ELLIE WARREN & SARA BEERY

10

THE PROMISE AND THE PITFALLS OF MACHINE LEARNING FOR CONSERVATION

'Even if you're not seeing a 99% accuracy rate, being able to say, 'Look, we only have to manually label 15% of our data now because of the ongoing work we're putting into this' should feel like a success, especially compared to time-consuming alternatives without machine learning tools.'

Sara Beery

In this case study, **Ellie Waren** and **Sara Beery** sit down for a conversation about the sky-high hype and inevitable disappointment when it comes to our expectations for machine learning in conservation.

MegaDetector detects humans, vehicles, and animals in camera trap data.

Machine learning is often touted as conservation technology's silver bullet, the tool that will make our work infinitely easier, faster, and more effective. But those who work with machine learning can tell you from experience that it's far from a magic solution, and in fact, the hype surrounding machine learning's potential can make its failures feel that much worse.

As someone who is far from an expert in machine learning, I sat down with AI for Conservation expert Sara Beery, part of the team that created Microsoft AI for Earth's MegaDetector, to discuss machine learning's unique challenges, and how learning from those challenges can help make this tool more useful and accessible to conservationists and ecologists.

In speaking with Sara, one common theme keeps rising to the surface of our conversation: when machine learning tools fail to deliver consistent results, i.e. when a model achieves very high accuracy on a prototype dataset but doesn't work in the field, the cause is often that the prototype data wasn't representative of the end use case. This means that when the model "fails" it's really being asked to do something significantly outside the scope of what it has been trained to do. And because many of us don't yet understand exactly what machine learning is capable of, we're more likely to buy into hype and sky-high expectations, resulting in a feedback loop that leads us to expect near-perfect performance, and then feel disappointed by the inevitable letdown.

"A huge part of the problem comes down to the data we're using to train and test these models," says Sara. "Beginners expect that training a ML model is the challenging part, but really, the training isn't that hard. The real challenge is that data curated for ecology tends to be project-specific, covering limited geographic areas or taxonomic groups, and collected from project-specific sensors. To build a one-size-fits-all machine learning model, you would need to collect a dataset that covers all possible use cases - which in a changing world is essentially impossible."

So how does that relate to our expectations for machine learning, built around incredible promises like 99% accuracy? "In almost every paper that promises those kinds of results for the ecological community at large, the data they have built their model on will not support their broad claims. It is entirely possible to attain 99% accuracy in a highly controlled setting - fixed sensors, time periods, sensor placement strategies, etc. It's not that they're not getting great results, and it's not that the machine learning model itself is doing anything wrong. The problem is that it's never going to work that well for anyone else or any other project, making it "fail" when used in the real world. They're not testing the model in a way that shows how it will work for other potential users, and the media hype makes it seem like 99% accuracy for one project means 99% accuracy for everyone,

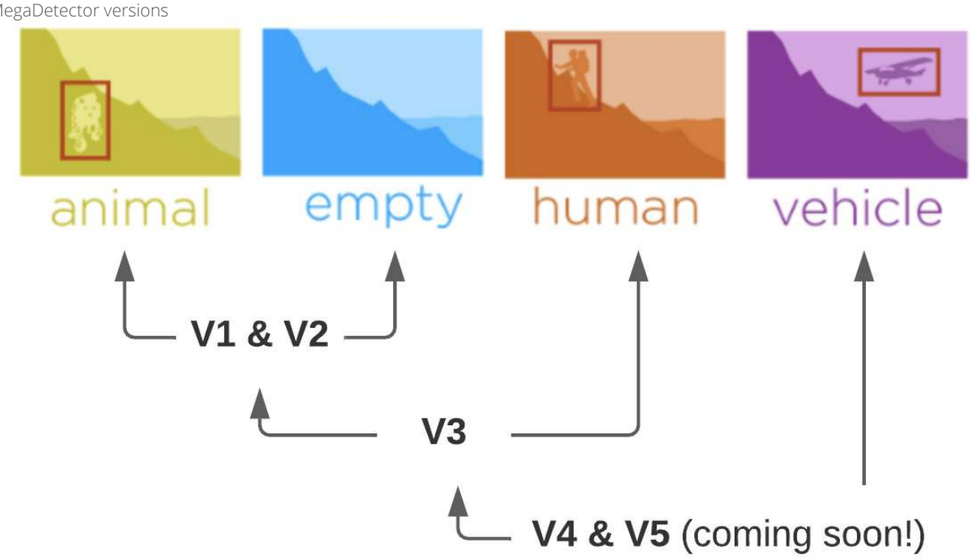
with no additional effort. That disconnect between expectations and reality is what makes us perceive anything less than almost perfect results in actual practice as a failure, when in fact, a slightly-less-perfect machine learning model can still save you a lot of time and effort."

