



# Advanced Big Data: Cloud Machine Learning

Dan Zaratsian

AI/ML Solutions Architect, Gaming @ Google

[d.zaratsian@gmail.com](mailto:d.zaratsian@gmail.com)

<https://github.com/zaratsian>

# Logistics

- Homework Assignments
- Consent to share students' GenAI use cases
- Upcoming Class Content
- Meeting after class



# GenAI on Google Cloud for Living Games

Some thoughts & insights from GDC  
(Game Developer Conference)



# Overview of GenAI on Google Cloud

Choose the GenAI model, infra, and supporting services that meet your requirements.

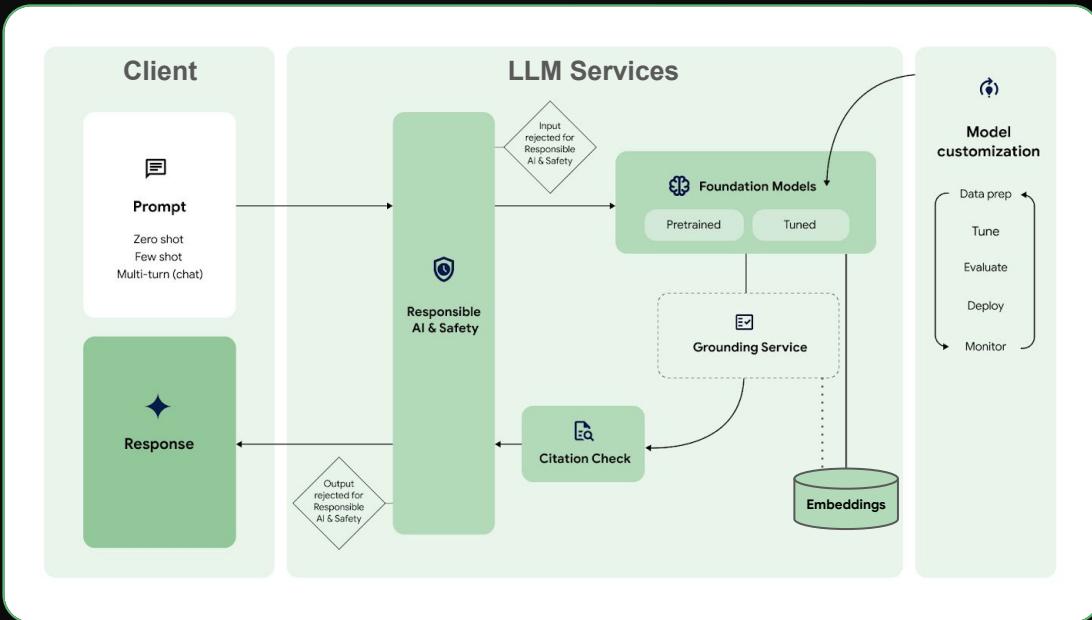
## Scalable Infra

## Google Serving Infrastructure

- ▷ Managed Endpoints
- ▷ One-click deploy
- ▷ Serving on GKE
- ▷ Run on GPUs, TPUs
- ▷ Llama 2
- ▷ Stable Diffusion

### 3rd Party:

- ▷ Anthropic Claude 2

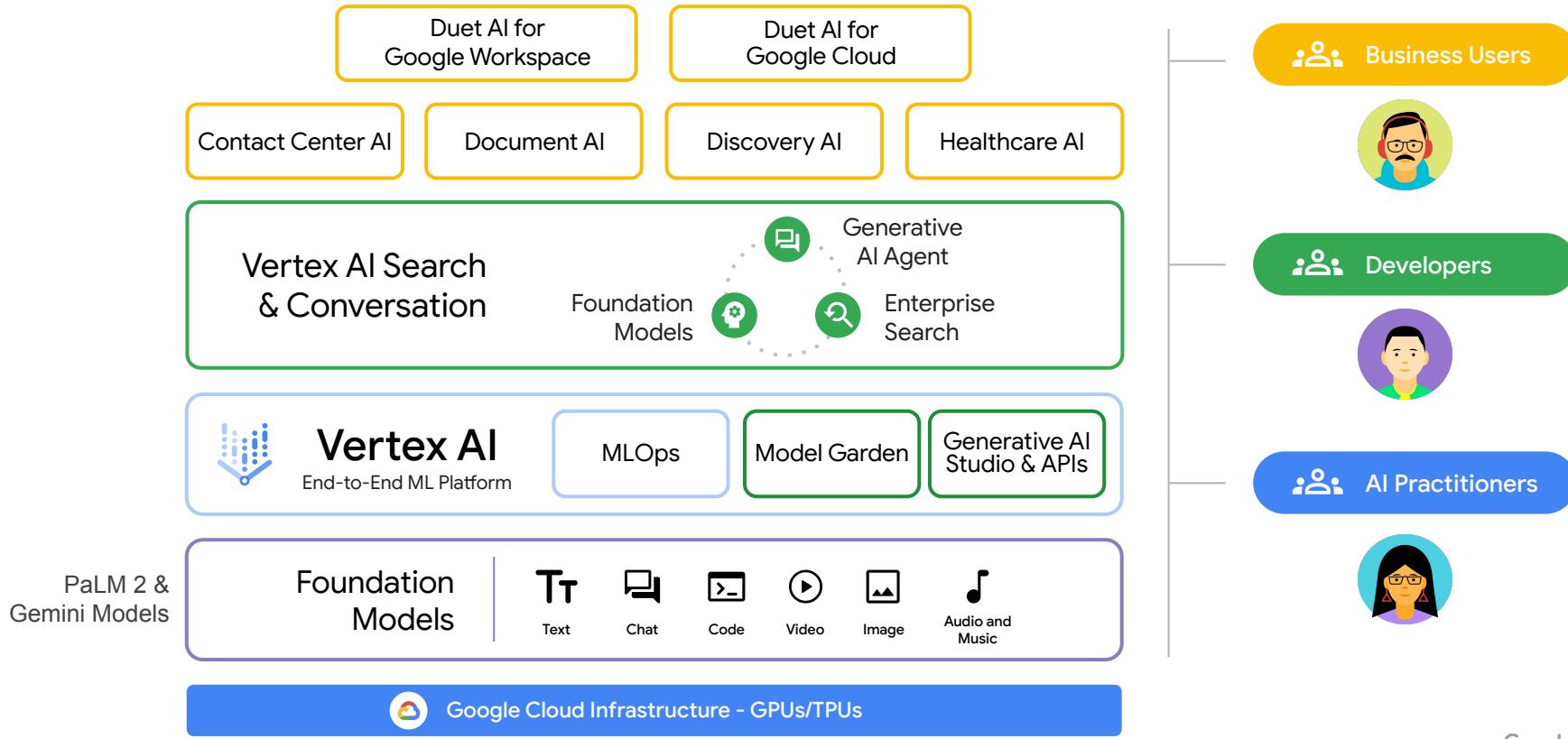


# TOPICS

- **Session 1: Course Intro, Trends, and Approach to AI/ML**
- **Session 2: SQL and NoSQL**
- **Session 3: Distributed ML with Spark and Tensorflow**
- **Session 4: Cloud Generative AI Services and Architectures**
- **Session 5: Cloud Machine Learning Services**
- **Session 6: Serverless ML, Architectures, and Deploying ML**

# Vertex AI: Cloud AI Portfolio

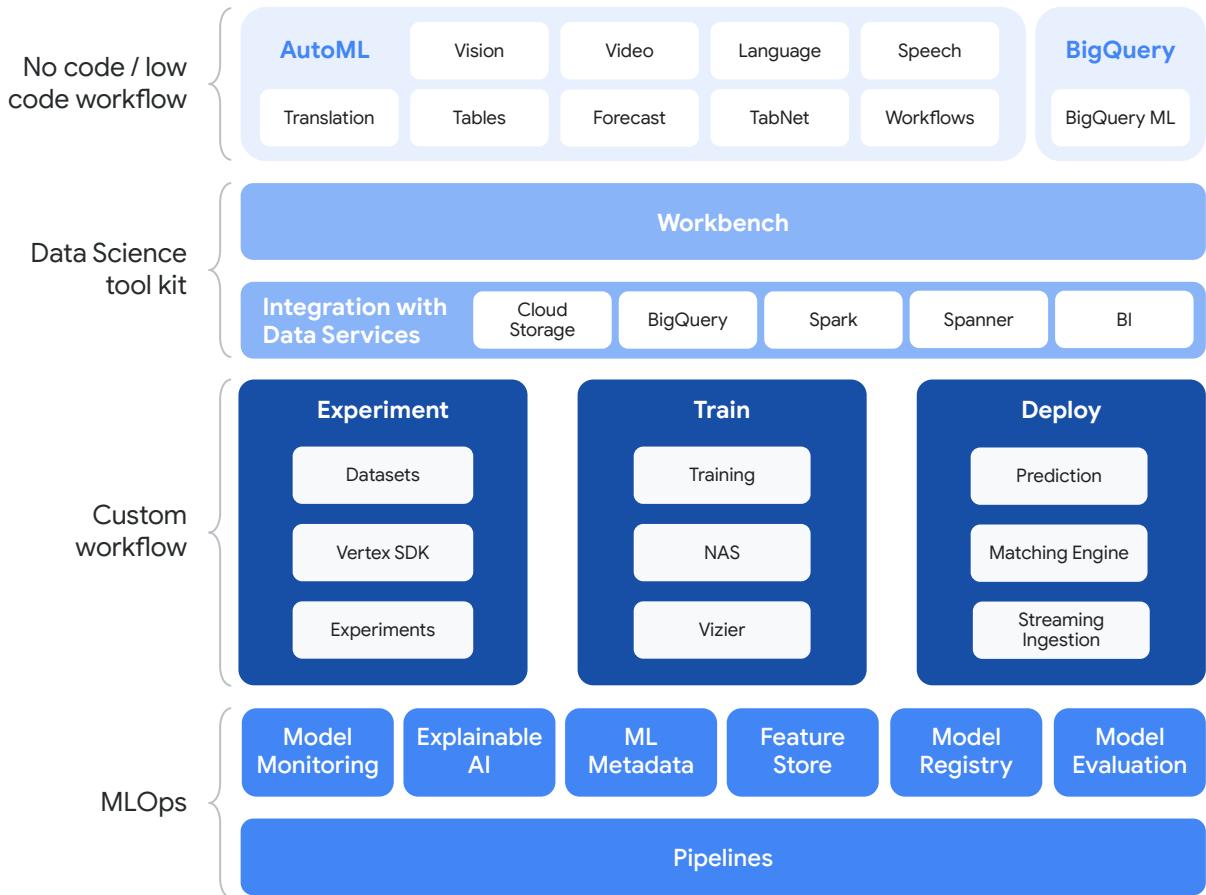
To support the needs of **Generative AI** centric enterprise development





