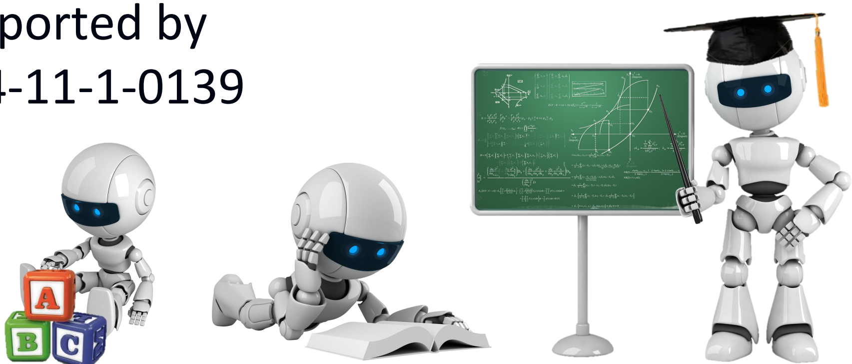
Active Task Selection for Lifelong Machine Learning

Paul Ruvolo
Olin College

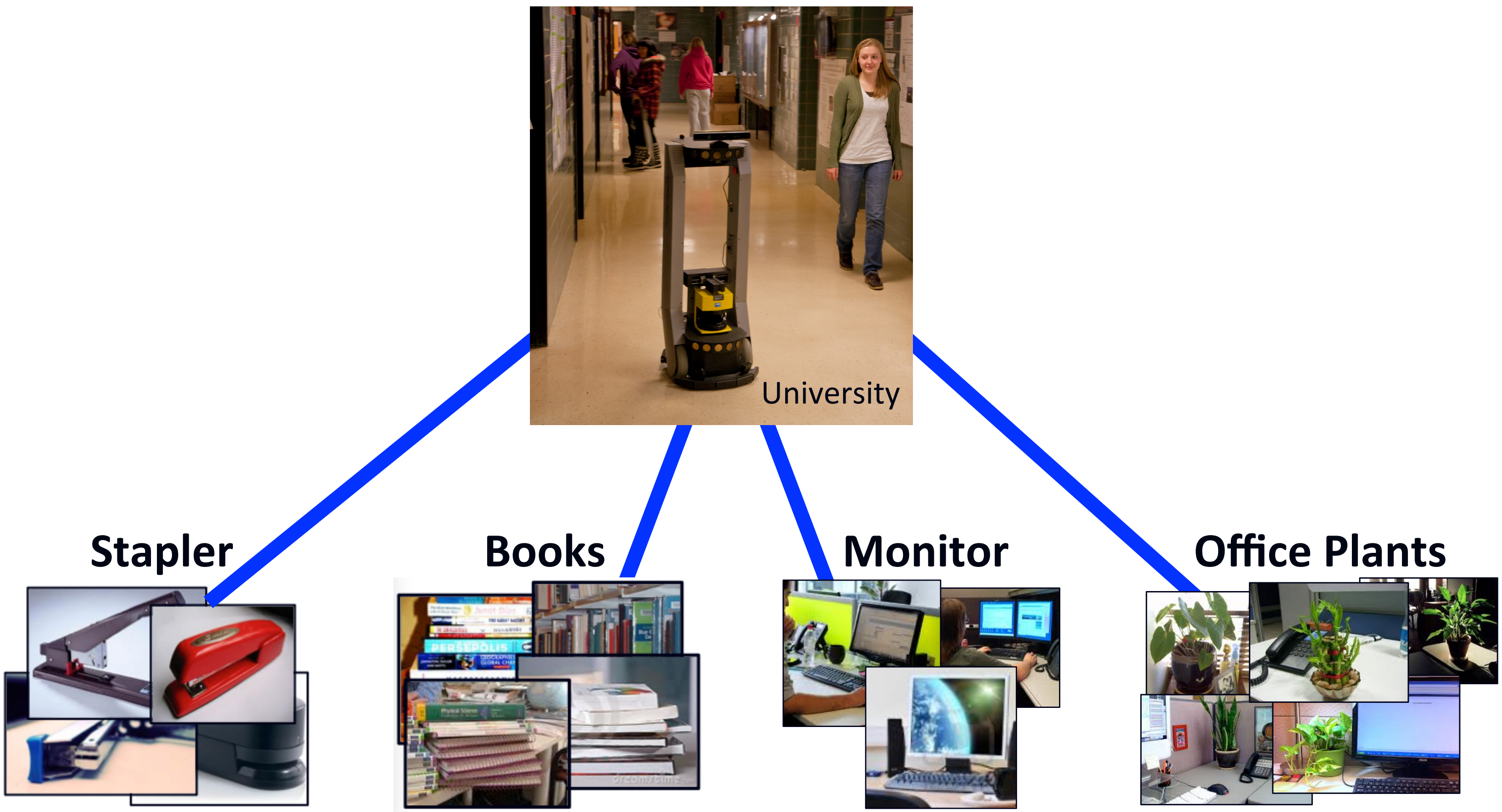
Eric Eaton
Univ. of Pennsylvania

This work was supported by
ONR Grant #N00014-11-1-0139

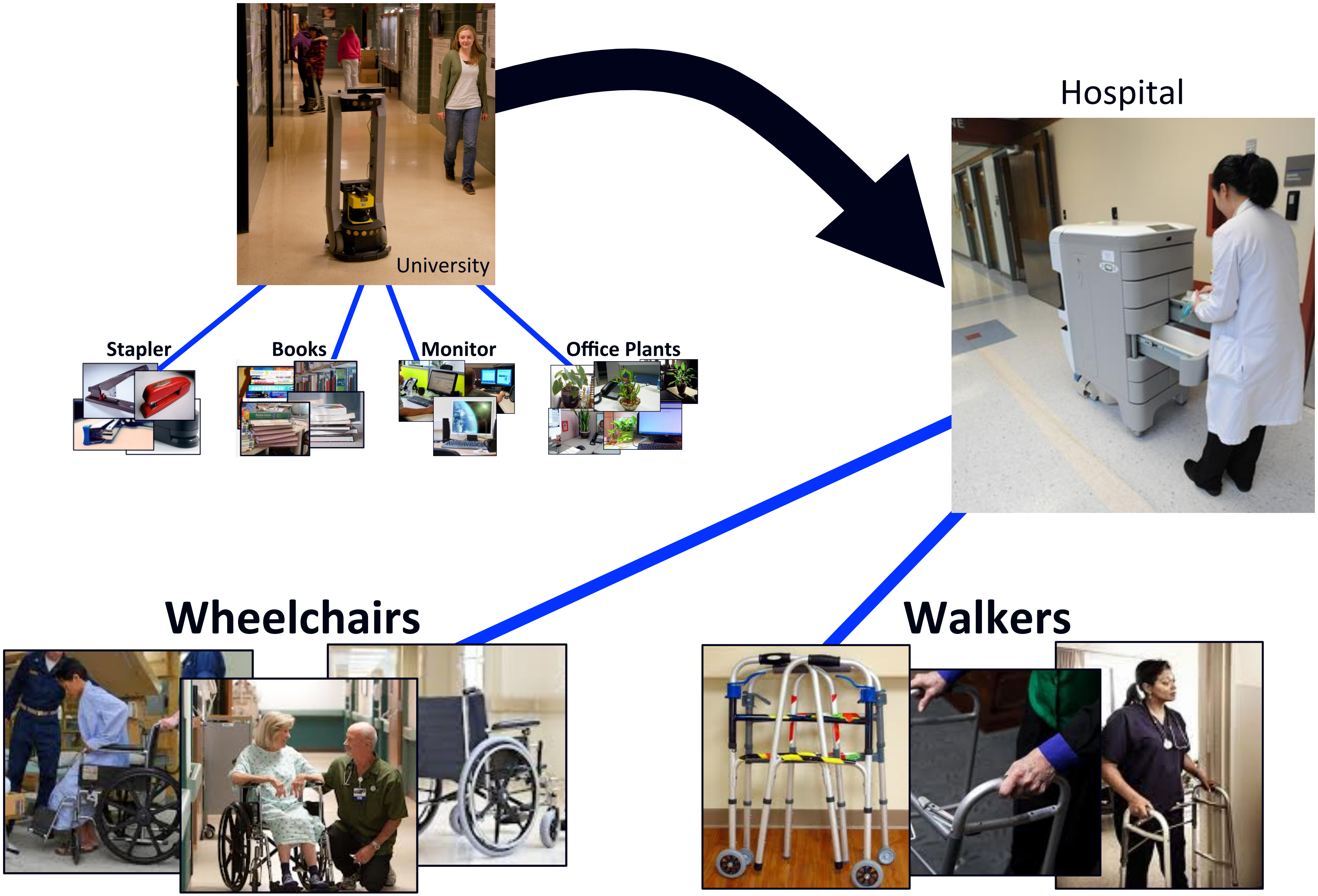


Motivation

Consider a robot tasked with learning to recognize many objects over an extended time period



Motivation



University
Stapler
Books
Monitor
Office Plants
Hospital
Wheelchairs
Walkers

Motivation

University

Hospital

Museum

. . .

