Advanced Big Data:
Cloud Machine Learning

Dan Zaratsian

AI/ML Solutions Architect, Gaming @ Google

d.zaratsian@gmail.com

https://github.com/zaratsian

Logistics

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

Google for Games

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.

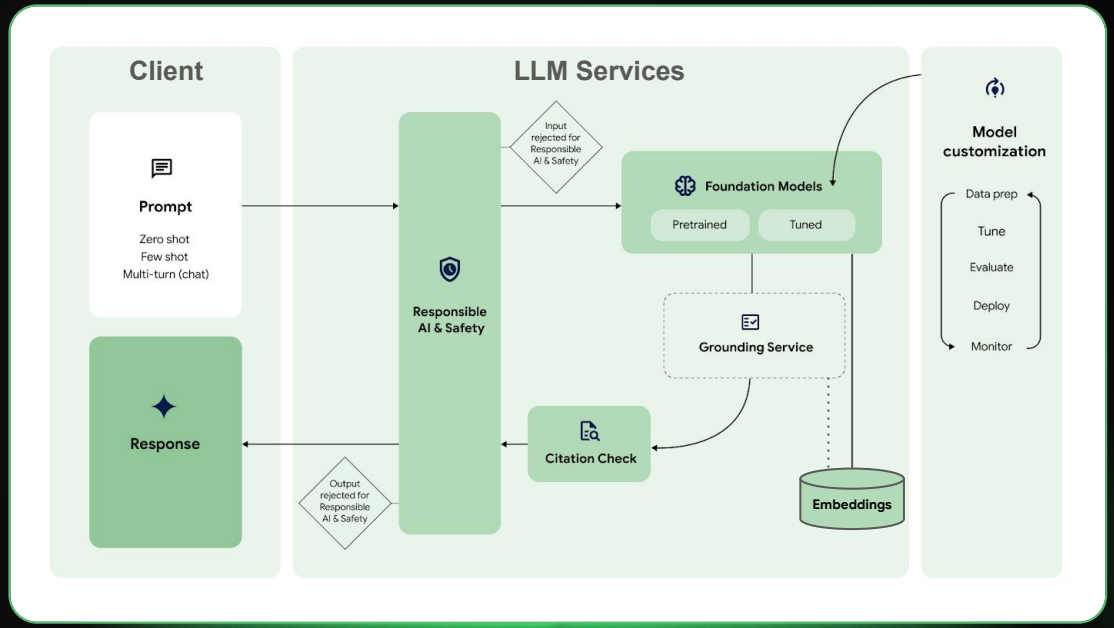
Scalable Infra

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

Duet AI for Google Workspace

Duet AI for Google Cloud

Contact Center AI

Document AI

Discovery AI

Healthcare AI

Vertex AI Search & Conversation

Generative AI Agent

Foundation Models

Enterprise Search

Vertex AI

End-to-End ML Platform

MLOps

Model Garden

Generative AI Studio & APIs

PaLM 2 & Gemini Models

Foundation Models

Text

Chat

Code

Video

Image

Audio and Music

Google Cloud Infrastructure - GPUs/TPUs

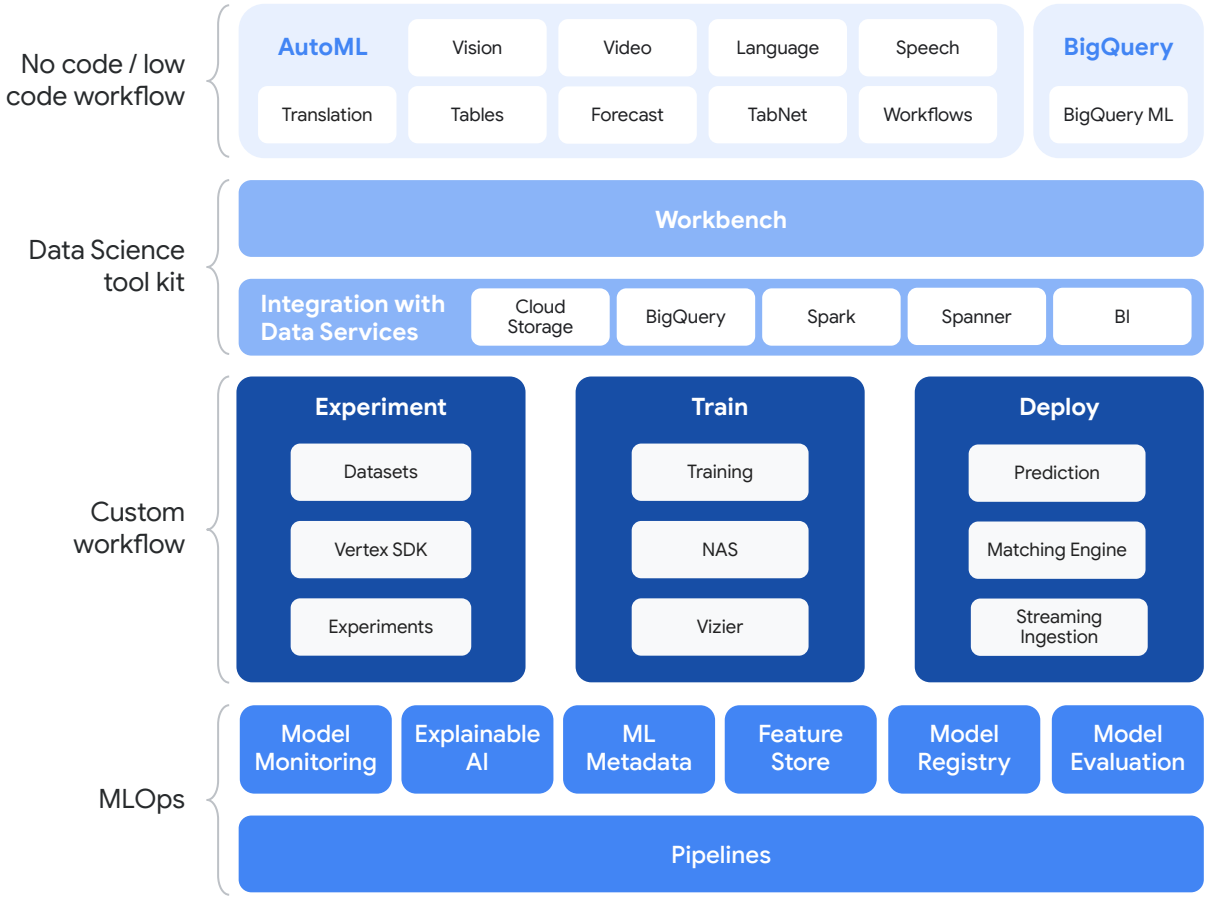
Business Users

Developers

AI Practitioners

Vertex AI

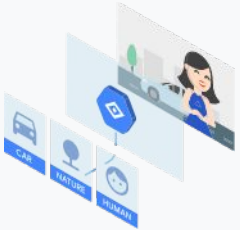
- Unified development and deployment platform for data science and machine learning
- Increase productivity of data scientists and ML engineers



Your Data + Pretrained Model

Your Data + Custom Model

Machine Learning APIs



Serverless

AutoML

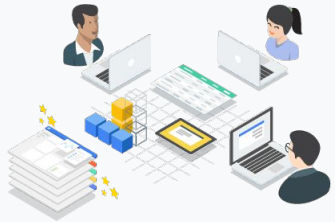
* Offline / Edge Compatible

Serverless

SQL-Based Machine Learning

Serverless

Vertex AI Platform



* Offline / Edge Compatible

Serverless Deployment

Developer-focused

Analyst

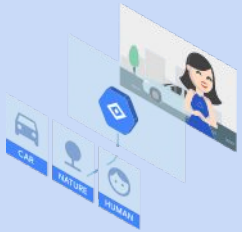
Data Scientist

Google Cloud

Your Data + Pretrained Model

Your Data + Custom Model

Machine Learning APIs



Serverless

AutoML

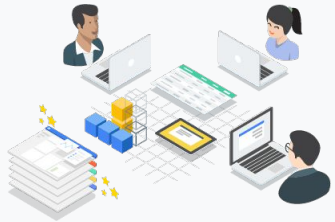
* Offline / Edge Compatible

Serverless

SQL-Based Machine Learning

Serverless

Vertex AI Platform



* Offline / Edge Compatible

Serverless Deployment

Developer-focused

Analyst

Data Scientist

Google Cloud

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

Let's say I'm a meteorologist ...

