The Team & Agenda

Camus Ma
AI/ML Lead (Google Cloud)

Ayan Kar
EMEA Head of Data & AI (Capgemini)

Luc Schamhart
Account Manager Nationale-Nederlanden

Turan Bulmus
AI/ML Practice Lead Benelux

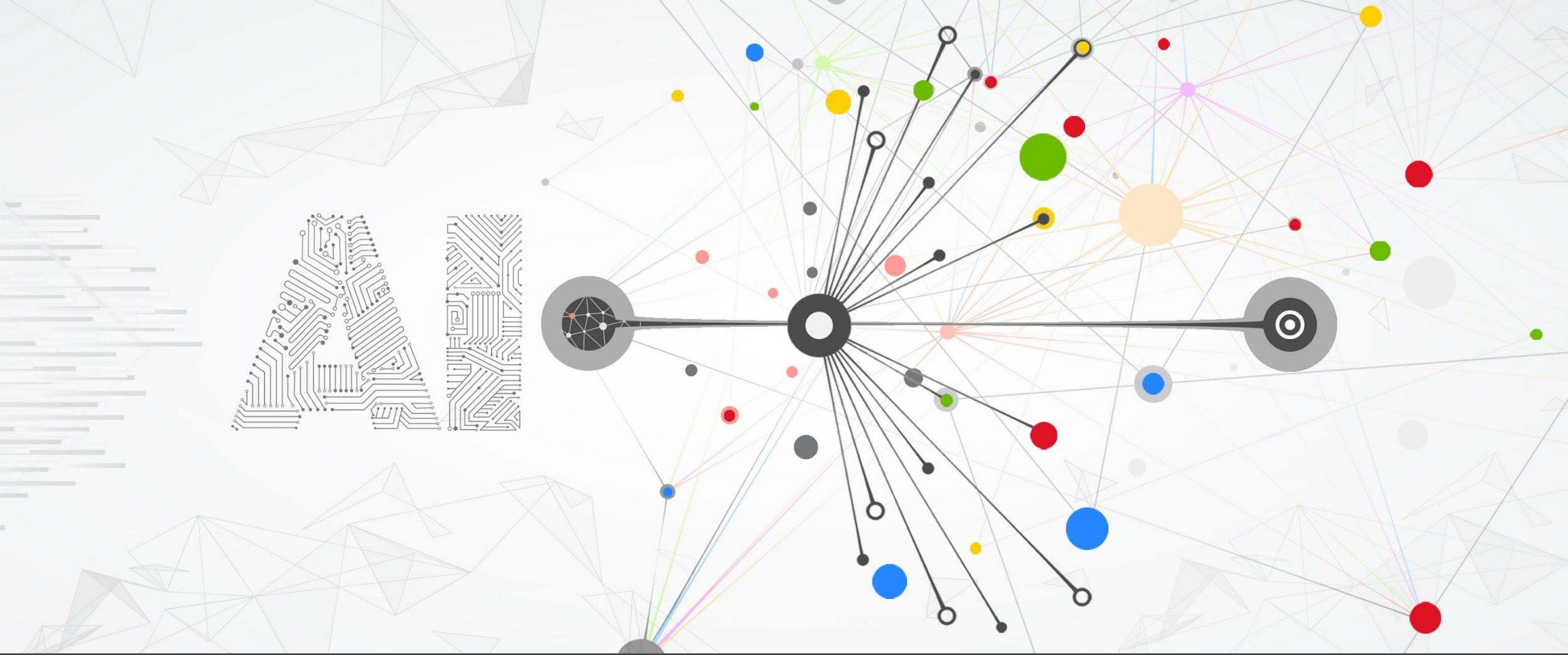
Marijn Markus
AI/ML Consultant (Capgemini)

Joost Carpaij
Insight/Data Consultant (Capgemini)

1. 10:00 - 10:25 - Google Cloud AI Adoption Framework - Camus
2. 10:25 - 10:50 - Applying AI governance & ML Opps.& Glassbox Demo - Ayan & Bikas

