BRYN MAWR
COLLEGE

ELLA: An Efficient Lifelong Machine Learning Algorithm

Paul Ruvolo

Eric Eaton

Bryn Mawr College
Computer Science Department

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

Overview

- 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

Overview

- 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. Optimizes performance over all 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

Overview

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

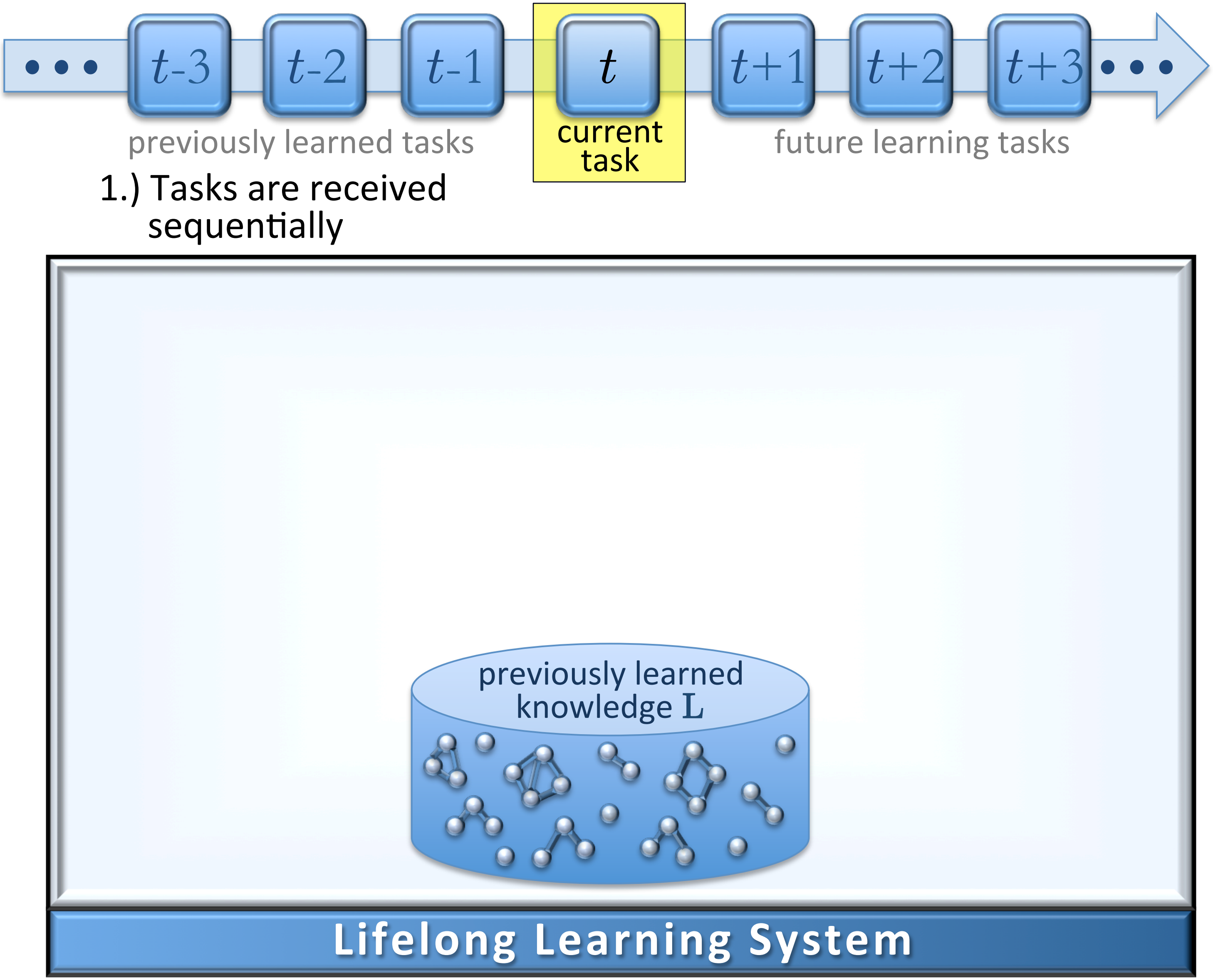
- **ELLA's Capabilities**:
  1. Learns tasks consecutively
  2. Transfers knowledge from previous tasks
  3. Optimizes performance over all tasks
  4. Theoretical guarantees on performance and convergence