## Vertex AI

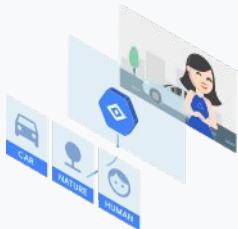
- Unified development and deployment platform for data science and machine learning
- Increase productivity of data scientists and ML engineers



Your Data + Pretrained Model

Your Data + Custom Model

## Machine Learning APIs



Serverless

## AutoML



\* Offline / Edge Compatible

Serverless

## SQL-Based Machine Learning



Serverless

## Vertex AI Platform



\* Offline / Edge Compatible

Serverless Deployment

Developer-focused

Analyst

Data Scientist

Your Data + Pretrained Model

Your Data + Custom Model

## Machine Learning APIs



Serverless

Developer-focused

## AutoML



\* Offline / Edge Compatible

Serverless

## SQL-Based Machine Learning



Serverless

## Vertex AI Platform



\* Offline / Edge Compatible

Serverless Deployment



# Vertex AI

## Pre-Trained Models

Generally available

Best in class tools allowing customers to leverage Google's leadership in AI to solve common problems



### Vision

-  Vision
-  AutoML Vision
-  Video Intelligence
-  AutoML Video Intelligence



### Language

-  Translation
-  AutoML Translation
-  Natural Language
-  AutoML Natural Language



### Conversation

-  Dialogflow
-  Speech-to-Text
-  Text-to-Speech
-  Speaker ID



### Structured data

- AutoML Tables
- TabNet
- Time Series Insights API
- Fleet Routing API
- Vertex AI Forecast



Let's say I'm a  
meteorologist

...

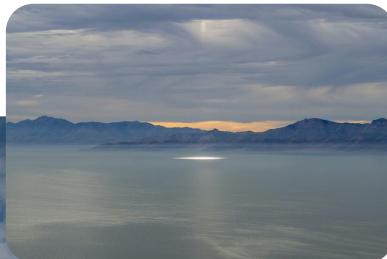


I want to predict  
weather trends  
and flight plans  
from images

# There are 10+ different types of clouds



# There are 10+ different types of clouds



# Let's try the Vision API



# Let's try the Vision API



# Vision AI (Image) Demo



Why Google Solutions Products Pricing Getting Started



Docs Support

English

Console



Cloud Vision API

Contact Us

## Vision AI

Benefits

Demo

Key features

Vision API and AutoML Vision customers

What's new

## Documentation

## Use cases

Vision product search

Document classification

Image search

## Compare features

## Pricing

## Take the next step

# Vision AI

Derive insights from your images in the cloud or at the edge with AutoML Vision or use pre-trained Vision API models to detect emotion, understand text, and more.

Go to console

Contact sales

- ✓ Use machine learning to understand your images with industry-leading prediction accuracy
- ✓ Train machine learning models that classify images by your custom labels using AutoML Vision
- ✓ Detect objects and faces, read handwriting, and build valuable image metadata with Vision API



VIDEO

Fortune 500 global power company AES drives green energy with AutoML Vision

02:30

## BENEFITS

### Detect objects automatically

Detect and classify multiple objects including the location of each object within the image. Learn more about object detection with [Vision API](#) and

### Gain intelligence at the edge

Use AutoML Vision Edge to build and deploy fast, high-accuracy models to classify images or detect objects at the edge, and trigger real-time actions based

### Reduce purchase friction

With Vision API's [vision\\_product\\_search](#), retailers can create an engaging mobile experience that enables customers to upload a photo of an item and immediately see a list of similar



# NLP Demo



Why Google Solutions Products Pricing Getting Started



Docs Support

English

Console



Cloud Natural Language

Contact Us

## Natural Language AI

Benefits

Demo

Key features

Our customers

### Documentation

Compare features

Pricing

Take the next step

# Natural Language AI

Derive insights from unstructured text using Google machine learning.

Get started

- ✓ Get insightful text analysis with machine learning that extracts, analyzes, and stores text
- ✓ Train high-quality machine learning custom models without a single line of code with AutoML
- ✓ Apply natural language understanding (NLU) to apps with Natural Language API



VIDEO

**Next '19: Learn how customers are using the latest Natural Language updates**

46:53

### BENEFITS

#### Insights from customers

Use entity analysis to find and label fields within a document—including emails, chat, and social media—and then sentiment analysis to understand customer opinions to find actionable product and UX insights.

#### Multimedia and multilingual support

Natural Language with [Speech-to-Text API](#) extracts insights from audio. [Vision API](#) adds optical character recognition (OCR) for scanned docs. [Translation API](#) understands sentiments in multiple

#### Extract key document entities that matter

Use custom entity extraction to identify domain-specific entities within documents—many of which don't appear in standard language models—without having to spend time or money on manual analysis.



Your Data + Pretrained Model

Your Data + Custom Model

## Machine Learning APIs



Serverless

## AutoML



\* Offline / Edge Compatible

Serverless

## SQL-Based Machine Learning



Serverless

## Vertex AI Platform



\* Offline / Edge Compatible

Serverless Deployment

Developer-focused

Analyst

Data Scientist

# ...but what if I need more granular intelligence?



## Machine Learning APIs: Ready to Go



Cloud  
Vision API



Cloud  
Speech API



Cloud  
Translation API



Cloud Natural  
Language API



Cloud Video  
Intelligence API

## AutoML: Bring Your Own Data (We Do the Rest)



# AutoML Vision

Upload and label images



Train your model



Cloud AutoML

Evaluate

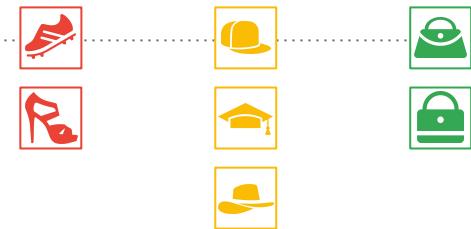
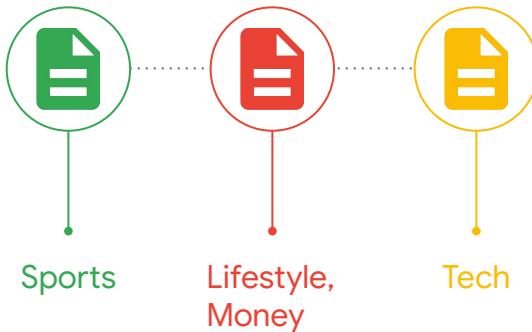


