Access Innovations, Inc.

Innovative • Effective • Customizable

EXCLUSIVE CHECKLIST REPORT

Succeeding With Machine Learning

Machine Learning Trends

By **Marydee Ojala,** Conference Program Director, Information Today, Inc.

Reading what Access Innovations, Inc.'s Heather Kotula had to say about Machine Learning (ML) took me back to when my children were little. It was the sudden silence that would get a parent's attention. One minute you're basking in the novelty of having some quiet time. The next you're suddenly aware that it's way *too* quiet and you jump out of your chair to see what the kids are up to. It particularly gets your attention when, just as you realize the silence has stretched out longer than it should have, you hear a loud crash followed by crying. In my case, it was usually nothing too dreadful, but I do recall the time a friend's 8-year-old son decided to climb the bookshelf in his room. Yup, it fell over on him. He was scared but not hurt. That need to check up on the kids becomes a knee-jerk reaction when prolonged silence occurs. The cliché of the calm before the storm comes to mind.

For those of you without children, you might have the same reaction in regard to your fur babies. Your dogs and cats can equally create a bit of mayhem if you're not paying attention. My dog once, apparently, decided she didn't like my new detergent and chewed a hole in my newly-laundered jeans. Or maybe she really did like it and thought it delicious. It's hard to know what was going on in her doggy brain. If your pets are fish, the worry about silence preceding disaster is moot. They're always quiet. Maybe I should get a fish.

NOT A GAME

When I read about ML or listen to presentations about ML, I'm fascinated by the number of explanations that begin by putting the technology into historical perspective. Authors and speakers like to trace the origins back to the 1950s and laud the advances by citing two major milestones—when IBM's Deep Blue beat Gary Kasparov at chess in 1997 and then when Google's DeepMind developed AlphaGo, which has gotten the best of several top human Go players 20 years later.

But it's important to note that ML is much more than games. ML has a multitude of applications with practical value. It can create metadata and enhance existing metadata for a wide range of purposes, develop controlled vocabularies, and do automatic indexing. It's not about winning at chess or Go, it's about real-world applications with important consequences. But as Kotula warns, do manage expectations, since Artificial Intelligence (AI) and ML are not the ultimate solution to every business problem you confront.

ADULT SUPERVISION NEEDED

Part of managing expectations involves the data sets a machine will learn from. It's helpful to think about training data sets for ML in the same way we think about supervising children. You can leave them by themselves for only so long, then the need for adult supervision kicks in. In ML, we have both supervised learning and unsupervised learning—with a middle ground that incorporates some of both. Access Innovations, Inc. has some reservations about completely unsupervised learning, recognizing the need to guard against obvious absurdities.

Unsupervised learning is what happens when the machine builds its own mathematical model based on whatever data inputs it receives. The hope is that the data sets will be sufficiently robust so that the machine will discover patterns valuable to its model creation. Think of it as my friend's son learning not to climb on the bookcase because the result is a painful fall. Learning from experience in order to gain insights from data, however, requires that the reactions to the experience be ones that accurately predict the future.

The importance of ML lies in its ability to see patterns that humans can't see. As the amount of available data escalates, the value of ML increases. Access Innovations, Inc.'s approach to ML, with its reliance on knowledge bases and an inference engine, exemplifies the ideal approach. Setting up a knowledge base requires an understanding of the nature of the included entities. It differentiates among people, places, and things, recognizing that someone with a surname of Park, for example, is a person, not a green space in the middle of a city with trees, shrubs, and walking paths. Henry Park is not Hyde Park; Cindy Park is not Central Park.

"ML has a multitude of applications with practical value. It can create metadata and enhance existing metadata for a wide range of purposes, develop controlled vocabularies, and do automatic indexing. It's not about winning at chess or Go, it's about real-world applications with important consequences."

It also knows about synonyms, so that it equates cat and feline, having been taught that these refer to the same animal, and it knows that a dog is not a feline. It goes beyond exact synonyms. Think of all the terms that could be applied to describe rain (drizzle, storm, showers, etc.), as well as the consequences of a severe rainstorm (flooding, accidents, overflowing river banks, etc.). ML can identify associated words and phrases relevant to the situation. Human intervention is only necessary to ensure that an irrelevant association is not created.