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

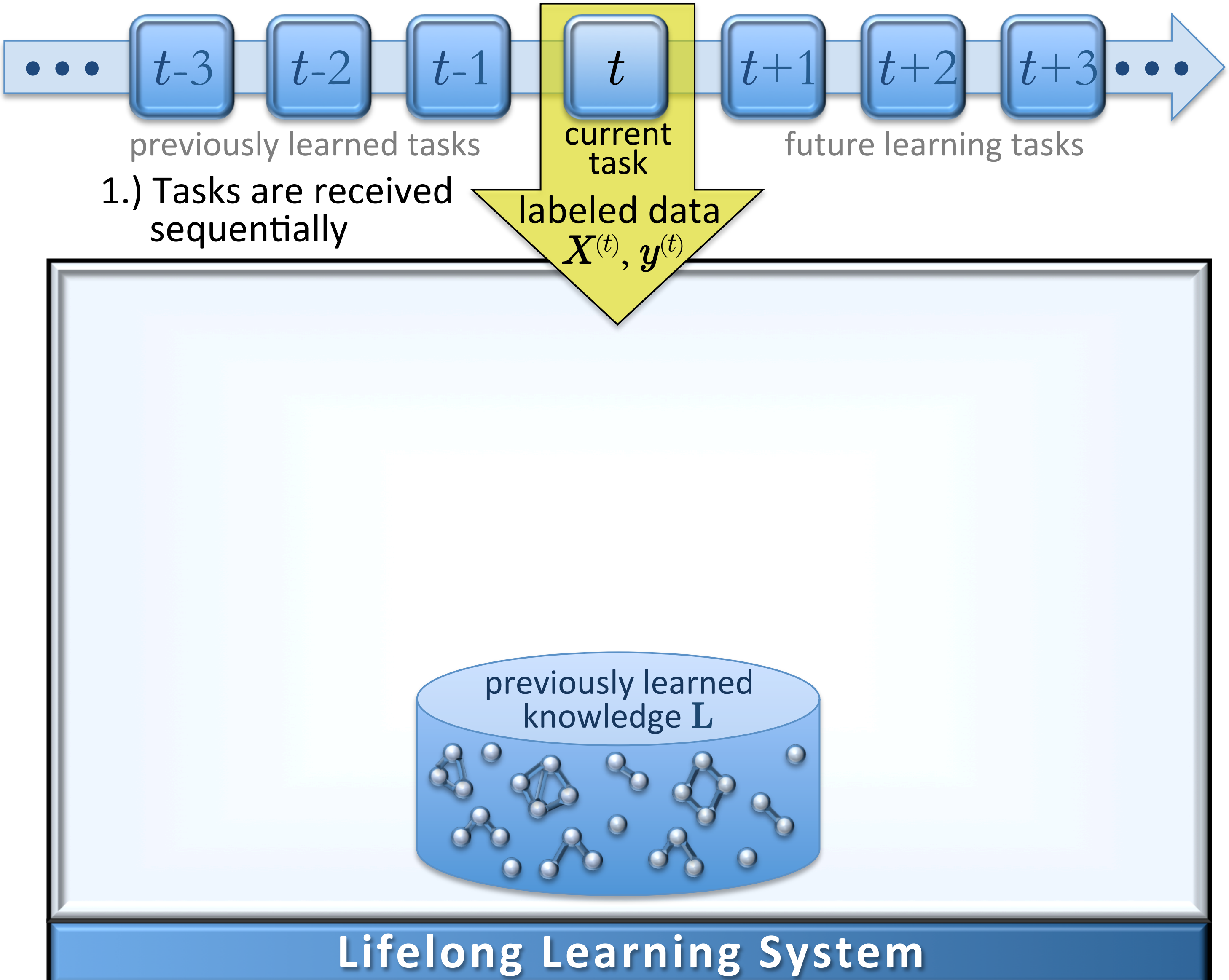
Lifelong learning includes elements of both transfer and multi-task learning