Image Classification  
Object detection  
Image segmentation

# AutoML Natural Language

Upload and label text

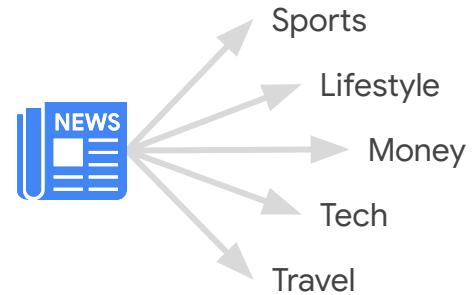


Train your model



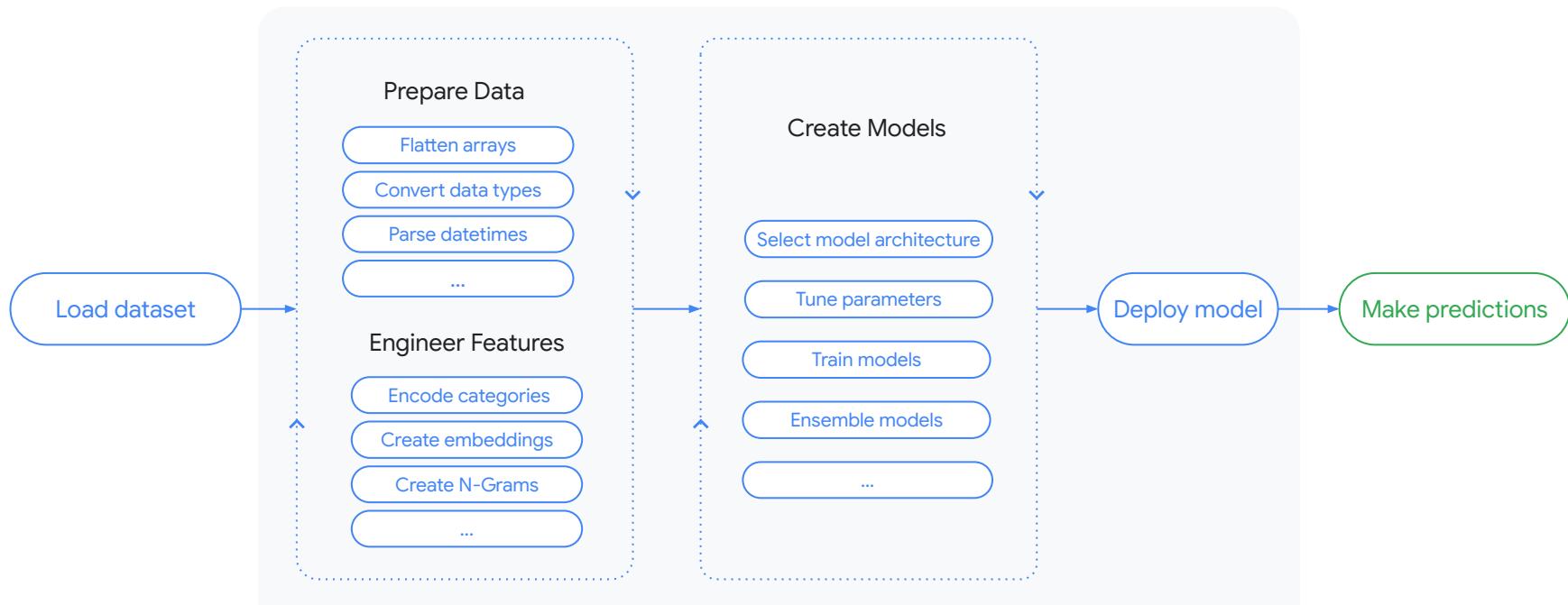
Cloud AutoML

Evaluate



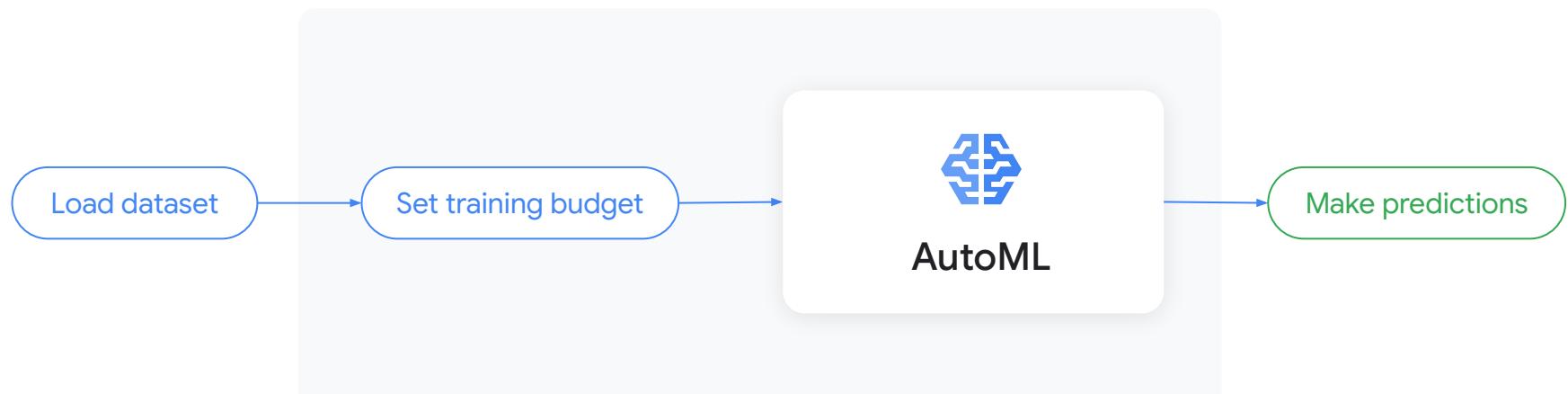
# Fastest path from data to value

## Traditional Machine Learning Workflow



# Fastest path from data to value

## AutoML Workflow



# Handle data as found in the wild

## Automated feature engineering for:



Numbers



Timestamps



Classes



Lists



Strings



Nested fields

## Resilient to + guardrails for:



Imbalanced data



Highly correlated features



Missing values



High cardinality features  
(like IDs)



Outliers

# High quality models for real world problems

## AutoML Scores Within 10% of Top Kaggle Scores

Mercari Price Suggestions Challenge

<7%

[2,382 teams](#)

Allstate claims severity

<0.3%

[3,045 teams](#)

KDD Cup 2014 - Predicting Excitement at Donors Choose

<10%

[472 teams](#)



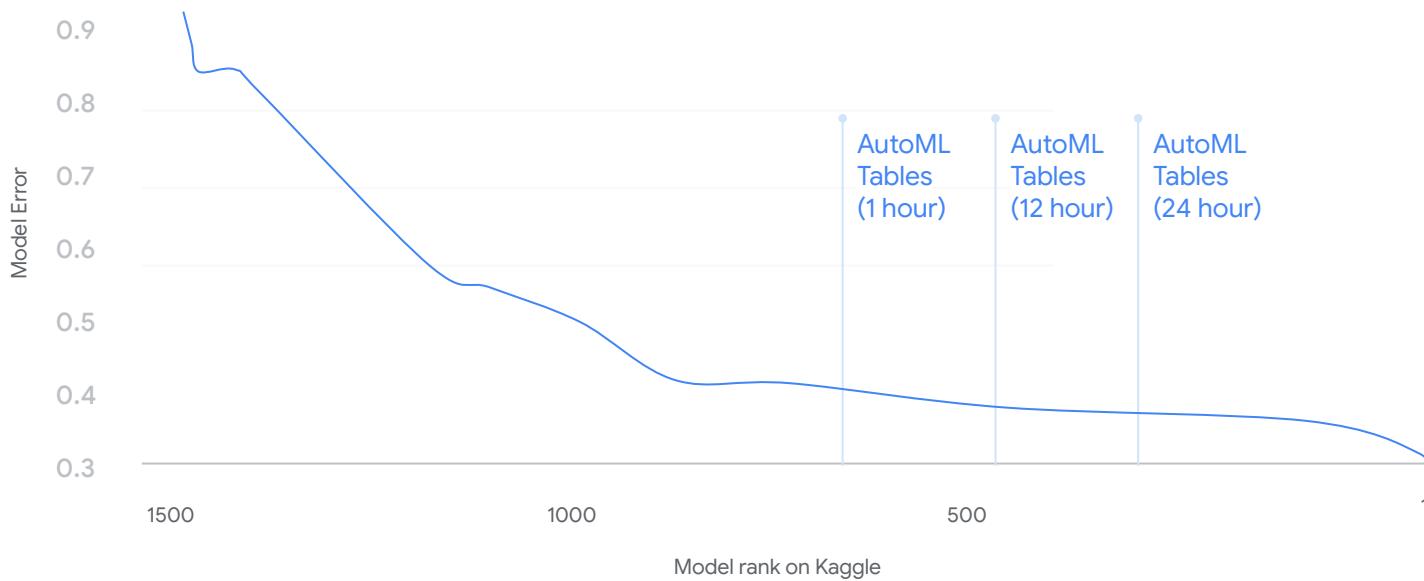
AutoML Score



Kaggle Top Score

# Getting to results quickly

## Mercari Price Suggestion Challenge



# Automated feature engineering

Best practice transformations for all data types

 **Numbers:** generate quantiles, log, z\_score transforms

 **Arrays of categories:** convert to lookup index, generate embeddings

 **Text:** tokenize, generate n-grams, create embeddings

 **Datetime:** extract year, month, day, weekday, categorize

 **Categories:** one-hot encoding, grouping, embeddings

 **Nested fields:** flatten, apply type transformations

and guardrails for

 **Imbalanced data**

 **Missing values**

## Vertex AI

## Dashboard

Dashboard

Datasets

Features

Labeling tasks

Workbench

Pipelines

Training

Experiments

Models

Endpoints

Batch predictions

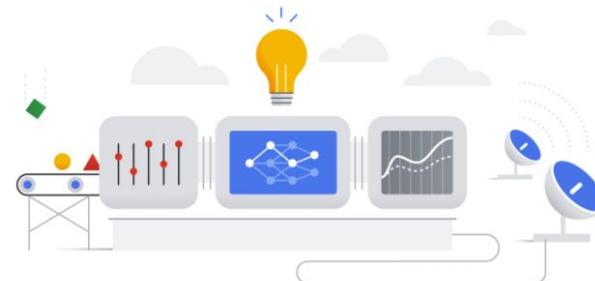
Metadata

Marketplace

## Get started with Vertex AI

Vertex AI empowers machine learning developers, data scientists, and data engineers to take their projects from ideation to deployment, quickly and cost-effectively. [Learn more](#)

Try an interactive tutorial to learn how to train, evaluate, and deploy a Vertex AI AutoML or custom-trained model

[VIEW TUTORIALS](#)

Region

us-central1 (Iowa)



## Recent datasets

- text-classification-ds 7 minutes ago
- bank-marketing-ds 13 minutes ago
- bank-marketing-data 27 minutes ago

[+ CREATE DATASET](#)

## Train your model

Train a best-in-class machine learning model with your dataset. Use Google's AutoML, or bring your own code.

