

Summarization in Sequential Data with Applications to Procedure Learning

Ehsan Elhamifar

**College of Computer and Information Science
Northeastern University**

Data deluge



> 400hrs videos / minute

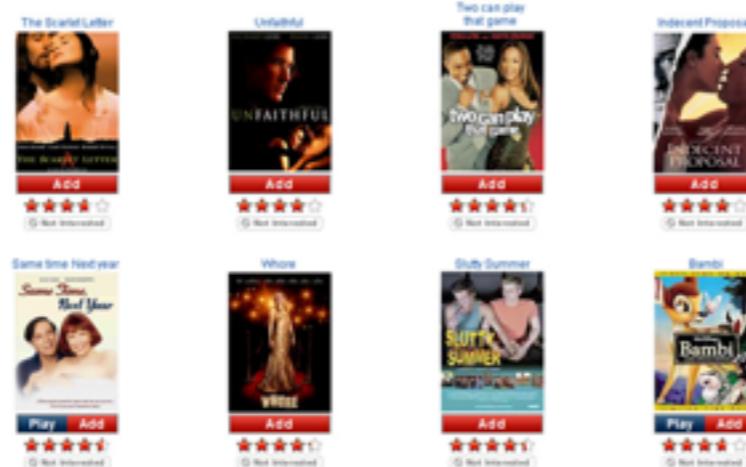


> 300M photos / day

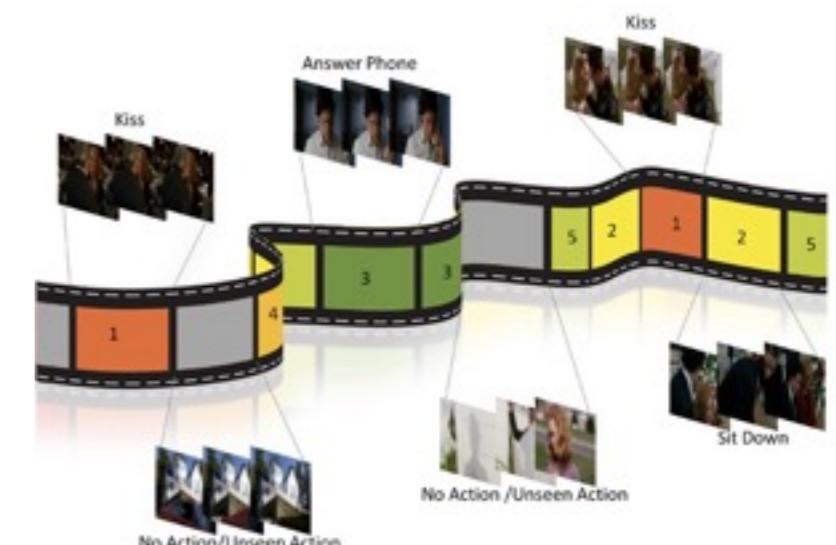


Congratulations! Movies we think You will ❤️

Add movies to your Queue, or Rate ones you've seen for even better suggestions.



24/7/365



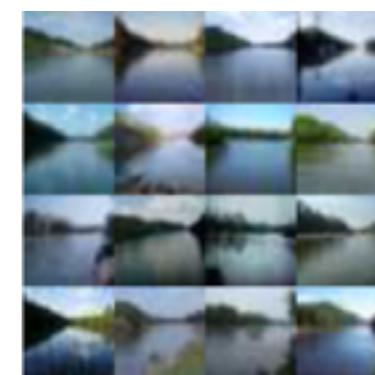
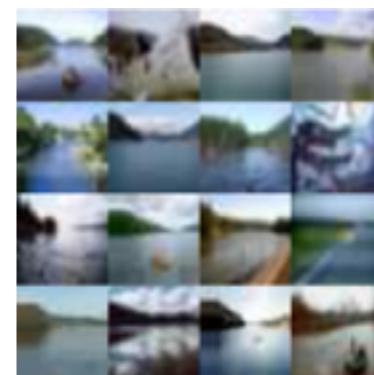
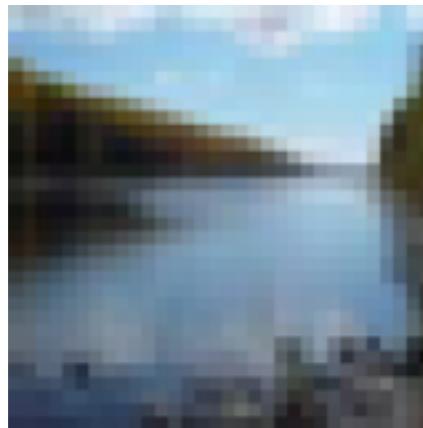
azing awesome bad band beautiful berlin best better book chocolate coffee come coming congrats cool creative ch definitely dinner divx doing editing email end epic-fu feeling feels finished food forward @frankjonen free frier me getting girl giving glad goes going google gr hah halo happy having head heading hear hearing ho setshow.com ill internet iphone @irinaslutsky isnt jetse left life listening little live lol long look looking looks ke makes making man maybe media missed morning m nice night nyc obama old omg online panel park party phone pixelodeon play playing post pretty reading rey right rock run said saw say says @seanbonner serio ls @spin @spytap start steve @stevewoolf stop strik ing teh tell thanks @thefemgeek theyre thing things t thx time tiny today tomorrow tonight @tonykatz try !O vote wait waiting want watch watching web week w wondering WORK working world writers writing wrong yay

Big data opportunities

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



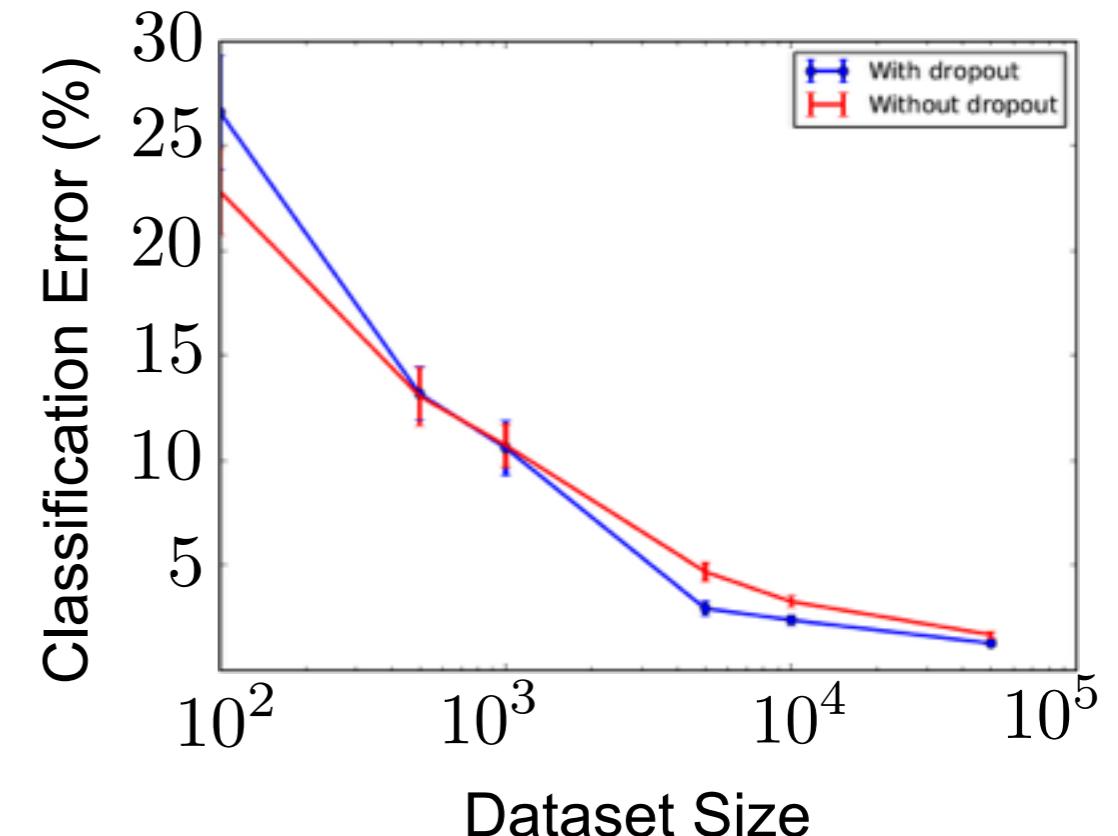
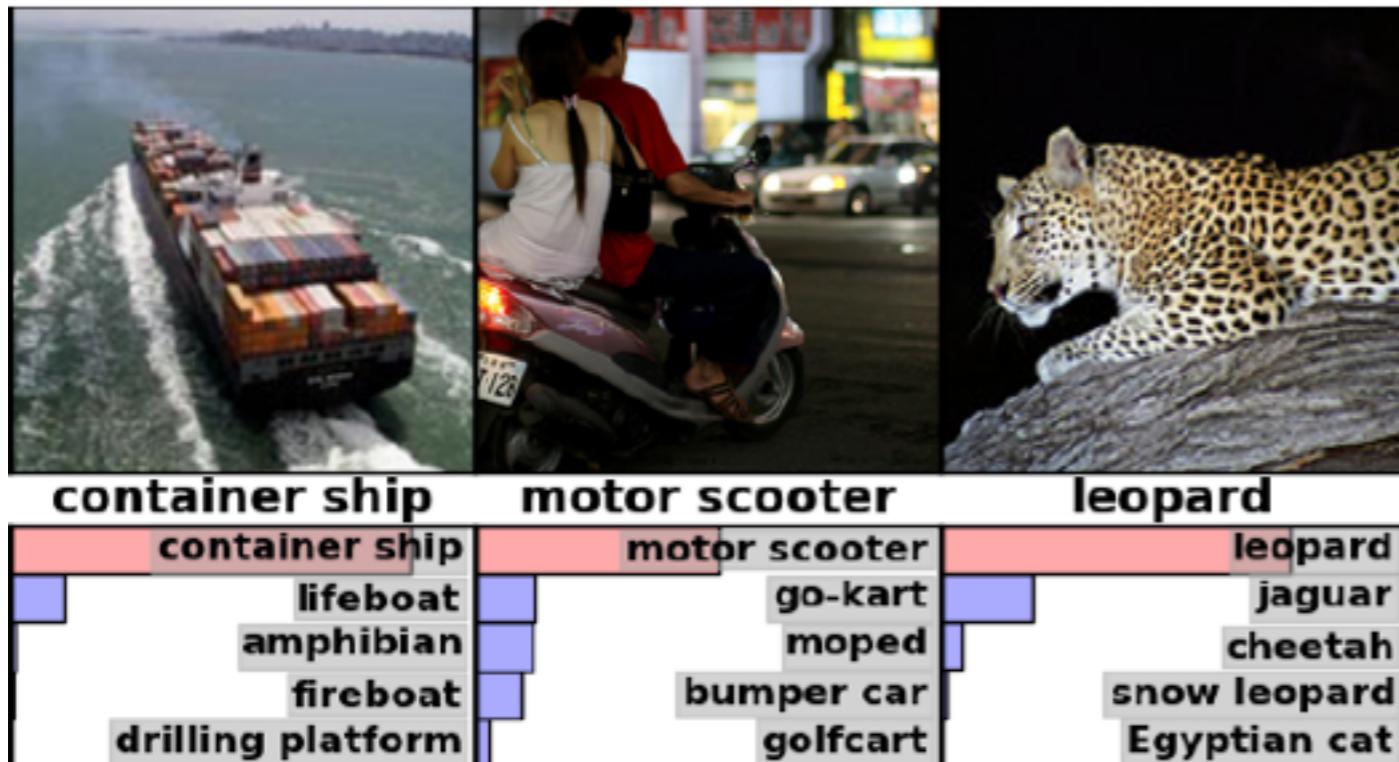
“It takes a large amount of data for our method to succeed. We saw **dramatic improvement** when moving from ten thousand to two million images.” [Hayes, Efros, SIGGRAPH ’07]



“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

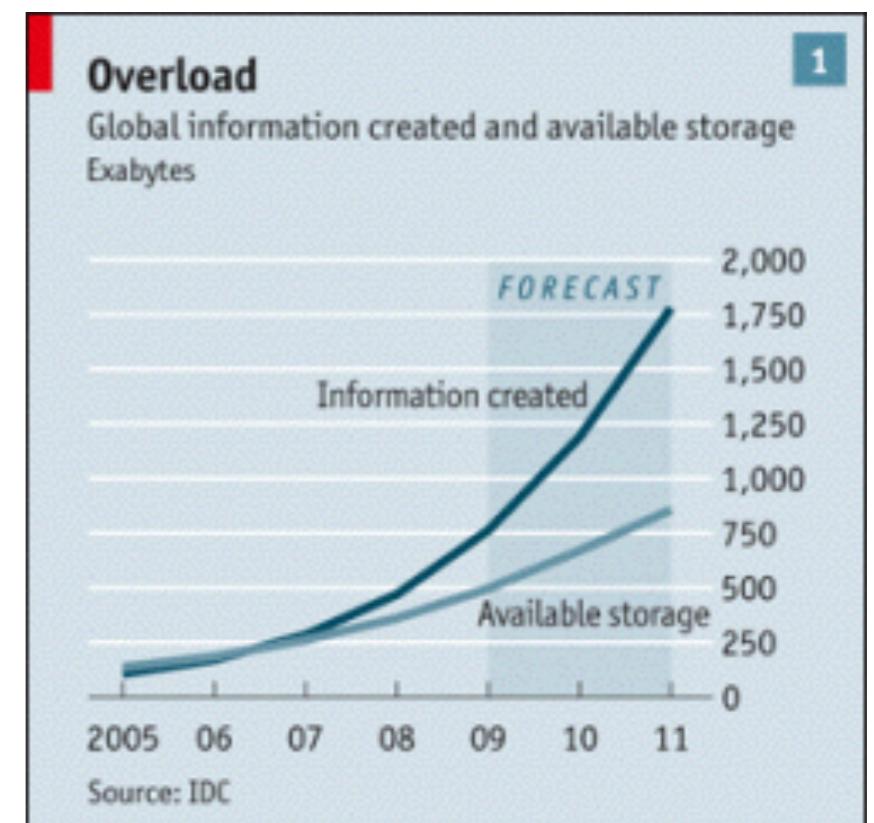
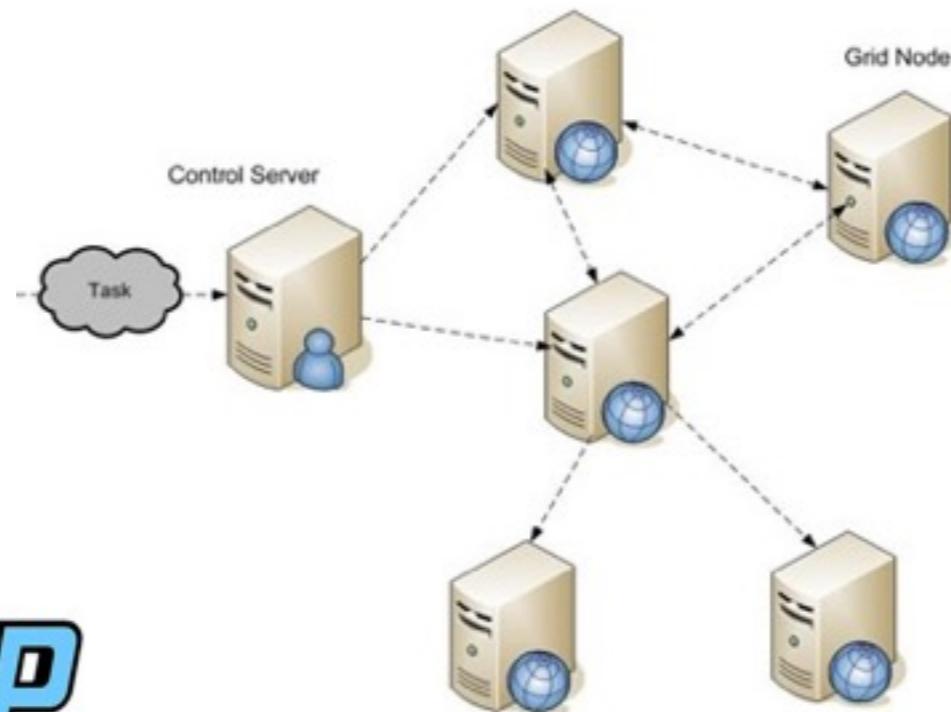
- Capturing modes and patterns **unseen** in small data
 - Increase learning and inference performance
- To **Deep Neural Networks**