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

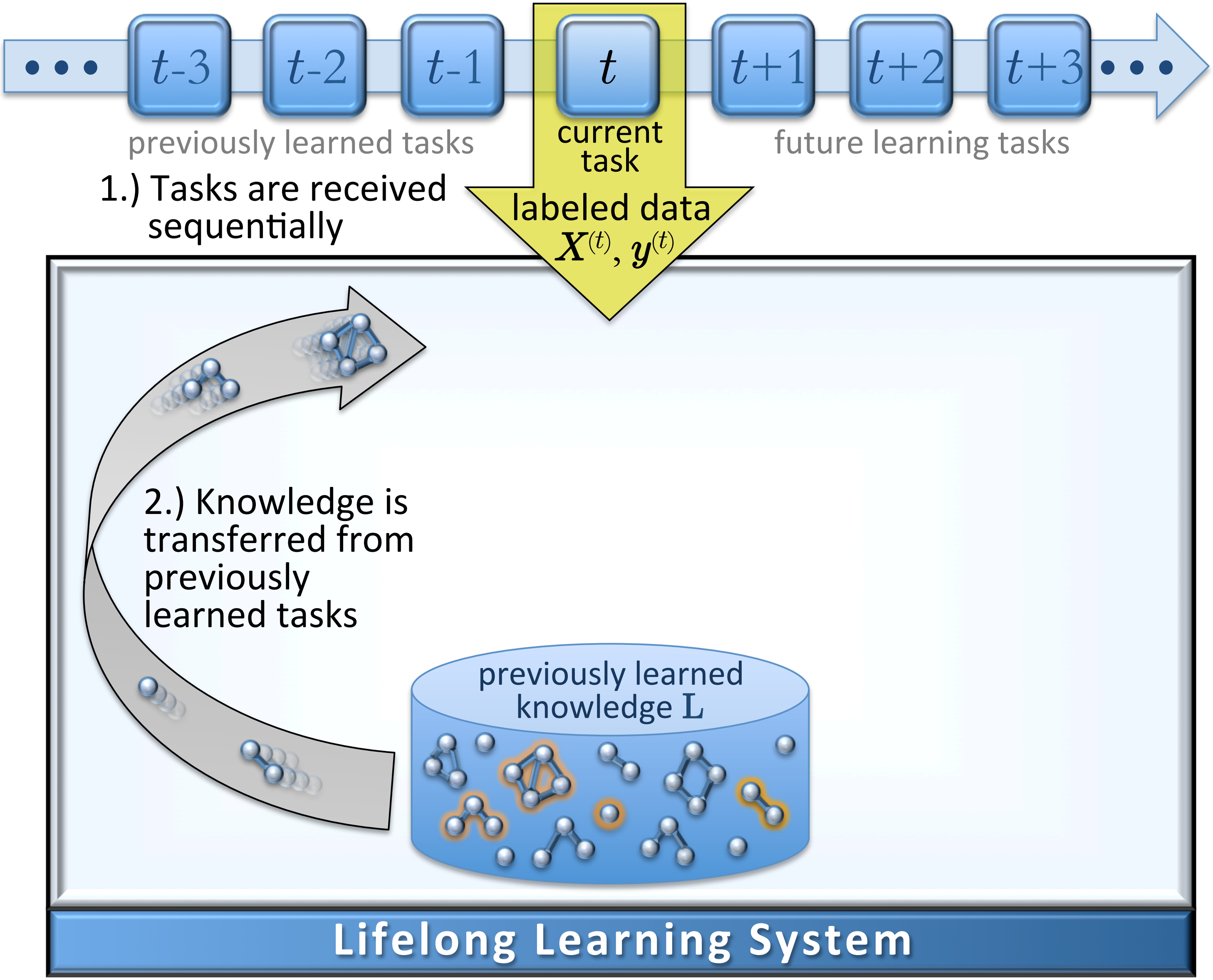
Lifelong Machine Learning



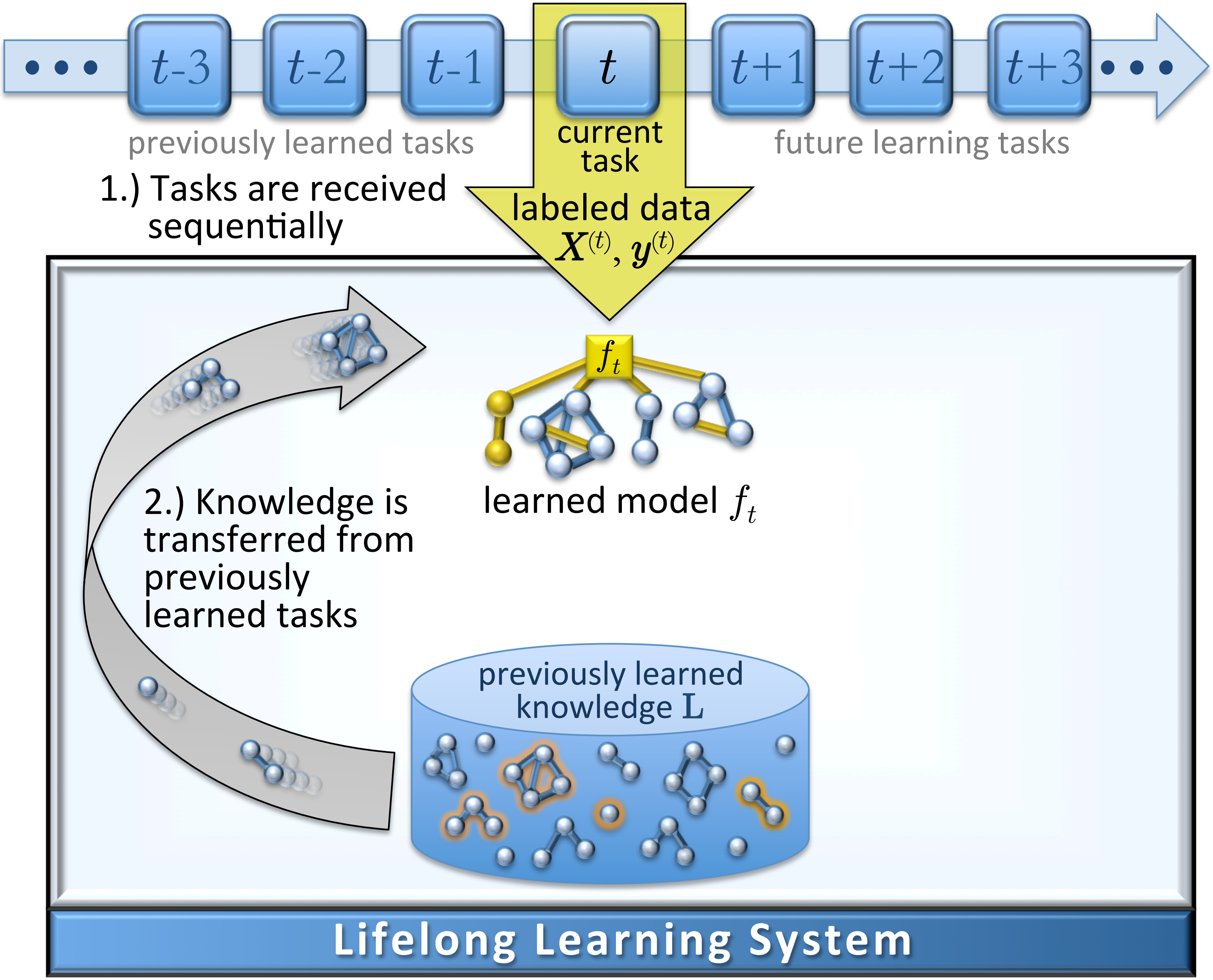
Lifelong Machine Learning



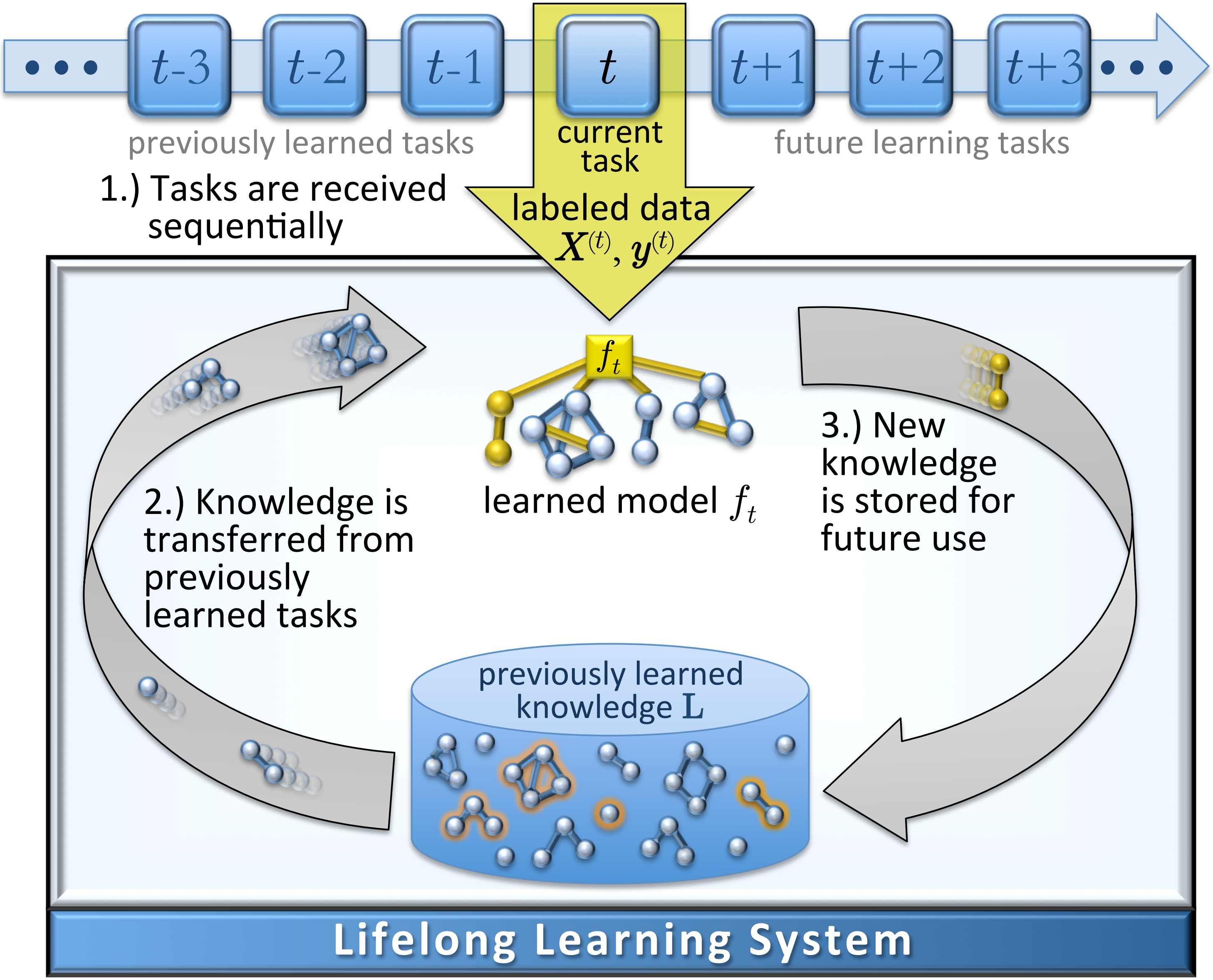
Lifelong Machine Learning



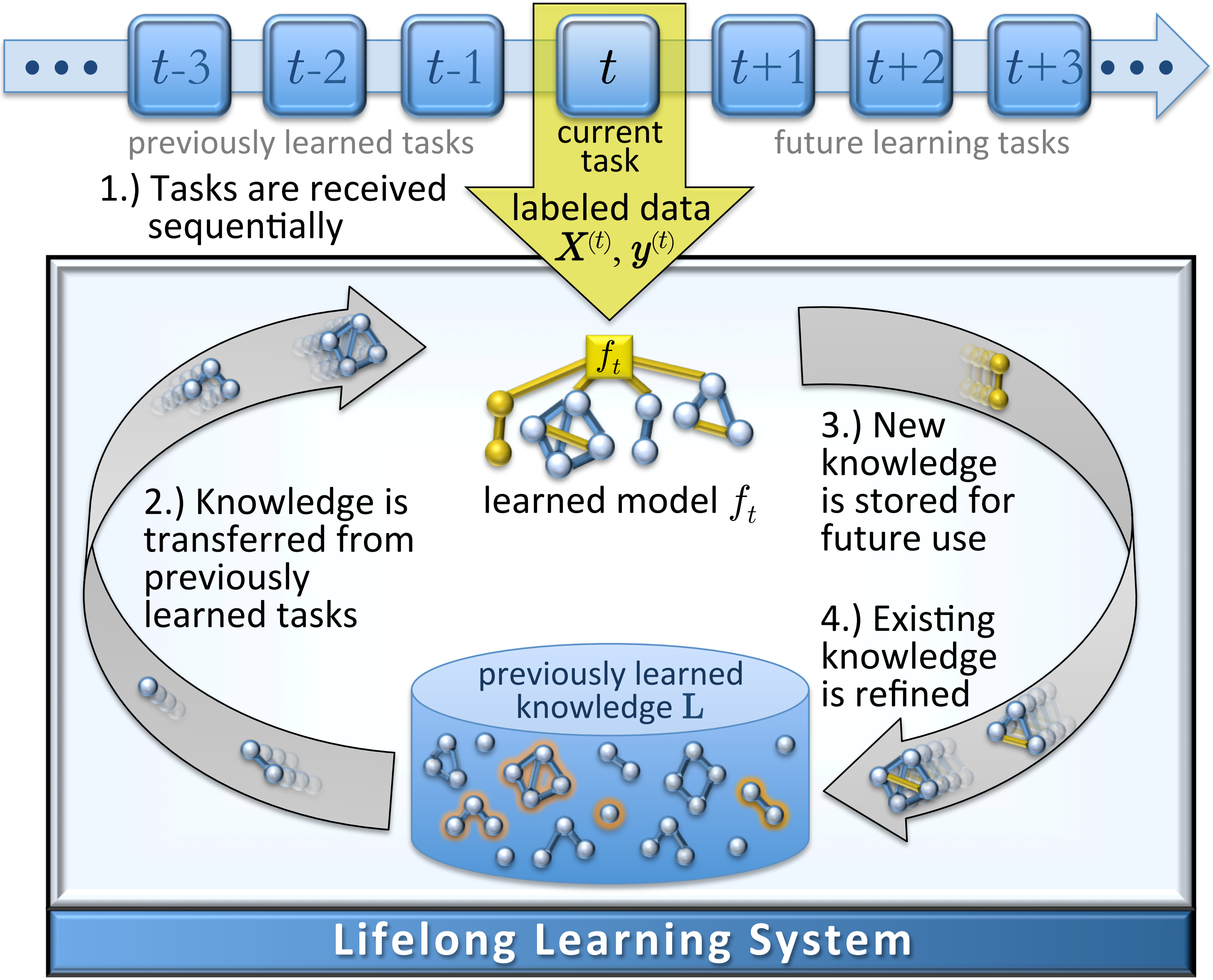
Lifelong Machine Learning



Lifelong Machine Learning



Lifelong Machine Learning



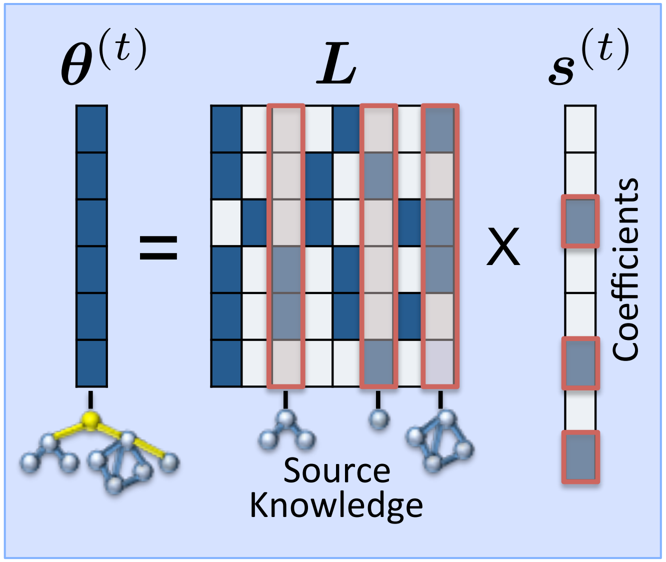
Task Structure Model

- 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 $\mathbf{L}$

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