10:50 - 11:00 - Stretch legs

3. 11:00 - 11:45 - Google Cloud Vertex Framework -Turan (incl. demo)
4. 11:45 - 11:55 - Experience @ NN (Life & Pensions) Joost
5. 11:55 - 12:20 - Cool Demo's Marijn
6. 12:20 - 13:00 - Lunch



Google Cloud’s AI Adoption Framework

Gain a competitive advantage through AI

Enterprises that invest in building industry-specific AI solutions are proven to be better positioned as future global economic leaders

Companies that fully absorb AI could double their cash flow

2X more data-driven decisions

5X faster decisions than others

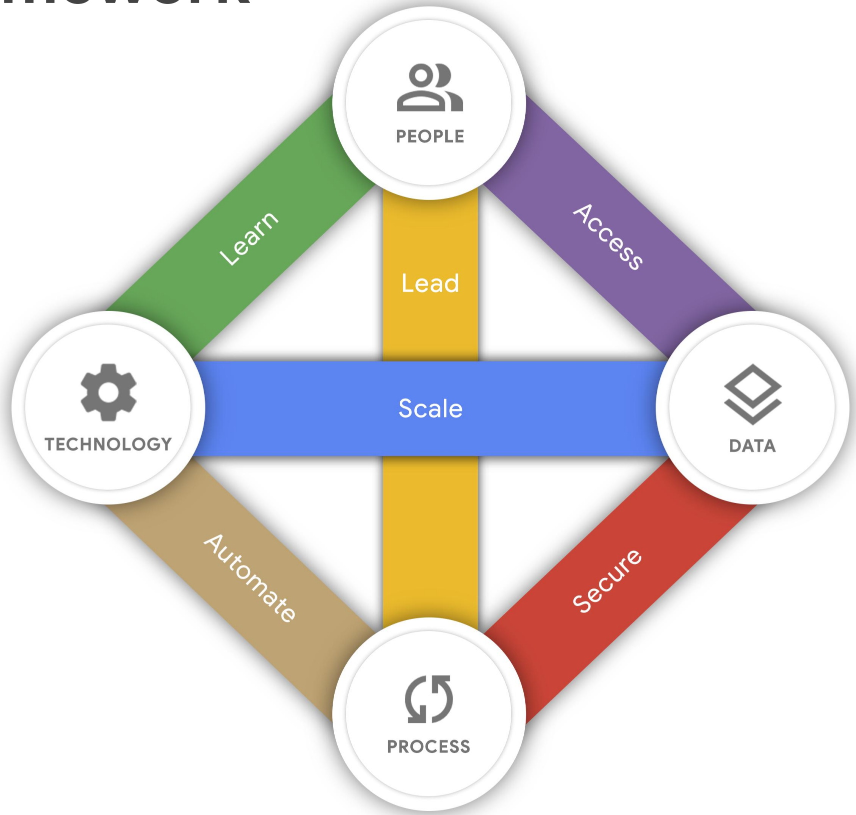
3X faster execution

Sources: McKinsey 2018 and MIT Tech Review 2017

The Cloud AI Adoption Framework

- A **guiding framework** for leaders who want to leverage the power of AI to transform their business

- A **tool to assess** where you are in your journey and **define where you want to be**



AI Maturity Themes

Lead

PEOPLE · PROCESS

How are the teams structured?

What is the level of executive sponsorship?

How is budget allocated for AI/ML projects?

Scale

DATA · TECHNOLOGY

How are cloud-based services provisioned?

How is capacity for workloads allocated?

Does an organization use accelerators?

Learn

PEOPLE · TECHNOLOGY

What the data and ML skill sets are required in the organisation?

How does an organization develop ML talent?

Secure

DATA · PROCESS

What controls are in place?

How does an organisation establish trust in it's AI capability?

Can you explain the decisions made by your AI systems?

Access

PEOPLE · DATA

How are datasets created, curated, and annotated?

Are they discoverable and reusable?

How are data and ML assets managed?

Automate

TECHNOLOGY · PROCESS

How models are continuously trained and deployed for serving?

How are model updates managed?