[+ TRAIN NEW MODEL](#)

## Get predictions

After you train a model, you can use it to get predictions, either online as an endpoint or through batch requests

[+ CREATE BATCH PREDICTION](#)

Your Data + Pretrained Model

Your Data + Custom Model

## Machine Learning APIs



Serverless

## AutoML



\* Offline / Edge Compatible

Serverless

## SQL-Based Machine Learning



Serverless

## Vertex AI Platform



\* Offline / Edge Compatible

Serverless Deployment

Developer-focused

Analyst

Data Scientist

Accessible Machine  
Learning for **Analysts**

# Machine Learning Using **SQL** in BigQuery

# Google BigQuery

Serverless cloud data warehouse with customers ranging from TB to 100+ PB



Easy scaling of storage and compute



Full ANSI SQL (inc. DML and DDL)



Built in security and encryption



High availability and durability



Cross-cloud Lakehouse



Real-time insights



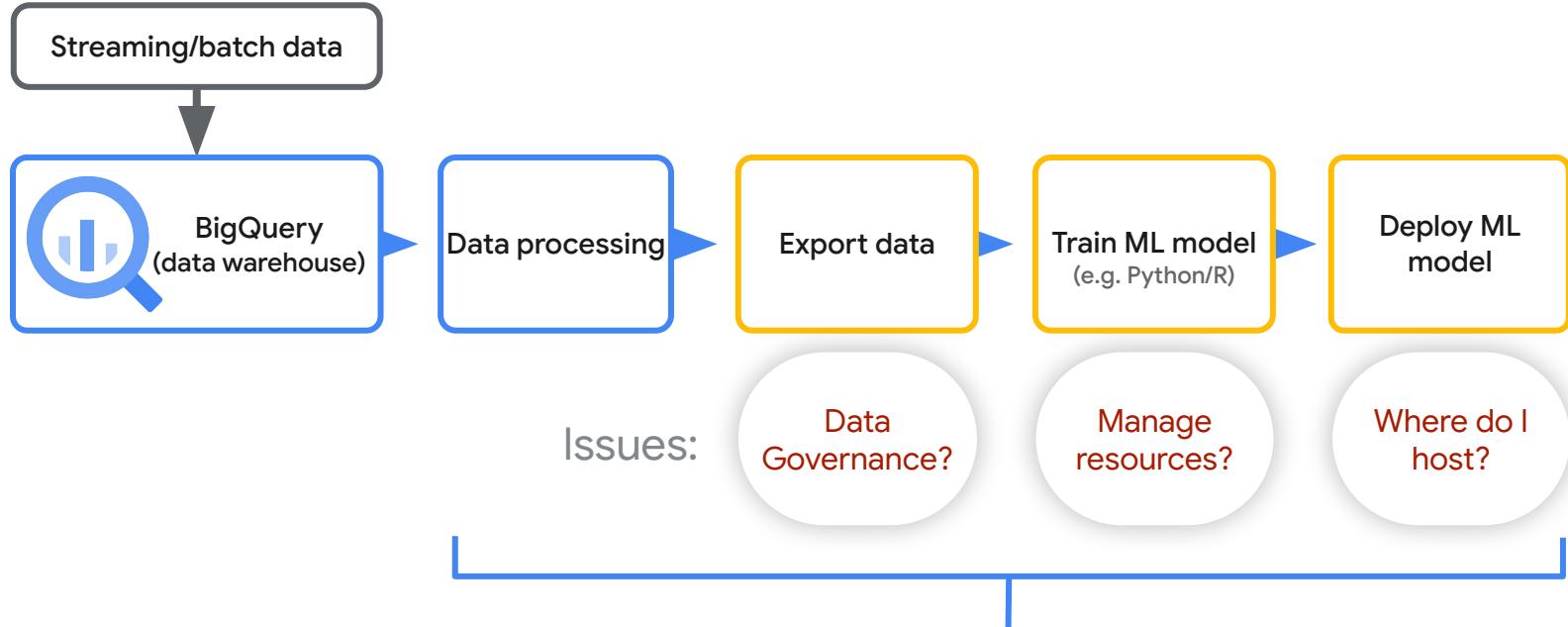
Built-in Machine Learning



Insights for everyone

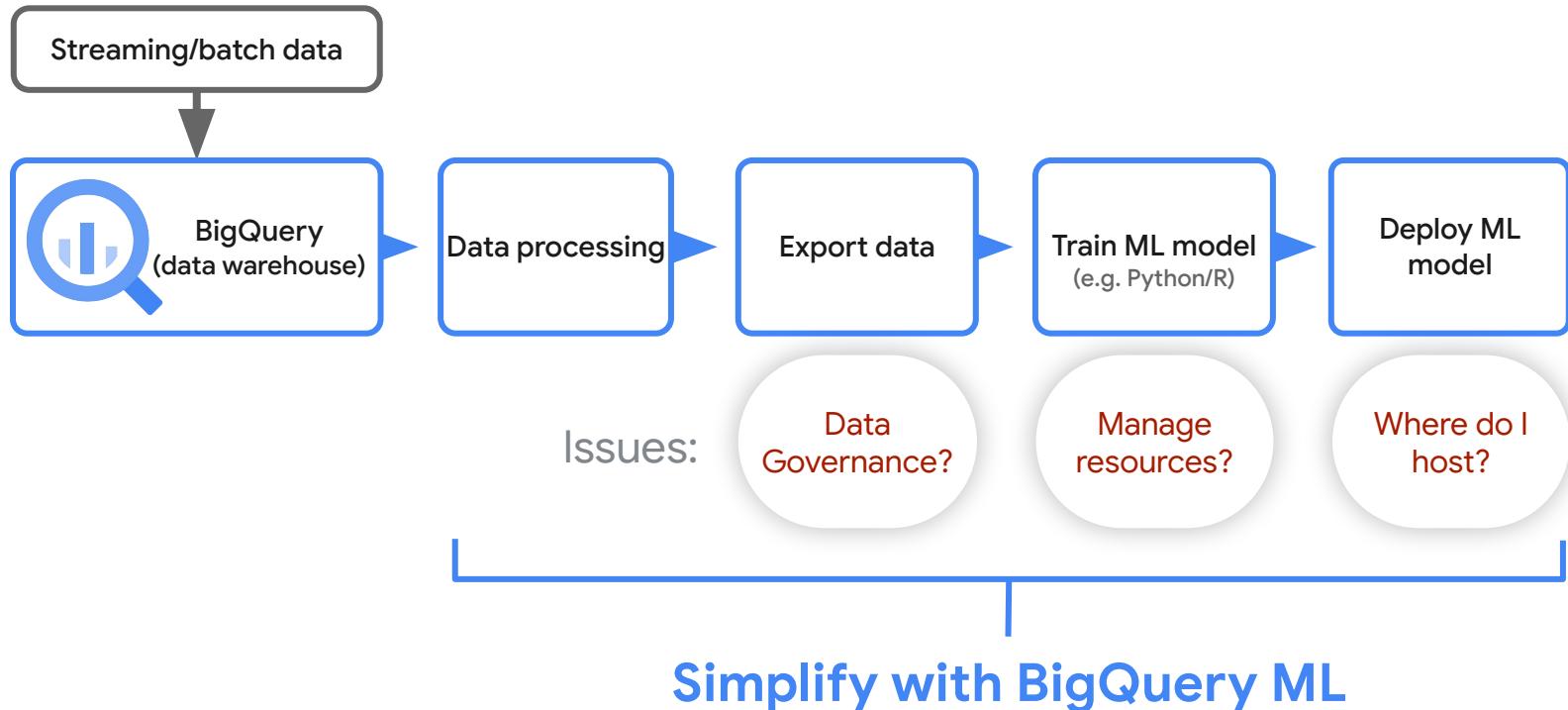


# Typical ML Workflow



Multiple products & roles can lead to unnecessary complexity & costs

# ML Workflow on BigQuery



# BigQuery ML supported models and features

## Classification

- Logistic regression
- DNN classifier (TensorFlow)
- XGBoost
- AutoML Tables
- Wide and Deep NNs

## Other Capabilities

- k-means clustering
- Time series forecasting (ARIMA+)
- Recommendation: Matrix factorization
- Anomaly Detection
- Explainable AI

## Regression

- Linear regression
- DNN regressor (TensorFlow)
- XGBoost
- AutoML Tables
- Wide and Deep NNs

## ML Ops

- Import TensorFlow models for batch prediction and export BigQuery models for online prediction
- Hyperparameter Tuning using Vertex AI Vizier
- Integration with Vertex Model Registry and Vertex Managed Pipelines

# BigQuery ML

## What does model building look like

```
CREATE OR REPLACE MODEL my_models.car_accidents
OPTIONS
  (model_type='logistic_reg', num_iter=10,
  l2_reg=1, learn_rate=0.5, warm_start=true,
  Labels=[bad_accident]) AS
SELECT
  borough,
  contributing_factor_vehicle_1,
  contributing_factor_vehicle_2,
  on_street_name,
  cross_street_name,
  vehicle_type_code1,
  vehicle_type_code2,
  zip_code,
  bad_accident
FROM ...
```

### What it does

Builds a model on data stored in BigQuery that predicts car accident severity based on the location, vehicles involved, and NYC borough.

