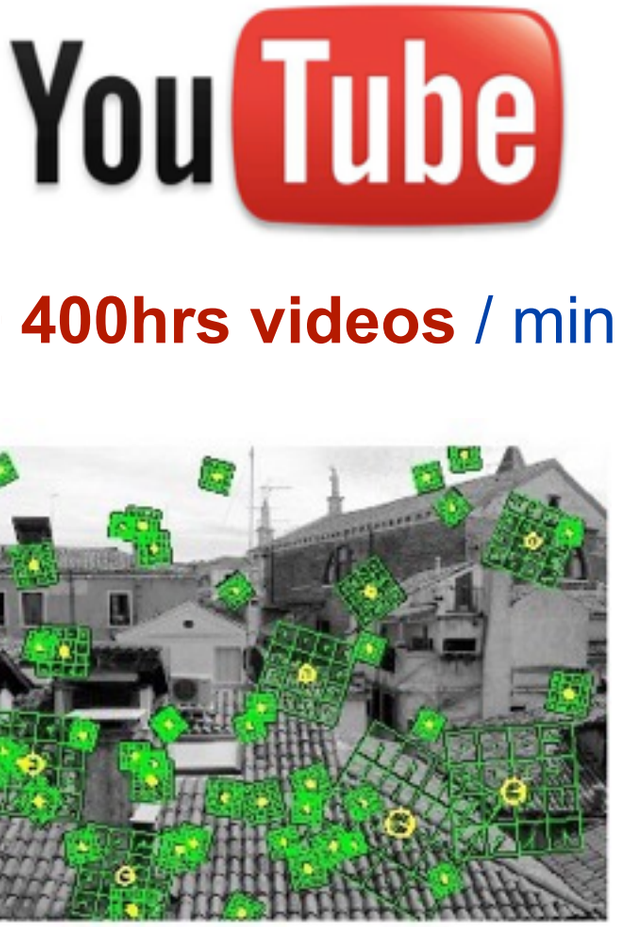
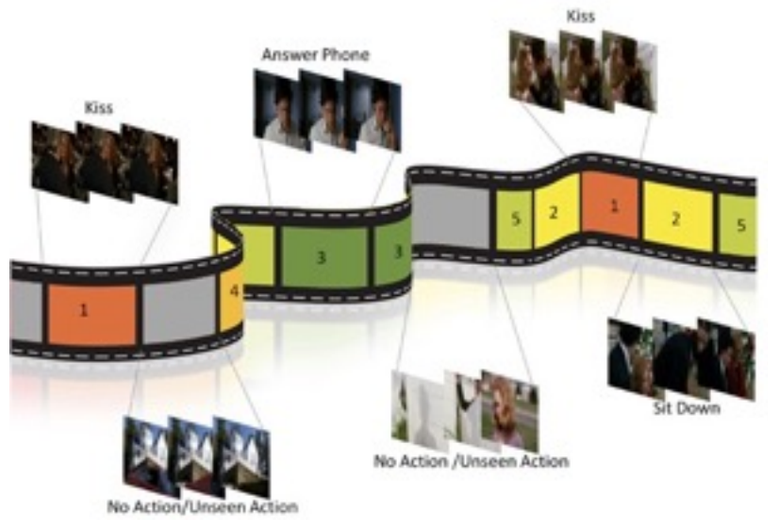
Summarization in Sequential Data with Applications to Procedure Learning

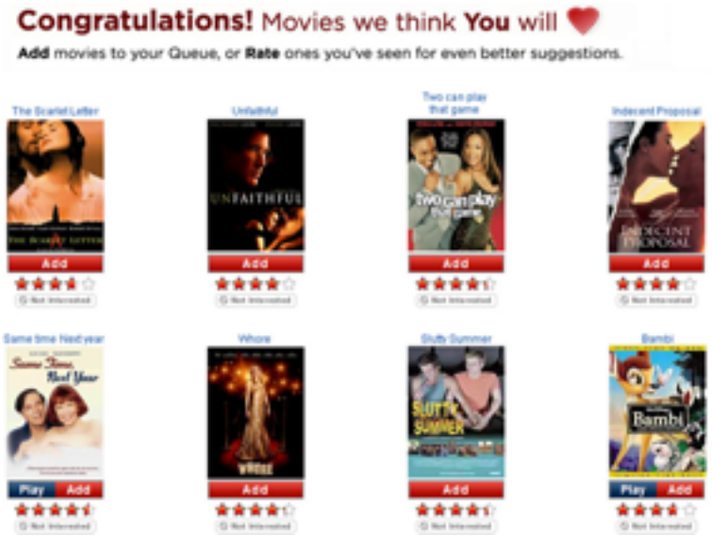
Ehsan Elhamifar

College of Computer and Information Science
Northeastern University

Data deluge







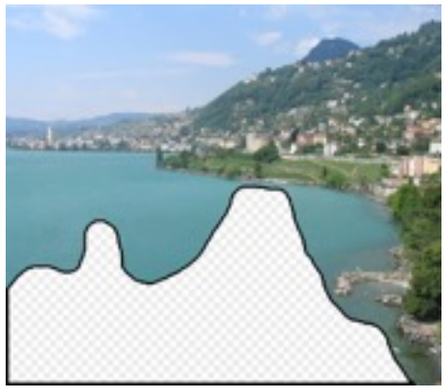
> 400hrs videos / minute

> 300M photos / day

24/7/365

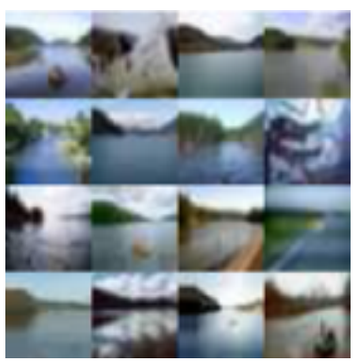
Big data opportunities

- Capturing modes and patterns **unseen** in small data
  - Increase learning and inference performance
- From **Nearest Neighbors**

"It takes a large amount of data for our method to succeed. We saw **dramatic improvement** when moving from ten thousand to two million images." **[Hayes, Efros, SIGGRAPH '07]**

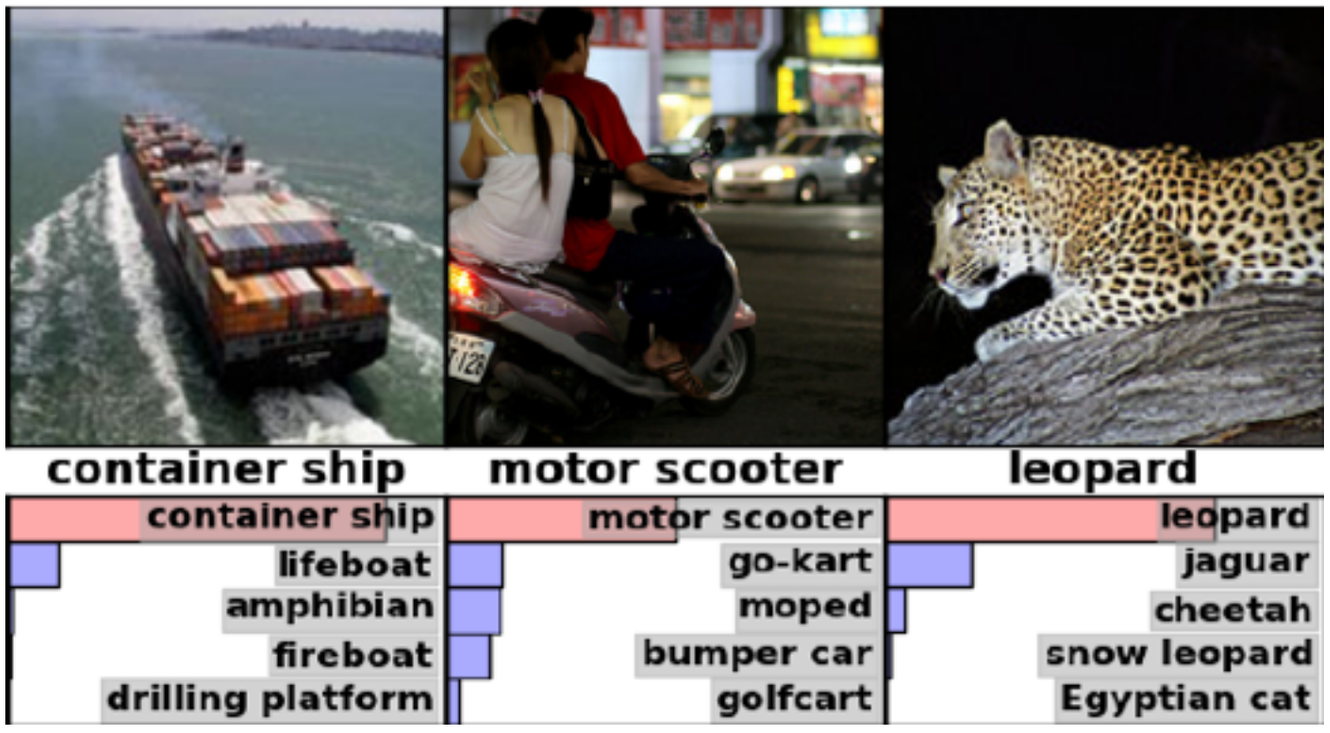


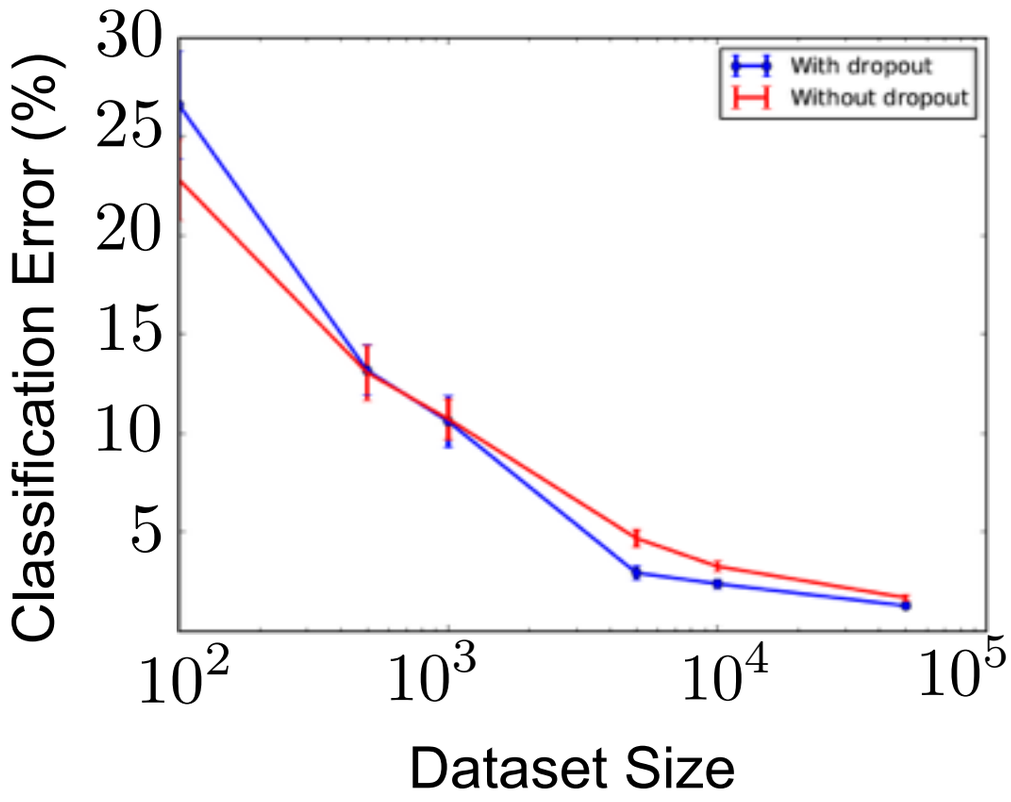



"As we increase the size of the dataset from 10^5 to the 10^8 images, the quality of the retrieved set **increases dramatically**." **[Torralba, Fergus, Freeman, PAMI '08]**

Big data opportunities

- Capturing modes and patterns **unseen** in small data
  - Increase learning and inference performance
- To **Deep Neural Networks**



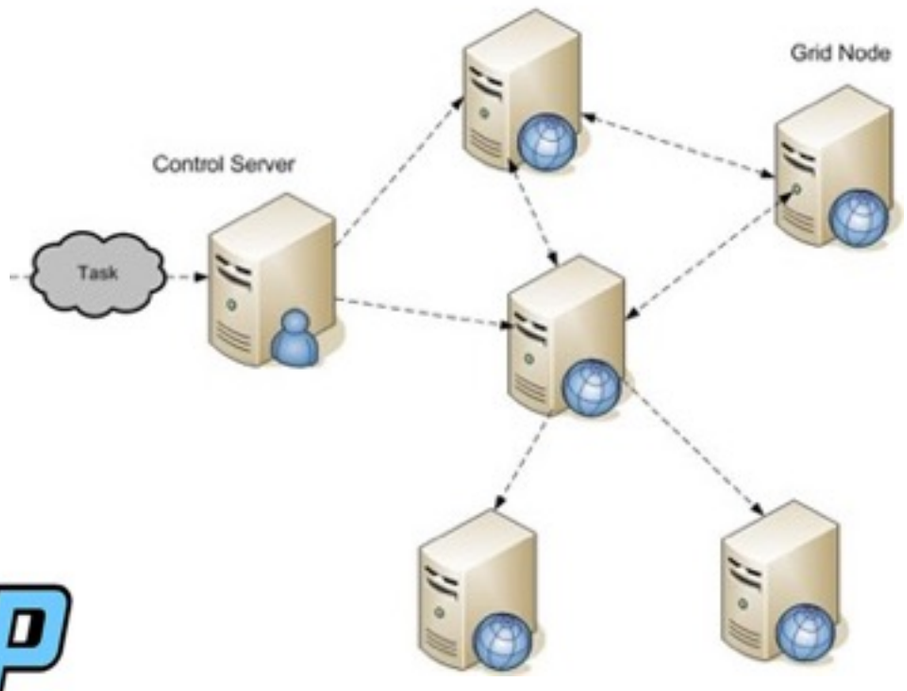


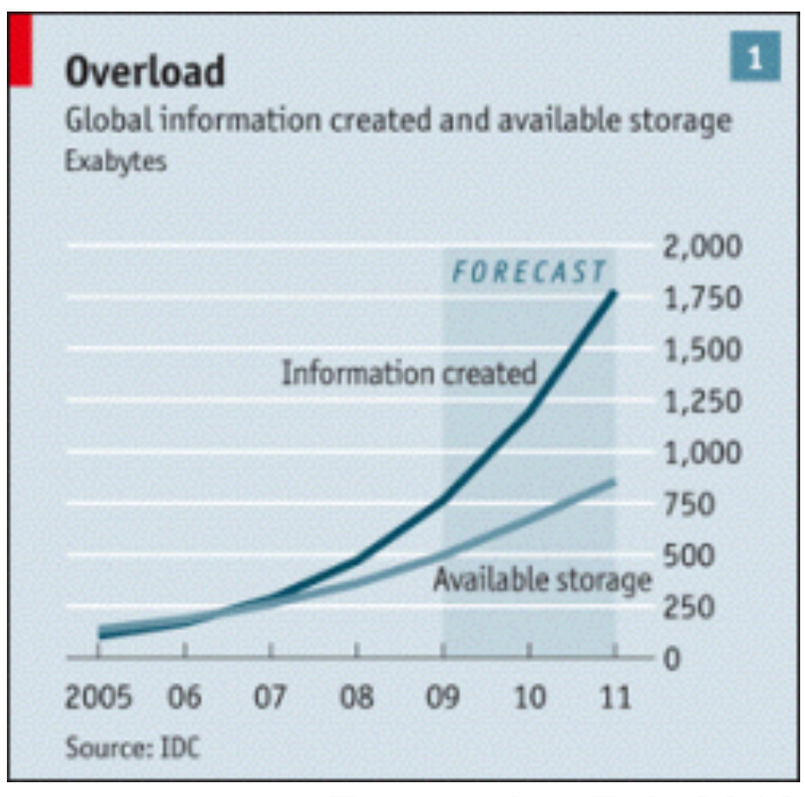
"From 100 to 100,000 training samples, classification error **drops from 25% to less than 5%**." **[Srivastava-Hinton-Krizhevsky-Sutskever-Salakhutdinov '14]**

