# Forty Years of Machine Learning: Metaphors, Myths, and Challenges

**Pat Langley**

Institute for the Study of
Learning and Expertise

Center for Design Research
Stanford University

This research was supported by ONR Grant N00014-20-1-2643 and by AFOSR Grant FA9550-20-1-0130. Thanks to many colleagues for discussions that led to these ideas.

# Introductory Remarks

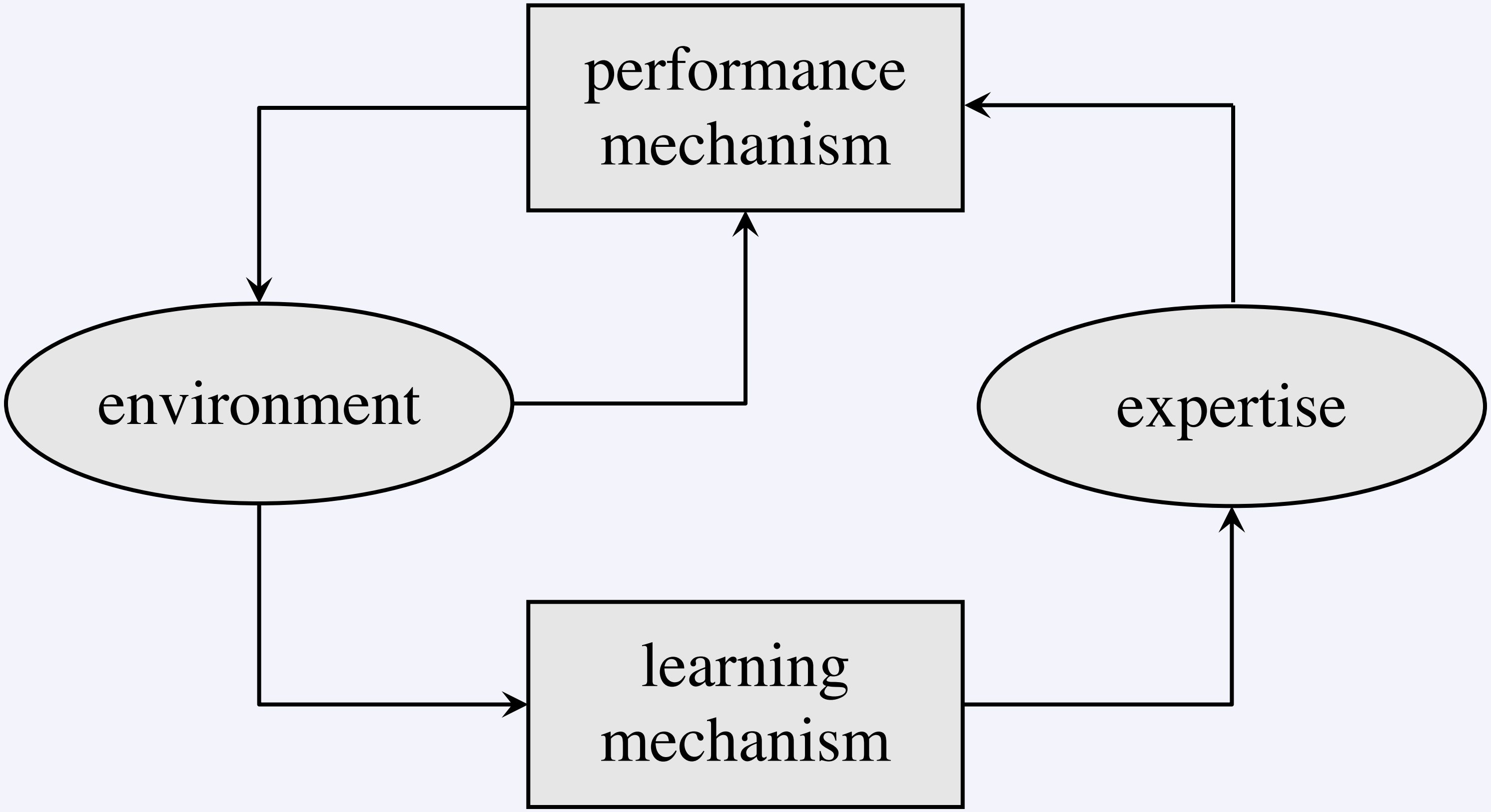
# What is Machine Learning?