“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

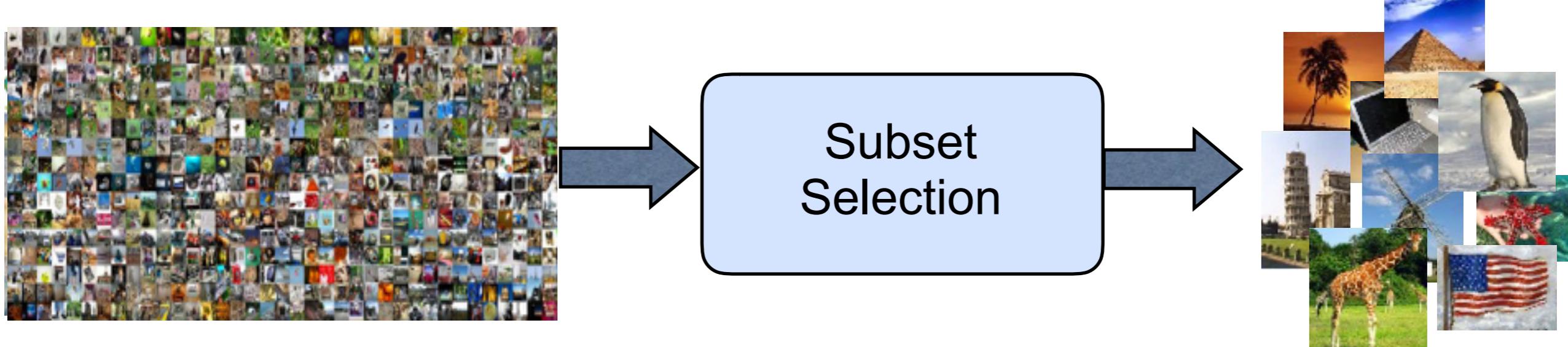
- Capturing modes and patterns **unseen** in small data
 - Increase learning and inference performance: from NNs to DNNs
- Limiting **computational** and **memory** resources
 - We have **more data** than capacities of our resources



Economist, Feb 2010

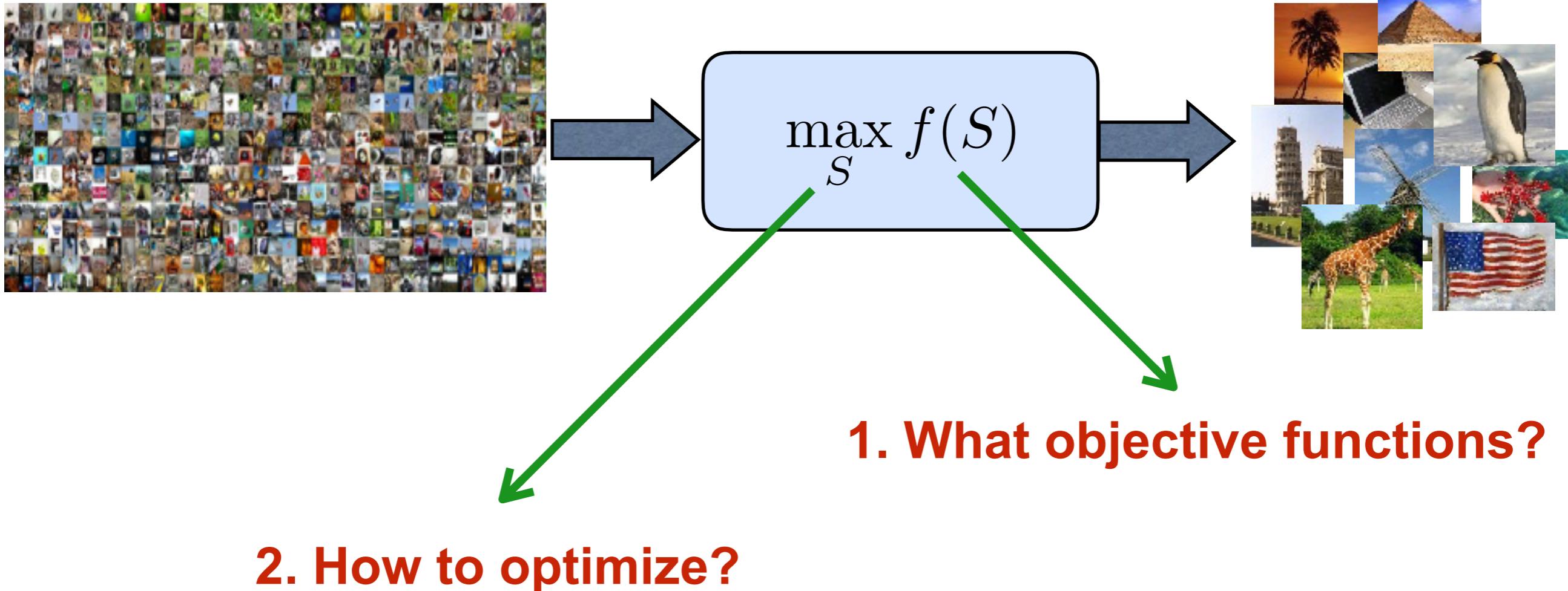
Subset selection

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

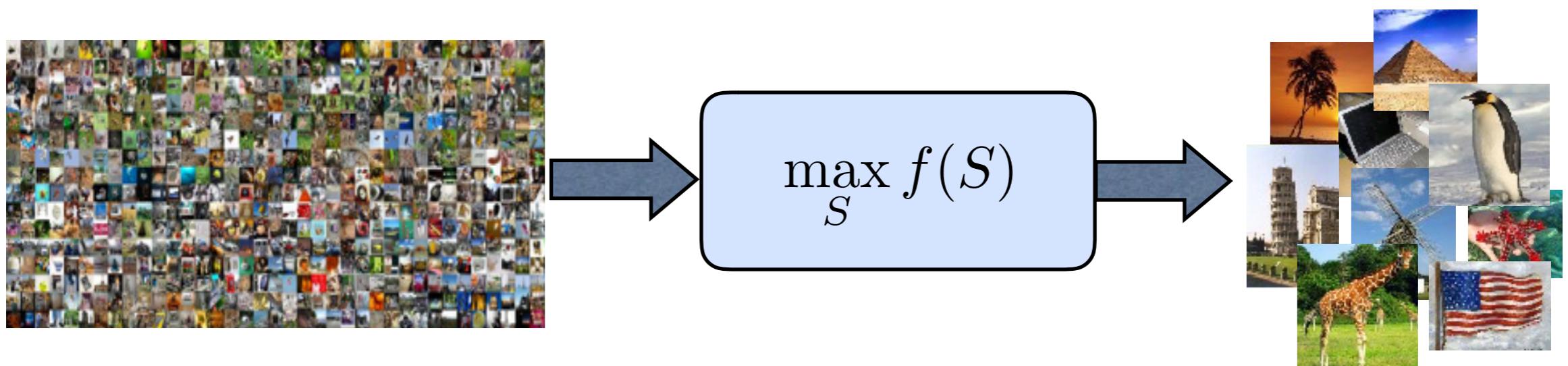


Subset selection components

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



This Talk



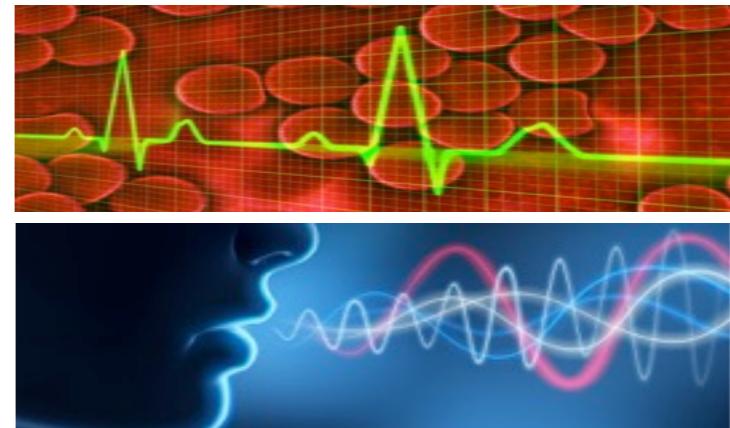
- Subset selection is **NP-hard**, hence, **approximate** algorithms
 - **Better summarization**: closer to **global optimum**
- Focus on one class of objective functions: **Facility Location**
 - Develop **efficient sparse optimization**
 - Show it can **achieve global optimum (under certain condition)**

This Talk

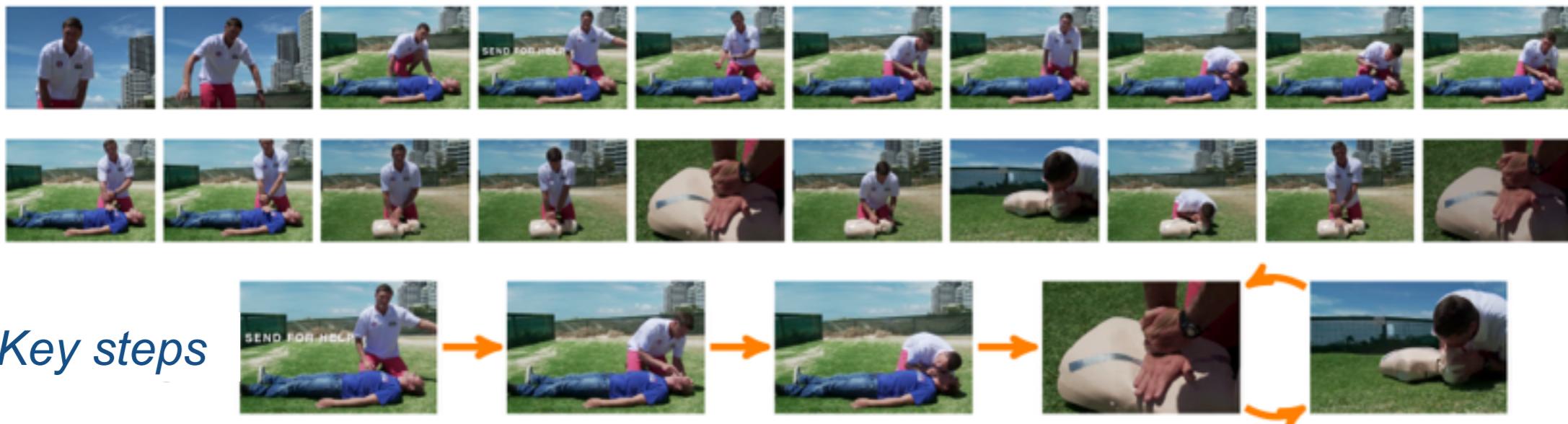
- **Sequential data** (video, text, signals,...): large part of ‘Big Data’



(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

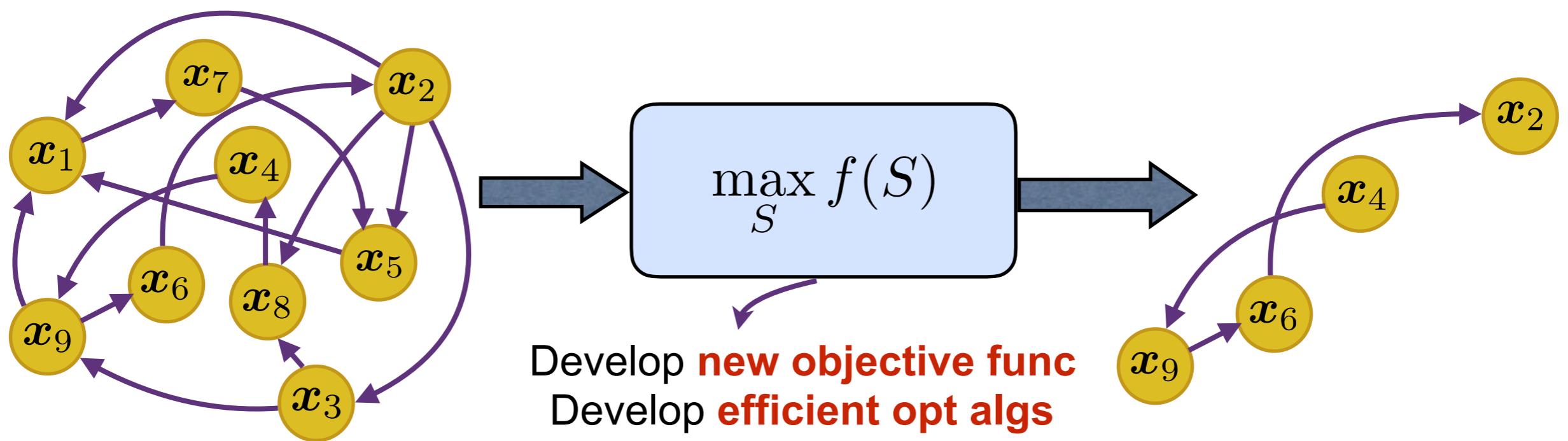


- Sequential data have **structured dependencies**
 - **Instructional videos**: ordering of key steps



This Talk

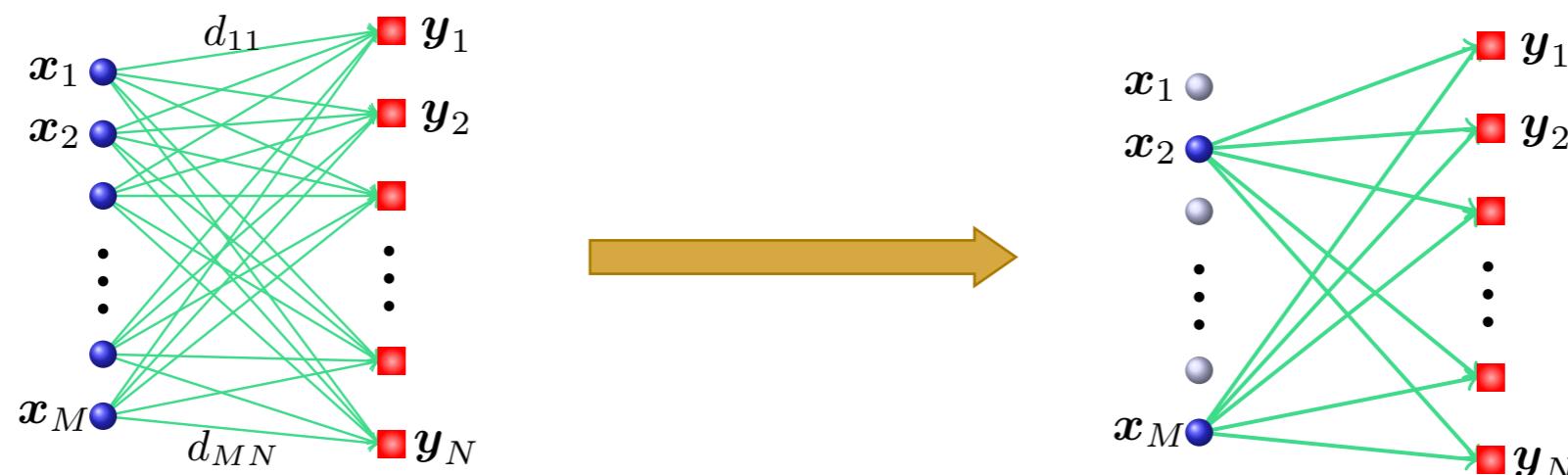
- **Sequential data** (video, text, signals,...): large part of ‘Big Data’
- Sequential data have **structured dependencies**
 - Existing SS: data is a **bag of randomly permutable items**
- **Generalize facility location** to handle **seq summarization**



