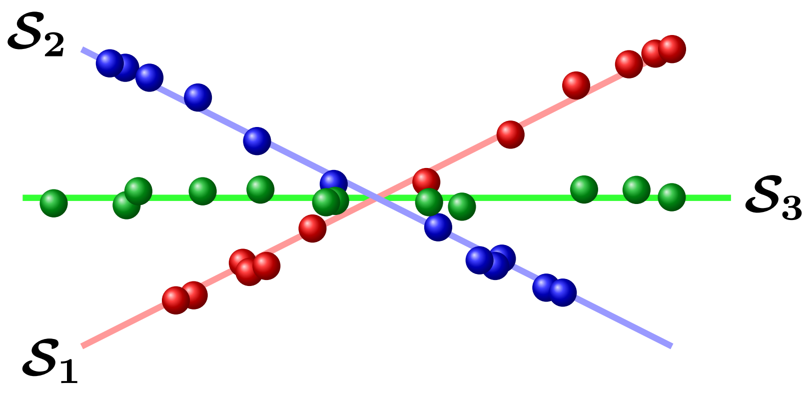
No need to have many points; Need a few but well distributed!



E. Elhamifar and R. Vidal, ICASSP 2010; E. Elhamifar and R. Vidal, IEEE Trans. PAMI, 2013

Clustering noisy data

- All points contaminated by noise



$$\widetilde{\boldsymbol{Y}} = \boldsymbol{Y} + \boldsymbol{Z} \qquad z_{ij} \sim \mathcal{N}(0, \sigma^2/n) \ \text{i.i.d.}$$

- Self-expressiveness implies

$$\underbrace{\boldsymbol{y}_i = \boldsymbol{Y}\boldsymbol{c}_i}_{\boldsymbol{c}_i \text{ sparse}} \Longrightarrow \widetilde{\boldsymbol{y}}_i = \widetilde{\boldsymbol{Y}}\underbrace{\boldsymbol{c}_i}_{\text{sparse}} + \underbrace{(\boldsymbol{z}_i - \boldsymbol{Z}\boldsymbol{c}_i)}_{\text{perturbation}}$$

- Solve Lasso

$$\min \ \lambda \|\boldsymbol{c}_i\|_1 + \frac{1}{2}\|\widetilde{\boldsymbol{y}}_i - \widetilde{\boldsymbol{Y}}\boldsymbol{c}_i\|_2^2 \quad \text{s.t.} \quad c_{ii} = 0$$

λ → ?
$\widetilde{\boldsymbol{y}}_i$, $\widetilde{\boldsymbol{Y}}$ → corrupted

M. Soltanolkotabi, E. Elhamifar and E. Candes, Annals of Stats, 2014; E. Elhamifar, M. Soltanolkotabi, S. Sastry, IEEE TSP, 2015

Robust SSC

- **Theorem:** Assume noise-free data is drawn uniformly at random from the intersection of each subspace and hypersphere. Apply the two-step procedure to $\widetilde{\boldsymbol{y}} \in \mathcal{S}_i$. Under some assumptions, if

$$\max_{j \neq i} \sqrt{\mathrm{Ave}(\cos^2(\boldsymbol{\theta}_{ij}))} \lesssim (\log N)^{-1}$$

with high prob, a) no false discovery, b) about subspace dim nonzeros.

- Algorithm: two-step approach

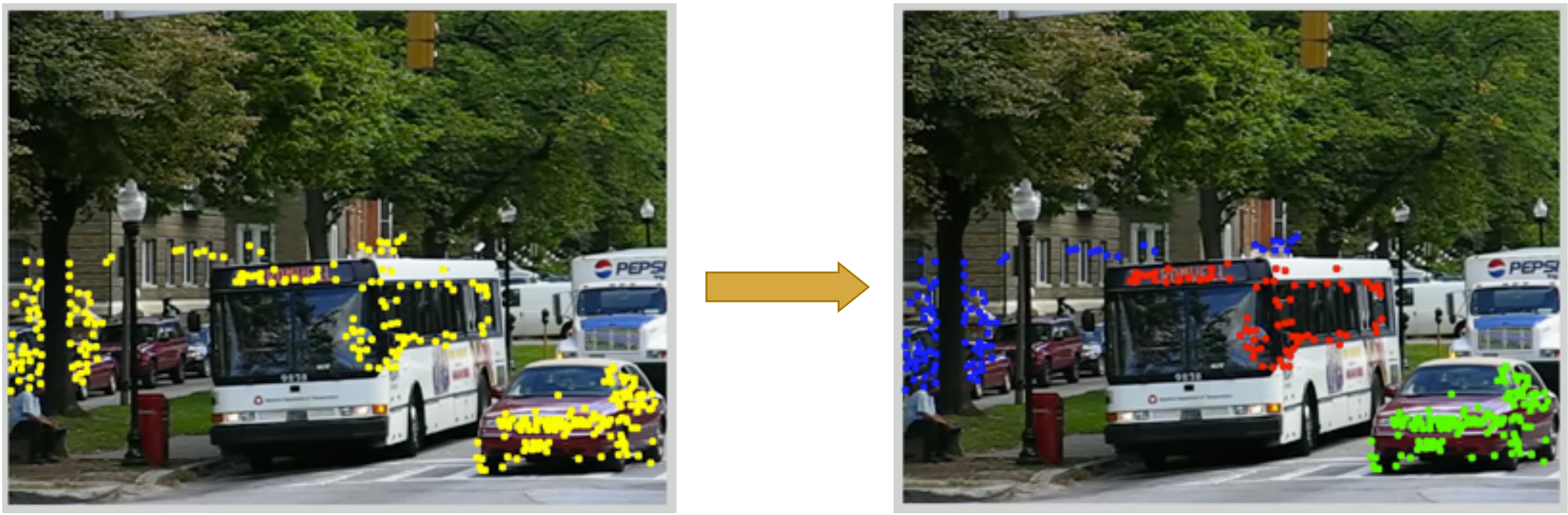
1) $\boldsymbol{\beta}_i^* = \arg\min_{\boldsymbol{\beta}_i} \|\boldsymbol{\beta}_i\|_1 \quad \mathrm{s.\,t.} \quad \|\widetilde{\boldsymbol{y}}_i - \widetilde{\boldsymbol{Y}}\boldsymbol{\beta}_i\|_2 \leq 2\sigma, \;\; \beta_{ii} = 0$ → $\lambda_i = \dfrac{1}{4\,\|\boldsymbol{\beta}_i^*\|_1}$

2) $\boldsymbol{c}_i^* = \operatorname{argmin}_{\boldsymbol{c}_i} \lambda_i \|\boldsymbol{c}_i\|_1 + \frac{1}{2}\|\,\widetilde{\boldsymbol{y}}_i - \widetilde{\boldsymbol{Y}}\boldsymbol{c}_i\|_2^2 \quad \mathrm{s.\,t.} \quad c_{ii} = 0$

data dependent!

M. Soltanolkotabi, E. Elhamifar and E. Candes, Annals of Stats, 2014; E. Elhamifar, M. Soltanolkotabi, S. Sastry, IEEE TSP, 2015

Application: motion segmentation

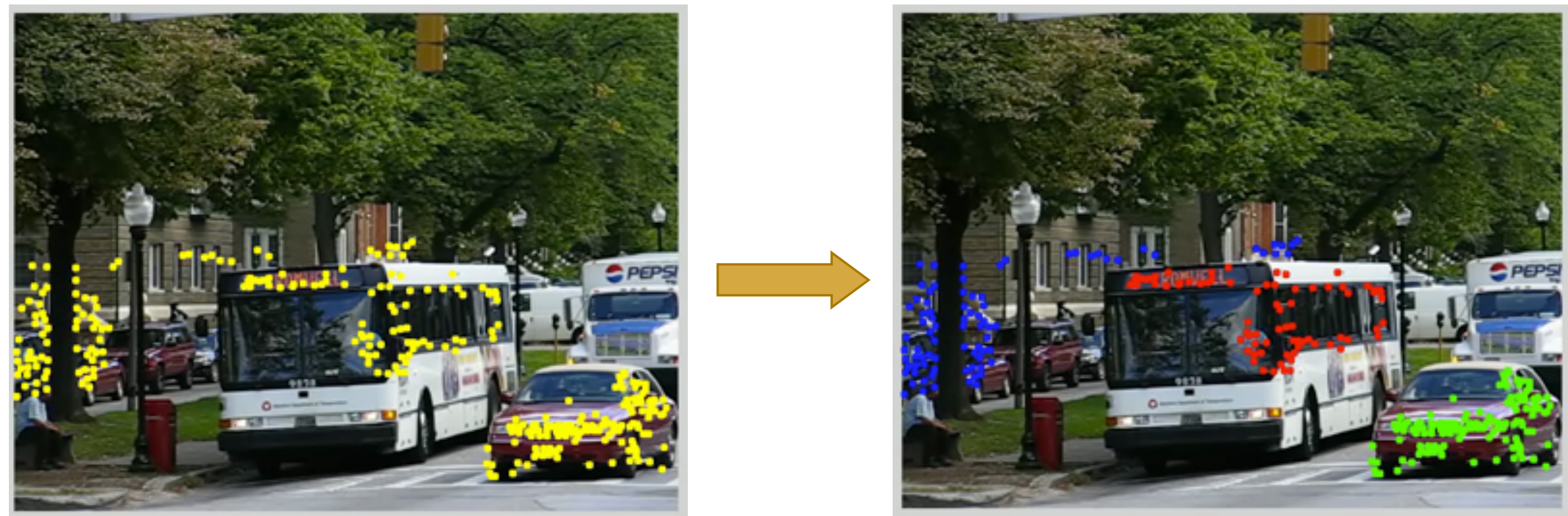


- Given feature trajectories of multiple rigid motions
- Find segmentation into underlying motions

Tomasi-Kanade'92, Sugaya-Kanatani'04, Vida-Ma-Sastry'05, Elhamifar-Vidal'09, Chen-Lerman'09, Zhao-Medioni'11

