



# Custom Model Building with AutoML

# Custom Model Building with AutoML

01

Why AutoML?

02

AutoML Vision

03

AutoML Natural Language

04

AutoML Tables

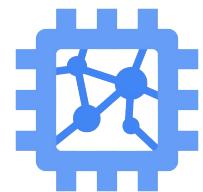


# Custom Model Building with AutoML

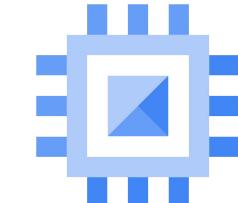
- 01 Why AutoML?
- 02 AutoML Vision
- 03 AutoML Natural Language
- 04 AutoML Tables



# Create and deploy custom models with AutoML



Cloud TPUs



Compute Engine



Dataproc



Google Kubernetes Engine

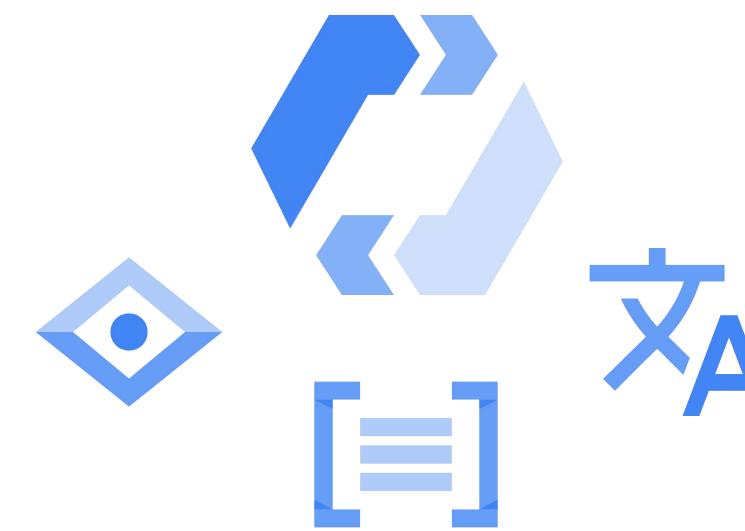


Vertex AI



BigQuery ML

## AutoML

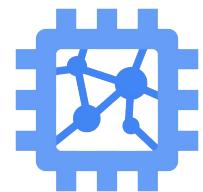
[Build a Custom Model](#)[Build Custom Model  
\(codeless\)](#)Cloud  
Translation API

Vision API

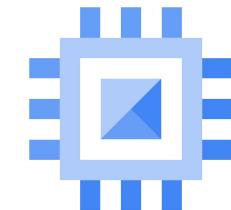
Speech-to-Text  
APIVideo  
Intelligence APIData Loss  
Prevention APIText-to-Speech  
APICloud Natural  
Language API

Dialogflow

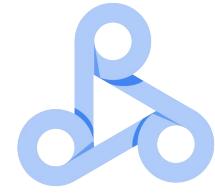
# Create and deploy custom models with AutoML



Cloud TPUs



Compute Engine



Dataproc



Google Kubernetes Engine

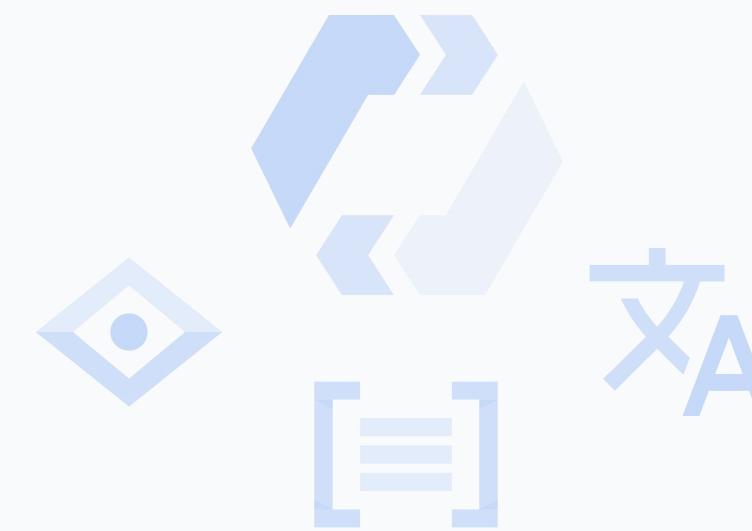


Vertex AI



BigQuery ML

## AutoML



Build a Custom Model

Build Custom Model  
(codeless)

Call a Pretrained Model

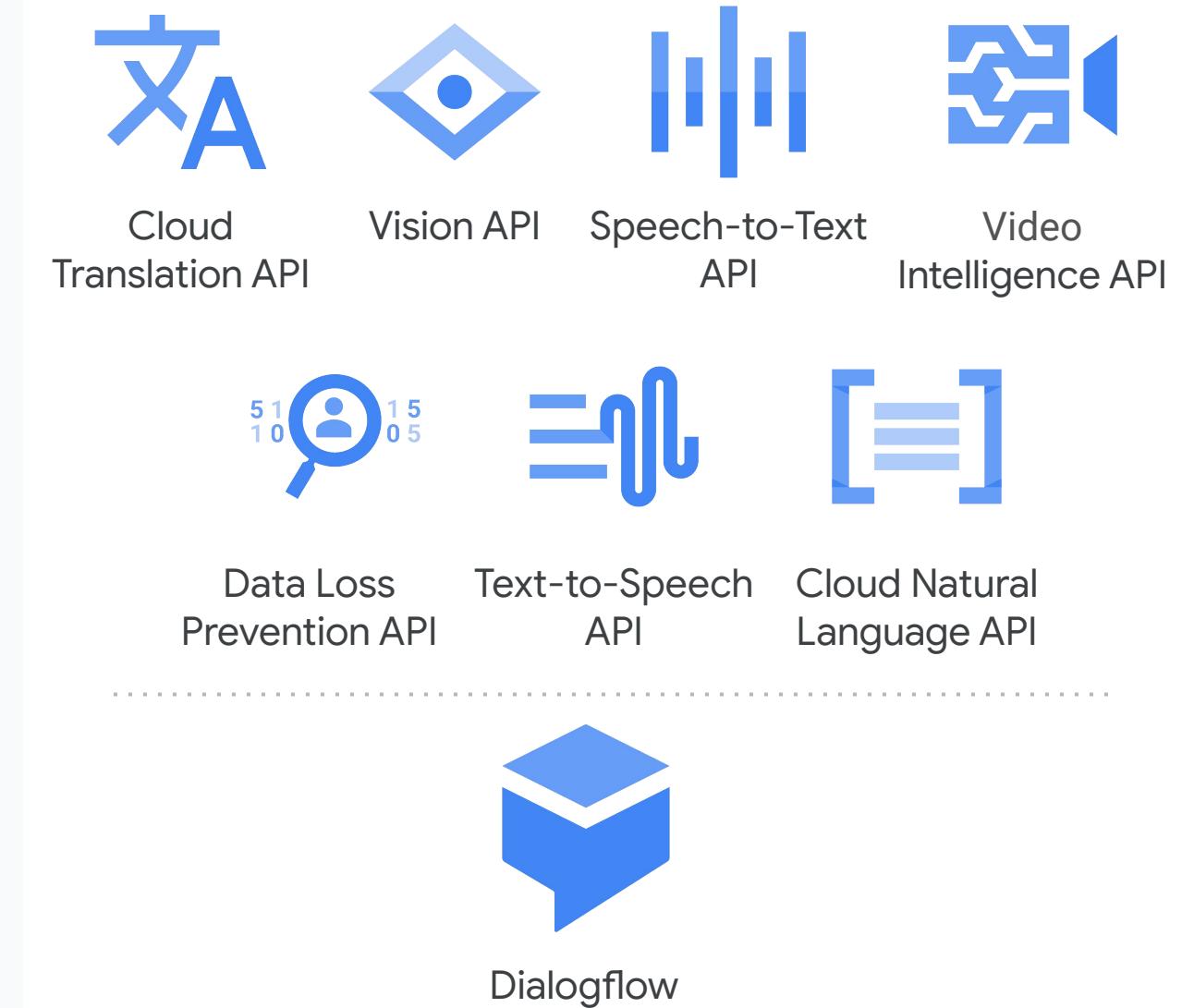
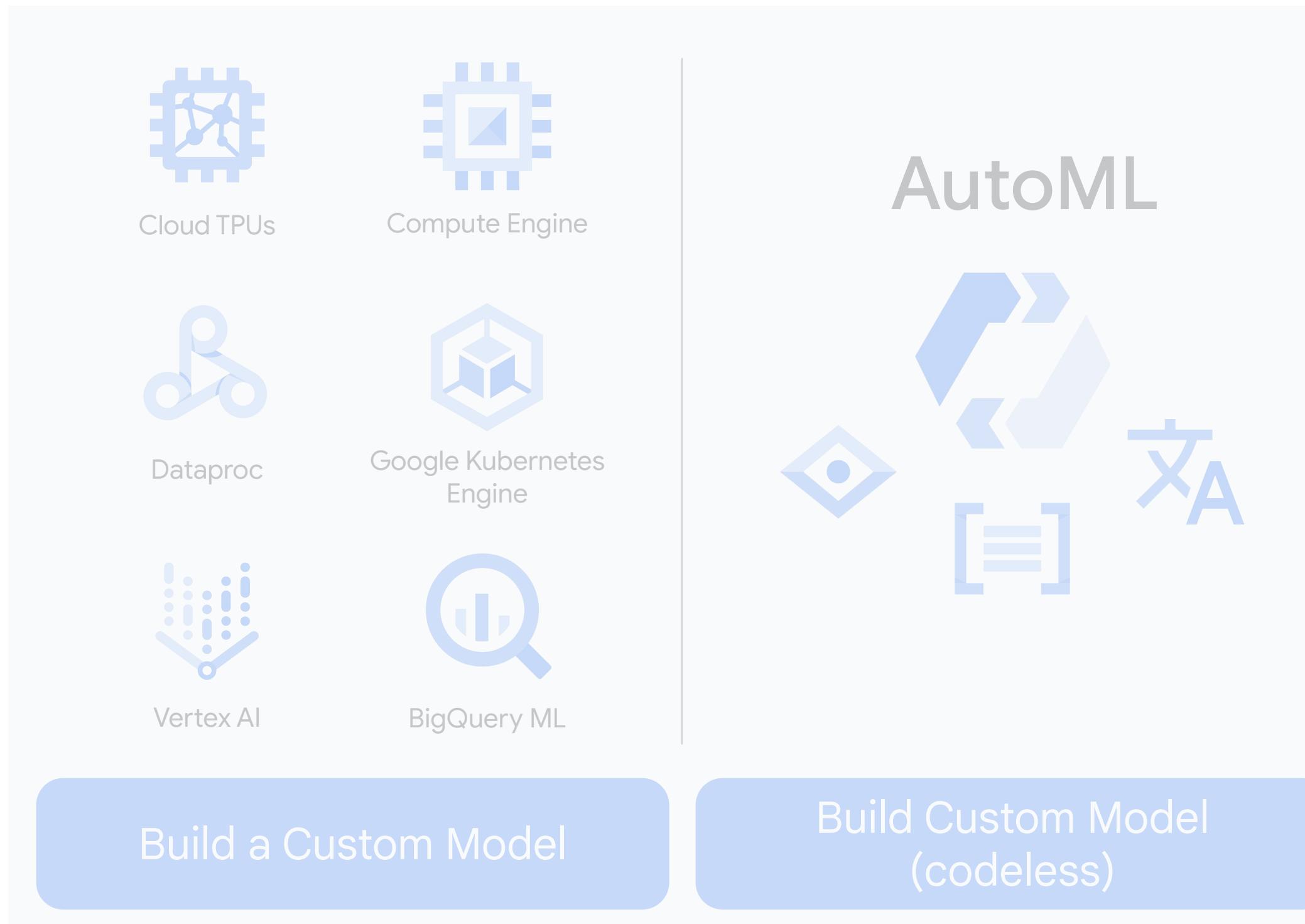
Cloud  
Translation API

