All Vision: A Platform for Fast and Scalable Unsupervised Learning

Unsupervised Learning: Advancing the Frontier of Applied ML

How Unsupervised Learning Supports Supervised Learning

Supervised machine learning has proven transformative for targeted capabilities like known entity recognition and risk scoring. Despite its successes, supervised learning faces limits due to its requirement for cost-prohibitive, time-consuming data labeling processes. On top of the costs to labeling, it's difficult to train supervised models for recognizing complicated behaviors, and when new events arise or real world data changes, models fail to function as intended. Furthermore, supervised models are sensitive to drifts in domain data which affect the underlying logic on which the algorithms train. Unsupervised learning presents the key to unlocking the full value within a company's data feeds.

Unsupervised learning leverages statistical reasoning to break down data into segments of like points, without the need for labels upfront. With unsupervised learning, organizations can power applications to rapidly assign and improve training labels, detect emerging trends in areas like fraud, and track for changes in domain data that affect model performance. The tooling for using unsupervised learning stands far behind that available for supervised learning, however. Data scientists often need to spend 6 months or more implementing, selecting and optimizing the right unsupervised learning model for a dataset. As a result, few organizations make effective use of unsupervised learning.

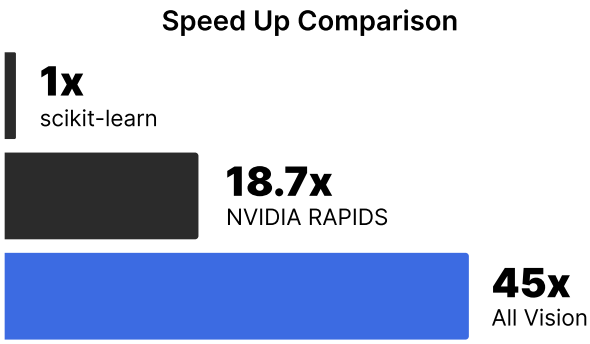
The following limitations create the most problems in unsupervised learning:

1. Available implementations of unsupervised and semi-supervised algorithms train slowly
2. As a consequence of slow training, complete inability to efficiently find optimal hyperparameters
3. Modeling teams rely on high engineering overhead in building supporting infrastructure to enable iteration and deployment of unsupervised models
4. There's a lack of metrics and tools to compare models across unsupervised and semi-supervised types
5. Fragmented communication between data teams and business stakeholders delays model development

The All Vision platform provides the most robust platform to address these challenges and make unsupervised learning accessible to any data scientist.

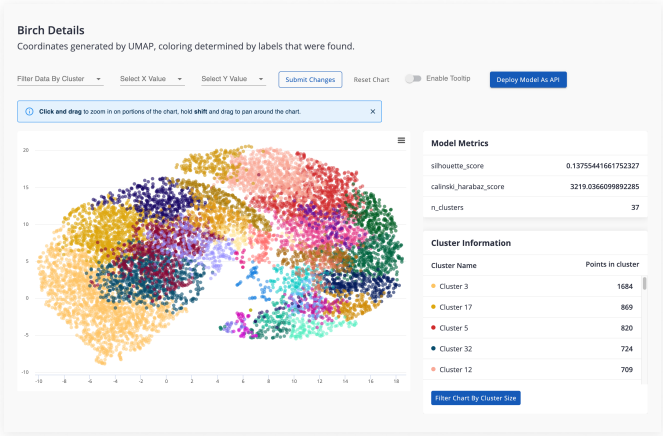
Model Optimization

Nearly all unsupervised learning models that are available open source today in libraries such as scikit-learn or spark's clustering library utilize CPU resources. This limits the extent to which these models can harness parallelizing operations. For unsupervised learning, this is a major bottleneck as most of the operations underpinning unsupervised learning are just array operations. NVIDIA RAPIDS has added GPU support for three models, yet the overwhelming majority of available models remain constrained to CPU support. The All Vision team has worked to optimize model runs in several different ways. First, **the team has reimplemented popular algorithms such as HDBSCAN to take on GPU support** so they can run in a fraction of the time they can on CPUs. This makes it possible to process large datasets and high-dimensional data modalities that have historically been infeasible to assess with unsupervised learning.



Second, **the All Vision platform runs multiple models at the same time, paving the way for the team to identify common computations among different models.** The All Vision team has cached repeated computations so that multiple models repeating the same computation share the results, rather than running through the calculation separately.

