

Shifting from Model-Centric to Data-Centric MLOps



01 Introduction

For organizations working with Machine Learning (ML), the journey from research and development to deployment can be quite challenging. Moving from controlled experiments to real-world applications presents a common set of challenges. It's no longer just about building ML models; it's about creating systems that can adapt to changing data and user needs.

This is where MLOps (Machine Learning Operations) comes in. MLOps provides a valuable framework for ML engineers and data scientists to navigate the complexities of taking ML processes into production. It helps immediately with resource efficiency and cost savings. It also accelerates the deployment of AI solutions, which is crucial in today's fast-paced business environment.

Traditionally, the focus in machine learning was primarily on the model. However, the landscape has shifted from being Model-Centric to Data-Centric. This shift has significant implications for the MLOps framework itself.

To address this shift, Datasaur plays a crucial role in extending MLOps platforms to emphasize 'Data Quality' over 'Model Architecture Quality'. We've simplified the process for users to analyze and work with their data directly within their MLOps workflow by integrating features such as [ML-Assisted Labeling](#) and [Datasaur Dinamic](#), with popular MLOps platforms like [AWS Comprehend](#), [Google Vertex AI](#), and [Azure](#).

In this article, we'll delve deeper into the concept of MLOps, its recent evolution, and demonstrate how Datasaur integrates with and supports this framework.

02 MLOps Definition

MLOps, short for Machine Learning Operations, represents a structured set of processes within the machine learning lifecycle. It encompasses everything from selecting the appropriate machine learning model to continuously iterating between data collection and deploying models into production. MLOps plays a crucial role in maintaining transparency and efficiency throughout this journey by meticulously tracking each step in the machine learning lifecycle.

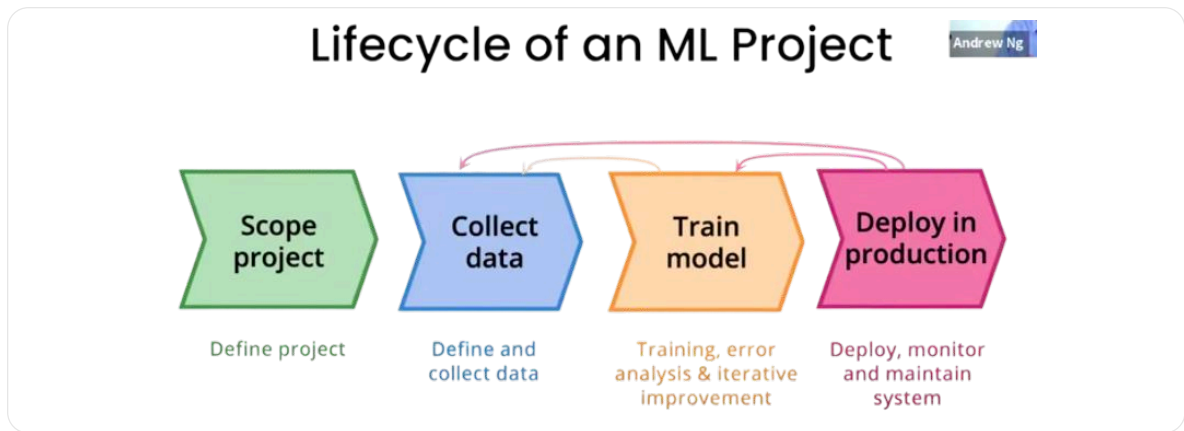


Fig. 1: [A Chat with Andrew on MLOps: From Model-centric to Data-centric AI](#)

MLOps systems can be found in popular machine learning platforms such as [AWS Comprehend](#), [Google Vertex AI](#), and [Azure ML](#). For example, AWS Comprehend offers a comprehensive MLOps system known as [Comprehend Flywheel](#).

In this integrated approach, MLOps streamlines the entire machine learning process, making it more efficient and transparent, ultimately leading to better AI solutions.