Google Cloud

There are 10+ different types of clouds

Google Cloud

There are 10+ different types of clouds

Google Cloud

Let’s try the Vision API

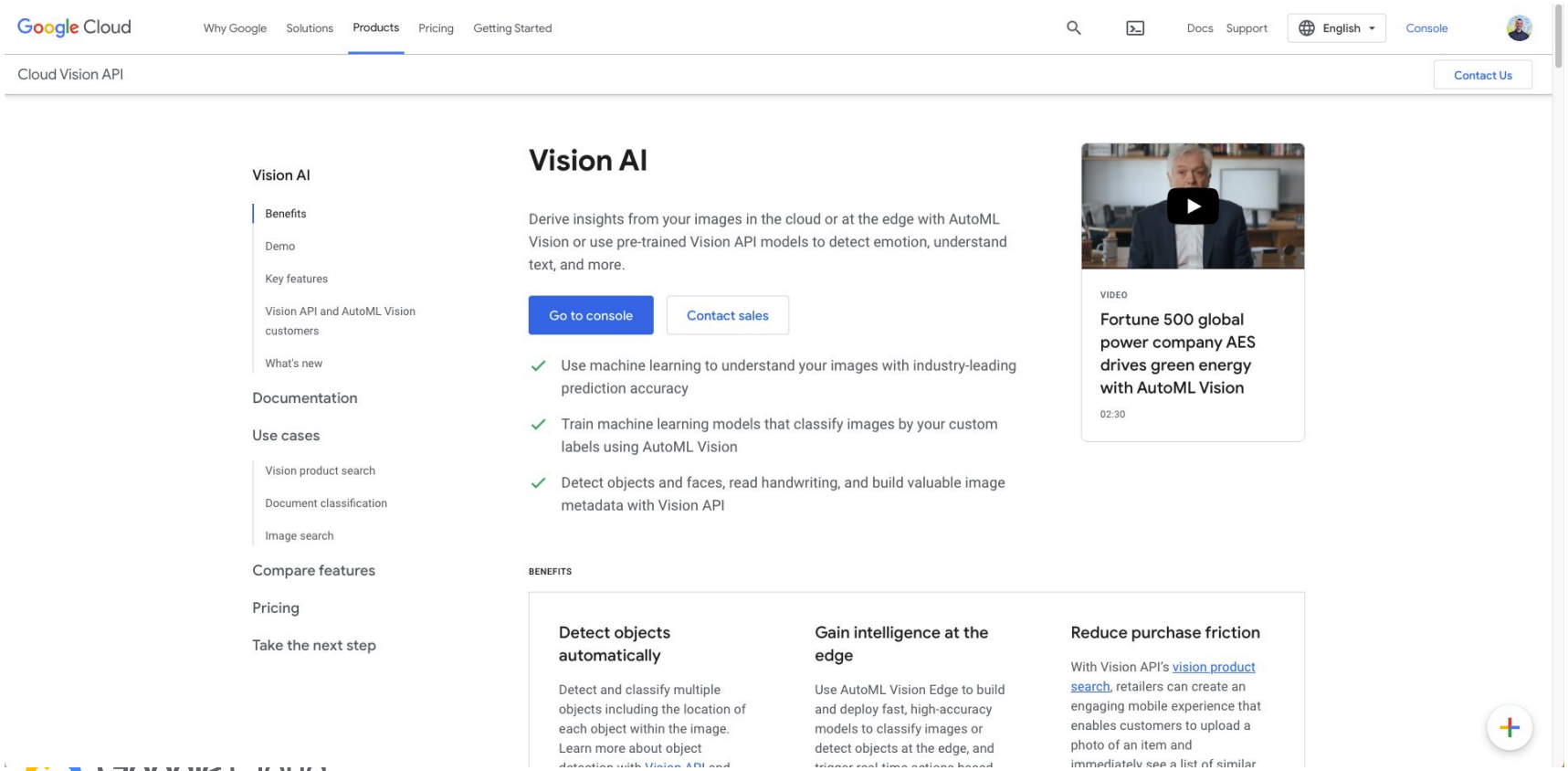
Google Cloud

Let’s try the Vision API

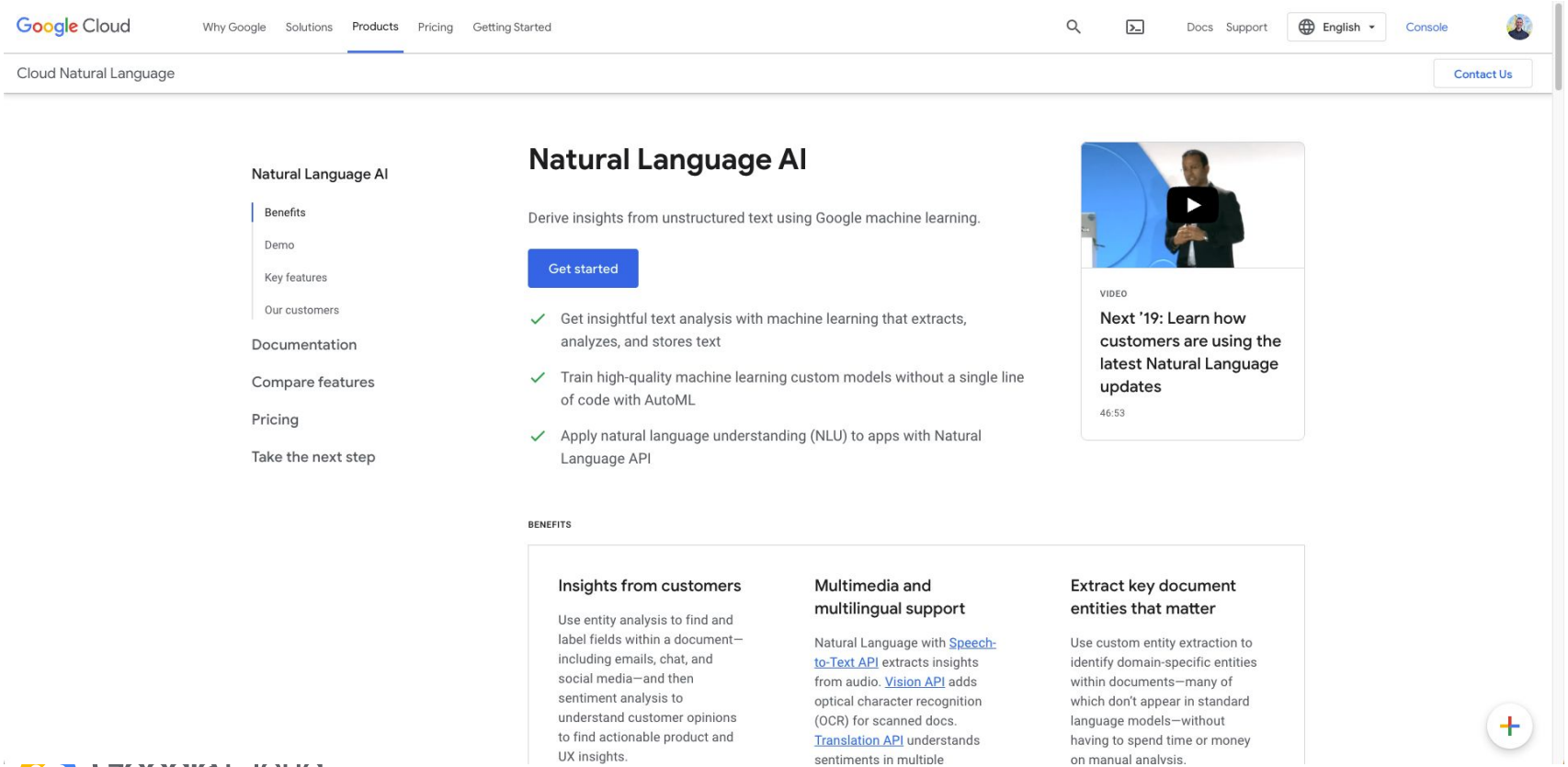
Sky	99%
Cloud	98%
Daytime	97%
Blue	97%
Atmosphere	93%

Google Cloud

Vision AI (Image) Demo



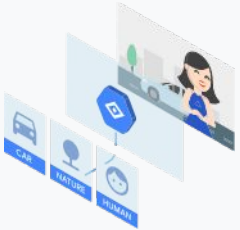
NLP Demo


Google Cloud
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

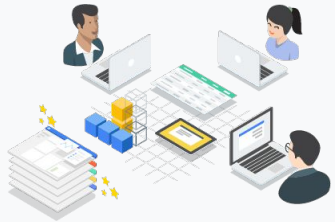
* Offline / Edge Compatible

Serverless

SQL-Based Machine Learning

Serverless

Vertex AI Platform



* Offline / Edge Compatible

Serverless Deployment

Developer-focused

Analyst

Data Scientist

Google Cloud

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

Sky	99%
Cloud	98%
Daytime	97%
Blue	97%
Atmosphere	93%

Google Cloud

Machine Learning APIs:
Ready to Go

Cloud
Vision API

Cloud
Speech API

Cloud
Translation API

Cloud Natural
Language API

Cloud Video
Intelligence API

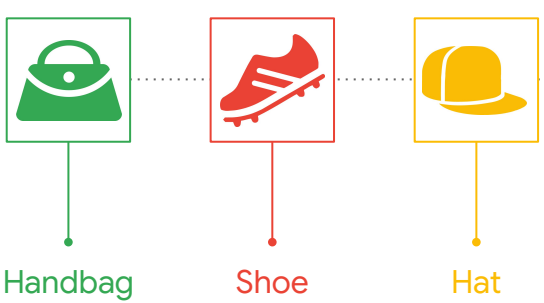
AutoML:
Bring Your Own Data
(We Do the Rest)