Big data challenges

- Capturing modes and patterns **unseen** in small data
  - Increase learning and inference performance: from NNs to DNNs

- Limiting **computational** and **memory** resources
  - We have **more data** than capacities of our resources

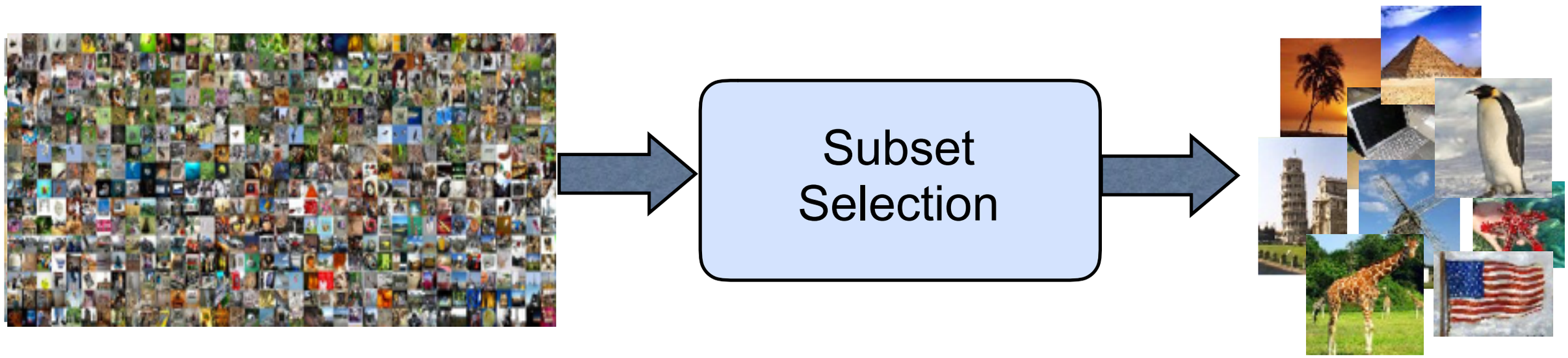




Economist, Feb 2010

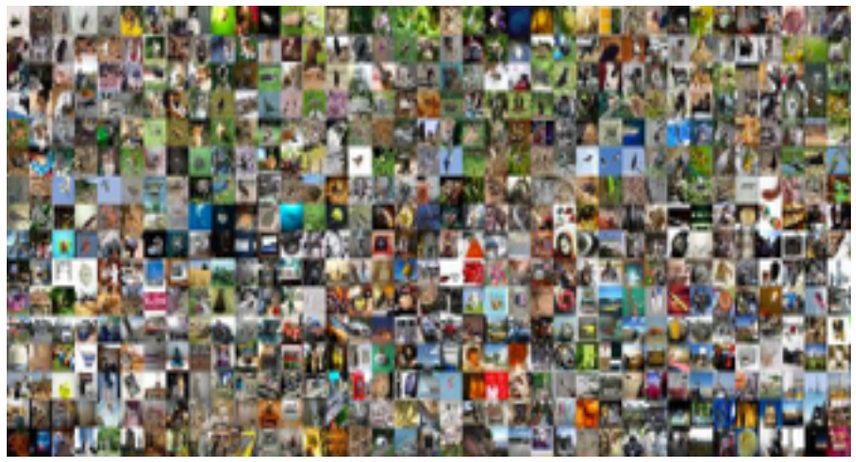
Subset selection

- Find **small data** capturing **statistical properties** of large data



Subset selection components

- How to **characterize informativeness** of items?
- How to recover the **most informative** items?

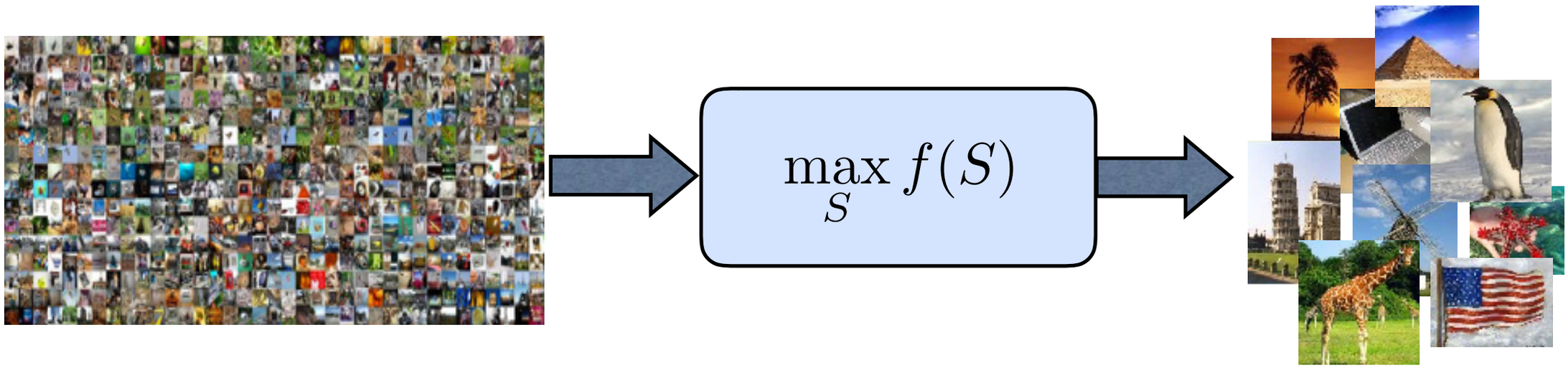


$$\max_S f(S)$$

1. What objective functions?

2. How to optimize?

This Talk



- Subset selection is **NP-hard**, hence, **approximate** algorithms
  - **Better summarization**: closer to **global optimum**

- Focus on one class of objective functions: **Facility Location**

  ➡ Develop **efficient sparse optimization**

  ➡ Show it can **achieve global optimum (under certain condition)**

This Talk

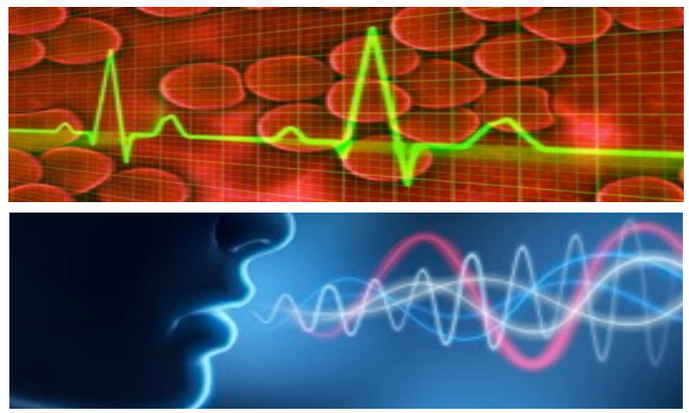
- **Sequential data** (video, text, signals,…): large part of **'Big Data'**

(CNN) -- A Japanese rocket roared into orbit early Friday (Thursday afternoon ET) carrying what NASA calls its most precise instrument yet for measuring rain and snowfall.

The Global Precipitation Measurement (GPM) satellite is the first of five earth science launches NASA has planned for 2014. The 4-ton spacecraft is the most sophisticated platform yet for measuring rainfall, capable of recording amounts as small as a hundredth of an inch an hour, said Gail Skofronick Jackson, GPM's deputy project scientist.

The $900 million satellite is a joint project with the Japanese space agency JAXA, and it lifted off from Tanegashima Space Center at 3:37 a.m. Friday (1:37 p.m. Thursday ET). In a little over a half hour, it had reached orbit, deployed its solar panels and began beaming signals back to its controllers, NASA said.

Also, once fully activated, GPM will use both radar and microwave instruments to detect falling snow for the first time. It will also combine data from other satellites with its own readings, beaming

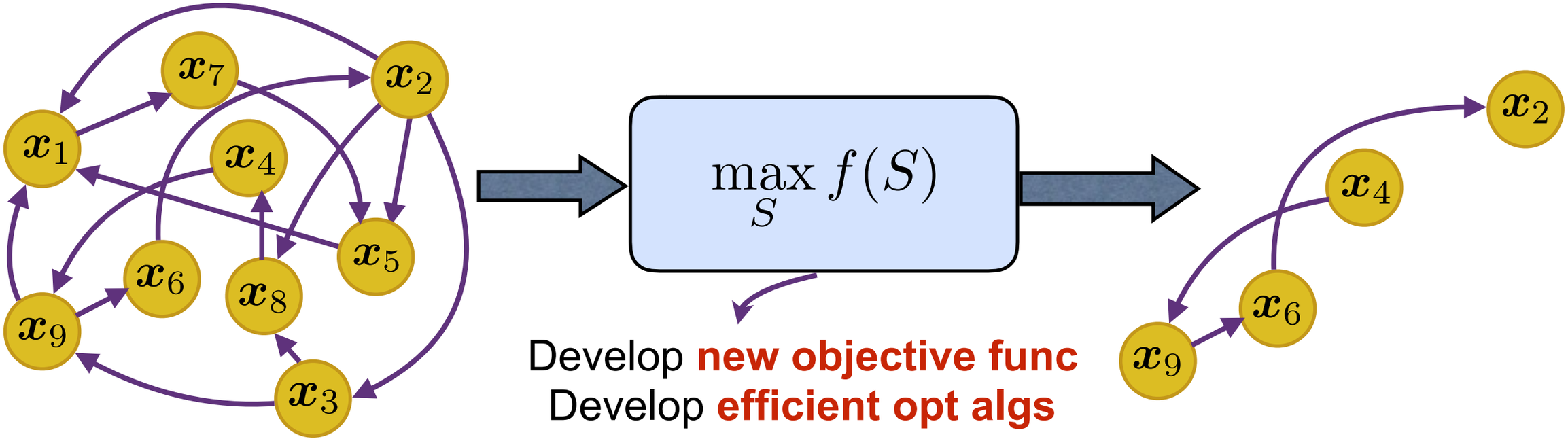


- Sequential data have **structured dependencies**
  - **Instructional videos**: ordering of key steps

Key steps

This Talk

- **Sequential data** (video, text, signals,…): large part of **'Big Data'**
- Sequential data have **structured dependencies**
  - Existing SS: data is a **bag of randomly permutable** items
- **Generalize facility location** to handle **seq summarization**

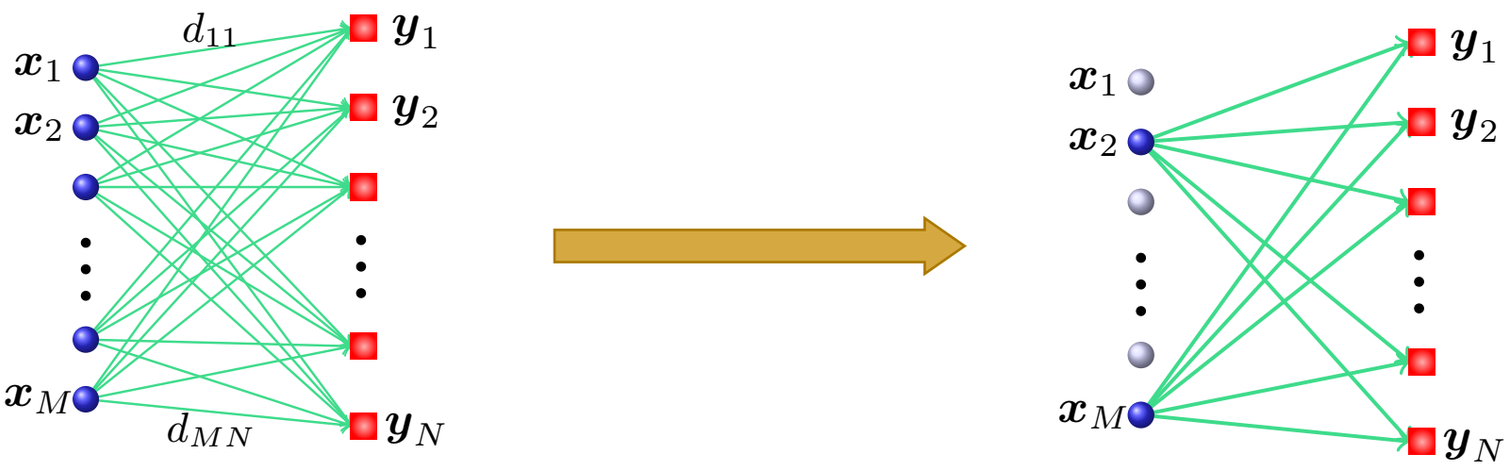


Subset selection from dissimilarities

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

  source: $\mathbb{X}$ target: $\mathbb{Y}$

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



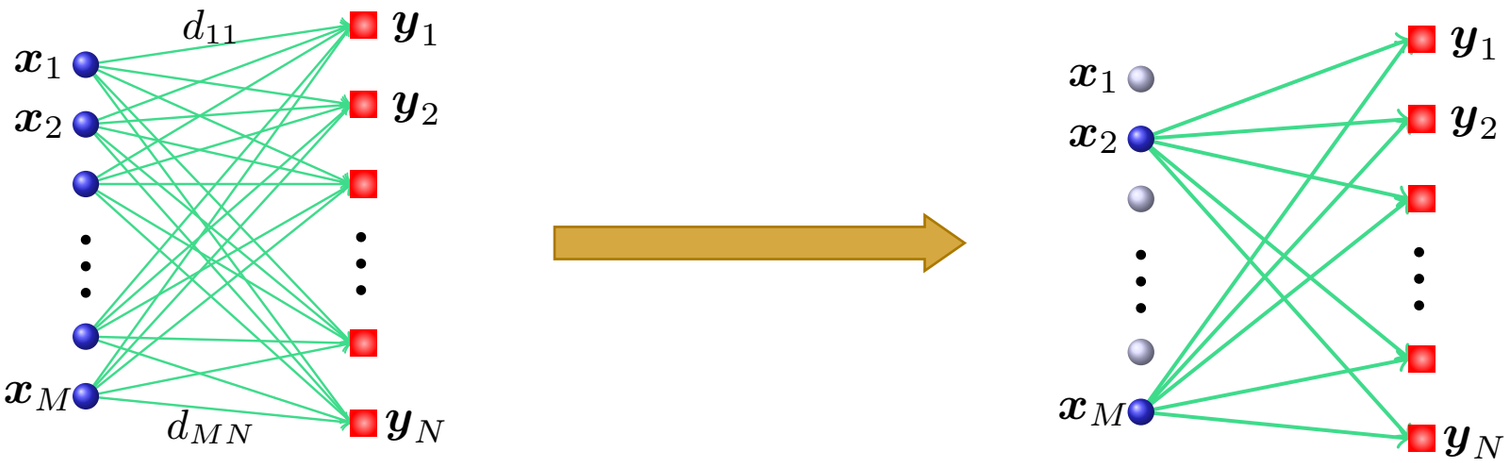
- $d(\boldsymbol{x}_i, \boldsymbol{y}_j) = d_{ij}$: how well $\boldsymbol{x}_i$ represents $\boldsymbol{y}_j$
  - $\mathbb{X}$ = data, $\mathbb{Y}$ = data $\longrightarrow$ $d(\cdot, \cdot)$ = Euclidean/geodesic distance
  - $\mathbb{X}$ = models, $\mathbb{Y}$ = data $\longrightarrow$ $d(\cdot, \cdot)$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