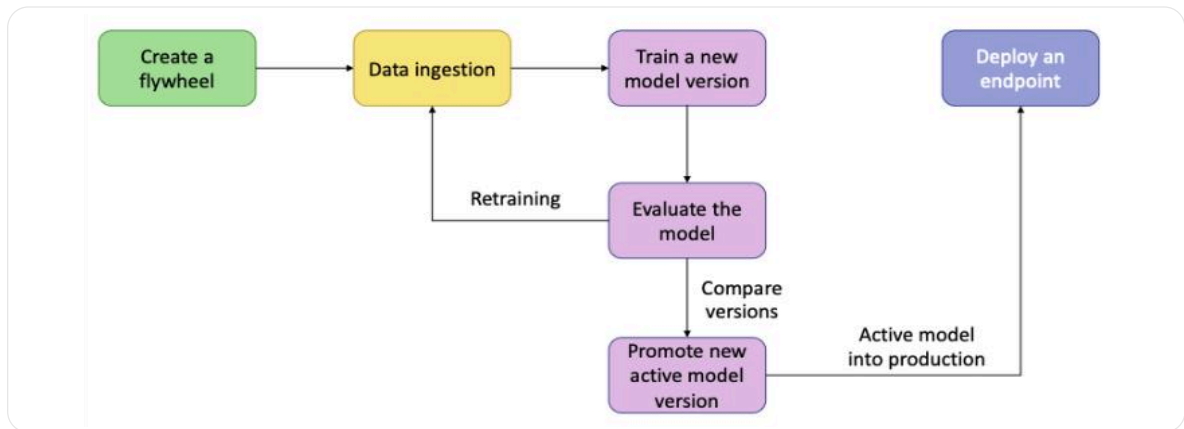


Fig. 2: Deploy an endpoint flow

03 MLOps Model-Centric vs Data-Centric

In the past, MLOps primarily focused on meticulously logging every model's hyperparameters during the training process. However, with the evolving landscape of AI shifting from a model-centric perspective to a data-centric one, MLOps has also transformed to support tools and practices that prioritize data-centric AI.

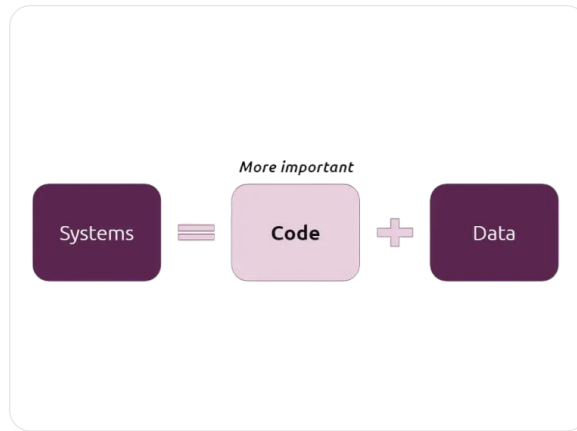


Fig. 3A: Model-Centric focus

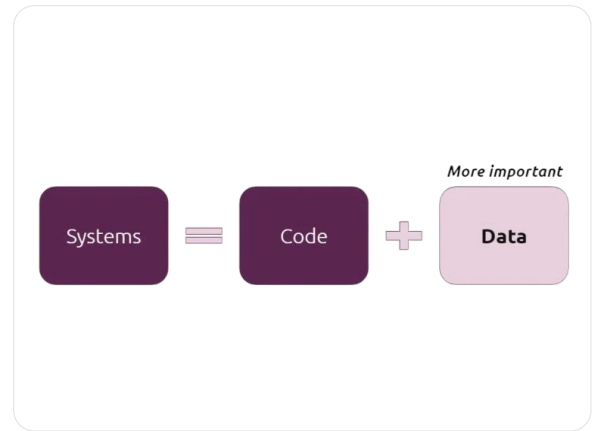


Fig. 3B: Data-Centric focus

The fundamental distinction between model-centric and data-centric AI lies in their development approaches. Model-centric AI centers around model architecture, loss functions, and hyperparameters during training. In contrast, data-centric AI directs its debugging and improvement efforts towards enhancing the model with “high-quality” datasets.

In essence, a data-centric AI approach adheres to several crucial principles:

- Data labeling
- Data augmentation
- Error analysis
- Data versioning

As highlighted by [Andrew Ng](#) in his popular [presentation](#), monitoring data performance, distribution, and flow during development is an effective means of ensuring consistent data quality. Data quality monitoring serves critical purposes, including assessing performance and facilitating data flow for retraining.