Motivation

University

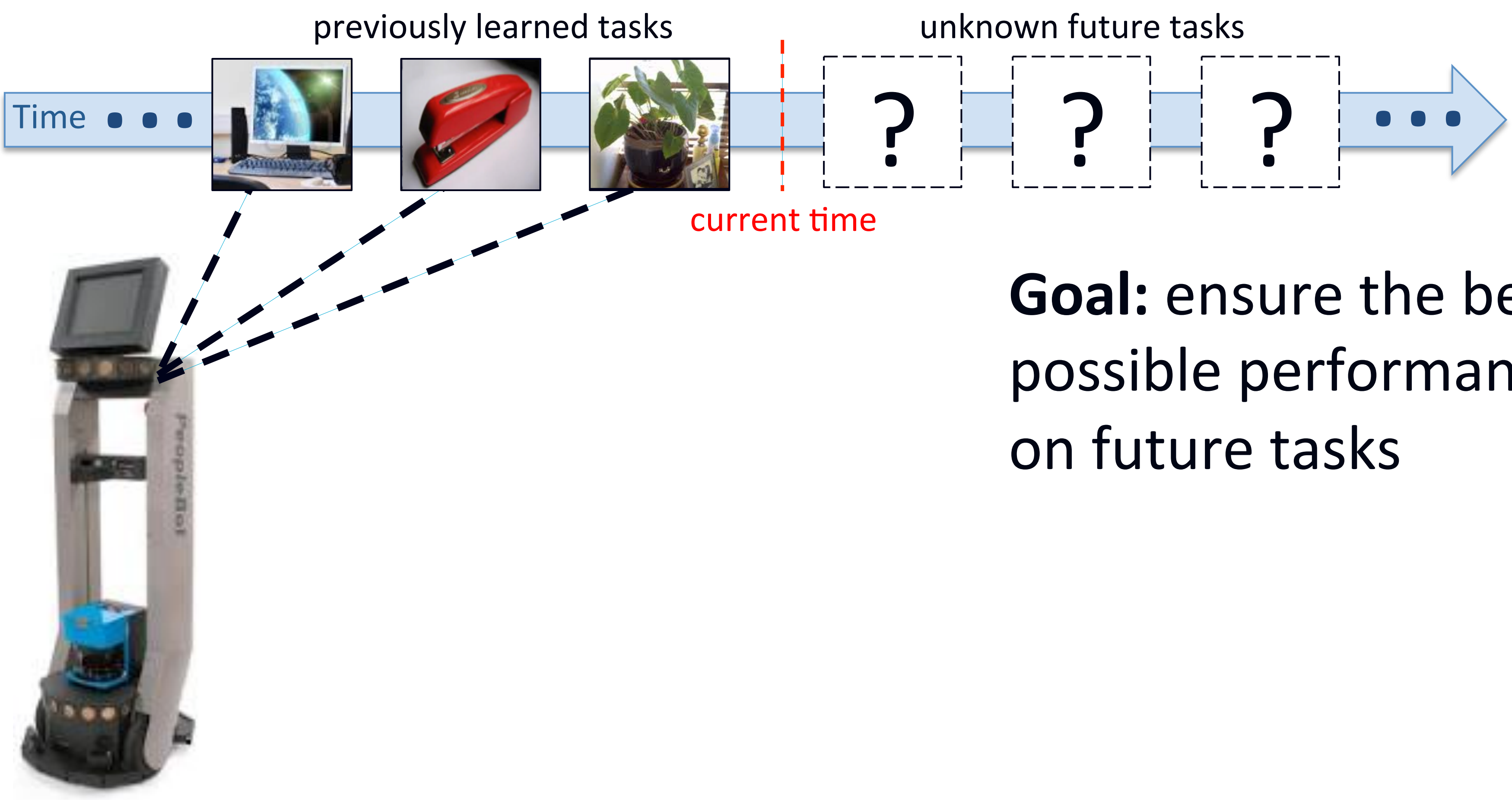
Hospital

Museum

…

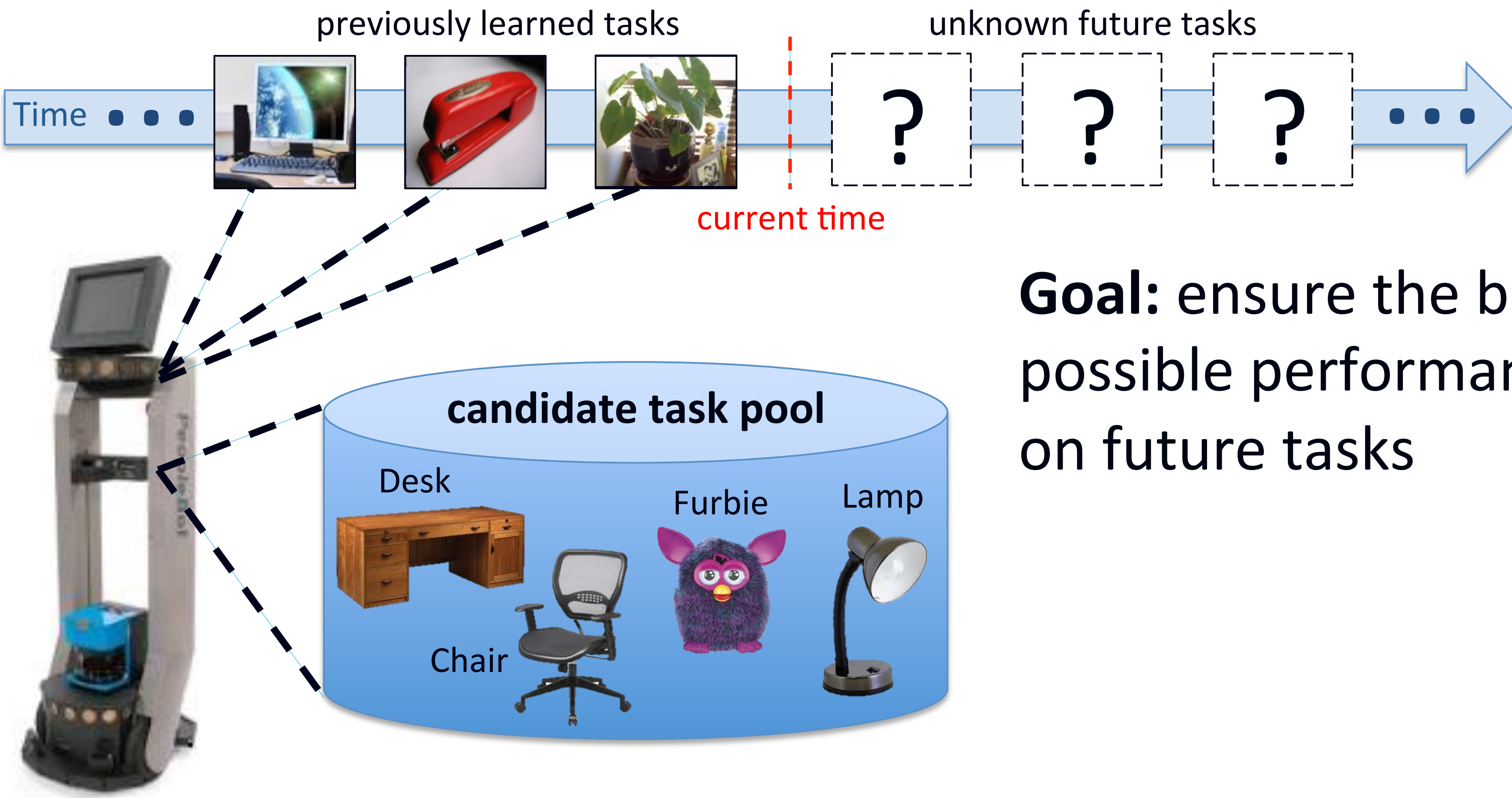
Lifelong learning enables the agent to build continually on its knowledge

Task Selection Problem



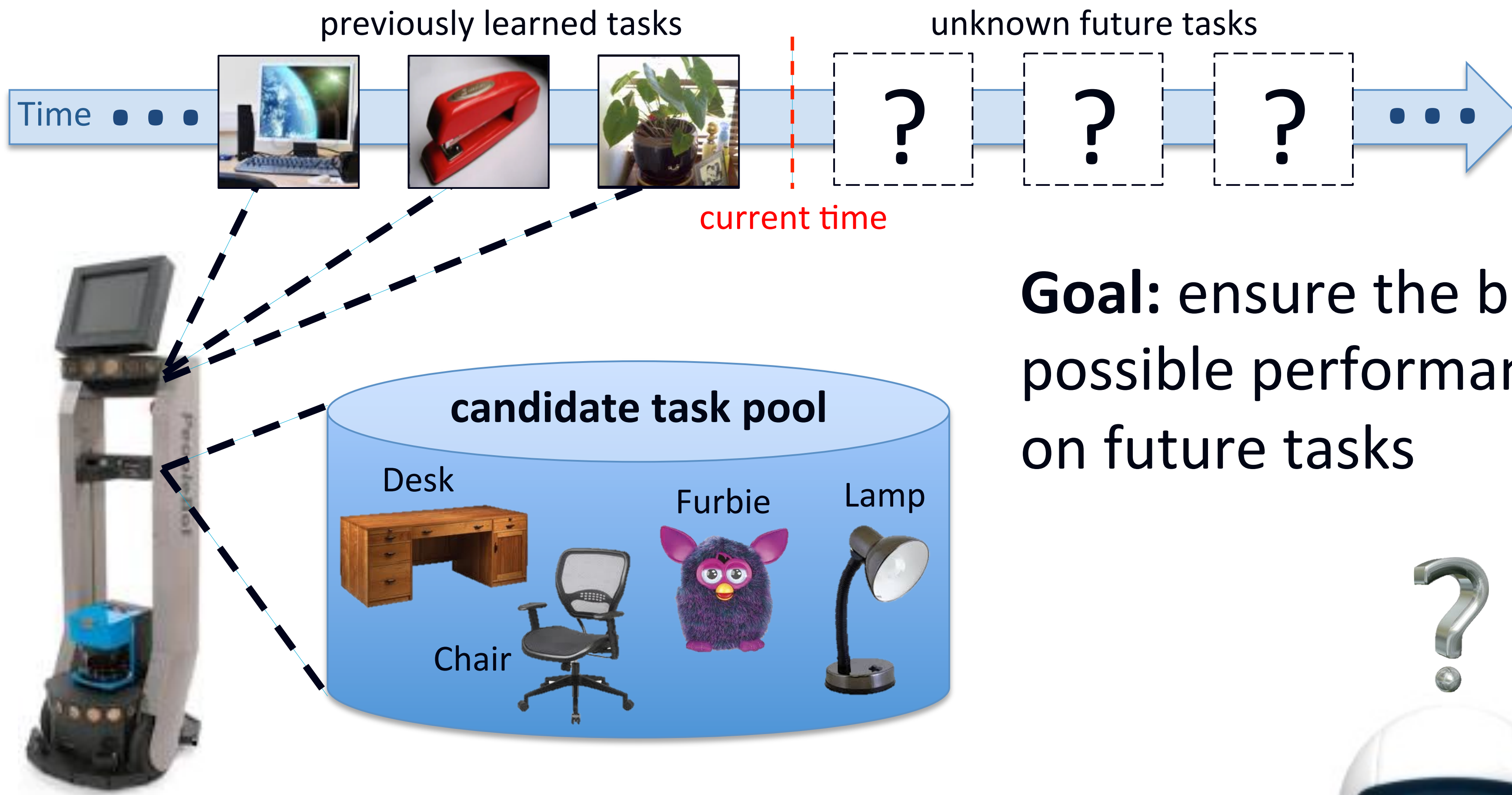
Goal: ensure the best possible performance on future tasks

Task Selection Problem



Goal: ensure the best possible performance on future tasks

Task Selection Problem



Goal: ensure the best possible performance on future tasks

Which task should it choose to learn next?