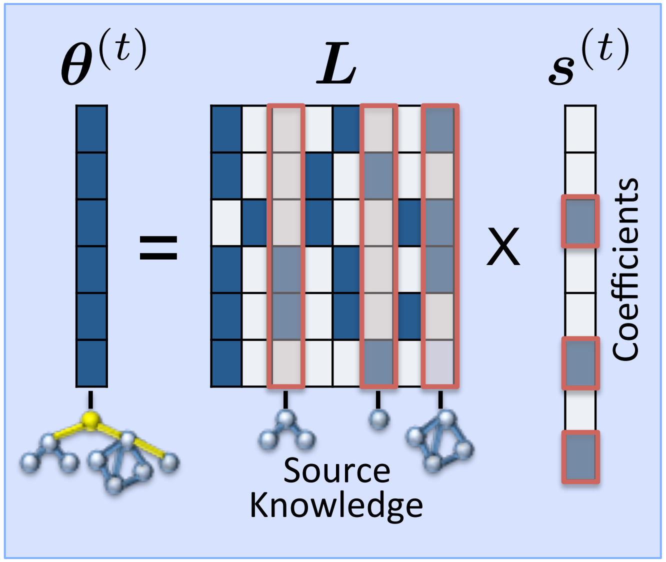
Task Structure Model

- 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 $\mathbf{L}$

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



Objective Function:

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

Efficient Lifelong Learning

Objective Function:

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

Problem 1: The complexity of the inner summation scales linearly with the number of training instances

Our solution: Replace the model-fit-to-data term with the second-order Taylor expansion around the optimal single task model:

$$g_T(\mathbf{L}) = \frac{1}{T}\sum_{t=1}^{T} \min_{\boldsymbol{s}^{(t)}} \left\{ \|\boldsymbol{\theta}^{(t)} - \mathbf{L}\boldsymbol{s}^{(t)}\|_{\mathbf{D}^{(t)}}^2 + \mu \|\boldsymbol{s}^{(t)}\|_1 \right\} + \lambda \|\mathbf{L}\|_{\mathsf{F}}^2$$

where, $\boldsymbol{\theta}^{(t)} = \arg\min_{\boldsymbol{\theta}} \frac{1}{n_t}\sum_{i=1}^{n_t} \mathcal{L}\left(f\left(\boldsymbol{x}_i^{(t)}; \boldsymbol{\theta}\right), y_i^{(t)}\right)$

$\boldsymbol{D}^{(t)}$ 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

Objective Function:

$$g_T(\mathbf{L}) = \frac{1}{T}\sum_{t=1}^{T}\min_{\boldsymbol{s}^{(t)}}\left\{\|\boldsymbol{\theta}^{(t)} - \mathbf{L}\boldsymbol{s}^{(t)}\|^2_{\mathbf{D}^{(t)}} + \mu\|\boldsymbol{s}^{(t)}\|_1\right\} + \lambda\|\mathbf{L}\|^2_{\mathsf{F}}$$

Problem 2: The complexity of the outer summation grows linearly with the number of tasks T

Our solution: Optimize $\boldsymbol{s}^{(t)}$ only when training on task t and not on any other tasks

- We prove that the penalty for not re-optimizing the other $\boldsymbol{s}^{(t)}$'s vanishes as T gets large