TRAINING THE DATA SETS

ML is not the science fiction notion of a robot taking your job or a cyborg out to kill you. Instead, it's adding to your human ability to make inferences from textual material at a grand scale. That's the basic underpinning of an inference engine. Avoiding errors, such as the faulty reasoning that leads to equating dogs and cats, relies on training the data set fed to the machine. The kernel of AI, as Kotula points out, is the imitation of human thinking rather than the replacement of it. It is obvious to a human being that a dog is not a cat, but the computer may very well not grasp the difference. One four-legged mammal is just like another four-legged mammal to the

"In the early days of ML, data sets were controlled. This sometimes led to biased learning. Historically, if a particular job had only been held by men, the data sets reflected that pattern, leading the machine to conclude that women should be excluded from the hiring pool. This is bias that can be corrected by human intervention. However, the bias needs to be made evident to the humans training the data set."

machine. What computers lack is the human concept of common sense. We can look at a dog and know it's not a cat. The computer can't do that.

It's good to remember, also, that unlike the sci-fi concept of AI, computers don't experience emotion. Our machines don't feel sad or happy, they simply reflect the patterns they've detected. Having a human-in-the-loop when it comes to implementing an ML project is essential to gaining reliable, replicable results. And there's no point in employing ML if you don't want reliable, replicable results. That's the object of the exercise, isn't it?

In the early days of ML, data sets were controlled. This sometimes led to biased learning. Historically, if a particular job had only been held by men, the data sets reflected that pattern, leading the machine to conclude that women should be excluded from the hiring pool. This is bias that can be corrected by human intervention. However, the bias needs to be made evident to the humans training the data set.

Reverting to Kotula's observation that toddlers learn by mimicking the humans around them, the analogy holds true for ML training sets. Just as children parrot their parents' behavior, good or bad, training sets can have built-in biases that skew the results created by the machine doing the learning. Feed a machine bad data and you will inevitably receive predictions that go wrong.

Recently, we have seen some spectacular failures in unsupervised learning. The main culprit is opening up the data sets to the general public rather than feeding the machine a smaller, controlled data set. It sets the scene for manipulation. The most striking example of ML gone wrong is Microsoft's Tay chatbot. Tay's Twitter persona invited anyone on Twitter to help it learn language. It learned—but it learned all the wrong things. In just one day, it was using racist, misogynistic, and anti-Semitic language, fed to it by internet trolls. To some, this proves the fallacy of relying on the "wisdom of crowds." These crowds didn't impart any wisdom to Tay, which was a catastrophic failure. Rather than fix it, Microsoft pulled the plug.

Tay received a lot of attention in ML circles. It's a perfect example of what Kotula means when she writes, "Left alone, the unpredictable results can expand exponentially, leading to outputs that are unreliable and not replicable." To return momentarily to the toddler analogy, what would happen if someone repeatedly told a child that the color blue was red? The child sees a blue ball but is told it's red. A blue book cover? Red. This type of miseducation is unlikely to happen in the real world, as other adults will correct the perception of blue as red and the child will learn the correct words for colors. But without adult supervision, as could happen in the ML world, the child/computer equates blue with red and might misidentify a Blue Jay as a Cardinal.

The key takeaway is the difference between intended and unintended bias. Even with smaller, vetted training sets, the possibility of bias remains. Amazon ran into trouble a few years ago when its attempt to analyze resumes to predict hiring decisions displayed gender bias. Google touted its ML abilities to predict flu but soon realized its predictions did not jibe with the reality of the flu statistics. Some bias could occur based on geography. A public school in the U.K. is very different from a public school in the U.S.

TRENDS IN MACHINE LEARNING

Coming into its own as one of the most important AI technologies, ML has moved well beyond its science fiction-based antecedents. As Kotula notes, science fiction may have "magnified our expectations," but the sometimes-magical thinking implicit in sci-fi story lines gets closer to reality all the time. The algorithms that drive ML projects have become ever more sophisticated and relevant to solving real-world problems.

ML has gained prominence in recent years. It holds great promise in many areas, particularly in science, technology, and health. Before jumping on the ML bandwagon, however, it's best to take Kotula's comment that data is messy as your mantra. Just as toddlers love to play in puddles, often getting a serious amount of mud on their clothing, data is rarely presented as absolutely pristine, needing only to be entered into a machine so that the ML system can accurately process it. No, it's much more likely, almost a dead certainty, that the data will require cleaning and organizing before it can be usable. Data that is incomplete, unstructured, lacking in governance, and generally of poor quality will not produce superior, or even reliable, results.

There is no reason to think that reliance on ML will decrease going forward. As Big Data becomes the norm, the volume of available data from which a machine can learn accelerates. This, in turn, provides many opportunities for successfully implementing ML. When looking for success in ML, it's important to recognize that it's not a black box. You don't just input data and expect miraculous results. Oversight and advance planning are essential. Concentrate on what you want to accomplish. You need human supervision for your knowledge base and your inference engine and you need to guard against unrealistic expectations. Then you can achieve success with your ML project. ▮

Marydee Ojala is conference program director for Information Today, Inc. She works on conferences such as Enterprise Search & Discovery, which is co-located with KMWorld, and WebSearch University, among others. She is a frequent speaker at U.S. and international information professional events. In addition, she moderates the popular KMWorld webinar series. Ojala is based in Indianapolis, Indiana and can be reached at marydee@infotoday.com.