Vision API

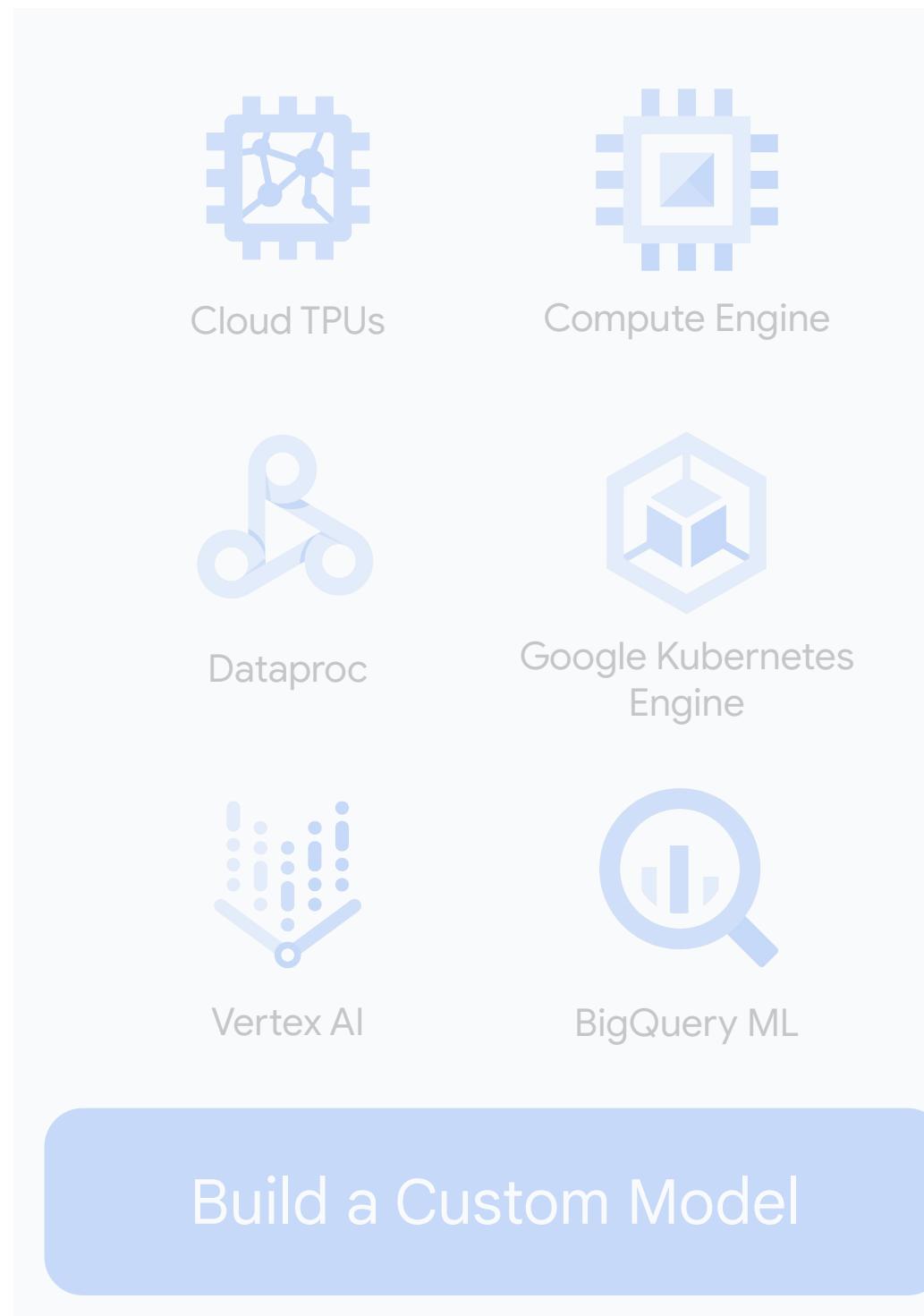
Speech-to-Text  
APIVideo  
Intelligence APIData Loss  
Prevention APIText-to-Speech  
APICloud Natural  
Language API

Dialogflow

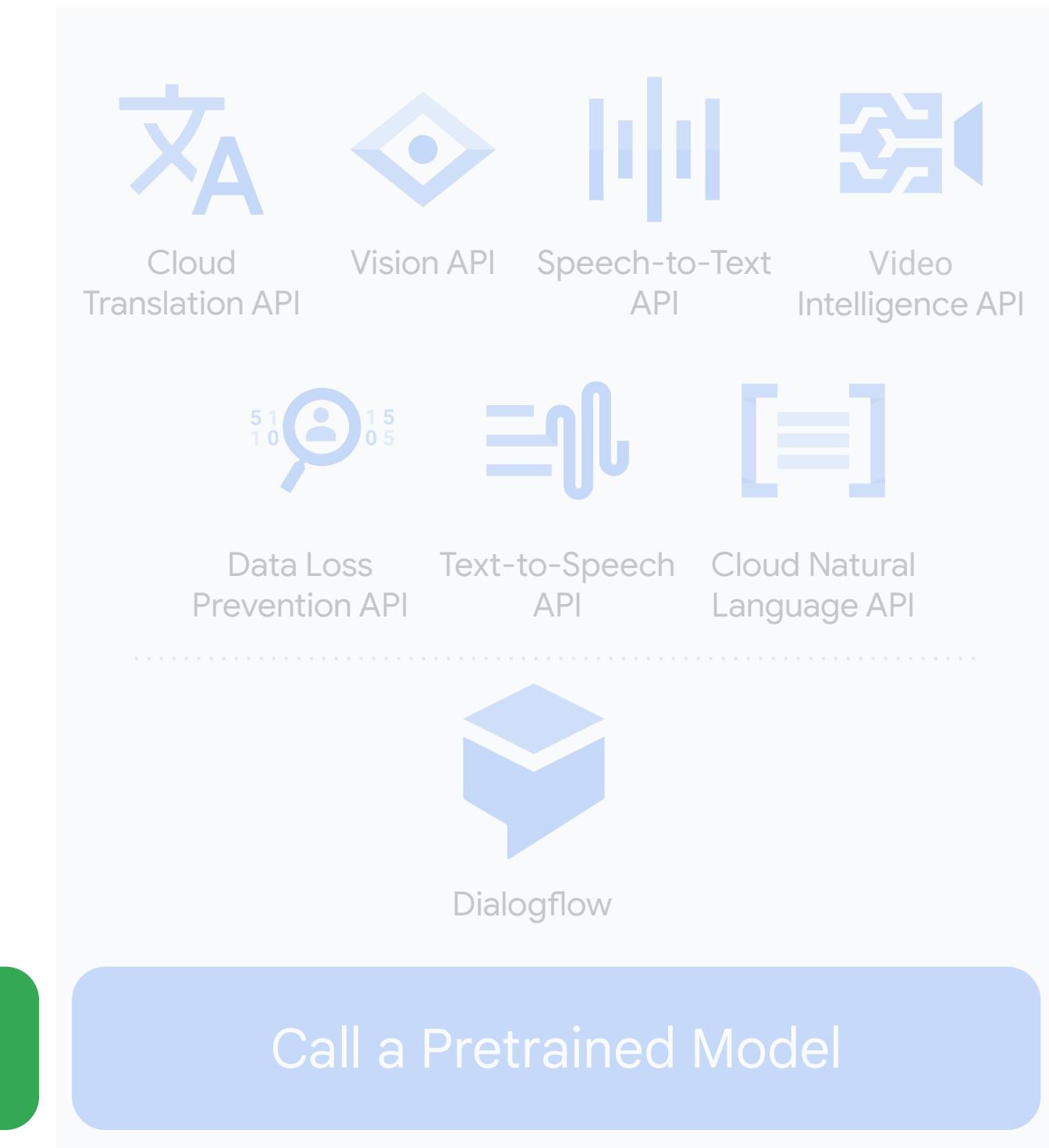
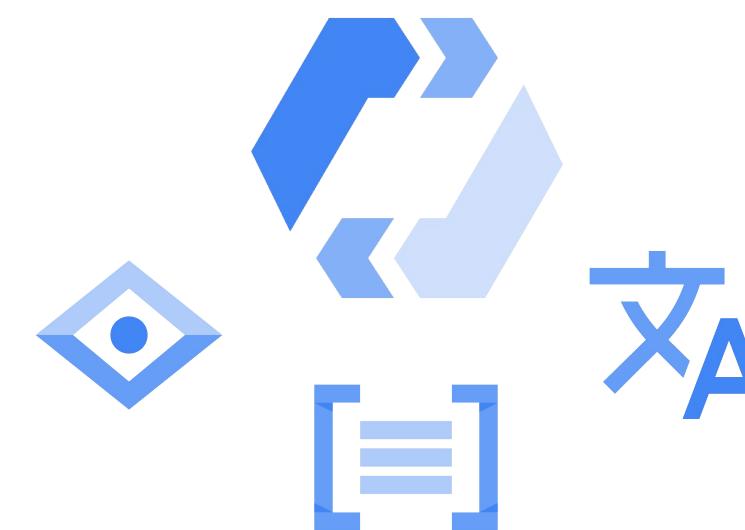
# Create and deploy custom models with AutoML



