# Lifelong Machine Learning Systems: Beyond Learning Algorithms

**Daniel L. Silver**
Jodrey School of Computer Science
Acadia University,
Wolfville, Nova Scotia, Canada B4P 2R6

**Qiang Yang and Lianghao Li**
Department of Computer Science and Engineering,
Hong Kong University of Science and Technology,
Clearwater Bay, Kowloon, Hong Kong

## Abstract

Lifelong Machine Learning, or LML, considers systems that can learn many tasks from one or more domains over its lifetime. The goal is to sequentially retain learned knowledge and to selectively transfer that knowledge when learning a new task so as to develop more accurate hypotheses or policies. Following a review of prior work on LML, we propose that it is now appropriate for the AI community to move beyond learning algorithms to more seriously consider the nature of systems that are capable of learning over a lifetime. Reasons for our position are presented and potential counter-arguments are discussed. The remainder of the paper contributes by defining LML, presenting a reference framework that considers all forms of machine learning, and listing several key challenges for and benefits from LML research. We conclude with ideas for next steps to advance the field.

## Introduction

Over the last 25 years there have been significant advances in machine learning theory and algorithms. However, there has been comparatively little work on systems that use these algorithms to learn a variety of tasks over an extended period of time such that the knowledge of the tasks is retained and used to improve learning.

This position paper argues that it is now appropriate to more seriously consider the nature of systems that are capable of learning, retaining and using knowledge over a life time. In accord with (Thrun 1997), we call these *lifelong machine learning*, or LML systems. We advocate that a systems approach is needed, taken in the context of an agent that is able to acquire knowledge through learning, retain or consolidate such knowledge, and use it for inductive transfer when learning new tasks.

We argue that LML is a logical next step in machine learning research. The development and use of inductive bias is essential to learning. There are a number of theoretical advances in AI that will be found at the point where machine learning meets knowledge representation. There are numerous practical applications of LML in areas such as

web agents and robotics. And our computing and communication systems now have the capacity to implement and test LML systems.

This paper reviews prior work on LML that uses supervised, unsupervised or reinforcement learning methods. This work has gone by names such as constructive induction, incremental and continual learning, explanation-based learning, sequential task learning, never ending learning, and most recently learning with deep architectures. We then present our position on the move beyond learning algorithms to LML systems, detail the reasons for our position and discuss potential arguments and counter-arguments. We then take some initial steps to advance LML research by proposing a definition of LML and a reference framework for LML that considers all forms of machine learning. We complete the paper by listing several key challenges for and benefits from LML research and conclude with two ideas for advancing the field.

## Prior Work on LML

Their exists prior research in supervised, unsupervised and reinforcement learning that consider systems that learn domains of tasks over extended periods of time. In particular, progress has been made in machine learning systems that exhibit aspects of knowledge retention and inductive transfer.

### Supervised Learning

In the mid 1980s Michalski introduced the theory of *constructive inductive learning* to cope with learning problems in which the original representation space is inadequate for the problem at hand (Michalski 1993). New knowledge is hypothesized through two interrelated searches: (1) a search for the best representational space for hypotheses and (2) a search for the best hypothesis within the current representational space. The underlying principle is that new knowledge is easier to induce if search is done using the *right* representation.

In 1989 Solomonof began work on *incremental learning* (Solomonoff 1989). His system was primed on a small, incomplete set of primitive concepts, that are able to express the solutions to the first set of simple problems. When the machine learns to use these concepts effectively it is given more difficult problems and, if necessary, additional primitive concepts needed to solve them, and so on.

In the mid 1990s, Thrun and Mitchell worked on a lifelong learning approached they called *explanation-based neural networks* (Thrun 1996). EBNN is able to transfers knowledge across multiple learning tasks. When faced with a new learning task, EBNN exploits domain knowledge of previous learning tasks (back-propagation gradients of prior learned tasks) to guide the generalization of the new one. As a result, EBNN generalizes more accurately from less data than comparable methods. Thrun and Mitchell apply EBNN transfer to autonomous robot learning when a multitude of control learning tasks are encountered over an extended period of time (Thrun and Mitchell 1995).