Subset selection from dissimilarities

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

  source: $\mathbb{X}$ target: $\mathbb{Y}$

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

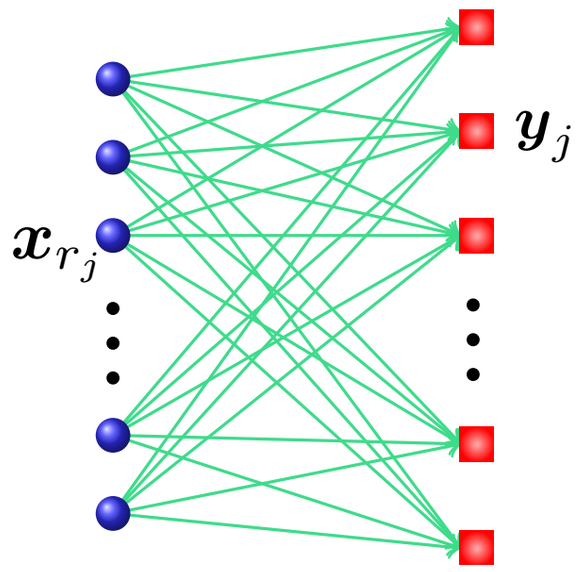


- $d(\boldsymbol{x}_i, \boldsymbol{y}_j) = d_{ij}$: how well $\boldsymbol{x}_i$ represents $\boldsymbol{y}_j$

Not necessarily metric:
asymmetric,
violate triangle ineq

Dissimilarity-based Sparse Subset Selection

- Let $r_j \in \{1, \ldots, M\}$ be index of representative of $\boldsymbol{y}_j$
- **Approach**: define & maximize **potential function**



$$\Psi(r_1, \ldots, r_N) \triangleq \Phi_{\text{enc}}(r_1, \ldots, r_N) \times \Phi_{\text{card}}(r_1, \ldots, r_N)$$

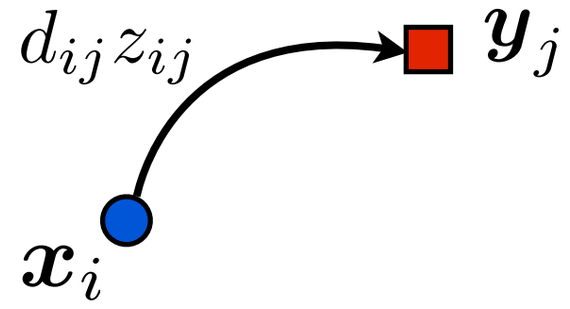
$$\Phi_{\text{enc}}(r_1, \ldots, r_N) \triangleq \prod_{j=1}^{N} e^{-d_{r_j, j}}$$ → **Prefer lower encoding**

$$\Phi_{\text{card}}(r_1, \ldots, r_N) \triangleq e^{-\lambda |\{r_1, \ldots, r_N\}|}$$ → **Prefer small # representatives**

- Max Ψ over $\{r_1, \ldots, r_N\} \in \{1, \ldots, M\}^N$: **combinatorial**

DS3 formulation

- Develop a **sparse binary optimization**
- Define variables $z_{i,j} \triangleq \mathrm{I}(r_j = i) \in \{0, 1\}$



$$\Psi = \prod_{i=1}^{M} \prod_{j=1}^{N} \exp(-d_{i,j} z_{i,j}) \times \exp\Big(-\lambda \sum_{i=1}^{M} \mathrm{I}(\|[z_{i,1} \ldots z_{i,N}]\|_p) \Big)$$

- Minimizing $-\log \Psi$:

Nonconvex

$$\min_{\{z_{i,j}\}} \sum_{i=1}^{M} \sum_{j=1}^{N} d_{i,j} z_{i,j} + \lambda \sum_{i=1}^{M} \mathrm{I}\left(\|[z_{i,1} \ \ldots \ z_{i,N}]\|_p\right)$$

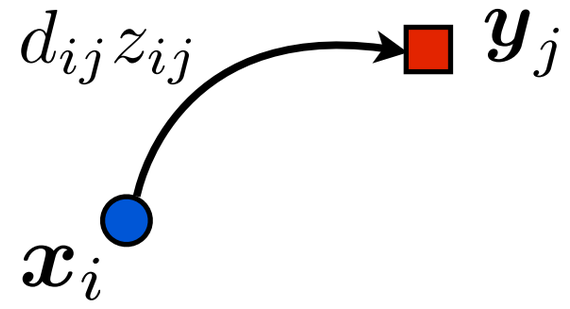
$$\text{s.t.} \quad z_{i,j} \in \{0, 1\}, \ \sum_{i=1}^{M} z_{i,j} = 1, \ \forall\, i, j$$

~ Facility location

E. Elhamifar, G. Sapiro, S. Sastry, TPAMI, 2016

DS3 formulation

- Develop a **sparse binary optimization**
- Define variables $z_{i,j} \triangleq \mathrm{I}(r_j = i) \in \{0, 1\}$



$$\Psi = \prod_{i=1}^{M} \prod_{j=1}^{N} \exp(-d_{i,j} z_{i,j}) \times \exp\Big(-\lambda \sum_{i=1}^{M} \mathrm{I}(\|[z_{i,1} \dots z_{i,N}]\|_p) \Big)$$

- Minimizing $-\log \Psi$:

Convex

$$\min_{\{z_{i,j}\}} \sum_{i=1}^{M} \sum_{j=1}^{N} d_{i,j} z_{i,j} + \lambda \sum_{i=1}^{M} \big\| \begin{bmatrix} z_{i,1} & \cdots & z_{i,N} \end{bmatrix} \big\|_p$$

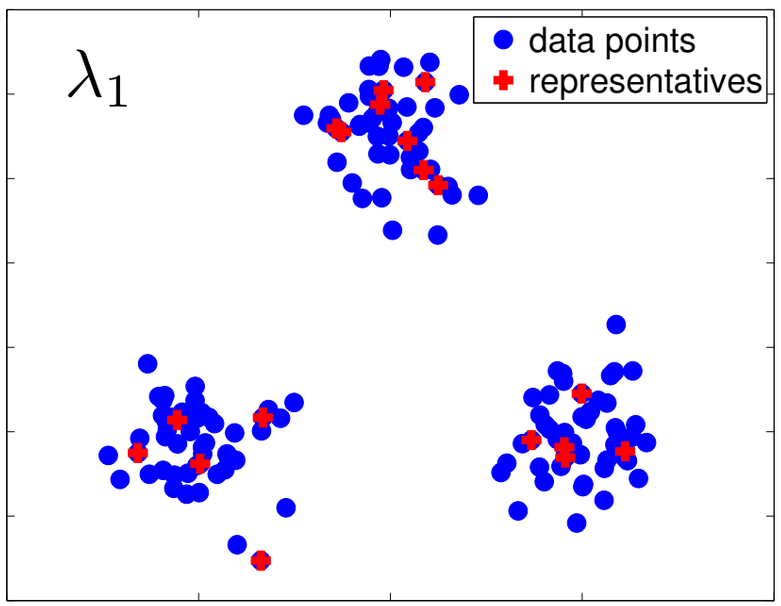
$$\text{s.t.} \quad z_{i,j} \geq 0, \quad \sum_{i=1}^{M} z_{i,j} = 1, \quad \forall\, i, j$$

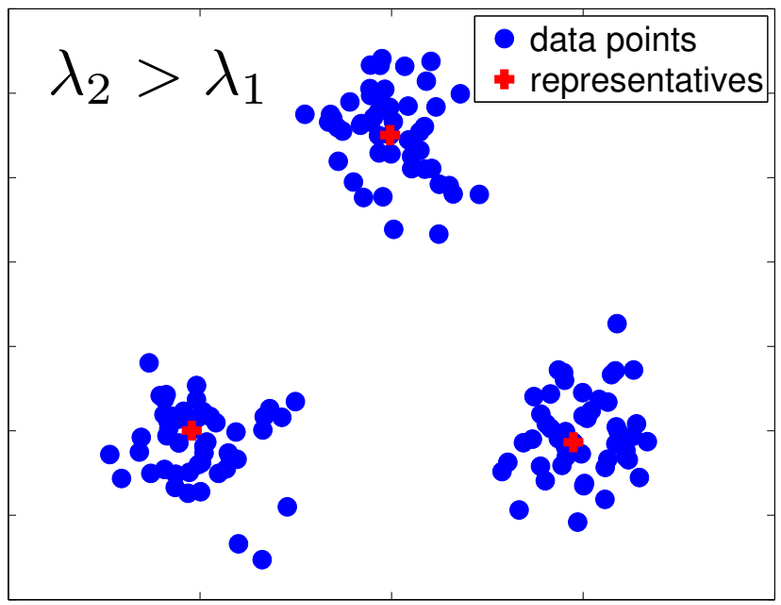
$$p \in \{2, \infty\}$$

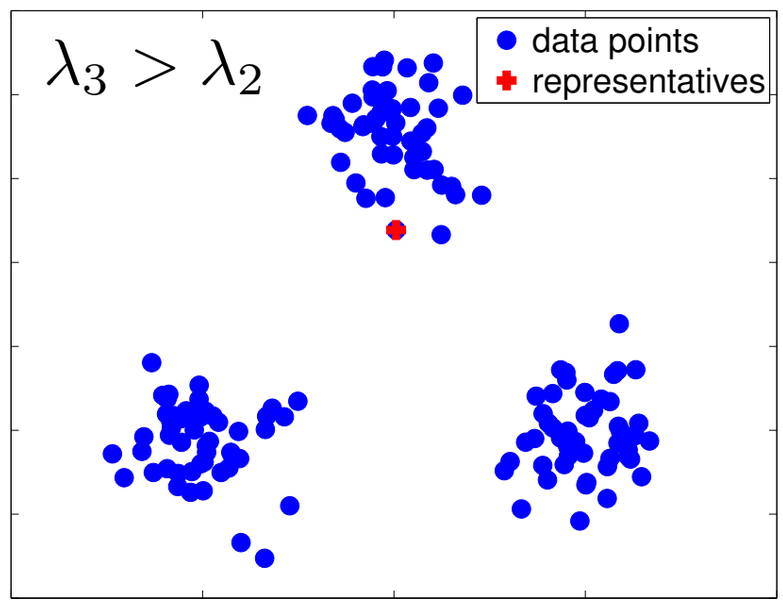
DS3

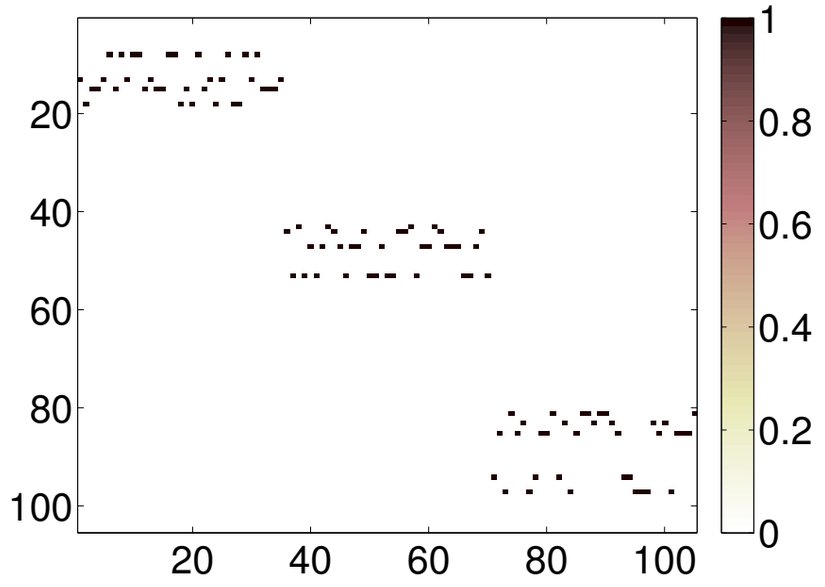
Toy example

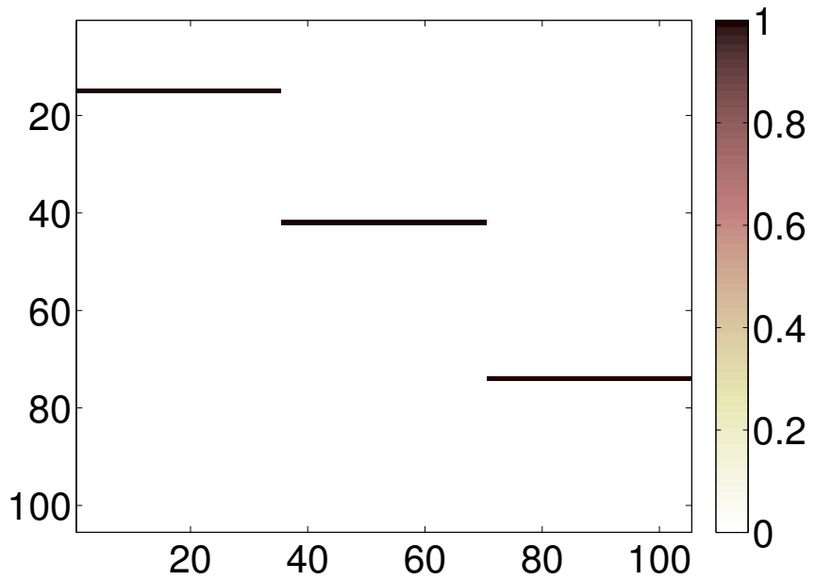
- Source = target = {data points}

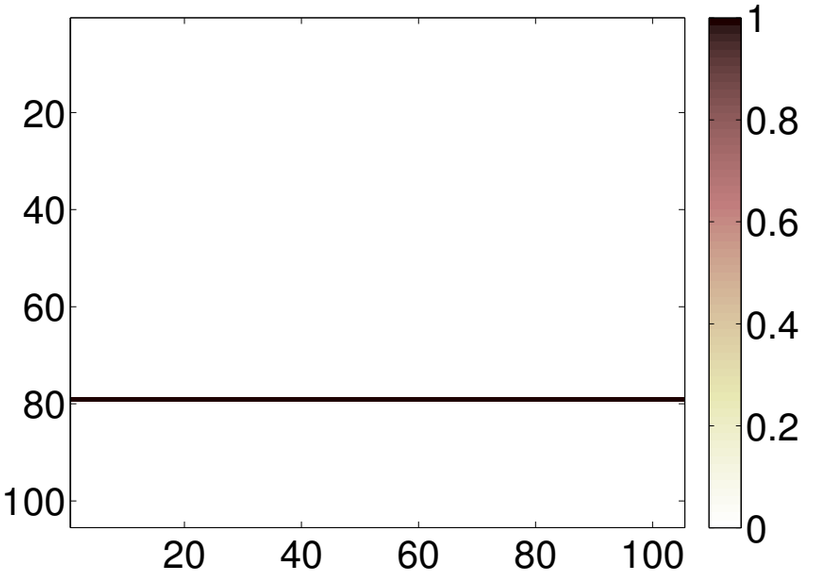












$$\min_{\{z_{i,j}\}} \sum_{i=1}^{M}\sum_{j=1}^{N} d_{i,j} z_{i,j} + \lambda \sum_{i=1}^{M} \left\| \begin{bmatrix} z_{i,1} & \cdots & z_{i,N} \end{bmatrix} \right\|_p \quad \text{s.t.} \quad z_{i,j} \geq 0, \ \sum_{i=1}^{M} z_{i,j} = 1, \ \ \forall\, i,j$$

