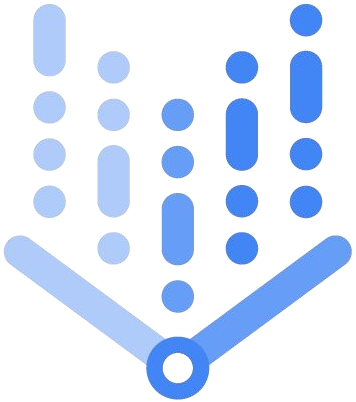
Vertex AI for ML Workloads

Vertex AI for ML Workloads

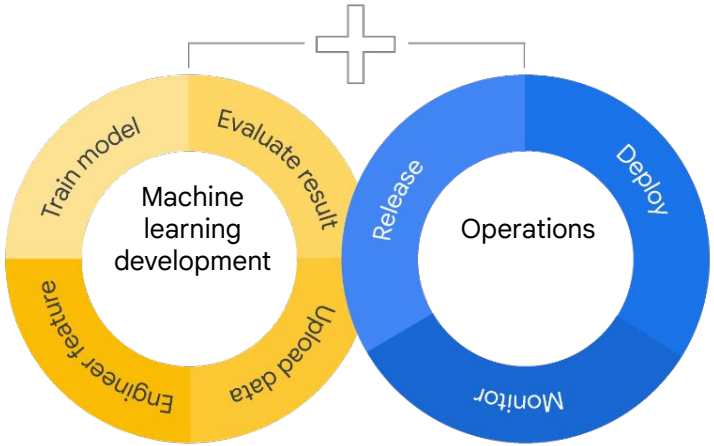


Vertex AI

Google Cloud's fully-managed, unified machine learning platform

Improve models with MLOps

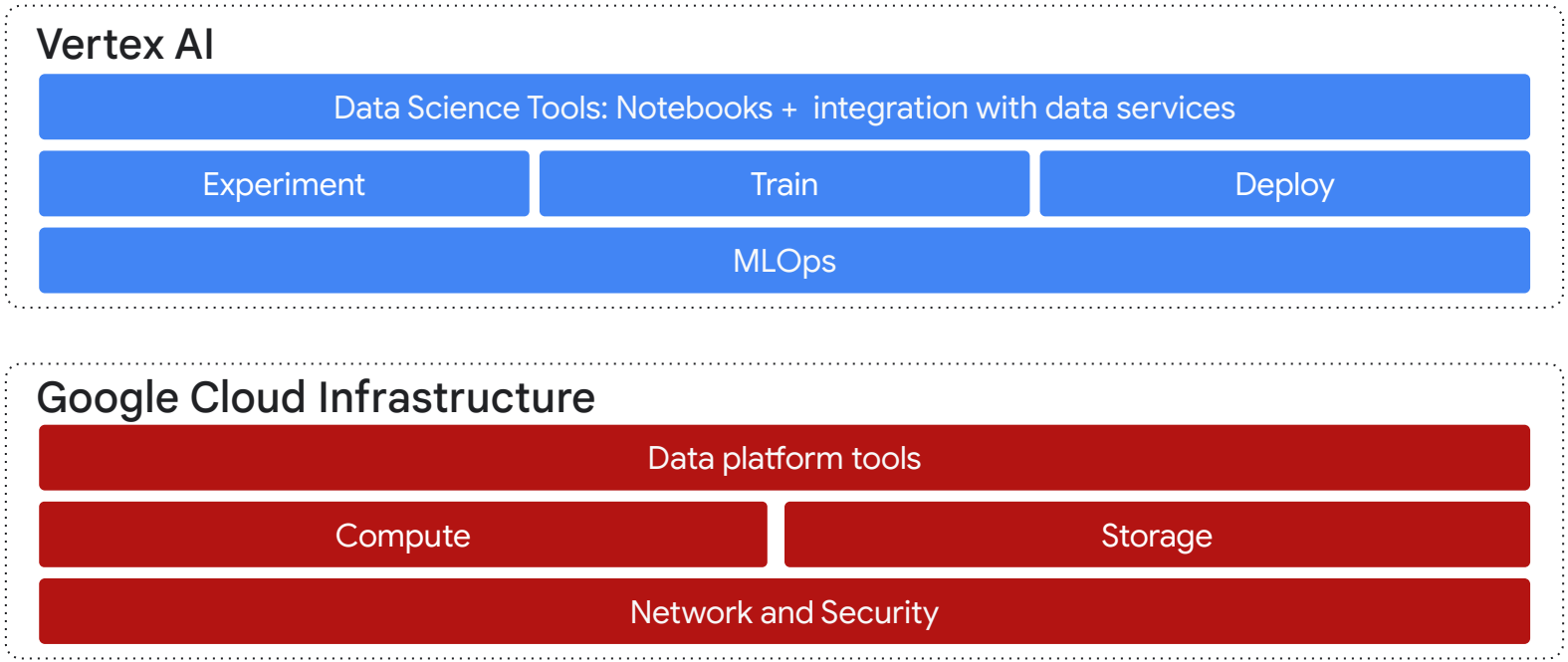
Models need to evolve over time, which requires a process for managing change