# Create and deploy custom models with AutoML



## AutoML

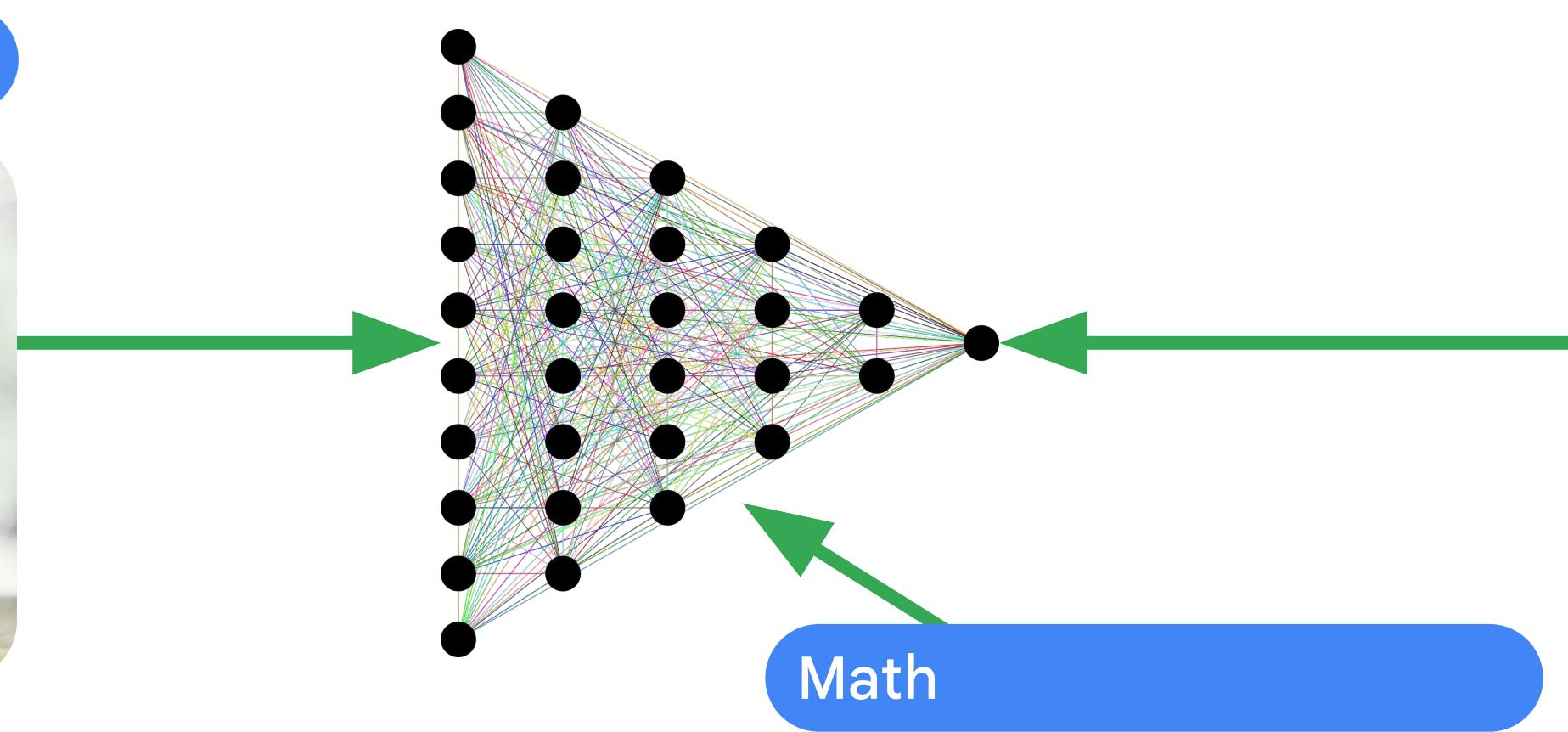


Build a Custom Model

Build Custom Model  
(codeless)

Call a Pretrained Model

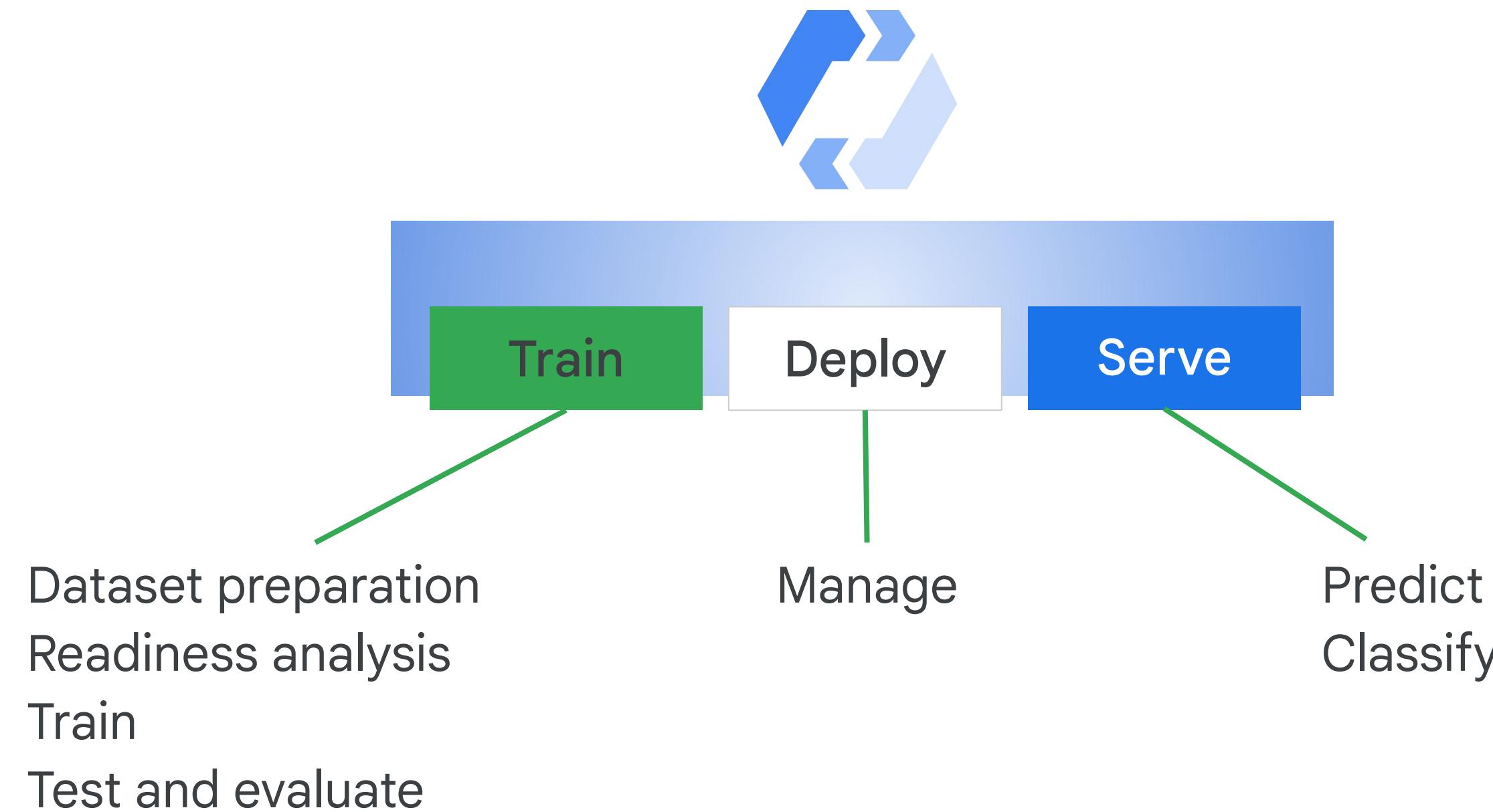
# Training high-quality, custom ML models requires a lot of effort and expertise



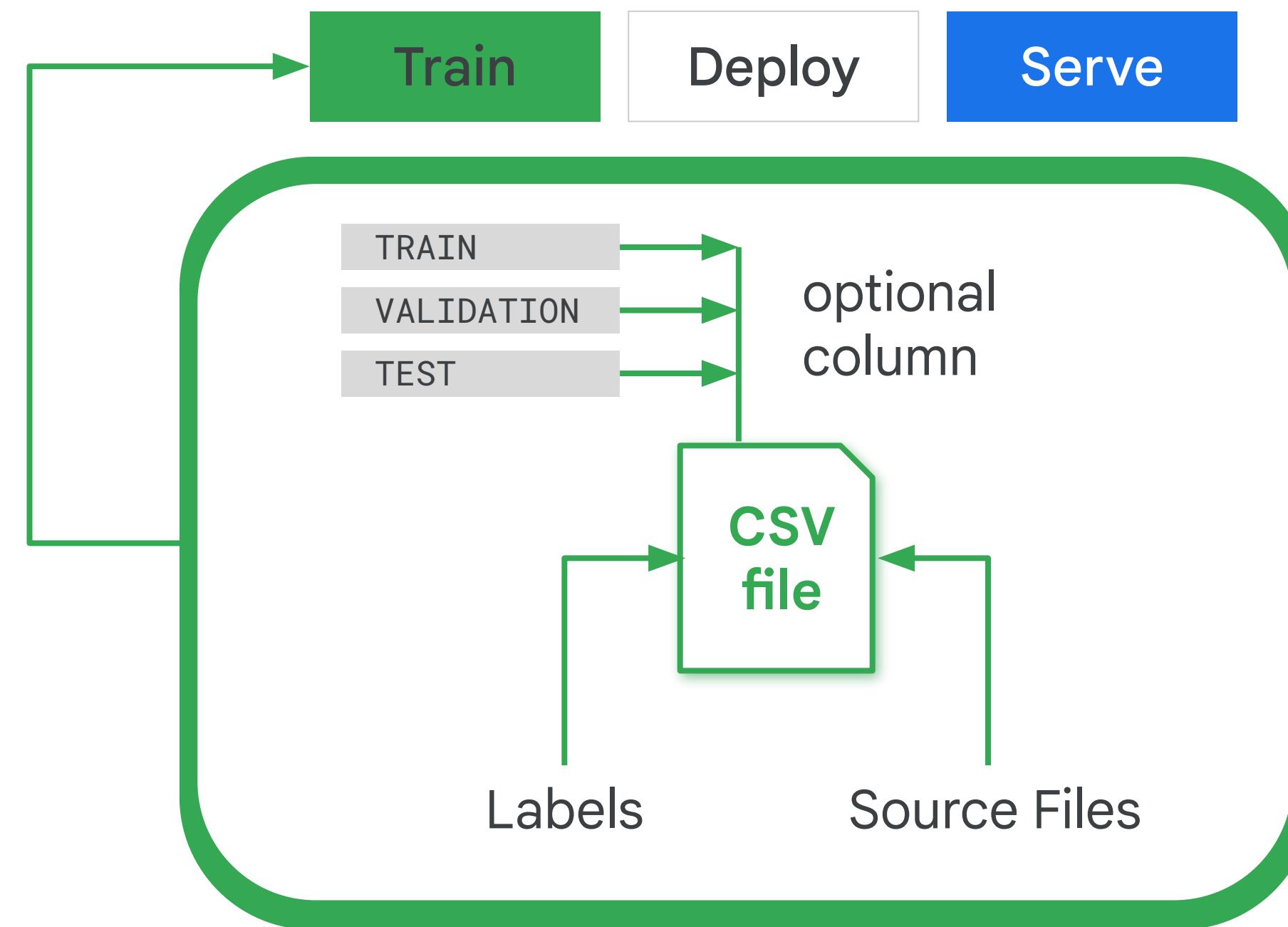
Math

$$\begin{aligned} F : I \rightarrow \mathbb{R}, \quad x \mapsto \int_a^x f(t) dt \\ \int_a^b f(x) dx = F(b) - F(a) \end{aligned}$$

