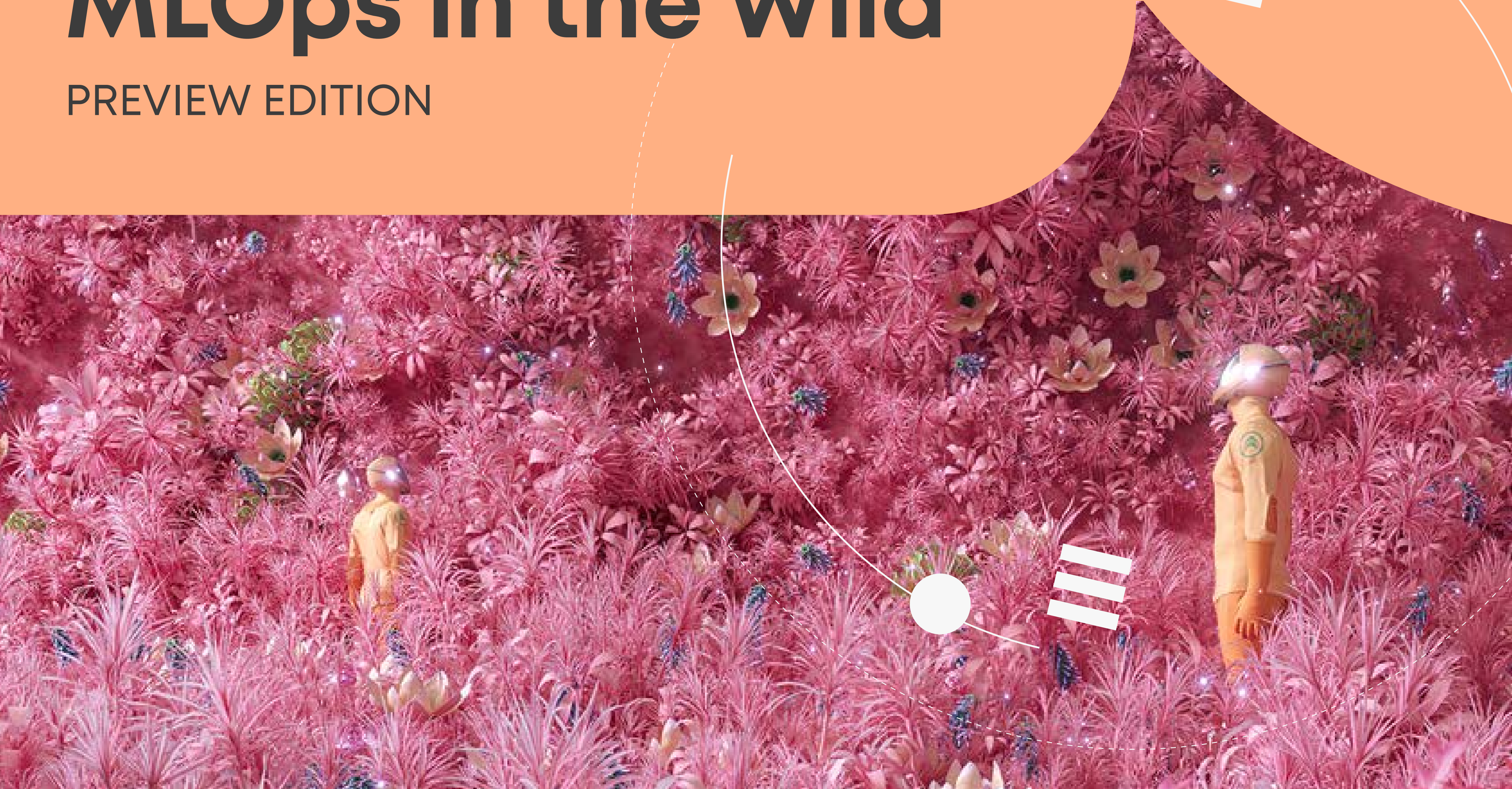
Valohai

MLOps in the Wild

PREVIEW EDITION



Foreword

The MLOps scene is seeing tremendous growth from all sides. Data scientists and machine learning engineers are looking for better ways to work. Looking at Google search volume for MLOps shows a growth trend that really kicked off in late 2019 and continues to accelerate. In November 2020, we released an eBook on MLOps and so far we've had over 2300 individuals download it. This enthusiasm exceeded our expectations and highlights for us the demand for machine learning operations.