AutoML Vision

Upload and label images

Train your model

Evaluate



Cloud AutoML

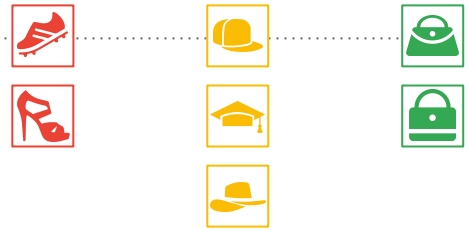
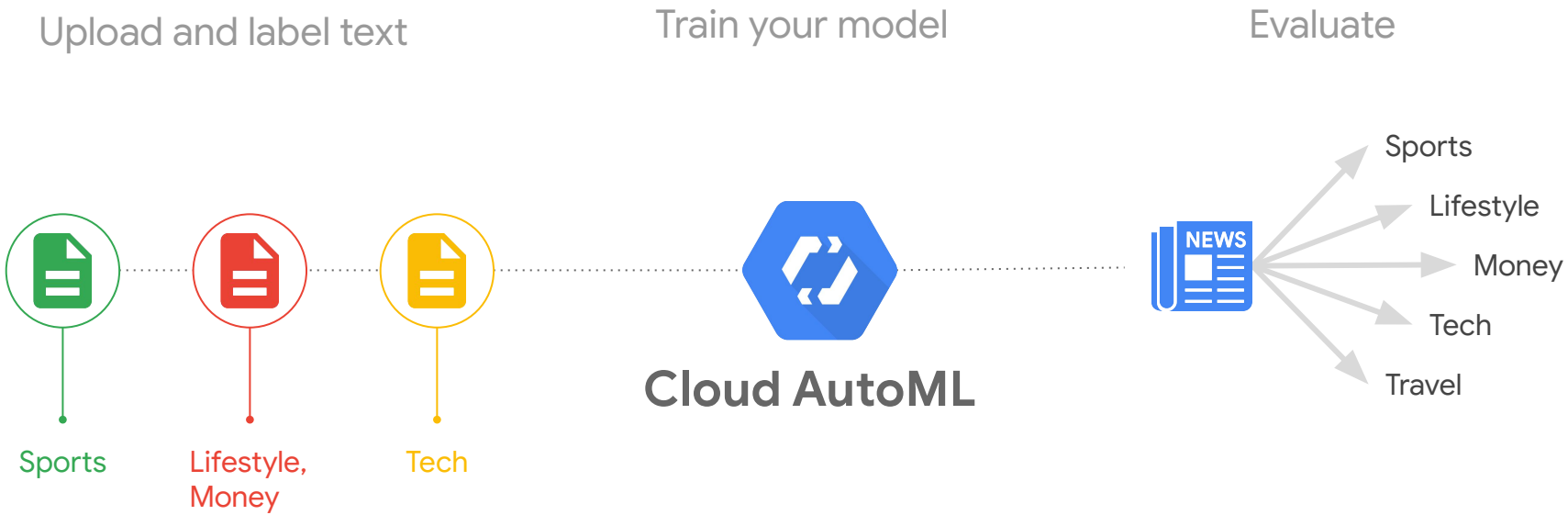