Introduction

- We present two methods for active task selection toward general knowledge acquisition:
  - Information Maximization approach
  - Diversity approach

Introduction

- We present two methods for active task selection toward general knowledge acquisition:
  - Information Maximization approach
  - Diversity approach
- We show how to focus InfoMax selection toward a specific future task for targeted knowledge acquisition

Introduction

- We present two methods for active task selection toward general knowledge acquisition:
  - Information Maximization approach
  - Diversity approach

- We show how to focus InfoMax selection toward a specific future task for targeted knowledge acquisition

Active task selection accelerates knowledge acquisition in a lifelong learning setting

Outline

- **Introduction**
- Efficient Lifelong Learning Algorithm
- Active Task Selection
- Targeted Active Task Selection

Outline

- Introduction
- **Efficient Lifelong Learning Algorithm**
- Active Task Selection
- Targeted Active Task Selection

Efficient Lifelong Learning Algorithm

[Ruvolo & Eaton, ICML'13]

- ELLA is a method for online multi-task learning in a lifelong learning setting

	Transfer Learning	Batch Multi-Task Learning
Optimizes performance over	Target task	All tasks
Learns tasks consecutively	Yes, efficiently	Very inefficiently
Computational cost	Low	High

Lifelong learning includes elements of both transfer and multi-task learning

Efficient Lifelong Learning Algorithm

[Ruvolo & Eaton, ICML'13]

- ELLA is a method for online multi-task learning in a lifelong learning setting

- **ELLA's Capabilities**:
  1. Learns tasks consecutively
  2. Transfers knowledge from previous tasks
  3. Complexity independent of the number of tasks
  4. Theoretical guarantees on performance and convergence

	Transfer Learning	**Batch Multi-Task Learning**
Optimizes performance over	Target task	All tasks
Learns tasks consecutively	Yes, efficiently	Very inefficiently
Computational cost	Low	High

Lifelong learning includes elements of both transfer and multi-task learning

Efficient Lifelong Learning Algorithm

[Ruvolo & Eaton, ICML'13]

- ELLA is a method for online multi-task learning in a lifelong learning setting

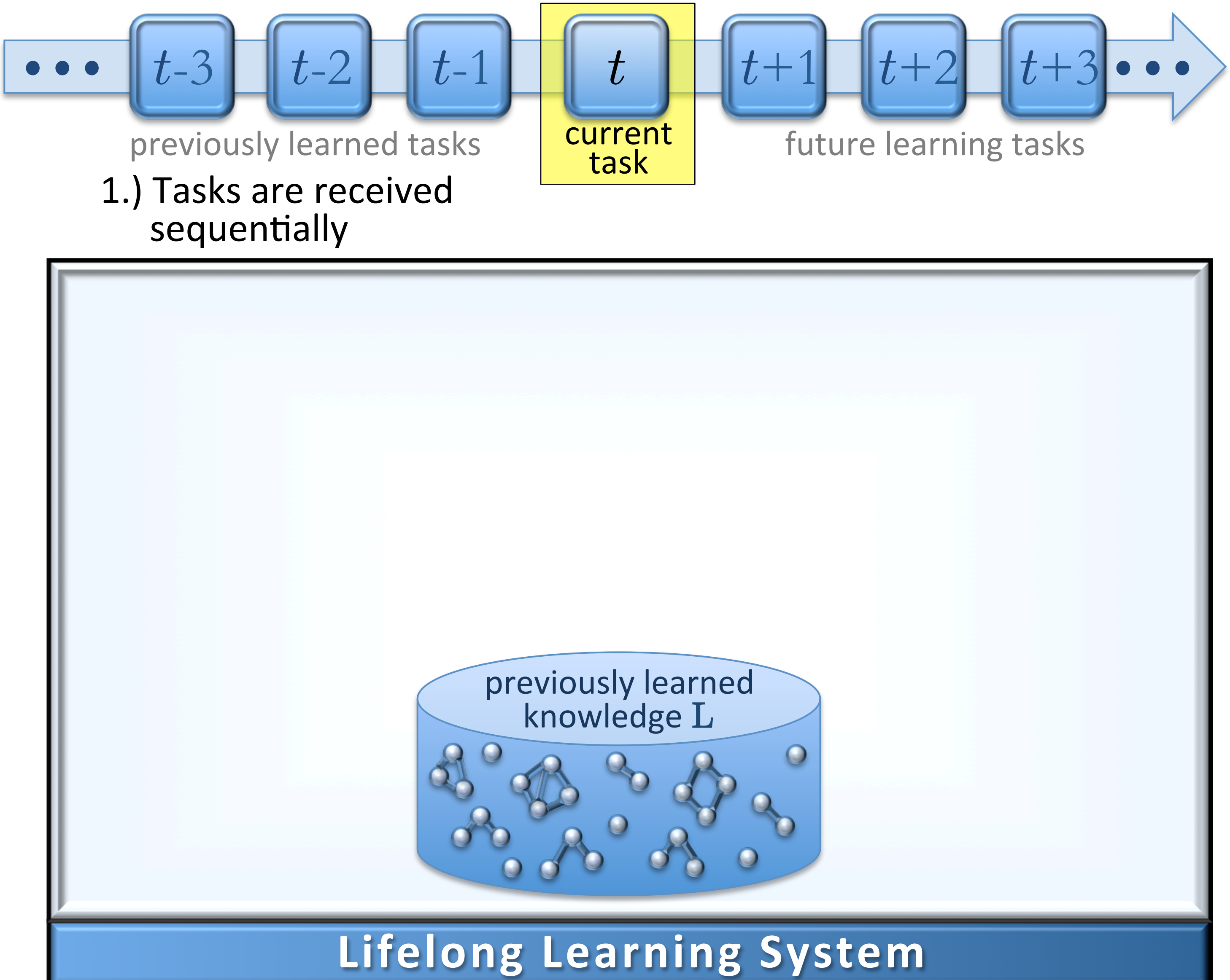
- **ELLA's Capabilities**:
  1. Learns tasks consecutively
  2. Transfers knowledge from previous tasks
  3. Complexity independent of the number of tasks
  4. Theoretical guarantees on performance and convergence

	Transfer Learning	**Batch Multi-Task Learning**
Optimizes performance over	Target task	All tasks
Learns tasks consecutively	Yes, efficiently	Very inefficiently
Computational cost	Low	High

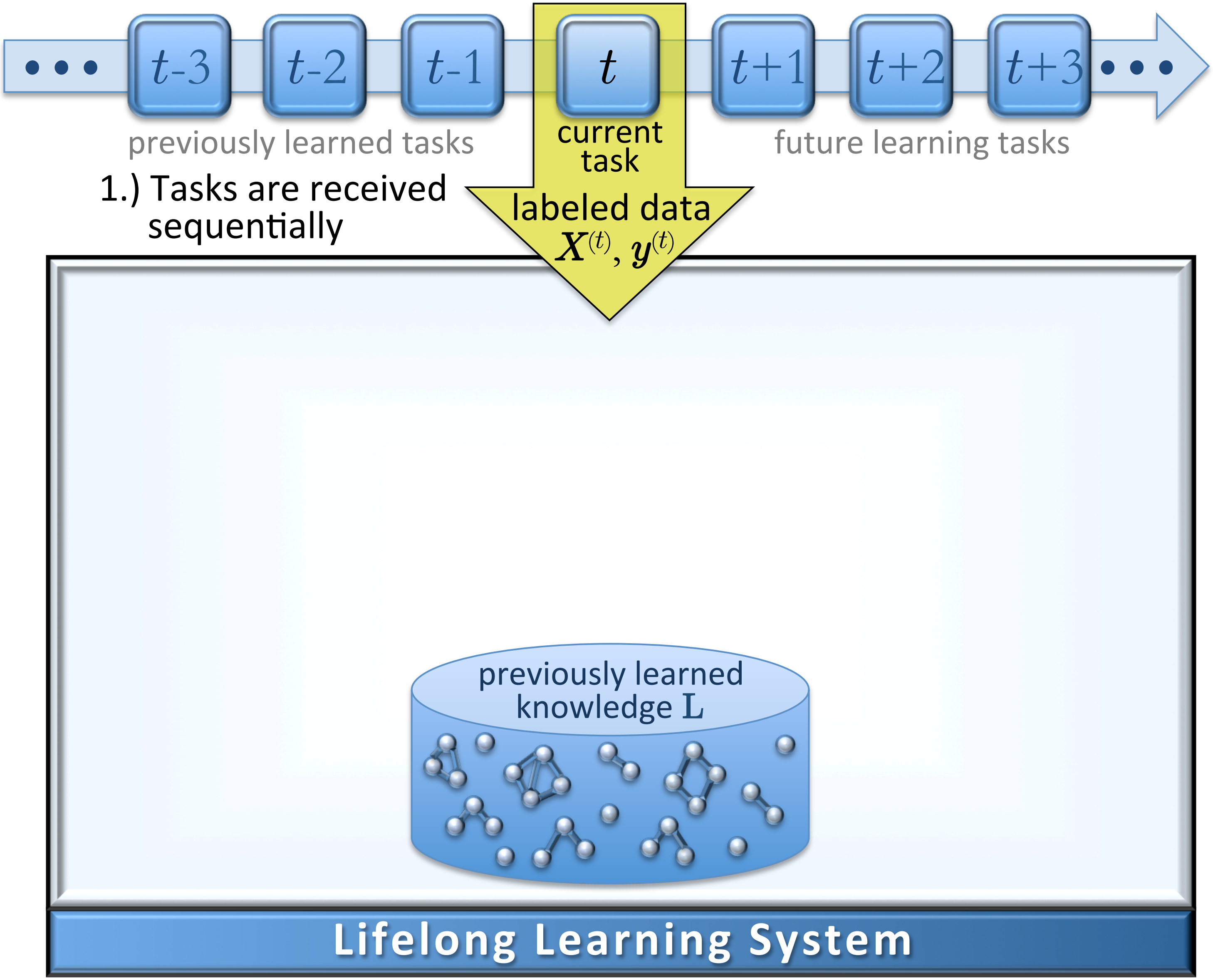
Lifelong learning includes elements of both transfer and multi-task learning

ELLA has equivalent accuracy to batch multi-task learning, but is over 1,000x faster and can learn online