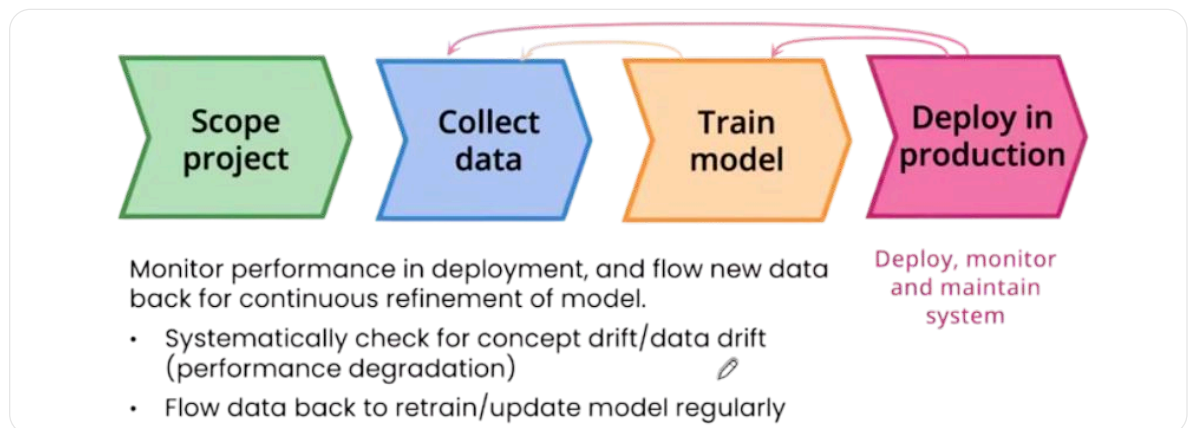



Fig. 4: [A Chat with Andrew on MLOps: From Model-centric to Data-centric AI](#)



The MLOps framework must be bolstered with one or more data-centric principles because, in data-centric model development, maintaining consistent data quality is paramount at every stage of the machine learning lifecycle.

The fundamental distinction between model-centric and data-centric AI lies in their development approaches. Model-centric AI centers around model architecture, loss functions, and hyperparameters during training. In contrast, data-centric AI directs its debugging and improvement efforts towards enhancing the model with “high-quality” datasets.

In essence, a data-centric AI approach adheres to several crucial principles:

- Data labeling
- Data augmentation
- Error analysis
- Data versioning

As highlighted by [Andrew Ng](#) in his popular [presentation](#), monitoring data performance, distribution, and flow during development is an effective means of ensuring consistent data quality. Data quality monitoring serves critical purposes, including assessing performance and facilitating data flow for retraining.

04 How Datasaur Helps in Data-centric MLOps

Build high-quality datasets efficiently.

To begin model development, a high-quality dataset is needed during the model's initial training to deliver accuracy. Datasaur offers a comprehensive NLP labeling platform that facilitates best practices during the data labeling process. It empowers labelers to achieve high data quality through a variety of features and extensions including:

- [Automated Review Capabilities](#): an extension to review labeling results. It is designed to identify conflicts amongst labelers and flag inconsistencies, ensuring issues are caught before model training.
- [Inter-Annotator Agreement](#): this metric is instrumental in assessing label consistency amongst annotators, ensuring that data labeling remains consistent and reliable.
- Automatic incorrect label detection: this tool takes on the task of identifying and rectifying labeling errors automatically, streamlining the data quality control process.

Datasaur also offers a variety of automated tools to help save valuable time and reduce budget:

- [ML-Assisted Labeling](#): Datasaur integrates with existing NLP and LLM libraries such as spaCy and OpenAI to predict labels for the given dataset. It also allows you to connect your own model via API.
- [Data Programming](#): Datasaur expands on the popular open-source Snorkel library to allow engineers to build annotation rules and heuristics in Python to automate labeling. For example, one could write a rule that any text containing the words “pizza” or “salad” should be labeled as “food”.
- [Datasaur Dinamic+Predictive Labeling](#): this feature takes a handful of existing labeled samples to predict how to label additional data.

We cover some of these features in greater detail below.

05 Integration with MLOps platforms

Datasaur enhances data-centric MLOps by emphasizing data quality monitoring. Datasaur integrates with leading MLOps platforms such as [AWS Comprehend](#), [Google Vertex AI](#), and [Azure](#). By doing so, users are empowered with the tools they need to ensure data quality at every stage of their MLOps workflow, within the framework of each of these platforms.