Image Classification
Object detection
Image segmentation

Google Cloud

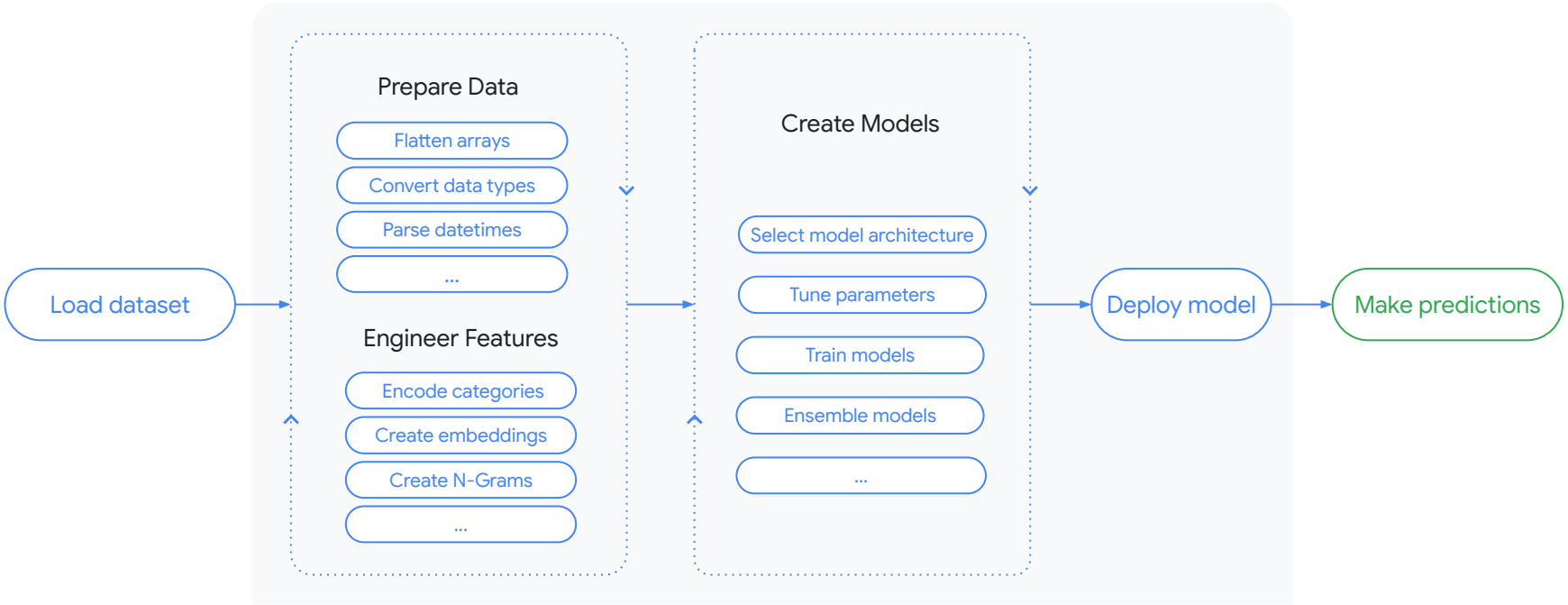
AutoML Natural Language



Google Cloud

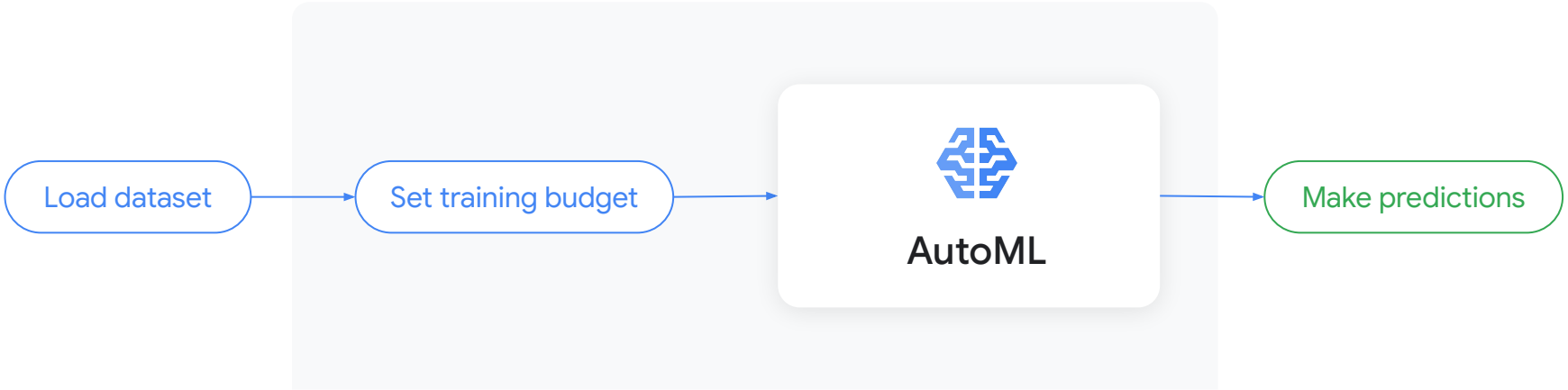
Fastest path from data to value

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

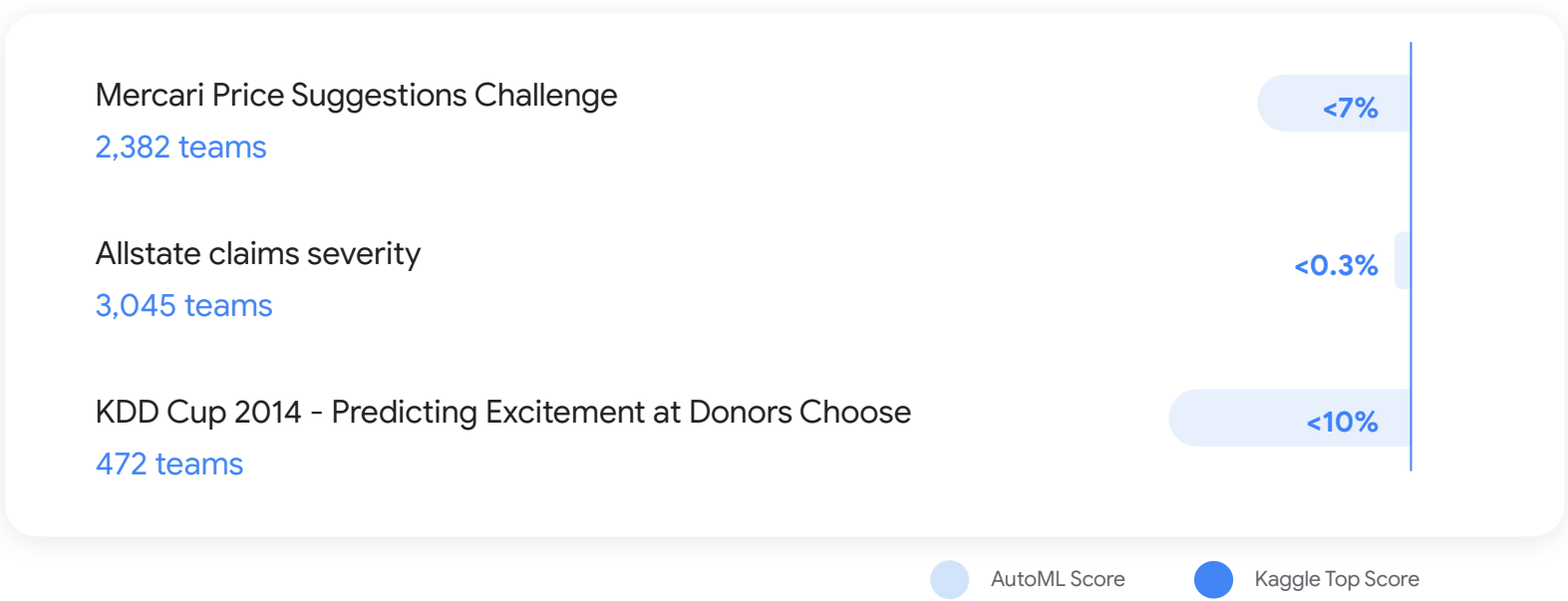
Resilient to + guardrails for:

- Imbalanced data
- Highly correlated features
- Missing values
- High cardinality features (like IDs)
- Outliers

Google Cloud

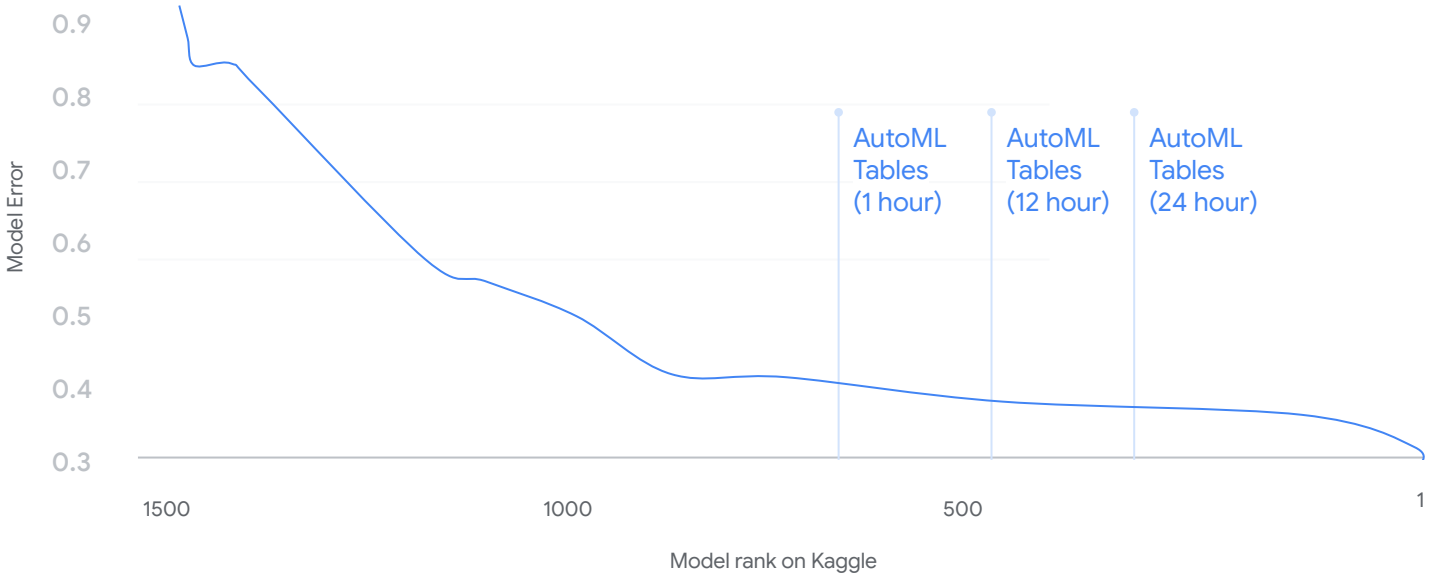
High quality models for real world problems

AutoML Scores Within 10% of Top Kaggle Scores



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

Google Cloud Platform | axial-module-342212 | Search Products, resources, docs (/)

Vertex AI

Dashboard

- Dashboard
- Datasets
- Features
- Labeling tasks
- Workbench
- Pipelines
- Training
- Experiments
- Models
- Endpoints
- Batch predictions
- Metadata