Subset selection from dissimilarities

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

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

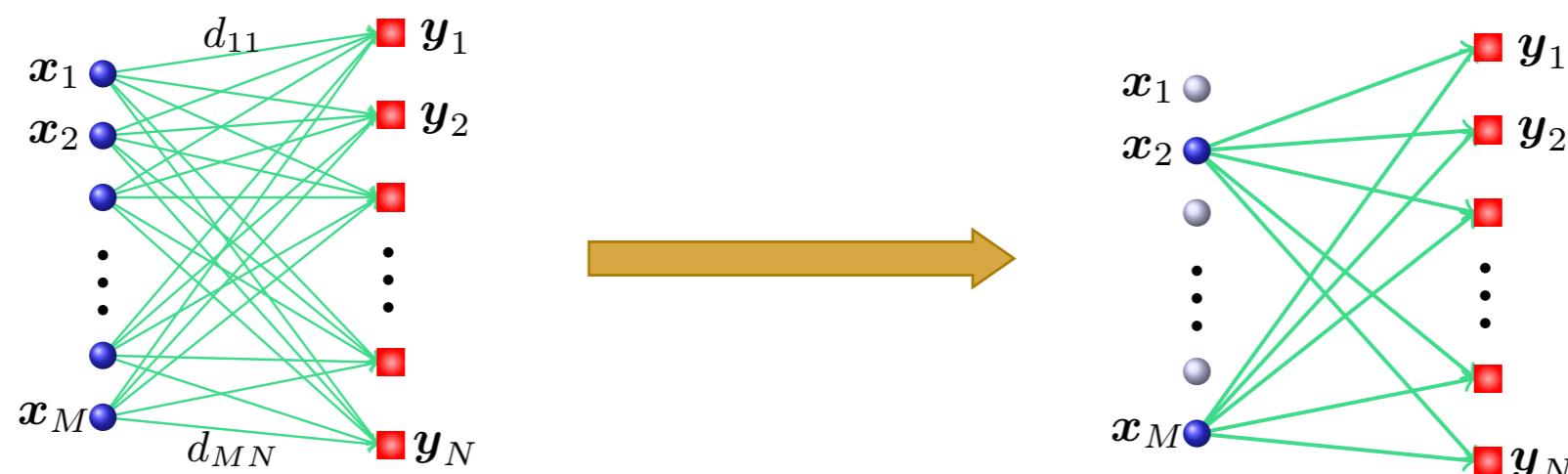


- $d(x_i, y_j) = d_{ij}$: how well x_i represents y_j
 - $\mathbb{X} = \text{data}, \mathbb{Y} = \text{data} \xrightarrow{} d(\cdot, \cdot) = \text{Euclidean/geodesic distance}$
 - $\mathbb{X} = \text{models}, \mathbb{Y} = \text{data} \xrightarrow{} d(\cdot, \cdot) = \text{representation/coding error}$

Subset selection from dissimilarities

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

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



- $d(x_i, y_j) = d_{ij}$: how well x_i represents y_j

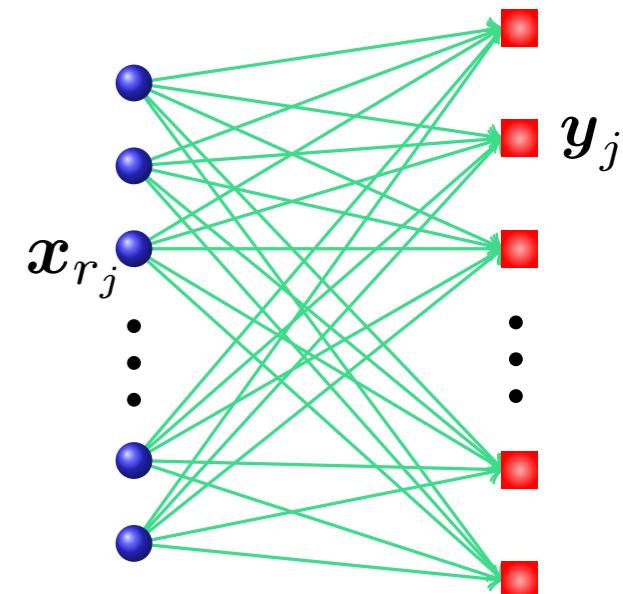
Not necessarily metric:
asymmetric,
violate triangle ineq



Dissimilarity-based Sparse Subset Selection

- Let $r_j \in \{1, \dots, M\}$ be index of representative of y_j
- **Approach:** define & maximize **potential function**

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



$$\Phi_{\text{enc}}(r_1, \dots, r_N) \triangleq \prod_{j=1}^N e^{-d_{r_j,j}}$$



Prefer lower encoding

$$\Phi_{\text{card}}(r_1, \dots, r_N) \triangleq e^{-\lambda |\{r_1, \dots, r_N\}|}$$

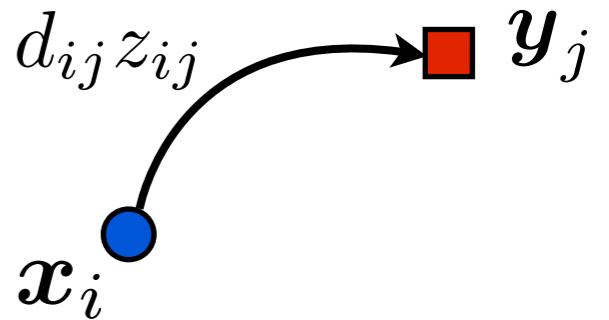


Prefer small # representatives

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

DS3 formulation

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



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

- Minimizing $-\log \Psi$:

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

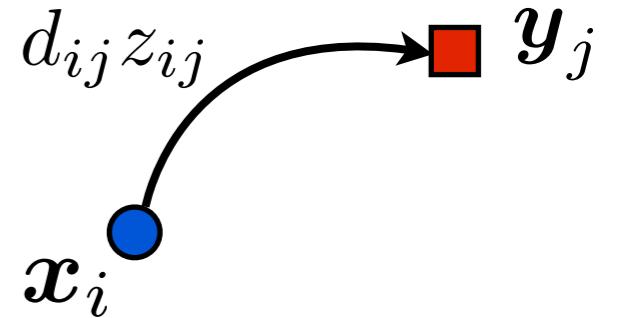
Nonconvex

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

~ Facility location

DS3 formulation

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



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

- Minimizing $-\log \Psi$:

Convex

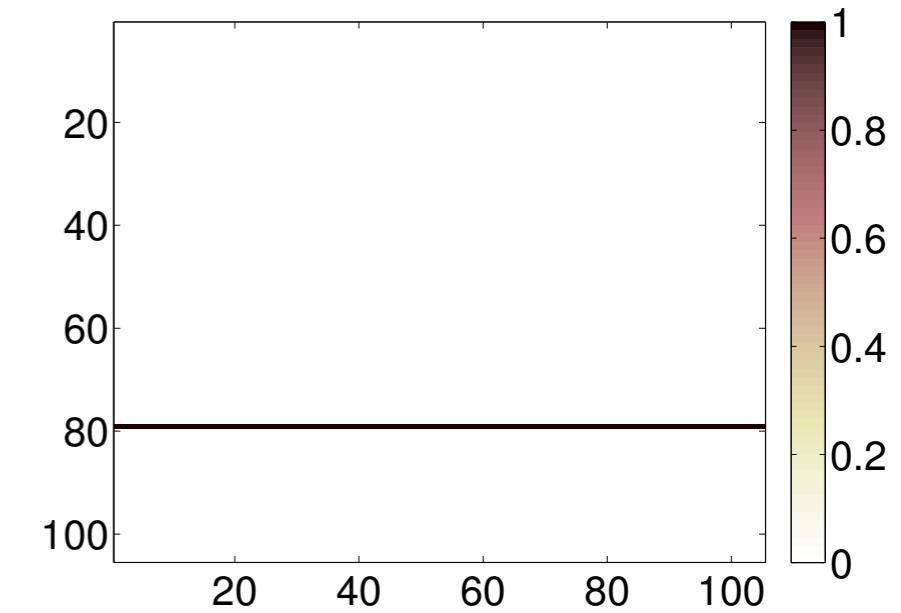
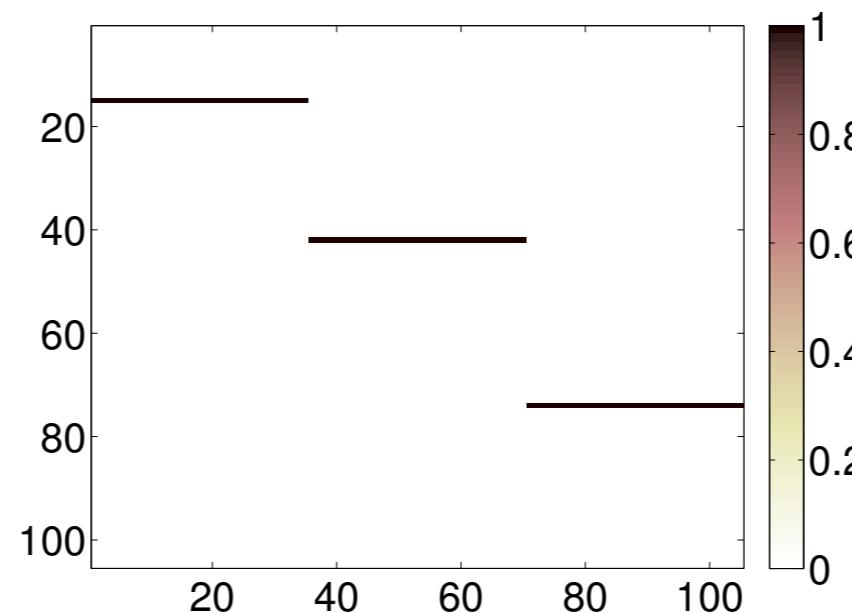
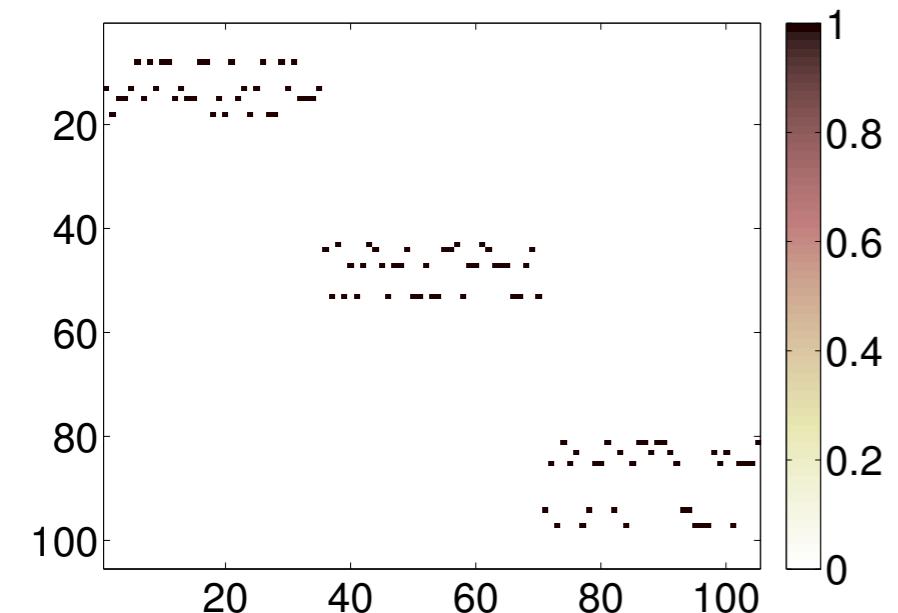
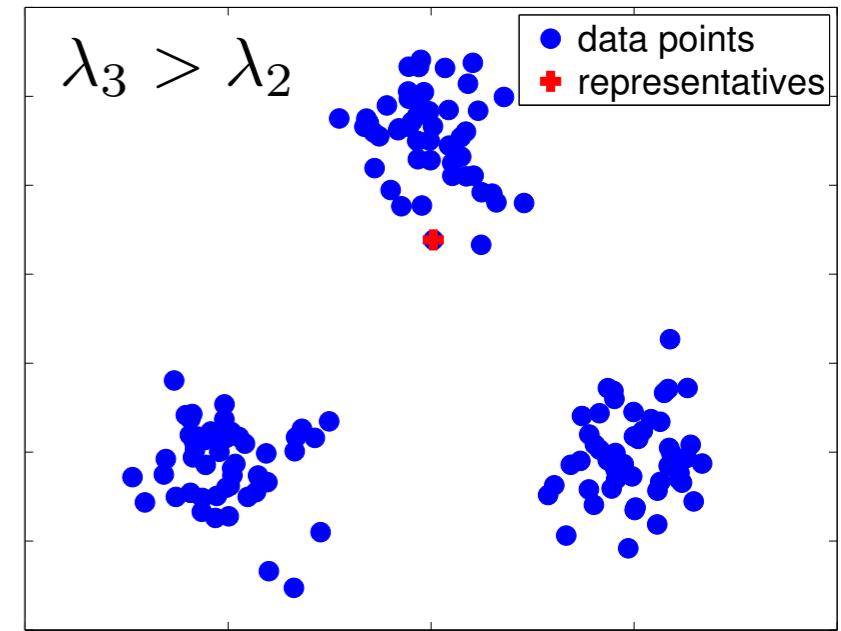
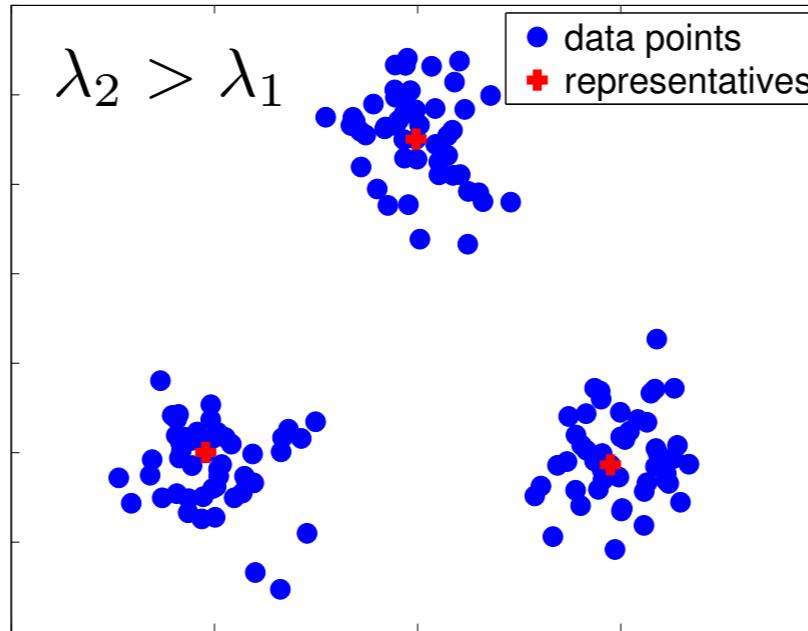
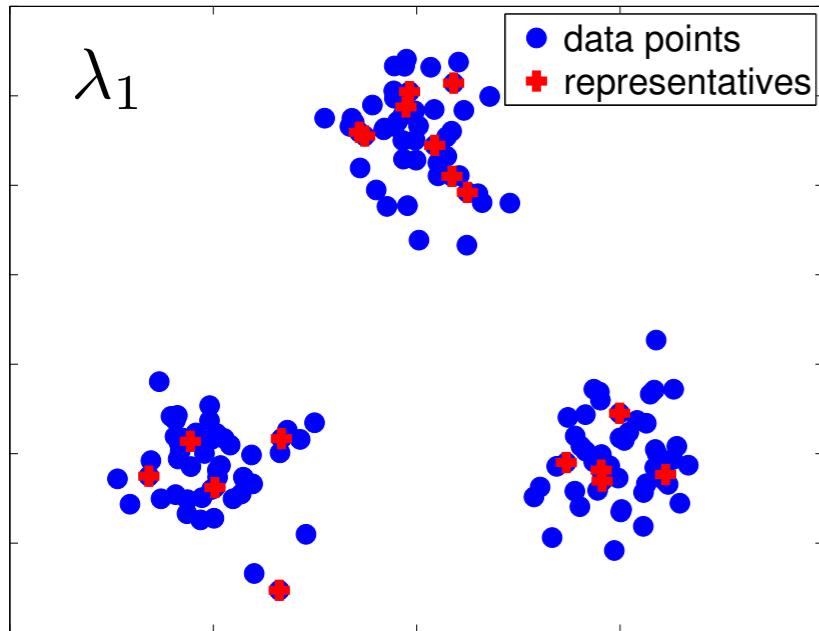
$$\begin{aligned} & \min_{\{z_{i,j}\}} \sum_{i=1}^M \sum_{j=1}^N d_{i,j} z_{i,j} + \lambda \sum_{i=1}^M \| [z_{i,1} \dots z_{i,N}] \|_p \\ & \text{s. t. } z_{i,j} \geq 0, \quad \sum_{i=1}^M z_{i,j} = 1, \quad \forall i, j \end{aligned}$$