Experiments: motion segmentation

- Hopkins 155 dataset
  - 155 sequences
  - 2 and 3 motions



- Clustering errors

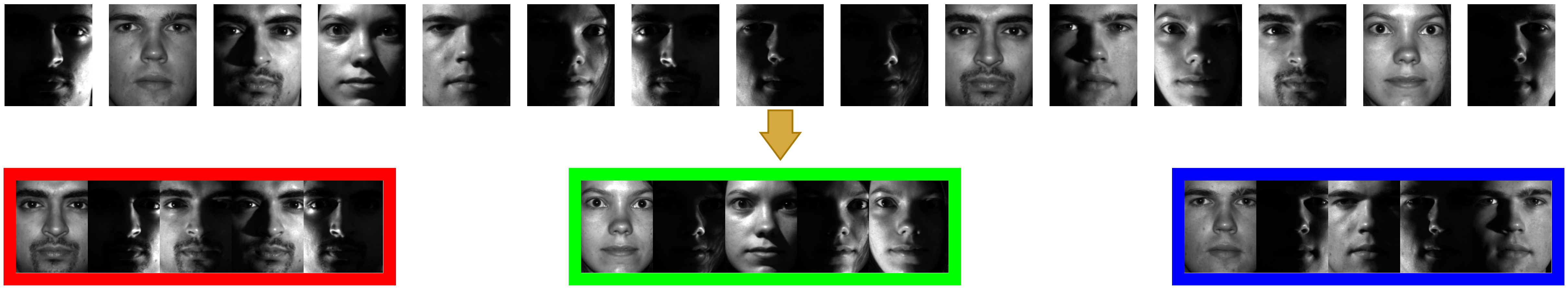
Algorithms	RANSAC	GPCA	MSL	LSA	SCC	LRR	LRSC	SSC
2 Motions								
Mean	5.56	4.59	4.14	4.23	2.89	4.10	3.69	**1.52**
Median	1.18	0.38	0.00	0.56	0.00	0.22	0.29	0.00
3 Motions								
Mean	22.94	28.66	8.23	7.02	8.25	9.89	7.69	**4.40**
Median	22.03	28.26	1.76	1.45	0.24	6.22	3.80	0.56

- nonconvex, local min
- k-NN based
- sensitive to noise
- exponential complexity

+ convex, provable
+ automatic selection
+ robust to noise
+ computationally efficient

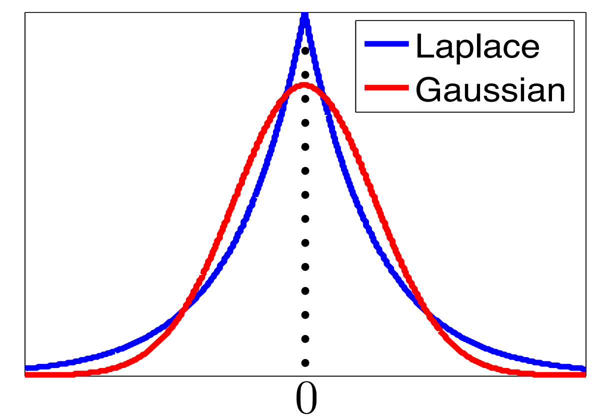
E. Elhamifar and R. Vidal, IEEE Trans. PAMI, 2013

Application: face clustering

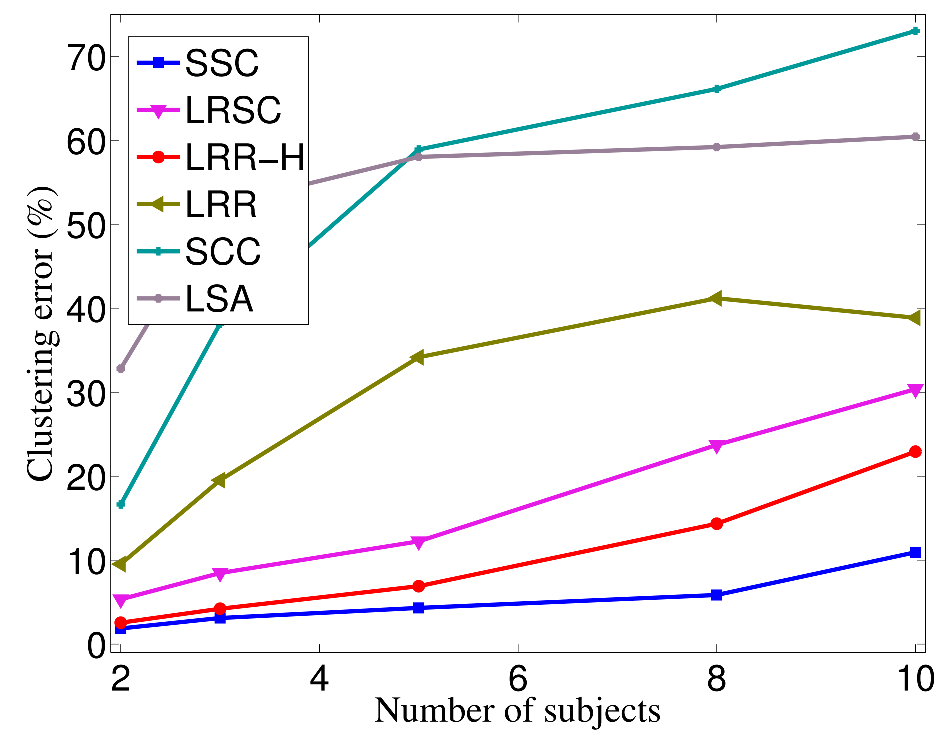


- Corruption by sparse errors $\tilde{\boldsymbol{y}}_i = \boldsymbol{y}_i + \boldsymbol{e}_i$

$$\min \lambda \|\boldsymbol{c}_i\|_1 + \|\tilde{\boldsymbol{y}}_i - \widetilde{\boldsymbol{Y}}\boldsymbol{c}_i\|_1 \quad \text{s.t.} \quad c_{ii} = 0$$



- SSC error on Ext YaleB faces
  - < **2.0%** for 2 subjects
  - < **11.0%** for 10 subjects



E. Elhamifar and R. Vidal, IEEE Trans. PAMI, 2013

Other extensions of SSC

- Extension to clustering and DR of nonlinear manifolds [Elhamifar-Vidal NIPS'11]

- Scaling to large datasets
  - Greedy algorithm, theory [Dyer-Sankaranarayanan-Baraniuk JMLR'13]
  - Sampling + more a compact dictionary [Peng-Zhang-Yi CVPR'13]

- Dealing with sequential and spatial data [Tierney-Gao-Guo CVPR'13, Pham et al CVPR'12]

- Enforcing block-diagonal structure on laplacian /adjacency [Feng-Lin et al CVPR'14]

- Connectivity of SSC graph [Nasihatkon-Hartley CVPR'11]

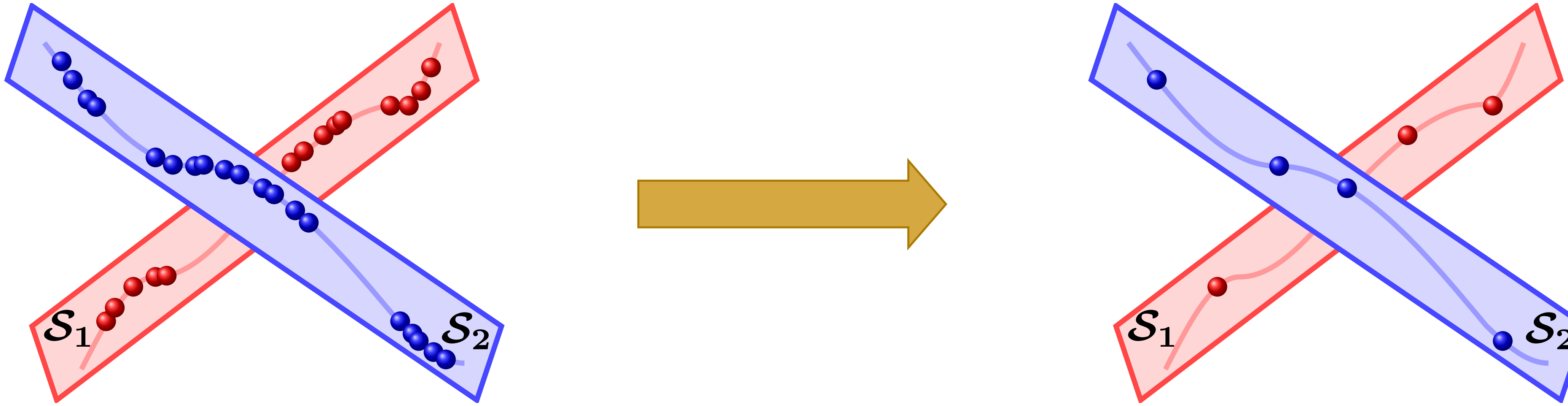
Conclusions

- Addressed **clustering** of data lying in **multiple subspaces**

- Proposed an **efficient algorithm** based on **sparse modeling**
  - Proved **theoretical guarantees** of the algorithm
  - Extended to deal with **corrupted data**
  - **Resolved challenges** of the state of the art
  - Showed it **performs well** in real-world problems