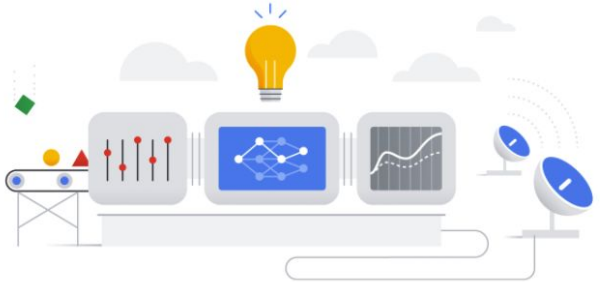
Marketplace

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

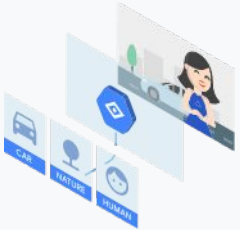
+ CREATE BATCH PREDICTION

Google Cloud

Your Data + Pretrained Model

Your Data + Custom Model

Machine Learning APIs



Serverless

AutoML

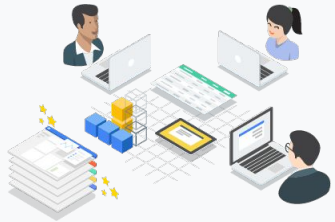
* Offline / Edge Compatible

Serverless

SQL-Based Machine Learning

Serverless

Vertex AI Platform



* Offline / Edge Compatible

Serverless Deployment

Developer-focused

Analyst

Data Scientist

Google Cloud

Accessible Machine Learning for **Analysts**

Machine Learning Using **SQL** in BigQuery

Google Cloud

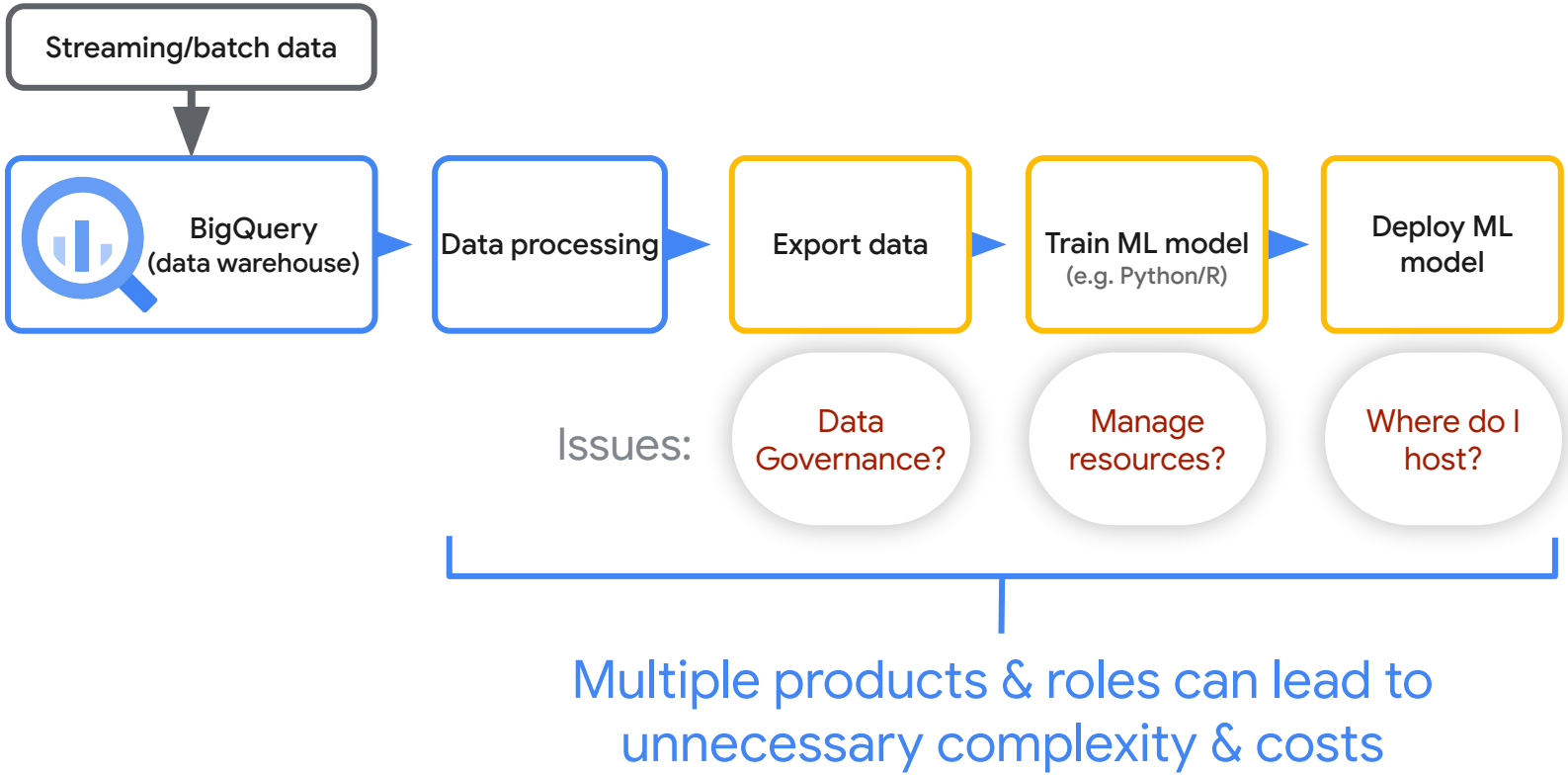
Google BigQuery

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

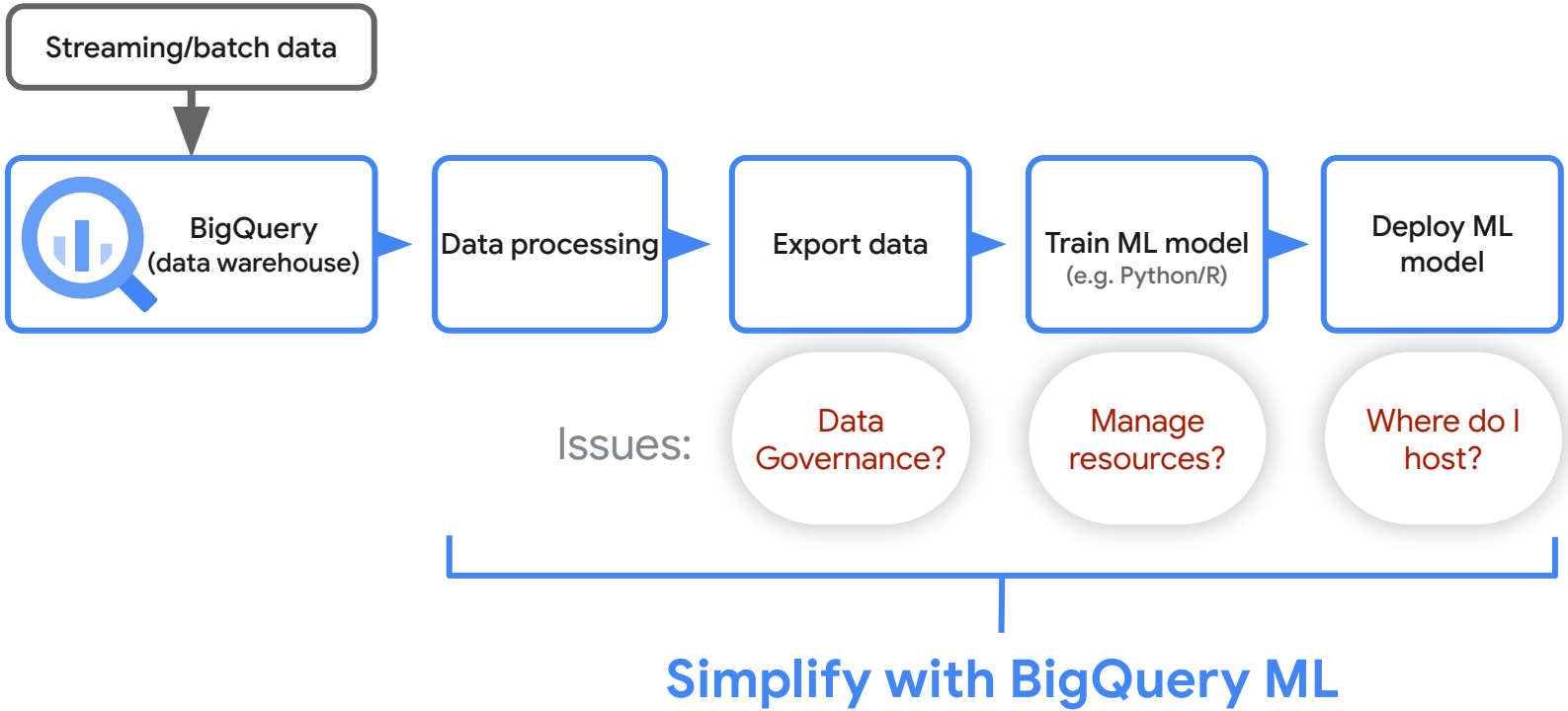
- Easy scaling of storage and compute
- Full ANSI SQL (inc. DML and DDL)
- Built in security and encryption
- High availability and durability

- **Cross-cloud** Lakehouse
- **Real-time** insights
- Built-in **Machine Learning**
- **Insights** for everyone

Typical ML Workflow



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

This is a partial full list of supported models and features. For complete details check out BQML Docs

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

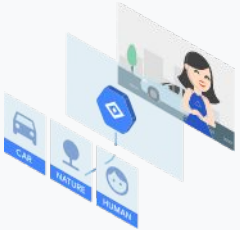
Can be returned to an application, browsed, or stored in BigQuery