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

DS3

Toy example

- Source = target = {data points}



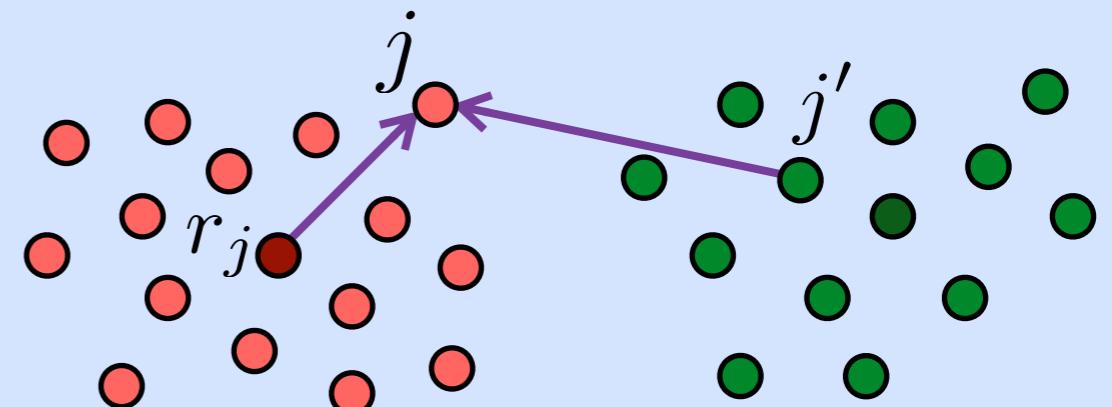
$$\min_{\{z_{i,j}\}} \sum_{i=1}^M \sum_{j=1}^N d_{i,j} z_{i,j} + \lambda \sum_{i=1}^M \| [z_{i,1} \ \cdots \ z_{i,N}] \|_p \quad \text{s.t.} \quad z_{i,j} \geq 0, \quad \sum_{i=1}^M z_{i,j} = 1, \quad \forall i, j$$

Theoretical analysis

- When is **convex relaxation exact?**

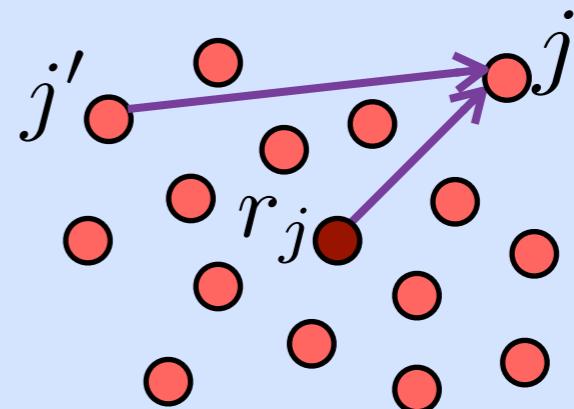
- **Theorem:** Let r_j be representative of cluster j via **non-convex optimization**. The convex relaxation is exact if

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



Clusters sufficiently separated

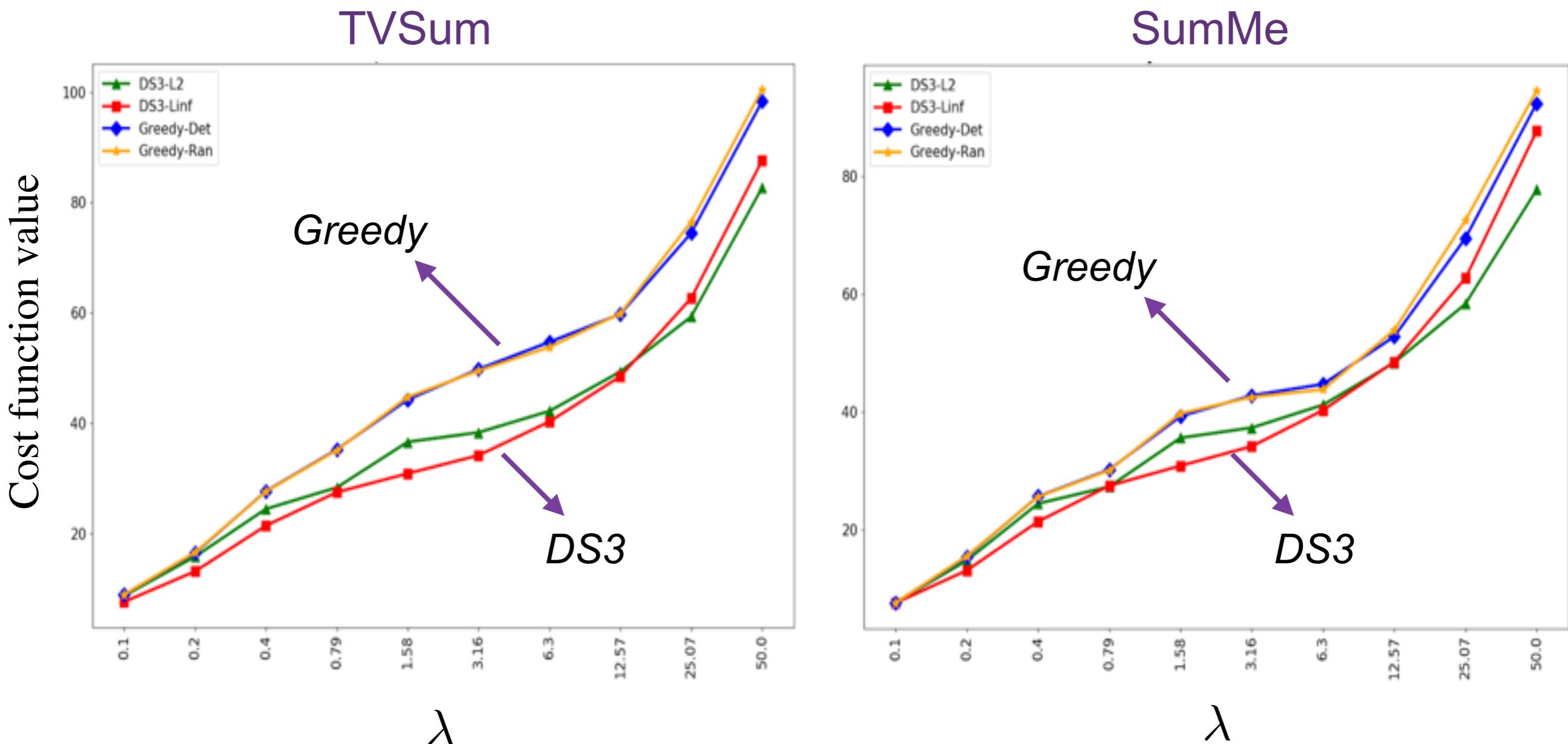
$$2) \quad d_{j', j} \leq \frac{\lambda}{N_{r_j}} + d_{r_j, j}$$