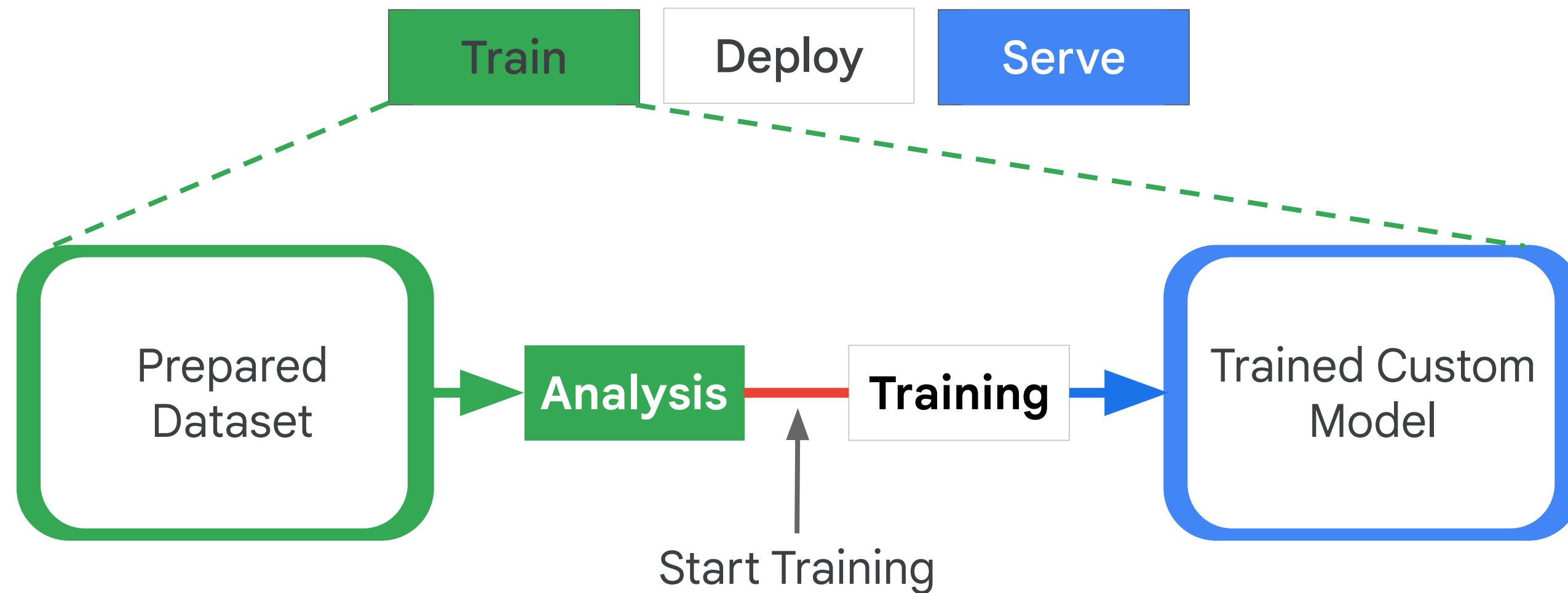
# AutoML follows a standard procedure that is divided into train, deploy, and serve phases



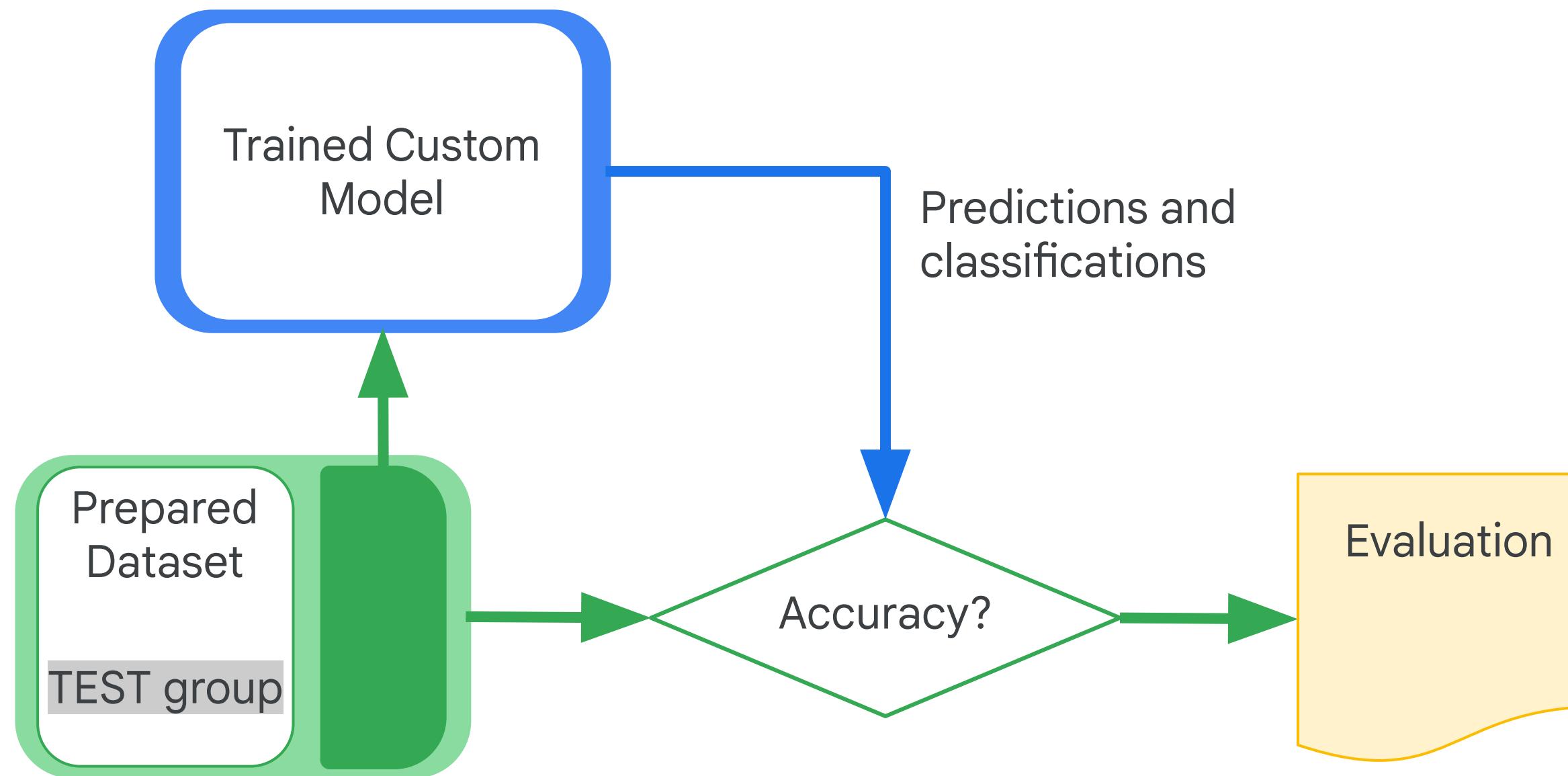
# AutoML uses a Prepared Dataset to train a Custom Model



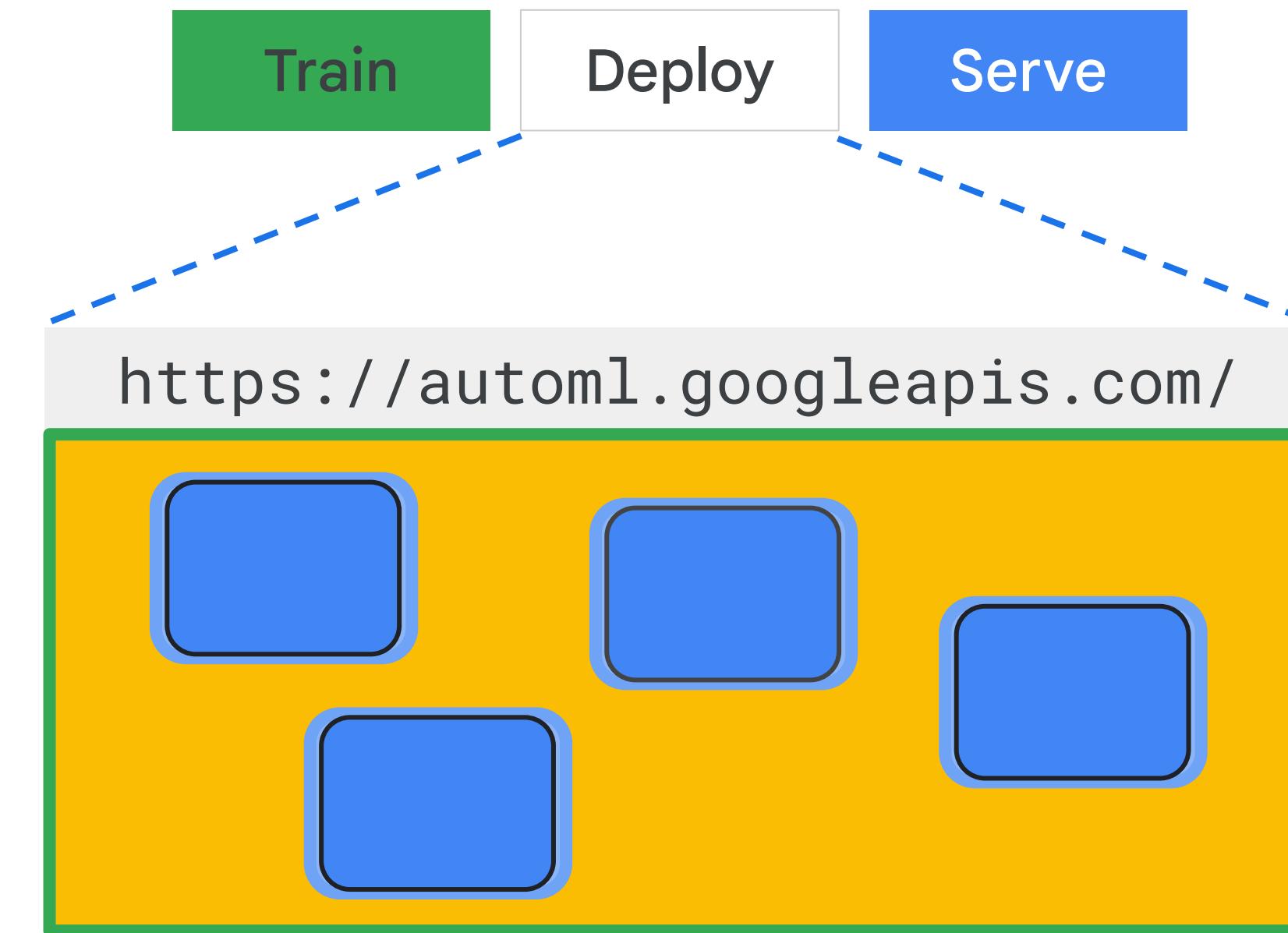
AutoML performs basic checks and a preliminary analysis of the Prepared Dataset to determine if there is enough information and if it is properly organized