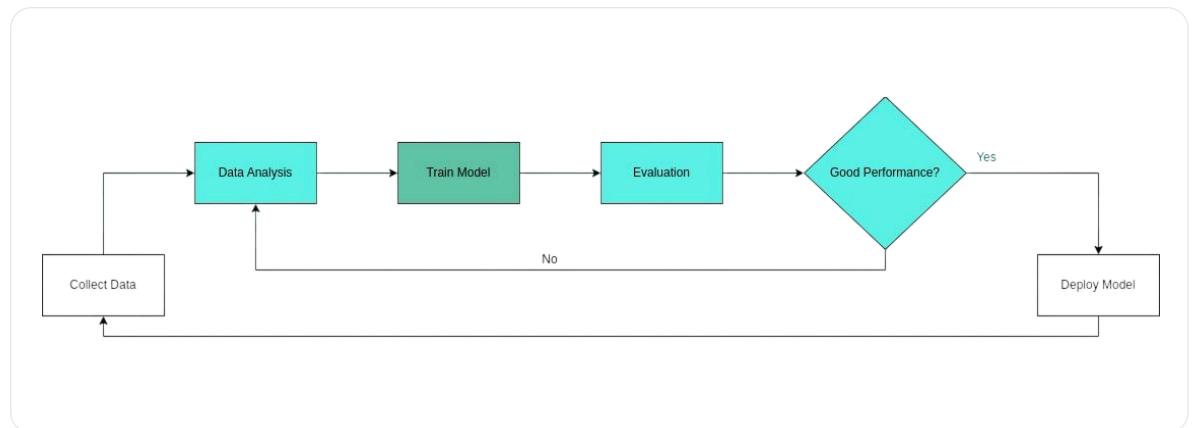
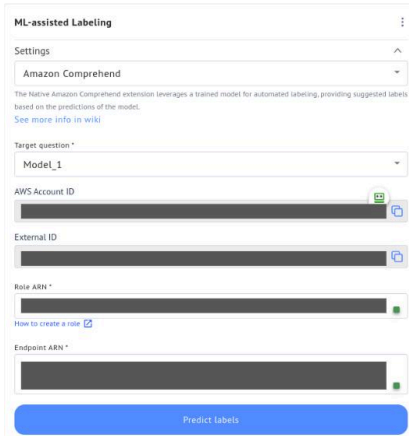


Fig. 5: ML lifecycle handled by MLOps

06 ML-Assisted Labeling

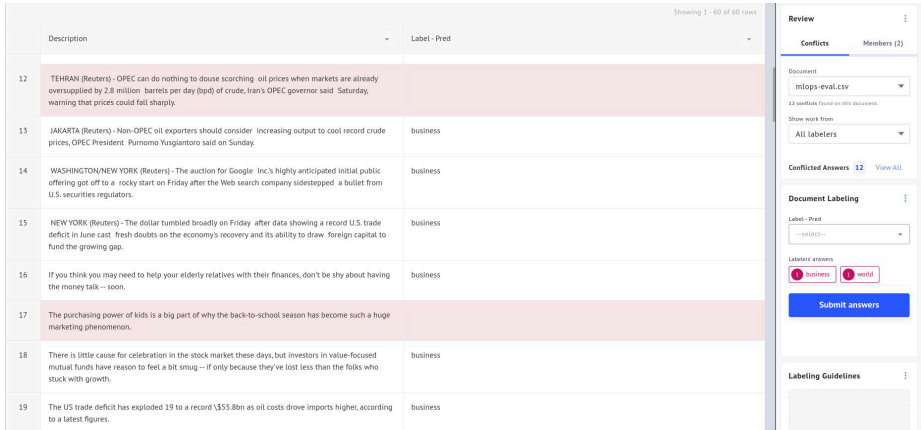
[ML-assisted Labeling](#) is an extension that labels data through custom model endpoint API calls. Users can make API calls to model endpoints then ask human annotators to analyze and evaluate the results of their model in Datasaur's interface. Some users are also utilizing this feature to apply predicted labels from two different models and comparing the outputs of two different models to streamline training.

For example, one could apply labels from OpenAI and from Cohere - these two LLMs may agree on most labels, but the areas where they disagree are data points of interest for a human to review.



The screenshot shows the 'ML-assisted Labeling' settings panel. It includes a 'Settings' section with a dropdown for 'Amazon Comprehend'. Below this is a 'Target question' dropdown set to 'Model_1'. There are input fields for 'AWS Account ID', 'External ID', 'Role ARN', and 'Endpoint ARN', each with a 'How to create a role' link. A 'Predict labels' button is at the bottom.