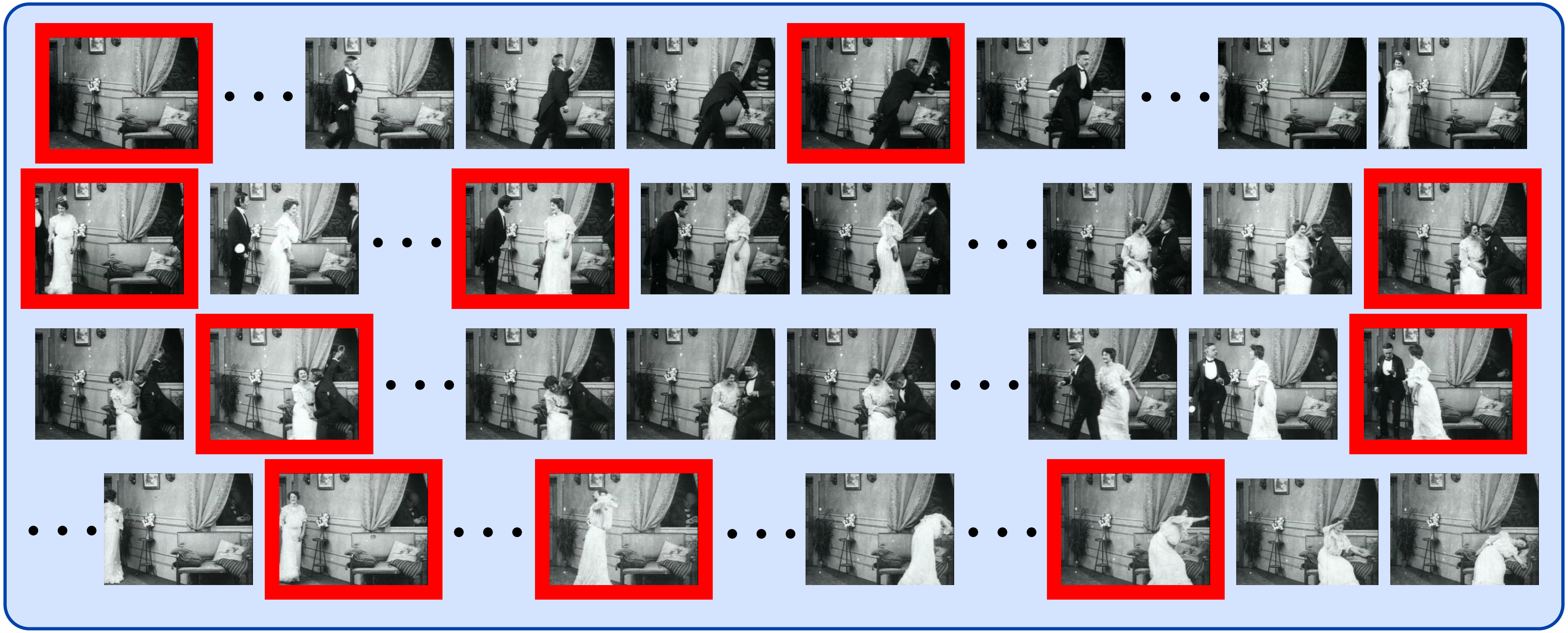
Sparse Subset Selection

Ehsan Elhamifar



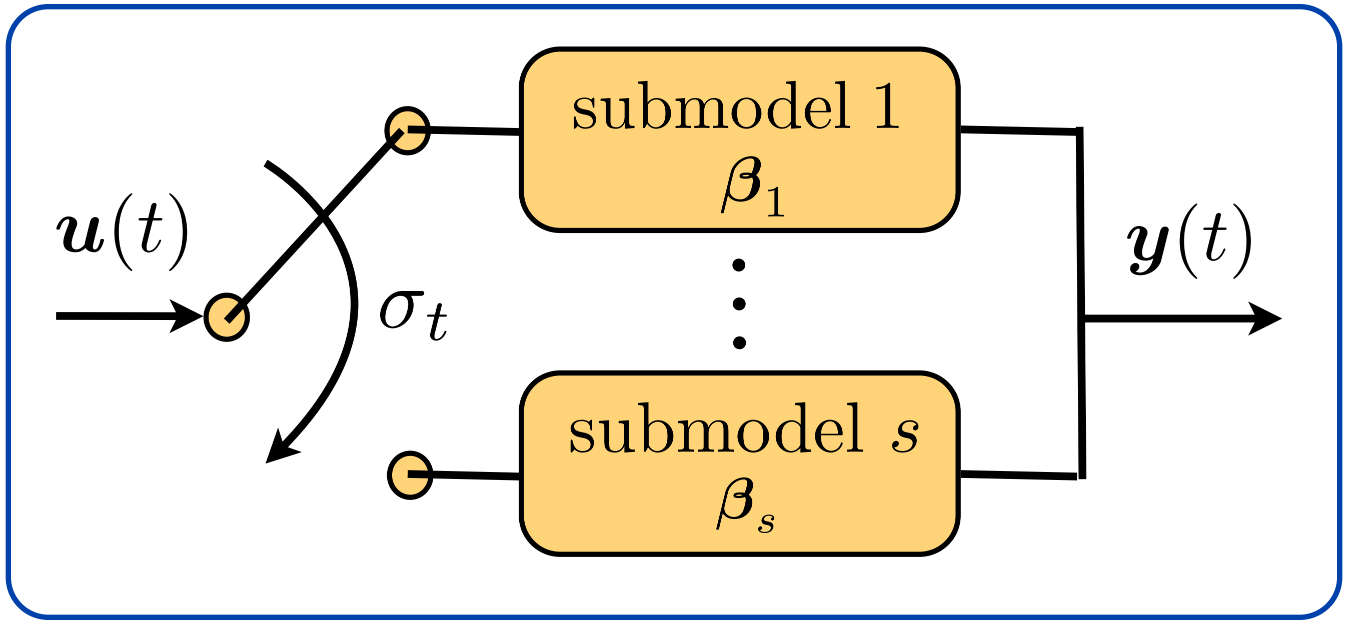
Finding representatives

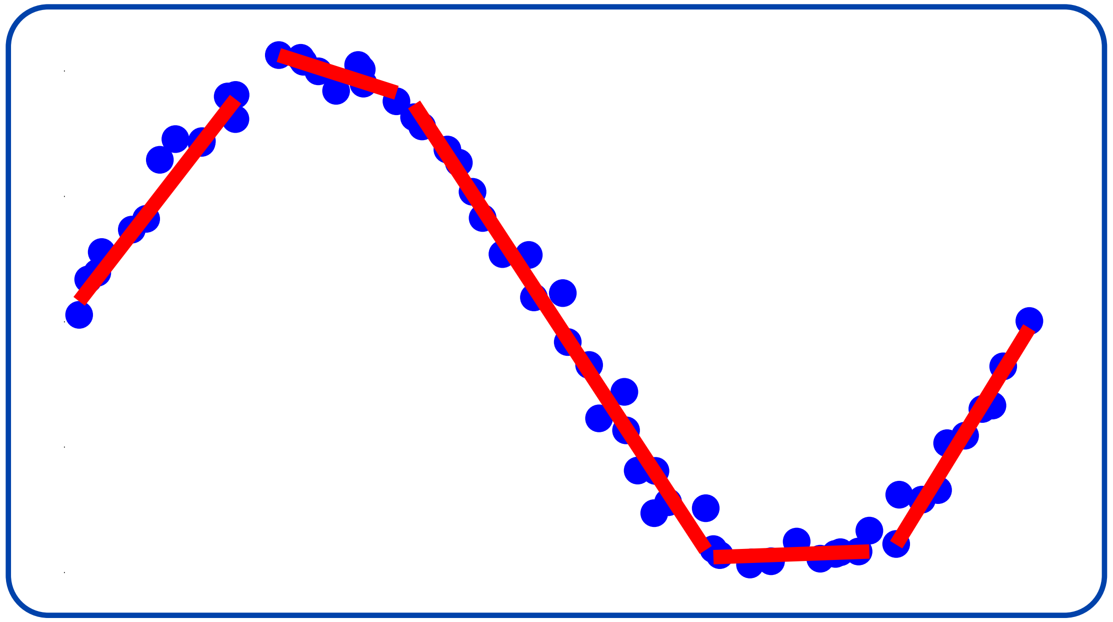
- A **subset** of data / models, **efficiently** representing the entire set
  - **Summarize** and **visualize** images/videos/text/web datasets



Finding representatives

- A **subset** of data / models, **efficiently** representing the entire set
  - **Summarize** and **visualize** images/videos/text/web datasets
  - Improve **computational time** and **memory**
  - Describe (complex) **nonlinear models**

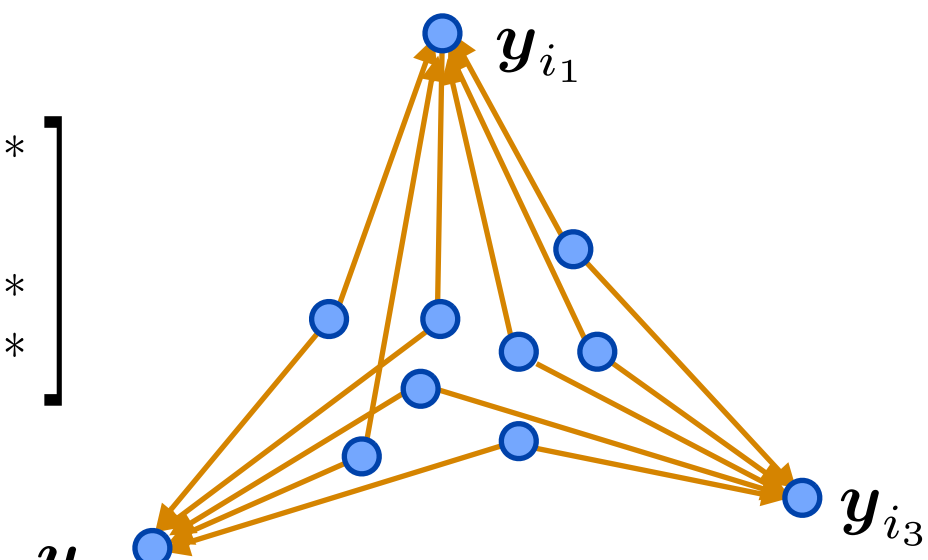




Column subset selection

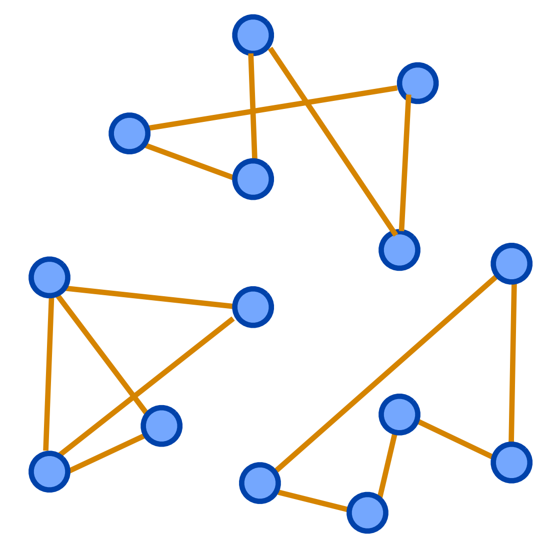
- Given $\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_N \in \mathbb{R}^n$, select a subset $\{\boldsymbol{y}_{i_1}, \ldots, \boldsymbol{y}_{i_k}\}$ that "well represent" the dataset

$$\underset{\boldsymbol{C}}{\operatorname{argmin}} \ \lambda \sum_{i=1}^{N} \|\boldsymbol{C}_{i*}\|_p + \frac{1}{2}\|\boldsymbol{Y} - \boldsymbol{Y}\boldsymbol{C}\|_F^2 \quad \text{s.t.} \quad \mathbf{1}^\top \boldsymbol{C} = \mathbf{1}^\top \qquad = \begin{bmatrix} * & * & & * & * & * & & * \\ & & & & & & & \\ * & & * & & * & * & * & * \\ & * & * & * & * & * & & * \end{bmatrix}$$