Since 1995, Silver *et al.* have proposed variants of *sequential learning and consolidation systems* using standard back-propagation neural networks (Silver and Poirier 2004; Silver, Poirier, and Currie 2008). A system of two multiple task learning networks is used; one for short-term learning using task rehearsal to selectively transfer prior knowledge, and a second for long-term consolidation using task rehearsal to overcome the stability-plasticity problem. *Task rehearsal* is an essential part of this system. After a task has been successfully learned, its hypothesis representation is saved. The saved hypothesis can be used to generate virtual training examples so as to rehearse the prior task when learning a new task. Knowledge is transferred to the new task through the rehearsal of previously learned tasks within the shared representation of the neural network. Similarly, the knowledge of a new task can be consolidated into a large domain knowledge network without loss of existing task knowledge by using task rehearsal to maintain the function accuracy of the prior tasks while the representation is modified to accommodate the new task.

Rivest and Schultz proposed *knowledge-based cascade-correlation* neural networks in the late 1990s (Shultz and Rivest 2001). The method extends the original cascade-correlation approach, by selecting previously learned sub-networks as well as simple hidden units. In this way the system is able to use past learning to bias new learning.

## Unsupervised Learning

To overcome the stability-plasticity problem of forgetting previous learned data clusters (concepts) Carpenter and Grossberg proposed ART (Adaptive Resonance Theory) neural networks (Grossberg 1987). Unsupervised ART networks learn a mapping between "bottom-up" input sensory nodes and "top-down" expectation nodes (or cluster nodes). The vector of new sensory data is compared with the vector of weights associated with one of the existing expectation nodes. If the difference does not exceed a set threshold, called the "vigilance parameter", the new example will be considered a member of the most similar expectation node. If the vigilance parameter is exceeded than a new expectation node is used and thus a new cluster is formed.

In (Strehl and Ghosh 2003), Trehl and Ghosh present a cluster ensemble framework to reuse previous partitionings of a set objects without accessing the original features. By using the cluster label but not the original features, the pre-existing knowledge can be reused to either create a single consolidated cluster or generate a new partitioning of the objects.

Raina et al. proposed the *Self-taught Learning* method to build high-level features using unlabeled data for a set of tasks (Raina et al. 2007). The authors used the features to form a succinct input representation for future tasks and achieve promising experimental results in several real applications such as image classification, song genre classification and webpage classification.

Carlson et al. (Carlson et al. 2010) describe the design and partial implementation of a *never-ending language learner*, or NELL, that each day must (1) extract, or read, information from the web to populate a growing structured knowledge base, and (2) learn to perform this task better than on the previous day. The system uses a semi-supervised multiple task learning approach in which a large number (531) of different semantic functions are trained together in order to improve learning accuracy.

Recent research into the learning of *deep architectures* of neural networks can be connected to LML (Bengio 2009). Layered neural networks of unsupervised Restricted Boltzman Machine and auto-encoders have been shown to efficiently develop hierarchies of features that capture regularities in their respective inputs. When used to learn a variety of class categories, these networks develop layers of common features similar to that seen in the visual cortex of humans.

Recently, Le et al. used the deep learning method to build high-level features for large-scale applications by scaling up the dataset, the model and the computational resources (Le et al. 2012). By using millions of high resolution images and very large neural networks, their system effectively discover high-level concepts like a cat's face and a human body. Experimental results on image classification show that their network can use its learned features to achieve a significant improvement in classification performance over state-of-the-art methods.

## Reinforcement Learning

Several reinforcement learning researchers have considered LML systems. In 1997, Ring proposed a lifelong learning approach called *continual learning* that builds more complicated skills on top of those already developed both incrementally and hierarchically (Ring 1997). The system can efficiently solve reinforcement-learning tasks and can then transfer its skills to related but more complicated tasks.

Tanaka and Yamamura proposed a lifelong reinforcement learning method for autonomous-robots by treating multiple environments as multiple-tasks (Tanaka and Yamamura 1999). Parr and Russell used prior knowledge to reduce the hypothesis space for reinforcement learning when the polices considered by the learning process are constrained by hierarchies (Parr and Russell 1997).