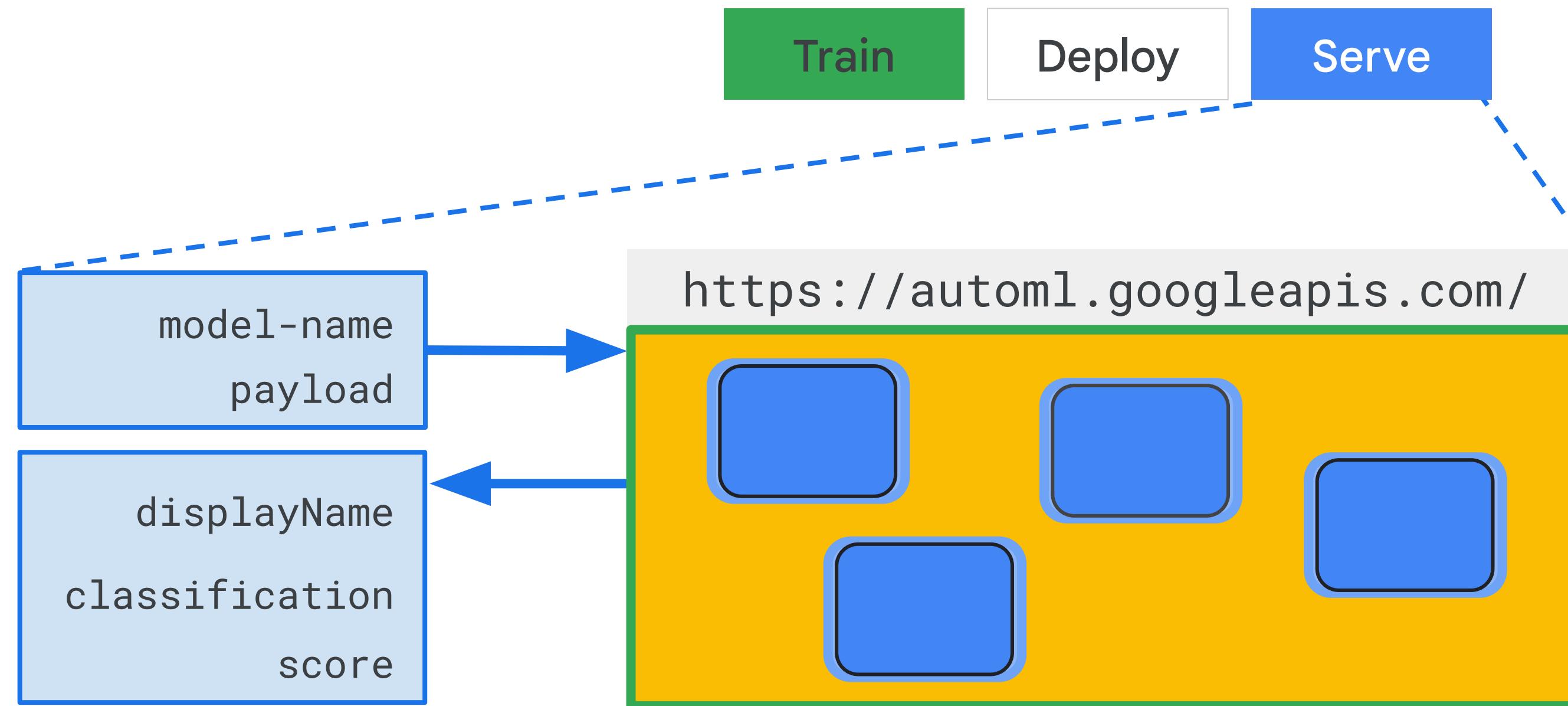
# Data from the TEST group is used to evaluate the Custom Model and to remove bias from the evaluation



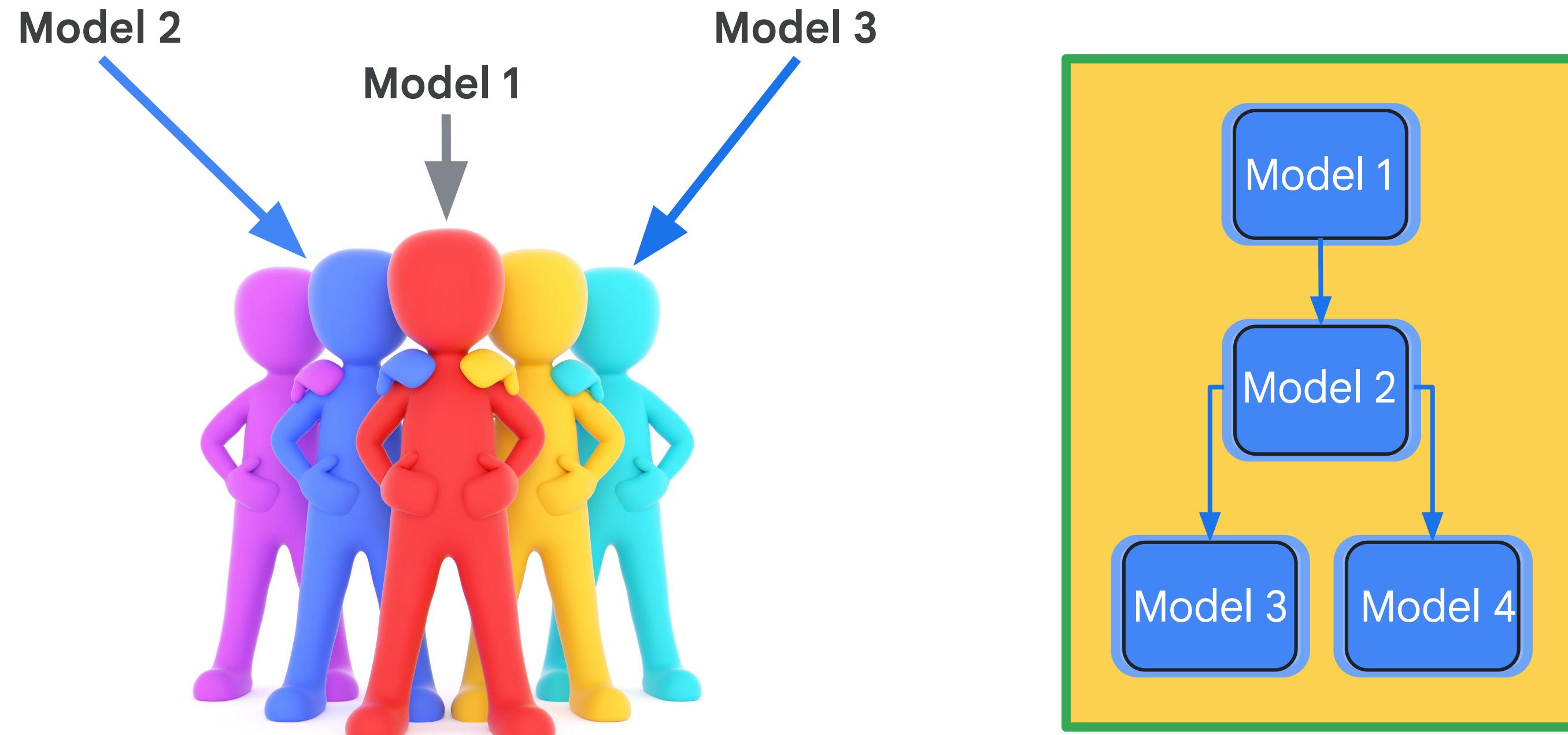
# There is nothing you need to do to deploy a trained model



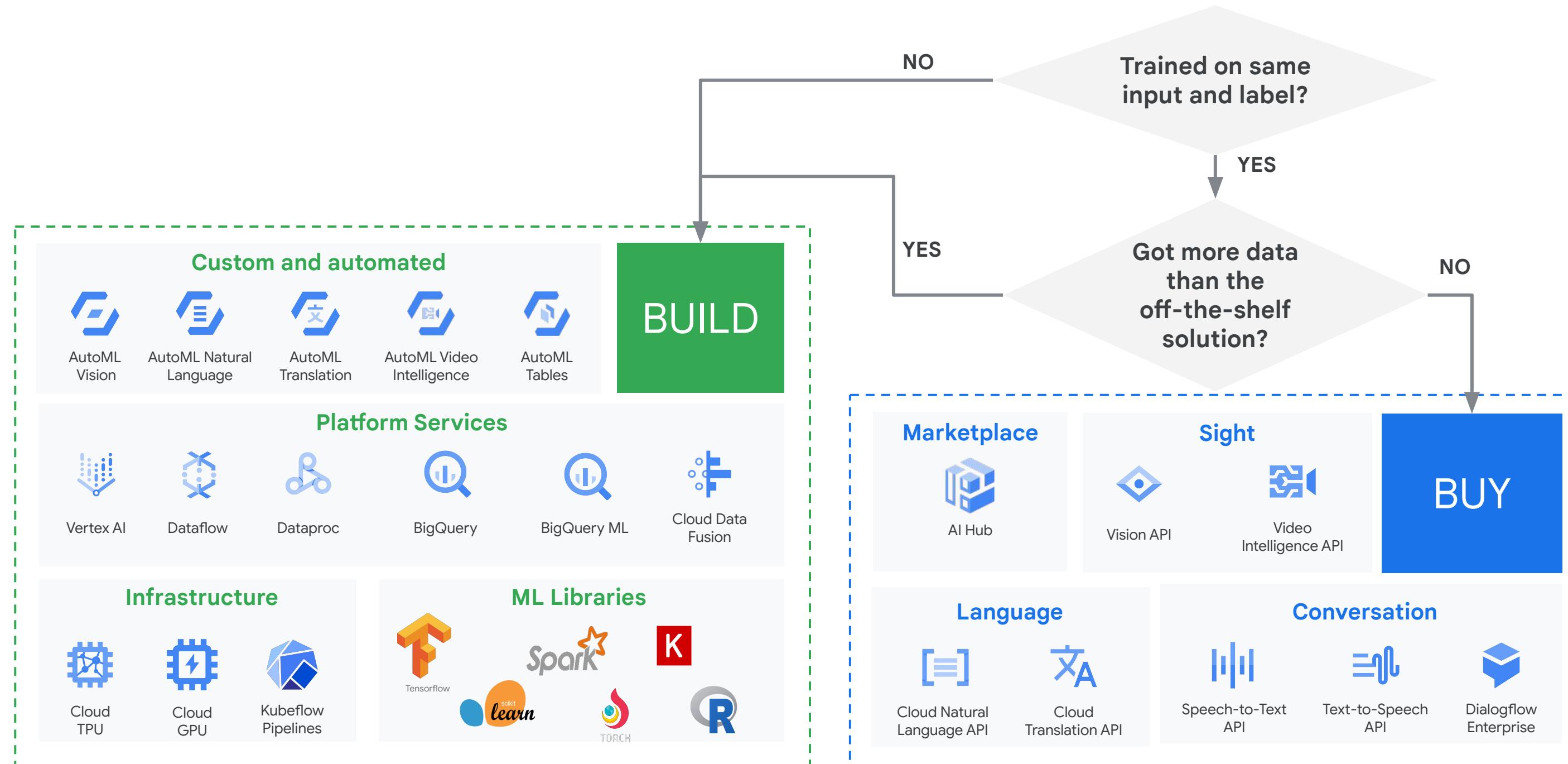
# Serve models using the Web UI, or from the command line using CURL to send a JSON-structured request