Sufficiently small radius

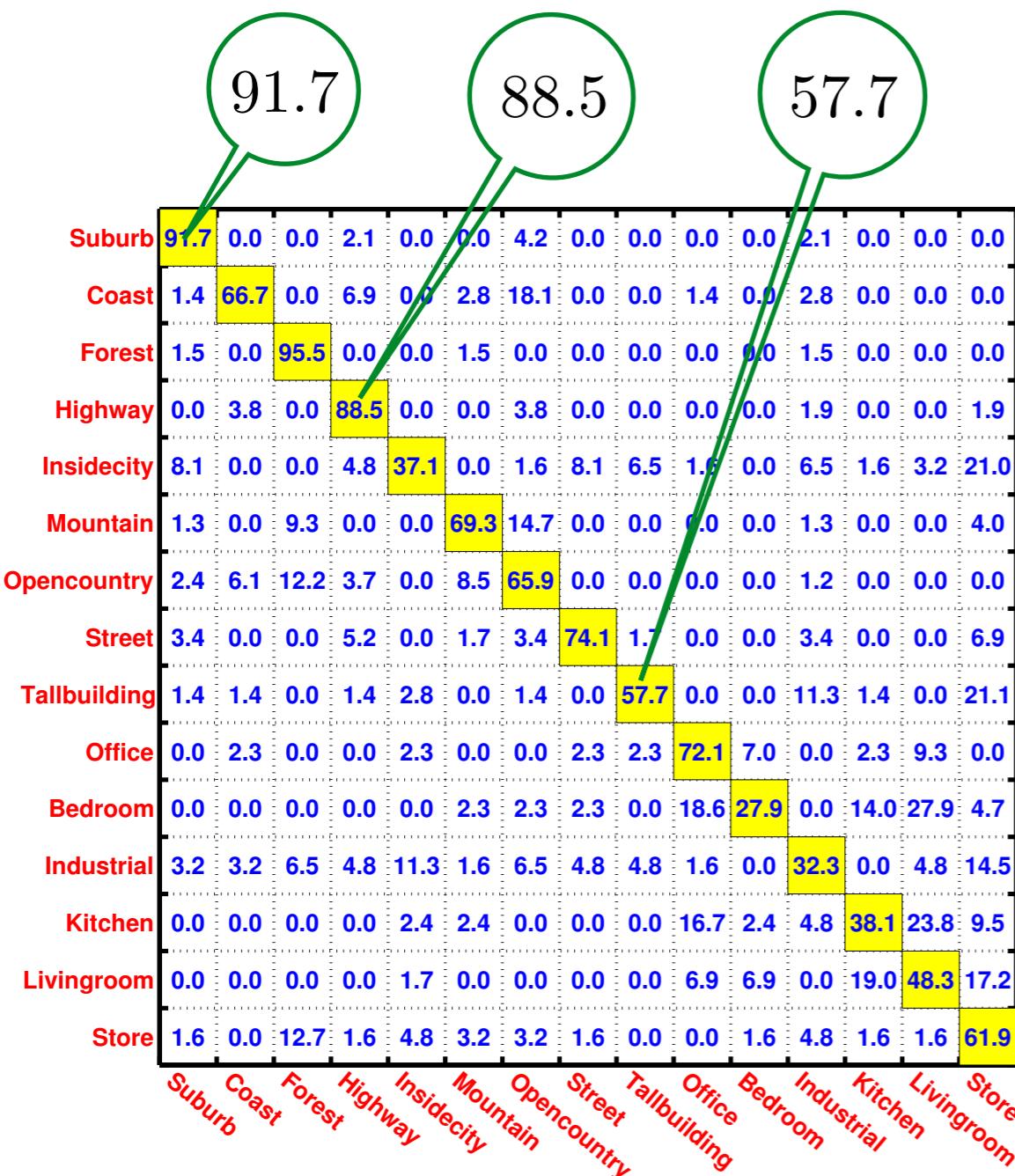
Comparison with greedy

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

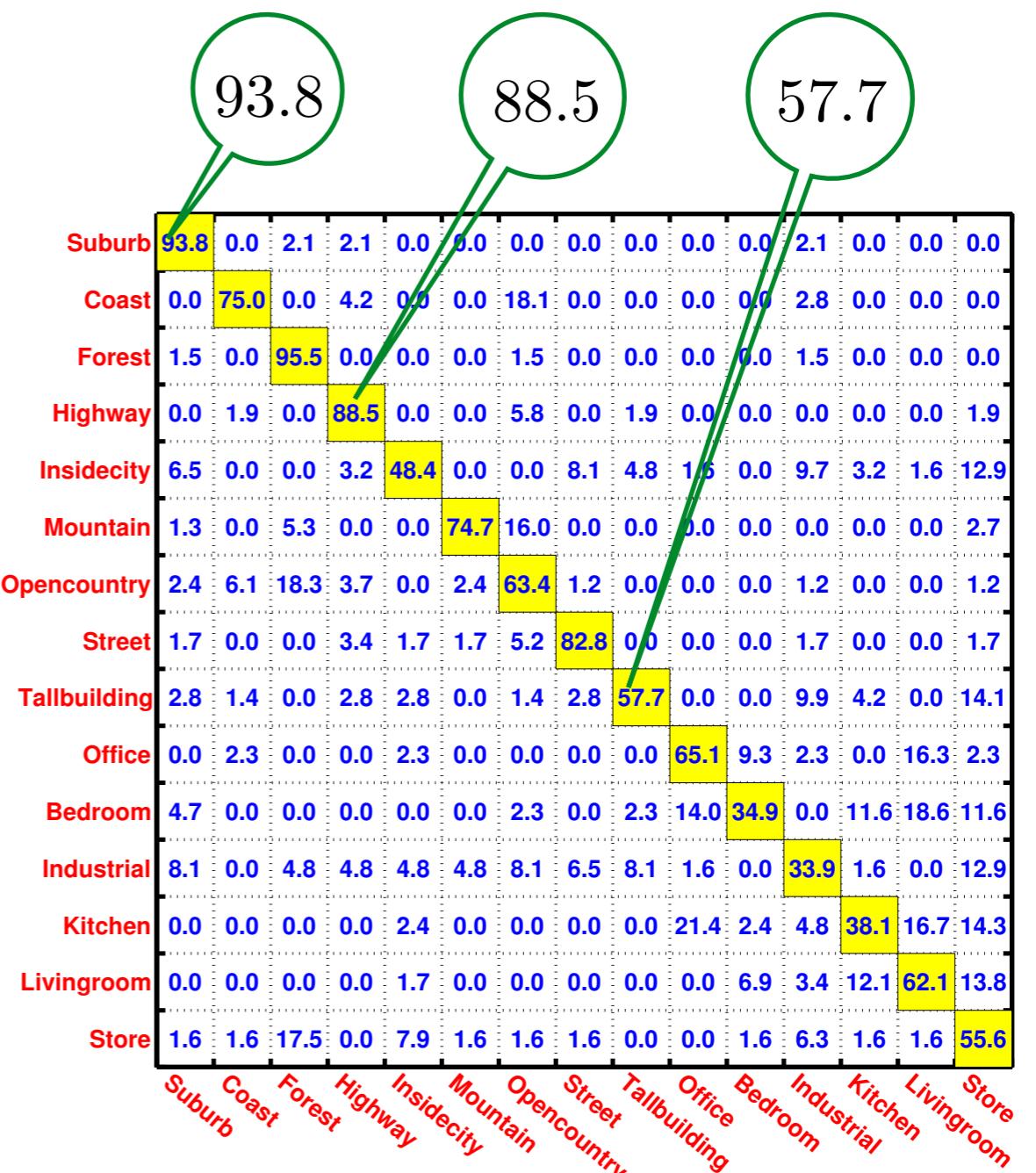


Recognition via representatives

- Select η fraction of training samples via DS3



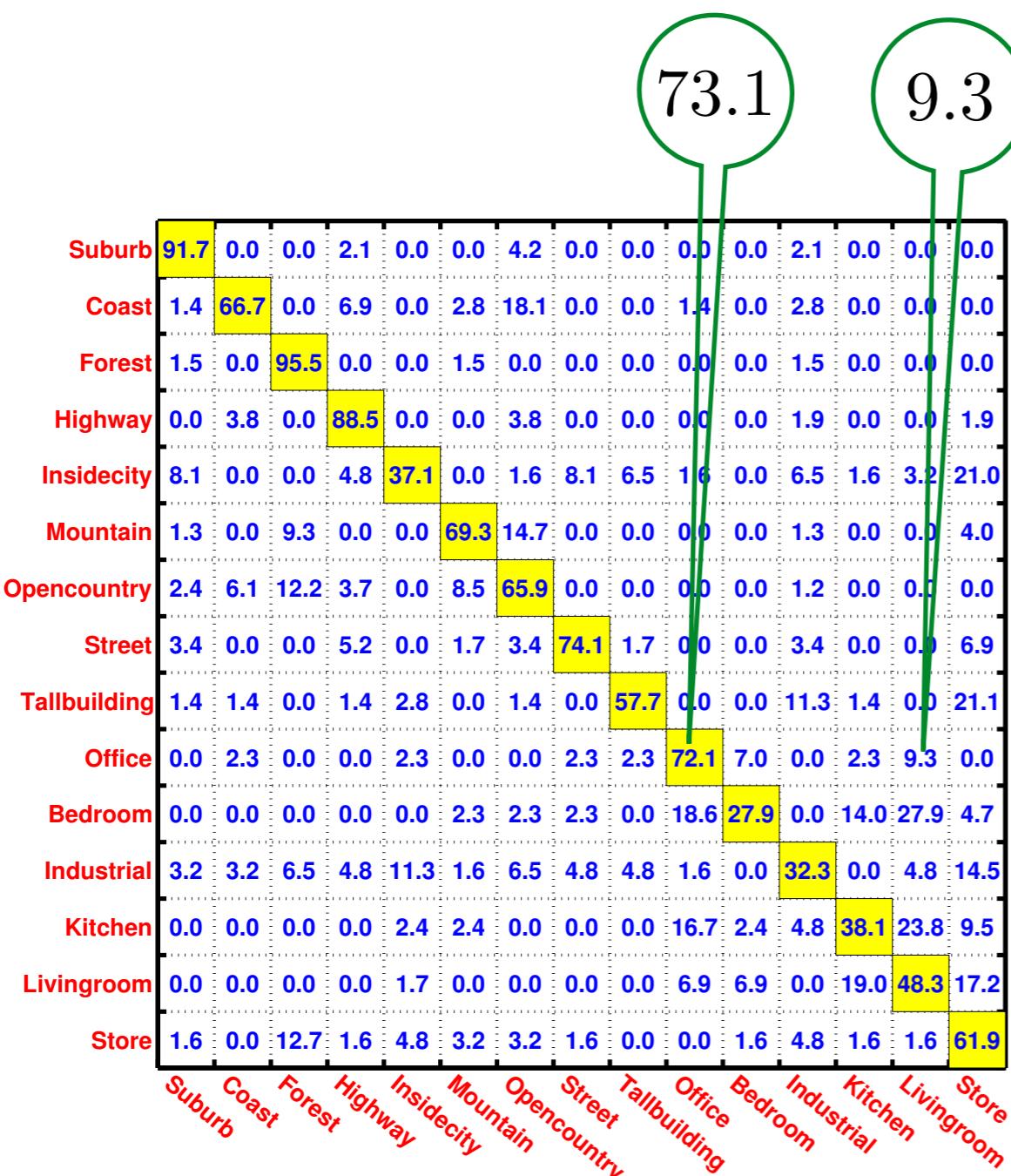
DS3 selecting 35% samples



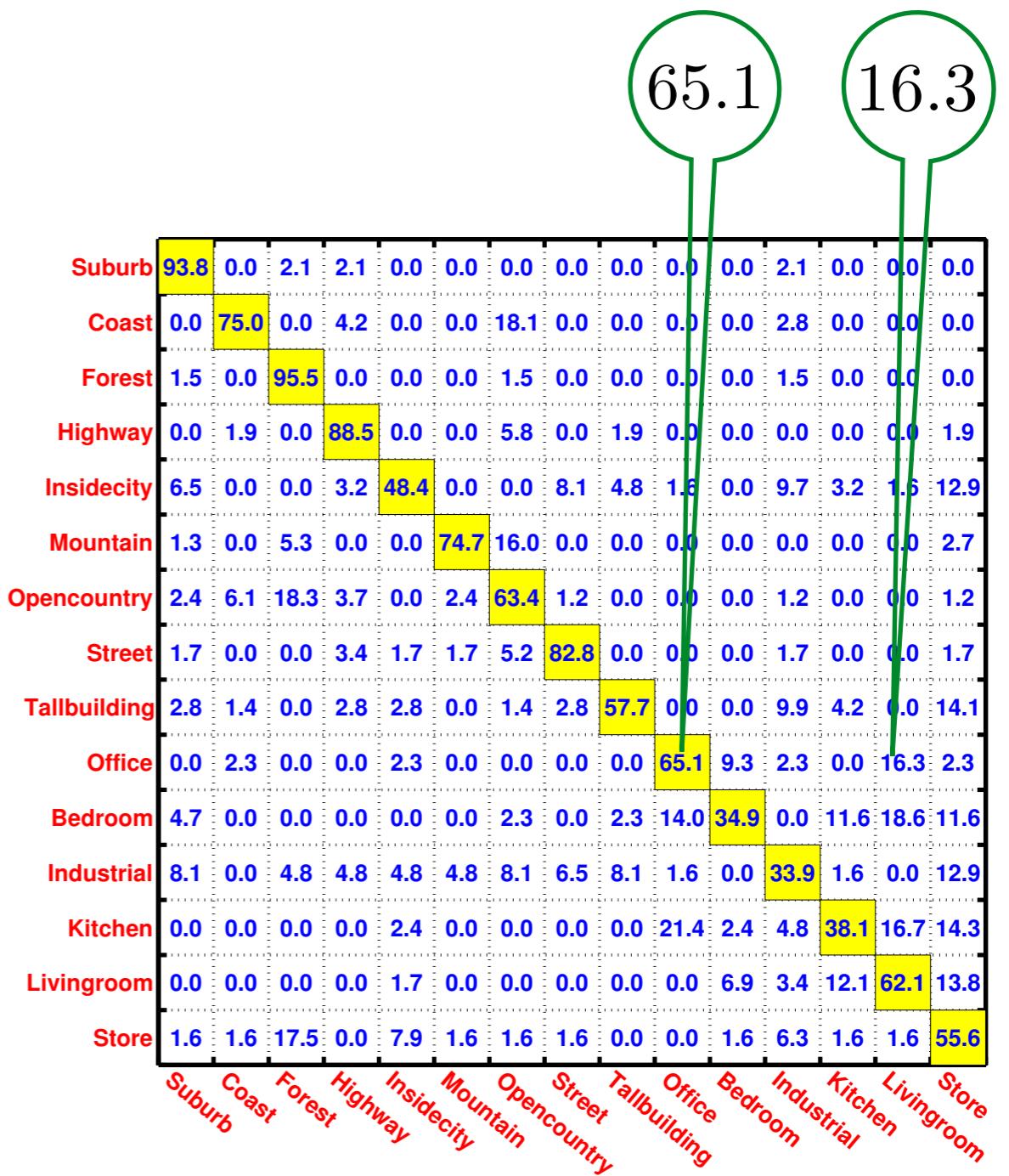
Trained via all samples

Recognition via representatives

- Select η fraction of training samples via DS3



DS3 selecting 35% samples



Trained via all samples

Recognition via representatives

- Select η fraction of training samples via DS3

	Suburb	Coast	Forest	Highway	InsideCity	Mountain	OpenCountry	Street	TallBuilding	Office	Bedroom	Industrial	Kitchen	LivingRoom	Store
Suburb	91.7	0.0	0.0	2.1	0.0	0.0	4.2	0.0	0.0	0.0	2.1	0.0	0.0	0.0	0.0
Coast	1.4	66.7	0.0	6.9	0.0	2.8	18.1	0.0	0.0	1.4	0.0	2.8	0.0	0.0	0.0
Forest	1.5	0.0	95.5	0.0	0.0	1.5	0.0	0.0	0.0	0.0	1.5	0.0	0.0	0.0	0.0
Highway	0.0	3.8	0.0	88.5	0.0	0.0	3.8	0.0	0.0	0.0	1.9	0.0	0.0	1.9	0.0
InsideCity	8.1	0.0	0.0	4.8	37.1	0.0	1.6	8.1	6.5	1.6	0.0	6.5	1.6	3.2	21.0
Mountain	1.3														
OpenCountry	2.4														
Street	3.4														
TallBuilding	1.4														
Office	0.0														
Bedroom	0.0	0.0	0.0	0.0	0.0	2.3	2.3	2.3	0.0	16.0	27.9	0.0	14.0	27.9	4.7
Industrial	3.2	3.2	6.5	4.8	11.3	1.6	6.5	4.8	4.8	1.6	0.0	32.3	0.0	4.8	14.5
Kitchen	0.0	0.0	0.0	0.0	2.4	2.4	0.0	0.0	16.7	2.4	4.8	38.1	23.8	9.5	
LivingRoom	0.0	0.0	0.0	0.0	1.7	0.0	0.0	0.0	6.9	6.9	0.0	19.0	48.3	17.2	
Store	1.6	0.0	12.7	1.6	4.8	3.2	3.2	1.6	0.0	0.0	1.6	4.8	1.6	1.6	61.9

9.3

	Suburb	Coast	Forest	Highway	InsideCity	Mountain	OpenCountry	Street	TallBuilding	Office	Bedroom	Industrial	Kitchen	LivingRoom	Store
Suburb	93.8	0.0	2.1	2.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Coast	0.0	75.0	0.0	4.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Forest	1.5	0.0	95.5	0.0	0.0	1.5	0.0	0.0	0.0	0.0	1.5	0.0	0.0	0.0	0.0
Highway	0.0	1.9	0.0	88.5	0.0	0.0	0.0	0.0	0.0	0.0	5.8	0.0	1.9	0.0	1.9
InsideCity	6.5	0.0	0.0	3.2	48.4	0.0	0.0	8.1	4.8	1.6	0.0	9.7	3.2	1.6	12.9

16.3

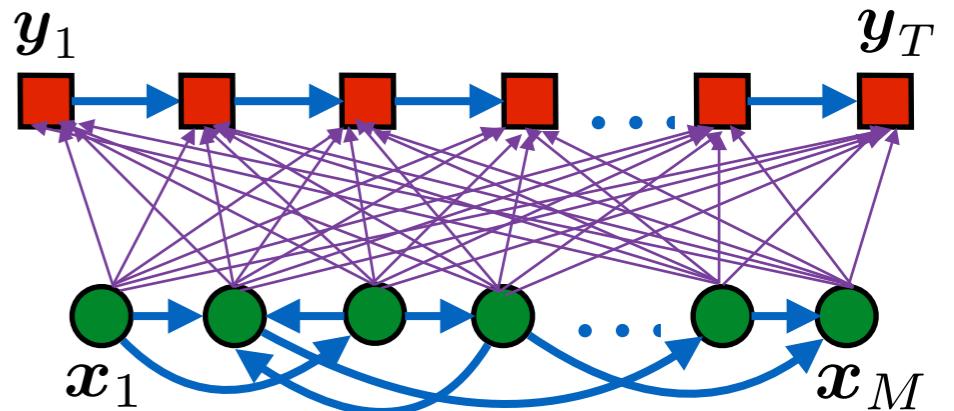
Representatives improve computational time and can improve performance

	Bedroom	Industrial	Kitchen	LivingRoom	Store	Suburb	Coast	Forest	Highway	InsideCity	Mountain	OpenCountry	Street	TallBuilding	Office	Bedroom
Bedroom	4.7	0.0	0.0	0.0	0.0	2.3	2.3	0.0	14.0	34.9	0.0	11.6	18.6	11.6	0.0	0.0
Industrial	8.1	0.0	4.8	4.8	4.8	8.1	8.1	6.5	1.6	0.0	33.9	1.6	0.0	12.9	0.0	0.0
Kitchen	0.0	0.0	0.0	0.0	2.4	2.4	0.0	0.0	21.4	2.4	4.8	38.1	16.7	14.3	0.0	0.0
LivingRoom	0.0	0.0	0.0	0.0	1.7	0.0	0.0	0.0	0.0	0.0	6.9	3.4	12.1	62.1	13.8	0.0
Store	1.6	1.6	17.5	0.0	7.9	1.6	1.6	1.6	1.6	0.0	1.6	6.3	1.6	1.6	55.6	0.0

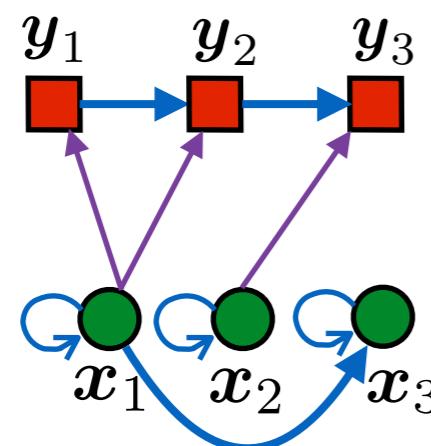
Sequential subset selection

- Incorporate **dynamics** of sequential data into SS

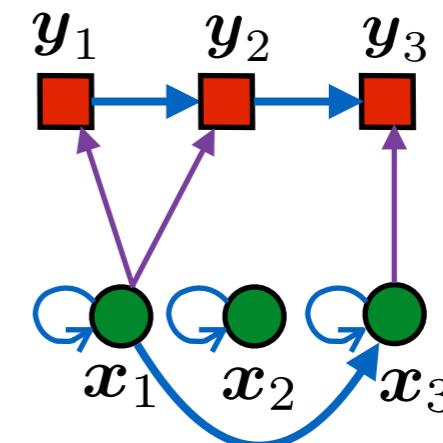
- Source set: with **transition dynamics**
- Target set: **sequential data**