Third, the All Vision team has performed extensive research and experimentation to **gauge how much compute is needed for specific dataset sizes** so that data scientists can avoid facing crashes to their compute instances or over-allocating compute for a given dataset. With all these optimizations, the All Vision platform runs models up to 25x faster than the leading alternative approaches. With All Vision, any data scientist can leverage unsupervised learning to its fullest potential.

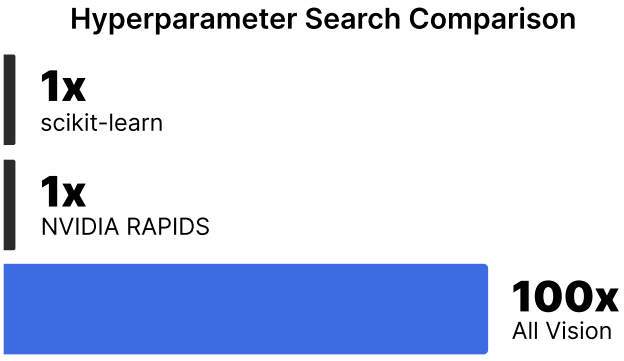


Hyperparameter Optimization

Oftentimes, finding and optimizing hyperparameters is the most difficult and time consuming step in performing unsupervised learning.

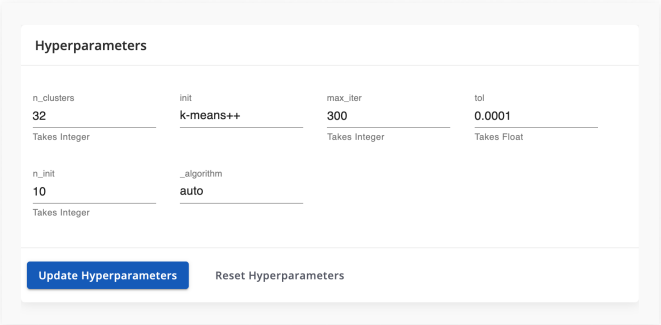
Whether it's finding n_clusters for k-means, finding eps for DBSCAN, or finding any other hyperparameter, getting the optimal hyperparameters for each cluster analysis model typically requires a large grid search involving high costs on compute. Popular libraries such as scikit-learn and NVIDIA RAPIDS make it easy to run a handful of unsupervised models, but they do not offer support for better hyperparameter identification and tuning. All Vision's clustering suite stands as the only automated resource for rapidly optimizing hyperparameters.

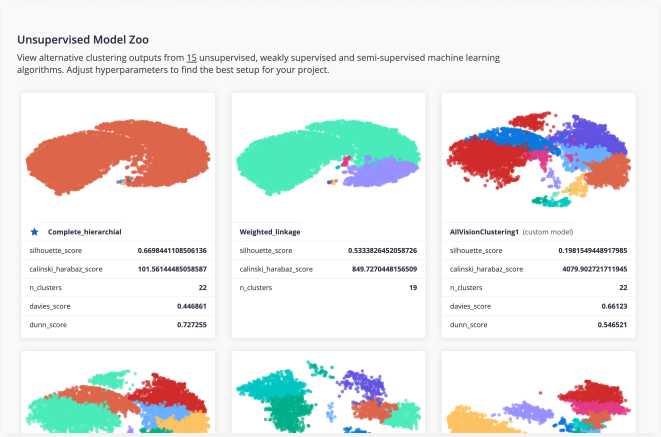
The All Vision team has performed three years of experimentation to develop significantly faster methods for finding and optimizing hyperparameters for the models in the All Vision model suite. The All Vision platform supports several distinct objective functions to optimize hyperparameter search for unsupervised learning. A user selects an objective function, then All Vision employs sequential model-based algorithm configuration (SMAC) to **quickly predict the best hyperparameters for a given dataset** and model using supervised learning. SMAC can speed up hyperparameter searches for each model by well over 100x, **taking place in minutes to hours rather than weeks.**



All Vision software **also employs cross model information sharing to speed up search times.** Essentially, this means that if one fast model can find some information intrinsic to a dataset before a slower model can, this information is shared to optimize the slower model so that it can converge faster on the right hyperparameters for the dataset.