What ML quality control are in place?

AI Maturity Phases

Tactical	Strategic	Transformational
• Focus on narrow, simple use cases • Good foundational skill set for core data wrangling and descriptive analytics • Benefit from improved actionable insights from data	• Several ML systems deployed and maintained in production • Degree of centralised coordination through an Advanced Analytics team • Customised models can be a source of competitive advantage	• Continuous ML training and serving (MLOps) • Culture where experimentation and learning is continuous • Cutting-edge research, a point of differentiation to attract the best talent

The Cloud AI Maturity Scale

Lead

PEOPLE

Lead

PROCESS

The theme concerns the extent to which your data scientists are supported by a mandate from leadership to apply ML to business use cases, and the degree to which the data scientists are cross-functional, collaborative, and self-motivated.

Tactical

- AI adoption driven by individual contributors
- "Heroic" project manager with team budget
- AI/ML link to business goals not always clear

Strategic

- Creating a centralized cross-functional advanced analytics team to establish common ML patterns and practices
- Senior executive sponsorship and dedicated budget by C-level for innovative projects
- Aligning AI efforts with business objectives and priorities

Transformational

- Endorsement and dedicated budget within each line of business
- Function-specific data science teams with domain expertise, in addition to the centralized advanced analytics team
- Innovation and research teams

Access

The theme concerns the extent to which your organization recognizes data management as a key element to enable AI and the degree to which data scientists can share, discover, and reuse data and other ML assets.

Tactical

- No asset sharing
- Isolated data islands
- Building a data lake

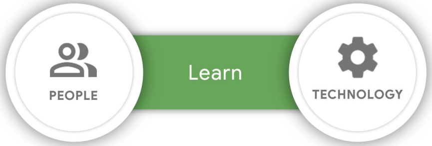
Strategic

- Managing an enterprise data warehouse
- Defining and sharing a unified data model
- Centralized data and ML asset management

Transformational

- Discovering, sharing, and reusing datasets and AI assets
- Standardized ML feature stores and datasets
- Encouraging contributions from across the organization

Learn



The theme concerns the quality and scale of learning programs to upskill your staff, hire outside talent, and augment your data science and ML engineering staff with experienced partners.

Tactical

- Self-motivated, isolated learning using online resources
- Third parties cover the skills gap in the organization
- No hiring for ML skills

Strategic

- Hiring data science and ML skills
- Organizing structured continuous training
- Strategic partner selected to provide consulting and specialized knowledge

Transformational

- Learning by embedding data scientists to the business function teams
- Hiring data science and ML talent for innovation with industry expertise
- Partnering to innovate, co-create, and augment technical resources

Secure

The theme concerns the extent to which you understand and protect your data and ML services from unauthorized and inappropriate access, in addition to ensuring responsible and explainable AI.

Tactical

- Implementing private networks with primitive IAM accessed and managed by a dedicated team
- Ensuring privacy through sensitive data classification and obfuscation
- Enabling data protection through encryption

Strategic

- Implementing principle of least privilege
- Exploring explainable AI techniques
- Investing in establishing AI ethics

Transformational

- IAM continuously monitored and improved
- Considering AI safety and robustness
- Developing fair ML systems

Scale

The theme concerns the extent to which you use cloud-native ML services that scale with large amounts of data and large numbers of data processing and ML jobs, with reduced operational overhead.

Tactical

- Dedicated hardware for cost control
- Working with a limited number of small datasets

Strategic

- Utilizing a fully managed serverless,data warehouse for ad hoc querying and data exploration
- Utilizing fully managed serverless data services for ingestion and processing
- Using fully managed serverless ML services for training and prediction serving

Transformational

- Operating an integrated ML experimentation and production platform
- Using specialized ML accelerators (GPUs, TPUs) on demand
- Orchestrating end-to-end data and ML pipelines

Automate

TECHNOLOGY Automate PROCESS

The theme concerns the extent to which you are able to deploy, execute, and operate technology for data processing and ML pipelines in production efficiently, frequently, and reliably.