Automatically transforms the data to make sure it's ready for ML

Stores the model in BigQuery for later use

Provides measures of accuracy and model quality

# BigQuery ML

## What does prediction look like?

```
SELECT *  
FROM ml.PREDICT(  
    MODEL my_models.car_accidents,  
    (  
        SELECT  
            borough,  
            contributing_factor_vehicle_1,  
            contributing_factor_vehicle_2,  
            on_street_name,  
            cross_street_name,  
            vehicle_type_code1,  
            vehicle_type_code2,  
            Zip_code  
        FROM ...  
    )  
);
```

### What it does

Uses the named model to predict outcomes from a table or SQL query

Automatically transforms the data using the preprocessing in the model

Can be returned to an application, browsed, or stored in BigQuery

Can be used in any SQL query



Your Data + Pretrained Model

Your Data + Custom Model

## Machine Learning APIs



Serverless

## AutoML



\* Offline / Edge Compatible

Serverless

## SQL-Based Machine Learning



Serverless

## Vertex AI Platform



\* Offline / Edge Compatible

Serverless Deployment

Developer-focused

Analyst

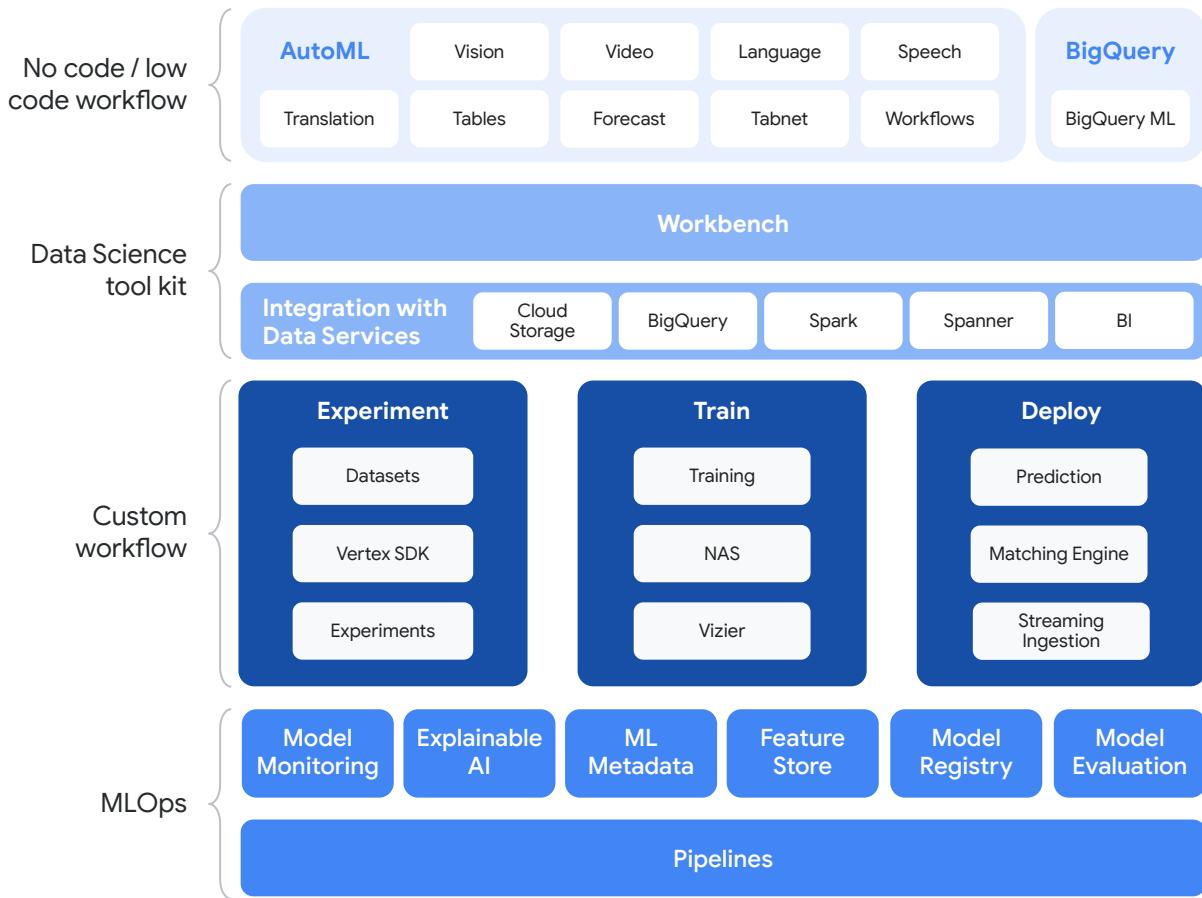
Data Scientist



# Vertex AI

A Unified ML Platform  
for Solving All Business Problems

- **Unified** development and deployment platform for machine learning **at scale**
- Increase **productivity** of data scientists and ML engineers
- Improve **time to value**



# Vertex AI Workbench

A one-stop surface for Data Science



## Fully managed compute with admin control

A Jupyter-based fully managed, scalable, enterprise-ready compute infrastructure with easily enforceable policies and user management



## Fast workflow for data tasks

Seamless visual and code-based integrations with data & analytics services



## At-your-fingertips integration

Load and share notebooks alongside your AI and data tasks. Run tasks without extra code

The screenshot shows the Google Cloud Platform interface for Vertex AI Workbench. On the left, there is a sidebar with various options: Dashboard, Datasets, Features, Labeling tasks, Workbench (which is selected), Pipelines, Training, Experiments, Models, Endpoints, Edge deployments, Batch predictions, and Metadata. The main area is titled "Workbench" and shows a list of "MANAGED NOTEBOOKS". The list includes:

Notebook name	Location	Access mode
managed-notebook-1643391246	us-central1-f	Single user only
managed-notebook-1643393492	us-central1-c	Single user only
managed-notebook-1647318683	us-central1-c	Single user only
mchrestkha-sandbox	us-central1-a	Single user only
nvidia-ncg	us-central1-f	Single user only
tf-mnist-ncg	us-central1-f	Single user only

Below this, a specific notebook named "mchrestkha-sandbox" is shown in a preview window. The preview window has tabs for File, Edit, View, Run, Kernel, Git, Tabs, Settings, and Help. It also shows a file browser with a directory structure: /, src, and tutorials. The last modified time for both is 7 minutes ago. To the right of the preview window, there is a "Notebook" section with icons for Python (Local), PySpark (Local), P (PySpark on cluster-5977-m), Python 3 on cluster-5977-m, Pytorch (Local), R (Local), and R on cluster-5977-m in us. Below that is a "Console" section with similar icons.

# Serverless Spark on Vertex AI Workbench

Spark for Data Science in one click

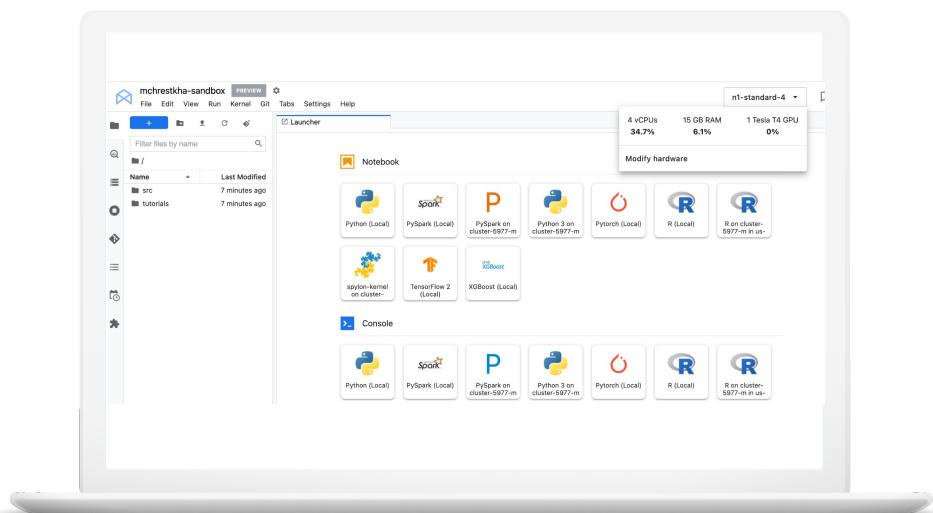


## Built-in security and authentication

Google Cloud security and user access are applied from Vertex AI to Spark.