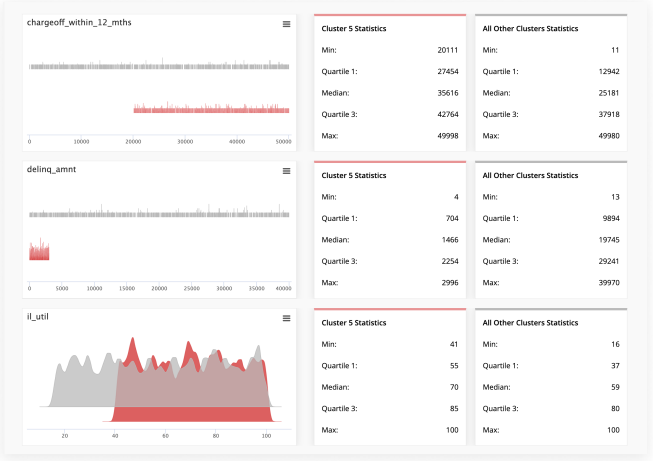
The All Vision team is also working on several optimizations beyond these two methods, in order to completely automate this step in unsupervised learning.



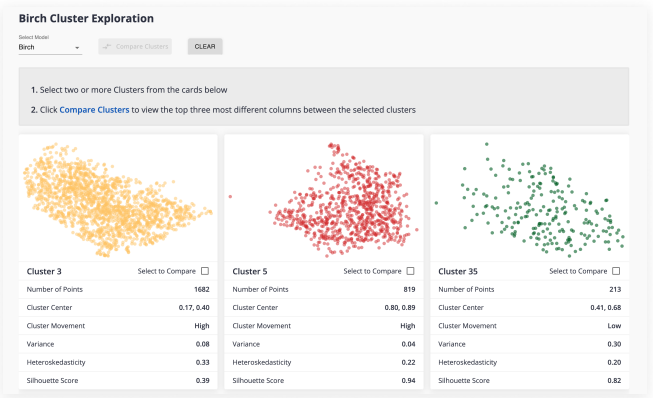


Cluster Explainability

Historically, one of the barriers to using unsupervised learning has been the lack of clarity on why clusters form as they do. Even after spending a large amount of time optimizing an unsupervised learning model, it can be very time consuming for data scientists to understand what makes each group unique and why the unsupervised learning model marked specific data points as similar. All Vision provides an easy way to **contextualize the differences among clusters and between individual clusters and the rest of the dataset**, eliminating the haze surrounding the outputs of clustering models.



All Vision software reads the characteristics of each group and sorts by difference in order of significance, so that users can automatically see the columns that most stand out within a cluster or between clusters. This capability makes it easy to explain to stakeholders what makes each cluster unique, and provides context to understand the underlying logic behind an unsupervised model's classifications.

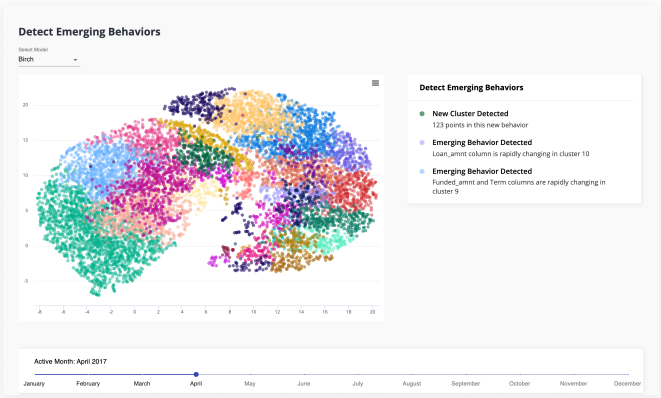


Tracking Cluster Drift

One of the largest problems in machine learning is converting performance on training data to successful performance on real-world data. Drifts in domain data are a major component of this problem. Real world events often change in ways that cause model performance to decline over time, because past models exclude, misclassify or misprioritize data relative to its most up to date circumstances.

All Vision leverages unsupervised learning to **track the emergence of new clusters and movement affecting**

current clusters to detect drift in any domain.



This enables experts to **receive notice when major changes happen**, so they know to adjust operations and make updates to models. Some practical applications of cluster drift detection include responding to new types of fraud, or addressing when an existing customer rapidly bifurcates. This capability is not present in any open source library – currently, the only method that comes close to replicating it is streaming k-means in the spark library. The All Vision platform employs several methods to perform the most thorough monitoring of cluster changes over time, by leveraging the predictive power of All Vision's model suite to find the best approach to track changing trends.