Tactical

- Ad hoc, manual data processing and ML model training and serving
- High-risk changes reviewed and deployed infrequently and manually

Strategic

- Automating (scheduled and event-driven) data pipelines
- Automating ML training and batch-prediction pipelines
- Managing logging, monitoring, and notifications

Transformational

- Implementing ML training pipelines with continuous integration and delivery
- Implementing ML prediction services with continuous integration and delivery
- Managing ML model registry, ML metadata, and ML artifacts

Next Steps

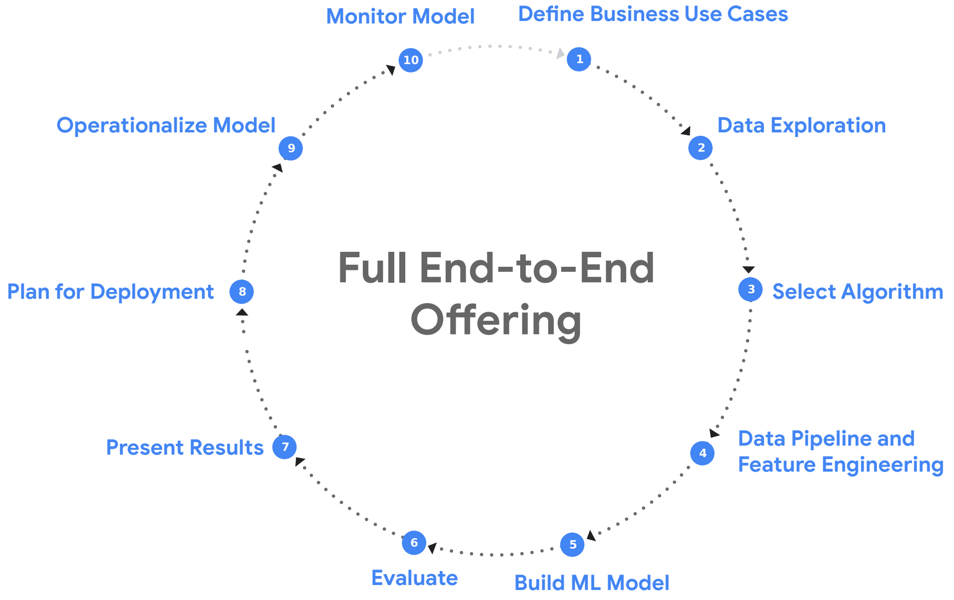
- **Understand where you are -** complete the maturity assessment
- **Set your goal** - where you do you want to go?
- **Create a group of leaders** - who are responsible for building your AI capability?
- **Devise a strategy** - based on the gaps, establish a plan for evolving your AI capability
- **We are here to help you every step of the way** - speak to your account representative about how you can engage Google

Questions?

Appendix

The transformational AI journey offering

AI Services provides a **complete end-to-end path** to accelerate your AI transformation



Address your business challenge from start to finish

- Build for production from the beginning
- Deep understanding of the problem and solution
- Trusted partner throughout the journey

Google Cloud Smart Analytics & AI

AI Solutions

Horizontal solutions
- Document AI
- Contact Center AI
- Dialogflow

Industry solutions
- Talent Solution
- Recommendation AI

Prebuilt ML APIs

Sight
- Vision
- Video Intelligence

Language
- Natural Language
- Translate

Conversation
- Speech-to-Text
- Text-to-Speech

AutoML

Sight
- Vision
- Video

Language
- Translate
- Natural Language

Structured Data
- Tables

AI Platform

Data Science and Machine Learning
- Notebooks
- Data Labeling
- Training
- Prediction
- Continuous evaluation
- Explainability
- Pipelines

Data Analytics & Management

Ingestion and Processing
- Cloud Pub/Sub
- Cloud Dataflow
- Cloud Dataproc
- Data Fusion

Storage and Analytics
- Cloud Storage
- BigQuery
- Cloud Bigtable
- Cloud SQL
- Data Catalog
- Data Studio