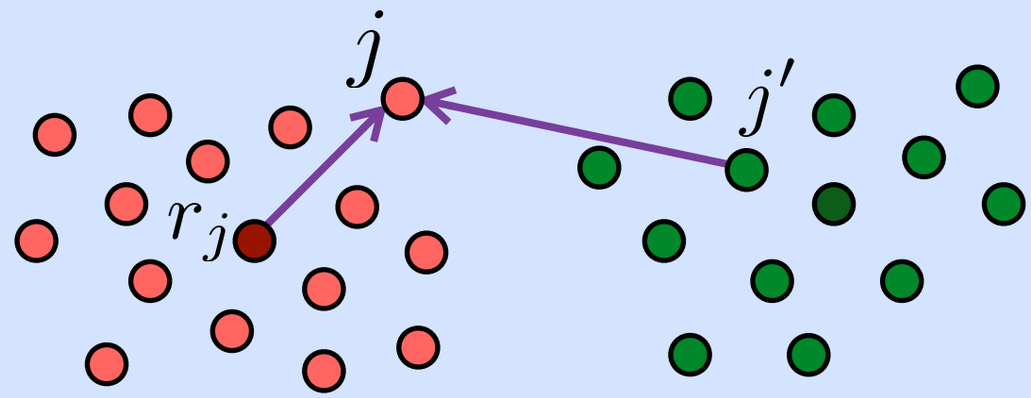
Theoretical analysis

- When is **convex relaxation exact?**

- **Theorem:** Let r_j be representative of j via **non-convex optimization**. The convex relaxation is exact if

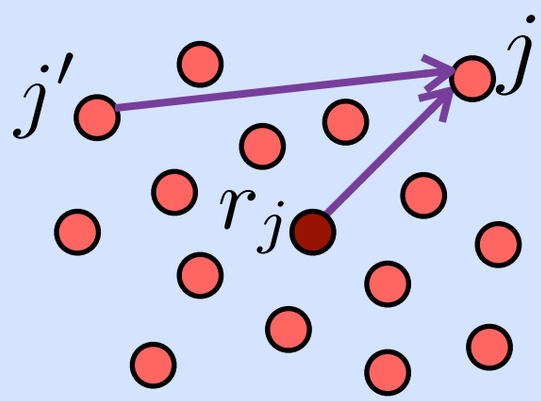
1) $\frac{\lambda}{N_{r_j}} + d_{r_j,j} < d_{j',j}$

Clusters sufficiently separated



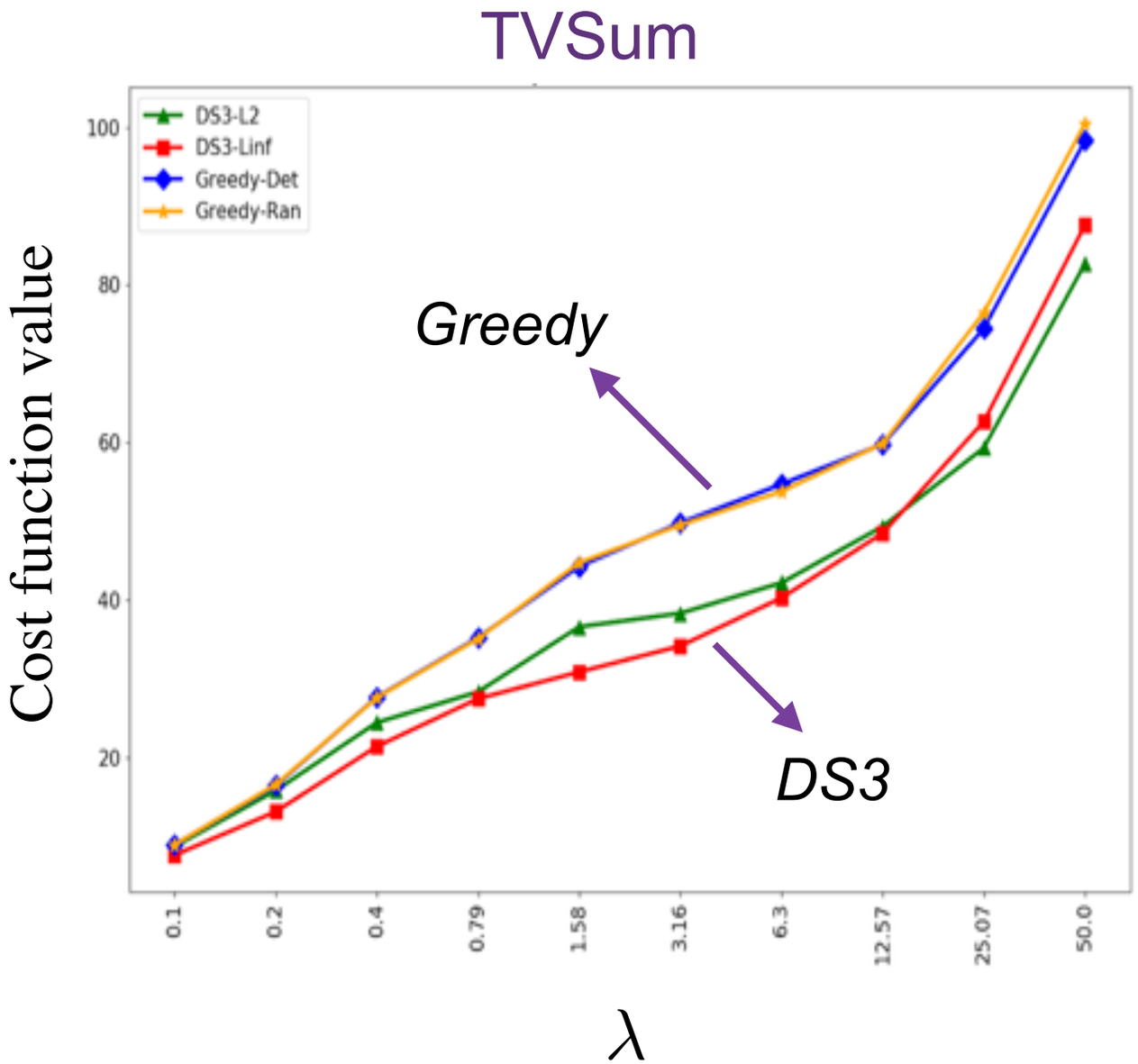
2) $d_{j',j} \leq \frac{\lambda}{N_{r_j}} + d_{r_j,j}$

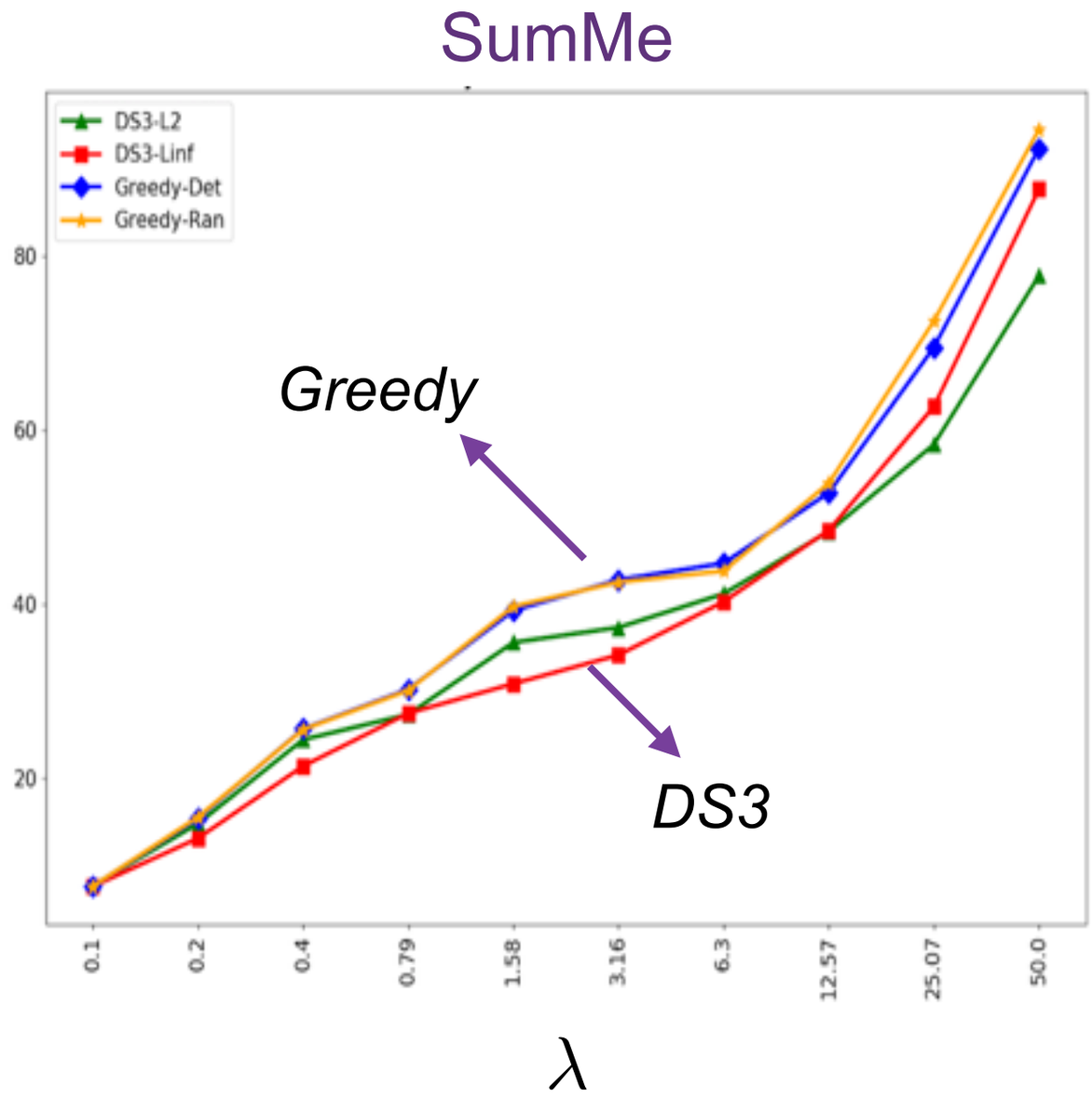
Sufficiently small radius



Comparison with greedy

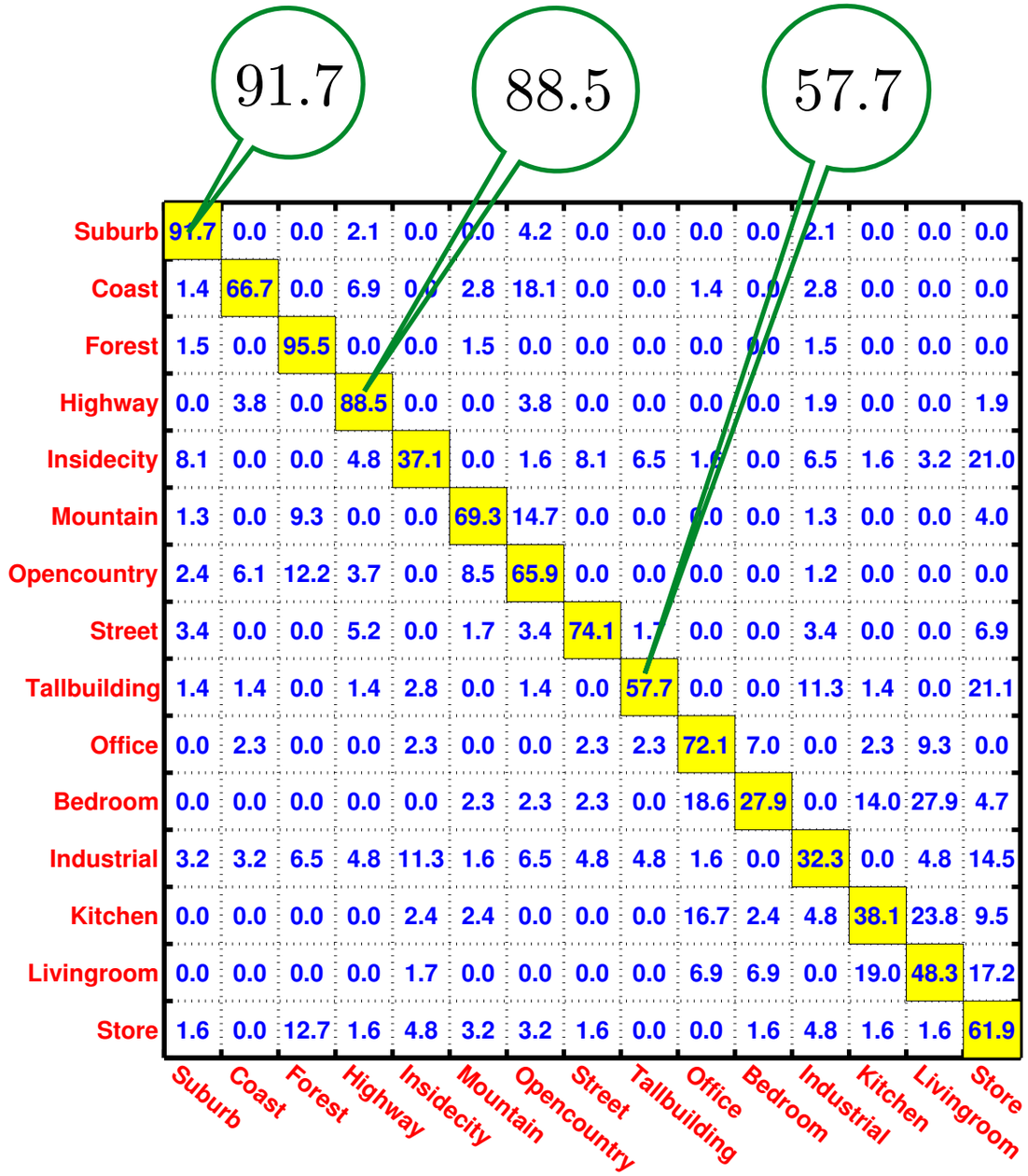
- Video summarization
  - Solve FL via Greedy and sparse coding (DS3)