Machine learning is what Simon (1976) has called a *science of the artificial*, in that it:

- Designs and constructs *artifacts*
- Examines the *behavior* of these entities
- Attempts to *understand* and *explain* this behavior
- Formulates *principles* to aid future design efforts

In this case, we are interested in *computational* artifacts that *improve* their *performance* based on *experience*.

# Elements of Learning Systems



performance mechanism

environment

expertise

learning mechanism

This diagram clarifies we can never describe learning in isolation. We must also specify the *environmental input*, the *representation of expertise*, and the *performance element* that uses it.

# Relation to Artificial Intelligence

Machine learning was originally launched as a subfield of AI that focused on learning.

- But after a decade, many researchers in the area had come to view it as a standalone discipline.
- And now many see machine learning as the only viable path to building intelligent systems.

These two perspectives ignore the need to study learning in the context of representation and performance.

Researchers who adopt them ignore concepts and tools that are needed to understand the mind fully.

# Metaphors for Machine Learning

Over the years, researchers have adopted different metaphors for machine learning, viewing it as:

- Automated knowledge acquisition
- Caching results of multi-step reasoning
- Search through a space of model structures
- Parameter estimation / optimization

These are not right, wrong, or mutually exclusive, but they offer different perspectives on learning.

Unfortunately, the last two metaphors have come to dominate the field at the expense of others.

# Myths About Machine Learning

# Expertise is Compiled Experience

A widespread but seldom stated view is that learning involves the *compilation* of experience into expertise.

- Clearly, learning revolves around the acquisition of expertise from experience.
- But this does *not* mean that this acquisition process is a form of compilation.

Some human expertise (e.g., motor skills) may be compiled, but much is *composed* at performance time.

Thus, idea that learned expertise *must* be compiled experience is no more than a popular myth.

# Learned Expertise is Opaque

Another common belief is that learning necessarily produces *opaque* expertise.

- This is linked to the notion of compilation, which produces uninterpretable binary code.
- Compiled code is so common in computer science, so many assume that learning should produce it.

But humans often learn structures (e.g., concepts, constraints) that can conveyed easily to others.

In other words, the popular view that learned expertise *must* be opaque is a second myth.

# Learning is a Batch Process

A third, somewhat less common, belief is that learning relies on off-line, *batch* processing.

- Thus, it requires that all of the training data be available at the outset of learning.
- When this situation holds, it means that statistics inherent in the data can drive learning.

However, human learning is on-line and incremental, and yet somehow it is remarkably effective.

This is another false assumption that has reduced exploration through the space of learning methods.

# Learning is Guided Entirely by Data

A fourth assumption is that data is the *only* information that can (or should) guide learning.

- This was partially a backlash against the early expert systems movement, which added knowledge manually.
- Another factor was rhetoric from the data mining paradigm, which emphasized the value of data.

But people *never* learn from scratch; they always acquire new knowledge in the context of existing structures.

This is another myth that has hindered research in machine learning for decades.

# Effective Learning Needs ‘Big Data’

A final widespread belief is that learning depends on *massive training sets* (and powerful computers to process them).

- This assumption is tied closely to the idea that learning must be driven by data alone.
- But it also comes from a *data fetish* promoted by companies whose business models benefit from it.

However, there are many cases where learning requires little data and where processing is inexpensive.

This final myth has warped the tasks and methods considered by machine learning researchers.

# History of Machine Learning
# (How We Got Here)

# AI and Pattern Recognition

Early research on AI and pattern recognition were linked, but by 1970 the two movements had diverged.

- Pattern recognition focused on *perceptual* tasks like classifying objects in images and recognizing words in speech.
  - Most work used *numeric* encodings and relied on *statistical learning* to induce classifiers.
- AI emphasized higher *cognitive* tasks like reasoning, planning, and language understanding.
  - Most work adopted symbolic notations and learning played only a minor role.