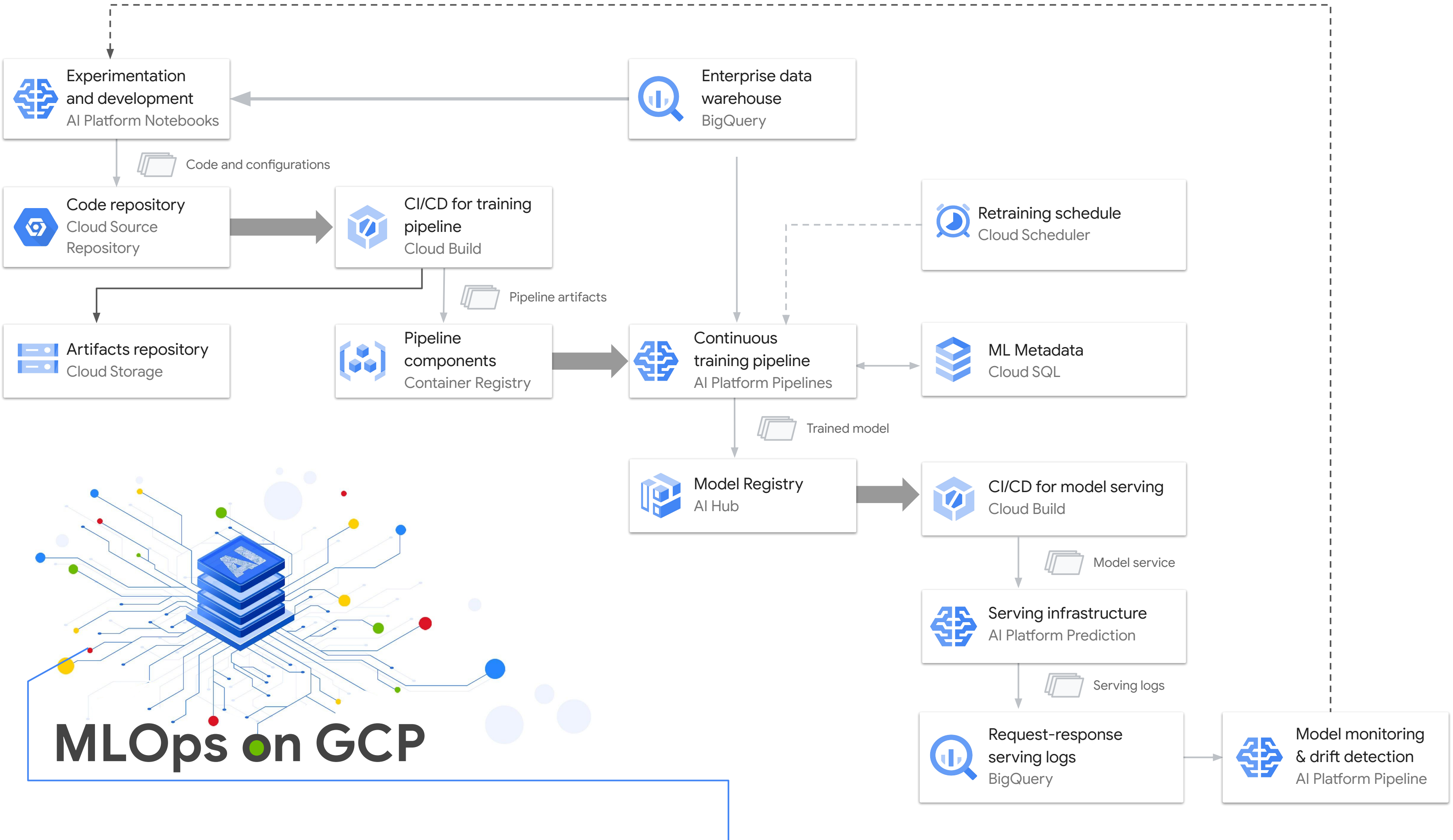
Foundation

Frameworks
- learn
- Spark

Compute
- Compute Engine
- Cloud GPU
- Cloud TPU

Orchestration
- Cloud scheduler
- Cloud Composer

Instrumentation
- Cloud Build
- Container Registry


Experimentation and development
AI Platform Notebooks
Enterprise data warehouse
BigQuery
Code and configurations
Code repository
Cloud Source Repository
CI/CD for training pipeline
Cloud Build
Retraining schedule
Cloud Scheduler
Pipeline artifacts
Artifacts repository
Cloud Storage
Pipeline components
Container Registry
Continuous training pipeline
AI Platform Pipelines
ML Metadata
Cloud SQL
Trained model
Model Registry
AI Hub
CI/CD for model serving
Cloud Build
Model service
Serving infrastructure
AI Platform Prediction
Serving logs
Request-response serving logs
BigQuery
Model monitoring & drift detection
AI Platform Pipeline
MLOps on GCP

ML Quality Control

- **Testing in development**
  - ✓ Data and feature testing
  - ✓ Model testing and debugging
  - ✓ Model evaluation
- **Testing in deployment**
  - ✓ Testing ML pipeline components integration
  - ✓ Validate model-infrastructure compatibility