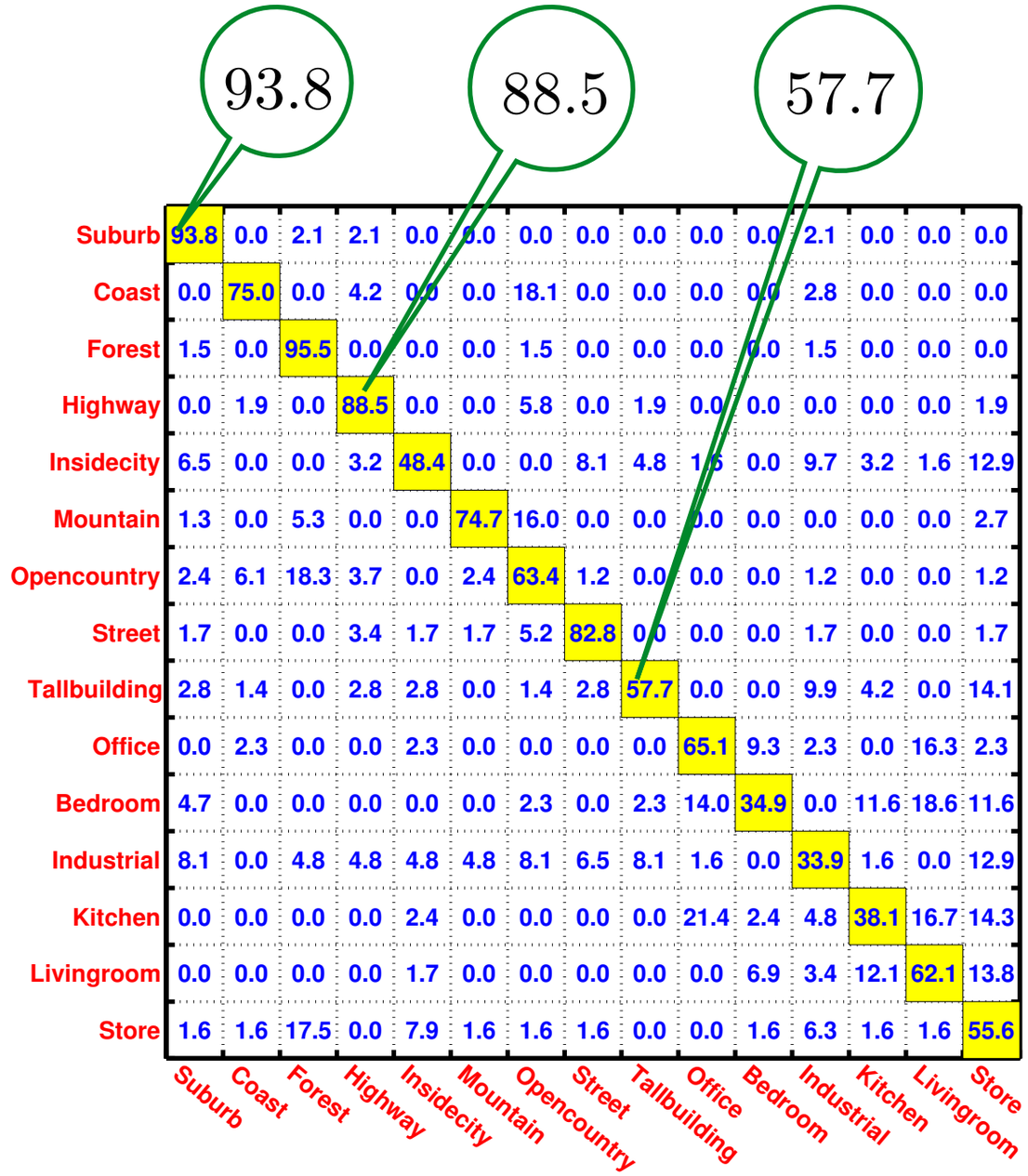


Recognition via representatives

- Select η fraction of training samples via DS3



DS3 selecting **35% samples**

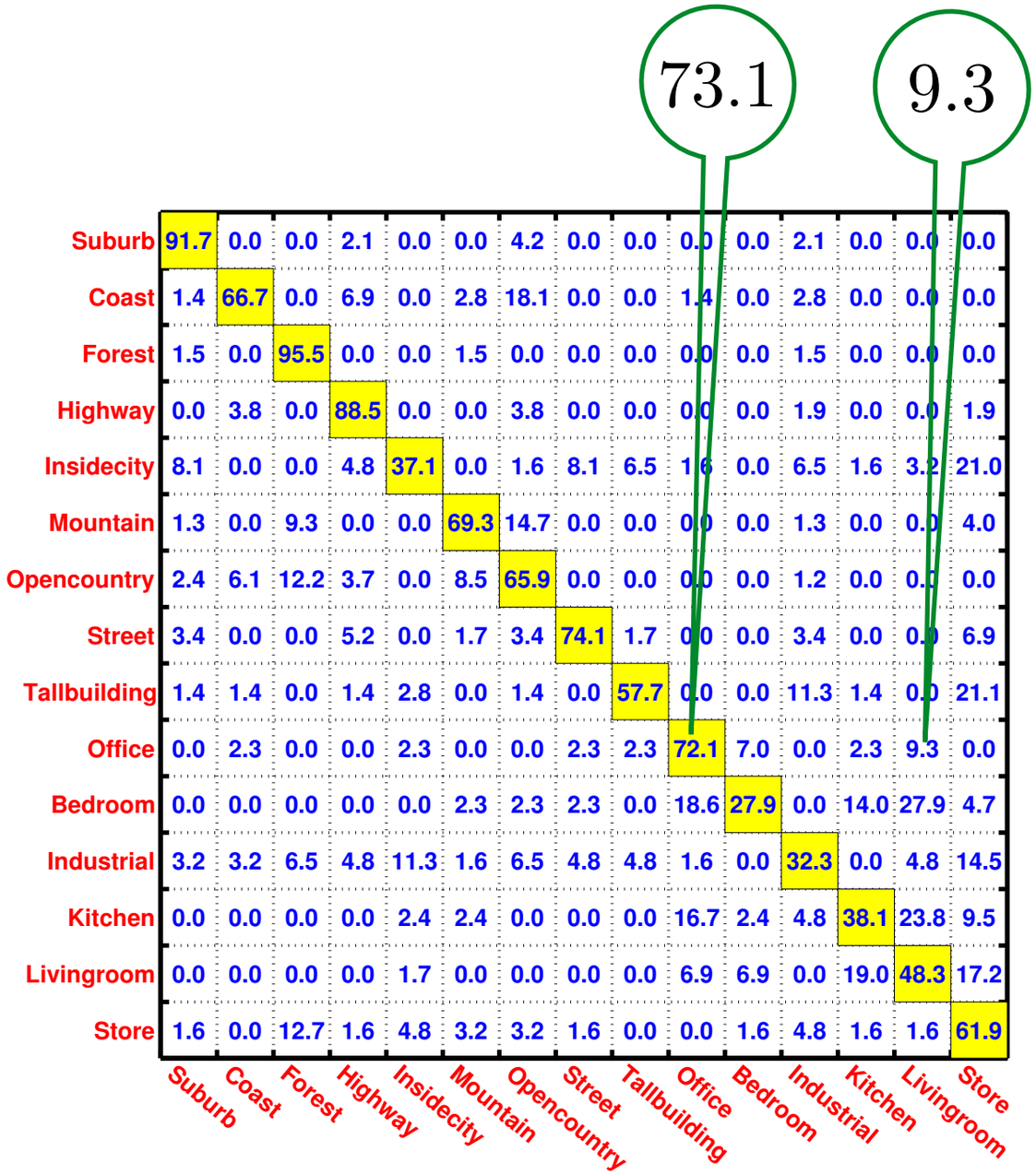


Trained via **all samples**

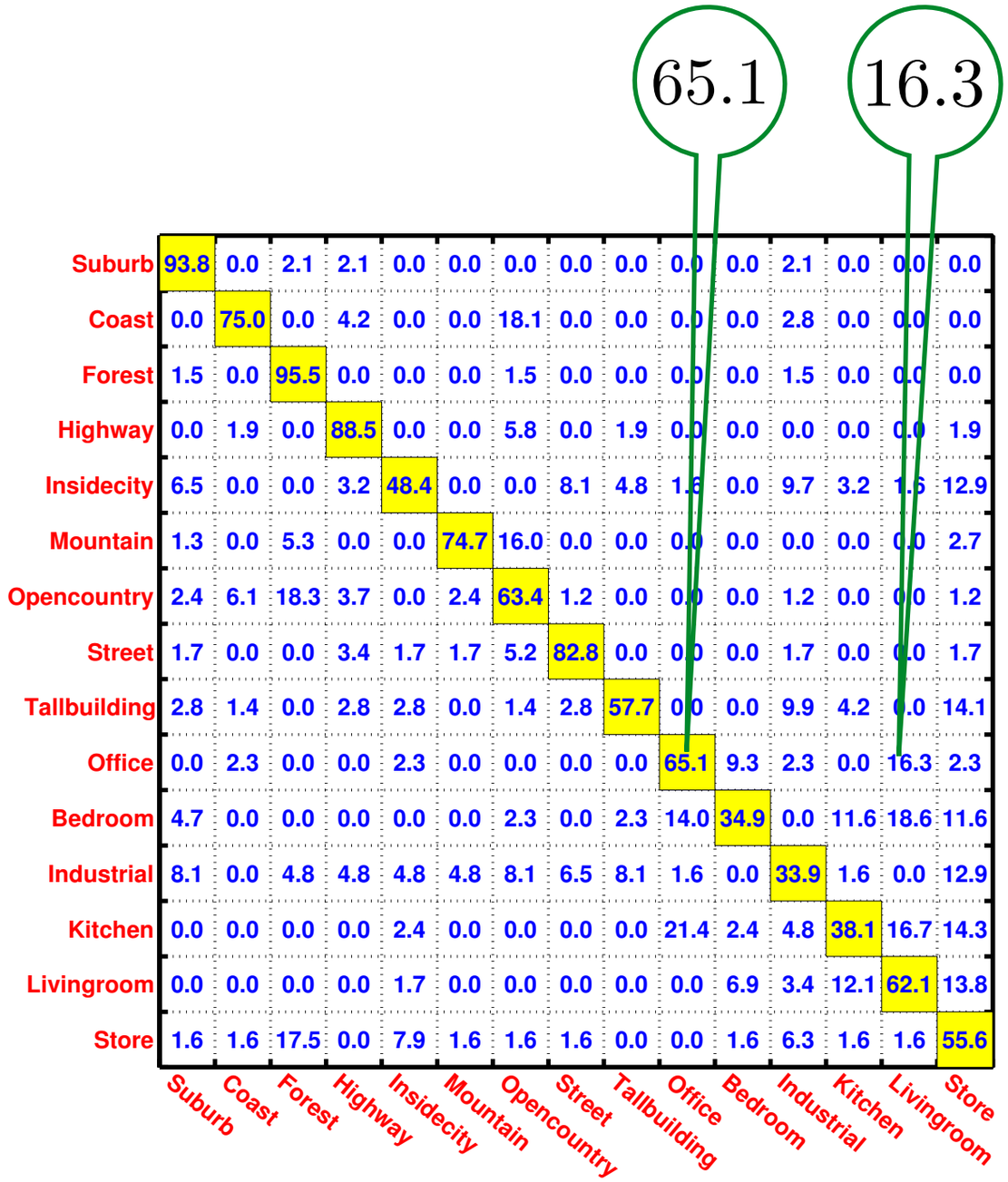
E. Elhamifar, G. Sapiro, S. Sastry, TPAMI, 2016

Recognition via representatives

- Select η fraction of training samples via DS3



DS3 selecting **35% samples**

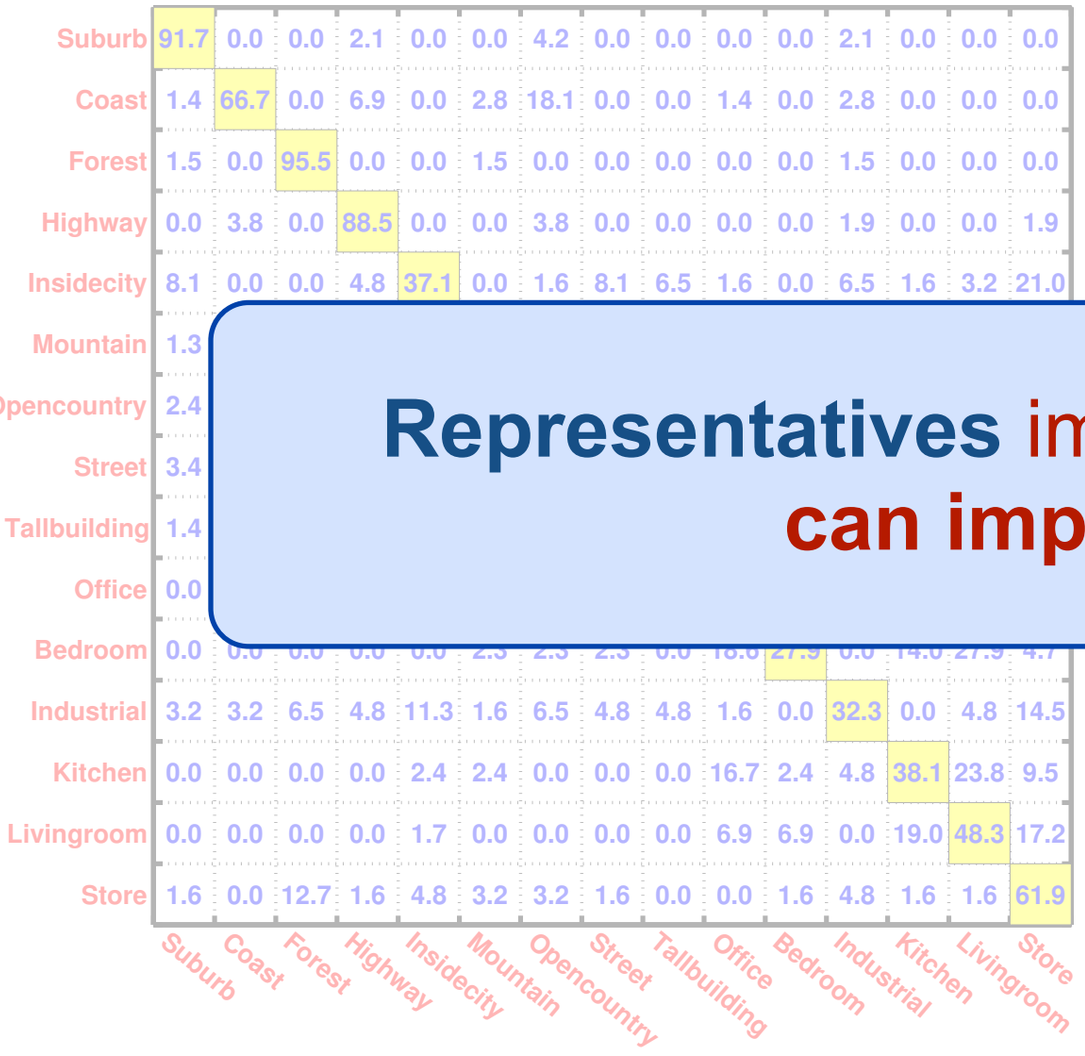


Trained via **all samples**

E. Elhamifar, G. Sapiro, S. Sastry, TPAMI, 2016

Recognition via representatives

- Select η fraction of training samples via DS3

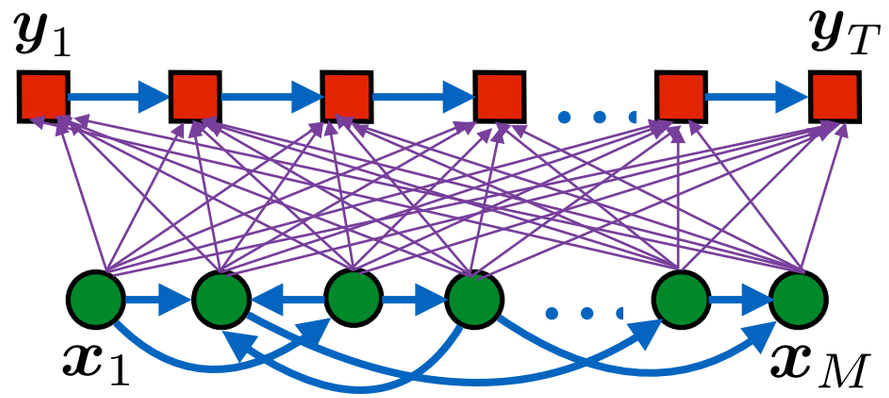


Representatives improve computational time and **can improve performance**

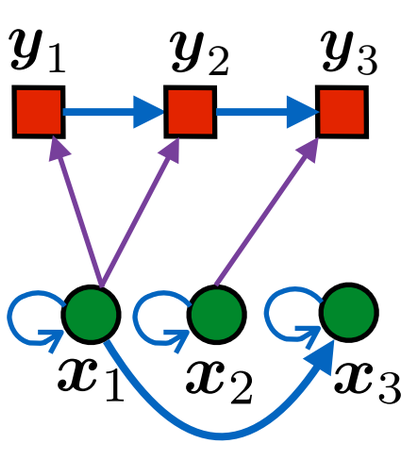
E. Elhamifar, G. Sapiro, S. Sastry, TPAMI, 2016

Sequential subset selection

- Incorporate **dynamics** of sequential data into SS
  - Source set: with **transition dynamics**
  - Target set: **sequential data**

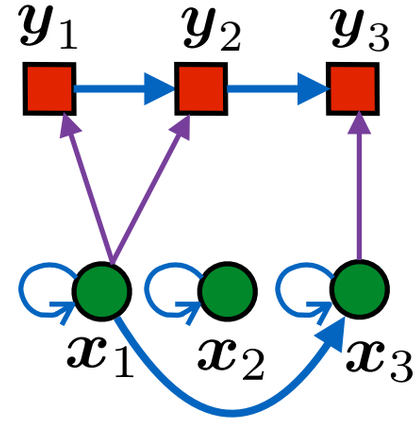


- **Goal:** find $(r_1, \ldots, r_T)$ with **high transition probability** that **well encodes the data** and has **small cardinality**