In (Sutton, Koop, and Silver 2007), Sutton et al. suggests that learning should continue during an agent's operations since the environment may change making prior learning insufficient. In their work, an agent is proposed to adapt to different local environments when encountering different parts of its world over an extended period of time. The experimental results suggest continual tracking of a solution can

achieve a better performance than learning a solution from only prior learning.

## Moving Beyond Learning Algorithms

Our position is that it is now appropriate for the AI community to seriously tackle the LML problem, moving beyond the development of learning algorithms and onto systems that learn, retain and use knowledge over a lifetime. The following presents our reasons for a call for wider research on LML systems.

### Inductive Bias is Essential to Learning

The constraint on a learning system's hypothesis space, beyond the criterion of consistency with the training examples, is called *inductive bias* (Mitchell 1980). Utgoff and Mitchell wrote in 1983 about the importance of inductive bias to concept learning from practical sets of training examples (Utgoff 1983). They theorized that learning systems should conduct their own search for an appropriate inductive bias using knowledge from related tasks of same domain. They proposed a system that could shift its bias by adjusting the operations of the modeling language. Since that time, the AI community has come to accept the futility of searching for a universal machine learning algorithm (Wolpert 1996). Our proposal to consider systems that retain and use prior knowledge as a source of inductive bias promotes this perspective.

### Theoretical Advances in AI: ML meets KR

In (Thrun 1997), Thrun writes "The acquisition, representation and transfer of domain knowledge are the key scientific concerns that arise in lifelong learning." We believe that knowledge representation will play an important a role in the development of LML systems. More specifically, the interaction between knowledge retention and knowledge transfer will be key to the design of LML agents. Lifelong learning research has the potential to make serious advances on a significant AI problem - the learning of *common background knowledge* that can be used for future learning, reasoning and planning. The work at Carnegie Mellon University on NELL is an early example of such research (Carlson et al. 2010).

### Practical Agents/Robots Require LML

Advances in autonomous robotics and intelligent agents that run on the web or in mobile devices present opportunities for employing LML systems. Robots such as those that go into space or travel under the sea must learn to recognize objects and make decisions over extended periods of time and varied environmental circumstances. The ability to retain and use learned knowledge is very attractive to the researchers designing these systems. Similarly, software agents on the web or in our mobile phones would benefit from the ability to learn more quickly and more accurately as they are challenged to learn new but related tasks from small numbers of examples.

### Increasing Capacity of Computers

Advances in modern computers provide the computational power for implementing and testing LML systems. The number of transistors that can be placed cheaply on an integrated circuit has doubled approximately every two years since 1970. This trend is expected to continue into the foreseeable future, with some expecting the power of computing systems to move to a log scale as computing systems increasingly use multiple processing cores. We are now at a point where an LML system focused on a constrained domain of tasks (*e.g.* product recommendation) is computationally tractable in terms of both computer memory and processing time. As an example, Google Inc. recently used 1000 computers, each with 16 cores, to train very large neural networks to discover high-level features from unlabeled data (Le et al. 2012).

## Counter Arguments

There are arguments that could be made against greater investment in LML research. Here we present two arguments and make an effort to counter them.

First, some could argue that machine learning should focus on the fundamental computation truths of learning and not become distracted by systems that employ learning theory. The idea is to stick to the field of study and leave the engineering of systems to others. Our response to this argument is that the retention of learned knowledge and its transfer would seem to be important constraints for the design of any learning agent; constraints that would narrow the choice of machine learning methods. Furthermore, it may directly inform the choice of representation used by machine learning algorithms.

A second potential argument is that LML is too wide an area of investigation with significant cost in terms of empirical studies. We agree that this has been a deterrent for many researchers. Undertaking repeated studies, where the system is tested on learning sequences of tasks, increases the empirical effort by an order of magnitude. However, because it is hard, does not make it impossible, nor does it decrease its relevance to the advance of AI. In recent years, there has been a growing appeal for a return to solving Big AI problems; LML is a step in this direction. As more researchers become involved, shared software and hardware tools, methodologies and best practises will begin to offset the increase in experimental complexity.

The next four sections make contributions toward advancing the field of LML by proposing a definition of Lifelong Machine Learning, presenting essential ingredients of LML, developing a general reference framework, and outlining a number of key challenges and benefits to LML research.