# Break up complicated problem into multiple models



# As a data engineer should you build or buy a solution?



# Custom Model Building with AutoML

01

Why AutoML?

02

**AutoML Vision**

03

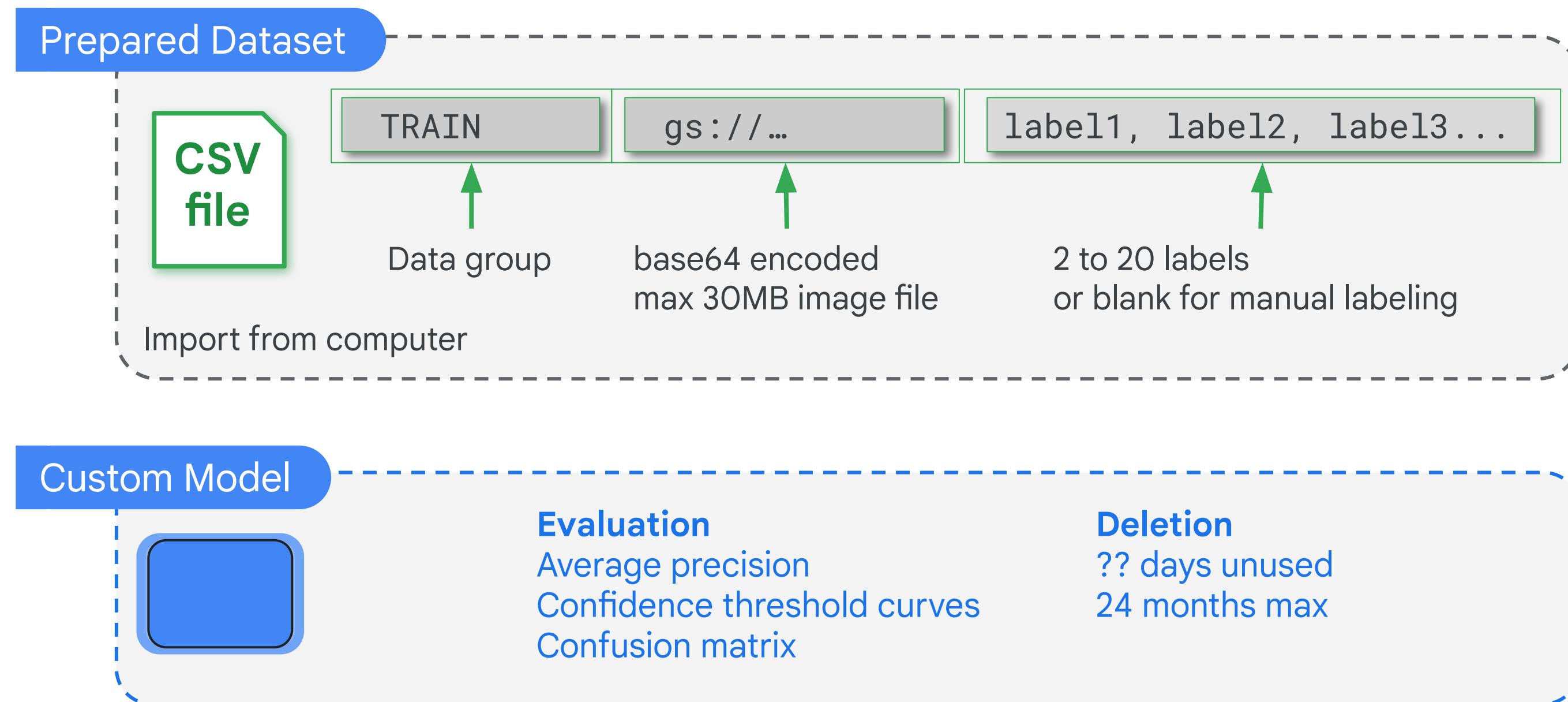
AutoML Natural Language

04

AutoML Tables



# AutoML Vision specializes in training models for image classification



# Improving Vision Custom Models



Train on examples similar to those you will classify

Low scores:  
Increase data

Perfect scores:  
Increase variety



- Verify labels are used consistently
- 100x images for most common labels than the least common labels
- Remove infrequently used labels

# Custom Model Building with AutoML

01

Why AutoML?

02

AutoML Vision

03

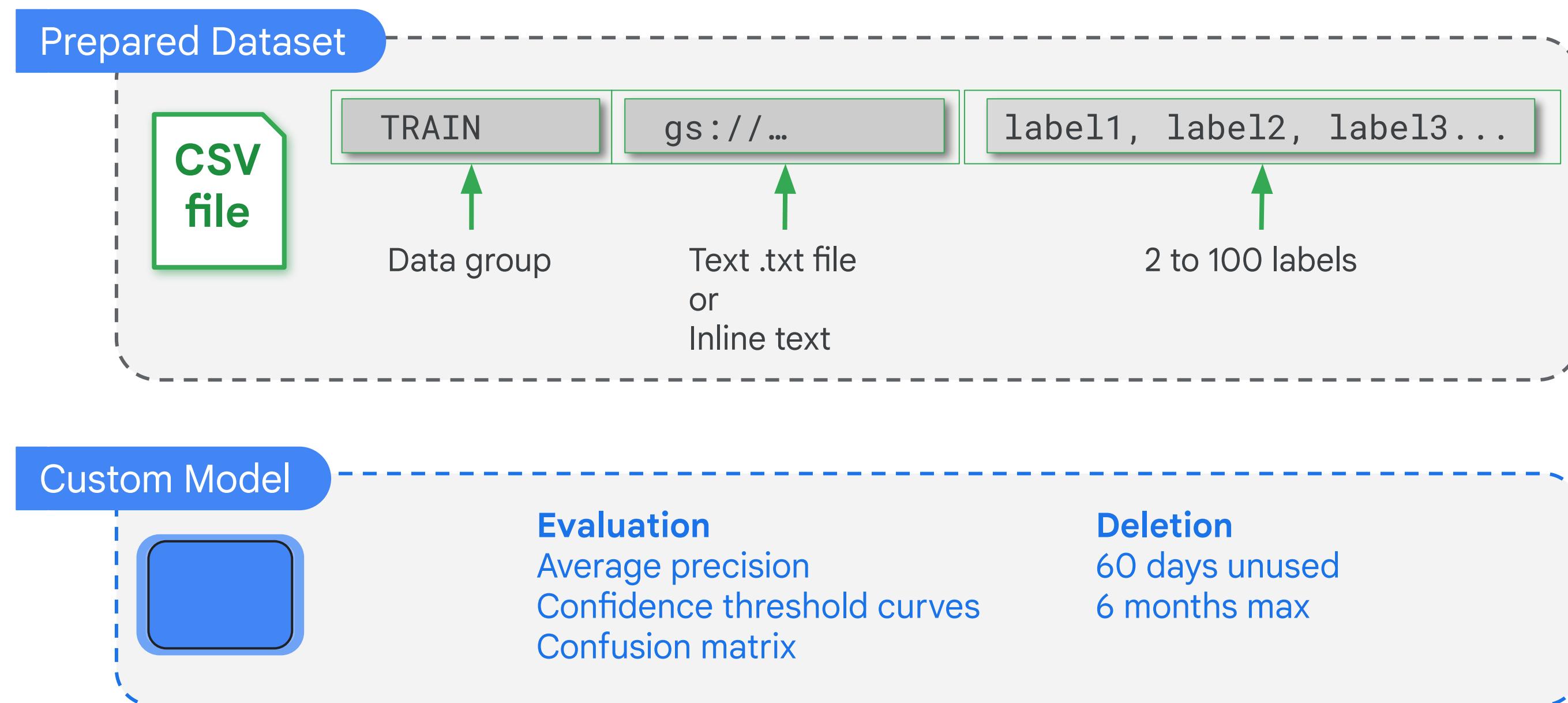
AutoML Natural Language

04

AutoML Tables



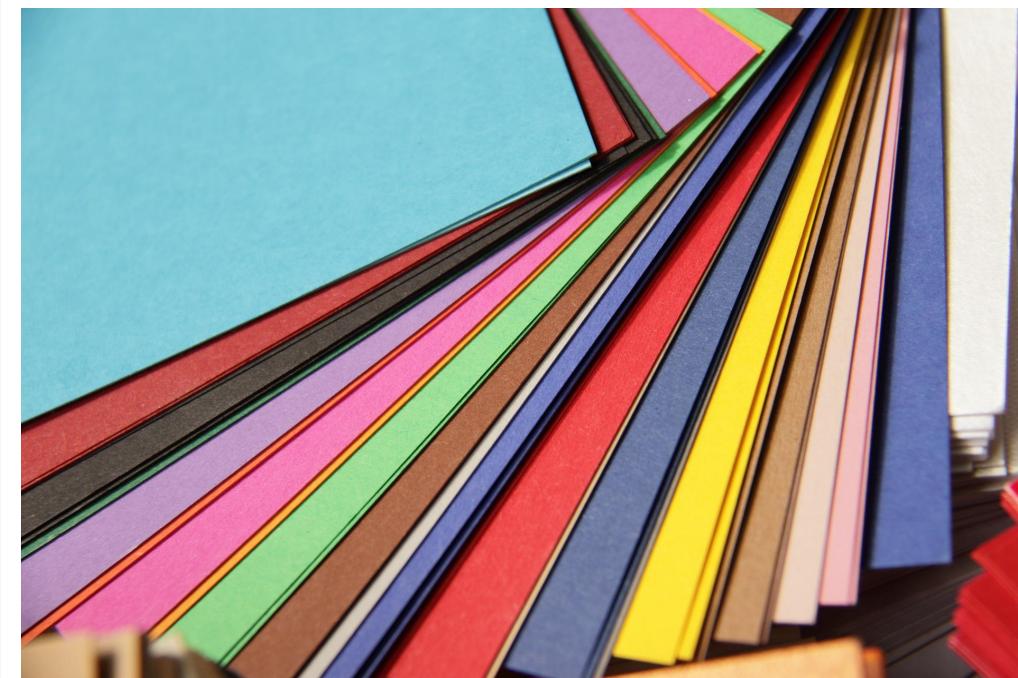
# AutoML Natural Language specializes in training models for text



# Improving Natural Language custom models



Add more documents



Increase document variety



Reduce the number of labels

# Custom Model Building with AutoML

01

Why AutoML?