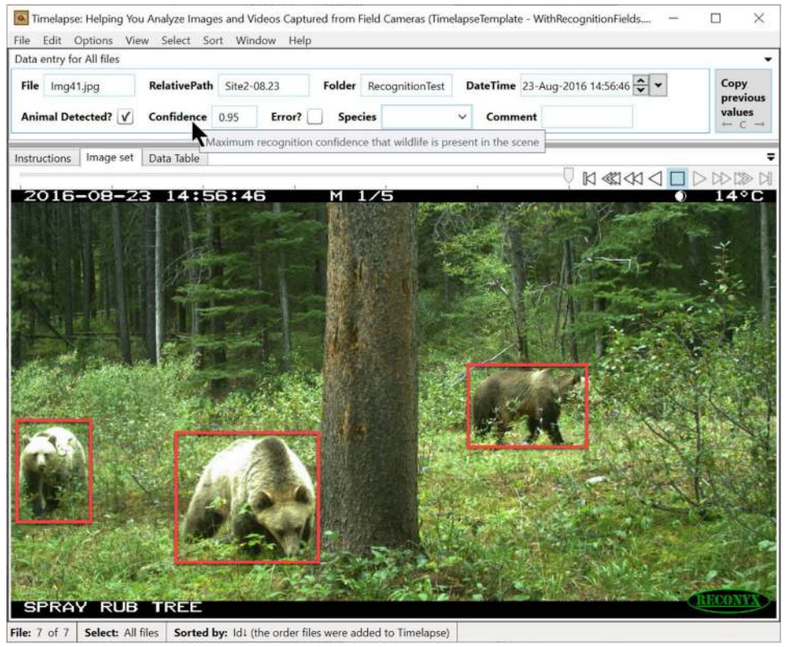
Tools like MegaDetector, which detects humans, vehicles, and animals (here "animal" is a single, generic class) in camera trap data, saves users time combing through massive amounts of empty photos. A significant amount of effort has gone into making the results of the model easy to analyze for any new project, vital when using a machine learning model in practice. The user-friendliness comes from MegaDetector's ability to pick out wildlife - any wildlife - from enormous datasets without needing to retrain the model for new projects. However, when it first became available to users, Sara says, "It took a lot of handholding for each group. We'd walk people through the process of using it, and there was a lot of uncertainty at first about the pain points and how to put it into practice most effectively. It's built to be easy to use and work as well as possible off the shelf, anywhere in the world and for any animal taxa. That said, every user has different needs and requirements depending on their study, so each user has to weigh their own pros and cons when it comes to how they use the model. No two user's needs are the same. For example, if you're surveying

> It's not that they're not getting great results, and it's not that the machine learning model itself is doing anything wrong. The problem is that it's never going to work that well for anyone else or any other project, making it "fail" when used in the real world.

something like invasive rodents on islands, any sighting of a rodent is highly significant, so you want to be sure that you're not missing detections. For this use case, you'd use a lower confidence threshold on model detections, which reduces the risk of missing one but requires humans to filter through a larger number of false positives. Because the MegaDetector doesn't predict the species of detected animals, if you're looking for one specific species that's elusive or rare in your area you're still going to be combing through large numbers of animal images in your dataset

MegaDetector versions





MegaDetector incorporated into TimeLapse

searching for your subject species. MegaDetector saves time by eliminating empty images that don't contain wildlife, but it doesn't mean you won't still have to do further data processing."

When considering what machine learning tools will most successfully meet your needs, Sara recommends considering your priorities, resources, and the risk associated with errors for your study. "With any trained machine learning model, you can pick an operating threshold that will trade off between recall, avoiding missed detections but resulting in more potential false positives to analyze, and precision, where your predicted results are more accurate, but you may have a higher risk of missing something important. Knowing which one of those options will lead to the right tradeoff between human processing effort and risk for your study is something you'll learn from experience, and it's based on knowing how to match your tools to the question you're trying to answer."