## Definition of Lifelong Machine Learning

**Definition:** *Lifelong Machine Learning, or LML, considers systems that can learn many tasks over a lifetime from one or more domains. They efficiently and effectively retain the knowledge they have learned and use that knowledge to more efficiently and effectively learn new tasks.*

## Effective and Efficient Retention

An LML system should resist the introduction and accumulation of erroneous knowledge. Only hypotheses with an acceptable level of generalization accuracy should be retained in long-term memory else it may take some time to be corrected. Similarly, the process of retaining a new hypothesis should not reduced its accuracy or that of prior hypotheses existing in long-term memory. In fact, the integration or consolidation of new task knowledge should increase the accuracy of related prior knowledge.

An LML system should provide an efficient method of retaining knowledge both in terms of time and space. The system must make use of its finite memory resources such that the duplication of information is minimized, if not eliminated. An LML system should also be computationally efficient when storing learned knowledge in long-term memory. Ideally, retention should occur online, however, in order to ensure efficient (consolidated) and effective retention (minimal error) this may not be possible.

## Effective and Efficient Learning

An LML system should produce a hypothesis for a new task that meets or exceeds the generalization performance of a hypothesis developed strictly from the training examples. Preferably, the transfer of prior knowledge from long-term memory should never develop less accurate models for a new task. An LML system should be able to select the most related prior knowledge to favourably bias the learning of a new task.

The use for prior knowledge by an LML system should not increase the computational time for developing a hypothesis for a new task as compared to using only the available training examples. Preferably, knowledge transfer within an LML system should reduce training time.

## Essential Ingredients for LML

Prior work suggests the following are essential elements for an LML agent: (1) the *retention* (or consolidation) of learned task knowledge; (2) the selective *transfer* of prior knowledge when learning new tasks; and (3) a *systems* approach that ensures the effective and efficient interaction of the retention and transfer elements.

Knowledge retention looks at LML from the knowledge representation perspective. Learned knowledge can be stored in various forms. The simplest method of retaining task knowledge is in functional form such as the training examples (Silver and Mercer 1996). An advantage of functional knowledge is the accuracy and purity of the knowledge (effective retention). A disadvantage of functional knowledge is the large amount (inefficient use) of storage space that it requires. Alternatively, the representation of an accurate hypothesis developed from the training examples can be retained. The advantages of representational knowledge are its compact size relative to the space required for the original training examples and its ability to generalize beyond those examples (efficient and effective retention).

Knowledge transfer looks at LML from the perspective of machine learning. Representational transfer involves the direct or indirect assignment of known task representation to the model of a new target task (Silver and Mercer 1996). In this way the learning system is initialized in favour of a particular region of hypothesis space of the modeling system (Ring 1993; Shavlik and Dietterich 1990; Singh 1992). Representational transfer often results in substantially reduced (efficient) training time with no loss in the generalization performance of the resulting hypotheses. In contrast to representational transfer, functional transfer employs the use of implicit pressures from training examples of related tasks (Abu-Mostafa 1995), the parallel learning of related tasks constrained to use a common internal representation (Baxter 1995; Caruana 1997), or the use of historical training information from related tasks (Thrun 1997; Naik and Mammone 1993). These pressures reduce the effective hypothesis space in which the learning system performs its search. This form of transfer has its greatest value in terms of developing more accurate (effective) hypotheses.

The systems approach emphasizes the interaction between knowledge retention and transfer learning and that LML is not just a new learning algorithm. It may benefit from a new learning algorithm or modifications to an existing algorithm, but it also involves the retention and organization of knowledge. We feel there is much to be learned in this regard from the writings of early cognitive scientists, artificial intelligence researchers and neuroscientists such as Albus, Holland, Newel, Langly, Johnson-Laird and Minsky. To emphasize this, consider that the form in which task knowledge is retained can be separated from the form in which it is transferred. For example, the retained hypothesis representation for a learned task can be used to generate functional knowledge in the form of training examples (Robins 1995; Silver and Mercer 2002). These training examples can then be used as supplementary examples that transfer knowledge when learning a related task.

## Framework for Lifelong Machine Learning

Figure 1 provides a general reference framework for an LML system that uses universal knowledge and knowledge of the task domain as a source of inductive bias (Silver and Mercer 2002; Yang et al. 2009). The framework is meant to encompass supervised, unsupervised, and reinforcement learning, and combinations thereof.