02

AutoML Vision

03

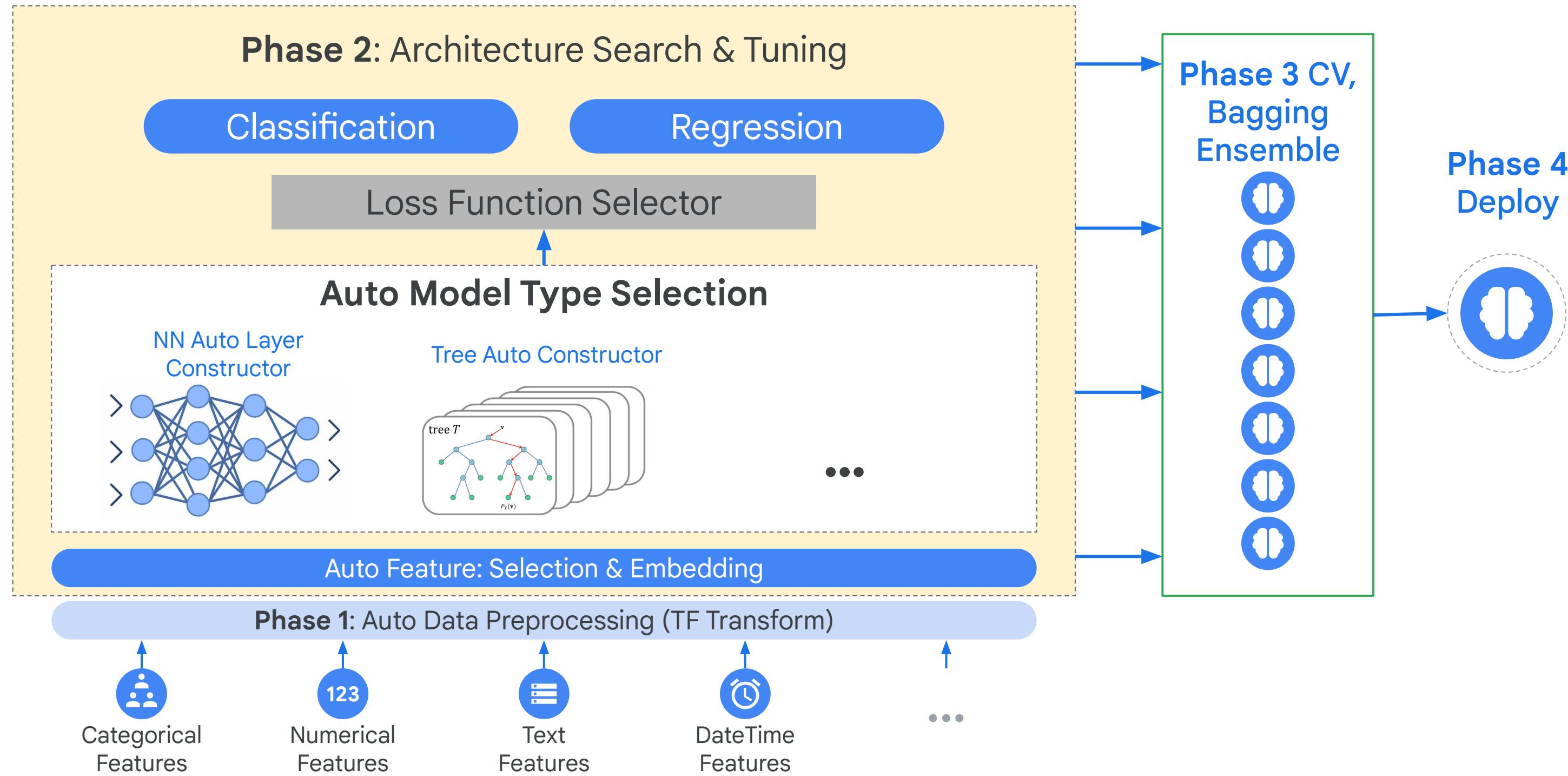
AutoML Natural Language

04

AutoML Tables



# AutoML Table is for structured data

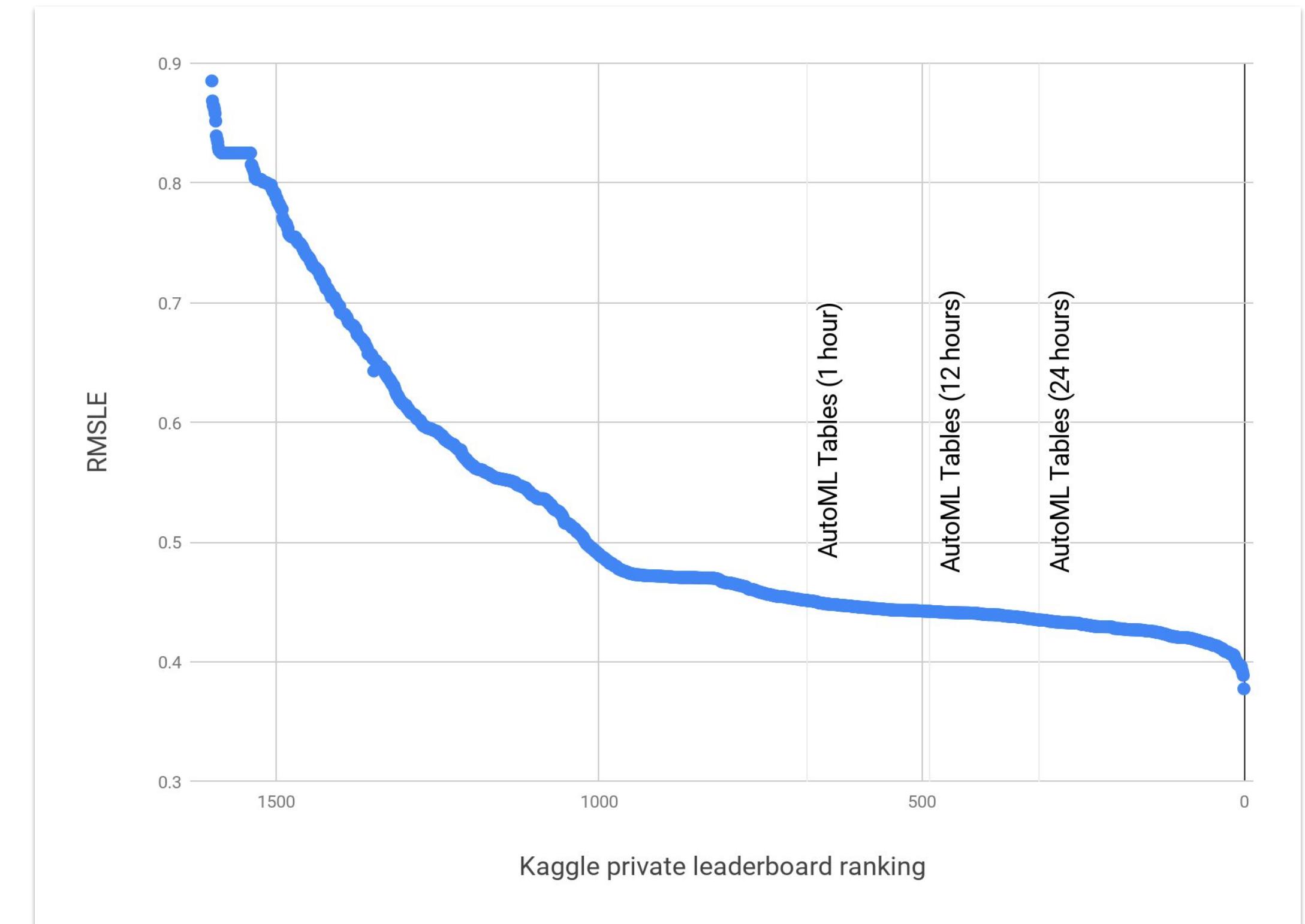


# Example: Mercari price suggestion challenge

Goal: Automatically suggest product prices to online sellers

Training data							
ID	Name	Item Condition	Categories	Brand name	Shipping	Item description	Price
0	MLB Cincinnati Reds T Shirt Size XL	3	Men, Tops, T-shirts		1	No description yet.	\$10
1	Razer BlackWidow Chroma Keyboard	3	Electronics, Computers & Tablets, Components & Parts	Razer	0	This keyboard is in great condition and works like it came out of the box. All of the ports are tested and work perfectly. The lights are customizable via the Razer Synapse app on your PC.	\$52
2	AVA-VIV Blouse	1	Women, Tops & Blouses, Blouse	Target	1	Adorable top with a hint of lace and a key hole in the back! The pale pink is a 1X, and I also have a 3X available in white!	\$10
3	Leather Horse Statues	1	Home, Home Décor, Home Décor Accents		1	New with tags. Leather horses. Retail for [rm] each. Stand about a foot high. They are being sold as a pair. Any questions please ask. Free shipping. Just got out of storage.	\$35

**AutoML Tables  
produced some  
of the best results  
on the challenge**



The easiest way  
to import data  
into AutoML  
  
Tables is through  
BigQuery

**IMPORT** SCHEMA ANALYZE TRAIN EVALUATE PREDICT

**Import your data**

AutoML Tables uses tabular data that you import to train a custom machine learning model. Your dataset must contain at least one input feature column and a target column. Optional columns can be added to configure parameters like the data split, weights, etc. [Preparing your training data](#)

**Table from BigQuery**  
The table must be in the US regional location

BigQuery project ID \*

BigQuery dataset ID \*

BigQuery table ID \*

**CSV from Cloud Storage**  
The bucket containing the CSV must be in the us-central1 region. [CSV formatting](#)

gs:// BROWSE

IMPORT ?