Representatives

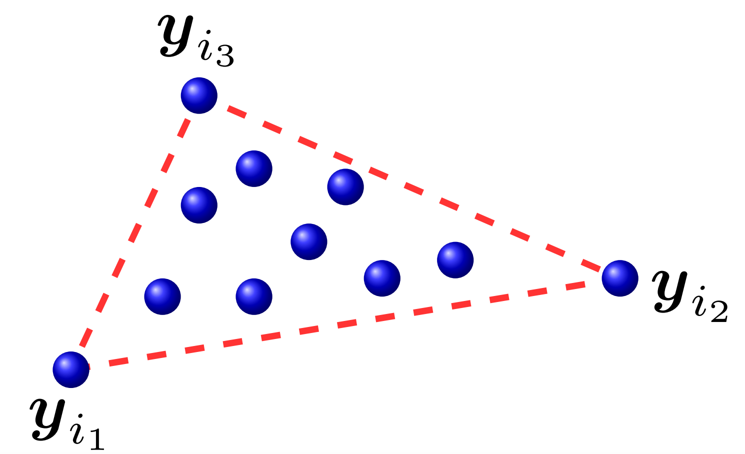
$$\underset{\boldsymbol{C}}{\operatorname{argmin}} \ \lambda \|\boldsymbol{C}\|_1 + \frac{1}{2}\|\boldsymbol{Y} - \boldsymbol{Y}\boldsymbol{C}\|_F^2 \quad = \begin{bmatrix} * & & * & & & & * \\ & & & & * & & \\ & & * & & & * & \\ & * & & & & * & \\ & & & * & & & \\ * & & & & & & \\ & & & & * & & * \\ & * & * & & & & \\ & & & * & & & \end{bmatrix}$$


SSC via L1 graph

E. Elhamifar, G. Sapiro and R. Vidal, CVPR 2012

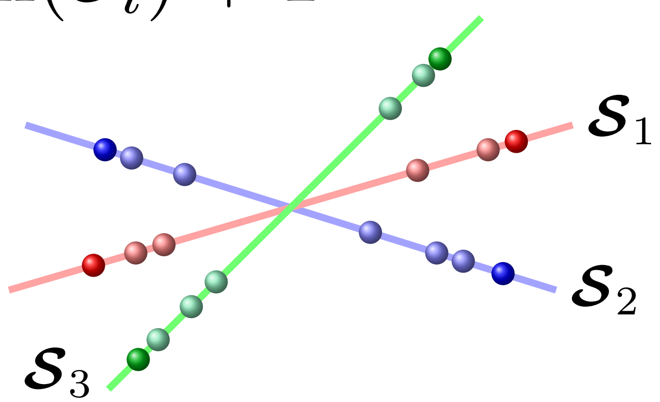
Column subset selection: theory

- **Theorem**: Let $\mathcal{H}$ be the convex hull of the columns of $\boldsymbol{Y}$ with k vertices. Assume the columns of $\boldsymbol{Y}$ lie in a (k-1)-dim. affine subspace. For p > 1, we obtain k representatives, corresponding to the vertices of $\mathcal{H}$.



$$\boldsymbol{C}^* = \boldsymbol{\Gamma} \begin{bmatrix} \boldsymbol{I}_k & \boldsymbol{\Delta} \\ \boldsymbol{0} & \boldsymbol{0} \end{bmatrix} \qquad \boldsymbol{\Delta} \in [0,1)^k$$

- **Theorem**: For points lying in a union of independent subspaces ($\dim(\oplus_i \mathcal{S}_i) = \sum_i \dim(\mathcal{S}_i)$), we obtain at least $\dim(\mathcal{S}_i) + 1$ representatives from each $\mathcal{S}_i$.



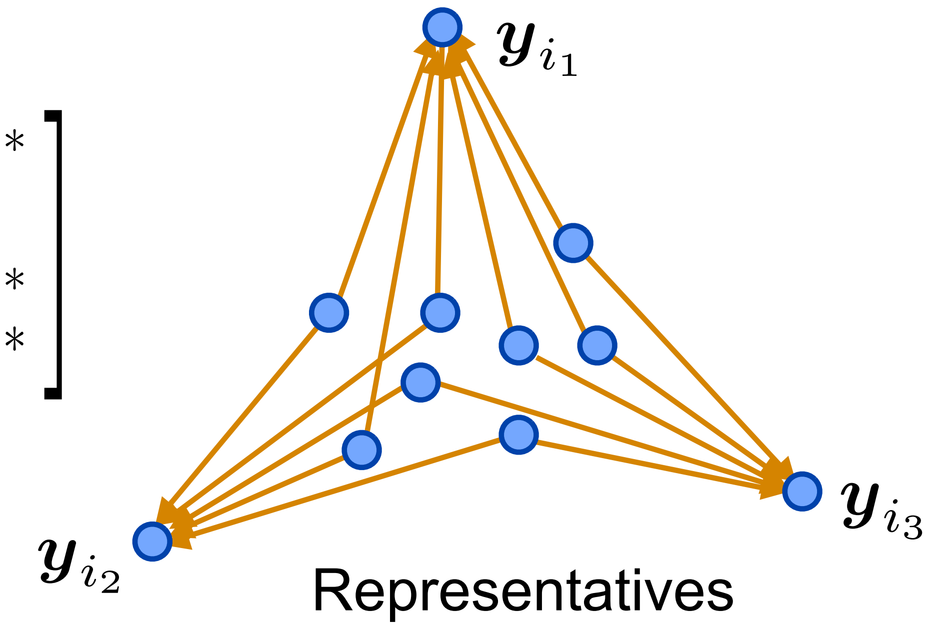
E. Elhamifar, G. Sapiro and R. Vidal, CVPR 2012

Column subset selection

- Given $\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_N \in \mathbb{R}^n$, select a subset $\{\boldsymbol{y}_{i_1}, \ldots, \boldsymbol{y}_{i_k}\}$ that "well represent" the dataset

$$\underset{\boldsymbol{C}}{\operatorname{argmin}} \ \lambda \sum_{i=1}^{N} \|\boldsymbol{C}_{i*}\|_p + \frac{1}{2}\|\boldsymbol{Y} - \boldsymbol{Y}\boldsymbol{C}\|_F^2 = \begin{bmatrix} * * & * * * & * \\ & & \\ * \ \ * & * * * * & * \\ * * * * * & & * \end{bmatrix}$$

$$\text{s.t.} \quad \mathbf{1}^\top \boldsymbol{C} = \mathbf{1}^\top$$



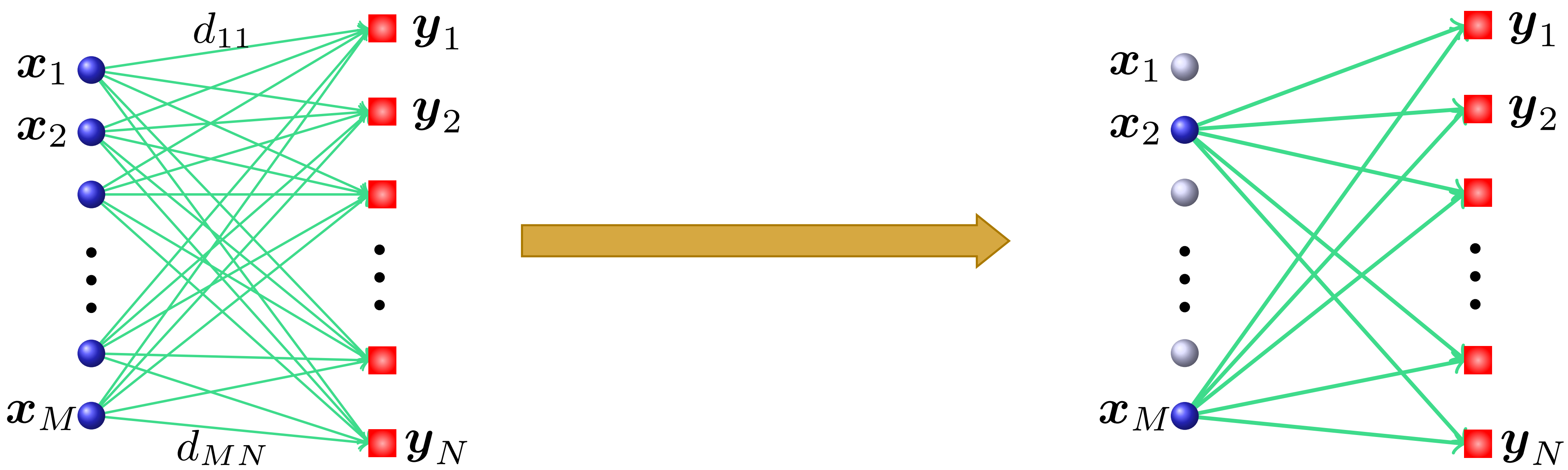
- What if data not in low-dim subspaces?
- What if no feature representation? e.g, social network graph
- What about summarization between two / multiple sets?