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

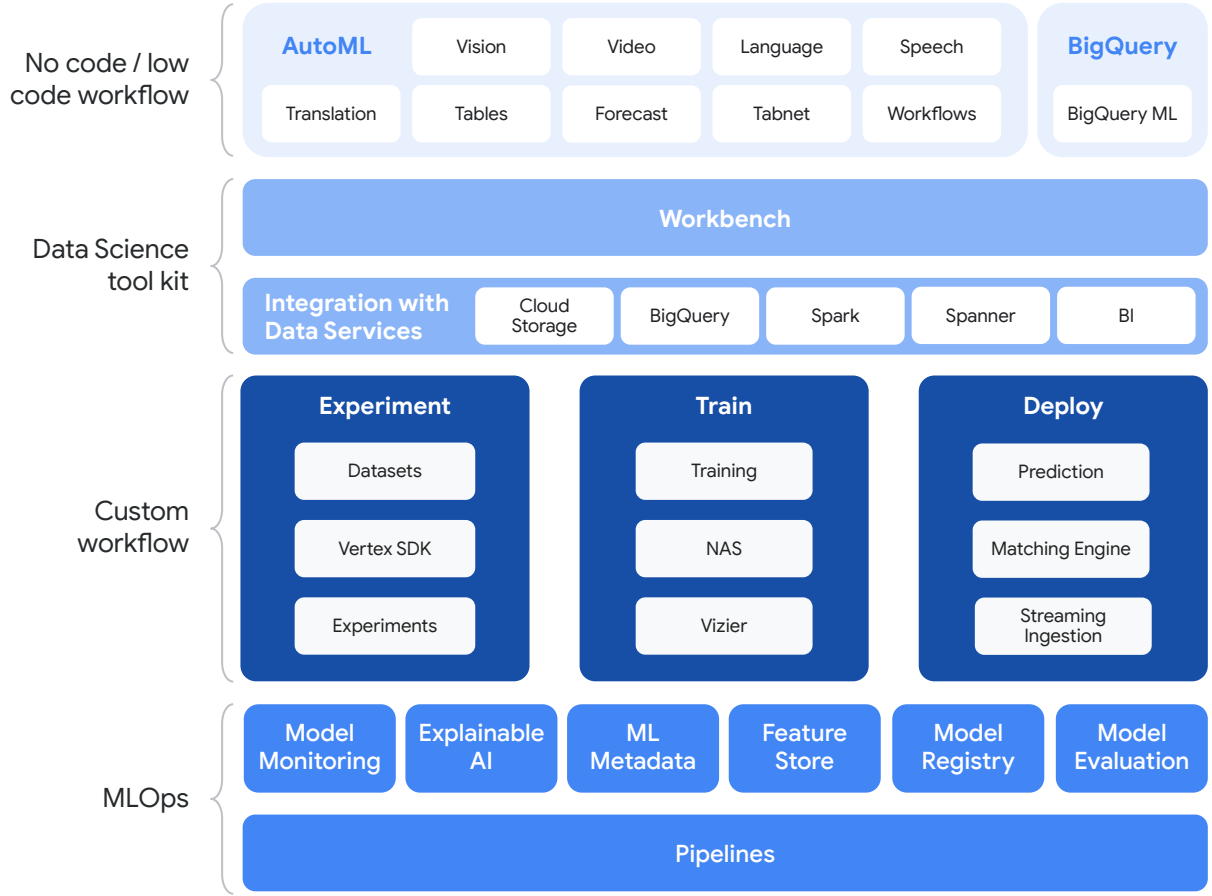
Data Scientist

Google Cloud

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

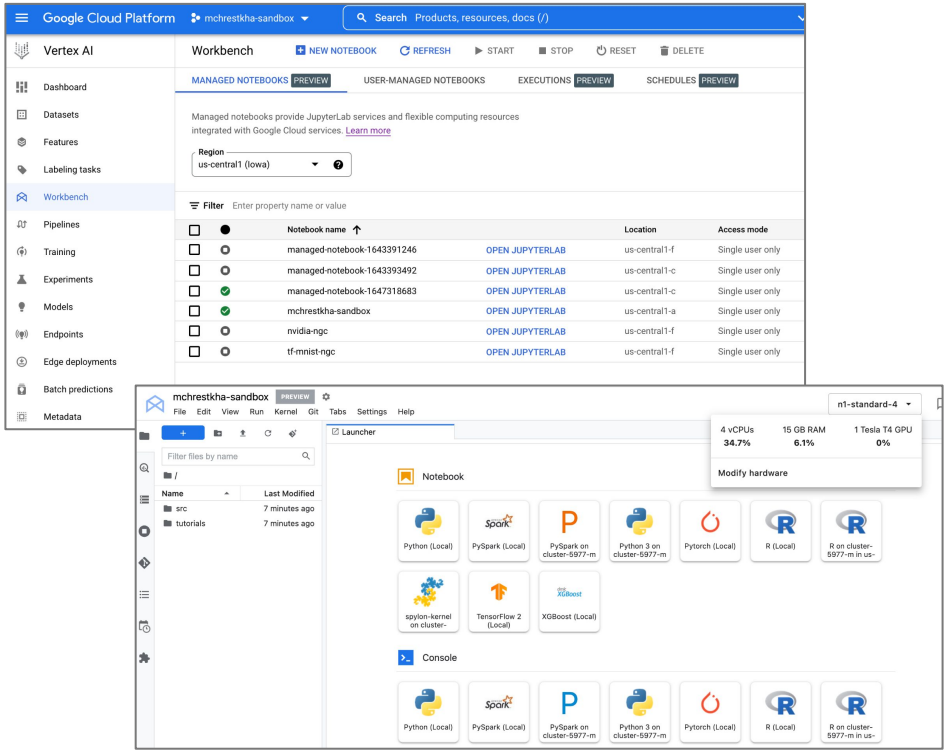
A Jupyter-based fully managed, scalable, enterprise-ready compute infrastructure with easily enforceable policies and user management