## Integrate Spark with MLOps

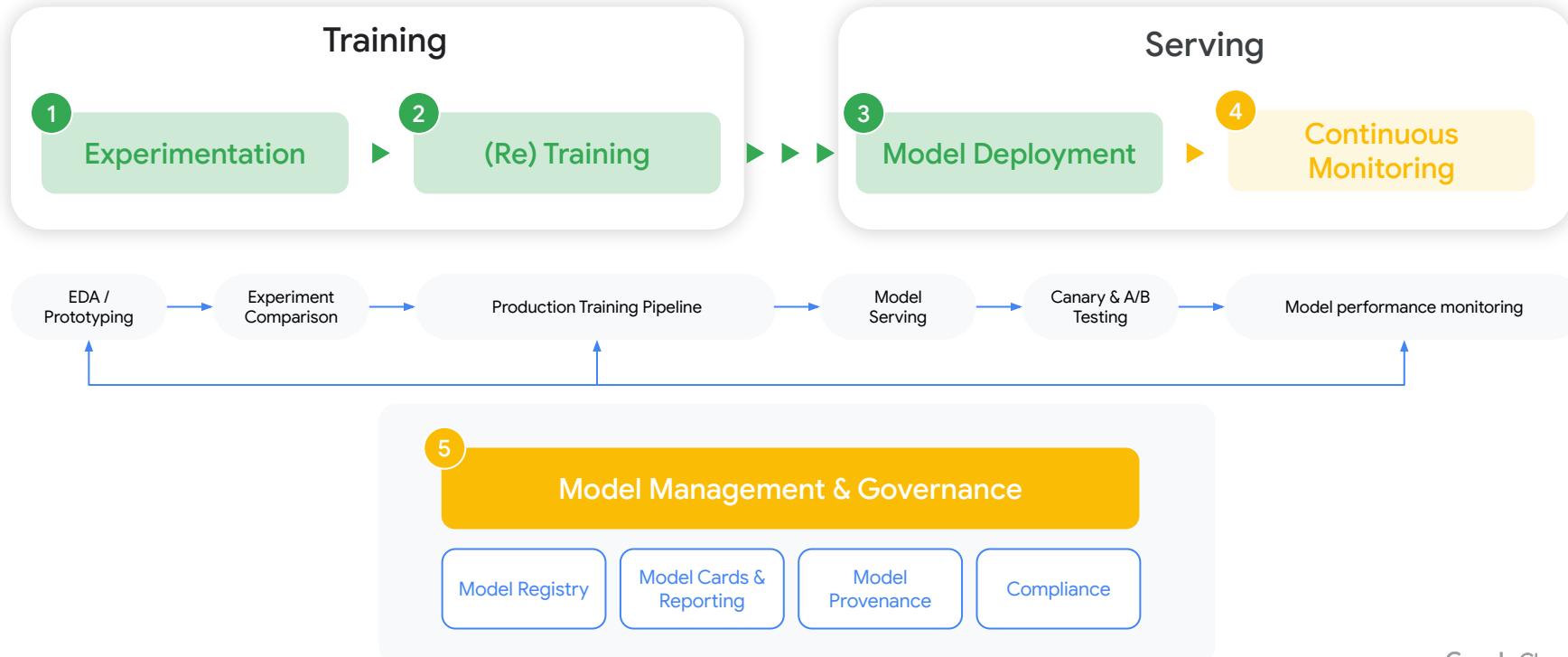
Execute Spark code through Kubeflow pipelines.





# Efficient and responsible AI requires end-to-end MLOps

Vertex AI's end-to-end MLOps enables data scientists and ML engineers to efficiently and responsibly **manage, monitor, govern, and explain** ML projects throughout the entire development lifecycle.



# Vertex AI Experiments

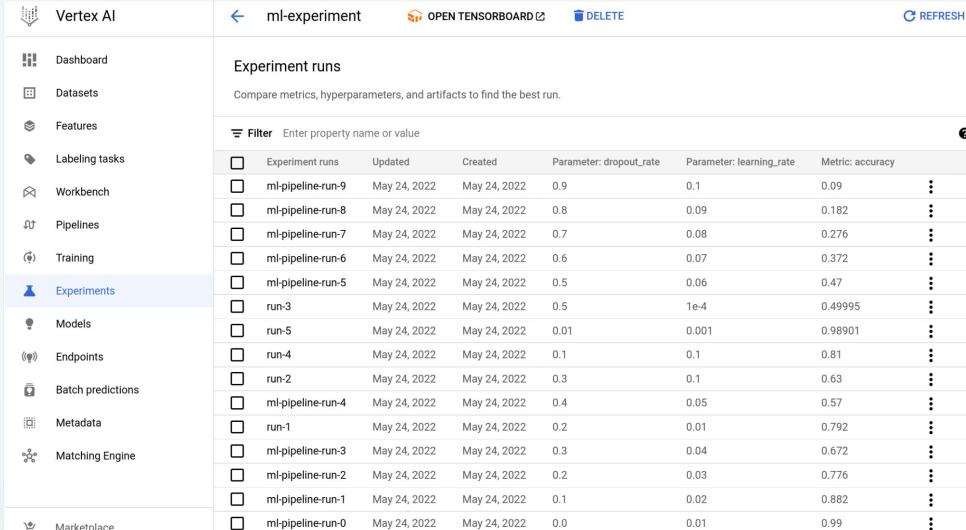
Track and compare multiple experiment runs and analyze key model metrics

**Vary and track parameters and metrics as you experiment.**

**Organize Vertex Pipeline runs and compare their parameters, metrics, and artifacts.**

**Track steps and artifacts to capture the lineage of experiments.**

**Compare Vertex Pipelines against Notebook experiments.**



The screenshot shows the Vertex AI Experiments interface. On the left is a sidebar with icons for Dashboard, Datasets, Features, Labeling tasks, Workbench, Pipelines, Training, Experiments (which is selected), Models, Endpoints, Batch predictions, Metadata, Matching Engine, and Marketplace. The main area has a header with back, forward, open tensorboard, delete, and refresh buttons. Below the header is a section titled "Experiment runs" with a sub-section "Compare metrics, hyperparameters, and artifacts to find the best run." A filter input field says "Enter property name or value". The main table lists 18 experiment runs with columns: Experiment runs, Updated, Created, Parameter: dropout\_rate, Parameter: learning\_rate, Metric: accuracy, and three-dot more actions columns. The runs are: ml-pipeline-run-9, ml-pipeline-run-8, ml-pipeline-run-7, ml-pipeline-run-6, ml-pipeline-run-5, run-3, run-5, run-4, run-2, ml-pipeline-run-4, run-1, ml-pipeline-run-3, ml-pipeline-run-2, ml-pipeline-run-1, and ml-pipeline-run-0. All runs were created and updated on May 24, 2022.

Experiment runs	Updated	Created	Parameter: dropout_rate	Parameter: learning_rate	Metric: accuracy	⋮
ml-pipeline-run-9	May 24, 2022	May 24, 2022	0.9	0.1	0.09	⋮
ml-pipeline-run-8	May 24, 2022	May 24, 2022	0.8	0.09	0.182	⋮
ml-pipeline-run-7	May 24, 2022	May 24, 2022	0.7	0.08	0.276	⋮
ml-pipeline-run-6	May 24, 2022	May 24, 2022	0.6	0.07	0.372	⋮
ml-pipeline-run-5	May 24, 2022	May 24, 2022	0.5	0.06	0.47	⋮
run-3	May 24, 2022	May 24, 2022	0.5	1e-4	0.49995	⋮
run-5	May 24, 2022	May 24, 2022	0.01	0.001	0.98901	⋮
run-4	May 24, 2022	May 24, 2022	0.1	0.1	0.81	⋮
run-2	May 24, 2022	May 24, 2022	0.3	0.1	0.63	⋮
ml-pipeline-run-4	May 24, 2022	May 24, 2022	0.4	0.05	0.57	⋮
run-1	May 24, 2022	May 24, 2022	0.2	0.01	0.792	⋮
ml-pipeline-run-3	May 24, 2022	May 24, 2022	0.3	0.04	0.672	⋮
ml-pipeline-run-2	May 24, 2022	May 24, 2022	0.2	0.03	0.776	⋮
ml-pipeline-run-1	May 24, 2022	May 24, 2022	0.1	0.02	0.882	⋮
ml-pipeline-run-0	May 24, 2022	May 24, 2022	0.0	0.01	0.99	⋮

## Monitor

# Proactively monitoring model performance with **Model Monitoring**



### Monitor and alert

Monitor signals for model's predictive performance (batch and online), and alert when those signals deviate.



### Diagnose

Help identify the cause for the deviation i.e. what changed, how and how much?