E. Elhamifar, G. Sapiro and R. Vidal, CVPR 2012

Subset selection using dissimilarities

- Given: dissimilarities $d : \mathbb{X} \times \mathbb{Y} \longrightarrow \mathbb{R}^{\geq 0}$

  source (↳ $\mathbb{X}$)   target (↳ $\mathbb{Y}$)

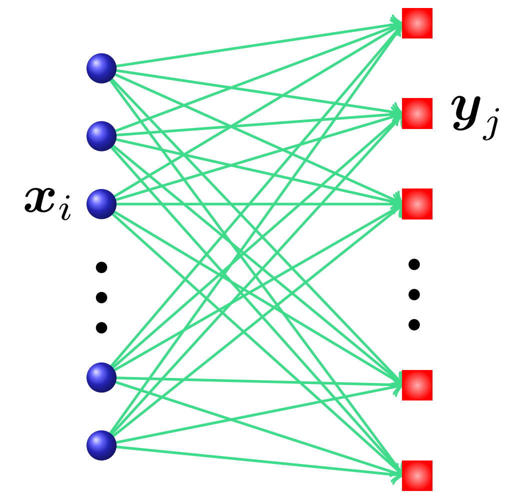


- Goal: select **a small subset** of $\mathbb{X}$ that **well represent** $\mathbb{Y}$ w.r.t. $d(\cdot,\cdot)$

- $d(\boldsymbol{x}_i, \boldsymbol{y}_j) = d_{ij}$: how well $\boldsymbol{x}_i$ represents $\boldsymbol{y}_j$
  - $\mathbb{X}$ = models, $\mathbb{Y}$ = data ⟶ $d(\cdot,\cdot)$ = representation/coding error
  - $\mathbb{X}$ = data, $\mathbb{Y}$ = data ⟶ $d(\cdot,\cdot)$ = Euclidean/ geodesic distance

Dissimilarity-based sparse subset selection (DS3)

- Let $\boldsymbol{D} = [d_{ij}]$, introduce $\boldsymbol{Z} = [z_{ij}]$
  - $z_{ij} = P(\boldsymbol{x}_i \ \text{rep.} \ \boldsymbol{y}_j)$



- To select few elements of $\mathbb{X}$ that well represent $\mathbb{Y}$, minimize

1) **Encoding** of $\mathbb{Y}$ via representatives

$$\sum_{i=1}^{M} \sum_{j=1}^{N} d_{ij} z_{ij} = \operatorname{tr}(\boldsymbol{D}^\top \boldsymbol{Z})$$

2) **Number** of representatives

$$\sum_{i=1}^{M} \mathrm{I}\left(\|\boldsymbol{Z}_{i*}\|_p\right)$$

- Solve the simultaneous sparse recovery program

$$\min_{\boldsymbol{Z}} \ \lambda \sum_{i=1}^{M} \|\boldsymbol{Z}_{i*}\|_p + \operatorname{tr}(\boldsymbol{D}^\top \boldsymbol{Z}) \quad \text{s.t.} \quad \boldsymbol{Z} \geq \boldsymbol{0}, \ \boldsymbol{1}^\top \boldsymbol{Z} = \boldsymbol{1}^\top$$

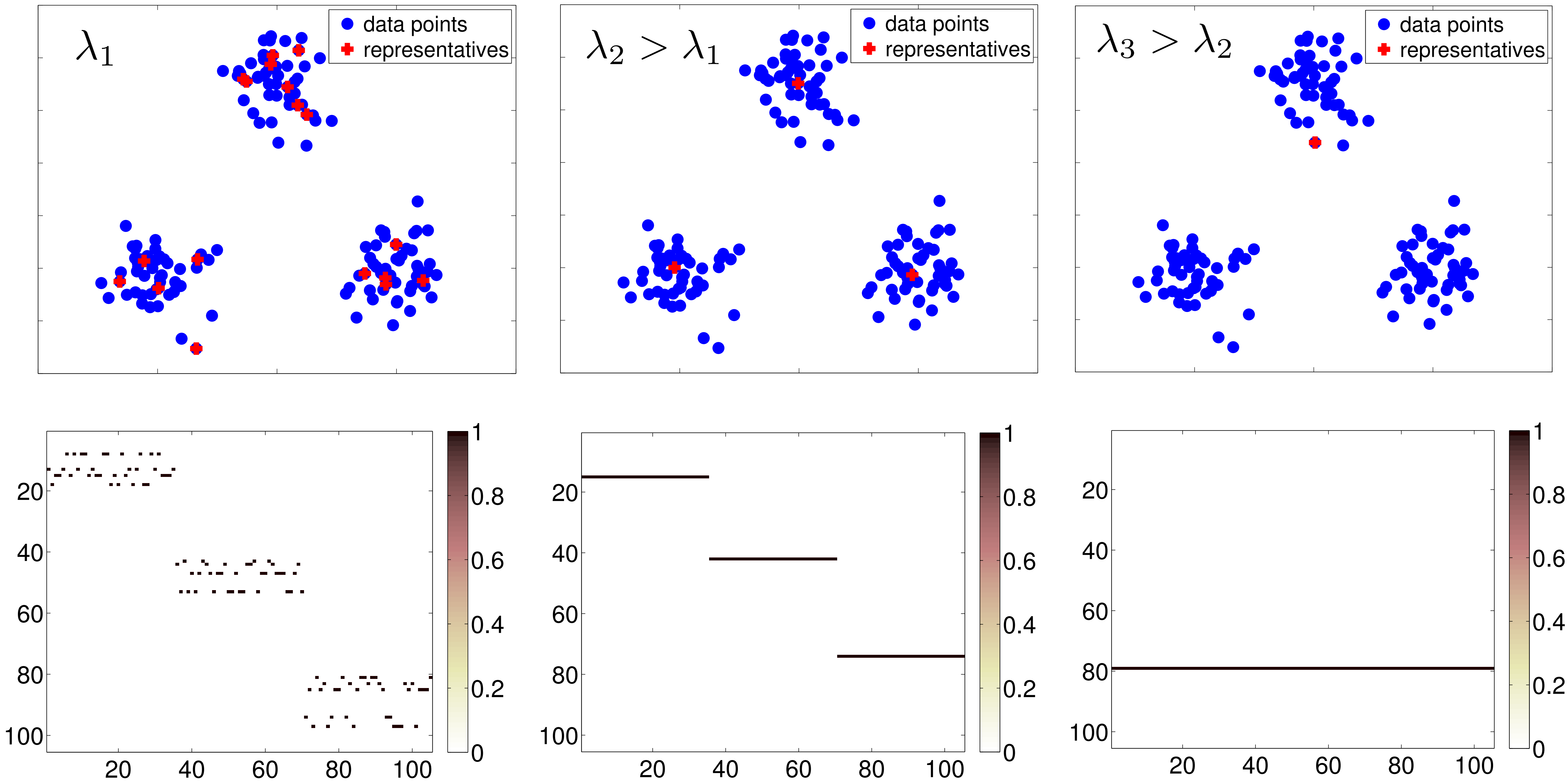
Convex

$p \in \{2, \infty\}$

E. Elhamifar, G. Sapiro and R. Vidal, NIPS, 2012; E. Elhamifar, G. Sapiro and S. Sastry, IEEE Trans. PAMI, 2015

Dissimilarity-based sparse subset selection (DS3)

- Identical source and target



Theoretical analysis

$$\min_{\boldsymbol{Z}} \; \lambda \sum_{i=1}^{M} \|\boldsymbol{Z}_{i*}\|_p + \mathrm{tr}(\boldsymbol{D}^\top \boldsymbol{Z}) \quad \mathrm{s.\,t.} \quad \boldsymbol{Z} \geq \boldsymbol{0},\; \boldsymbol{1}^\top \boldsymbol{Z} = \boldsymbol{1}^\top$$

- **Proposition 1**: Assume $\mathbb{X}$ and $\mathbb{Y}$ are identical. If λ is sufficiently large, only one representative is selected. If λ is sufficiently small, each point chooses itself as a representative.
  - $\lambda \geq \lambda_{\max,p}(\boldsymbol{D})$ ⟹ $\boldsymbol{Z} = \boldsymbol{e}_\ell \boldsymbol{1}^\top$, where $\ell = \operatorname{argmin}_i \boldsymbol{1}^\top \boldsymbol{D}_{i*}$
  - $\lambda \leq \lambda_{\min,p}(\boldsymbol{D})$ ⟹ $\boldsymbol{Z} = \boldsymbol{I}$

- We determine $[\lambda_{\min,p}(\boldsymbol{D}) \,,\, \lambda_{\max,p}(\boldsymbol{D})]$ to set the regularization

e.g., $\lambda_{\min,2}(\boldsymbol{D}) = \min_j \left(\min_{i \neq j} d_{ij} - d_{jj} \right) \,,\; \lambda_{\max,2}(\boldsymbol{D}) = \max_{i \neq \ell} \frac{\sqrt{N}}{2} \frac{\|\boldsymbol{D}_{i*} - \boldsymbol{D}_{\ell*}\|_2^2}{\boldsymbol{1}^\top (\boldsymbol{D}_{i*} - \boldsymbol{D}_{\ell*})}$

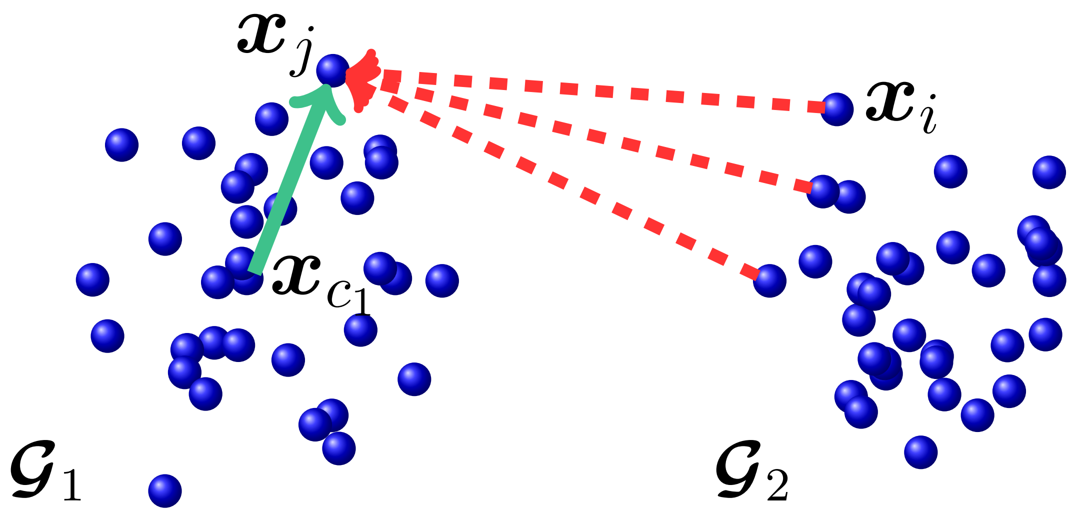
E. Elhamifar, G. Sapiro and S. Sastry, IEEE Trans. PAMI, 2015

Theoretical analysis

$$\min_{\boldsymbol{Z}} \; \lambda \sum_{i=1}^{M} \|\boldsymbol{Z}_{i*}\|_p + \mathrm{tr}(\boldsymbol{D}^\top \boldsymbol{Z}) \quad \mathrm{s.t.} \quad \boldsymbol{Z} \geq \boldsymbol{0}, \; \boldsymbol{1}^\top \boldsymbol{Z} = \boldsymbol{1}^\top$$

- **Proposition 2**: Assume $\mathbb{X}$ and $\mathbb{Y}$ are identical. Assume points partition into L groups. If $\lambda \leq \lambda_g(\boldsymbol{D})$, the optimal $\boldsymbol{Z}$ is such that
  - (1) each group will have representatives;
  - (2) points in each group select representatives from that group only.