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



Serverless Spark on Vertex AI Workbench

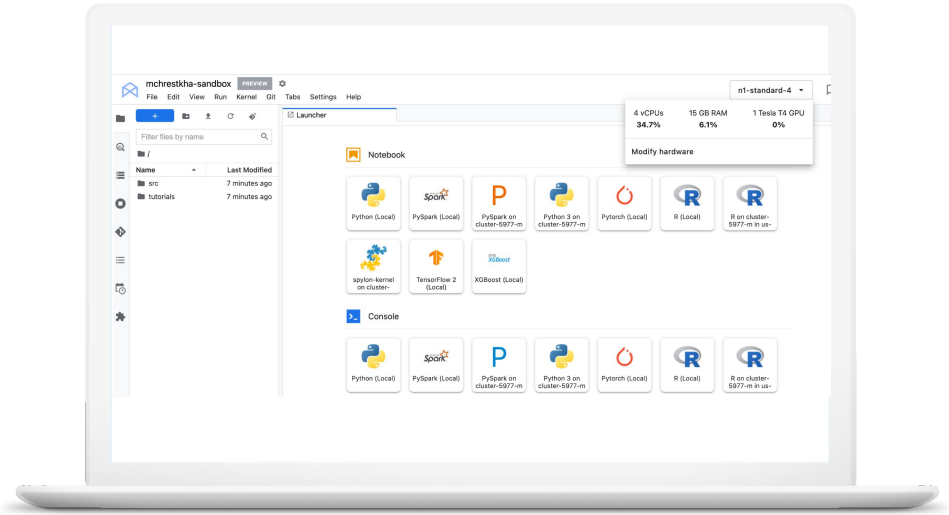
Spark for Data Science in one click

Built-in security and authentication

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

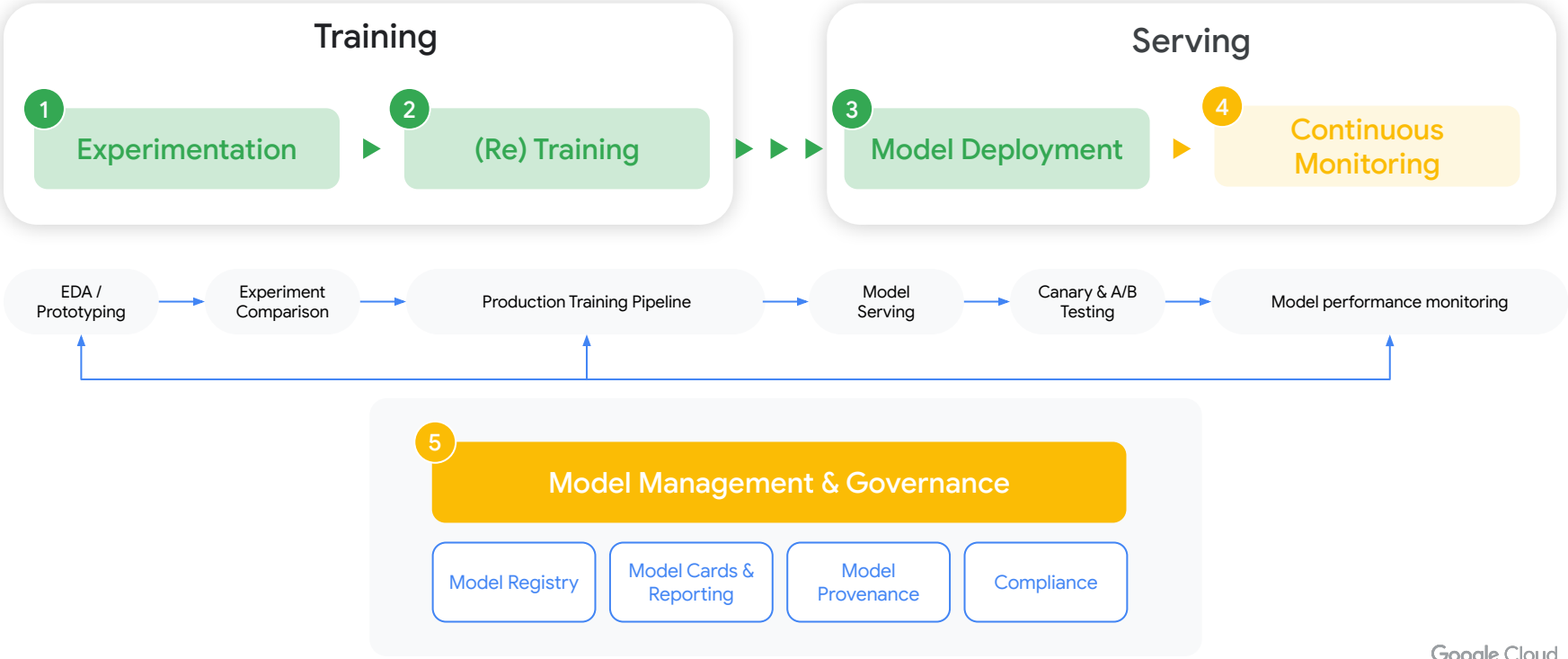
Integrate Spark with MLOps

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

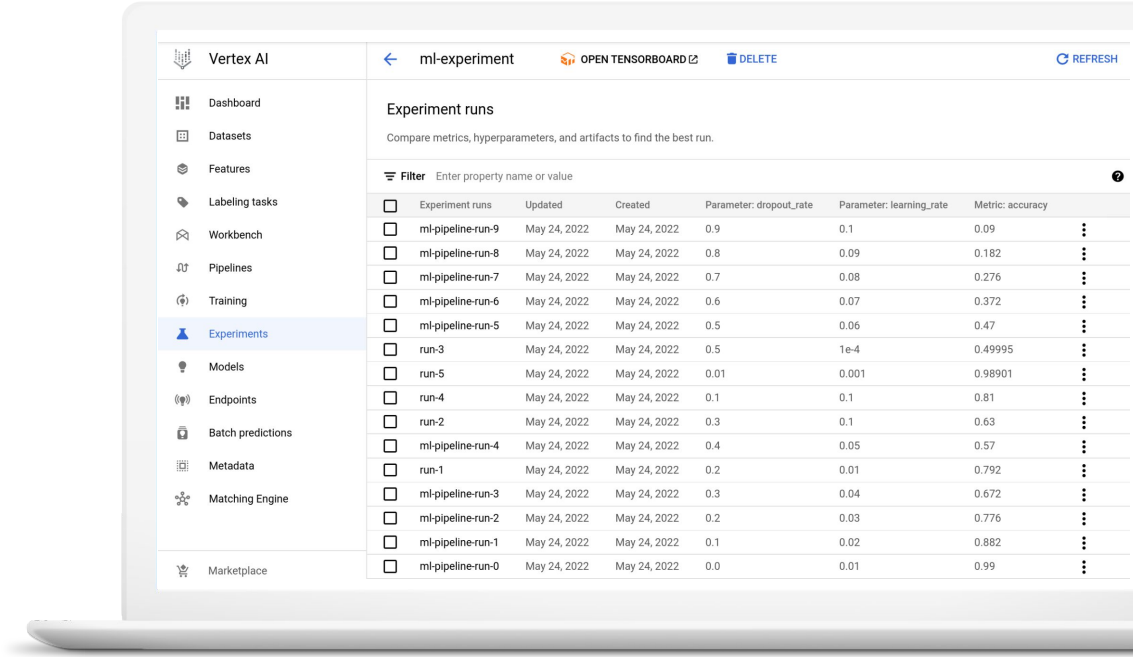
Track and compare multiple experiment runs and analyze key model metrics

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.



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

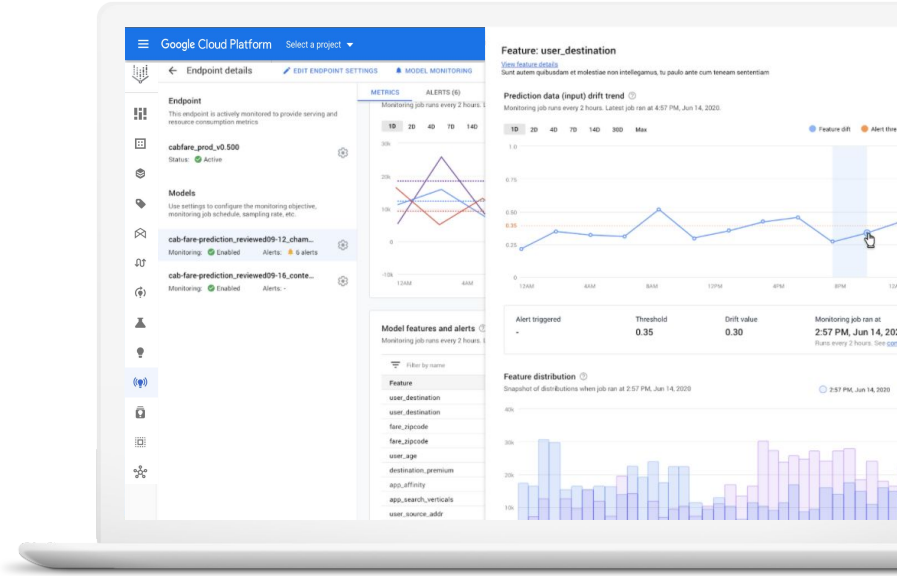
Help identify the cause for the deviation i.e. what changed, how and how much?

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).

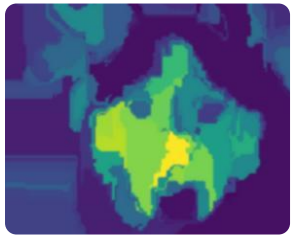
Built into multiple Vertex AI services

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

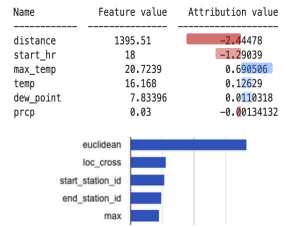
Flexible, fast & scalable

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

Images



Tabular

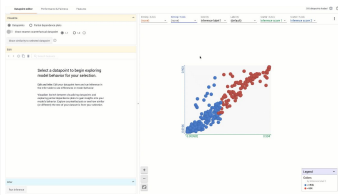


Text

The cake tastes delicious!