This demand is met, and probably at this point, exceeded by supply for different MLOps solutions. A community-led effort to collect all the different tools in the space produced a list of over 330 different MLOps tools. Although these tools address different problems and use cases, the list is simply staggering.

This brings us to this eBook: MLOps in the Wild. Many of the data science teams we speak to know they have a need for tooling but pinpointing what they need exactly is more difficult. The MLOps space is still in its infancy and how solutions are applied varies case by case. We felt that we could help by providing examples of how companies are working with tooling to propel their machine learning capabilities.

We wanted to keep these case studies short, skimmable and inspirational. Think of this as a lookbook for machine learning systems. You might find something that clicks and opens up exciting new avenues to organize your work – or even build entirely new types of products.

Table of Contents

This is a preview edition of the MLOps in the Wild and only contains the three first case studies.

As a reader of this edition, we would love to hear your feedback:
- What kind of cases would you like to see?
- Do you have a case you would like to contribute?

Email us at hello@valohai.com

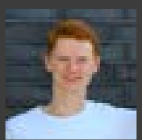
CLASSIFICATION, GERMANY

Automating custom model training for document processing

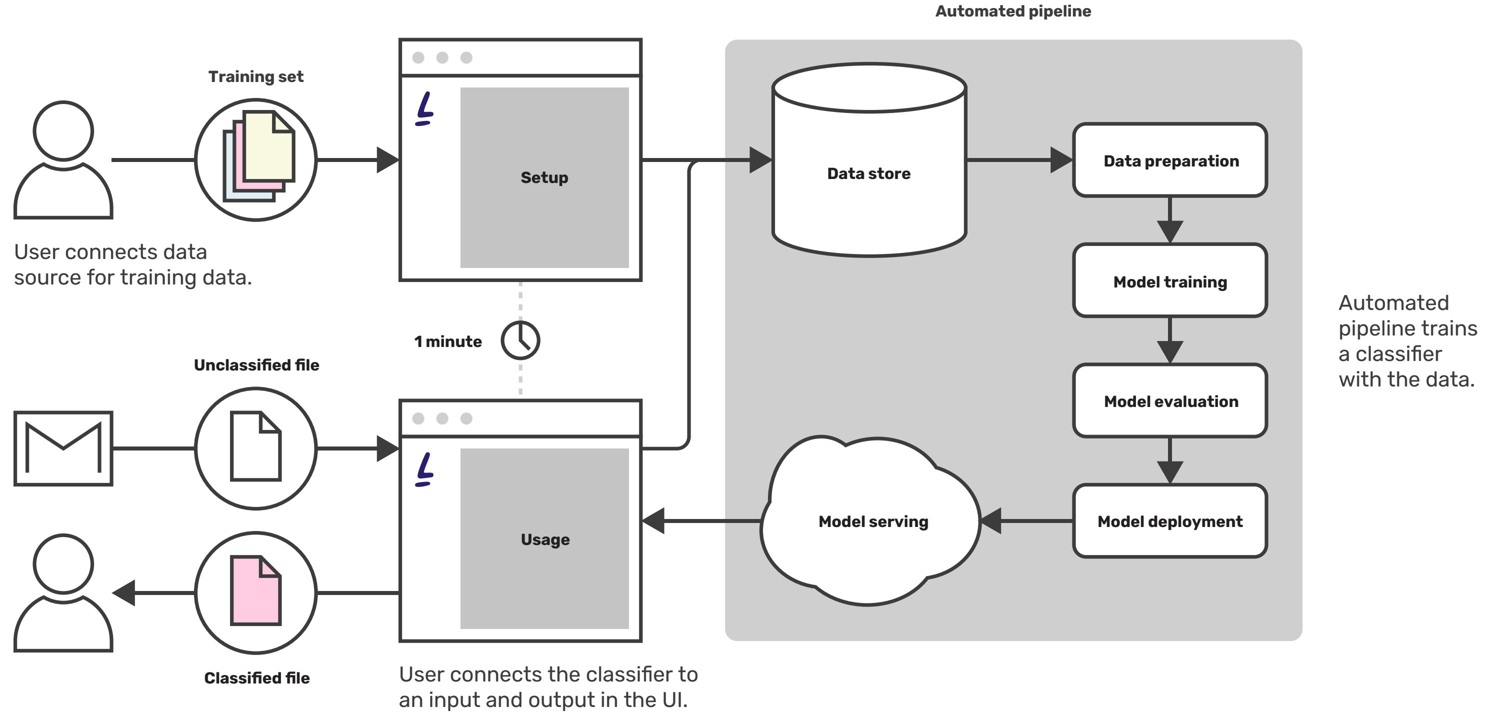
Levity enables companies to automate workflows specific to their business, from recognizing objects in microscopic images to automatically categorizing incoming documents for different internal workflows.