$(r_1, r_2, r_3) = (1, 1, 2)$

<

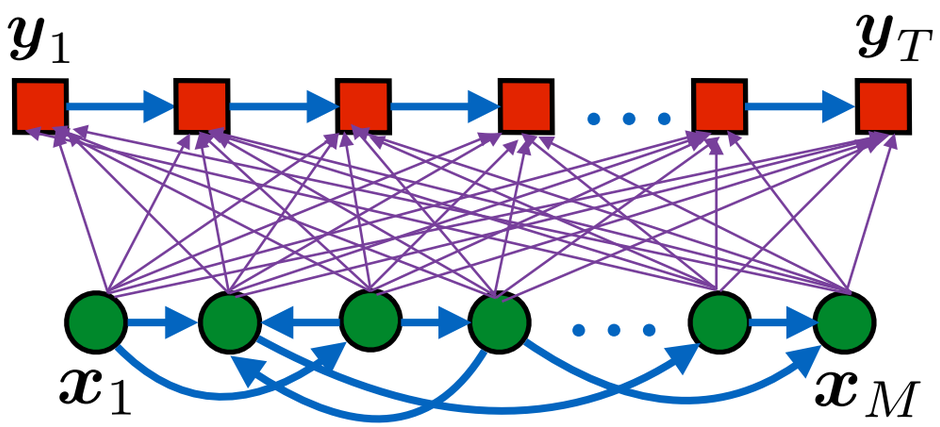


$(r_1, r_2, r_3) = (1, 1, 3)$

E. Elhamifar, C. De Paolis Kaluza, NIPS, 2017

Sequential facility location

- Let r_t be index of representative of $\boldsymbol{y}_t$



- Maximize global potential function

Prefer a sequence of compatible reps

$$\max_{(r_1,\ldots,r_T)} \Phi_{\text{enc}}(r_1,\ldots,r_N) \times \Phi_{\text{card}}(r_1,\ldots,r_N) \times \Phi_{\text{dyn}}(r_1,\ldots,r_N)$$

- **Dynamic potential**: n-th order Markov model

$$\Phi_{\text{dyn}}(r_1,\ldots,r_N) \triangleq \left(\prod_t p_t(\boldsymbol{x}_{r_t}|\boldsymbol{x}_{r_{t-1}},\ldots,\boldsymbol{x}_{r_{t-n}})\right)^{\beta}$$

E. Elhamifar, C. De Paolis Kaluza, NIPS, 2017

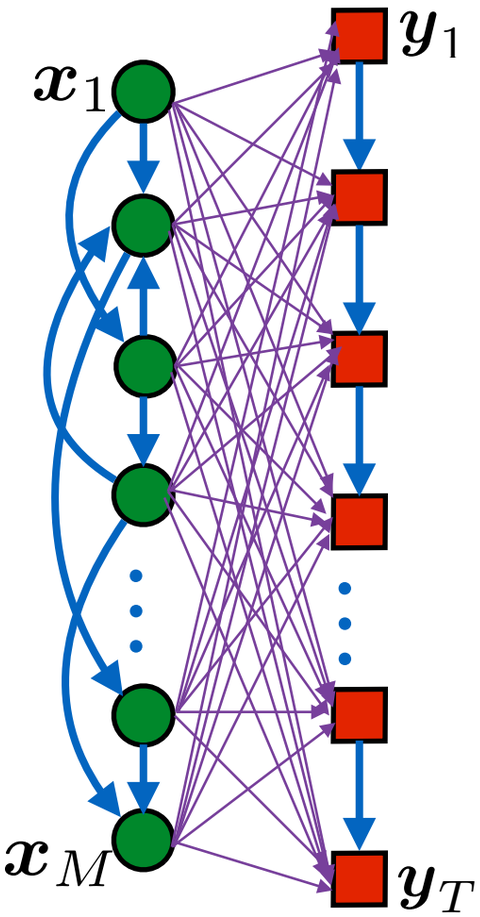
Sequential facility location

- **SeqFL**: Solve optimization on assignment variables $\{z_{i,t}\}$

$$\max_{\{z_{i,t}\}} \sum_{t=1}^{T}\sum_{i=1}^{M} -z_{i,t}d_{i,t} - \lambda \sum_{i=1}^{M} \|\begin{bmatrix} z_{i,1} & \cdots & z_{i,T} \end{bmatrix}\|_\infty$$

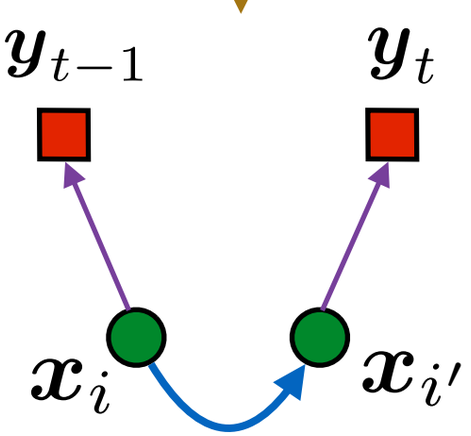
$$+ \beta\Big(\sum_{i=1}^{M} z_{i,1} \log p_1(\boldsymbol{x}_i) + \sum_{t=2}^{T}\sum_{i,i'=1}^{M} z_{i,t-1} z_{i',t} \log p(\boldsymbol{x}_{i'}|\boldsymbol{x}_i)\Big)$$

$$\text{s.t.} \quad z_{i,t} \in \{0,1\}, \quad \sum_{i=1}^{M} z_{i,t} = 1, \ \forall\, i,t.$$



$$z_{i,t-1} z_{i',t} = 1$$

$p(\boldsymbol{x}_{i'}|\boldsymbol{x}_i)$ should be large!



E. Elhamifar, C. De Paolis Kaluza, NIPS, 2017

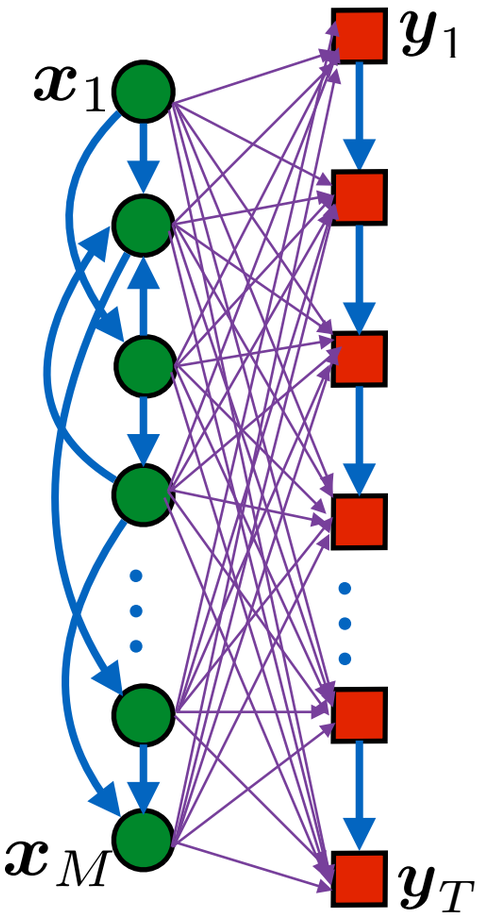
Sequential facility location

- **SeqFL**: Solve optimization on assignment variables $\{z_{i,t}\}$

$$\max_{\{z_{i,t}\}} \sum_{t=1}^{T}\sum_{i=1}^{M} -z_{i,t}d_{i,t} - \lambda \sum_{i=1}^{M} \|\begin{bmatrix} z_{i,1} & \cdots & z_{i,T} \end{bmatrix}\|_\infty$$

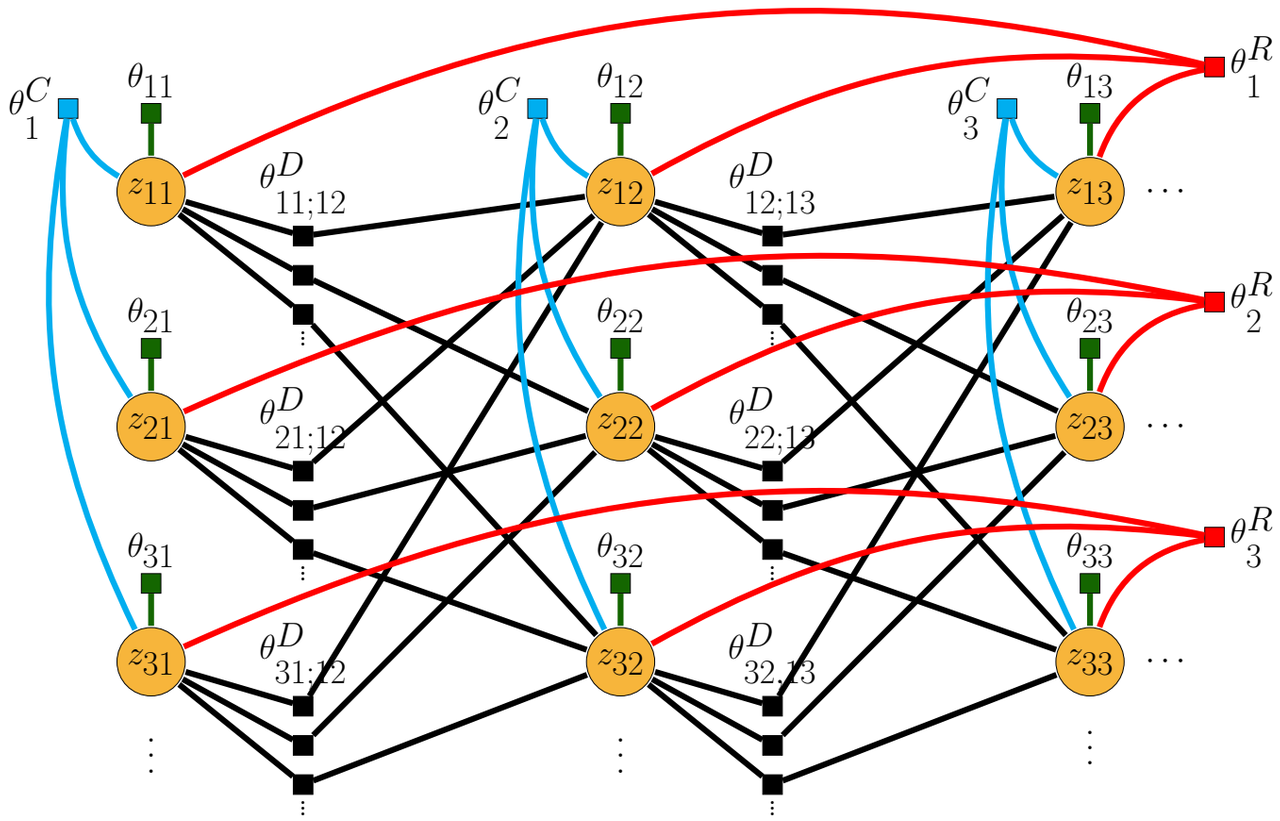
$$+ \beta\Big(\sum_{i=1}^{M} z_{i,1} \log p_1(\boldsymbol{x}_i) + \sum_{t=2}^{T}\sum_{i,i'=1}^{M} z_{i,t-1} z_{i',t} \log p(\boldsymbol{x}_{i'}|\boldsymbol{x}_i)\Big)$$

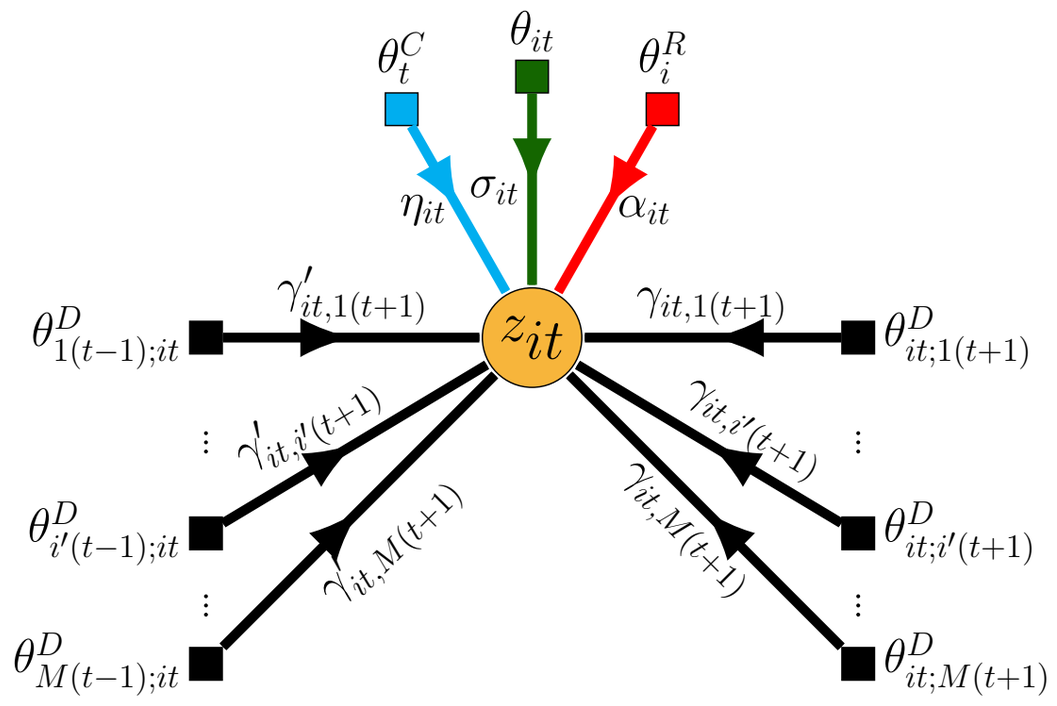
$$\text{s.t.} \quad z_{i,t} \in \{0,1\}, \quad \sum_{i=1}^{M} z_{i,t} = 1, \ \forall\, i,t.$$



- **Non-convex**: does not make sense (here) to relax!
  - Solve via **Max-Sum Message Passing**

SeqFL via message passing



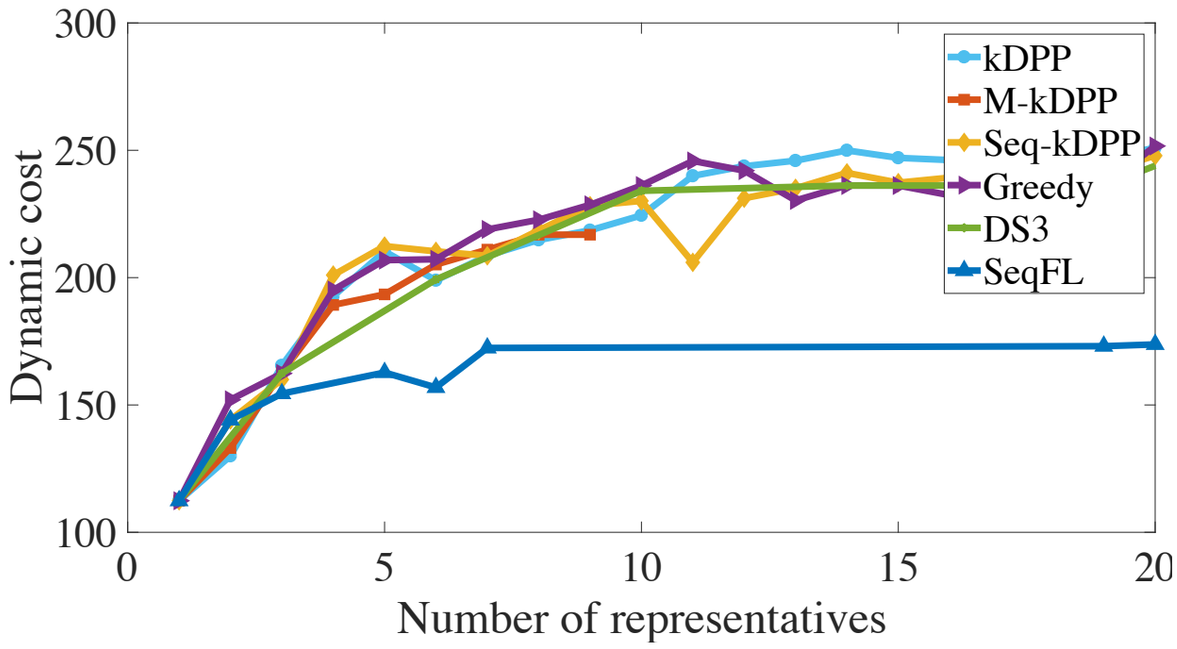


Encoding factor:	$\theta_{it}(z_{i,t})$	$= -\bar{d}_{i,t} z_{i,t}$
Cardinality factor:	$\theta^R_i(\boldsymbol{z}_{i,:})$	$= \begin{cases} -\lambda, & \|\boldsymbol{z}_{i,:}\|_\infty > 0 \\ 0, & \text{otherwise} \end{cases}$
Dynamic factor:	$\theta^D_{it-1;i',t}(z_{i,t-1}, z_{i',t})$	$= \log p(\boldsymbol{x}_{i'} \vert \boldsymbol{x}_i) z_{i,t-1} z_{i',t}$
Constraint factor:	$\theta^C_t(\boldsymbol{z}_{:,t})$	$= \begin{cases} 0, & \sum_{i=1}^M z_{i,t} = 1 \\ -\infty, & \text{otherwise.} \end{cases}$