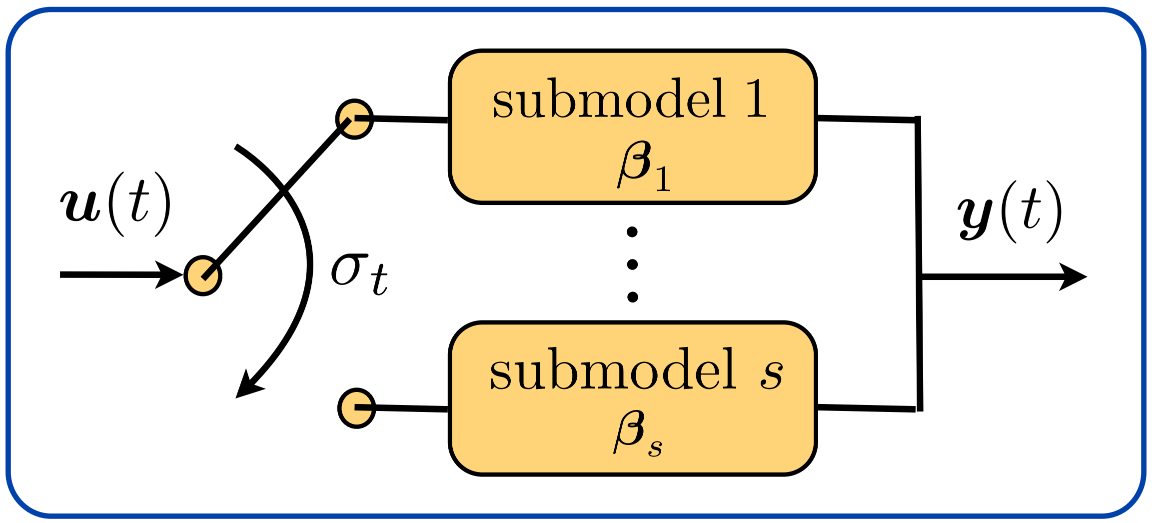
$$\lambda_g(\boldsymbol{D}) = \min_{k} \min_{j \in \mathcal{G}_k} \left(\min_{k' \neq k} \min_{i \in \mathcal{G}_{k'}} d_{ij} - d_{c_k j} \right)$$

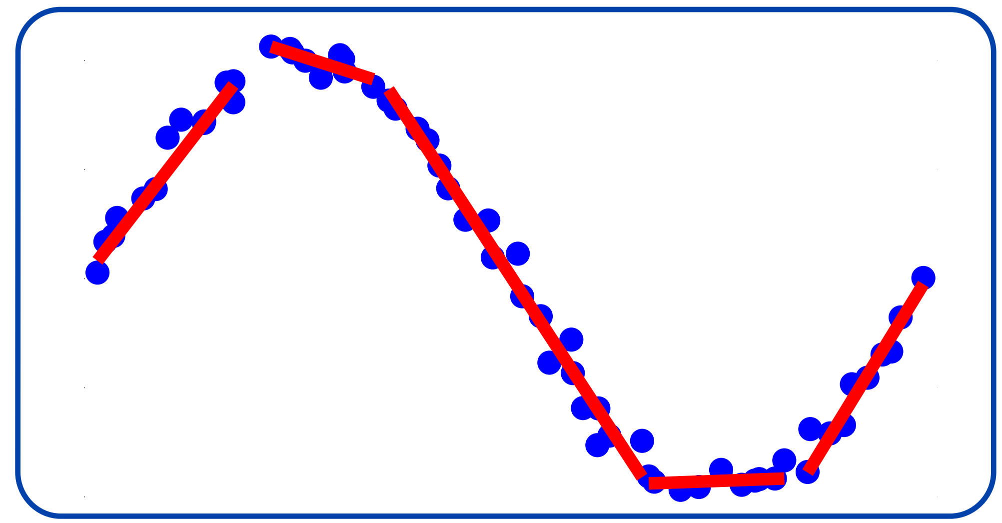


E. Elhamifar, G. Sapiro and S. Sastry, IEEE Trans. PAMI, 2015

DS3 applications: Learning nonlinear models

- **Nonlinear** dynamical systems as **switched linear models**
  - Human gaits / activities, motor control systems, ...





- Learning switched dynamical models: **Non-convex** & **NP-hard**

Our **convex** solution

$$\mathbb{X} = \{\hat{\boldsymbol{\beta}}_1, \ldots, \hat{\boldsymbol{\beta}}_M\}$$ = ensemble of models

$$\mathbb{Y} = \{(\boldsymbol{u}(1), \boldsymbol{y}(1)), \ldots, (\boldsymbol{u}(N), \boldsymbol{y}(N))\}$$

E. Elhamifar, G. Sapiro and S. Sastry, IEEE Trans. PAMI, 2015; E. Elhamifar, S. Burden and S. Sastry, IFAC, 2014

DS3 applications: Learning nonlinear models

- Experiments on segmentation of CMU motion capture data

  - Discrete-time SS model via subspace ID, snippets of length 100
  - d_{ij} = Euclidean norm of representation error of j-th snippet via i-th estimated submodel

Sequence number	1	2	3	4	5	6	7	8	9	10	11
# activities	4	8	7	7	7	10	6	9	4	4	7
SC error (%)	23.86	30.61	19.02	40.60	26.43	47.77	14.85	38.09	**9.02**	8.31	**3.47**
SBiC error (%)	22.77	22.08	18.94	28.40	29.85	30.96	30.50	24.78	13.03	12.68	23.68
Kmedoids error (%)	18.26	46.26	49.89	51.99	37.07	54.75	29.81	49.53	9.71	33.50	33.80
AP error (%)	22.93	41.22	49.66	54.56	37.87	50.19	37.84	48.37	9.71	26.05	23.84
DS3 error (%)	**5.33**	**9.90**	**12.27**	**19.64**	**16.55**	**14.66**	**12.56**	**11.73**	11.18	**3.32**	6.18

E. Elhamifar, G. Sapiro and S. Sastry, IEEE Trans. PAMI, 2015

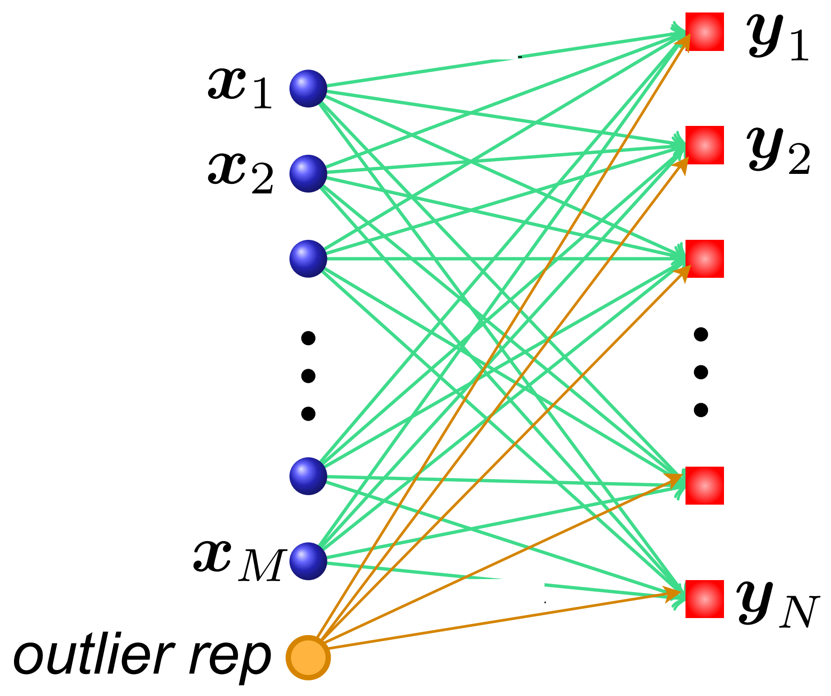
Dealing with outliers via DS3

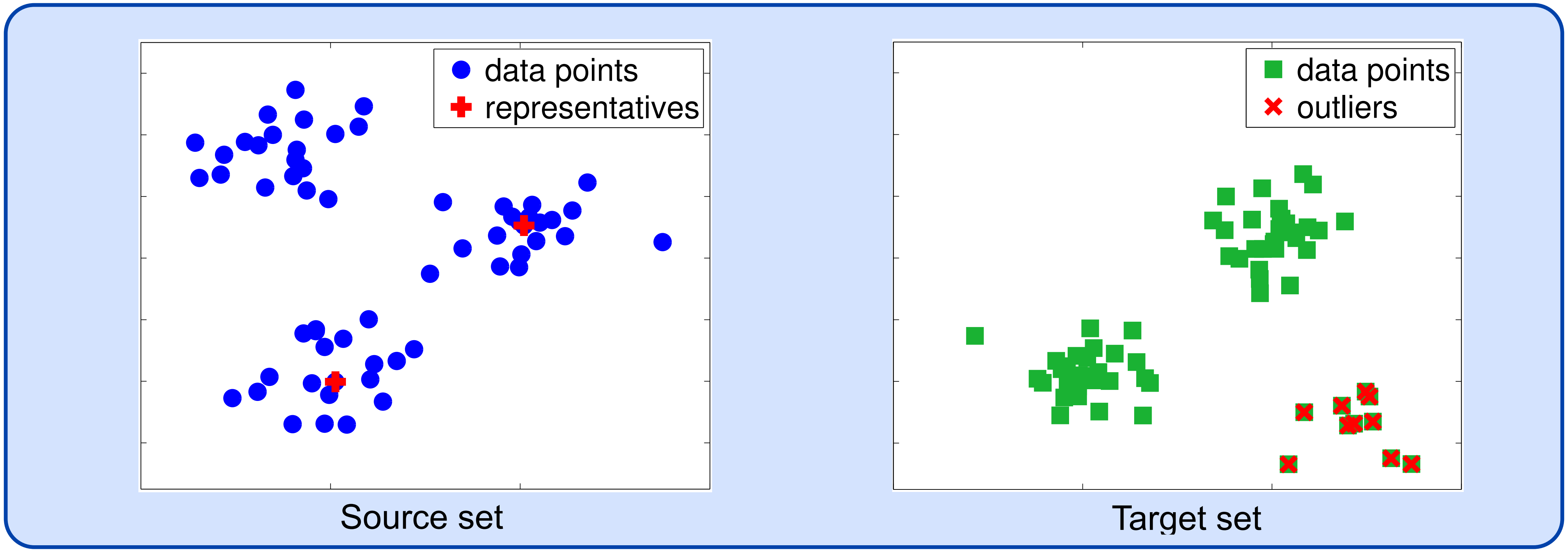
- Add **outlier representative** node to source