But to those without previous experience in meeting machine learning's intrinsic challenges, the perception of failure can creep in simply from being forced to consider these factors. "People get frustrated by the fact that off-the-shelf tools can't do exactly what they want immediately," says Sara. "People expect 99% precision because that's what they're seeing reported in papers. But your expectations for use should include investing the time to carefully analyze how well any off-the-shelf tool works for your data, learning how to fine-tune an existing model for your specific project if you need specific predictions (like species, gender, age, or behavior), and committing to iteratively fine-tuning that model and doing continual quality control for new seasons or new deployments."

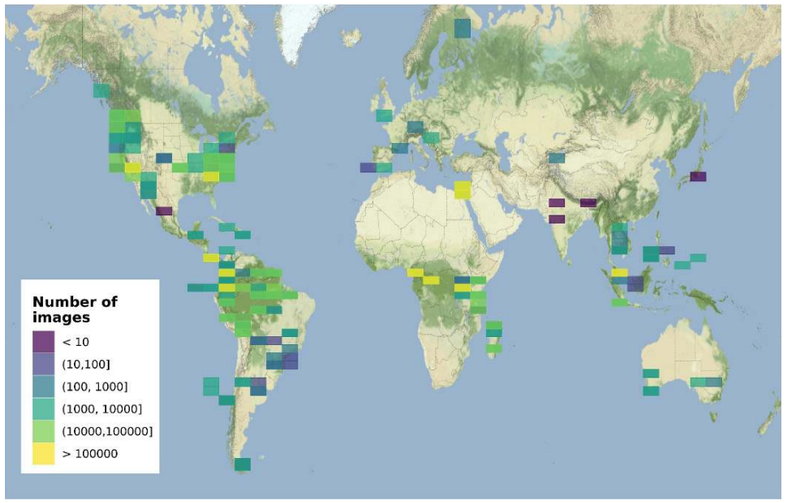
Adjusting your perspective can play a big role in helping you see the value of your own efforts and avoiding the urge to see a lack of perfection as failure. "Even if you're not seeing a 99% accuracy rate, being able to say, 'Look, we only have to manually label 15% of our data now because of the ongoing work we're putting into this' should feel like a success, especially compared to time-consuming alternatives without machine learning tools. For example, MegaDetector user Beth Gardener at the University of Washington told us, 'We had a big image processing party last week [...] Because of the MegaDetector; ~6 of us processed over 100,000 images in one day. That would have taken weeks or months before.' I count that as a

considerable success for AI, and I expect that as our research and our access to data improves that we will make experts increasingly efficient in processing their large datasets."

Frustration can also stem from lack of clarity about the common metrics for measuring machine learning's success. "Metrics like accuracy can be misleading depending on your data. For example, if you get 90% classification accuracy from a machine learning model predicting whether an image is of a dog or of a mountain lion, that sounds really great! But if your data is 90% dogs, your model can get that accuracy by predicting 'dog' for every single image. When data is imbalanced, sometimes a class-averaged metric is more interpretable, because it better captures how well the model is doing across all of the possible classes. However, optimizing for class-averaged metrics frequently result in models that make more errors on common species, and if a common species is 90% of your data that's a lot of errors. I recommend users break down performance across classes, and if applicable across sensor deployments, seasons, and regions, to better understand what the model is doing and to decide which images to trust the results for, and which images to send for additional human review."

When building machine learning tools, part of the learning process when it comes to ongoing maintenance and quality control revolves around accepting and dealing with "the drop-off." As Sara explains it, "With camera trap data, you'll never get as good of performance out of your machine learning model as you did last season. You'll always see a drop-off in performance, and if you don't realize that and put in the effort to re-analyze model performance, your results won't be what they should be, and you'll have a poorly-calibrated sense for the trustworthiness of those results. This process can be incredibly frustrating for people who don't realize that quality control and model training isn't a one-time thing."