  - ✓ Test model API
- **Testing in production**
  - ✓ Pipeline data and model validation
  - ✓ A/B testing and performance monitoring
  - ✓ Data drift and shift detection

Google Cloud

Machine Learning and MLOps on Google Cloud

Turan Bulmus
AI/ML Practice Lead Benelux
April 2022

Enable every company to be an AI company by reducing the challenges of AI model creation down to only the steps that require human judgement or creativity.

Build a portfolio of AI use cases

	Use AI out of the box	Deploy custom AI	Build end-to-end AI
Benefit	Maximize the value AI delivers into business workflows	Extract value from your data	Use AI on your data to differentiate your product
Effort	Low effort, high volume	Medium effort, customization	High effort, low volume
Requires	High quality AI building blocks and industry solutions	High quality baselines and ease-of-use	Powerful, easy-to-use platform that allows you to reuse pre-built and/or customizable components

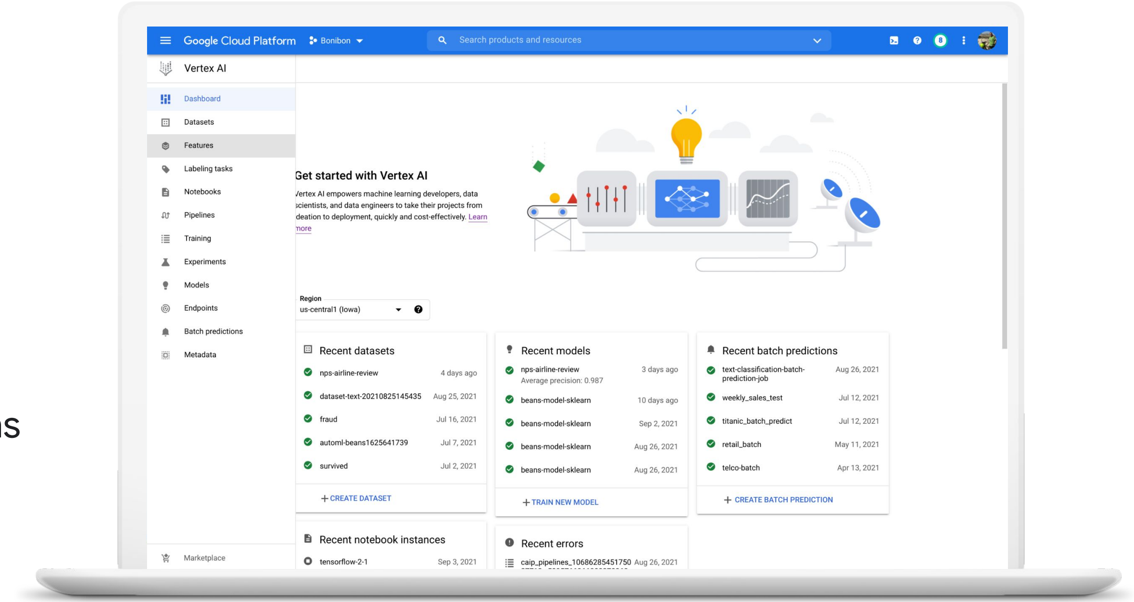
“Business value” generated from all three buckets
Need a unified platform that supports all three buckets