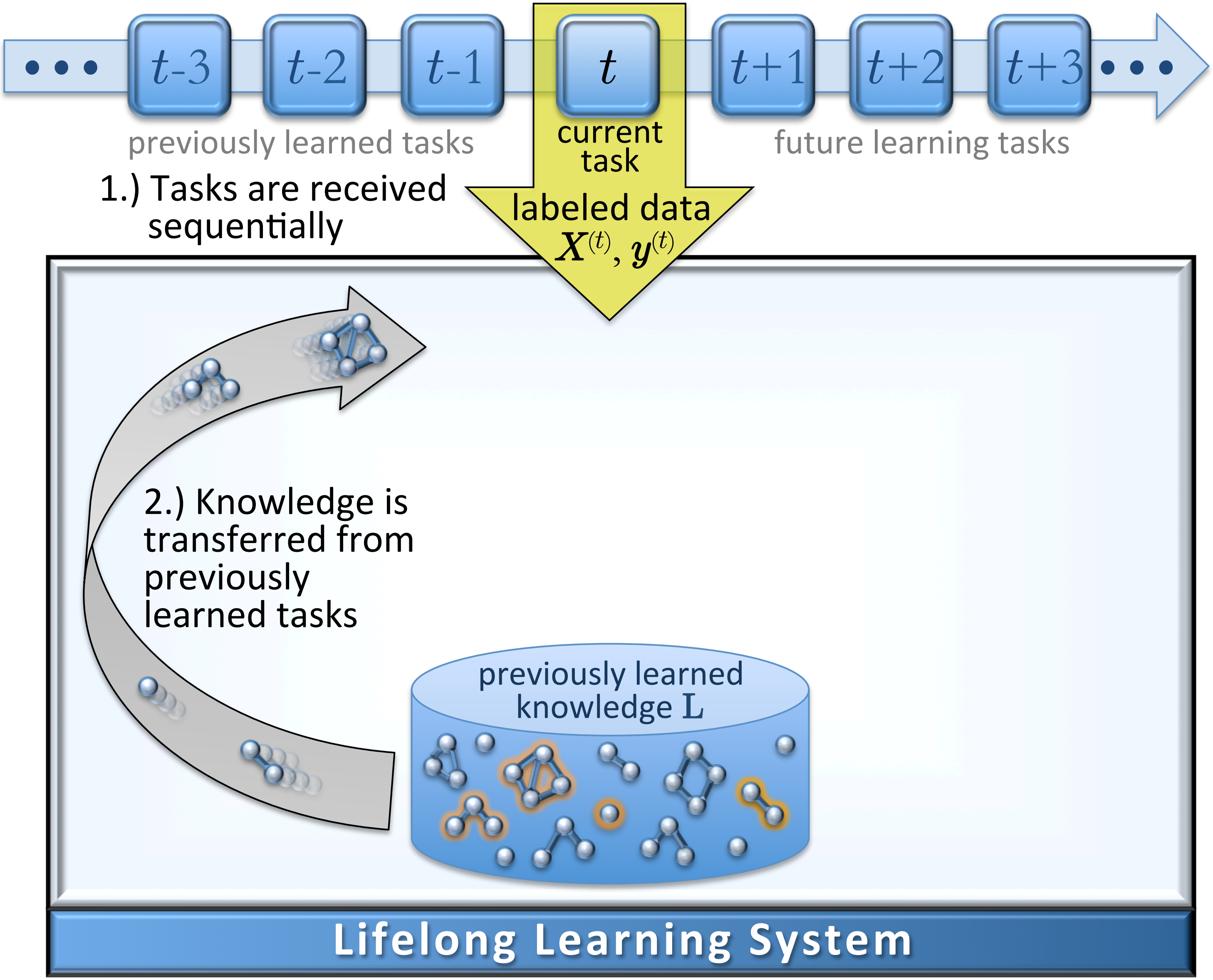
Lifelong Machine Learning



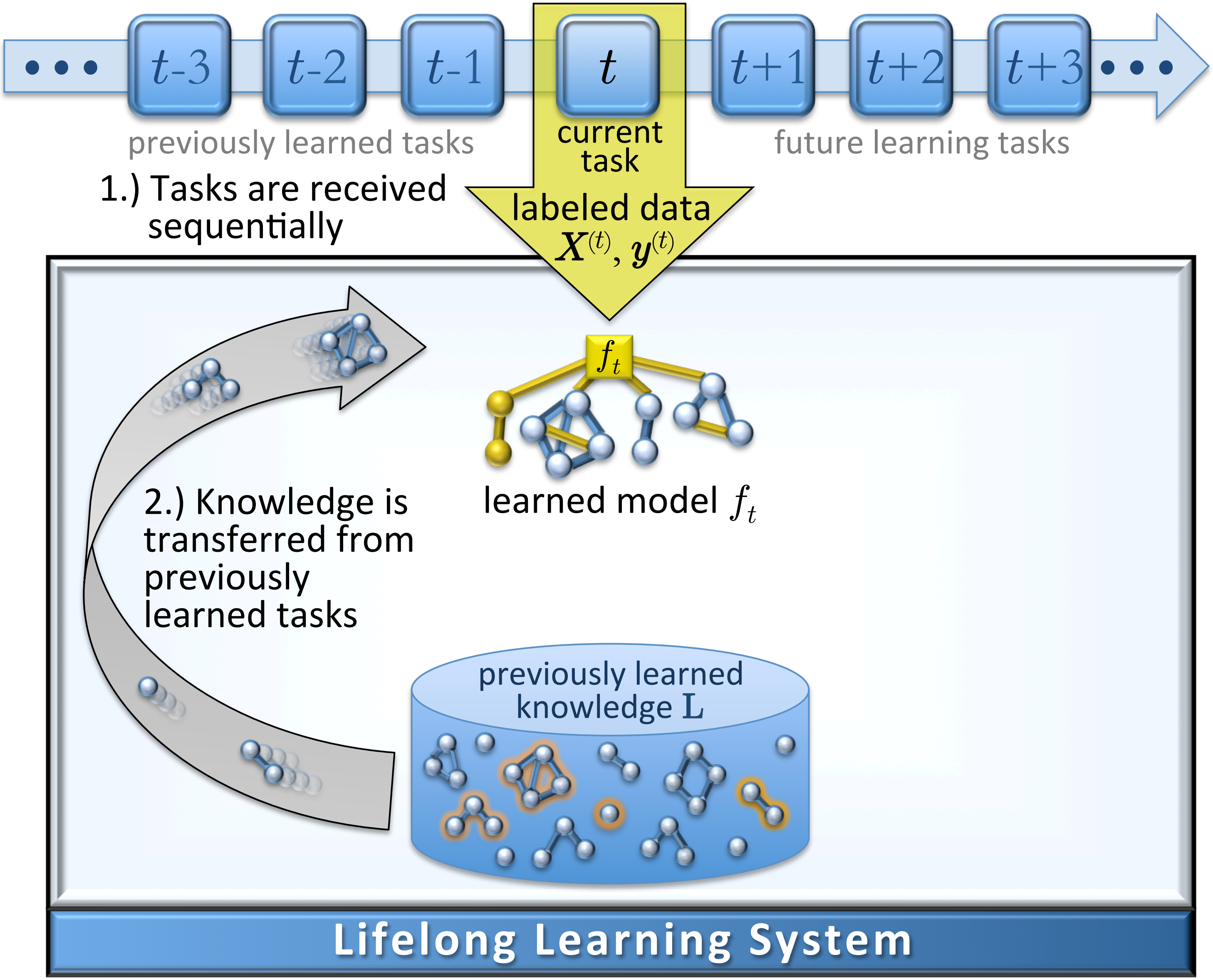
Lifelong Machine Learning



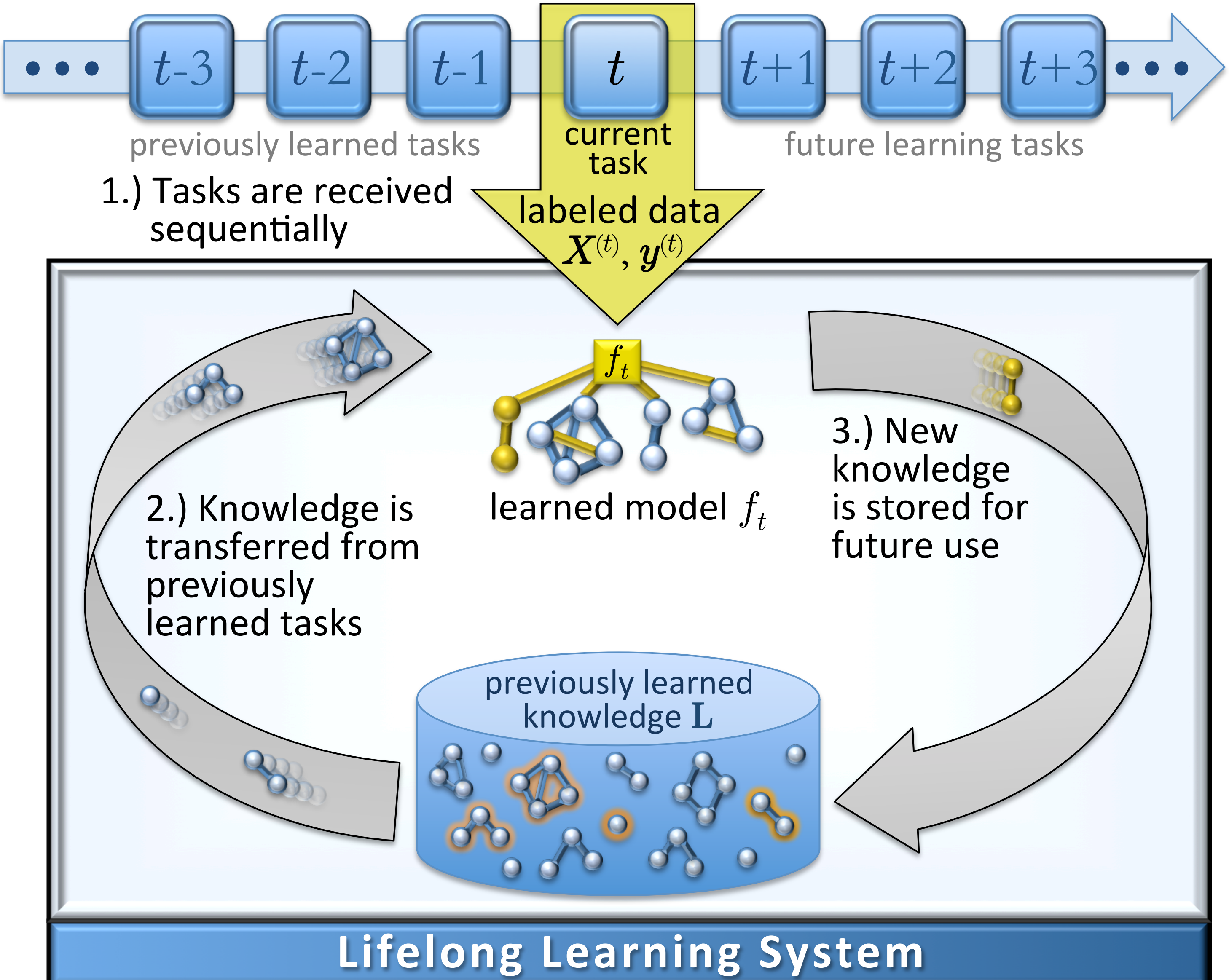
Lifelong Machine Learning



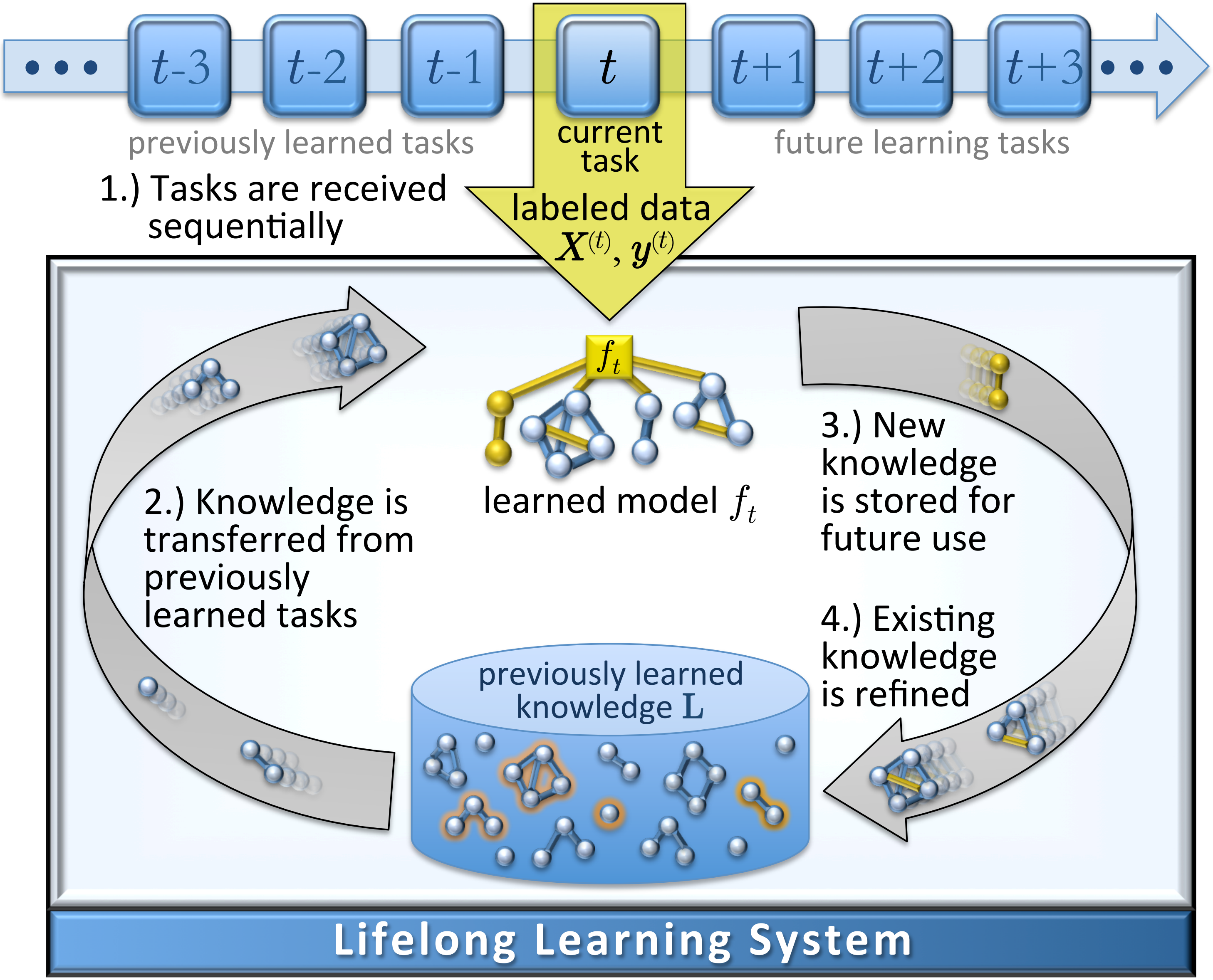
Lifelong Machine Learning



Lifelong Machine Learning



Lifelong Machine Learning



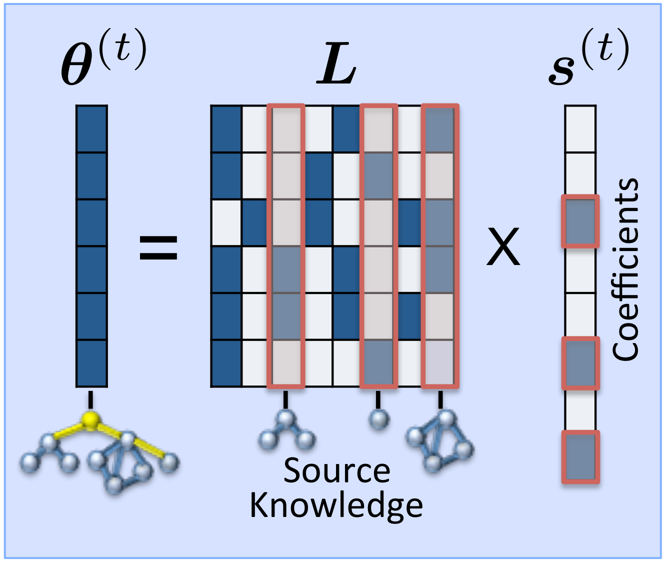
Task Structure Model

- ELLA fits a parametric model for each task t

$$f^{(t)}(\mathbf{x}) = f(\mathbf{x}; \boldsymbol{\theta}^{(t)}) \qquad \boldsymbol{\theta}^{(t)} \in \mathbb{R}^d$$

- The parameters $\boldsymbol{\theta}^{(t)}$ are linear combinations of a shared basis $\mathbf{L}$

$$\boldsymbol{\theta}^{(t)} = \mathbf{L}\boldsymbol{s}^{(t)} \qquad \mathbf{L} \in \mathbb{R}^{d \times k},\ \boldsymbol{s}^{(t)} \in \mathbb{R}^k$$



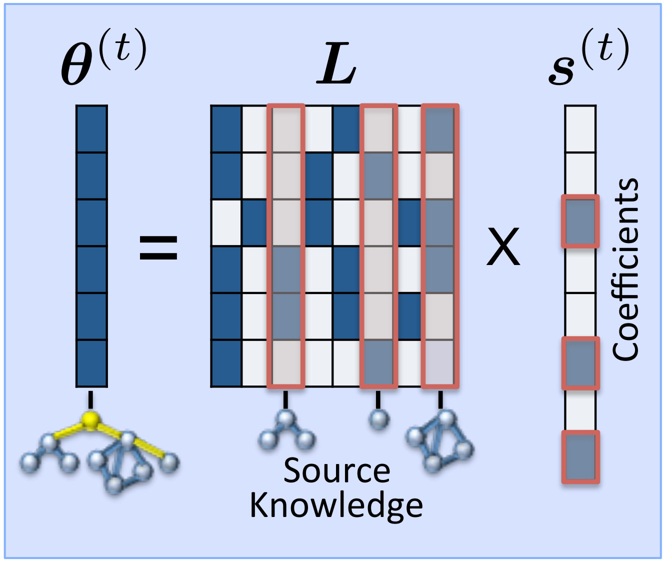
Task Structure Model

- ELLA fits a parametric model for each task t

$$f^{(t)}(\mathbf{x}) = f(\mathbf{x};\boldsymbol{\theta}^{(t)}) \qquad \boldsymbol{\theta}^{(t)} \in \mathbb{R}^d$$

- The parameters $\boldsymbol{\theta}^{(t)}$ are linear combinations of a shared basis $\mathbf{L}$

$$\boldsymbol{\theta}^{(t)} = \mathbf{L}\boldsymbol{s}^{(t)} \qquad \mathbf{L} \in \mathbb{R}^{d \times k},\ \boldsymbol{s}^{(t)} \in \mathbb{R}^k$$



Multi-Task Learning Objective Fn:

$$e_T(\mathbf{L}) = \frac{1}{T}\sum_{t=1}^{T}\min_{\boldsymbol{s}^{(t)}}\left\{\underbrace{\frac{1}{n_t}\sum_{i=1}^{n_t}\mathcal{L}\left(f\left(\boldsymbol{x}_i^{(t)};\mathbf{L}\boldsymbol{s}^{(t)}\right),y_i^{(t)}\right)}_{\text{model fit to data}} + \underbrace{\mu\|\boldsymbol{s}^{(t)}\|_1}_{\text{sparsity}}\right\} + \underbrace{\lambda\|\mathbf{L}\|_{\mathsf{F}}^2}_{\text{complexity}}$$

#tasks seen so far

Efficient Lifelong Learning Algorithm

[Ruvolo & Eaton, ICML'13]

MTL Objective Function:

$$e_T(\mathbf{L}) = \frac{1}{T}\sum_{t=1}^{T}\min_{\boldsymbol{s}^{(t)}}\left\{\frac{1}{n_t}\sum_{i=1}^{n_t}\mathcal{L}\left(f\left(\boldsymbol{x}_i^{(t)};\mathbf{L}\boldsymbol{s}^{(t)}\right),y_i^{(t)}\right)+\mu\|\boldsymbol{s}^{(t)}\|_1\right\}+\lambda\|\mathbf{L}\|_{\mathsf{F}}^2$$