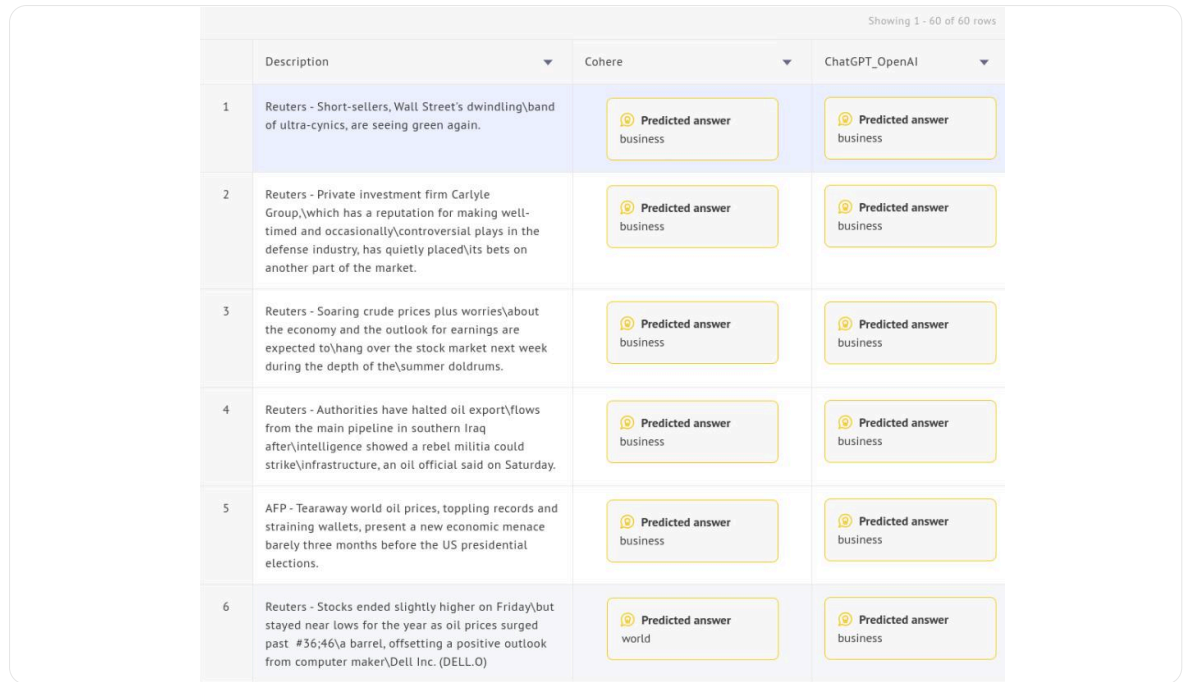
Fig. 6: ML-Assisted Labeling extension



The screenshot displays the Datasaur interface. A table shows data points with 'Description' and 'Label - Pred' columns. Rows 12, 15, and 17 are highlighted in red, indicating conflicts between the two models. The right sidebar contains a 'Review' panel with options for 'Conflicts' and 'Members (2)', a 'Document' dropdown, and a 'Document Labeling' section with a 'Label - Pred' dropdown and 'Submit answers' button.

	Description	Label - Pred
12	TEHRAN (Reuters) - OPEC can do nothing to douse scorching oil prices when markets are already oversupplied by 2.8 million barrels per day (bpd) of crude, Iran's OPEC governor said Saturday, warning that prices could fall sharply.	
13	JAKARTA (Reuters) - Non-OPEC oil exporters should consider increasing output to cool record crude prices, OPEC President Hamed Vahidpour said on Sunday.	business
14	WASHINGTON/NEW YORK (Reuters) - The auction for Google Inc.'s highly anticipated initial public offering got off to a rocky start on Friday after the Web search company sidestepped a bullet from U.S. securities regulators.	business
15	NEW YORK (Reuters) - The dollar tumbled broadly on Friday after data showing a record U.S. trade deficit in June cast fresh doubts on the economy's recovery and its ability to draw foreign capital to fund the growing gap.	business
16	If you think you may need to help your elderly relatives with their finances, don't be shy about having the money talk -- soon.	business
17	The purchasing power of kids is a big part of why the back-to-school season has become such a huge marketing phenomenon.	
18	There is little cause for celebration in the stock market these days, but investors in value-focused mutual funds have reason to feel a bit smug -- if only because they've lost less than the folks who stuck with growth.	business
19	The US trade deficit has exploded to a record \$55.8bn as oil costs drove imports higher, according to a latest figures.	business

Fig. 7: Datasaur shows rows where the two models conflict, allowing reviewers to directly skip to data points of interest