Team Size	3 people
Time Spent	4 months
Models Trained	1,000
Predictions Made	1,000,000
Technologies	Python Tensorflow Valohai GCP

Contact Person



Thilo Huellmann
CTO & Co-Founder
at Levity



User Story

In Levity, users are using a series of pre-built integrations to connect their labeled data, train a new classifier to automate their manual PDF, image, or text classification workflows.

After the user connects the training data, they specify how they want to receive the output of their custom classifier, for example directly into a Google Sheet.

The workflow is then connected to an input for new, unclassified data. From setup to usage, Levity can train and deploy a custom classifier within minutes for real world use.

Challenge

The central challenge Levity faced was how to build a system that delivers an excellent user experience that requires no knowledge of machine learning or associated terminology. The data upload needed to be dead simple so the user could populate a training data set for the custom classifier.

The other part of an excellent UX was that the trained, custom model would need to be usable quickly. The whole pipeline from data processing to training and deployment needed to happen within a few minutes or less. And of course, this has to work consistently for every user.

Solution

Part of the solution was designing the Levity product to require user input in the right places and in the right way. The user can adjust the accepted error rate to clarify that there may be outliers that are not represented in the training data.

The on-demand model training is at the core of the product, and therefore the MLOps platform needed to be deeply integrated. The Valohai API enabled just that allowing Levity to train and deploy models via an API call and even rollback previous versions of the model if needed as all the artifacts produced are stored and versioned.