Succeeding With Machine Learning

By **Heather Kotula,** VP, Marketing and Communications, Access Innovations, Inc.

Access Innovations, Inc. deployed its first client instance of its artificial intelligence (AI) system in 1995. While this isn't a Star Trek-esque "computer" that analyzes data and comes to independent conclusions, it does exactly what it is supposed to: make keyword suggestions based on inferences from the text of a document.

What do you think of when you hear "artificial intelligence" or "machine learning (ML)?" Science fiction is loaded with AIs that think—or attempt to think—like humans and have—or try to have—feelings. These are entertaining, sometimes wildly so, but they are still just fiction. Artificial intelligence doesn't have to have all the bells and whistles of robots that walk and talk and cavort on the Holodeck. At its kernel, what is artificial intelligence?

It starts with an expert system. Wikipedia defines an expert system as "a computer system that emulates the decision-making ability of a human expert."[1] Let me stress the word "emulate." To emulate is to "imitate with effort to equal or surpass."[2] Here I stress the word "imitate." In this context, the best definition is "to mimic."

There are two components to an expert system: a knowledge base and an inference engine. The knowledge base is made of up known facts and rules regarding those facts. The inference engine is a system that applies the rules in a knowledge base to the known facts and, through deductive reasoning, creates new facts. Look at this example:

- ✓ *Known fact #1: A dog is a mammal.*
- ✓ *Known fact #2: A cat is a mammal.*
- ✓ *New fact: Dogs and cats are mammals.*

This result shows that the system has mimicked human or natural intelligence and made a correct deduction.

However, can you see the potential pitfall here? If the inference engine isn't well-supervised, you could have this reasoning:

- ✓ *Known fact #1: A dog is a mammal.*
- ✓ *Known fact #1: A cat is a mammal.*
- ✓ *New fact: A dog is a cat.*

HUMAN-SUPERVISED AI CAN REDUCE ERRORS

This is why the artificial intelligence system developed by Access Innovations—Data Harmony—will always be human-supervised. Some of the techniques used in the subset of artificial intelligence known as machine learning rely on statistical inferences using neural nets, Bayesian statistics, vector analysis, and co-occurrence.[3] Our experience has demonstrated that these techniques, left unsupervised, often come to conclusions similar to our example above, wherein a dog is a cat. In order to avoid such erroneous conclusions, the Data Harmony software incorporates a number of other techniques.

> *"Insufficient data, unstructured data, and data with poor governance will result in unsatisfactory output, no matter the quality of the system and technology. With so-called "black box" systems, those that rely solely on statistical approaches, locating and fixing the sources of the problem is nearly impossible."*

For our knowledge bases, we include entities (people, places, and things), lexical variations (also known as synonyms), part-of-speech tagging, abbreviations, concept extraction, and Boolean style rules.

In our inference engine, we use salience via weighting, syntactic analysis (also known as parsing), semantic analysis, sentiment analysis, pragmatic analysis, grammar, lemmatization (also known as stemming), morphological analysis, sentence boundaries, punctuation marks, terminology extractions, term weighting, co-occurrence (here based on counts of occurrence), word parsing, and phrase parsing.

With this combination of approaches, we have a system that produces reliable, replicable results, so that dogs and cats are always mammals and dogs are never cats.

Machine learning is, in some ways, like a human toddler. Babies and small children learn by experiencing the world and they use all five senses to do it. I provide examples below that include everything except smell.

From birth, children are using sight to observe the world around them. They learn by watching people doing things and events unfolding. Their behavior will mimic that of those around them. A child will often "parrot" their parents' behavior, good, bad, or ugly.

Small children will put almost anything in their mouths to experience through taste. Imagine the temptation to pick up a discarded lollipop from the grocery store floor. They probably already realize it is going to be sweet. The attentive parent reacts by saying, no, that's yucky/dirty/has germs, we don't want to eat it, therefore teaching the child through sound. An inattentive parent may miss the beginning of this episode. What if some vinegar was spilled on

that yucky/dirty/has germs lollipop? The child will experience a negative taste, enforcing the concept that we don't put things on the floor into our mouths.