MLOps phases:

- Continuous integration (CI)
- Continuous training (CT)
- Continuous delivery (CD)

Vertex AI provides an easy-to-use platform



Machine learning workflow

You can use Vertex AI to manage the following stages:

Manage and serve data for faster model training.

Train an ML model on your data:

- Train the model
- Evaluate model accuracy
- Tune hyperparameters

Upload and store your models in Vertex AI.

Send batches of data to your models for prediction.

Deploy your trained model to an endpoint for online predictions.

Split traffic between models for testing.

Manage your models and endpoints.

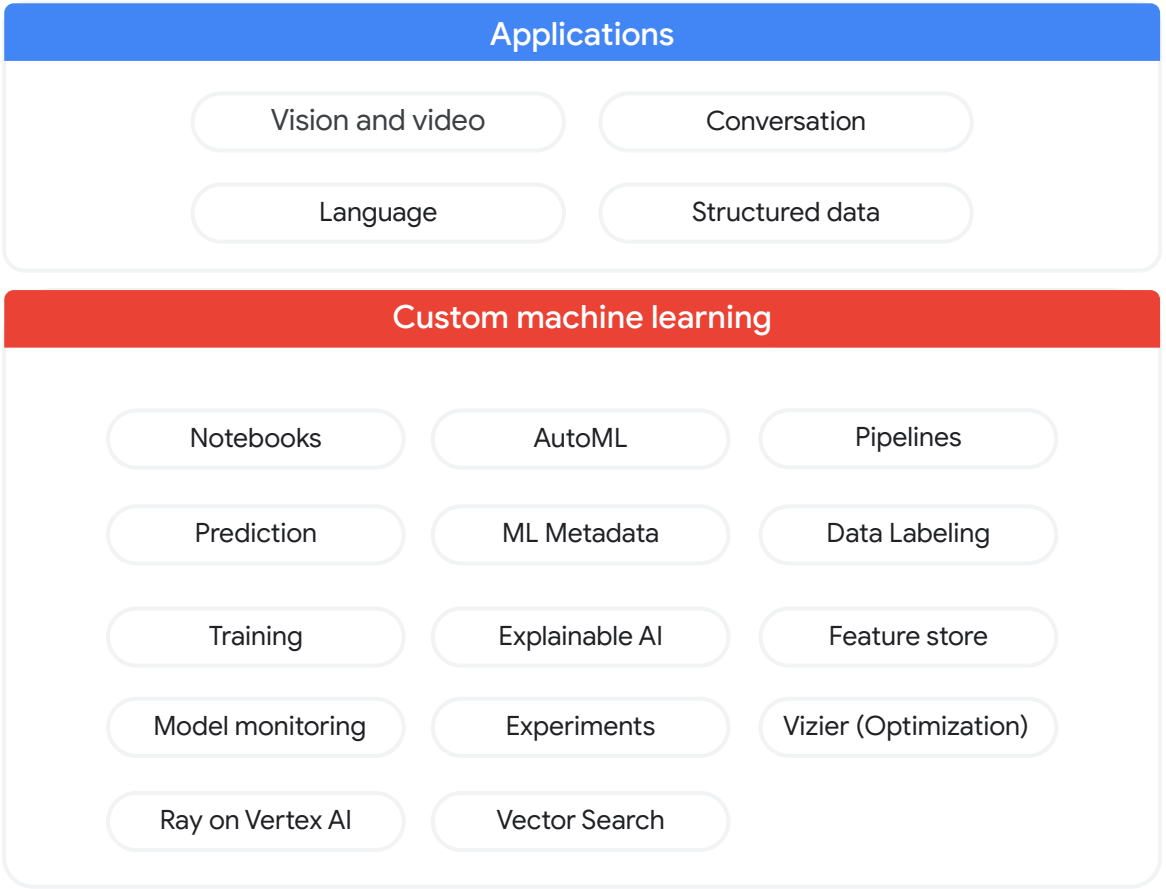
Vertex AI for ML Workloads

01 Vertex AI on Google Cloud

02 Vertex AI Model Workflow Tools

03 Vertex AI for Generative AI

Vertex AI



Notebooks on Vertex AI: Workbench and Colab Enterprise

Exploratory Analysis

Use industry-standard Jupyter notebooks to explore your data and further transform it at scale with other Google Cloud services.

Experiment Locally

Build and test your models locally, then migrate the training to use the full power of Google Cloud.

Collaborate

ML is not a solo task. Share notebooks with your team when you need more eyes on a task, or for inspiration when you find something interesting!

Colab Enterprise

67% of Fortune 100 companies already using Colab, the usual capabilities plus the enterprise-ready security, data controls, and reliability of Google Cloud.

Feature Store

01

Ingest Features

Batch? No problem.
Streaming? Of course.

02

Serve Features

Serve features for training.