Since the inductive bias of related domains can be similar, LML systems should leverage both *universal knowledge* from across related domains and *domain knowledge* from the current task domain as a source of inductive bias. Inductive bias can include sets of auxiliary features extracted from prior tasks. For example, Raina et al. (Raina et al. 2007) and Le et al. (Le et al. 2012) proposed to build high-level features using unlabeled data for a set of learning tasks. The constructed features are then used to generate a succinct input representation for future tasks.

As with a standard inductive learner, training examples (supervised, unsupervised, or reinforcement) are used to develop a hypothesis (or policy in the case of reinforcement learning) for a task. However, unlike a standard learning system, knowledge from each hypothesis is saved in long-term
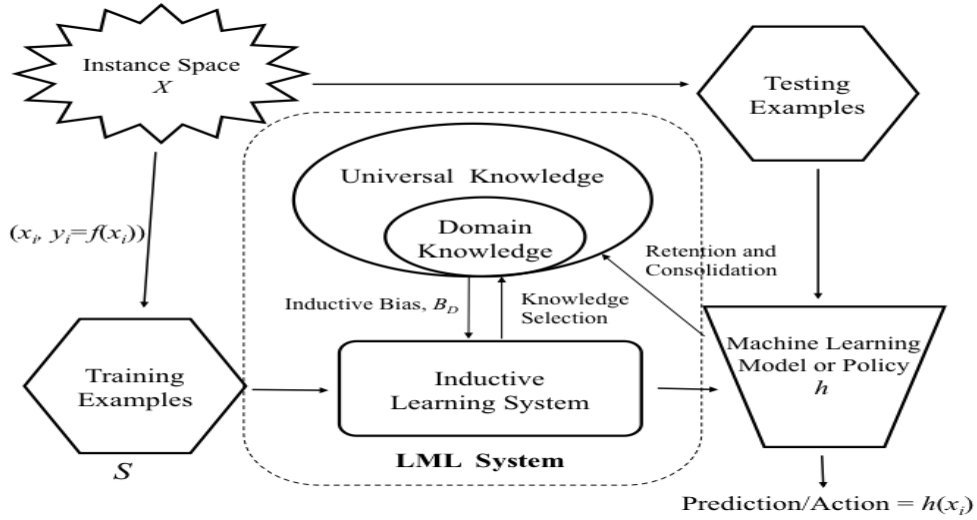
Figure 1: A framework for lifelong machine learning.

memory structure containing *universal and domain knowledge*. When learning a new task, aspects of this prior knowledge are selected to provide a beneficial inductive bias to the learning system. The result is a more accurate hypothesis or policy developed in a shorter period of time. The method relies on the transfer of knowledge from one or more prior secondary tasks, stored in universal or domain knowledge, to the hypothesis for a new primary task. The problem of selecting an appropriate bias becomes one of selecting the most related knowledge for transfer.

Formally, given a learning algorithm $L$, and an inductive bias $B_D$ given by universal and domain knowledge, the problem becomes one of finding a hypothesis or policy $h$, based on a set of examples $S$ of the form $(x_i)$ for unsupervised learning or $(x_i, y_i)$ from an input space space $X$ to an output / action space $Y$ for supervised and reinforcement learning, such that: $L \wedge B_D \wedge S \succ h$ where, for supervised and reinforcement learning, $h(x_i) = y_i$ for all $(x_i, y_i)$ in $X$ and $Y$. The relation is not one of entailment because it is possible that $B_D$ forms only a portion of all assumptions required to logically deduce $h$ given $S$.

## Challenges and Benefits

There are significant challenges and potential benefits for AI and brain sciences from the exploration of LML. The following captures several of these challenges and benefits.

### Type of Machine Learning

An earlier section presented prior work on LML that uses unsupervised, supervised, and reinforcement learning. An open question is which approach to learning or combination of approaches are best for LML. For example, recent work has shown the benefit of unsupervised training using many unlabelled examples as a source of inductive bias for supervised learning (Bengio 2009). The choice of learning type will dramatically affect the structure and function of an LML system. Combinations of approaches may be helpful for learning individual tasks but prove challenging for knowledge consolidation.