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



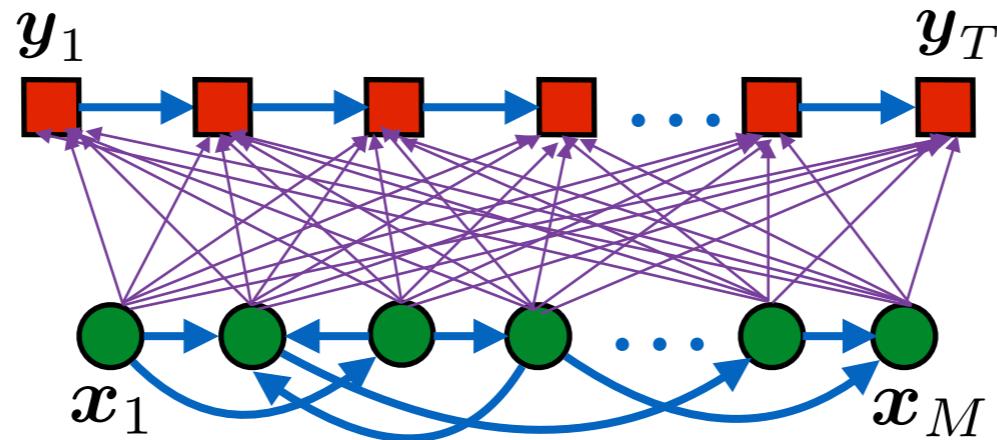
$$(r_1, r_2, r_3) = (1, 1, 2)$$



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

Sequential facility location

- Let r_t be index of representative of y_t



- Maximize global potential function

Prefer a sequence of compatible reps

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

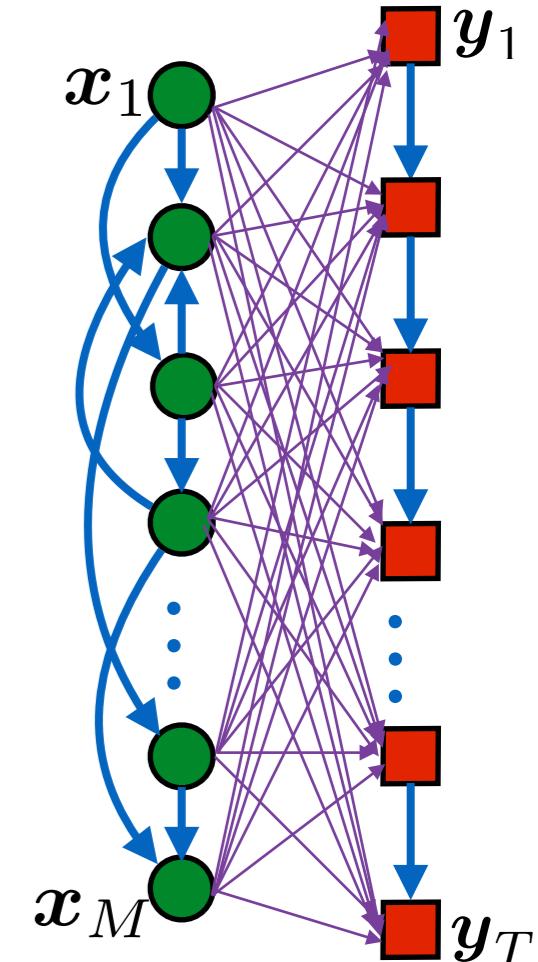
- Dynamic potential:** n-th order Markov model

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

Sequential facility location

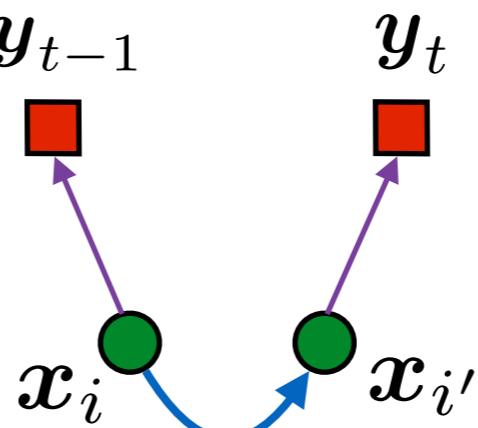
- **SeqFL:** Solve optimization on assignment variables $\{z_{i,t}\}$

$$\begin{aligned}
 \max_{\{z_{i,t}\}} \quad & \sum_{t=1}^T \sum_{i=1}^M -z_{i,t} d_{i,t} - \lambda \sum_{i=1}^M \| [z_{i,1} \cdots z_{i,T}] \|_\infty \\
 & + \beta \left(\sum_{i=1}^M z_{i,1} \log p_1(\mathbf{x}_i) + \sum_{t=2}^T \sum_{i,i'=1}^M z_{i,t-1} z_{i',t} \log p(\mathbf{x}_{i'} | \mathbf{x}_i) \right) \\
 \text{s. t.} \quad & z_{i,t} \in \{0, 1\}, \quad \sum_{i=1}^M z_{i,t} = 1, \quad \forall i, t.
 \end{aligned}$$



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

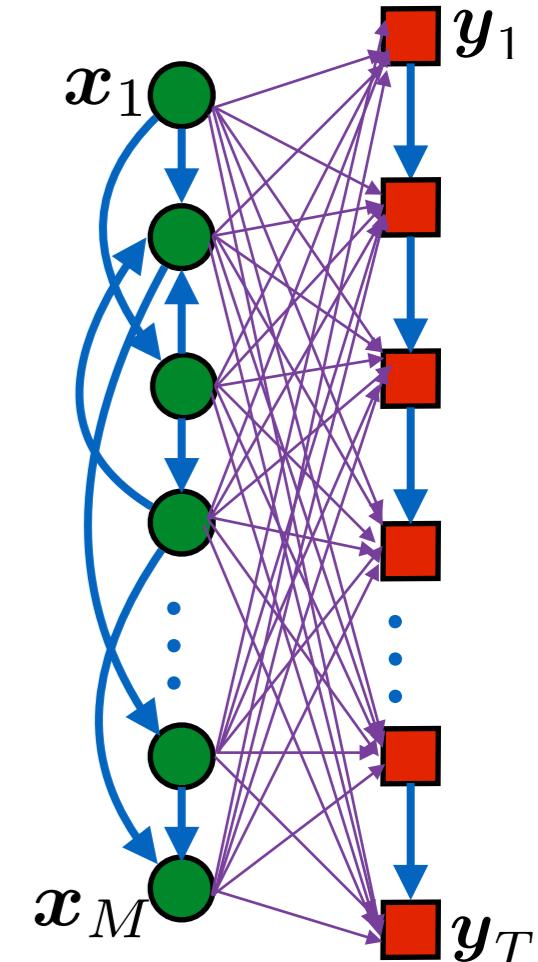
$p(\mathbf{x}_{i'} | \mathbf{x}_i)$ should be large!



Sequential facility location

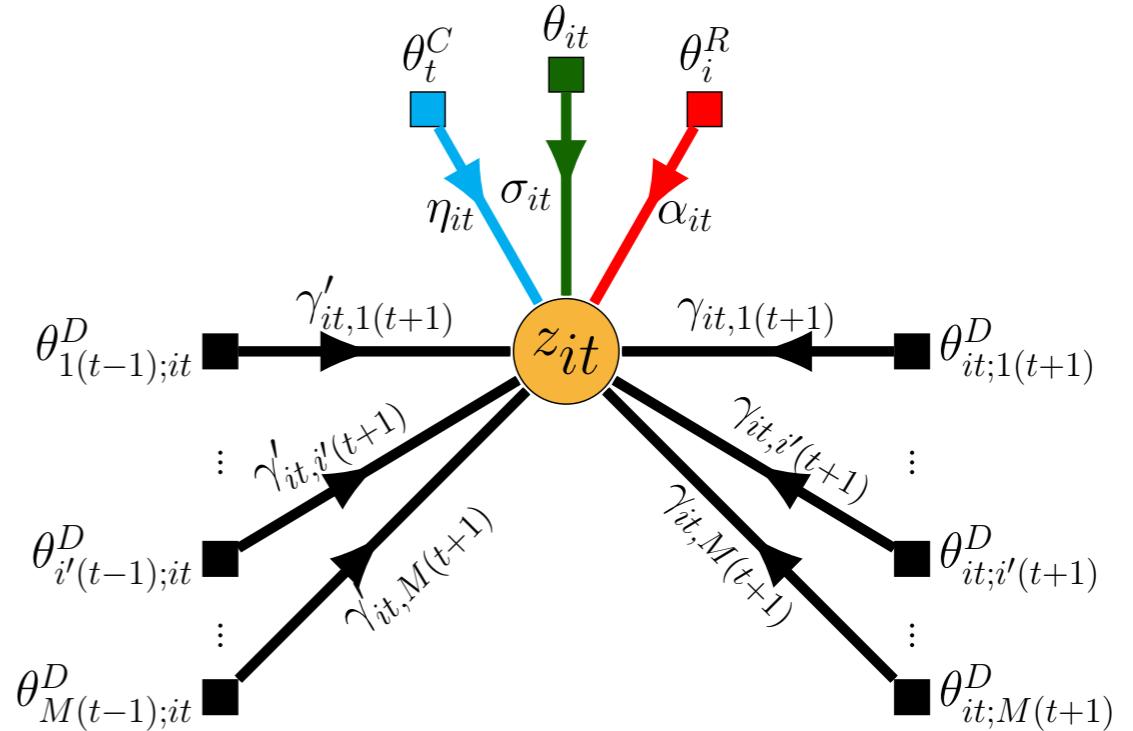
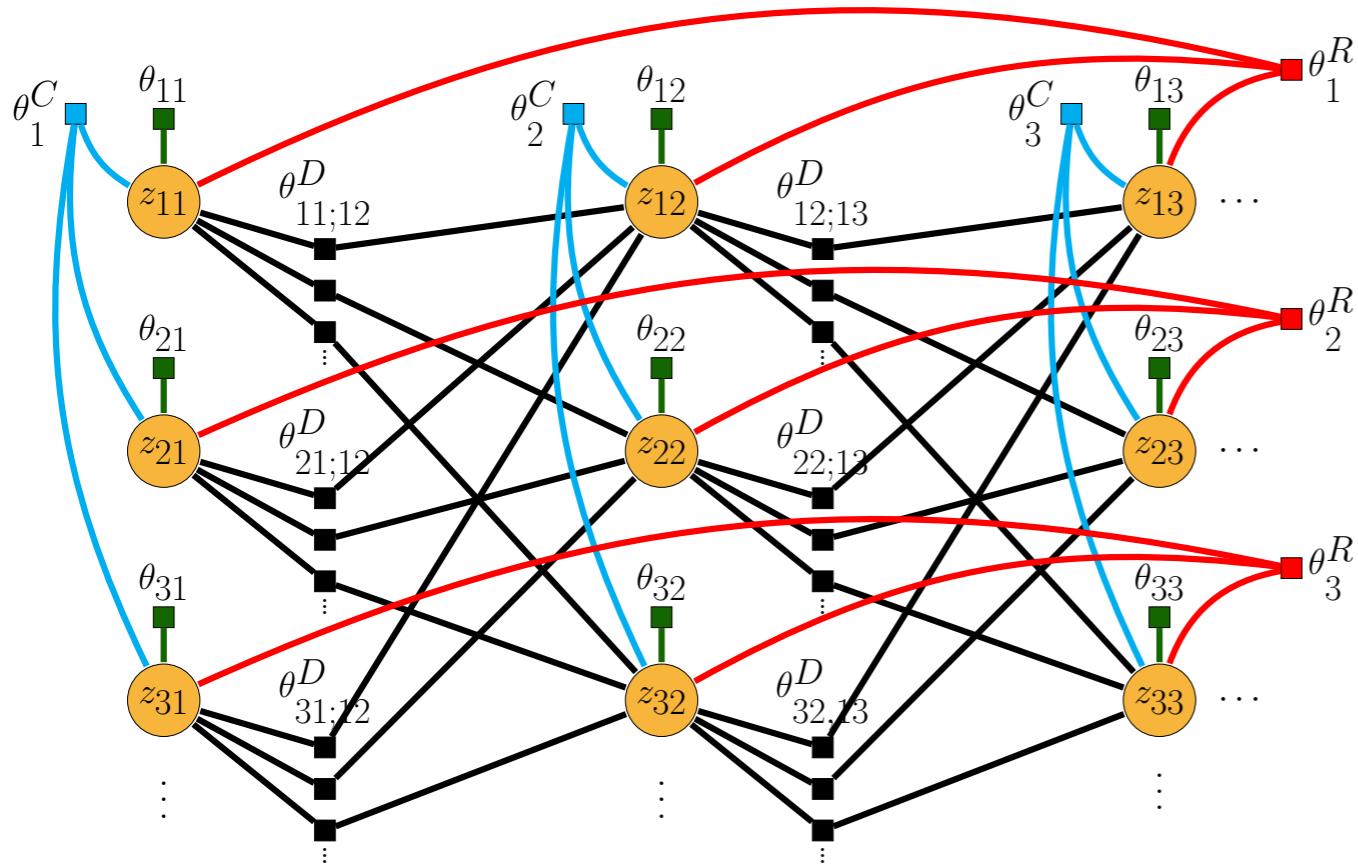
- **SeqFL:** Solve optimization on assignment variables $\{z_{i,t}\}$

$$\begin{aligned} \max_{\{z_{i,t}\}} \quad & \sum_{t=1}^T \sum_{i=1}^M -z_{i,t} d_{i,t} - \lambda \sum_{i=1}^M \| [z_{i,1} \cdots z_{i,T}] \|_\infty \\ & + \beta \left(\sum_{i=1}^M z_{i,1} \log p_1(\mathbf{x}_i) + \sum_{t=2}^T \sum_{i,i'=1}^M z_{i,t-1} z_{i',t} \log p(\mathbf{x}_{i'} | \mathbf{x}_i) \right) \\ \text{s. t.} \quad & z_{i,t} \in \{0, 1\}, \quad \sum_{i=1}^M z_{i,t} = 1, \quad \forall i, t. \end{aligned}$$



- **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:

$$\theta_i^R(z_{i,:}) = \begin{cases} -\lambda, & \|z_{i,:}\|_\infty > 0 \\ 0, & \text{otherwise} \end{cases}$$

Dynamic factor:

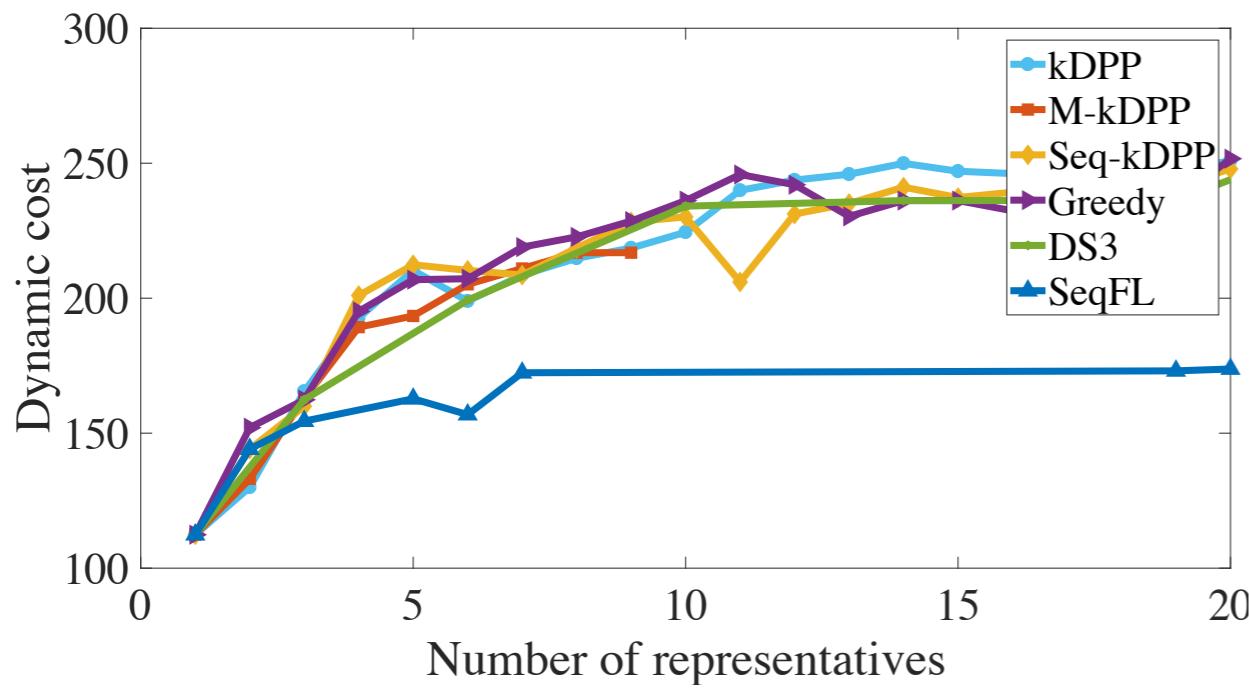
$$\theta_{it-1;i',t}^D(z_{i,t-1}, z_{i',t}) = \log p(\mathbf{x}_{i'} | \mathbf{x}_i) z_{i,t-1} z_{i',t}$$