ELLA: Given a new task t,

1. Train a single-task model $\boldsymbol{\theta}^{(t)}$ for task t
2. Reconstruct $\boldsymbol{\theta}^{(t)}$ in the current basis (LASSO)

$$\boldsymbol{s}^{(t)} \leftarrow \arg\min_{\boldsymbol{s}^{(t)}} \ell(\mathbf{L}_m, \boldsymbol{s}^{(t)}, \boldsymbol{\theta}^{(t)}, \boldsymbol{D}^{(t)})$$

3. Update the basis

$$\mathbf{L}_{m+1} \leftarrow \arg\min_{\mathbf{L}} \lambda\|\mathbf{L}\|_{\mathsf{F}}^2 + \underbrace{\frac{1}{T}\sum_{t=1}^{T}\ell\left(\mathbf{L},\boldsymbol{s}^{(t)},\boldsymbol{\theta}^{(t)},\boldsymbol{D}^{(t)}\right)}_{\text{in practice, } \mathbf{L} \text{ is constructed incrementally with each task}}$$

where $\ell\left(\mathbf{L},\mathbf{s},\boldsymbol{\theta},\mathbf{D}\right) = \mu\left\|\mathbf{s}\right\|_1 + \|\boldsymbol{\theta}-\mathbf{L}\mathbf{s}\|_{\mathbf{D}}^2$

$\boldsymbol{D}^{(t)}$ is ½ the Hessian of the single-task loss evaluated at $\boldsymbol{\theta}^{(t)}$

$\|\mathbf{x}\|_{\mathbf{D}}^2 = \mathbf{x}^\top\mathbf{D}\mathbf{x}$

Efficient Lifelong Learning Algorithm

[Ruvolo & Eaton, ICML'13]

- ELLA's per-task computational complexity is:
  1. Independent of the number of tasks T
  2. Independent of the numbers of training instances for previous tasks
- We have a variety of theoretical guarantees on ELLA's performance and convergence
- Online dictionary learning for sparse coding [Mairal et al ICML'09] is a special case of ELLA

Summary of Previous Results

[Ruvolo & Eaton, ICML'13]

ELLA achieves nearly identical accuracy to batch MTL:

Dataset	Problem Type	Batch MTL Accuracy	ELLA Relative Accuracy	OMTL Relative Accuracy	STL Relative Accuracy
Land Mine	Classification	0.7802 ± 0.013 (AUC)	$99.73 \pm 0.7\%$	$82.2 \pm 3.0\%$	$97.97 \pm 1.5\%$
Facial Expr.	Classification	0.6577 ± 0.021 (AUC)	$99.37 \pm 3.1\%$	$97.58 \pm 3.8\%$	$97.34 \pm 3.9\%$
Syn. Data	Regression	-1.084 ± 0.006 (-rMSE)	$97.74 \pm 2.7\%$	N/A	$92.91 \pm 1.5\%$
London Sch.	Regression	-10.10 ± 0.066 (-rMSE)	$98.90 \pm 1.5\%$	N/A	$97.20 \pm 0.4\%$