### Input / Output Type, Complexity and Cardinality

LML systems can vary based on the kinds of data it can work with: binary, real-valued, or vector. They can also vary based on their ability to deal with changing numbers of input attributes across tasks. For example certain tasks may not require the same or as many input attributes as another, particularly if they are from heterogenous task domains. This raises problems both for knowledge retention and transfer learning.

### Training Examples versus Prior Knowledge

A lifelong learning system must weigh the relevance and accuracy of retained knowledge along side the information resident in the available training examples for a new task. An estimate of the sample complexity of the new task will play a role here. Theories on how to select inductive bias and modify the representational space of hypotheses (Solomonoff 1989) will be of significant value to AI and brain science.

### Effective and Efficient Knowledge Retention

Mechanisms that can effectively and efficiently retain learned knowledge over time will suggest new approaches to common knowledge representation. In particular, methods of overcoming the stability-plasticity problem so as to integrate new knowledge into existing knowledge are of value to researchers in AI, cognitive science and neuroscience (Silver and Poirier 2004). Efficient long-term retention of learned knowledge should cause no loss of prior task knowledge, no loss of new task knowledge, and an increase in the accuracy of old tasks if the new task being retained is related. Furthermore, the knowledge representation approach should allow a lifelong learner to efficiently select the most effective prior

knowledge for inductive transfer during short-term learning. In general research in lifelong learning systems will see theories of transfer learning and knowledge representation enfluence and affect each other.

## Effective and Efficient Knowledge Transfer

The search for transfer learning methods that are both rapid (efficient) and develop accurate (effective) hypotheses is a challenging one. Transfer learning should produce a hypothesis for a new task that meets or exceeds the generalization performance of a hypothesis developed from only the training examples. There is evidence that the functional transfer somewhat surpasses that of representation transfer in its ability to produce more accurate hypotheses (Caruana 1997; Silver and Poirier 2004). Starting from a prior representation can limit the development of novel representation required by the hypothesis for a new task. Preferably, transfer learning should decrease the computational time and space for developing a hypothesis for the primary task as compared to using only the training examples. In practice, this reduction has been rarely observed because more computation is required to *index* into domain knowledge and memory requirements increase as aspects of prior knowledge are introduced. Research has shown that a representational form of knowledge transfer is often more efficient than a functional form but it rarely results in improved model effectiveness. (Silver and Poirier 2004).

## Scalability

Scalability is often the most difficult and important challenge for computer scientists. A lifelong learning system must be capable of scaling up to large numbers of inputs, outputs, training examples and learning tasks. Preferably, the space and time complexity of the learning system grows polynomially in all of these factors.

## Practicing a Task

A lifelong learning system should facilitate the practice of a task such that the generalization accuracy of the hypothesis for the task increases over time. But how can a lifelong learning system determine from the training examples that it is practicing a task it has previously learned versus learning a new but closely related task? Related work suggests that a system should not be explicit in this determination (Silver and Alisch 2005; Silver, Poirier, and Currie 2008); rather, the similarity of a set of training examples to that of prior domain knowledge should be implicit; each training example should be able to draw upon those aspects of domain knowledge that are most related. This suggests that domain knowledge should be seen as continuum as apposed to a set of disjoint tasks. A computation theory of how best to practice tasks will be useful to the fields of AI, psychology and education.

## Curriculum

The study of lifelong learning systems will provided insight into curriculum and training sequences that are beneficial for both humans and machines (Solomonoff 1989; Ring 1997).

This will be beneficial to robot and software agent training and will likely lead to the confirmation of and advances in human educational curriculum.

## Heterogenous Domains of Tasks

In cross-domain learning (Yang et al. 2009), by discovering the relation between the source and target domains, transfer learning methods use shared knowledge (i.e., common features, shared version spaces, etc.) to construct inductive bias for the task domain. In heterogeneous transfer learning, the key idea is to leverage the feature correspondence across heterogenous domains (i.e., bilingual dictionary, image and tags, music and Lyrics, etc.) to build an effective feature mapping for transferring knowledge. Although, much of initial LML research has focused on retention and transfer within a single domain of tasks, an important area of research will be LML systems that work across heterogenous domains.