A Unified ML Platform for Solving All Business Problems

Processing all sources of data including images, documents, tables, video

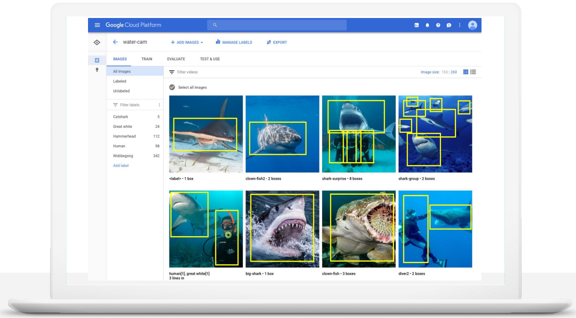
Vertex AI

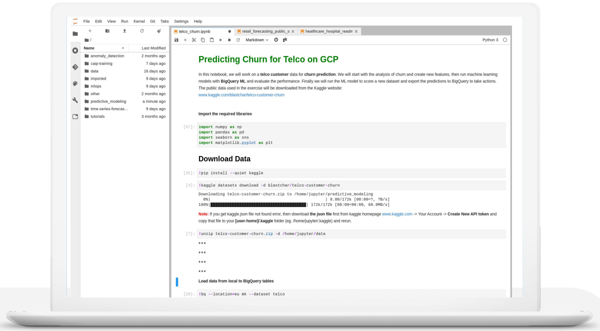
- One unified experience to create, deploy, and manage models over time, at scale
- Tools for all levels of expertise and for all types of data
- Accuracy and fairness of predictions and resulting decisions
- Flexible and secure



How to design your ML workflow?

Freedom of choice from no/low code to custom code





BigQuery ML

- Descriptive and predictive modeling on structured data
- Hyper-parameter tuning
- Feature engineering
- Explainability
- Simple SQL code

AutoML Models in Vertex AI

- Predictive modeling on structured & unstructured data
- Hyper-parameter tuning
- Feature engineering
- Explainability
- No code

End-to-end AI with Vertex AI

- Custom models on pre-built frameworks
- Noops, serverless training with hyperparameter tuning
- Explainability
- Custom code

Why organizations choose Vertex AI

Build on the Best of Google

Access to Google's continuously enhanced AI tools

A platform created upon the foundation of Google's pioneering AI research

Accelerate Time to Value

Make AI more accessible & useful

Flexible tools for streamlined and scalable collaboration across all levels of technical expertise