Batch MTL = [Kumar & Daumé III, ICML'12]

OMTL = [Saha et al, AISTATS'11]

Summary of Previous Results

[Ruvolo & Eaton, ICML'13]

ELLA achieves nearly identical accuracy to batch MTL:

Dataset	Problem Type	Batch MTL Accuracy	ELLA Relative Accuracy	OMTL Relative Accuracy	STL Relative Accuracy
Land Mine	Classification	0.7802 ± 0.013 (AUC)	$99.73 \pm 0.7\%$	$82.2 \pm 3.0\%$	$97.97 \pm 1.5\%$
Facial Expr.	Classification	0.6577 ± 0.021 (AUC)	$99.37 \pm 3.1\%$	$97.58 \pm 3.8\%$	$97.34 \pm 3.9\%$
Syn. Data	Regression	-1.084 ± 0.006 (-rMSE)	$97.74 \pm 2.7\%$	N/A	$92.91 \pm 1.5\%$
London Sch.	Regression	-10.10 ± 0.066 (-rMSE)	$98.90 \pm 1.5\%$	N/A	$97.20 \pm 0.4\%$

While obtaining speedups of:

- over 1,000x for learning all tasks

Dataset	Batch Runtime (seconds)	ELLA All Tasks (speedup)	ELLA New Task (speedup)	OMTL All Tasks (speedup)	OMTL New Task (speedup)	STL All Tasks (speedup)	STL New Task (speedup)
Land Mine	231 ± 6.2	$1{,}350 \pm 58$	$39{,}150 \pm 1{,}682$	22 ± 0.88	638 ± 25	$3{,}342 \pm 409$	$96{,}918 \pm 11{,}861$
Facial Expr.	$2{,}200 \pm 92$	$1{,}828 \pm 100$	$38{,}400 \pm 2{,}100$	948 ± 65	$19{,}900 \pm 1{,}360$	$8{,}511 \pm 1{,}107$	$178{,}719 \pm 23{,}239$
Syn. Data	$1{,}300 \pm 141$	$5{,}026 \pm 685$	$502{,}600 \pm 68{,}500$	N/A	N/A	$156{,}489 \pm 17{,}564$	1.6E6±1.8E5
London Sch.	715 ± 36	$2{,}721 \pm 225$	$378{,}219 \pm 31{,}275$	N/A	N/A	$36{,}000 \pm 4{,}800$	5.0E6±6.7E5

Batch MTL = [Kumar & Daumé III, ICML'12] **OMTL = [Saha et al, AISTATS'11]**

Summary of Previous Results

[Ruvolo & Eaton, ICML'13]

ELLA achieves nearly identical accuracy to batch MTL:

Dataset	Problem Type	Batch MTL Accuracy	ELLA Relative Accuracy	OMTL Relative Accuracy	STL Relative Accuracy
Land Mine	Classification	0.7802 ± 0.013 (AUC)	$99.73 \pm 0.7\%$	$82.2 \pm 3.0\%$	$97.97 \pm 1.5\%$
Facial Expr.	Classification	0.6577 ± 0.021 (AUC)	$99.37 \pm 3.1\%$	$97.58 \pm 3.8\%$	$97.34 \pm 3.9\%$
Syn. Data	Regression	-1.084 ± 0.006 (-rMSE)	$97.74 \pm 2.7\%$	N/A	$92.91 \pm 1.5\%$
London Sch.	Regression	-10.10 ± 0.066 (-rMSE)	$98.90 \pm 1.5\%$	N/A	$97.20 \pm 0.4\%$

While obtaining speedups of:

- over 1,000x for learning all tasks
- over 38,000x for learning each new task

Dataset	Batch Runtime (seconds)	ELLA All Tasks (speedup)	ELLA New Task (speedup)	OMTL All Tasks (speedup)	OMTL New Task (speedup)	STL All Tasks (speedup)	STL New Task (speedup)
Land Mine	231 ± 6.2	$1{,}350 \pm 58$	$39{,}150 \pm 1{,}682$	22 ± 0.88	638 ± 25	$3{,}342 \pm 409$	$96{,}918 \pm 11{,}861$
Facial Expr.	$2{,}200 \pm 92$	$1{,}828 \pm 100$	$38{,}400 \pm 2{,}100$	948 ± 65	$19{,}900 \pm 1{,}360$	$8{,}511 \pm 1{,}107$	$178{,}719 \pm 23{,}239$
Syn. Data	$1{,}300 \pm 141$	$5{,}026 \pm 685$	$502{,}600 \pm 68{,}500$	N/A	N/A	$156{,}489 \pm 17{,}564$	1.6E6$\pm$1.8E5
London Sch.	715 ± 36	$2{,}721 \pm 225$	$378{,}219 \pm 31{,}275$	N/A	N/A	$36{,}000 \pm 4{,}800$	5.0E6$\pm$6.7E5

Batch MTL = [Kumar & Daumé III, ICML'12] **OMTL = [Saha et al, AISTATS'11]**

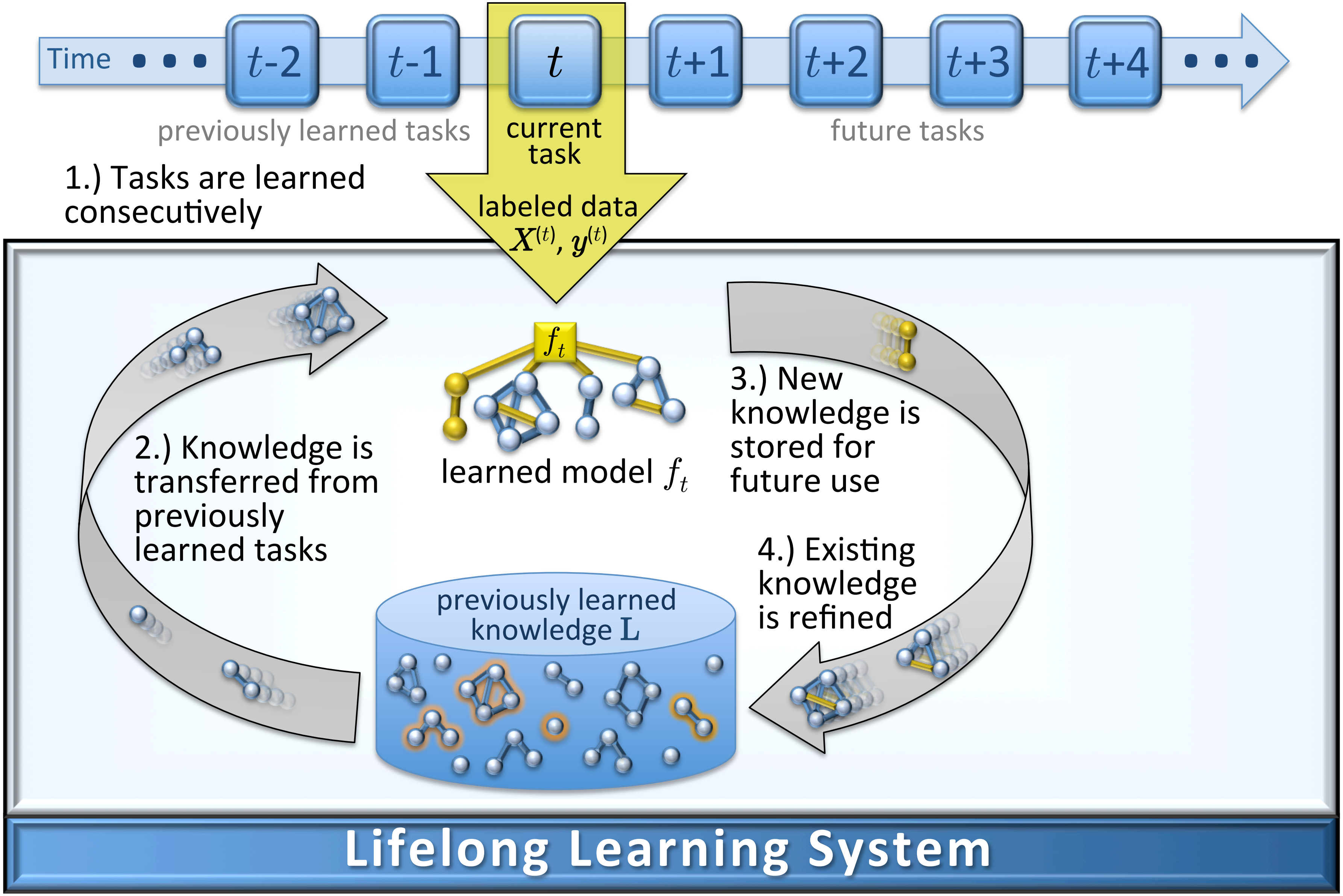
Outline

- Introduction
- **Efficient Lifelong Learning Algorithm**
- Active Task Selection
- Targeted Active Task Selection