# Start by setting the features/label that will be used for training

Column name	Variable type	Nullability
Age	Numeric	Nullable
Job	Categorical	Nullable
MaritalStatus	Categorical	Nullable
Education	Categorical	Nullable
Default	Categorical	Nullable
Balance	Numeric	Nullable
Housing	Categorical	Nullable
Loan	Categorical	Nullable
Contact	Categorical	Nullable
Day	Categorical	Nullable
Month	Categorical	Nullable
Duration	Numeric	Nullable
Campaign	Categorical	Nullable
PDays	Numeric	Nullable
Previous	Numeric	Nullable
POutcome	Categorical	Nullable
<input checked="" type="checkbox"/> Deposit	Target	Categorical

# Next, do some data validation to ensure you're not passing junk into your model

**IMPORT   SCHEMA   ANALYZE   TRAIN   EVALUATE   PREDICT**

**⚠️ Not up to date. Click the "Continue" button on the Schema tab to regenerate statistics.**

		Filter instances						
		Feature name ↑	Type	Missing ?	Distinct values ?	Correlation with Target ?	Mean ?	
All features	Numeric	Age	Numeric	0%	77	0.065	40.936	
		Balance	Numeric	0%	7,168	0.095	1,362.272	
		Categorical	Campaign	Categorical	0%	48	0.083	---
			Contact	Categorical	0%	3	0.144	---
			Day	Categorical	0%	31	0.122	---
	Default		Categorical	0%	2	0.028	---	
	Deposit		Categorical	0%	2	---	---	
	Duration		Numeric	0%	1,573	0.333	258.163	
	Education		Categorical	0%	4	0.071	---	
	Housing		Categorical	0%	2	0.117	---	
	Job		Categorical	0%	12	0.134	---	
	Loan		Categorical	0%	2	0.073	---	
	MaritalStatus		Categorical	0%	3	0.059	---	
	Month		Categorical	0%	12	0.245	---	
	PDays	Numeric	0%	559	0.181	40.198		
	POutcome	Categorical	0%	4	0.313	---		
	Previous	Numeric	0%	41	0.181	0.58		

Rows per page: 50 ▾ 1 – 17 of 17 < >

**Details** X

**Distribution**

cellular (29285)  
unknown (13020)  
telephone (2906)

28.8%  
64.8%

**Top correlated features to Contact**

Month  
Housing  
Day  
POutcome  
Previous  
Job  
Education  
Age  
Campaign

Carry out some experiments in BigQuery ML to set some base metrics for model performance

You can allocate a budget when training the model

**Train your model**

**Model name \***  
banking\_20190410095716

**Training budget**  
Enter a number between 1 and 72 for the maximum number of node hours to spend training your model. If your model stops improving before then, AutoML Tables will stop training and you'll only be charged for the actual node hours used. [Training pricing guide](#)

Budget \* maximum node hours ?

**Input feature selection**  
By default, all other columns in your dataset will be used as input features for training (excluding target, weight, and split columns).

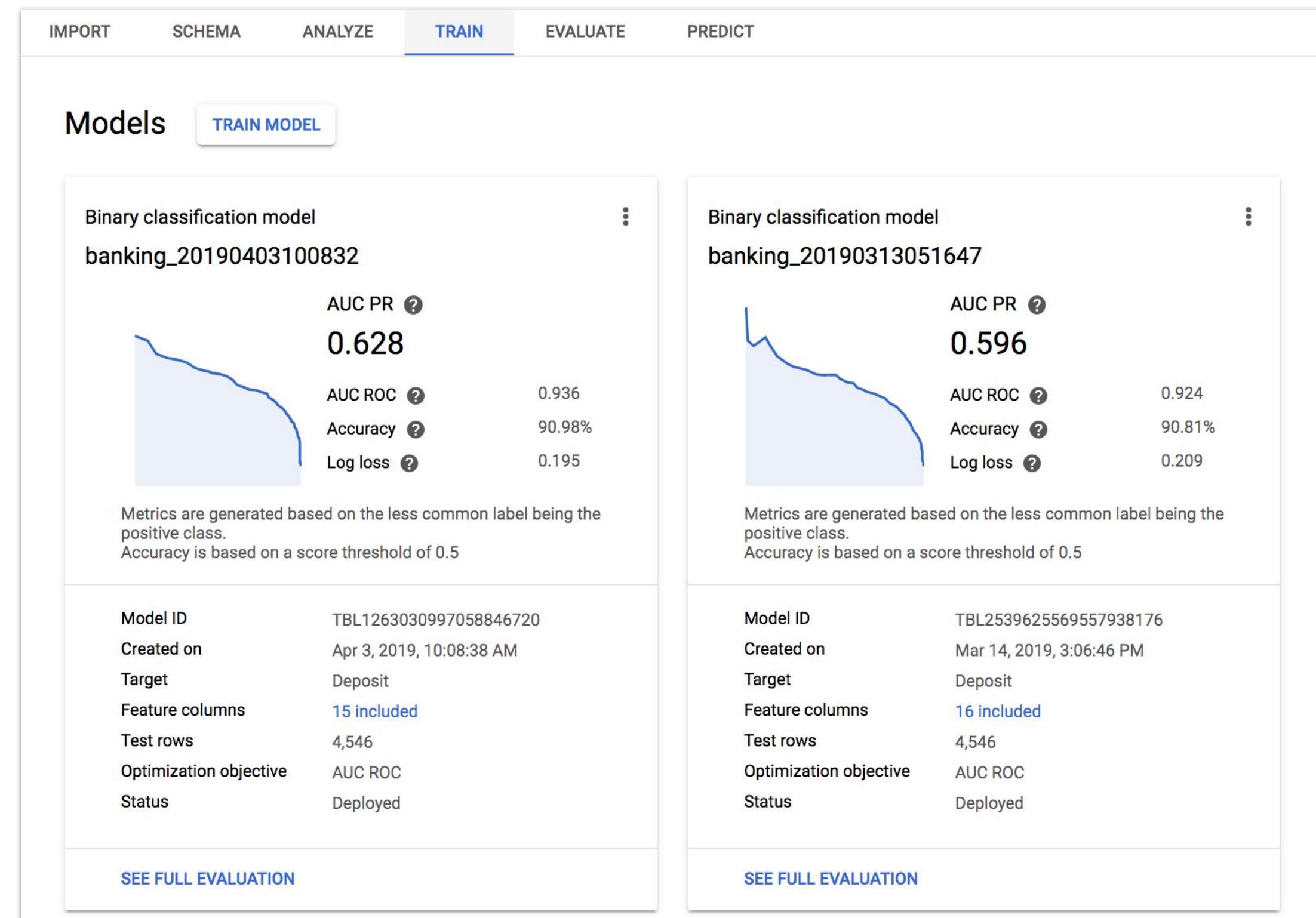
16 feature columns \*  
All columns selected ▾

**Summary**  
Model type: Binary classification model  
Data split: Automatic  
Target: Deposit  
Input features: 16 features  
Rows: 45,211 rows

**Optimization objective ▾**  
Depending on the outcome you're trying to achieve, you may want to train your model to optimize for a different objective. [Learn more](#)

TRAIN MODEL CANCEL

# Inspect the training metrics across multiple models



# Check how model performs against test data to gauge how well it will generalize in the wild

**EVALUATE**

Model: banking\_20190403100832

Binary classification model  
Apr 3, 2019, 10:08:38 AM

Target	Feature columns	Optimized for	AUC PR	AUC ROC	Accuracy	Log loss
Deposit	15 included 4,546 test rows	AUC ROC	0.628	0.936	91.0%	0.195

Metrics are generated using the least-common class as the positive class. Accuracy based on score threshold of 0.5

→ EXPORT PREDICTIONS ON TEST DATASET TO BIGQUERY

You have up to 30 days to export your test dataset to BigQuery

Filter labels: 2

1	Score threshold: 0.50
2	F1 score: 0.557 Accuracy: 91.0% (4,136/4,546) Precision: 64.3% (258/401) True positive rate (Recall): 49.1% (258/525) False positive rate: 0.036 (143/4,021)

The score threshold determines the minimum level of confidence needed to make a prediction positive. [Learn more about model evaluation](#)

Precision vs Recall: AUC: 0.628

True positive rate vs False positive rate: AUC: 0.936

# Integrate your trained model into your applications

The screenshot shows the Google Cloud AI Platform Predict interface. At the top, there are tabs: IMPORT, SCHEMA, ANALYZE, TRAIN, EVALUATE, and PREDICT. The PREDICT tab is selected. Below it, there are two buttons: BATCH PREDICTION and ONLINE PREDICTION, with ONLINE PREDICTION being the active one. A dropdown menu labeled 'Model' contains the entry 'banking\_20190403100832'. A success message states: 'Your model was deployed and is available for online prediction requests. Your model size is 1,131.127 MB. [Learn more](#)'.

**Test and use your model**

Online prediction deploys your model so you can send real-time REST requests to it. Online prediction is useful for time-sensitive predictions (for example, in response to an application request). [Learn more](#)

Online prediction pricing is based on the size of your model and the length of time your model is deployed. [View pricing guide](#)

Predict label	Prediction result
Deposit	1 Confidence score: 0.992
	2 Confidence score: 0.008

```
5 "values": [
6   "technician",
7   "married",
8   "secondary",
9   "no",
10  "52",
11  "no",
12  "no",
13  "cellular",
14  "12",
15  "aug",
16  "96",
17  "2"]
```

# How to choose between BigQuery ML, AutoML and a custom model

Model type	BigQuery ML	AutoML	Custom deep learning model
How	SQL in BigQuery for ML on structured data	AutoML uses neural architecture search and best-of-class model architectures for the specific problem	Keras with a TensorFlow backend, trained on Cloud ML Engine
Best if you are a	Data analyst who can wrangle data with SQL	Developer who can create the dataset in the required format	ML Engineer who knows Python and knows deep learning, NLP techniques
How long it takes an experienced practitioner	About an hour	About a day	A week to a month
Most of this time is spent in	Writing SQL	Waiting for job to finish	Coding Python and experimentation with ML
Cloud computing costs	Low	Medium	Medium to high depending on size of data, number of experiments, etc.
Accuracy	Moderate to high, mostly depending on the size of your dataset	High	Low if you don't know what you are doing; extremely high if you employ appropriate architectures and have a large-enough dataset

# Summary

- AutoML can be used to create powerful ML models without any coding.
- Use AutoML Vision when you have image data.
- Use AutoML Natural Language when you have text data.
- Use AutoML Tables when you have structured data.