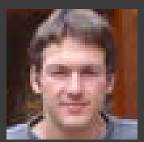
COMPETITION, WORLDWIDE

Scoring backend for the NeurIPS black-box optimization competition

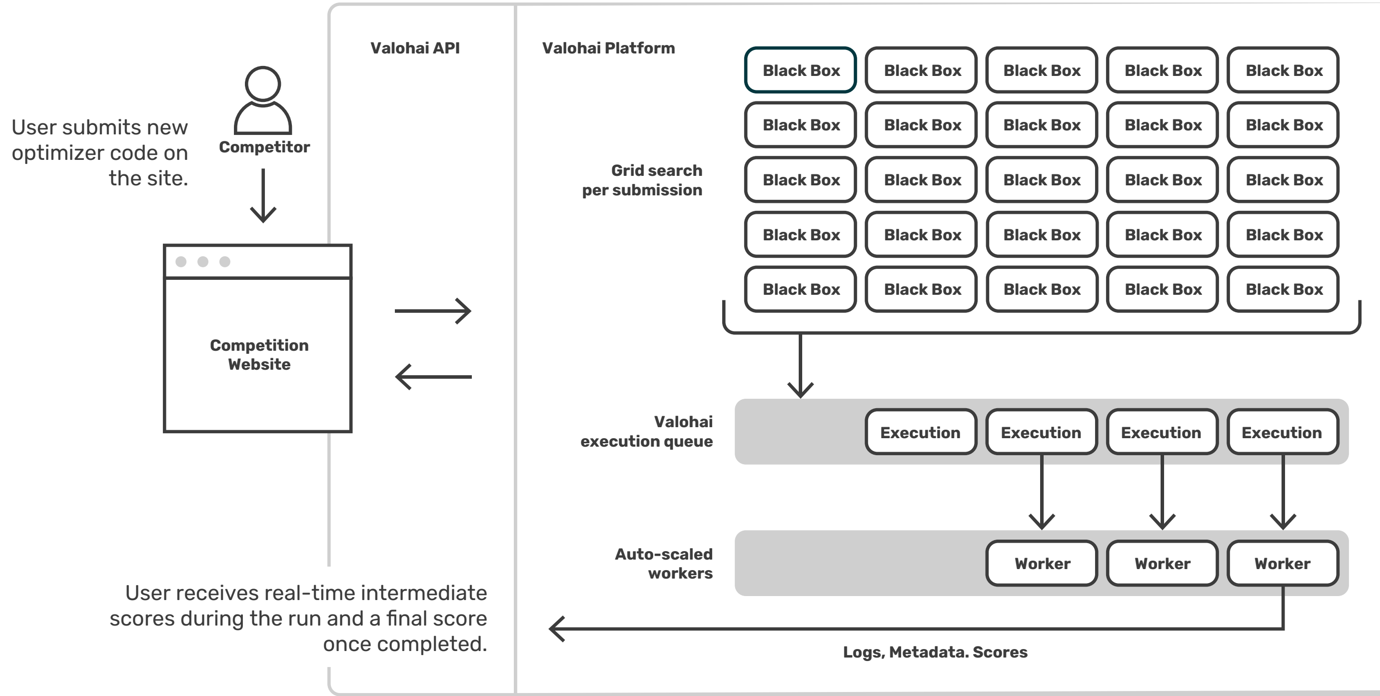
Scoring black-box optimizers using ML models in real-time with auto-scaling to handle the peak traffic and huge computing requirements of a popular NeurIPS competition.

Team Size	4 people
Time Spent	3.5 months
Models Trained	277,000,000
Technologies	Python Bayesmark Valohai

Contact Person



Ryan Turner
Senior Research Scientist at Twitter



User Story

In the NeurIPS 2020 black-box optimization competition, the participants submit an optimizer that needed to be scored immediately. To accurately score a black-box optimizer, the backend uses the optimizer to train several ML models for thousands of iterations.

The participants get intermediate scores while the optimizer is being scored to improve the stickiness of the competition. Waiting for a score for too long would encourage participants to leave the competition prematurely.

Challenge

The challenge for the backend was simply the massive amount of computing resources required to score an optimizer. For optimal user experience, the scoring needed to happen fast enough and with real-time intermediate scores.

The resources were absolutely required to be on-demand. The scoring queue load was unpredictable, and static idle worker machines would have been way too costly.

Solution

Valohai provided a robust backend where resources could automatically scale alongside the master scoring queue load. More than 300 worker machines were crunching the numbers during peak hours – and at quieter moments, not a single one.

As a bonus, the Valohai platform provided intuitive access to inspect the scoring tasks. This helped a lot in a competition where the participants were not allowed to see the errors. Organizers could still easily comment on problems through manual channels as they had access to full history.