Efficient Lifelong Learning Algorithm

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

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

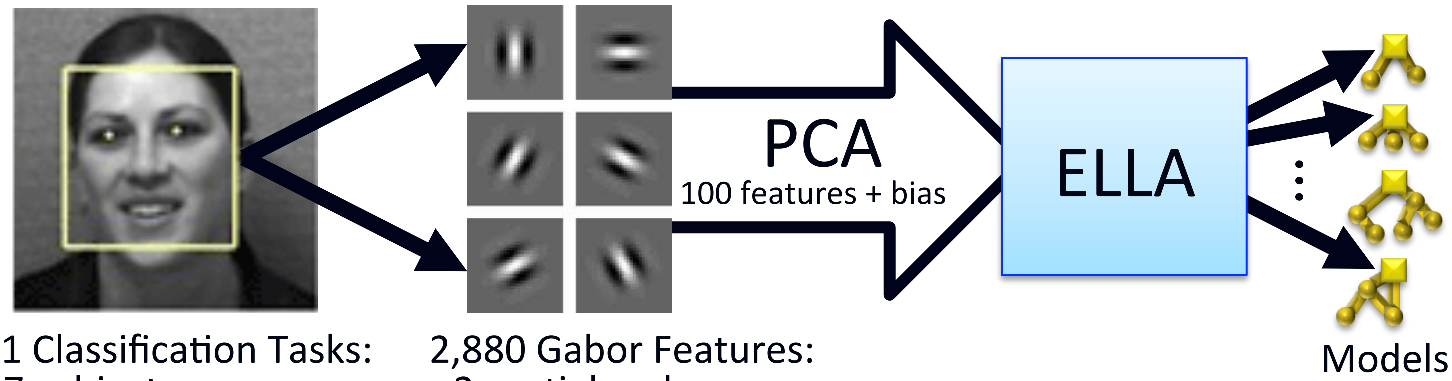
$\|\mathbf{x}\|_{\mathbf{D}}^2 = \mathbf{x}^\top\mathbf{D}\mathbf{x}$

Efficient Lifelong Learning

- 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 show 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

Applications

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

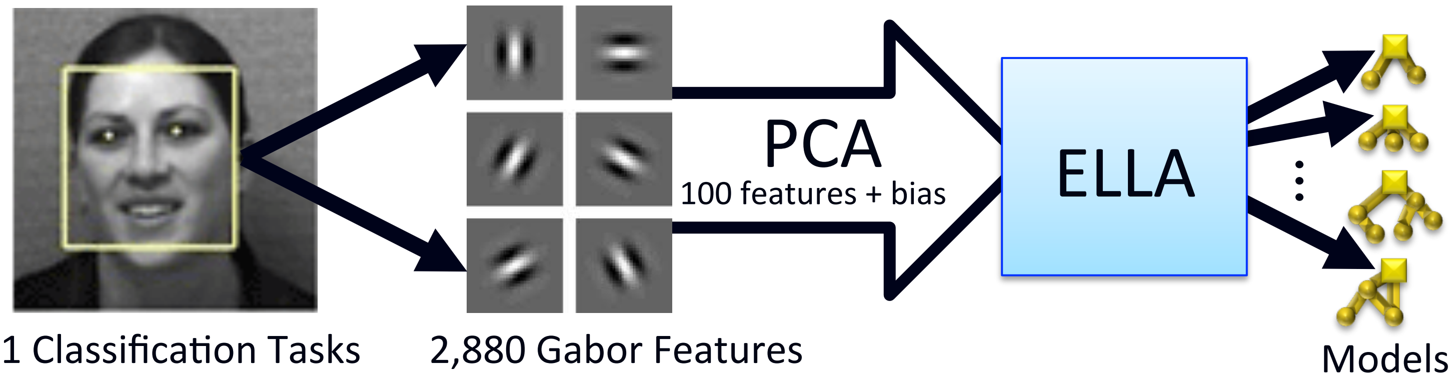


21 Classification Tasks:
- 7 subjects
- 3 action units
- 450-999 images each

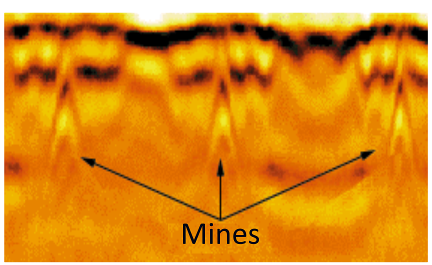
2,880 Gabor Features:
- 2 spatial scales
- 4 orientations
- 576 locations

Applications

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

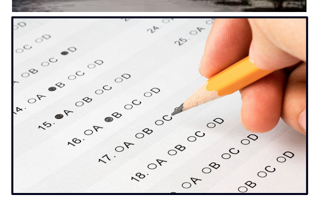


Land Mine Detection from radar images [Xue et al. 2007]