This separation of objectives, representations, and approaches continued until recently.

# AI’s Early Dislike of Learning

Early AI research was strongly linked to cognitive psychology, with many ideas developed jointly.

- Two of AI’s founders, Allen Newell and Herbert Simon, saw themselves as computational psychologists.
- Behaviorism, which focused on learning for sensorimotor tasks, had dominated American psychology for decades.
- In response, the cognitive revolution of the 1950s emphasized high-level human behavior and denigrated learning.

Some work on learning continued in AI and psychology, but it played a minor role in both fields.

# AI's Rediscovery of Learning

By the late 1970s, some had realized theories of intelligence were incomplete without accounts of learning.

- Research on concept attainment, grammar acquisition, and strategy learning began to explore the topic.
- This interest led to a series of workshops on learning in 1980, 1983, 1985, and 1987, with associated edited books.
- The initial organizers – Ryszard Michalski, Jaime Carbonell, and Tom Mitchell – dubbed the subfield *machine learning*.

The movement was small and often viewed as a fringe element of the AI community, but it was now on the map.

# Motivations for a New Subfield

There were multiple reasons to foster a subfield of machine learning in the 1980s, including:

- Automating the arduous construction of expert systems
- Understanding the processes that underlie human learning
- Lack of respect in AI community for work on learning
- Young researchers who were eager to have an impact

Members of the new community often had different aims, but there was also substantial overlap in interests.

# Launching a New Journal

Of course, the new subfield needed a place to publish its results and communicate progress:

- The journal *Artificial Intelligence* and meetings like *IJCAI* and *AAAI* were not very supportive of learning.
- The obvious solution was to establish a new journal, which we decided to call *Machine Learning*.
- I volunteered to serve as the Executive Editor and Michalski, Carbonell, and Mitchell signed on as Editors.

The first volume of the journal, published by Kluwer, appeared in 1986, without any great fanfare (except the *color*).

# Systems vs. Algorithms

Early work in machine learning, and AI generally, involved complex (named) *systems* with distinct components.

As the subfield progressed, researchers shifted their attention to *algorithms* that:

- Focused on narrow, well-defined aspects of learning
- Could be specified clearly in succinct pseudocode
- Were modeled after formal work in computer science

Although machine learning remained mainly empirical, this change coincided with interest in formal analyses.

# Experimental Evaluation

Machine learning began as an empirical discipline, with most researchers building running programs.

- Early studies demonstrated them on tens of examples and then reported the learned structures (e.g., rules).
- Gradually, the community recognized that it was important to measure how learning affected performance.
- Despite resistance, by the early 1990s, experimental studies with such measures became the default.

The reviewing process in journals and conferences aided this shift, but ‘public comments’ were also important.

Initial studies showed that learning improved performance on some task, *not* how it compared to other methods.

# The UCI Repository

Around 1987, David Aha at UCI began to collect data sets to support empirical studies of machine learning.

- This emphasized supervised induction for classification, the most common problem under study.
- The initial collection was small and the data sets had many issues, but its effect was profound.
- Not only could researchers now test their systems on many domains; they could *compare* them to other methods.

The UCI repository helped transformed machine learning into an empirical discipline, but it also had negative effects.

# Early Comparative Studies

AI and machine learning have always suffered from polemical stances of questionable scientific value.

Rhetoric in the 1980s stated that ‘symbolic’ and ‘connectionist’ methods were suited for entirely different settings, but:

- The UCI repository made it possible to test this assumption.
- Multiple teams (e.g., Mooney et al., 1989) compared decision-tree induction and neural networks empirically.
- Results showed that neither was always better, but, even more important, *they could be applied to the same tasks*.

Some had difficulty accepting these results, but people soon realized the different frameworks were comparable.

# Redefining the Field

The insights about experimental evaluation led naturally to a redefinition of the field:

- *Machine learning is the study of computational methods for improving performance based on experience.*

The field had initially focused on learning rules, decision trees, grammars, and other symbolic structures.