$$e_j = P\left(\text{outlier node} \leftarrow \boldsymbol{y}_j\right)$$

- Solve the optimization

$$\min_{\boldsymbol{Z},\boldsymbol{e}} \; \lambda \sum_{i=1}^{M} \|\boldsymbol{Z}_{i*}\|_p + \operatorname{tr}\left(\begin{bmatrix} \boldsymbol{D} \\ \boldsymbol{d}_o \end{bmatrix}^\top \begin{bmatrix} \boldsymbol{Z} \\ \boldsymbol{e}^\top \end{bmatrix} \right)$$
$$\text{s.t.} \quad \boldsymbol{1}^\top \begin{bmatrix} \boldsymbol{Z} \\ \boldsymbol{e}^\top \end{bmatrix} = \boldsymbol{1}^\top, \quad \begin{bmatrix} \boldsymbol{Z} \\ \boldsymbol{e}^\top \end{bmatrix} \geq \boldsymbol{0}$$

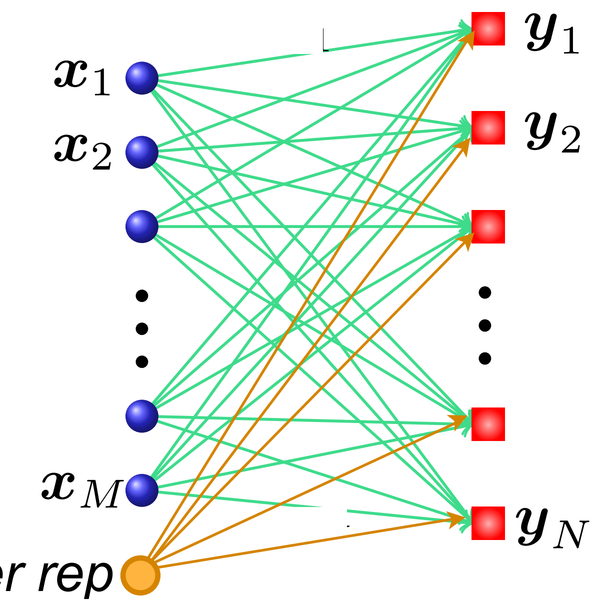




E. Elhamifar, G. Sapiro and S. Sastry, IEEE Trans. PAMI, 2015

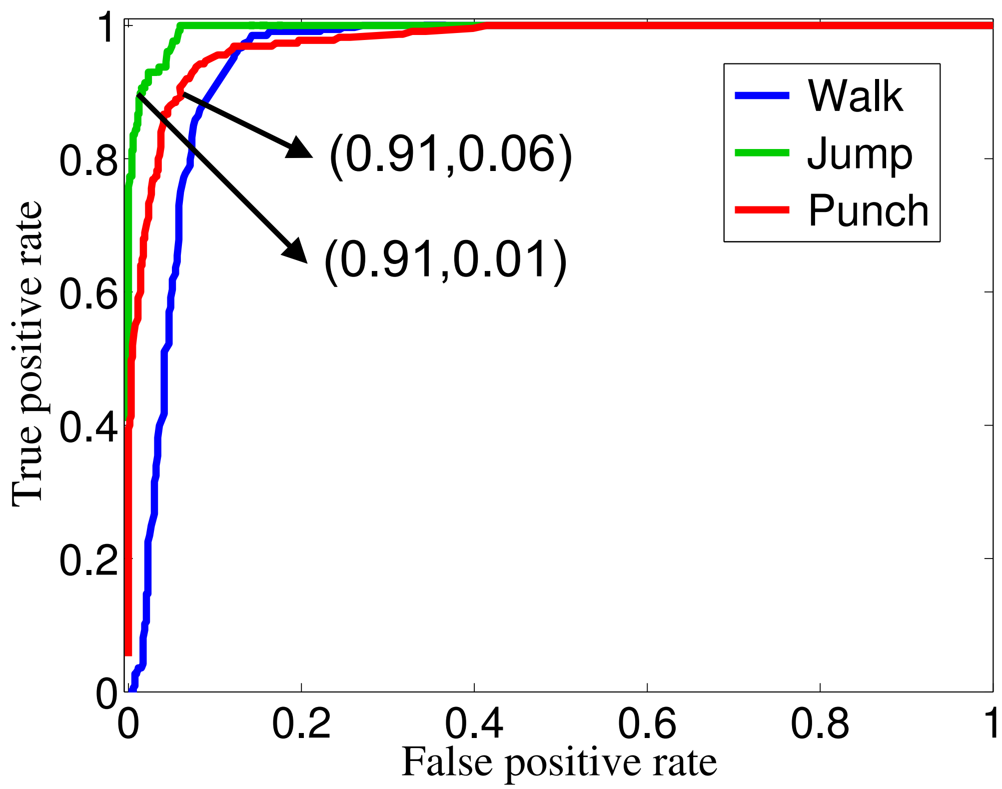
Dealing with outliers via DS3: experiment