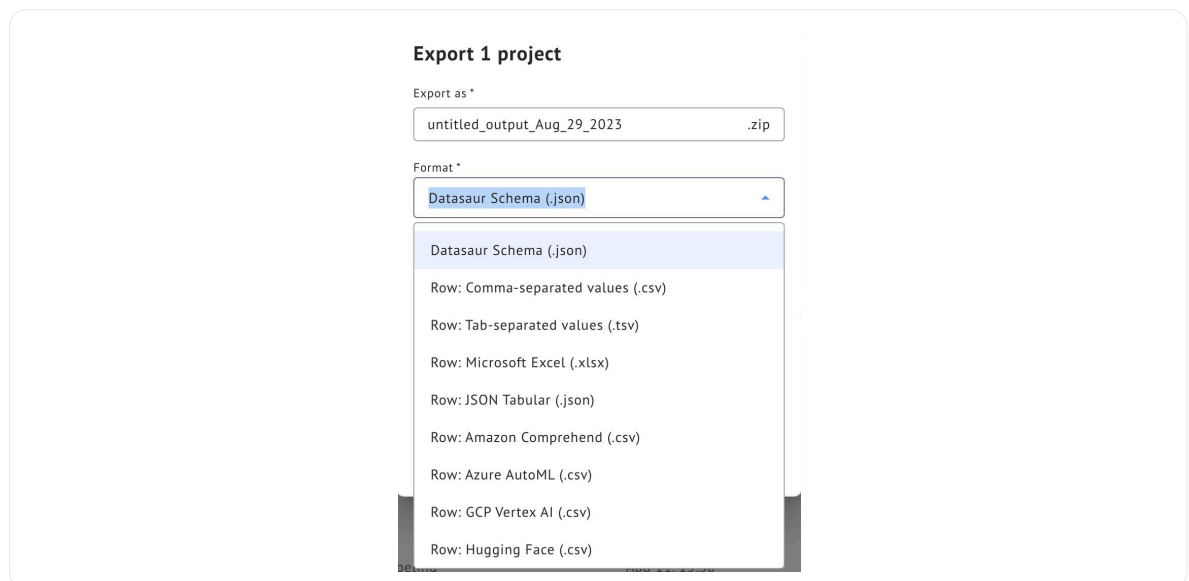


Showing 1 - 60 of 60 rows			
	Description	Cohere	ChatGPT_OpenAI
1	Reuters - Short-sellers, Wall Street's dwindling\band of ultra-cynics, are seeing green again.	Predicted answer business	Predicted answer business
2	Reuters - Private investment firm Carlyle Group,\which has a reputation for making well-timed and occasionally\controversial plays in the defense industry, has quietly placed\its bets on another part of the market.	Predicted answer business	Predicted answer business
3	Reuters - Soaring crude prices plus worries\about the economy and the outlook for earnings are expected to\hang over the stock market next week during the depth of the\summer doldrums.	Predicted answer business	Predicted answer business
4	Reuters - Authorities have halted oil export\flows from the main pipeline in southern Iraq after\intelligence showed a rebel militia could strike\infrastructure, an oil official said on Saturday.	Predicted answer business	Predicted answer business
5	AFP - Tearaway world oil prices, toppling records and straining wallets, present a new economic menace barely three months before the US presidential elections.	Predicted answer business	Predicted answer business
6	Reuters - Stocks ended slightly higher on Friday\but stayed near lows for the year as oil prices surged past \$36\46\barrel, offsetting a positive outlook from computer maker\Dell Inc. (DELL.O)	Predicted answer world	Predicted answer business

Fig. 8: ML-Assisted Labeling compares labels from two different models

07 Directly Consumable Data Export Formats

Datasaur supports support multiple MLOps platforms' data formats in our data export, enabling data scientists to immediately use the labeled data for model training with no post-processing required.