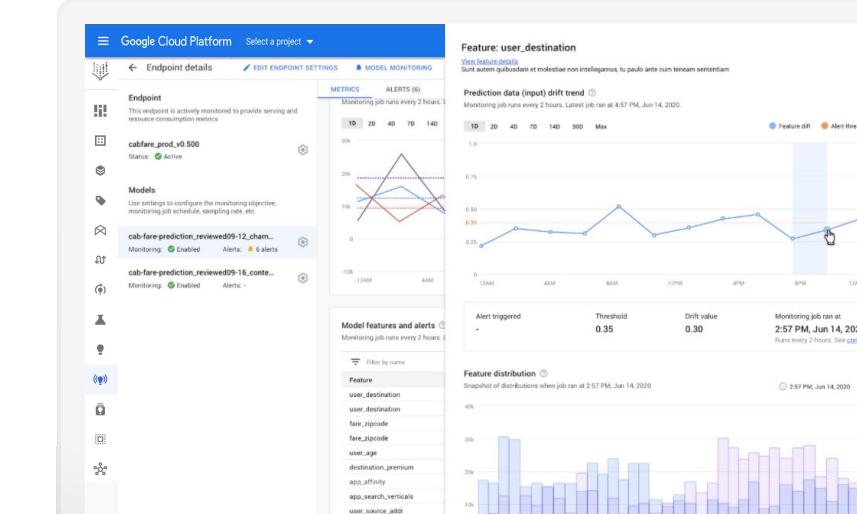
### Update Model

Trigger model re-training pipeline or collect relevant training data to address performance degradation.



### Integrated with Feature Store

Monitor and set up alerts for Feature Store performance and resource utilization, and track how much a feature's value distribution changes over time





## Govern

# Manage and govern your ML models with **Feature Store**, **ML Metadata**, **Model Registry**, and **Model Evaluation**

Preview

## Feature Store

- Share and **reuse** ML features across use cases
- Serve ML Features **at scale** with **low latency**
- Alleviate training serving skew

## ML Metadata

- Automatically track inputs / outputs to all components
- Track custom metadata **directly from your code**
- Visualize, analyze, and compare detailed ML lineage

## Model Registry

- Register, organize, track, and **version** your trained and deployed ML models.
- Govern** the model launch process
- Maintain model documentation and reporting

## Model Evaluation

- Iteratively **run model evaluations** on new datasets at scale
- Visualize and compare model evaluations to identify the **best model for prod deployment**
- Assess the performance of models** on different slices and evaluated annotations



## Explain

# Reveal the ‘why’ behind your model & predictions with Explainable AI



### Robust, actionable explanations

“Feature attributions” show you which input features are most important to your model overall and for specific predictions (using Sampled Shapley, Integrated Gradients, and XRAI).



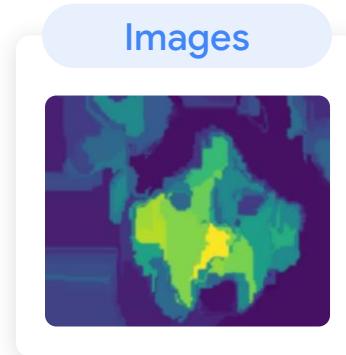
### Built into multiple Vertex AI services

Get explanations easily through Vertex AI Prediction, AutoML Tables, and Vertex AI Workbench



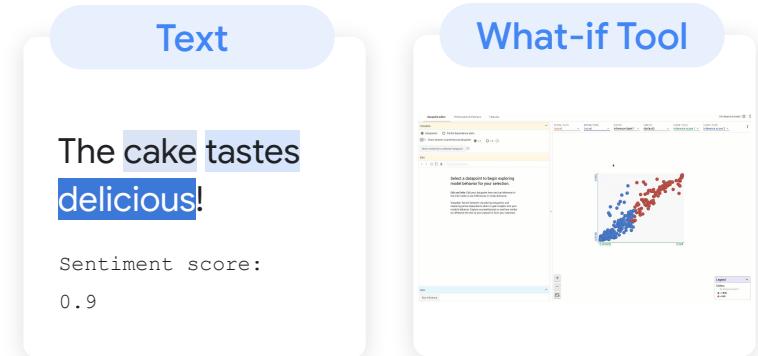
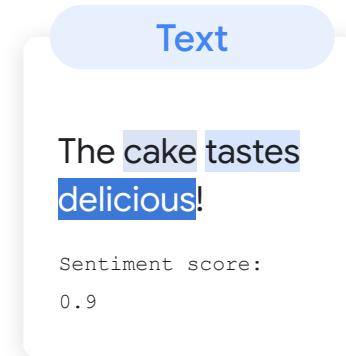
### Flexible, fast & scalable

Supports tabular, image & text models from any ML framework. Fully managed, serverless, and significantly faster than open-source.



### Tabular

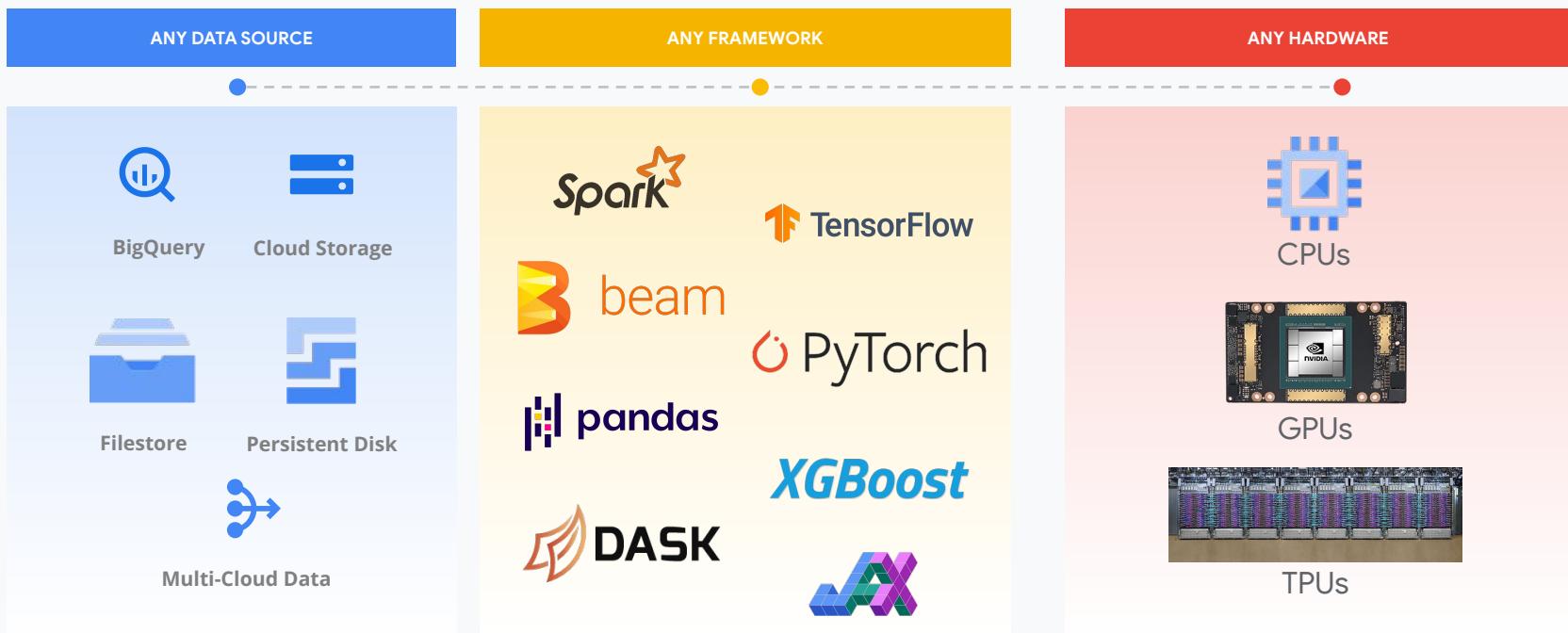
Name	Feature value	Attribution value
distance	1395.51	-2.44478
start_hr	18	-1.29039
max_temp	20.7239	0.690506
temp	16.168	0.12629
dew_point	7.83396	0.0110318
precipitation	0.03	-0.00134132
euclidean		
loc_cross		
start_station_id		
end_station_id		
max		