Imagine the same child getting ready to poke the pet dog in the eye with something sharp. The parent can use verbal guidance again, saying no, we don't poke the dog with something sharp. I hope they will add that this action would cause pain to the dog and that's not a nice thing to do. Or perhaps the parent is in the other room and not aware of the situation unfolding. The child pokes the dog in the eye with that sharp object. Depending on the temperament of the dog, several things could happen. The dog could bark or cry, causing the parent to appear and give additional verbal/sound guidance. The dog could run and hide to nurse the injury, meaning that ample time passes between the incident and the guidance—if it happens at all. Lastly, the dog could bite the child, causing the child to cry out. In this event, the child gets guidance through touch that equates to pain. One hopes that the child will learn that poking the dog in the eye with something sharp leads to the child being punished by the parents.

Machine learning can only grow through the sound sense, translated into written or coded messages. Here, we are the parents, guiding and supervising the development of the system. The system can make inferences based on the input, but left unsupervised, the results are unpredictable. Left alone, the unpredictable results can expand exponentially, leading to outputs that are unreliable and not replicable.

HUMAN INPUT IS ESSENTIAL FOR ML AND AI

Data is messy. According to *Towards Data Science*, "Data scientists spend about 80% of their time cleaning and organizing the data."[4] That's a phenomenal amount of time for these brilliant and highly paid data scientists. Imagine if they could spend 80% of their time developing systems instead. Insufficient data, unstructured data, and data with poor governance will result in unsatisfactory output, no matter the quality of the system and technology. With so-called "black box" systems, those that rely solely on statistical approaches, locating and fixing the sources of the problem is nearly impossible. These systems will output unreproducible results if the data is run in an iterative fashion. Human-supervised approaches allow for identifying and resolving key problems. When this is built into the process, corrections are made and iterative outputs improve dramatically.

The landscape is still experiencing massive changes. In an article by Louis Columbus in *Forbes*, he states, "The global machine learning market was valued at $1.58B in 2017 and is expected to reach $20.83B in 2024, growing at a Compound Annual Growth Rate of 44.06% between 2017 and 2024."[5] While the investment in technology—hardware, software, and code—is massive, we are finally seeing that companies with famous ML and AI systems concede that humans still have to be in the loop. In a 2017 interview with IBM Vice President Ed Harbour, he says, "Watson technologies are *trained by humans* to understand information specific to different industries, specialties, and languages—in other words, Watson learns in an expert way, not just a general way. This involves training the system to recognize patterns by feeding it large amounts of labeled data and then *working with human experts* to refine the answers. Through successive rounds of input and feedback from *subject-matter experts*, Watson's understanding and responses improve."[6] This is something that Access Innovations has known and been doing for 25 years!

Focusing on the goal and less so on the technology is key to succeeding with machine learning and artificial intelligence. In a post by The Enterprisers Project, Stephanie Overby refers to this as "Shiny Things Disease."[7] In the same post, Anil Vijayan, vice president at Everest Group, is quoted saying "At this point in time, many enterprises have inflated expectations from AI solutions." As I mentioned at the beginning, this is not a Star Trek-esqe computer. We're a long way from that kind of capability, but science fiction has very likely magnified our expectations.

There is a lot written about "Why Machine Learning and AI Fail," and very few true success stories—at least for now. Adopting some cautionary procedures should generate some of those longed-for success stories soon:

1. **Manage your expectations**
2. **Understand what AI and ML can and can't do**
3. **Focus on the goal and let the technology follow**
4. **Grok your system and the approaches it uses**
5. **Clean your input data**
6. **Apply common (human) sense—also known as natural intelligence—and review the output**

The Data Harmony software from Access Innovations has had productive systems, implemented and operational, across a wide variety of organizations, for over two decades. These systems don't have the flash and drama of IBM Watson and other AIs, but *they are working*. ▮

[1] https://en.wikipedia.org/wiki/Expert_system
[2] https://www.dictionary.com/browse/emulate
[3] In this interpretation, we define co-occurrence as two unique text strings occurring in the same document.
[4] https://towardsdatascience.com/whats-tidy-data-how-to-organize-messy-datasets-in-python-with-melt-and-pivotable-functions-5d52daa996c9
[5] https://www.forbes.com/sites/louiscolumbus/2020/01/19/roundup-of-machine-learning-forecasts-and-market-estimates-2020/#1f4a5df45c02
[6] https://www.fool.com/investing/2017/08/30/ibm-watson-everything-you-ever-wanted-to-know.aspx
[7] https://enterprisersproject.com/article/2020/3/why-ai-projects-fail-8-reasons?MessageRunDetailID=1548412674&PostID=12259128&utm_medium=email&utm_source=rasa_io