## Acquisition and Use of Meta-knowledge

Most LML systems will need to collect and retain meta-knowledge of their task domains. For example, it may be necessary to estimate the probability distribution over the input space so as to manufacture appropriate functional examples from retained task representation (Silver and Mercer 2002). Alternatively, it may be necessary to retain characteristics of the learning process for each task (Thrun 1996).

## Applications in Agents and Robots

Software agents and robots can make good use of lifelong learning systems, or at least provide useful test platforms for empirical studies (Thrun 1996). Agents and robots will naturally need to learn new but related tasks. This will provide opportunities to try different methods of retaining and consolidating task knowledge. The agents constrained input and output domains provide an environment to test the impact of curriculum and the practice of tasks in a controlled manner. Lifelong learning can also be used to overcome the *cold-start* problem exhibited by personal agents that employ user modeling (Lashkari, Metral, and Maes 1994). LML can be used to boot-strap a new user model by transferring knowledge from the model of another, related user.

## Conclusion and Next Steps

This paper has provided an initial survey of prior work on LML systems and takes the position that it is time for the AI community to more seriously tackle the LML problem. We call for a move beyond the development of learning algorithms and onto systems that learn, retain and use knowledge over a lifetime. Our reasons for this include the importance of retaining and selecting inductive bias, the potential for theoretical advances in AI at the point where machine learning and knowledge representation meet, the practical applications of LML in areas such as web agents and robotics, and the computing capacity that is now available to researchers. We then proposed a definition of LML and a reference framework that considers all forms of machine

learning. Finally, the paper lists several key challenges for and potential benefits from LML research.

As next steps toward advancing the research and application of LML, we would like to suggest two action items. First that the AI community consider a grand challenge that will help further define the field of LML and raise the profile of research in this important area. LML has been used in application areas that vary from predicting heart disease at various hospitals (Silver, Poirier, and Currie 2008) to robot soccer (Kleiner et al. 2002). One or more exciting challenges can surely be developed. Second, that an open source project be established similar to the WEKA project that will allow researchers to share knowledge and collaborate on systems that retain, consolidate and transfer learned knowledge.

# References

Abu-Mostafa, Y. S. 1995. Hints. *Neural Computation* 7:639–671.

Baxter, J. 1995. Learning internal representations. *Proceedings of the Eighth International Conference on Computational Learning Theory*.

Bengio, Y. 2009. Learning deep architectures for ai. *Foundations and Trends in Machine Learning* 2(1):1–127.

Carlson, A.; Betteridge, J.; Kisiel, B.; Settles, B.; Jr., E. R. H.; and Mitchell, T. M. 2010. Toward an architecture for never-ending language learning. In Fox, M., and Poole, D., eds., *AAAI*. AAAI Press.

Caruana, R. A. 1997. Multitask learning. *Machine Learning* 28:41–75.

Grossberg, S. 1987. Competitive learning: From interactive activation to adaptive resonance. *Cognitive Science* 11(1):23–63.

Kleiner, E.; Dietl, M.; Nebel, B.; and Freiburg, U. 2002. Towards a life-long learning soccer agent. In *In Proc. Int. RoboCup Symposium 02*, 119–127.

Lashkari, Y.; Metral, M.; and Maes, P. 1994. Collaborative interface agents. In *In Proceedings of the Twelfth National Conference on Artificial Intelligence*, 444–449. AAAI Press.

Le, Q.; Ranzato, M.; Monga, R.; Devin, M.; Chen, K.; Corrado, G.; Dean, J.; and Ng, A. 2012. Building high-level features using large scale unsupervised learning. In *International Conference in Machine Learning*.

Michalski, R. 1993. Learning = inferencing + memorizing. *Foundations of Knowledge Acquistion: Machine Learning* 1–41.

Mitchell, T. M. 1980. The need for biases in learning generalizations. *Readings in Machine Learning* 184–191. ed. Jude W. Shavlik and Thomas G. Dietterich.

Naik, D., and Mammone, R. J. 1993. Learning by learning in neural networks. *Artificial Neural Networks for Speech and Vision*.

Parr, R., and Russell, S. 1997. Reinforcement learning with hierarchies of machines. In *Advances in Neural Information Processing Systems 10*, 1043–1049. MIT Press.