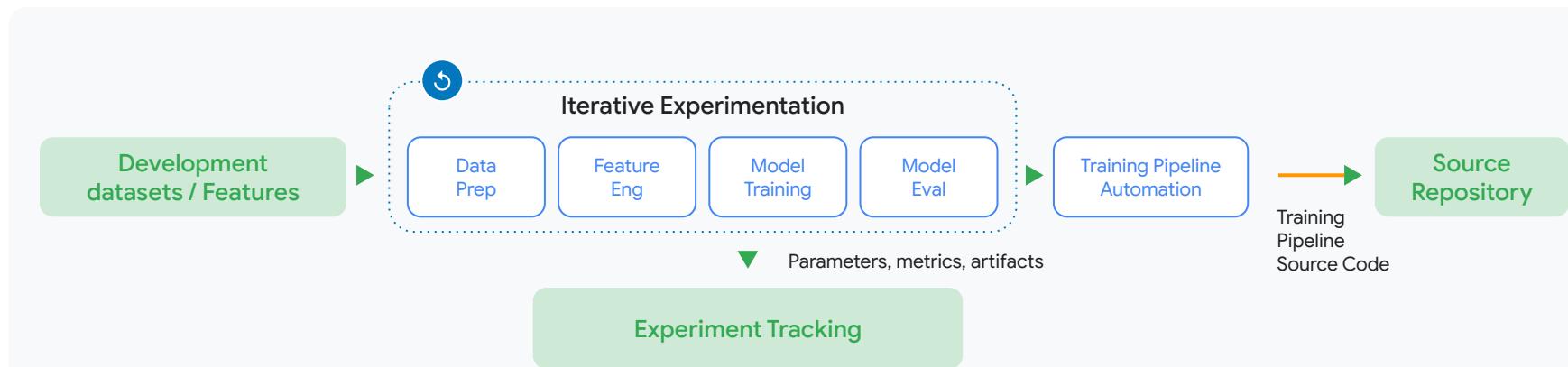


# Accelerate the **velocity** of models to production

While maintaining the **flexibility** of ML frameworks and **compute options**



# Experimentation



## Tools and services

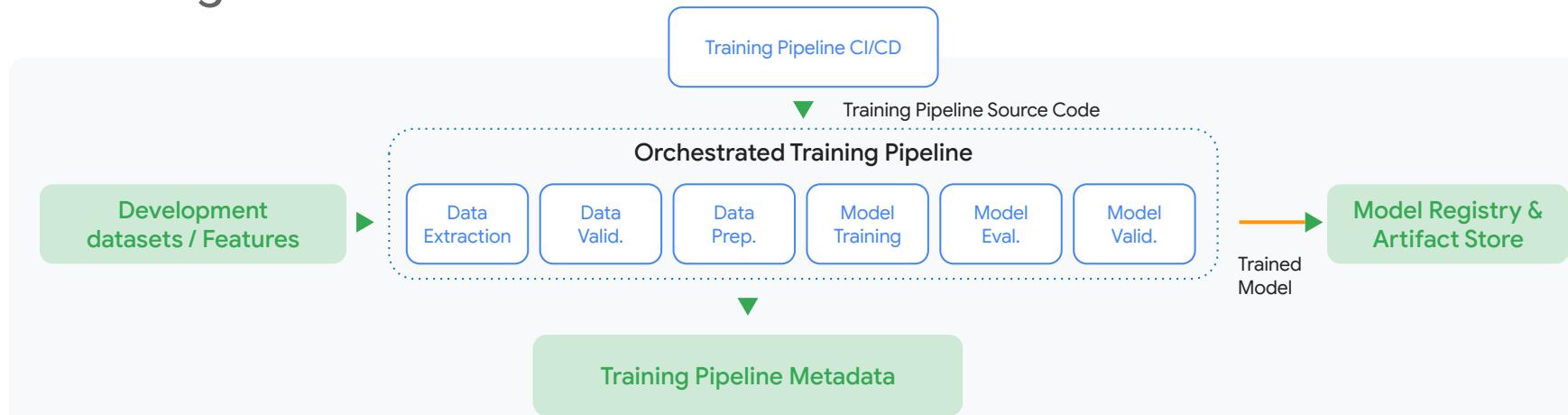
- Notebooks
  - Vertex Training
  - AutoML in Vertex AI
  - BQML
- Vertex Tensorboard
  - Vertex Pipelines
  - Vertex Feature Store
  - What-if Tool

## Key Artifacts

- Development datasets
- Features
- Experiments
- Parameters, metrics



# Training



## Tools and services

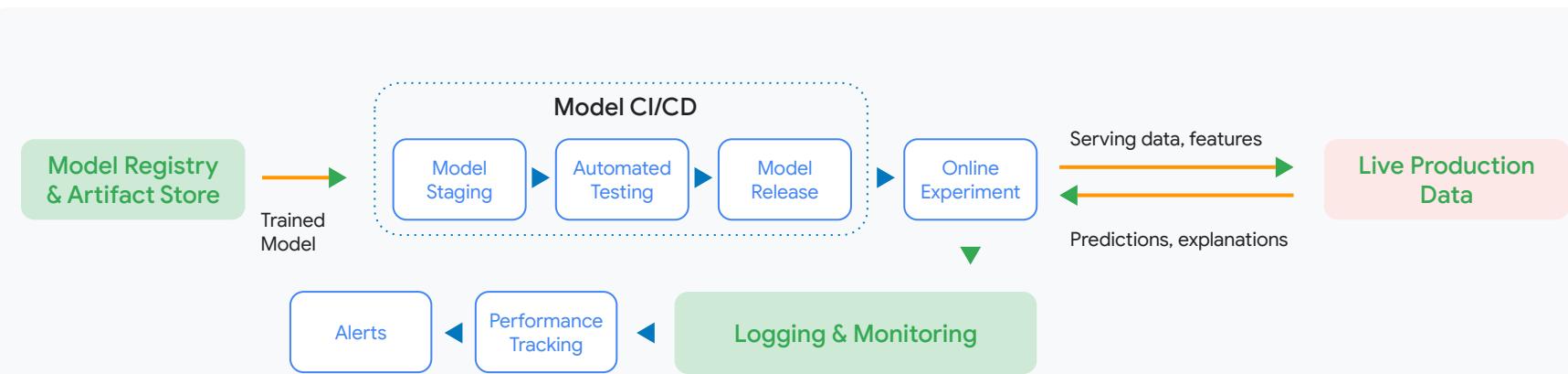
- Vertex Pipelines
  - Vertex Training
  - Cloud Build
  - Artifact Store
- Vertex Explainable AI
  - Vertex ML Metadata
  - Vertex Feature Store

## Key Artifacts

- Training datasets
- Features
- Pipeline source code & containers
- Pipeline Metadata
- Models



# Model Deployment with Monitoring



## Tools and services

- Vertex Prediction
  - Vertex Pipelines
  - Vertex Explainable AI
  - Artifact Store
- Vertex Model Monitoring
  - Vertex ML Metadata
  - Vertex Feature Store

## Key Artifacts

- Deployed Models
- Production Serving Data
- Features
- Online Predictions



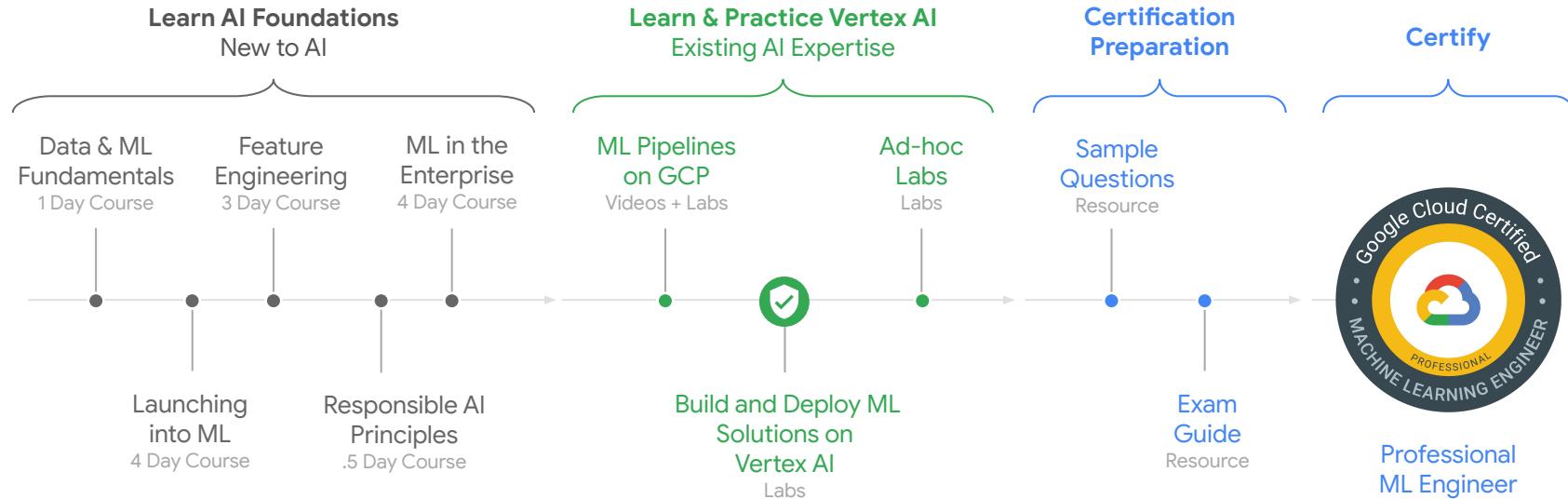
# Training on Vertex AI

Train with	Data Analyst	Software Developer	Data Scientist	Use when
Vertex AI Training			✓	<ul style="list-style-type: none"><li>You need more flexibility and customization than the criteria listed below for BigQuery ML or AutoML.</li><li>You're already running training on-premises or another cloud, and you need consistency across the platforms.</li></ul>
AutoML	✓	✓		<ul style="list-style-type: none"><li>Your problem fits into one of the types AutoML supports. Offers a point-and-click workflow.</li><li>Natural Language or Video models are served from Google Cloud. While Vision and Tables support edge / downloadable models.</li></ul>
BigQuery ML	✓	✓	✓	<ul style="list-style-type: none"><li>All your data is contained in BigQuery.</li><li>Users are most comfortable with SQL.</li><li>The set of <a href="#">models available in BigQuery ML</a> matches the problem you're trying to solve.</li></ul>



# Vertex AI certification pathway

Machine Learning and Artificial Intelligence | Role: ML Engineer



Thanks!

# Advanced Big Data: Cloud Machine Learning

**Dan Zaratsian**

AI/ML Solutions Architect, Gaming @ Google