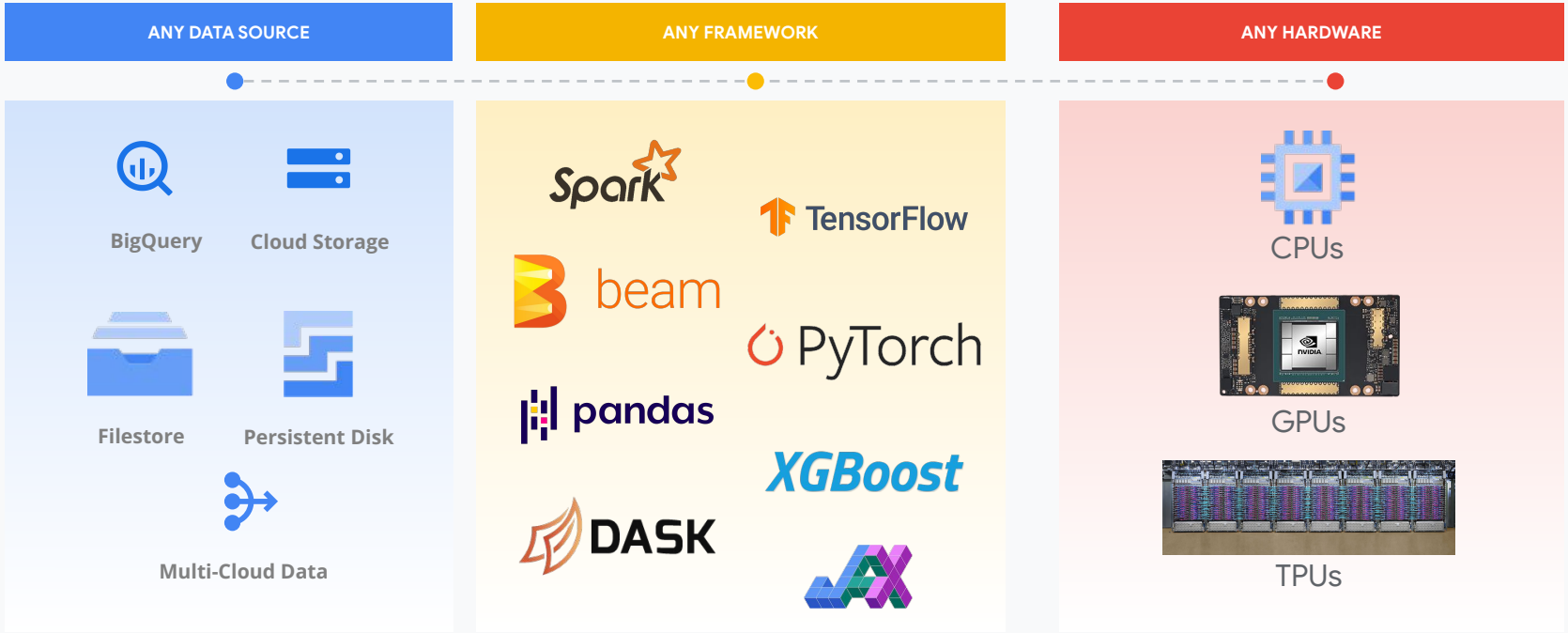
Sentiment score:
0.9

What-if Tool

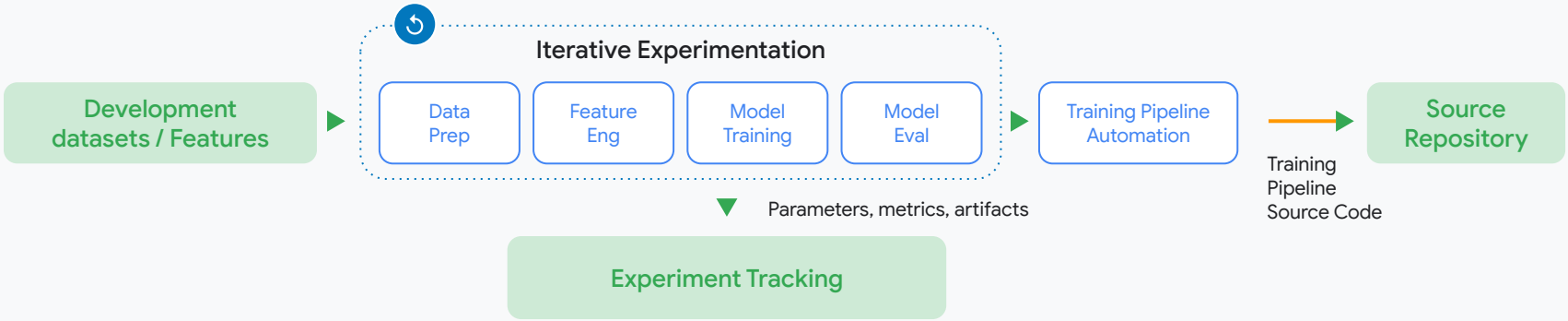


Accelerate the velocity of models to production

While maintaining the flexibility of ML frameworks and compute options



Experimentation



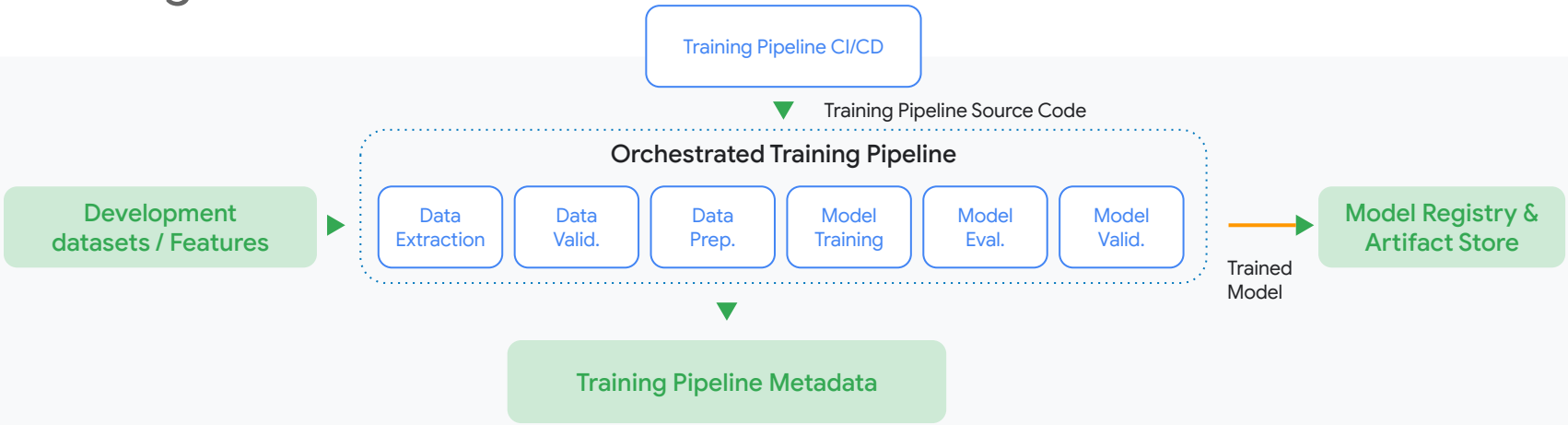
Tools and services

- 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

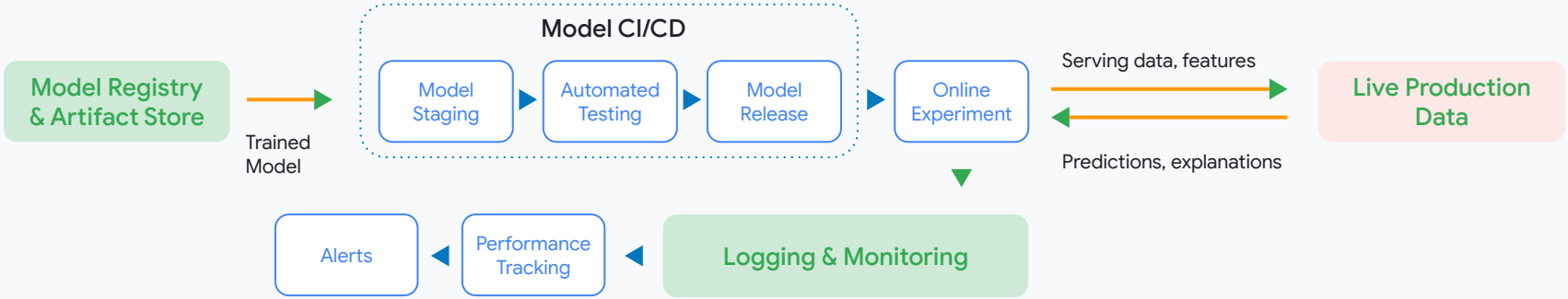
- Vertex Pipelines
- Vertex Training
- Cloud Build
- Artifact Store
- Vertex Explainable AI
- Vertex ML Metadata
- Vertex Feature Store

Key Artifacts

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

Google

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

Google

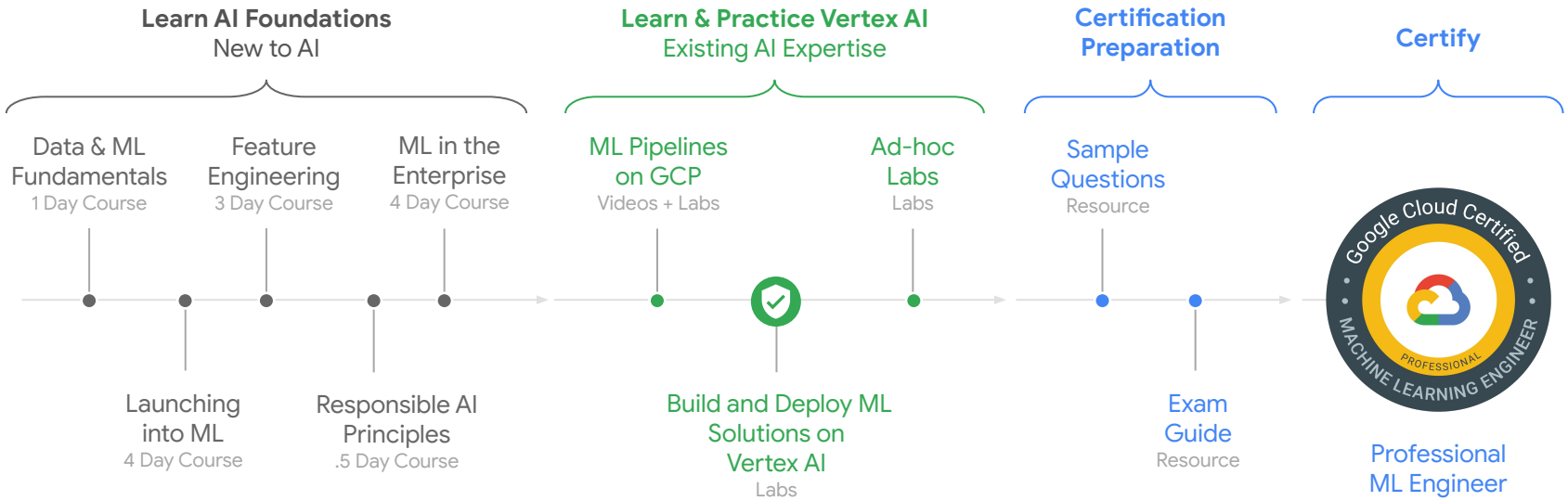
Training on Vertex AI

Train with	Data Analyst	Software Developer	Data Scientist	Use when
Vertex AI Training			✓	• You need more flexibility and customization than the the criteria listed below for BigQuery ML or AutoML. • You're already running training on-premises or another cloud, and you need consistency across the platforms.
AutoML		✓	✓	• Your problem fits into one of the types AutoML supports. Offers a point-and-click workflow. • Natural Language or Video models are served from Google Cloud. While Vision and Tables support edge / downloadable models.
BigQuery ML	✓	✓	✓	• All your data is contained in BigQuery. • Users are most comfortable with SQL. • The set of models available in BigQuery ML matches the problem you're trying to solve.

Google

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