- The original call for submissions to *Machine Learning* made this emphasis very explicit.
- But the new definition implied that we should include methods from pattern recognition and *even* neural networks.

There was some resistance to this idea, but the logic seemed irrefutable and gradually won out.

# Broadening the Community

In response, the field's leaders undertook efforts to broaden the community (e.g., special issues).

By the early 1990s, the journal and the new refereed conference (ICML) published papers on learning:

- *Rules, decision trees, logical formulae*
- *Case libraries / nearest neighbors*
- *Multilayer neural networks*
- *Probabilistic models*
- *Hybrid frameworks* (e.g., multivariate trees, Cobweb)

Each researcher had a favored approach, but there was mutual respect and the greater diversity was a healthy development.

# Insights About Simplicity

Research on machine learning initially emphasized complex techniques (e.g., Michalski’s AQ systems).

But experiments with the UCI repository revealed surprising results for very simple methods, including:

- *Classic perceptrons*
- *Decision stumps*
- *Naive Bayesian classifiers*

These results produced genuine insights about the nature of learning but encouraged focus on algorithms over systems.

# The Bias-Variance Tradeoff

One explanation for the effectiveness of simple methods was the bias-variance tradeoff:

- Some errors in learned models come from inherent *bias*
    - E.g., the limited representational power of naive Bayes
- Other errors come from *variance* in the learning method
    - Slight variations in training data let to different results

There is a tradeoff between these two sources of error, with some methods favoring one and some the other.

Simple techniques like decision-stump induction and naive Bayes have low variance but high bias.

# Successful Applications

The subfield’s connection to expert systems led to a healthy interest in applications of machine learning.

An ICML-93 workshop showed it had produced – often from small data sets – *deployed* systems for tasks like:

- *Recommending decisions for credit card requests (UK)*
- *Diagnosing electric motor pumps (Italy)*
- *Classifying sky objects in telescopic surveys (USA)*
- *Monitoring quality of rolling emulsions (Slovenia)*
- *Reducing banding in printing presses (USA)*
- *Predicting recurrence of breast cancer (Sweden)*

Langley and Simon (1995) reviewed these applications and factors that led to success (Hint: *not* the induction methods).

# The Data-Mining Movement

The late 1980s saw a growing interest in induction from data sets being collected in commerce.

The led to the data-mining movement, which used ideas from machine learning but:

- *Emphasized efficient processing of large training sets*
- *Encouraged the generation of interpretable models*
- *Incorporated methods from databases and statistics*

The first KDD conference took place in 1995 and a related journal launched shortly afterward.

The business community had a major presence in data mining, while machine learning remained largely academic.

# Machine Learning and the Web

Early ML applications faced challenges obtaining data and delivering results to customers.

The advent of the World Wide Web made data far easier to collect and services much easier to deliver:

- Search engines, on-line shopping sites, and other offerings changed the landscape entirely.
- Recommender systems in particular had data collection and machine learning built into their designs.

These changes fostered work on learning from large databases and discouraged studies of data-efficient methods.

# Increased Focus on Statistics

By 2005, machine learning had become highly enamored of statistical techniques, including:

- *Bayesian estimation*
- *Kernel methods*
- *Ensemble learning (e.g., random forests)*
- *Constrained optimization*
- *Statistical relational learning*

This trend dated back to the mid-1990s, but now it became the field's dominant theme.

Approaches to learning that did not fit this schema received almost no attention from the community.

# An Obsession with Metrics

Machine learning applications had become widespread, but academics needed their own criteria for progress.

Unfortunately, many researchers adopted unenlightened uses of empirical studies by:

- Treating a small group of data sets as ‘benchmarks’
- Adopting obfuscating metrics like AUC and F1
- Pursuing mindless ‘bake offs’ among techniques
- Encouraging competitions with prizes to winners

These ignored the purpose of experiments: *to gain scientific insights and to understand strengths and weaknesses*.

# Deep Learning: Pros

The past decade has seen great interest in so-called ‘deep learning’ approaches.