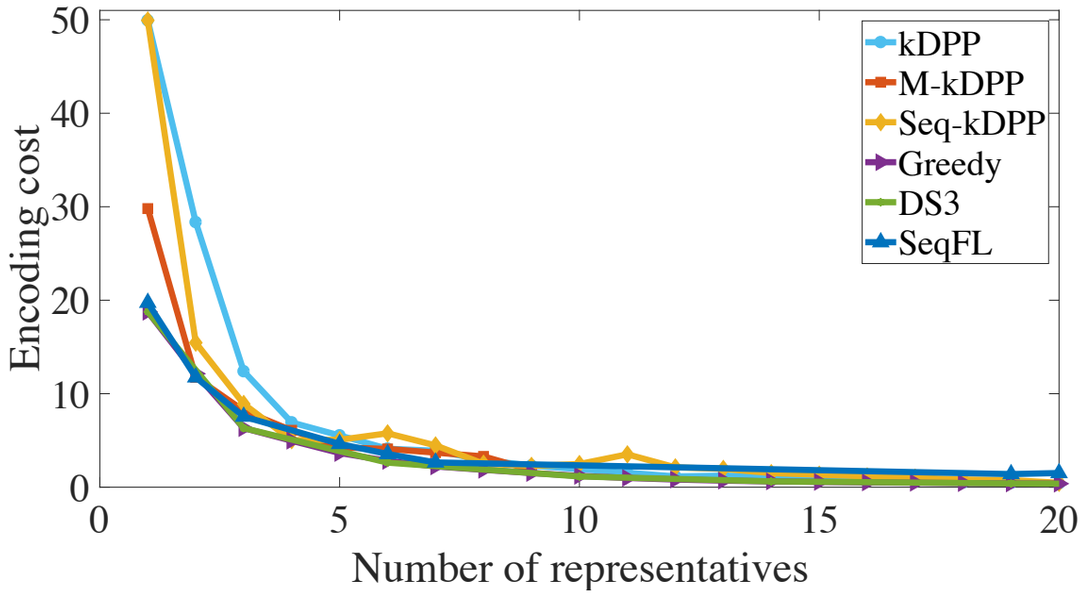
E. Elhamifar, C. De Paolis Kaluza, NIPS, 2017

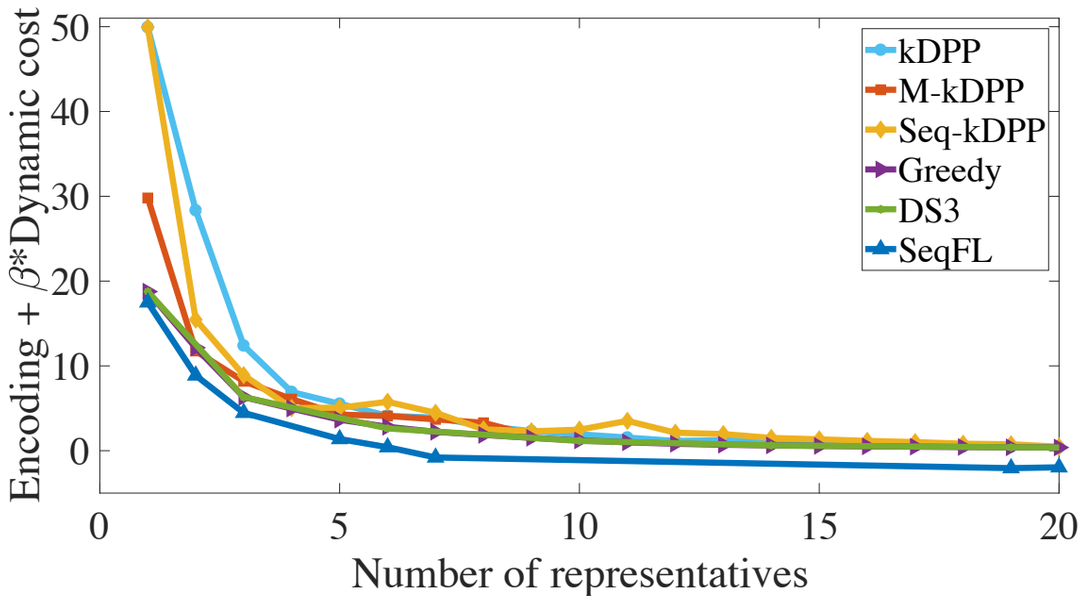
Synthetic experiments

- Data from HMM with M = 50 states and T = 100



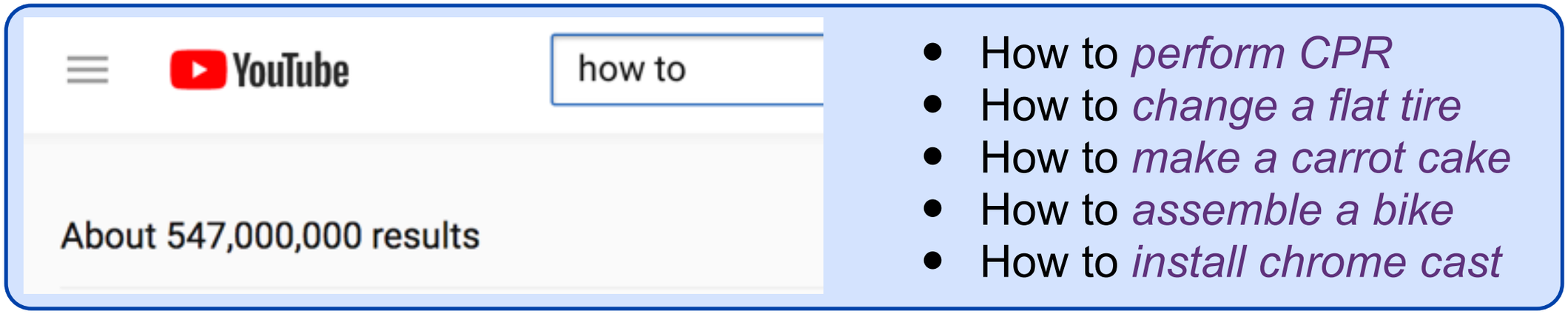
Sequence of reps in SeqFL:
compatible with dynamic model

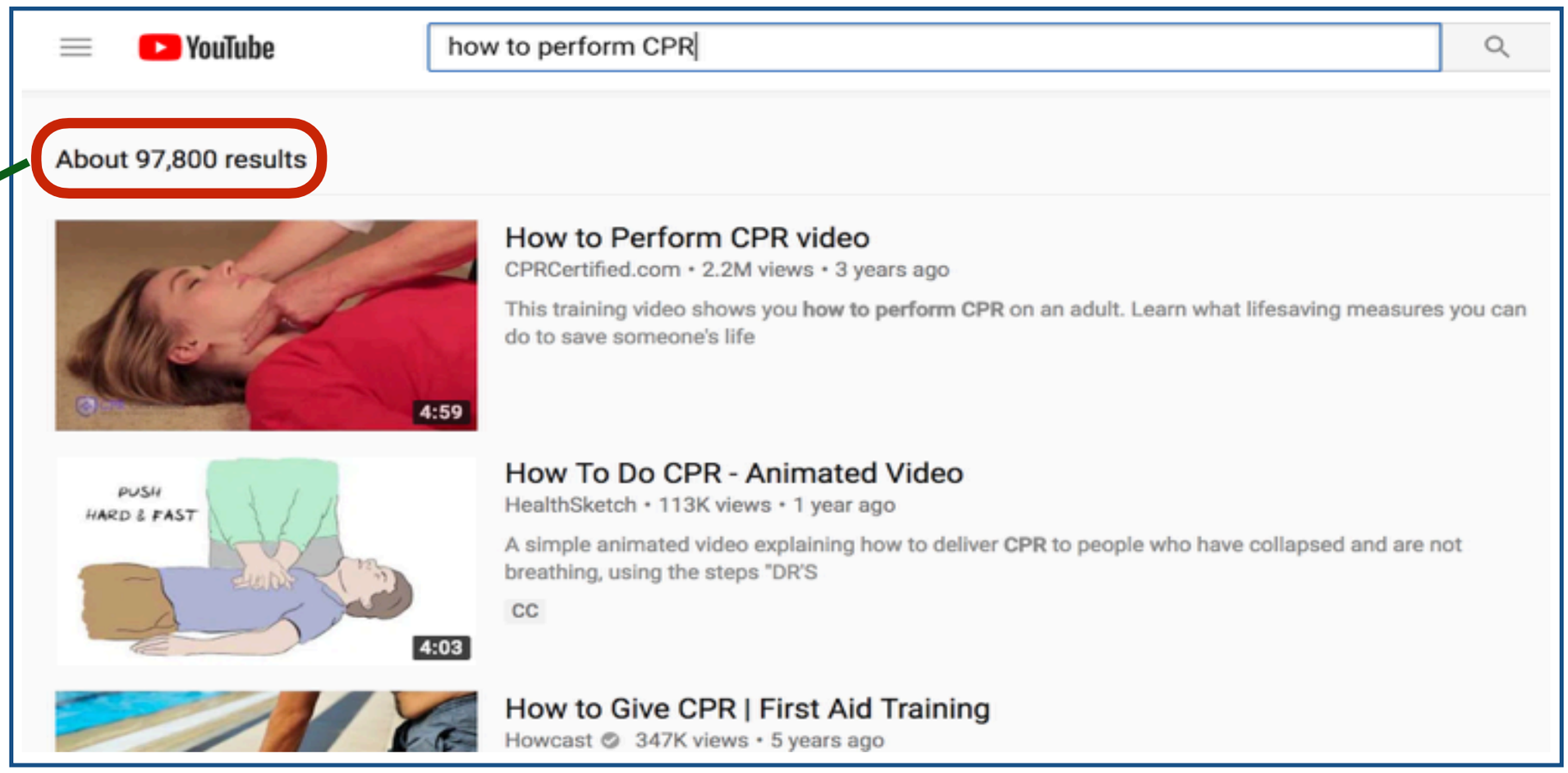




Procedure learning

- We learn **how to do** tasks using **instructional videos**

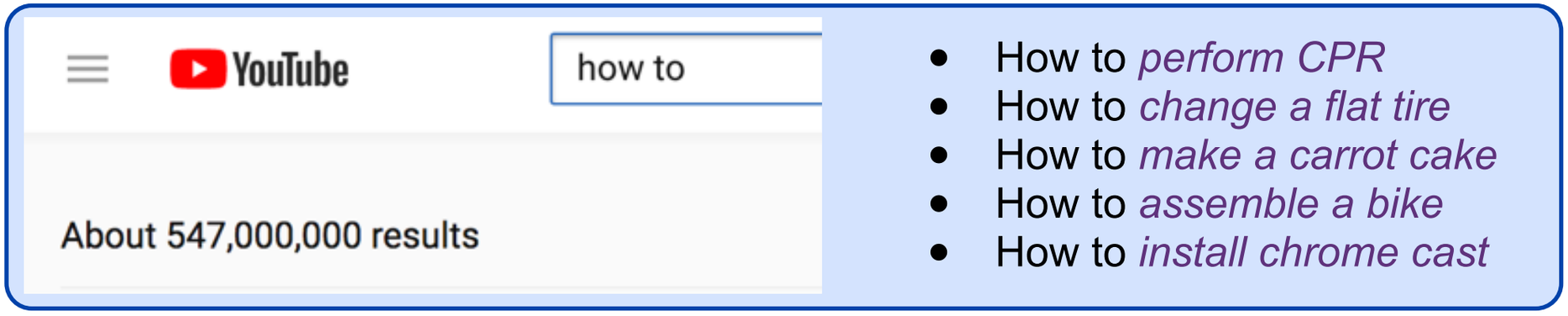




Large # videos

Procedure learning

- We learn **how to do** tasks using **instructional videos**



- Goal: **Learn grammar** of complex tasks using video data

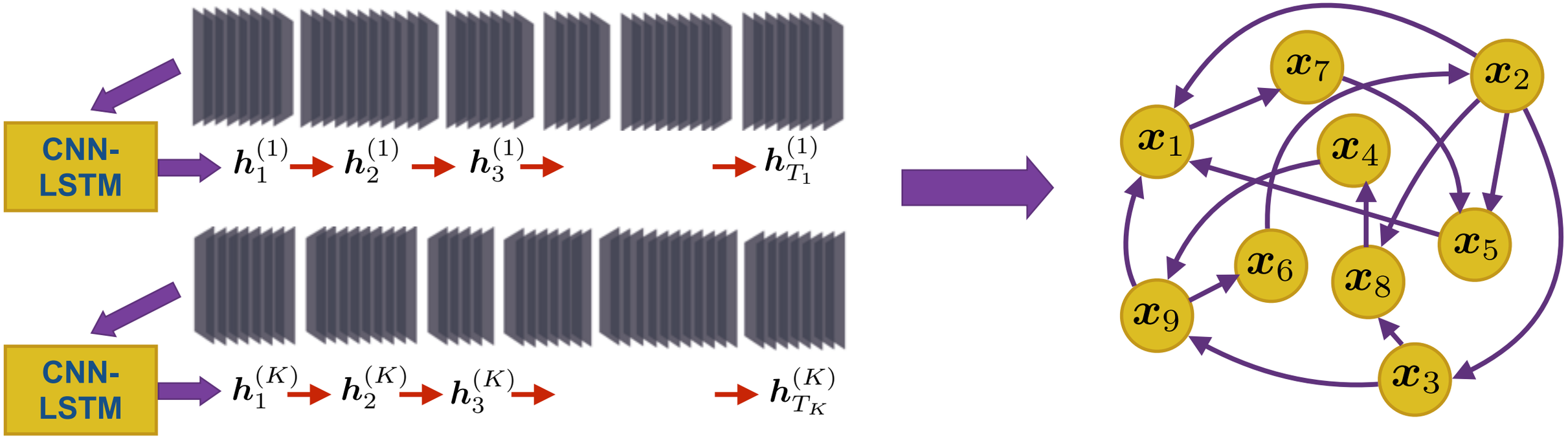
procedure:

Procedure learning via SeqFL

1) Extract **superframes** (Gygli et al ’14) from each video

2) Extract **deep features** for each superframe

3) Build a dynamic model: **1st order HMM**



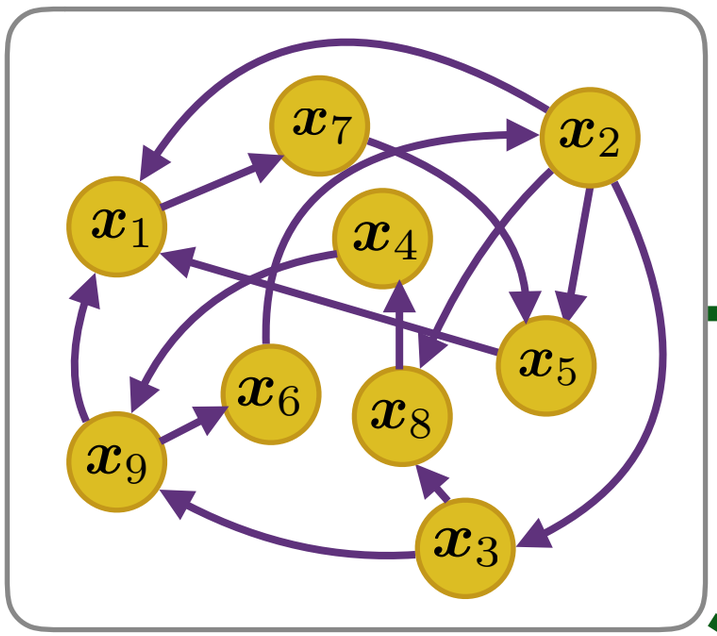
Procedure learning via SeqFL

1) Extract **superframes** (Gygli et al ’14) from each video

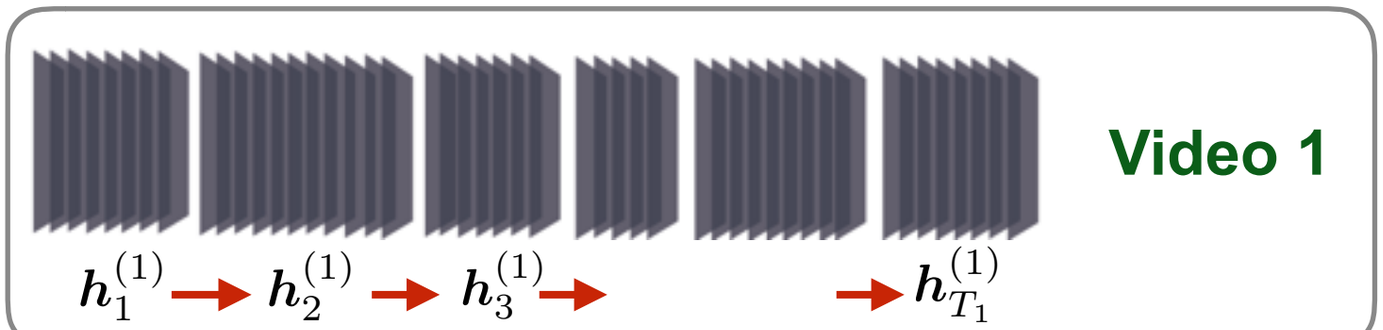
2) Extract **deep features** for each superframe

3) Build a dynamic model: **1st order HMM**

4) **SeqFL** (X = HMM, Y = test videos)

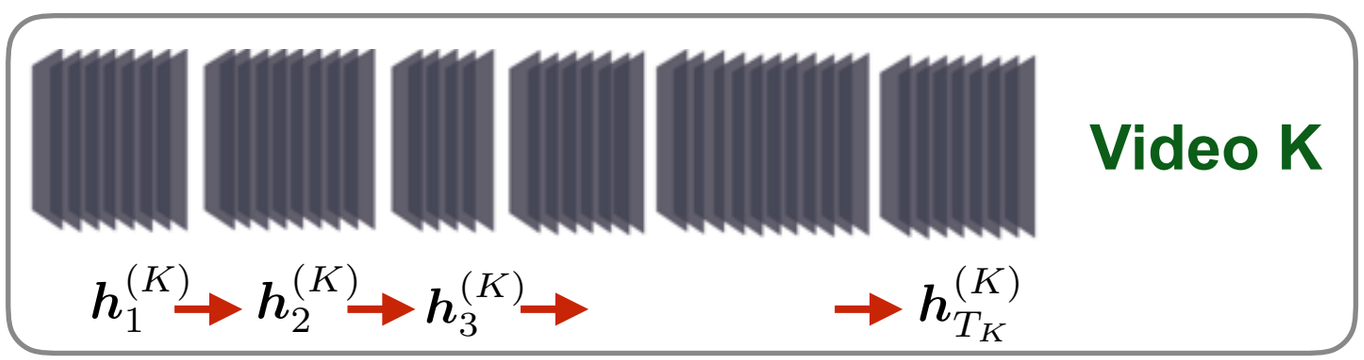


SeqFL



$r_1^{(1)}, r_2^{(1)}, r_3^{(1)}, \ldots, r_{T_1}^{(1)}$

SeqFL



$r_1^{(K)}, r_2^{(K)}, r_3^{(K)}, \ldots, r_{T_K}^{(K)}$

E. Elhamifar, C. De Paolis Kaluza, NIPS, 2017

Procedure learning via SeqFL

1) Extract **superframes** (Gygli et al ’14) from each video

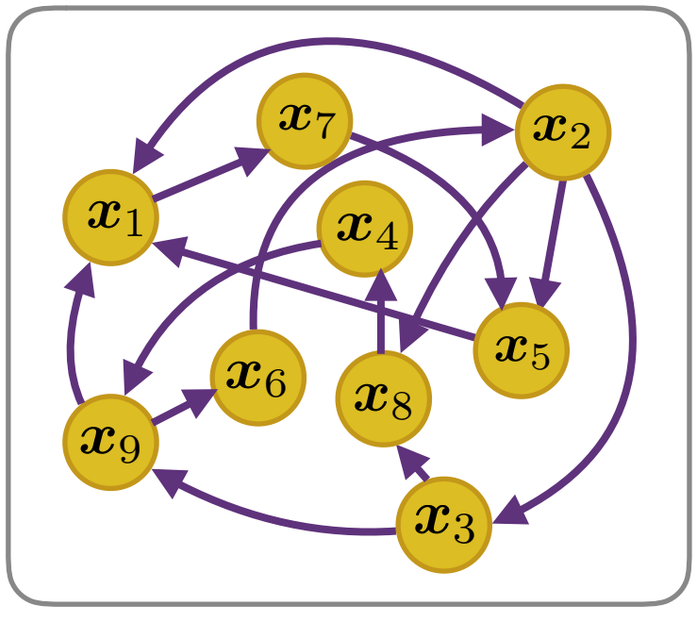
2) Extract **deep features** for each superframe

3) Build a dynamic model: **1st order HMM**

4) **SeqFL** (X = HMM, Y = test videos)

5) **Align sequences** of representatives

6) **Map** representatives to **actions**



$r_1^{(1)}, r_2^{(1)}, r_3^{(1)}, \ldots, r_{T_1}^{(1)}$ **Video 1 summary**

$r_1^{(K)}, r_2^{(K)}, r_3^{(K)}, \ldots, r_{T_K}^{(K)}$ **Video K summary**

$$x_5 \rightarrow x_2 \rightarrow x_9 \rightarrow x_3 \rightarrow x_8 \rightarrow x_3 \rightarrow x_8$$

Check Response → Open Airway → Check Breathing → Give Breath → Give Compression → Give Breath → Give Compression

E. Elhamifar, C. De Paolis Kaluza, NIPS, 2017

Experimental results

- **Instructional video dataset** (Alayrac et al CVPR ’16)
  - **5 tasks**, **30 videos/task**, labels of key steps given

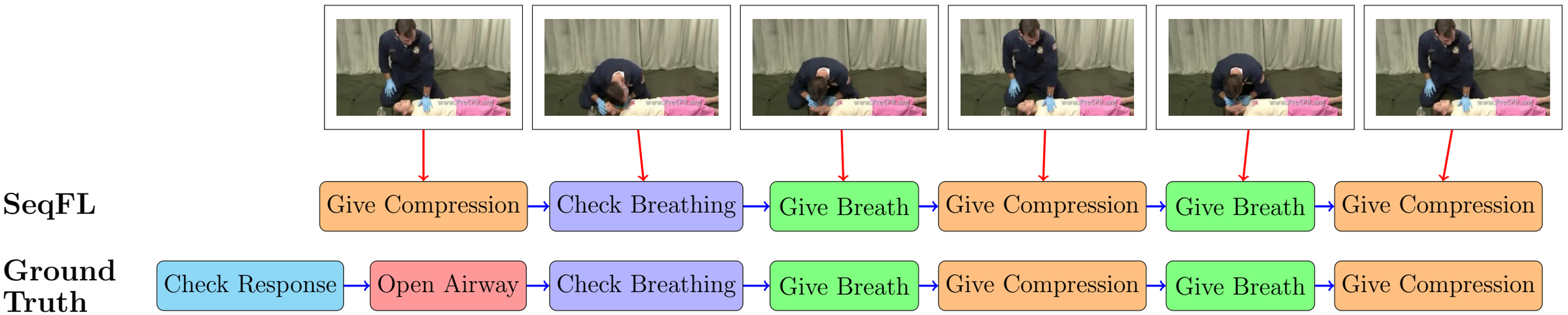
Task		kDPP	M-kDPP	Seq-kDPP	DS3	SeqFL
Change tire	(P, R)	(0.56, 0.50)	(0.55, 0.60)	(0.44, 0.40)	(0.56, 0.50)	(0.60, 0.60)
	F-score	0.53	0.57	0.42	0.53	**0.60**
Make coffee	(P, R)	(0.38, 0.33)	(0.50, 0.44)	(0.63, 0.56)	(0.50, 0.56)	(0.50, 0.56)
	F-score	0.35	0.47	**0.59**	0.53	0.53
CPR	(P, R)	(0.71, 0.71)	(0.71, 0.71)	(0.71, 0.71)	(0.71, 0.71)	(0.83, 0.71)
	F-score	0.71	0.71	0.71	0.71	**0.77**
Jump car	(P, R)	(0.50, 0.50)	(0.56, 0.50)	(0.56, 0.50)	(0.50, 0.50)	(0.60, 0.60)
	F-score	0.50	0.53	0.53	0.50	**0.60**
Repot plant	(P, R)	(0.57, 0.67)	(0.60, 0.50)	(0.57, 0.67)	(0.57, 0.67)	(0.80, 0.67)
	F-score	0.62	0.55	0.62	0.62	**0.73**
All tasks	(P, R)	(0.54, 0.54)	(0.58, 0.55)	(0.58, 0.57)	(0.57, 0.59)	(0.67, 0.63)
	F-score	0.54	0.57	0.57	0.58	**0.65**

No dynamics **with dynamics**

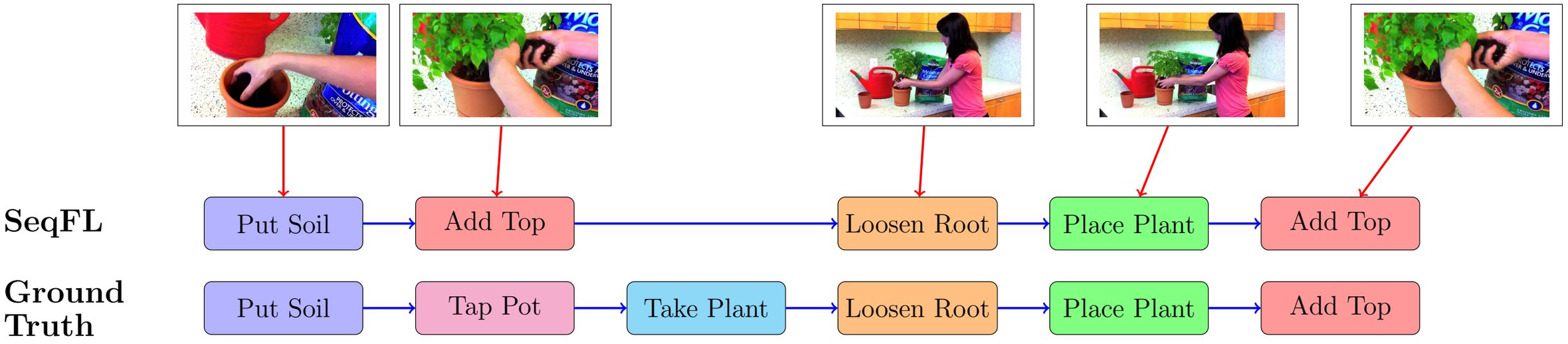
E. Elhamifar, C. De Paolis Kaluza, NIPS, 2017

Experimental results

- How to **'perform CPR'**



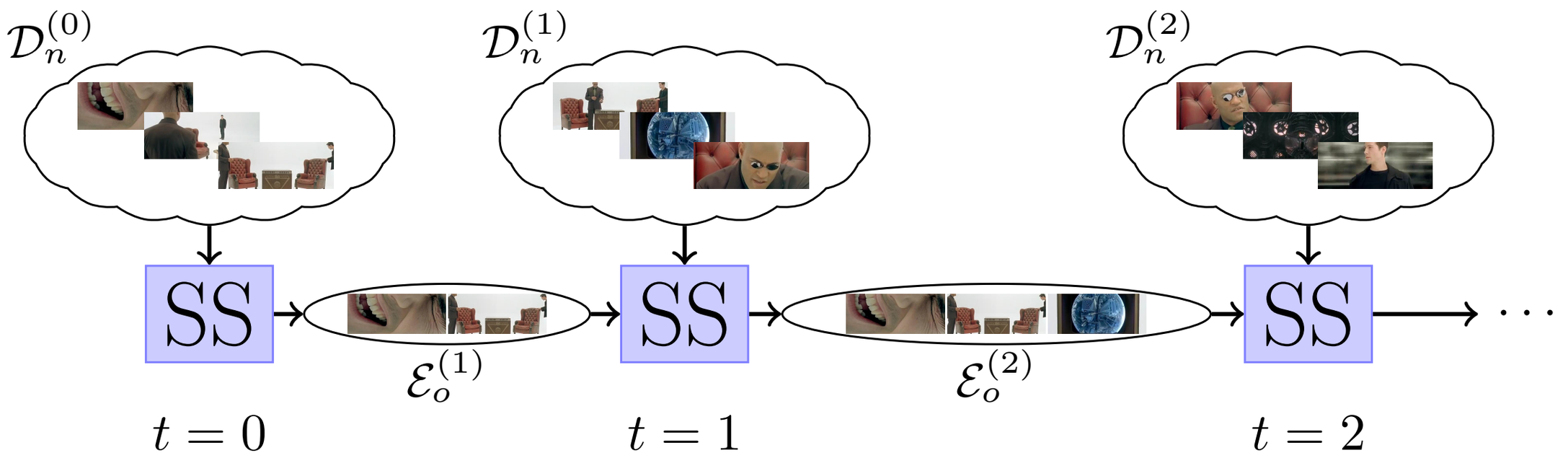
- How to **'repot a plant'**



E. Elhamifar, C. De Paolis Kaluza, NIPS, 2017

Online summarization

- **Sequential data**: grow **incrementally**, collecting all and running SS **infeasible**
- Approach: **incremental subset selection**



- Instead of running SS on $\bigcup_t \mathcal{D}_n^{(t)}$, run SS on $\mathcal{E}_o^{(t)} \cup \mathcal{D}_n^{(t)}$ at t
- Proved exactness of the online SS

Summary and ongoing work

- Important to have summarization close to **global optimum**
  - Developed **efficient sparse optimization** for **facility location**
  - Showed **exactness** under appropriate condition

- Important to handle **sequential data**: **compatible seq of reps**
  - Developed **SeqFL** for sequential summarization
  - Addressed **procedure learning** from **instructional videos**

Thanks!

Codes: http://www.ccs.neu.edu/home/eelhami/codes.htm