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

Empirical Results

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]**

Empirical Results

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±58	39,150±1,682	22±0.88	638±25	3,342±409	96,918±11,861
Facial Expr.	2,200±92	1,828±100	38,400±2,100	948±65	19,900±1,360	8,511±1,107	178,719±23,239
Syn. Data	1,300±141	5,026±685	502,600±68,500	N/A	N/A	156,489±17,564	1.6E6±1.8E5
London Sch.	715±36	2,721±225	378,219±31,275	N/A	N/A	36,000±4,800	5.0E6±6.7E5

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

OMTL = [Saha et al, AISTATS'11]

Empirical Results

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:

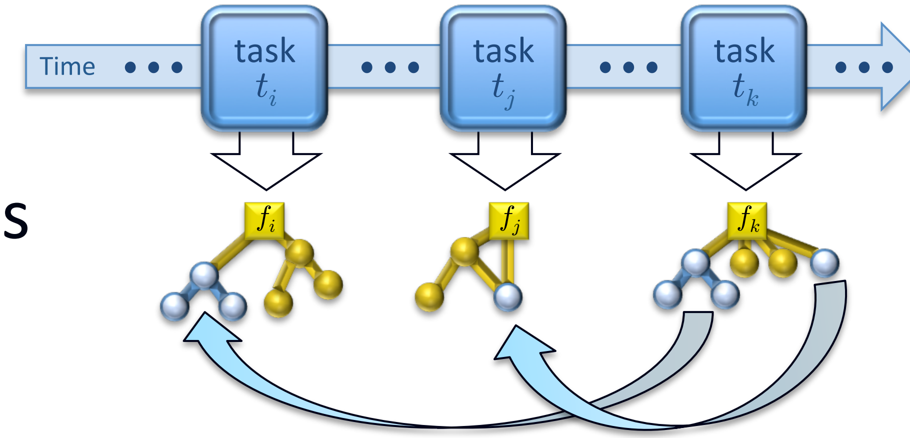
- over 1,000x for learning all tasks
- over 38,000x for learning each new task

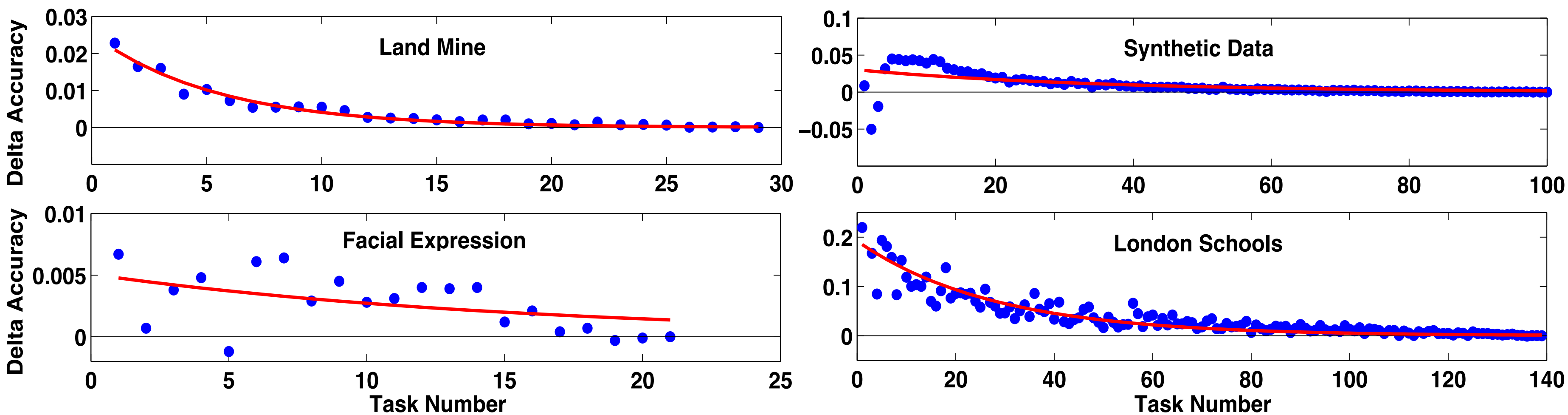
Dataset	Batch Runtime (seconds)	ELLA All Tasks (speedup)	ELLA New Task (speedup)	OMTL All Tasks (speedup)	OMTL New Task (speedup)	STL All Tasks (speedup)	STL New Task (speedup)
Land Mine	231±6.2	1,350±58	39,150±1,682	22±0.88	638±25	3,342±409	96,918±11,861
Facial Expr.	2,200±92	1,828±100	38,400±2,100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

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

Reverse Transfer in ELLA

- Earlier task models improve from later learning <u>without retraining</u> on the earlier tasks





Thank you!

Code for ELLA is available at cs.brynmawr.edu/~eeaton

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