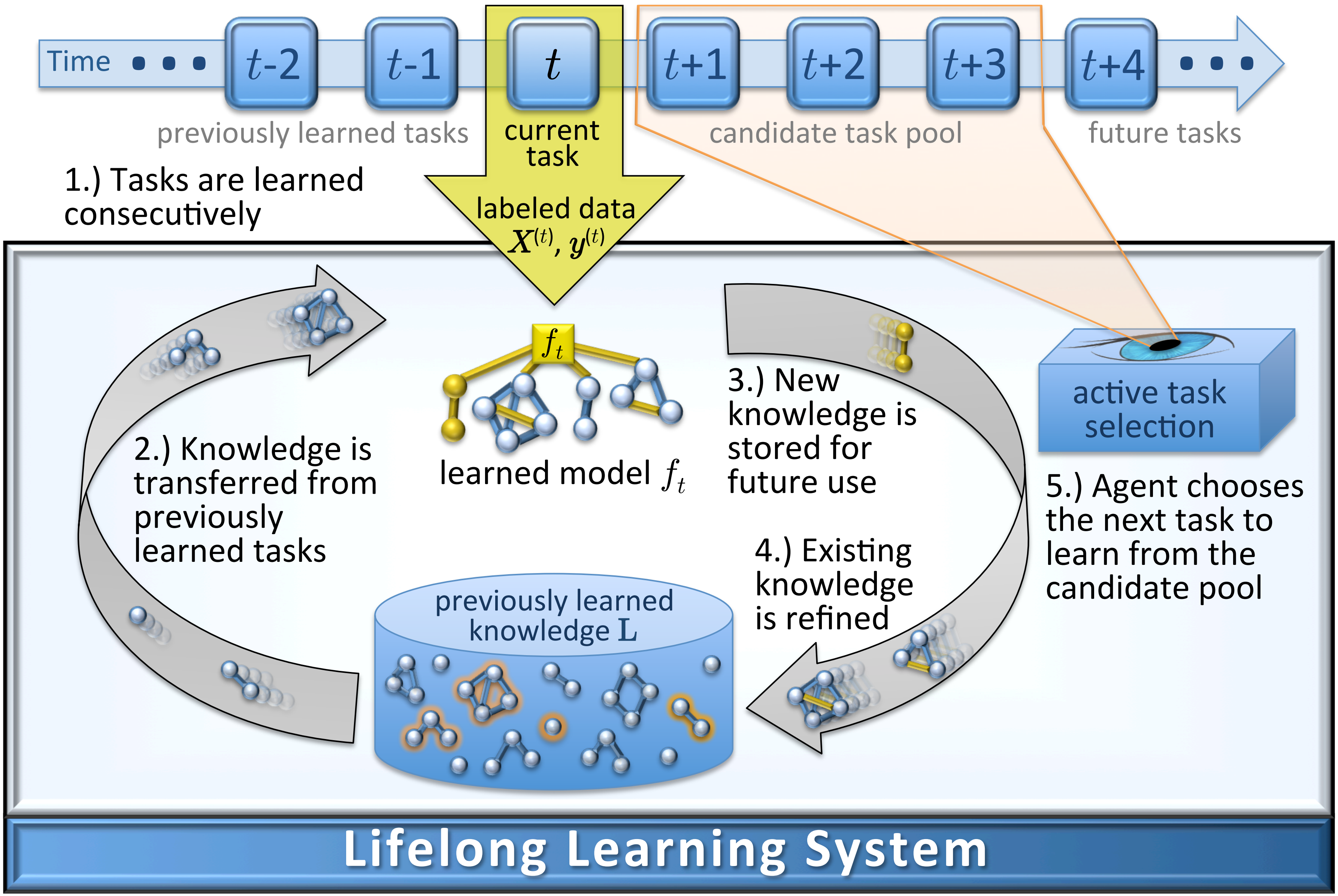
Outline

- Introduction
- Efficient Lifelong Learning Algorithm
- **Active Task Selection**
- Targeted Active Task Selection

Active Task Selection



Active Task Selection



Information Maximization Approach

- **Idea:** Select the task that maximizes the information gain on $\mathbf{L}$

current information available

$$t_{\text{next}} = \arg\min_{t} \iint \underbrace{p(\boldsymbol{\theta}^{(t)} = \mathbf{u}, \mathbf{D}^{(t)} = \mathbf{V} | \mathcal{I}_m)}_{\text{probability of model for task } t} \times \underbrace{H\left[\mathbf{L} | \boldsymbol{\theta}^{(t)} = \mathbf{u}, \mathbf{D}^{(t)} = \mathbf{V}, \mathcal{I}_m\right]}_{\text{differential entropy of } \mathbf{L}} d\mathbf{u}\, d\mathbf{V}$$

Information Maximization Approach

- **Idea:** Select the task that maximizes the information gain on $\mathbf{L}$

$$t_{\text{next}} = \arg\min_{t} \iint \underbrace{p(\boldsymbol{\theta}^{(t)} = \mathbf{u}, \mathbf{D}^{(t)} = \mathbf{V} \mid \mathcal{I}_m)}_{\text{probability of model for task } t} \times \underbrace{H\left[\mathbf{L} \mid \boldsymbol{\theta}^{(t)} = \mathbf{u}, \mathbf{D}^{(t)} = \mathbf{V}, \overset{\substack{\text{current information available}\\ \downarrow}}{\mathcal{I}_m}\right]}_{\text{differential entropy of } \mathbf{L}} d\mathbf{u}\, d\mathbf{V}$$

- To compute this efficiently, we
  1. Approximate the model probability using a Dirac delta function around the optimal single task model ($\hat{\boldsymbol{\theta}}^{(t)}$, $\hat{\mathbf{D}}^{(t)}$)
  2. Use a Laplace approximation of $\mathbf{L}$'s density as a multivariate Gaussian for the entropy term

Information Maximization Approach

- **Idea:** Select the task that maximizes the information gain on $\mathbf{L}$

$$t_{\text{next}} = \arg\min_{t} \iint \underbrace{p(\boldsymbol{\theta}^{(t)} = \mathbf{u}, \mathbf{D}^{(t)} = \mathbf{V} | \mathcal{I}_m)}_{\text{probability of model for task } t} \times \underbrace{H\left[\mathbf{L} | \boldsymbol{\theta}^{(t)} = \mathbf{u}, \mathbf{D}^{(t)} = \mathbf{V}, \overset{\substack{\text{current information available} \\ \downarrow}}{\mathcal{I}_m}\right]}_{\text{differential entropy of } \mathbf{L}} d\mathbf{u}\, d\mathbf{V}$$

- To compute this efficiently, we
  1. Approximate the model probability using a Dirac delta function around the optimal single task model ($\hat{\boldsymbol{\theta}}^{(t)}$, $\hat{\mathbf{D}}^{(t)}$)
  2. Use a Laplace approximation of $\mathbf{L}$'s density as a multivariate Gaussian for the entropy term
- This yields the following InfoMax task selection rule:

$$t_{\text{next}} = \arg\min_{t \in \{T+1, \ldots, T_{\text{pool}}\}} \ln \left| \boldsymbol{\Sigma}^{(t)} \right|$$

$$\boldsymbol{\Sigma}^{(t)} = \mathrm{Cov}\left[\mathrm{vec}(\mathbf{L}) | \boldsymbol{\theta}^{(t)} = \hat{\boldsymbol{\theta}}^{(t)}, \mathbf{D}^{(t)} = \hat{\mathbf{D}}^{(t)}, \mathcal{I}_m\right]$$

Diversity Approach

- **Idea**: Focus on candidate tasks that are poorly encoded by the current $\mathbf{L}$
  - This encourages $\mathbf{L}$ to serve as an effective basis for a variety of tasks

Diversity Approach

- **Idea**: Focus on candidate tasks that are poorly encoded by the current $\mathbf{L}$
  - This encourages $\mathbf{L}$ to serve as an effective basis for a variety of tasks
- Select the candidate task that the current $\mathbf{L}$ is doing the <u>worst</u> job solving:

$$t_{\text{next}} = \underset{t \in \{T+1,\ldots,T_{\text{pool}}\}}{\arg\max} \underbrace{\min_{\mathbf{s}} \ell\left(\mathbf{L}_m, \mathbf{s}, \hat{\boldsymbol{\theta}}^{(t)}, \hat{\mathbf{D}}^{(t)}\right)}_{\text{loss in reconstructing model for task } t}$$

Diversity Approach

- **Idea**: Focus on candidate tasks that are poorly encoded by the current $\mathbf{L}$
  - This encourages $\mathbf{L}$ to serve as an effective basis for a variety of tasks

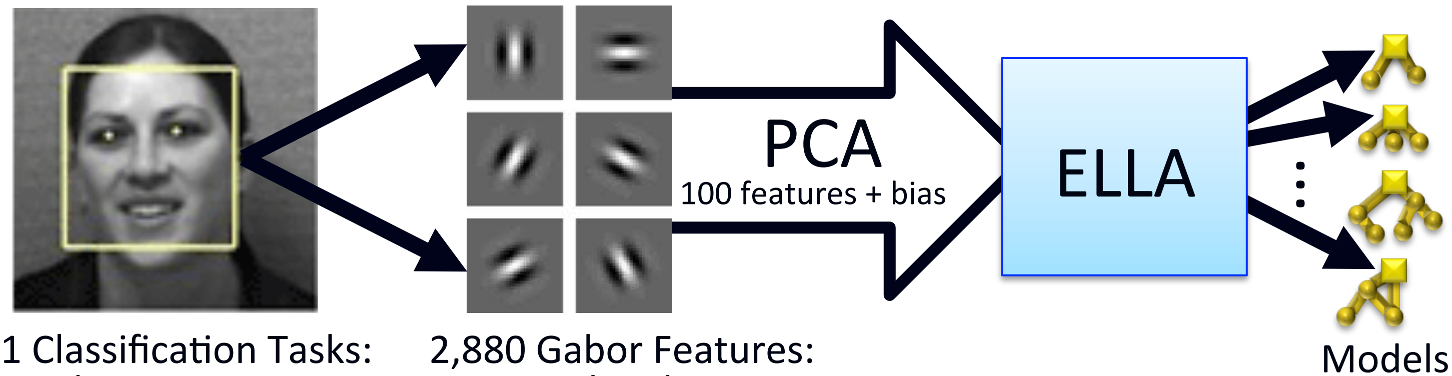
- Select the candidate task that the current $\mathbf{L}$ is doing the worst job solving:

$$t_{\text{next}} = \underset{t \in \{T+1,\ldots,T_{\text{pool}}\}}{\arg\max} \underbrace{\min_{\mathbf{s}} \ \ell\left(\mathbf{L}_m, \mathbf{s}, \hat{\boldsymbol{\theta}}^{(t)}, \hat{\mathbf{D}}^{(t)}\right)}_{\text{loss in reconstructing model for task } t}$$

  - We also explore a probabilistic version (Diversity++) that chooses a task proportionally to its reconstruction loss

Applications

Facial Expression Recognition: identify presence of facial action units (#5 upper lid raiser, #10 upper lip raiser, #12 lip corner pull)

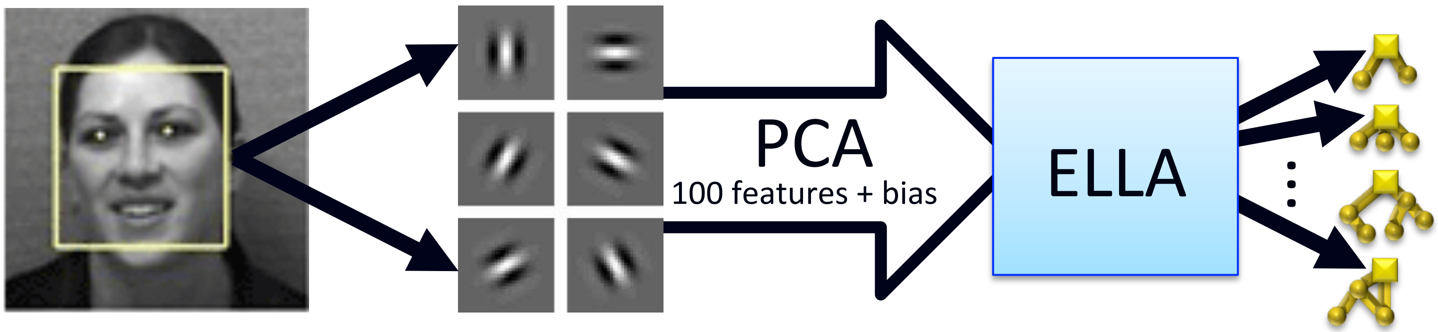


21 Classification Tasks:
- 7 subjects
- 3 action units
- 450-999 images each

2,880 Gabor Features:
- 2 spatial scales
- 4 orientations
- 576 locations

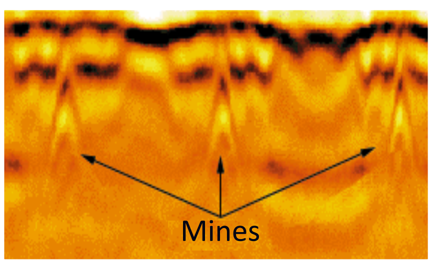
Applications

Facial Expression Recognition: identify presence of facial action units (#5 upper lid raiser, #10 upper lip raiser, #12 lip corner pull)