- Exclude one activity when estimating LDS ensemble

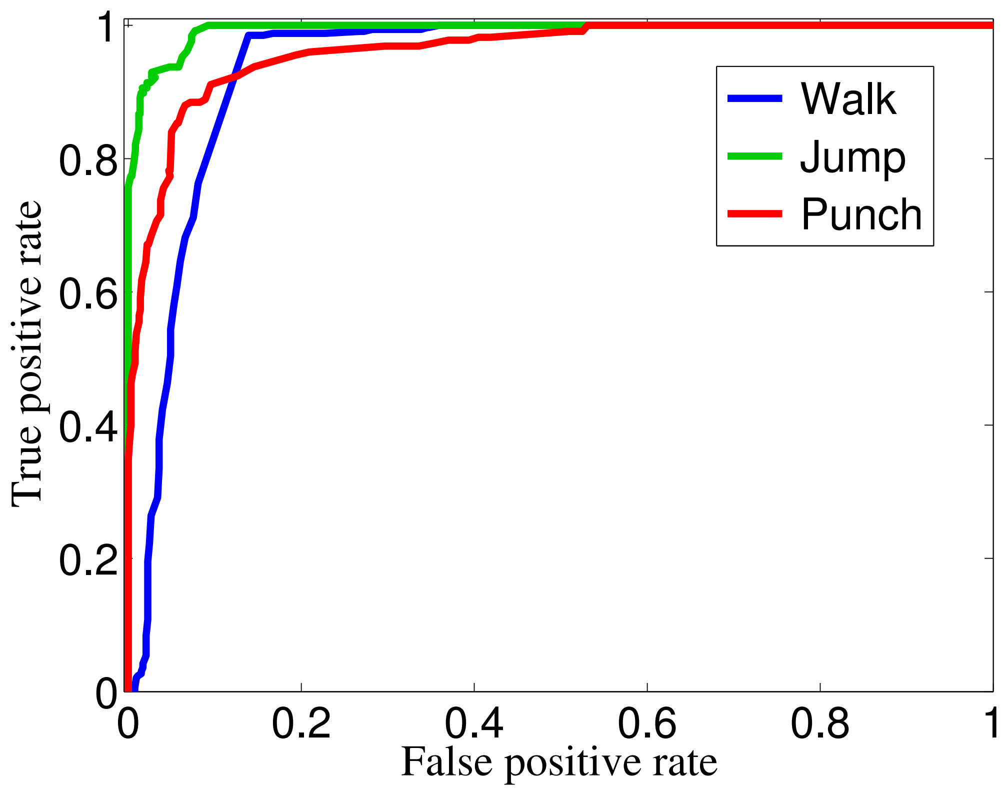


- Set outlier node weights $w_j = \beta\, e^{-\frac{\min_i d_{ij}}{\tau}}$

$\tau = 0.1$



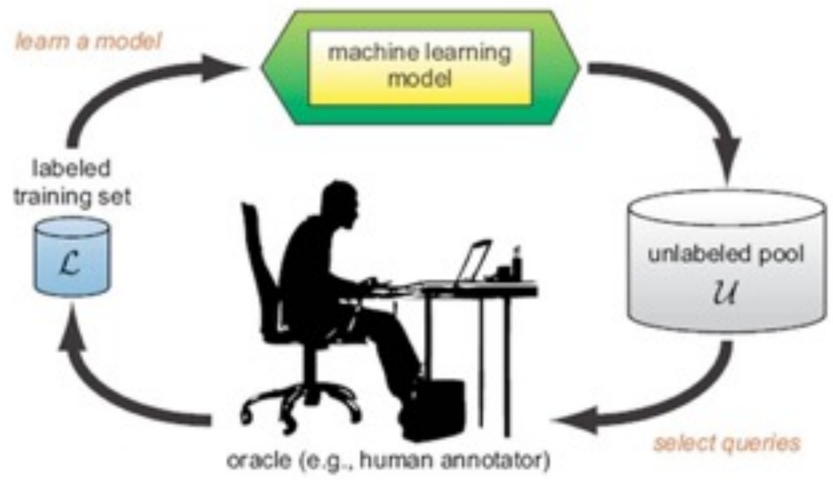
$\tau = 1.0$

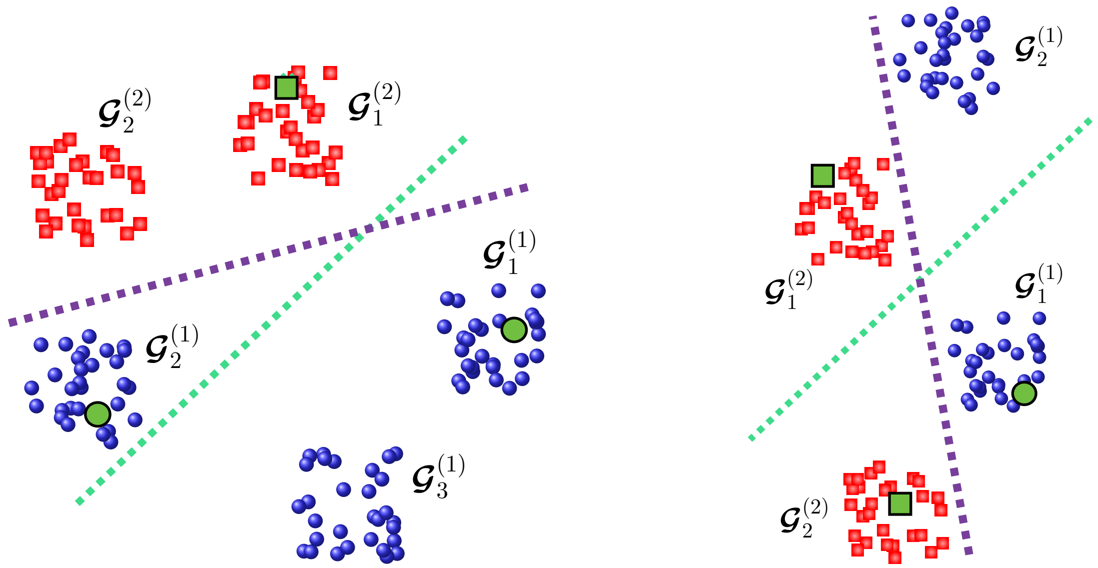


E. Elhamifar, G. Sapiro and S. Sastry, IEEE Trans. PAMI, 2015

DS3 applications: Active learning

- Successively annotate the most informative unlabeled samples





- For $\mathbb{X}$ = $\mathbb{Y}$ = { unlabeled samples }, solve

$$\min_{\boldsymbol{Z}} \ \lambda \|\boldsymbol{W}\boldsymbol{Z}\|_{1,p} + \operatorname{tr}(\boldsymbol{D}^\top \boldsymbol{Z}) \ \text{ s.t. } \ \boldsymbol{Z} \geq \boldsymbol{0}, \ \boldsymbol{1}^\top \boldsymbol{Z} = \boldsymbol{1}^\top \qquad \boldsymbol{W} \triangleq \operatorname{diag}(w_1, w_2, \ldots)$$

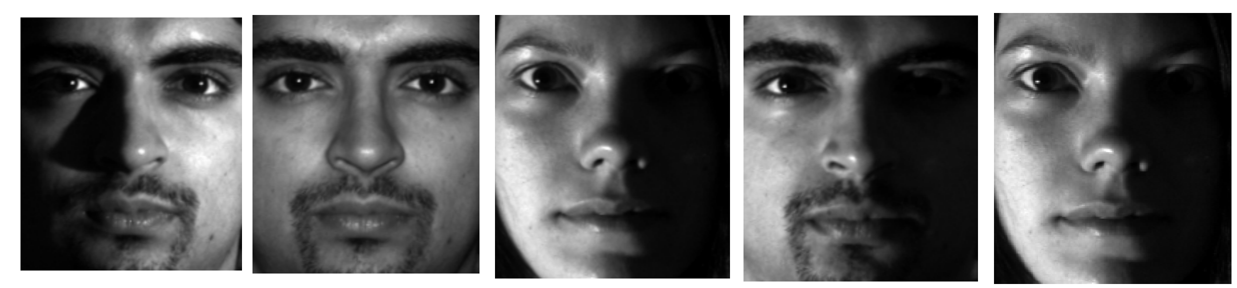
$$w_i \triangleq \min\{\sigma - (\sigma - 1)\frac{E(\boldsymbol{p}_i)}{\log_2(L)}, \ \sigma - (\sigma - 1)\frac{\min_{j\in\mathcal{L}} d_{ji}}{\max_{k\in\mathcal{U}} \min_{j\in\mathcal{L}} d_{jk}}\}$$

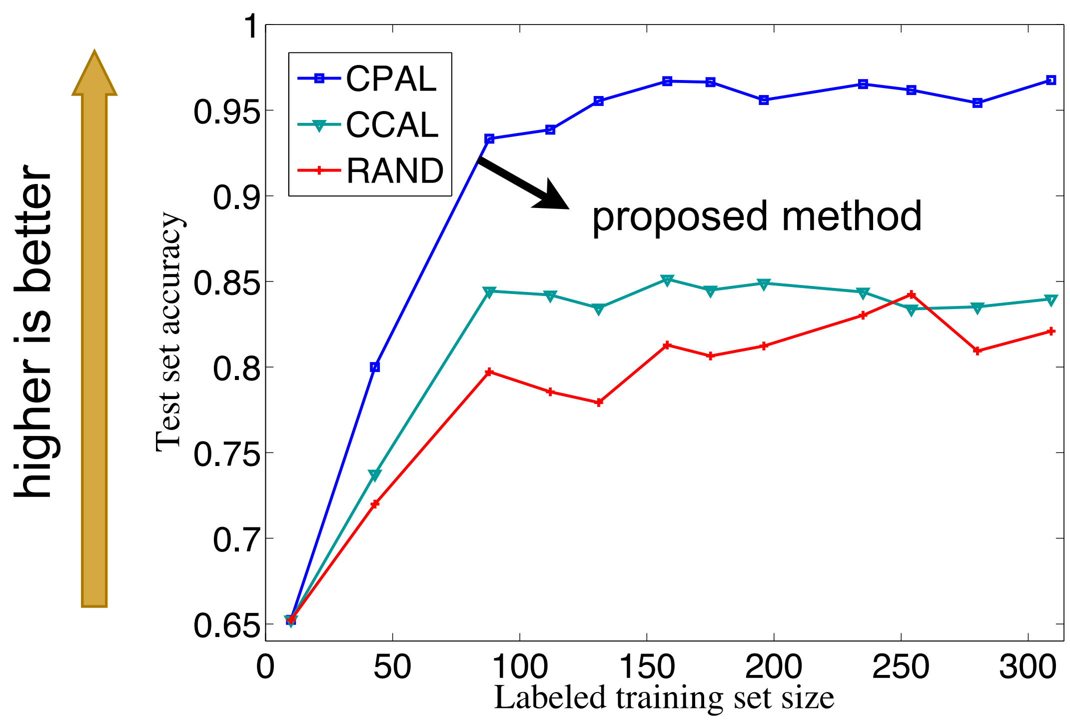
classifier uncertainty confidence ↗ ↖ sample diversity confidence

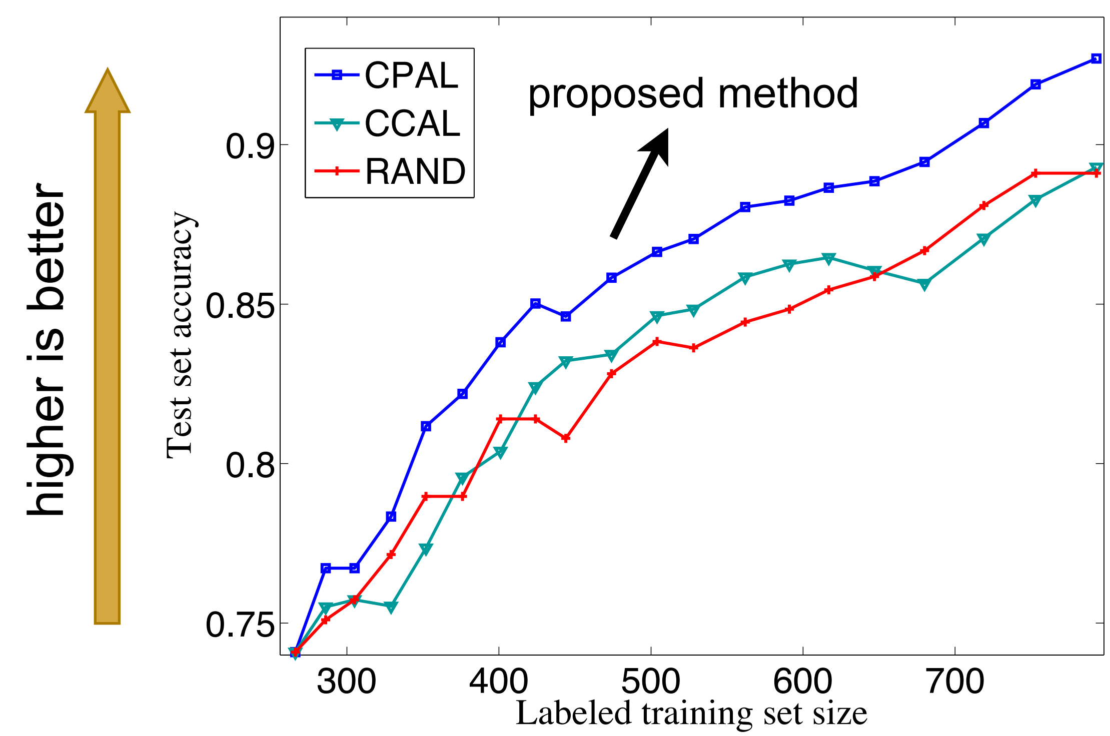
E. Elhamifar, G. Sapiro, A. Yang and S. Sastry, ICCV, 2013

DS3 applications: Active learning

- Pedestrian vs non-pedestrian
  - classifier: SVM
  - dissimilarity: HOG χ^2 distances
- Face recognition
  - classifier: SRC
  - dissimilarity: Euclidean dist.







E. Elhamifar, G. Sapiro, A. Yang and S. Sastry, ICCV, 2013

Conclusions

- Studied the problem of **subset selection**
  - Feature-space representations
  - Pairwise similarities
- Proposed convex programs using **simultaneous sparse recovery**
  - Extended to **deal with outliers**
- Proved the solution recovers **representatives from each group** and **correctly clusters** data
- Addressed several problems effectively
  - **Active learning**
  - **Learning nonlinear dynamical models**
  - **Segmentation of time-series data**

Thanks!

Codes: http://www.eecs.berkeley.edu/~ehsan.elhamifar/code.htm

Collaborators:

Shankar Sastry (UCB)

Emmanuel Candes (Stanford)

Allen Yang (UCB)

Mahdi Soltanolkotabi (USC)

Guillermo Sapiro (Duke)

Rene Vidal (JHU)

Sam Burden (UCB)