Serve features for prediction.

03

Share Features

Don't reinvent the feature. Share it!

Training: AutoML

Image	Video	Text	Tabular
• Classification • Object detection	• Action recognition • Classification • Object tracking	• Classification • Entity extraction • Sentiment analysis	• Classification • Regression • Forecasting

Training: Custom

01

Prebuilt Runtimes

TensorFlow, PyTorch, scikit-learn, and XGBoost environments are provided.

02

Custom containers

Need more control of the environment for your training? Hand Vertex AI a container image!

03

Hyperparameter tuning

Prebuilt or custom, tune your model's hyperparameters as part of the training job.

04

Distributed training

Models can also be trained on large datasets using distributed infrastructure with Vertex AI.

Training: Extra services

ML Metadata

Store training process artifacts, configurations, and metrics for posterity and repeatability.

Experiments

Compare different models' performances across a range of trial configurations.

Vizier

Explore hyperparameter values systematically to help find optimal configurations.

Model Serving

01

Model Registry

Store and version trained models.

02

Predictions

Batch or streaming, Vertex AI can provide compute power to meet your model's needs.

03

Explainable AI

Discover which features contribute most to predictions.

04

Model Monitoring

Track model inputs for statistical similarity to training data.

Vertex AI Pipelines

Directed Acyclic Graphs (DAG)

Automate, monitor, and experiment with interdependent parts of a ML workflow by building a DAG of steps.