Constraint factor:

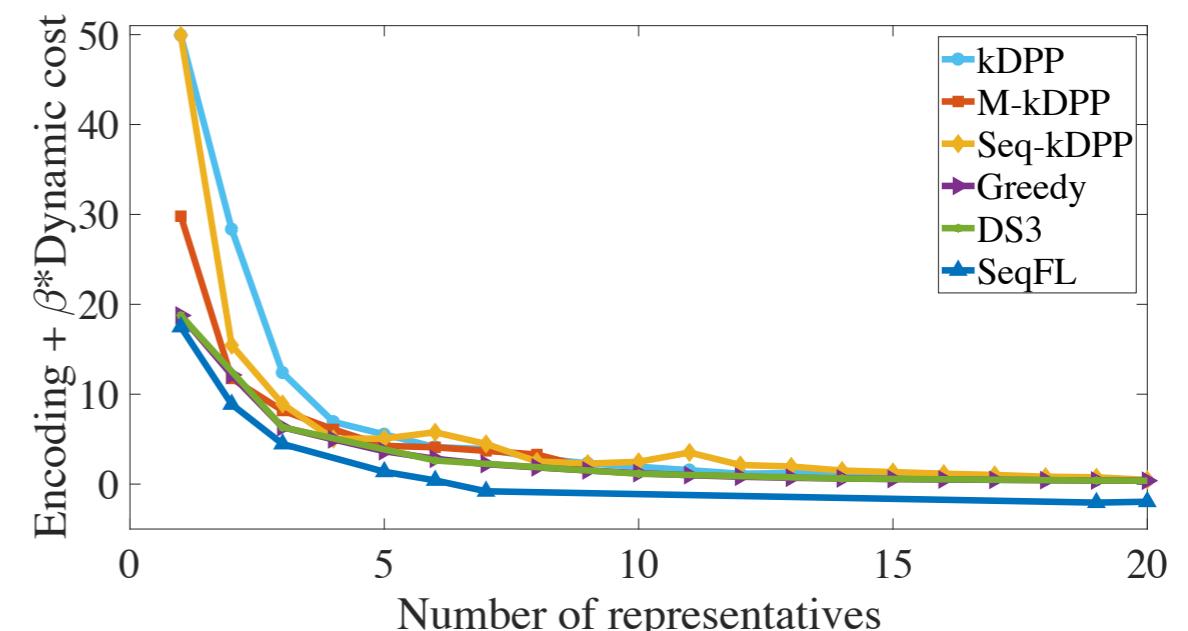
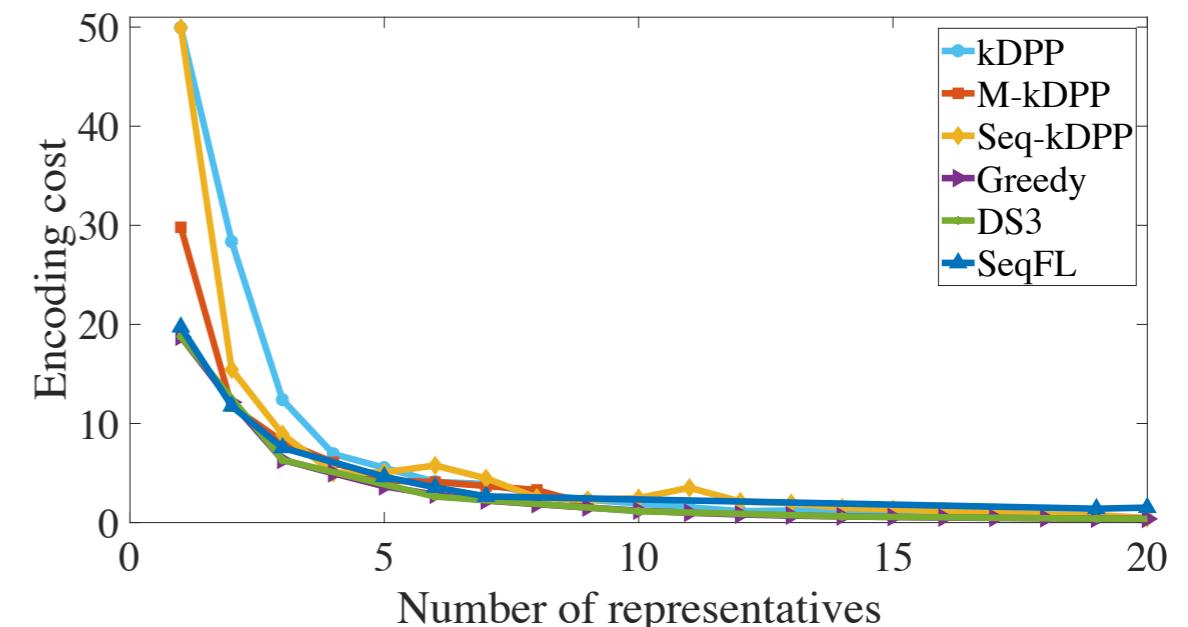
$$\theta_t^C(z_{:,t}) = \begin{cases} 0, & \sum_{i=1}^M z_{i,t} = 1 \\ -\infty, & \text{otherwise.} \end{cases}$$

Synthetic experiments

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

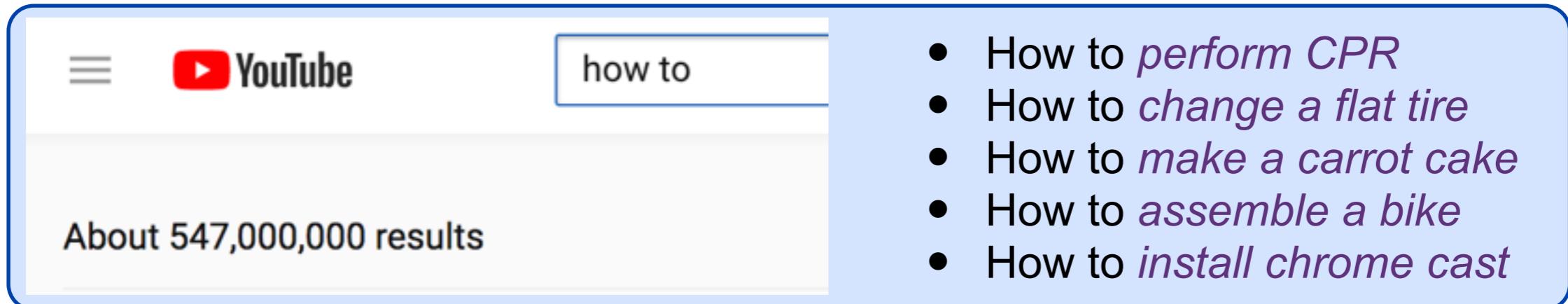


Sequence of reps in SeqFL:
compatible with dynamic model



Procedure learning

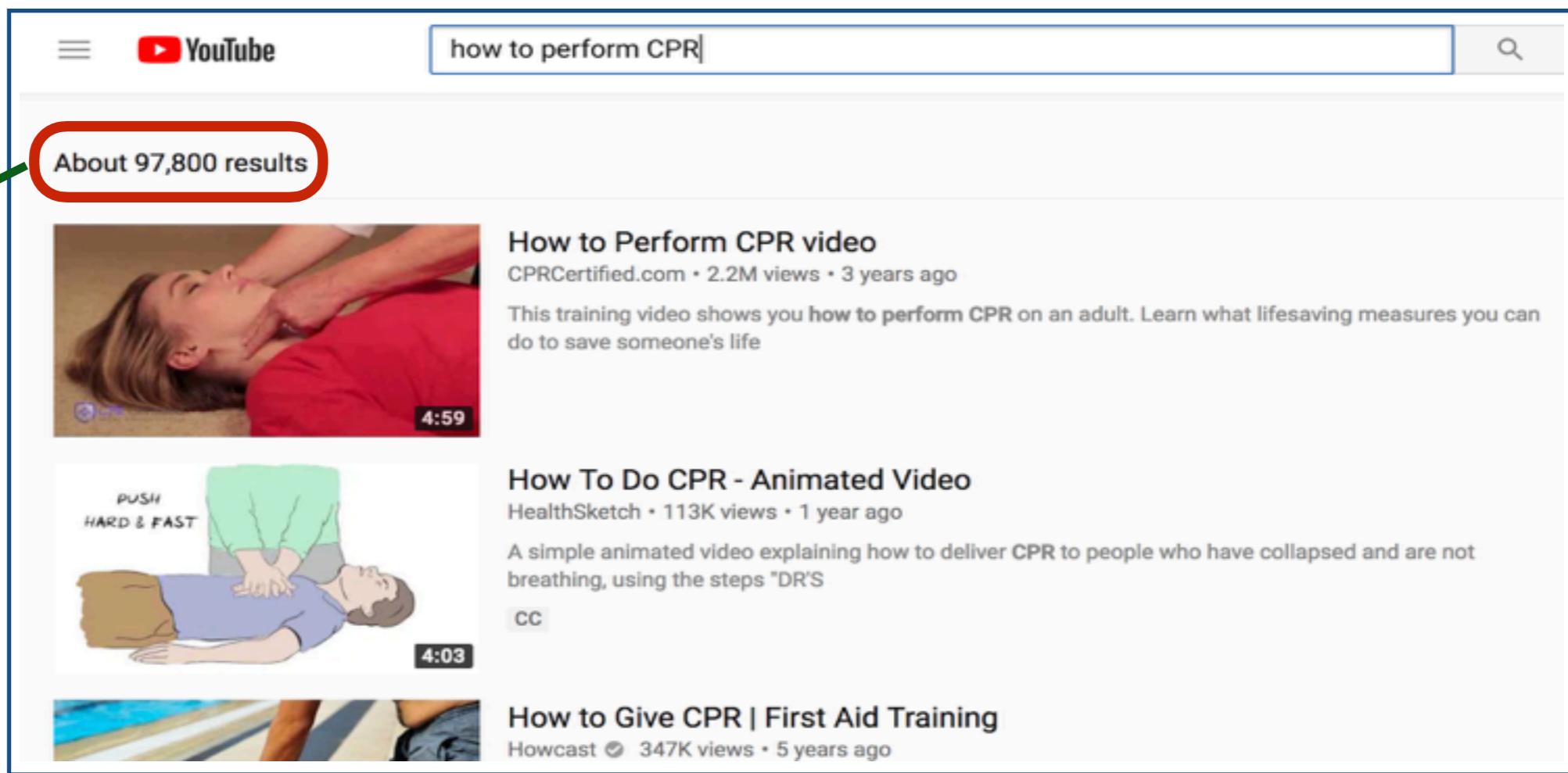
- We learn **how to do** tasks using **instructional videos**



A screenshot of a YouTube search results page. The search bar contains the text "how to". Below the search bar, it says "About 547,000,000 results". To the right of the search bar, there is a list of five items:

- How to *perform CPR*
- How to *change a flat tire*
- How to *make a carrot cake*
- How to *assemble a bike*
- How to *install chrome cast*

Large
videos

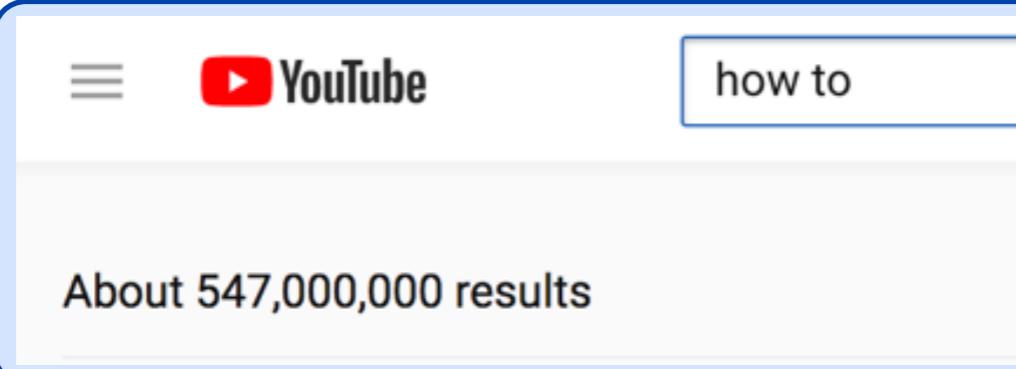


A screenshot of a YouTube search results page for "how to perform CPR". The search bar contains the text "how to perform CPR". Above the search bar, it says "About 97,800 results". A red box highlights this result count. To the right of the search bar, there are three video thumbnails:

- How to Perform CPR video**
CPRCertified.com • 2.2M views • 3 years ago
This training video shows you how to perform CPR on an adult. Learn what lifesaving measures you can do to save someone's life
- How To Do CPR - Animated Video**
HealthSketch • 113K views • 1 year ago
A simple animated video explaining how to deliver CPR to people who have collapsed and are not breathing, using the steps "DR'S"
- How to Give CPR | First Aid Training**
Howcast • 347K views • 5 years ago

Procedure learning

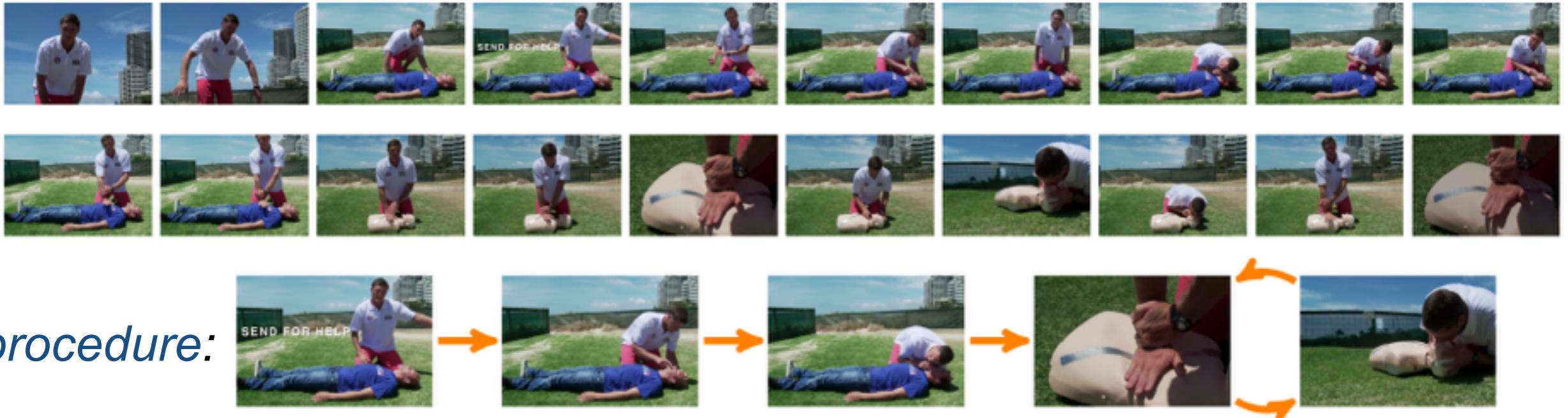
- We learn **how to do** tasks using **instructional videos**