Raina, R.; Battle, A.; Lee, H.; Packer, B.; and Ng, A. Y. 2007. Self-taught learning: transfer learning from unlabeled data. In *Proceedings of the 24th international conference on Machine learning*, ICML '07, 759–766. New York, NY, USA: ACM.

Ring, M. 1993. Learning sequential tasks by incrementally adding higher orders. *Advances in Neural Information Processing Systems 5* 5:155–122. ed. C. L. Giles and S. J. Hanson and J.D. Cowan.

Ring, M. B. 1997. Child: A first step towards continual learning. In *Machine Learning*, 77–104.

Robins, A. V. 1995. Catastrophic forgetting, rehearsal, and pseudorehearsal. *Connection Science* 7:123–146.

Shavlik, J. W., and Dietterich, T. G. 1990. *Readings in Machine Learning*. San Mateo, CA: Morgan Kaufmann Publishers.

Shultz, T. R., and Rivest, F. 2001. Knowledge-based cascade-correlation: using knowledge to speed learning. *Connect. Sci.* 13(1):43–72.

Silver, D. L., and Alisch, R. 2005. A measure of relatedness for selecting consolidated task knowledge. *Proceedings of the 18th Florida Artificial Intelligence Research Society Conference (FLAIRS05)* 399–404.

Silver, D. L., and Mercer, R. E. 1996. The parallel transfer of task knowledge using dynamic learning rates based on a measure of relatedness. *Connection Science Special Issue: Transfer in Inductive Systems* 8(2):277–294.

Silver, D. L., and Mercer, R. E. 2002. The task rehearsal method of life-long learning: Overcoming impoverished data. *Advances in Artificial Intelligence, 15th Conference of the Canadian Society for Computational Studies of Intelligence (AI'2002)* 90–101.

Silver, D. L., and Poirier, R. 2004. Sequential consolidation of learned task knowledge. *Lecture Notes in AI, Canadian AI'2004* 217–232.

Silver, D. L.; Poirier, R.; and Currie, D. 2008. Inductive tranfser with context-sensitive neural networks. *Machine Learning* 73(3):313–336.

Singh, S. P. 1992. Transfer of learning by composing solutions for elemental sequential tasks. *Machine Learning*.

Solomonoff, R. J. 1989. A system for incremental learning based on algorithmic probability. In *Probability, Proceedings of the Sixth Israeli Conference on Artificial Intelligence, Computer Vision and Pattern Recognition*, 515–527.

Strehl, A., and Ghosh, J. 2003. Cluster ensembles — a knowledge reuse framework for combining multiple partitions. *J. Mach. Learn. Res.* 3:583–617.

Sutton, R. S.; Koop, A.; and Silver, D. 2007. On the role of tracking in stationary environments. In *Proceedings of the 24th international conference on Machine learning*, ICML '07, 871–878. New York, NY, USA: ACM.

Tanaka, F., and Yamamura, M. 1999. An approach to lifelong reinforcement learning through multiple environments. In *Proc. of the 6th European Workshop on Learning Robots (EWLR-6)*, 93–99.

Thrun, S., and Mitchell, T. M. 1995. Lifelong robot learning. *Robotics and Autonomous Systems* 15:25 – 46.

Thrun, S. 1996. *Explanation-Based Neural Network Learning: A Lifelong Learning Approach*. Boston, MA: Kluwer Academic Publishers.

Thrun, S. 1997. Lifelong learning algorithms. *Learning to Learn* 181–209.1 Kluwer Academic Publisher.

Utgoff, P. E. 1983. Adjusting bias in concept learning. In *Proceedings of IJCAI-1983*, 447–449.

Wolpert, D. H. 1996. The lack of a priori distinctions between learning algorithms. *Neural Comput.* 8(7):1341–1390.

Yang, Q.; Chen, Y.; Xue, G.-R.; Dai, W.; and Yu, Y. 2009. Heterogeneous transfer learning for image clustering via the social web. In *Proc. of the Joint Conf. of the 47th Annual Meeting of the ACL and the 4th Int. Joint Conf. on NLP of the AFNLP: Volume 1*, ACL '09, 1–9. Stroudsburg, PA, USA: Assoc. for Computational Linguistics.