To demonstrate the constant need to practice diligence and good quality control habits, Sara shares a story from her work with Wildlife Insights, a machine learning platform that seeks to tackle the challenge of robust, global species identification by curating diverse camera trap data from around the world, and simultaneously provides users with a powerful platform for data management and analytics. "This is a case where we thought we'd done our due diligence. We learned that just because something isn't a problem the first time, or second time, or many, many times in a row, that doesn't mean it'll never be a problem." The issue arose when the team started to analyze the performance of a new model version before its release. The new version included a large amount of training data from a new projects, and to everyone's surprise, for no apparent reason, the



The distribution of data (images) contributed to Wildlife Insights. Created by Fabiola Iannarilli, Yale University, on behalf of Wildlife Insights

PROMISES AND PITFALLS OF MACHINE

Tools like MegaDetector, which detects humans, vehicles, and animals (here "animal" is a single, generic class) in camera trap data, saves users time combing through massive amounts of empty photos.

model began frequently predicting the presence of domestic cats. Lots and lots of cats, on images that were clearly deer or dogs or cows, species the model had handled very well in the past.

"In what seemed like a sudden catastrophic failure, we were getting detections of cats in what felt like almost every photo from certain camera trap projects." To understand what went wrong, we must go back to a potential issue that Sara thought had been thoroughly investigated. "Every camera trap brand has its own logo or watermark on photos. We'd previously wondered if all those different logos would bias our machine learning models or impact performance. But in earlier testing, we found that different logos in different locations in the image frame didn't seem to throw off generalization or make a difference in our results."

The mysterious abundance of cats turned out to be the result of a large project from an urban area that captured lots of cats and used only Bushnell cameras. By some chance, most other projects already in WI were using other camera trap models, so the algorithm learned the "easy" association that an orange Bushnell logo meant the image contained a cat. The team was reminded the hard way how machine learning models will always take the easy way out, and memorize spurious correlations in the data if possible. "This model was so good at everything else, but was making these weird cat errors that just didn't make sense! It took us a while to figure out what had gone wrong. We might've realized that the Bushnell logo was causing the problem sooner if we hadn't already invested time into analyzing whether logos caused issues in previous versions of the model. Because we'd already invested that time and energy, it was easier for us to overlook it because it wasn't on our radar anymore as a potential problem.

Ecologists and conservationists will need to develop an intuition about machine learning, and that comes with use and experience. The goal is to break down knowledge barriers and make it more accessible for ecologists to use practically.

But that's a good example of accepting that just because your model doesn't have a problem with something now, it doesn't mean it'll never have a problem. Don't trust the results of any model without corroborating them; otherwise, you won't recognize those problems when they do pop up." Because the Wildlife Insights team caught the error and were able to determine the cause, they were able to retrain their model after cropping out logos and verify that it fixed the issue. The new model version no longer has a love affair with cats.

With more established and familiar types of conservation technology, like camera traps themselves, the idea of failure may be easier to digest. After all, hardware can malfunction, especially when exposed to the elements and unpredictable wildlife. But with all machine learning's hype as the future of conservation tech, our own expectations may be setting machine learning up to fail. And that's unfair. Machine learning will very likely play a huge role in conservation's future, particularly as tools like MegaDetector and Wildlife Insights make it more and more user-friendly. But like we've come to accept mishaps with hardware, we need to accept machine learning's realities and current limitations in order to realize its full eventual potential.

Equally important is recognizing that machine learning's current capabilities are not its ultimate destination. This technology, like all technologies, will only improve and become more accurate and accessible over time. And with increased accessibility, Sara sees a bright future full of promise for machine learning. "Ecologists and conservationists will need to develop an intuition about machine learning, and that comes with use and experience. The goal is to break down knowledge barriers and make it more accessible for ecologists to use practically. If you give people the skills to experiment with machine learning, you open the door to innovative ideas, and new, exciting human-AI solutions for conservation and sustainability challenges."

ABOUT THE AUTHORS

ELLIE WARREN
WILDLABS COORDINATOR, WWF

Ellie Warren creates content and supports the conservation technology community at WILDLABS through virtual events, fellowships, and community engagement. She currently works as WILDLABS Coordinator at World Wildlife Fund, and has a background in English, nonfiction writing, and screenwriting.

SARA BEERY
PHD CANDIDATE, CALTECH

Sara's research focuses on building computer vision methods that enable global-scale biodiversity monitoring. She works closely with Microsoft AI for Earth and Wildlife Insights (via Google Research) where she helps turn her research into impactful tools for the ecological community. She seeks to break down knowledge barriers between fields. She founded the AI for Conservation slack community (+600 members), and is the founding director for the Caltech Summer School on Computer Vision Methods for Ecology.