PRELIGENS

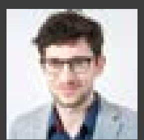
EARTH OBSERVATION, FRANCE

Democratizing resources within a large data science team

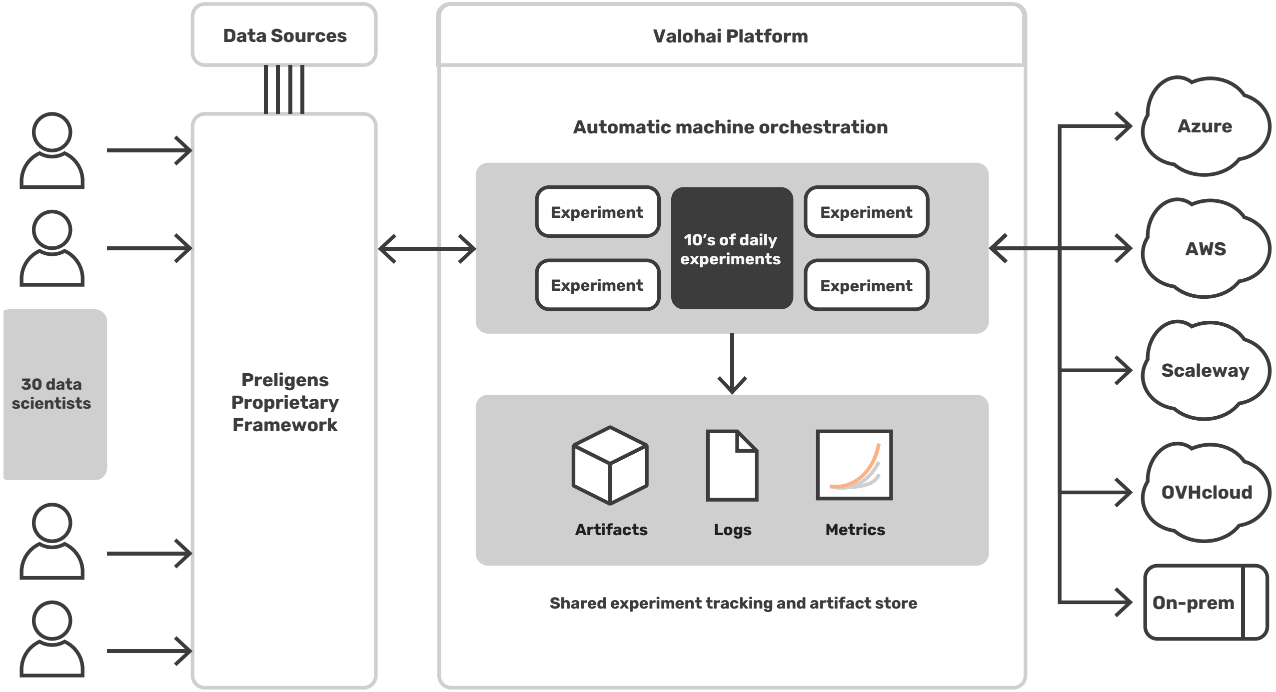
Preligens develops AI solutions to analyze geospatial data, such as recognizing and tracking objects from satellite imagery. They have one of the largest R&D teams in this field in Europe.

Team Size	30 people
Daily Experiments	>10 experiments
Avg. Experiment Runtime	2 weeks
Technologies	Valohai

Contact Person



Renaud Allioux
Co-Founder & CTO
at Preligens



Preligens can utilize cloud and on-premise machines flexibly depending on which is the most economical.

This can be done without any extra DevOps because the platform automatically handles the machine orchestration.

User Story

Preligens has a large, growing data science team that works on various models embedded in their geospatial intelligence product.

When a new data scientist starts at Preligens, they want to be effective as soon as possible. The data scientist is onboarded to the shared MLOps platform on day 1 to ensure that s(he) has access to past experiments.

The newly onboarded data scientist usually starts with running their first experiment in a sandbox project on the platform. The simple example teaches how to utilize cloud machines without any help from IT or DevOps.

Challenge

Preligens needed a flexible infrastructure layer to facilitate a large team doing daily deep learning experiments involving satellite imagery. These experiments may take days to complete, and therefore data scientists have to use remote hardware, i.e. cloud or on-premise machines with GPUs suited for the job. For this reason, Preligens needed a multi-cloud approach.

Preligens wanted to integrate the MLOps platform with their proprietary framework, which is used to facilitate work and centralize data, parameters, and other resources. For this reason, the infrastructure layer needed to be fully controllable through an API.

Solution

Preligens adopted Valohai to bring together the whole team under a single tool. Three main features helped solve the team's challenges.

Shared experiment tracking allowed data scientists to collaborate. Each experiment is automatically stored and reproducible by anyone.

With the platform handling the infrastructure layer, data scientists could easily access remote resources without DevOps knowledge.

And finally, organization management allowed leadership to measure and manage costs.