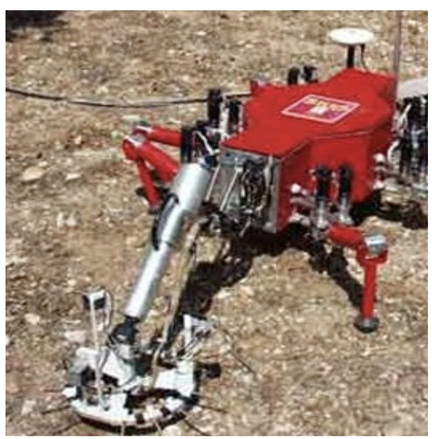


21 Classification Tasks 2,880 Gabor Features Models

Land Mine Detection from radar images [Xue et al. 2007]

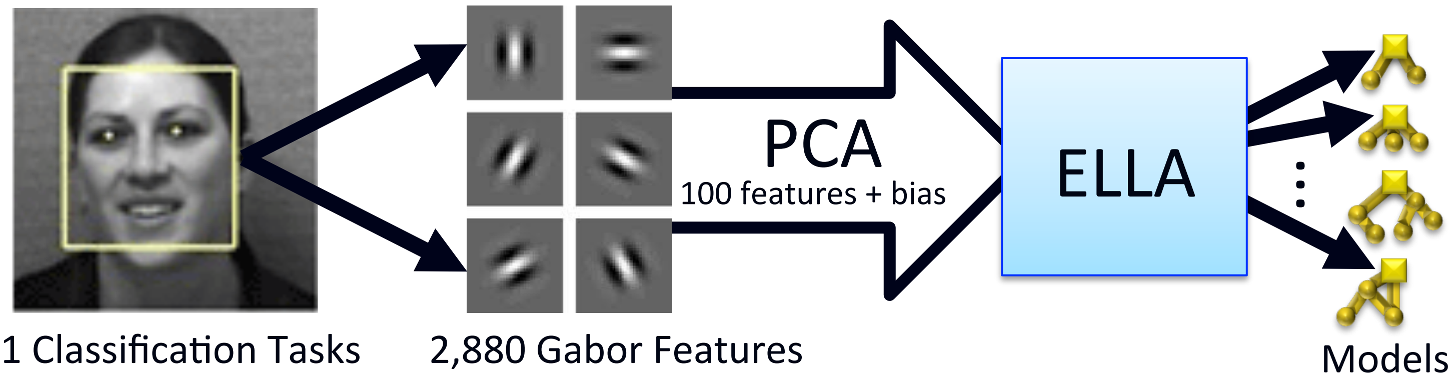


29 Classification Tasks:
- 29 regions
- 2 terrain types
- 14,820 instances total

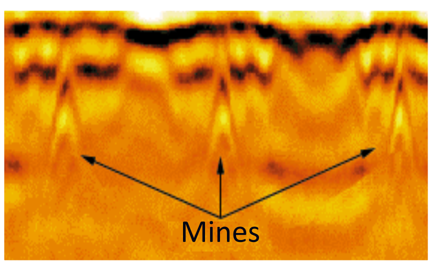


Applications

Facial Expression Recognition: identify presence of facial action units (#5 upper lid raiser, #10 upper lip raiser, #12 lip corner pull)

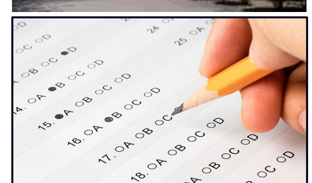


Land Mine Detection from radar images [Xue et al. 2007]



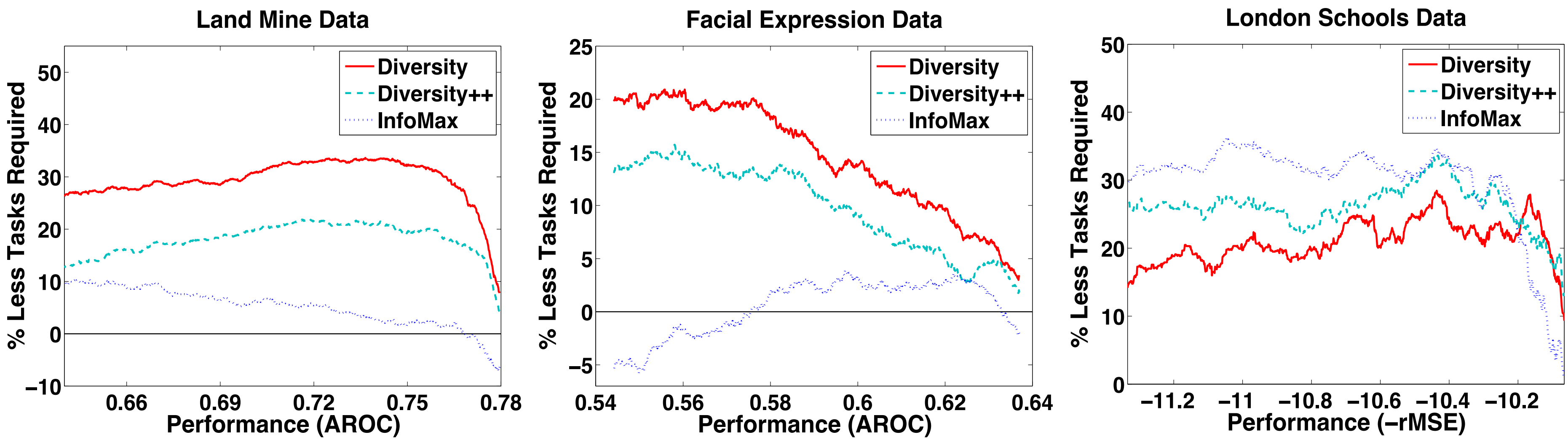
29 Classification Tasks:
- 29 regions
- 2 terrain types
- 14,820 instances total

Exam Score Prediction for London schools [Kumar et al. 2012]



139 Regression Tasks:
- 139 schools
- 15,362 students total
- 4 school-specific features
- 3 student-specific features
- Exam year + bias term

Active Task Selection Results



Plots show the relevant efficiency (in #tasks) as compared to random task selection

Average Task Reduction (%)

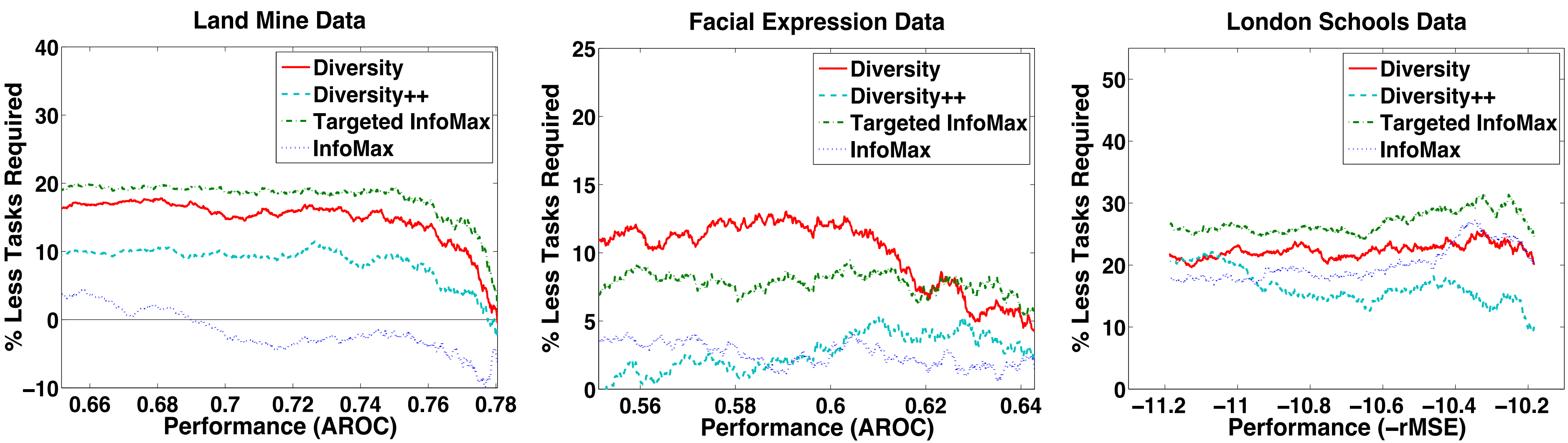
Data Set	InfoMax	Diversity	Diversity++
Land Mine	5.1±3.7	**29.4±4.1**	18.1±3.0
Facial Expr.	0.5±2.6	**14.6±5.1**	9.9±4.0
Syn. Data	10.2±7.9	**20.2±6.7**	17.0±5.9
London Sch.	**29.8±6.8**	21.0±3.1	26.2±3.1

Targeted InfoMax Task Selection

- The general InfoMax selection method tries to maximize the information gain on $\mathbf{L}$
  - Focuses on an *unknown* set of future tasks

- What if we are working toward learning a specific target set of future tasks?
  - Can improve performance by *targeting* InfoMax toward those tasks

Targeted Task Selection Results



Plots show the relevant efficiency (in #tasks) as compared to random task selection

Average Task Reduction (%)

Data Set	Targeted InfoMax	InfoMax	Diversity	Diversity++
Land Mine	**17.9±2.7**	-1.7±3.0	14.9±3.2	8.5±2.5
Facial Expr.	7.8±0.7	2.6±0.8	**10.0±2.5**	2.7±1.3
Syn. Data	**38.4±7.5**	11.4±5.6	19.9±4.9	16.6±5.0
London Sch.	**26.9±1.8**	20.1±2.8	22.3±1.1	16.4±2.7

Conclusions

- We presented two approaches to active task selection in a lifelong learning setting
  - Diversity approach is cheap and effective
  - InfoMax works well for targeted knowledge acquisition

- **Future work:** integrating with instance-based active learning and guidance from a teacher

Active task selection accelerates knowledge acquisition in a lifelong learning setting

Thank you!

Code for ELLA & active task selection is available at
cs.brynmawr.edu/~eeaton

This work was supported by
ONR Grant #N00014-11-1-0139