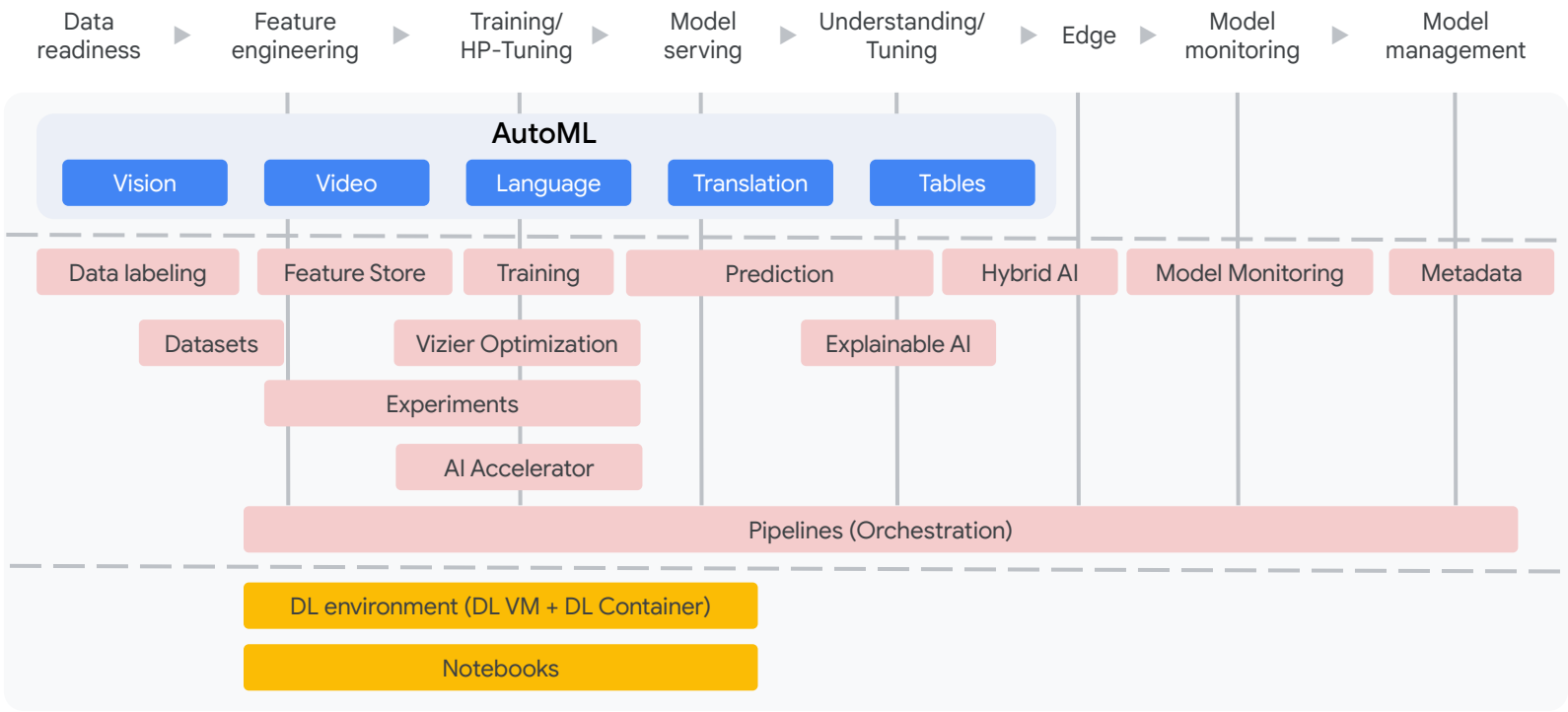
Scalability

Google Cloud serverless power for running your workflows.

Reusability

Pipelines can accept variable inputs so are easy to reuse for different, but similar, tasks.