Export 1 project

Export as *

untitled_output_Aug_29_2023 .zip

Format *

Datasaur Schema (.json)

Datasaur Schema (.json)

Row: Comma-separated values (.csv)

Row: Tab-separated values (.tsv)

Row: Microsoft Excel (.xlsx)

Row: JSON Tabular (.json)

Row: Amazon Comprehend (.csv)

Row: Azure AutoML (.csv)

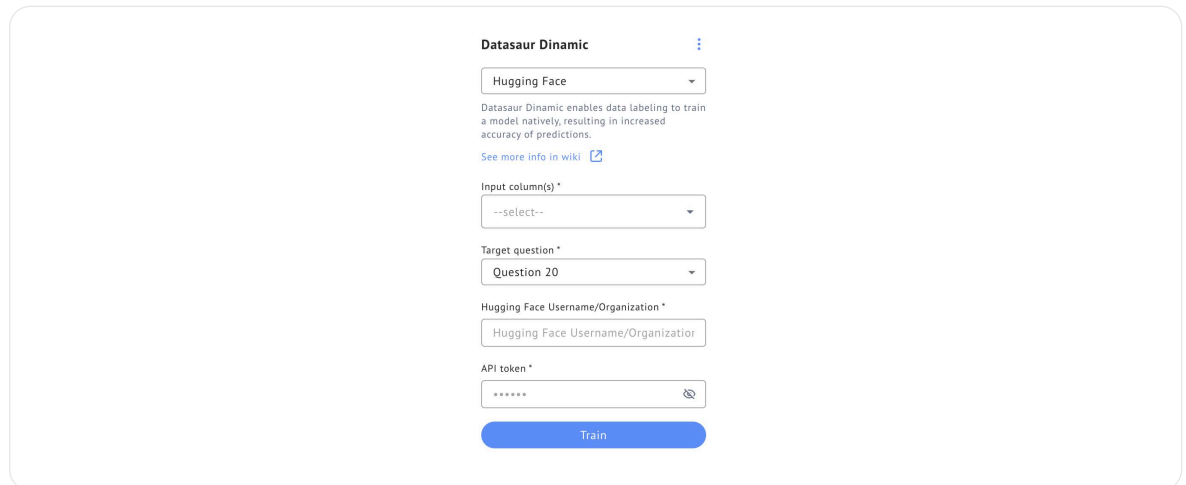
Row: GCP Vertex AI (.csv)

Row: Hugging Face (.csv)

Fig. 9: We support exported csv in Amazon Comprehend, Azure AutoML, and GCP Vertex AI format