In addition to well-known performance gains, unexpected benefits have included:

- An increased emphasis on systems over algorithms
- A growing realization that we are not ‘drowning in data’
- No more convergence proofs for reinforcement learning

Overall, these changes to the field’s perspectives have been healthy developments.

# Deep Learning: Cons

Unfortunately, the excitement about deep learning has also strengthened beliefs in common myths:

- Expertise is opaque, compiled experience
- Learning is a batch process guided only by data
- Effective learning requires massive data sets

Equally devastating – like an invasive plant species – it has nearly wiped out other approaches.

This headlong pursuit has drastically reduced the intellectual diversity in both research and education.

# Challenges for Machine Learning

# Human-Like Learning

Fortunately, there are viable alternatives to the dominant view.

One option is to mimic *human learning*, which provides strong constraints on computational systems.

- These constraints can serve as steps in a *computational gauntlet* that machine learners must traverse.
- Many of these features are documented in the psychological literature, but others are blindingly obvious.

They suggest a very different way to build learning systems that has much in common with the field’s original vision.

# Acquire Modular Structures

One basic feature of human learning (Bower, 1981) concerns the nature of acquired content:

- ***Learning involves the acquisition of modular cognitive structures.***

This does not specify details about these structures; only that expertise consists of discrete mental elements.

Candidates include concepts, production rules, exemplars, chunks, and even stimulus-response pairs.

But each contrasts sharply with the idea that learning produces a single compiled structure.

# Learn Composable Structures

A second characteristic is enabled by the first one and often associated with it closely:

- ***Learned cognitive structures can be composed during performance.***

That is, relevant elements of expertise are accessed and then combined as needed to produce behavior.

Generative grammars (Chomsky, 1965) and classic rule-based systems offer examples of composable structures.

These differ from the large structures (e.g., neural networks or decision trees) produced by most statistical methods.

# Learn in a Piecemeal Manner

Another feature involves how people process experiences and create structures. In particular:

- ***Expertise is acquired in a piecemeal manner, with one element added at a time.***

Humans learn one structure and then another, continuing until they achieve broad coverage.

They do not learn complex models en masse, as done by most methods for statistical induction.

This does not mean they never revisit elements created earlier, but each structure is learned reasonably independently.

# Acquire Expertise Incrementally

Another processing constraint focuses not on the knowledge elements but on the training cases:

- ***Learning is an incremental activity that processes one experience at a time.***

This is related to on-line learning, but also requires processing these stimuli only once, or at least rarely.

Incremental processing is often associated with piecemeal learning, but they are orthogonal issues.

Bottom-up induction of context-free grammars (Wolff, 1980) is piecemeal but batch; naive Bayes is the opposite.

# Guide Learning with Knowledge

This dependence on previous experience leads to a broader statement about the mechanisms at work:

- ***Learning is guided by knowledge that aids interpretation of new experiences.***

Because acquisition is piecemeal and incremental, it takes place in the presence of structures added earlier.

The influence takes different forms depending on the types of structures created (taxonomies, composites).

Knowledge-guided learning receives little attention in modern data-intensive paradigms.

# Learn from Few Experiences

A final feature of human learning, enabled by both incremental and knowledge-guided processing, is that:

- ***Cognitive structures are acquired and refined rapidly, from small numbers of training cases.***

The claim is not that we acquire all expertise in a domain from a few instances, but that we learn modular elements this way.

This relates to the idea of *learning rate* in psychology; human learning curves are often very rapid.

Again, this differs from statistical induction’s emphasis on learning from thousands or millions of items.

# Summary Remarks

We have seen that machine learning, despite many advances, has had its path warped by myths that:

- *Expertise is opaque, compiled experience* and *learning is a batch process guided only by massive data*.

But a viable alternative is to develop systems that learn in far more human-like ways by:

- *Acquiring modular, composable structures in a piecemeal, incremental way, aided by knowledge, from little data*.

Early work on machine learning built on these ideas, leading to both theoretical insights and application successes.

***I challenge other researchers to take the road less traveled.***