A screenshot of a YouTube search results page. The search bar contains the text "how to". Below the search bar, it says "About 547,000,000 results". To the right of the search bar, there is a list of five bullet points representing search results:

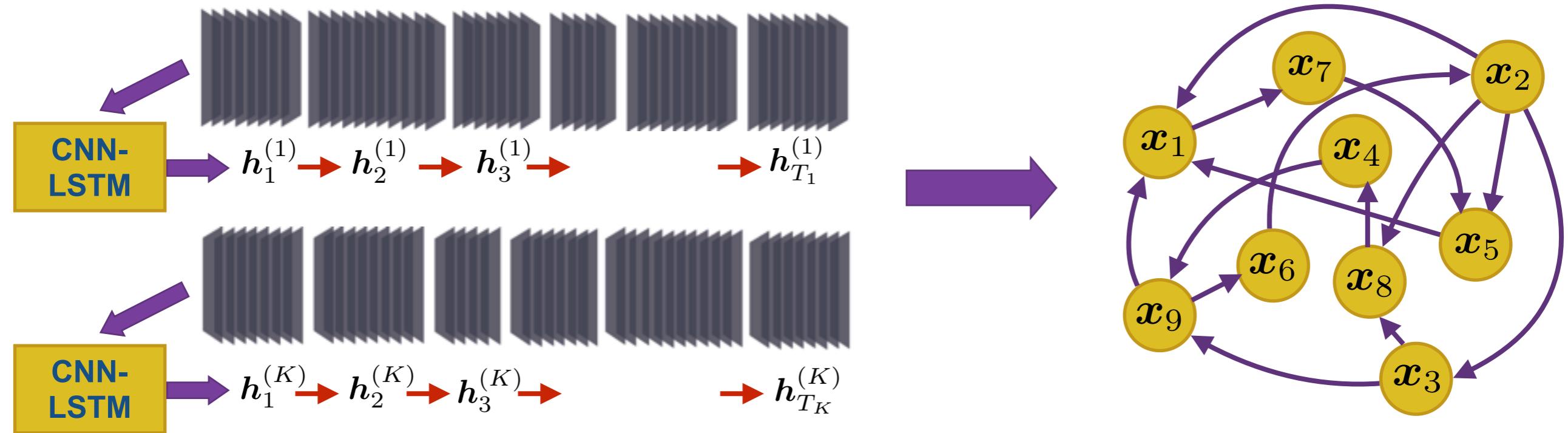
- How to perform CPR
- How to change a flat tire
- How to make a carrot cake
- How to assemble a bike
- How to install chrome cast

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



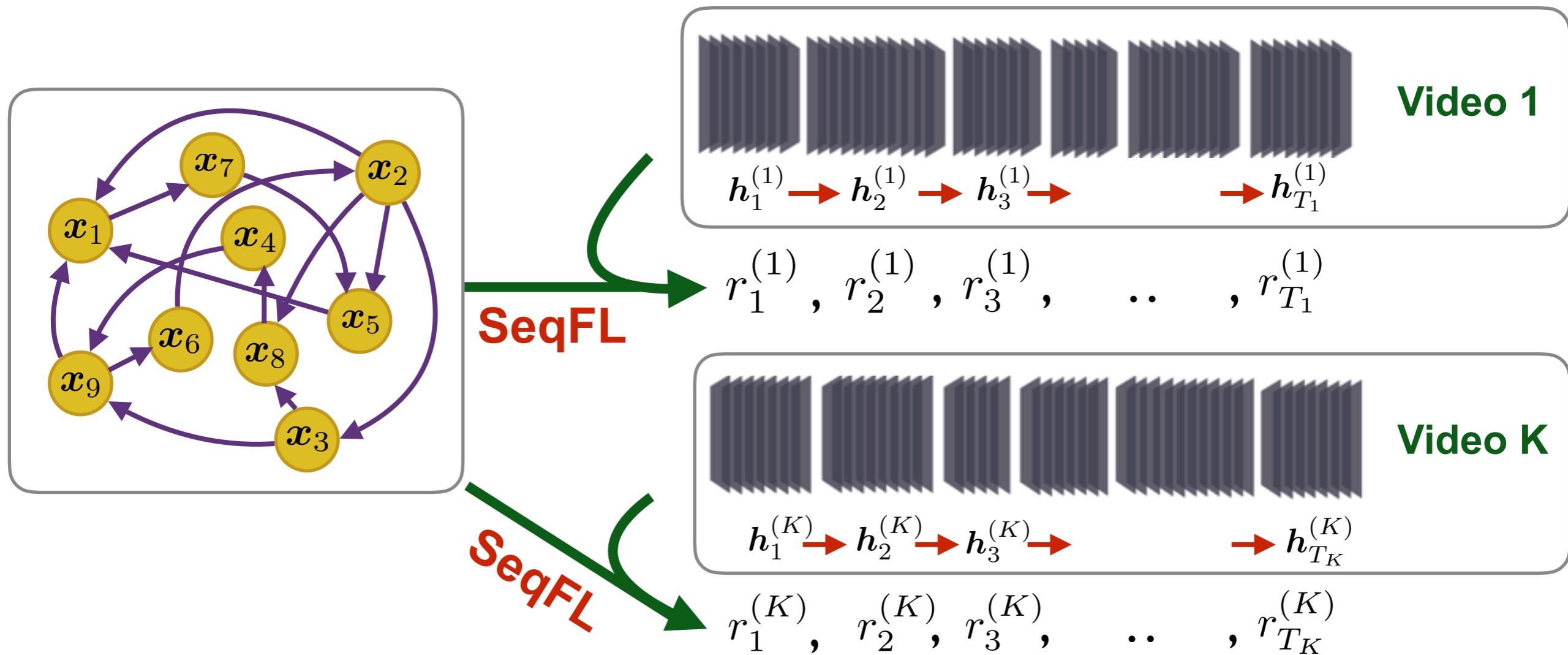
Procedure learning via SeqFL

- 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**



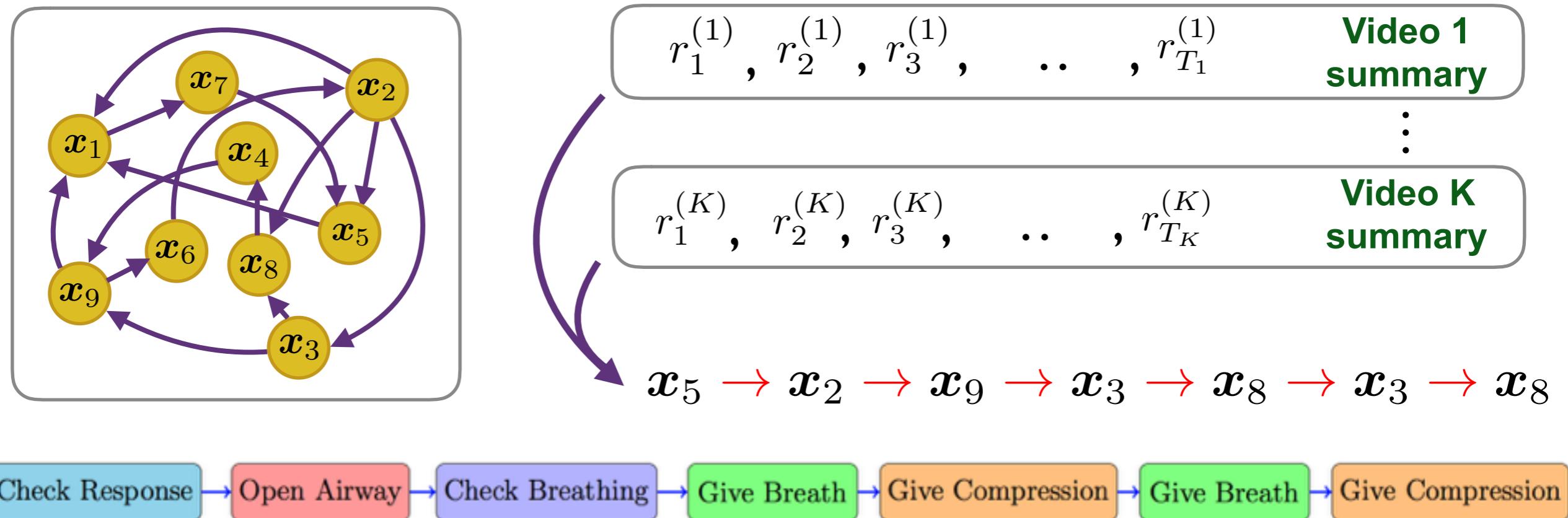
Procedure learning via SeqFL

- 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** ($X = \text{HMM}$, $Y = \text{test videos}$)



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** ($X = \text{HMM}$, $Y = \text{test videos}$)
- 5) **Align sequences** of representatives
- 6) **Map representatives to actions**



Experimental results

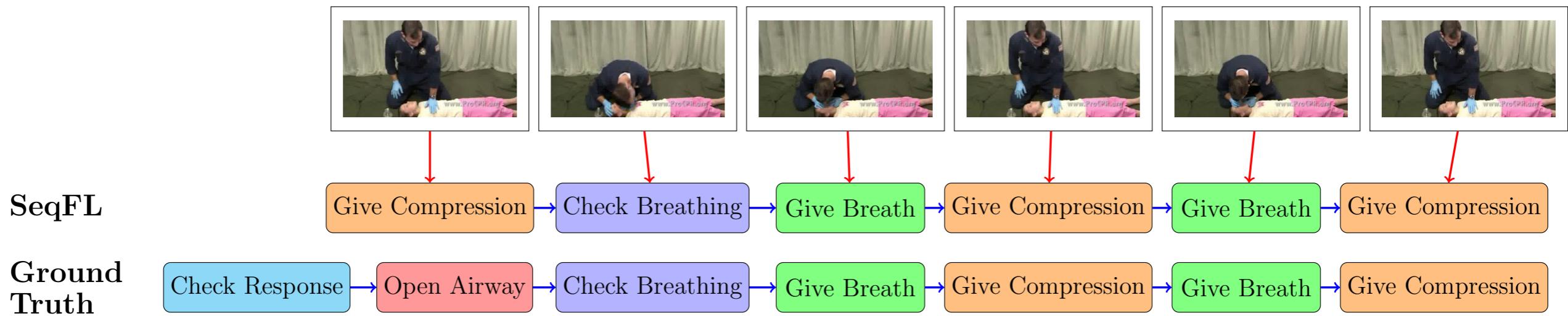
- **Instructional video dataset** (Alayrac et al CVPR '16)
 - **5 tasks, 30 videos/task**, labels of key steps given

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

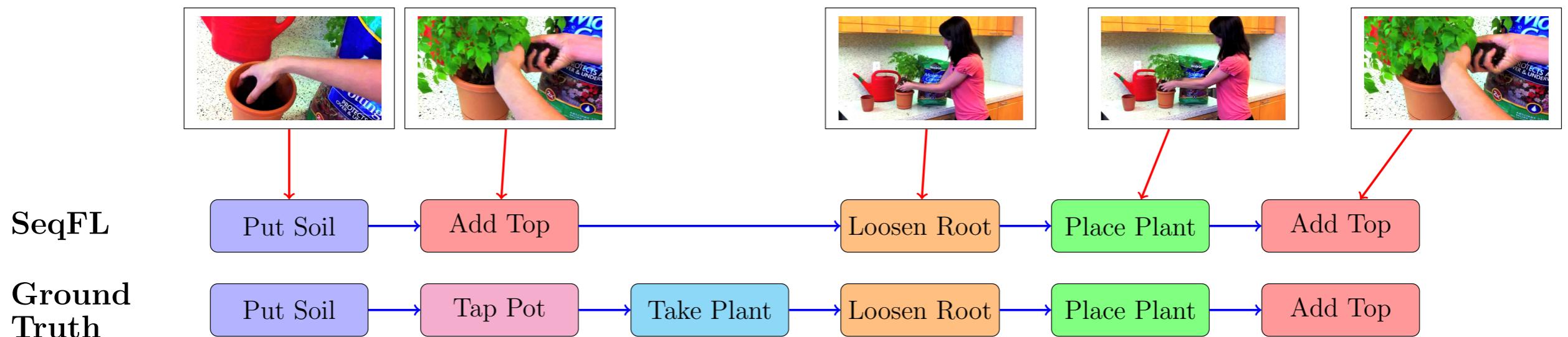
← **No dynamics** → **with dynamics**

Experimental results

- How to ‘**perform CPR**’

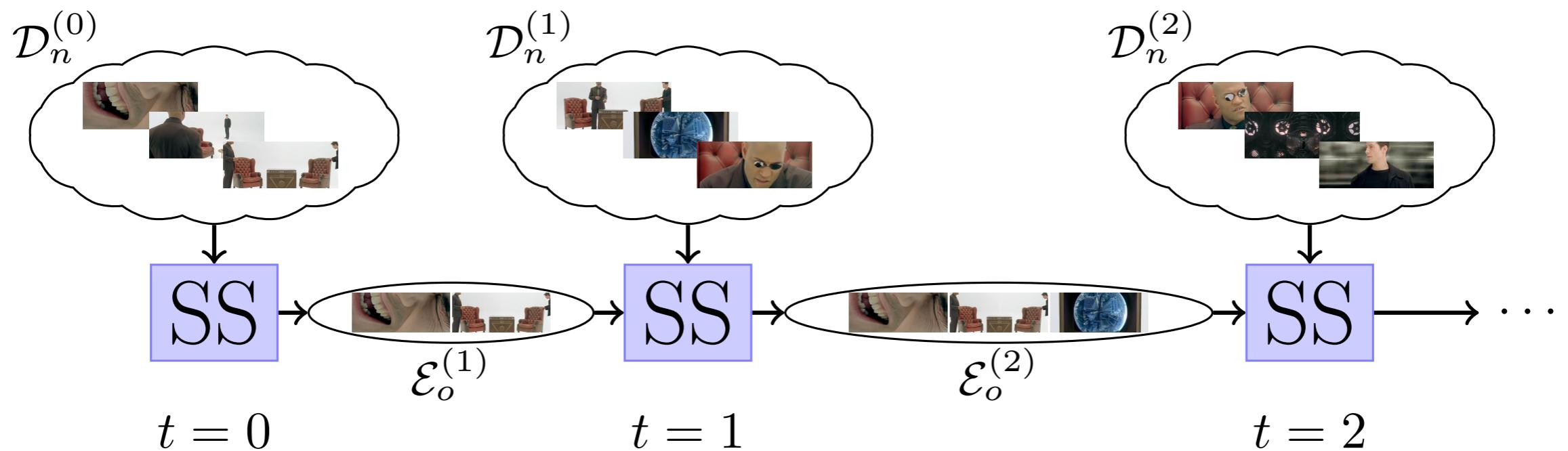


- How to ‘**repot a plant**’



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>