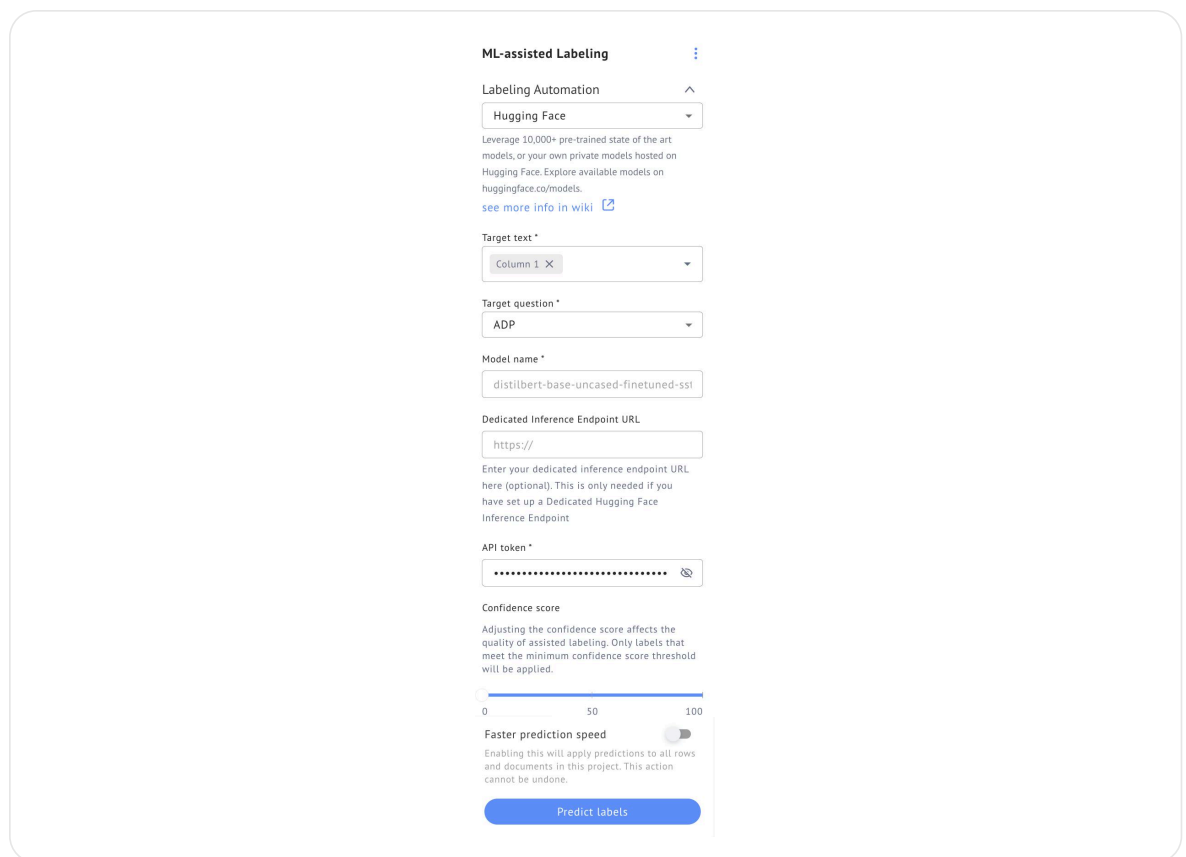
08 Datasaur Dinamic

[Datasaur Dinamic](#) is an extension to integrate the labeling process with training providers. It leverages powerful autoML solutions such as Hugging Face Auto Train and AWS Sagemaker to train models off of existing labels. These models can be immediately deployed to production or used to predict labels for upcoming new datasets.



The screenshot shows the 'Datasaur Dinamic' configuration panel. It features a dropdown menu set to 'Hugging Face'. Below this, a text box explains that Datasaur Dinamic enables data labeling to train a model natively, resulting in increased accuracy of predictions, with a link to 'See more info in wiki'. The 'Input column(s) *' dropdown is set to '--select--'. The 'Target question *' dropdown is set to 'Question 20'. The 'Hugging Face Username/Organization *' text field contains 'Hugging Face Username/Organizator'. The 'API token *' field is masked with asterisks and has a copy icon. A blue 'Train' button is at the bottom.

Fig. 10A: Datasaur Dinamic extension



The screenshot shows the 'ML-assisted Labeling' configuration panel. It features a dropdown menu set to 'Hugging Face'. Below this, a text box explains that it leverages 10,000+ pre-trained state-of-the-art models, or your own private models hosted on Hugging Face, with a link to 'see more info in wiki'. The 'Target text *' dropdown is set to 'Column 1 X'. The 'Target question *' dropdown is set to 'ADP'. The 'Model name *' text field contains 'distilbert-base-uncased-finetuned-sst'. The 'Dedicated Inference Endpoint URL' text field contains 'https://'. Below this, a text box explains that the dedicated inference endpoint URL is optional and only needed if you have set up a Dedicated Hugging Face Inference Endpoint. The 'API token *' field is masked with asterisks and has a copy icon. A 'Confidence score' section includes a slider from 0 to 100. Below the slider, a 'Faster prediction speed' toggle is turned on, with a text box explaining that enabling it will apply predictions to all rows and documents in the project, which cannot be undone. A blue 'Predict Labels' button is at the bottom.

Fig. 10B: ML-Assisted extension



09 Summary

MLOps has long been the foundation for developing updatable machine learning systems. It has evolved from merely logging model hyper parameters (model-centric) to a comprehensive focus on tracking and maintaining data quality throughout every phase of ML development (data-centric). With the adoption of data-centric MLOps, the process of upgrading and debugging ML models has become more interpretable.

In practical terms, the key to leveraging data-centric MLOps is having access to a tool that can consistently provide and maintain high-quality datasets. We have explored how Datasaur can play a pivotal role in supporting data-centric MLOps. This support is facilitated through our labeling tools and seamless integration with leading MLOps platforms like AWS Comprehend, GCP Vertex AI, and Azure ML.

Datasaur leads the charge in advancing Machine Learning Operations (MLOps) towards a data-centric focus. Through seamless integration with prominent MLOps platforms like AWS Comprehend, GCP Vertex AI, and Azure, the Datasaur platform empowers users with a suite of tools facilitating effortless dataset analysis, evaluation, and curation.

A commitment to technology stack flexibility allows users to effortlessly build models through Datasaur Dinamic or through a preferred integrated MLOps platform via native data export formats. Datasaur is dedicated to simplifying and enhancing the MLOps journey, placing data quality and efficiency at the forefront of every cycle.

About Datasaur

Datasaur is a private LLM provider and data labeling platform designed for companies to build their AI ecosystem with ease and efficiency. It assists organizations and universities in setting up custom LLMs and annotating data more efficiently and accurately through automation, quality control, and human-in-the-loop workflows. For more information, visit www.datasaur.ai.

Schedule a demo