Vertex AI Model Workflow



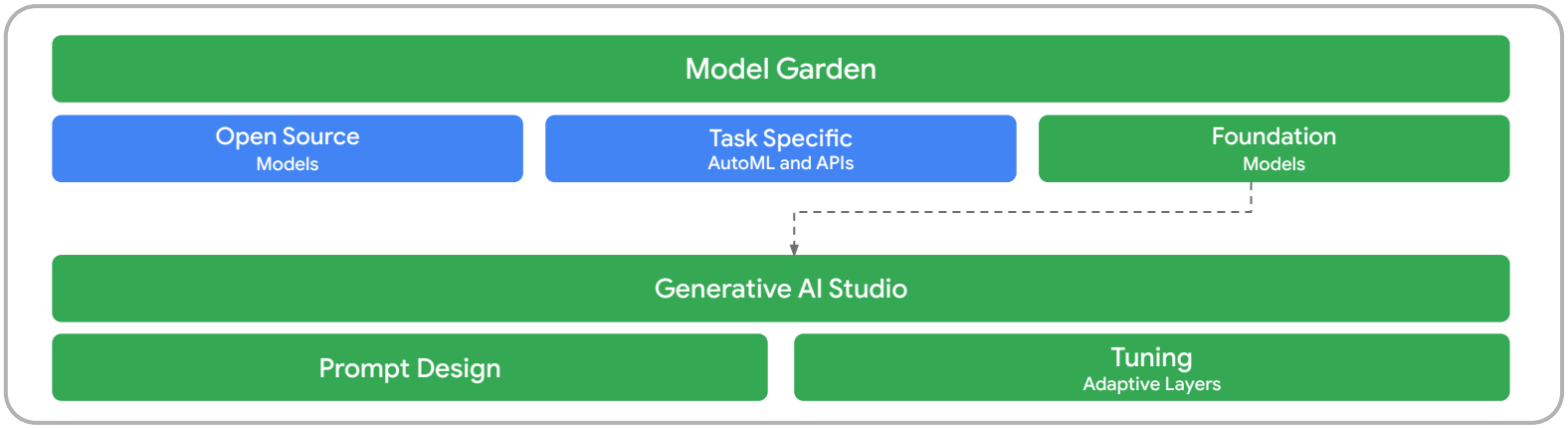
Vertex AI for ML Workloads

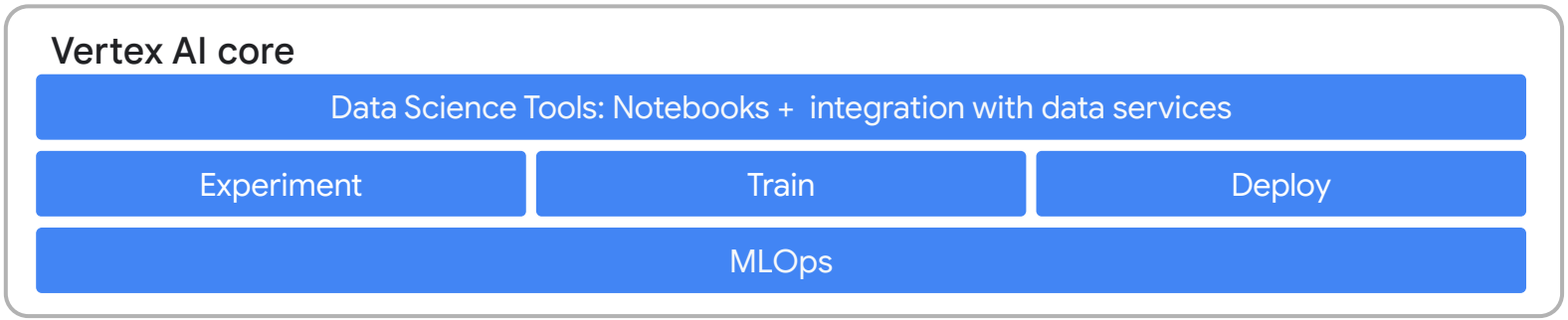
01 Vertex AI on Google Cloud

02 Vertex AI Model Workflow Tools

03 Vertex AI for Generative AI

Generative AI with Vertex AI





Model Garden

Foundation	Pre-trained APIs	Open Source	Third-Party
Google's state-of-the-art multimodal models	Google's pre-trained, task-specific models	Enterprise-ready open source models	Coming soon!

Generative AI Studio

Explore

Test models, build prototypes, and iterate on designs. No coding or deployment required!

Tune

Bring your own data to tune the model outputs.

Deploy

Custom models are in Model Registry, ready to be deployed!

Data Governance

We [Google] don’t use data that you provide us to train our own models without your permission.

[Google AI/ML Privacy Commitment]

Responsible AI

https://ai.google/responsibility

01	Be socially beneficial.
02	Avoid creating or reinforcing unfair bias.
03	Be built and tested for safety.
04	Be accountable to people.
05	Incorporate privacy design principles.
06	Uphold high standards of scientific excellence.
07	Be made available for uses that accord with these principles.