MLOps capabilities make practitioners jobs easier

Trust & Responsibility

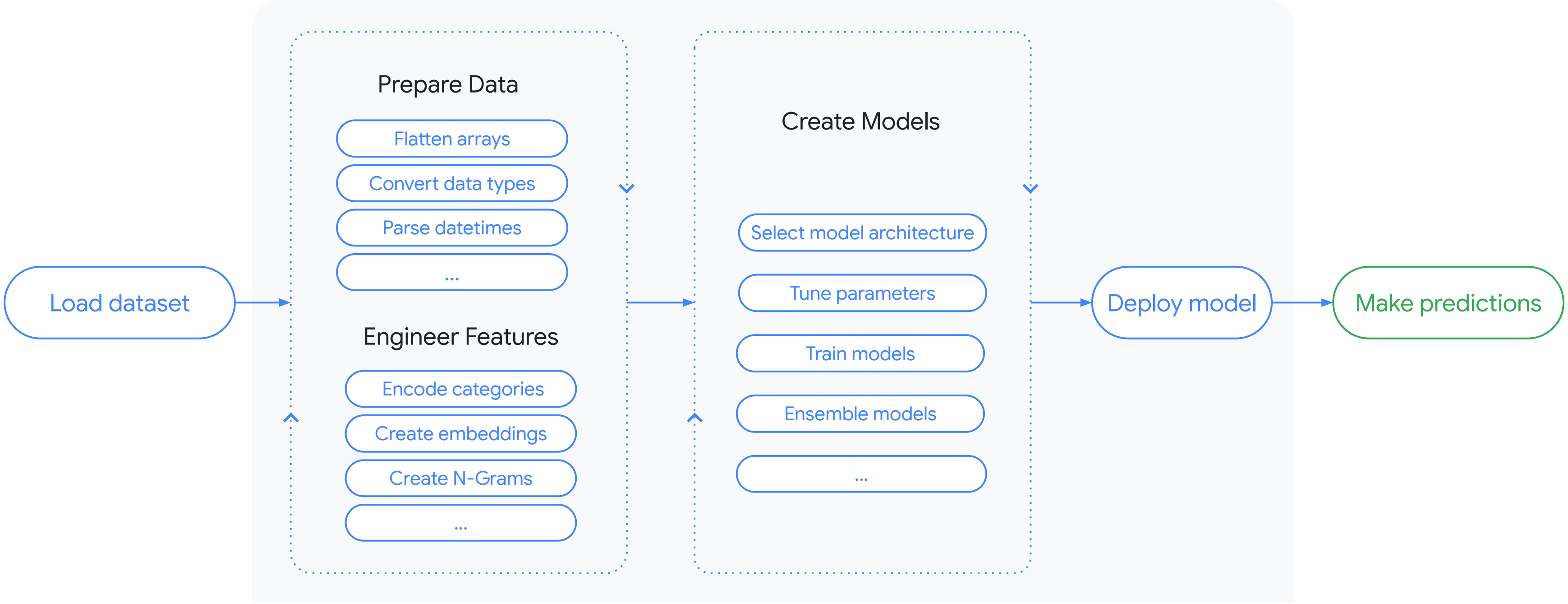
Responsible AI, Security & Sustainability

Leverage Google Cloud's secure and sustainable infrastructure

Google's AI Principles review process empowers users to trust that the tools are built with ethical governance

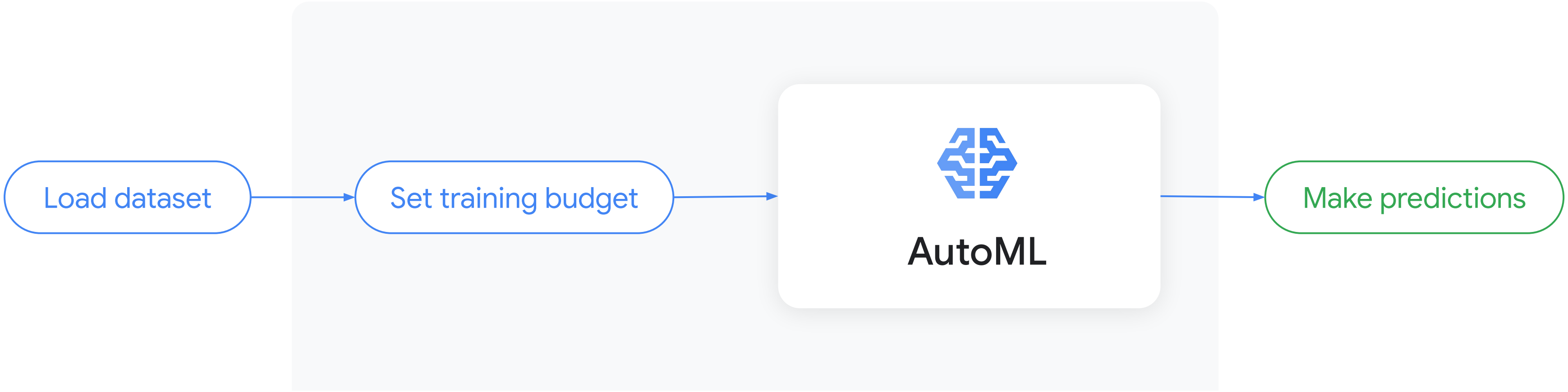
AutoML - Fastest path from data to value

Traditional Machine Learning Workflow



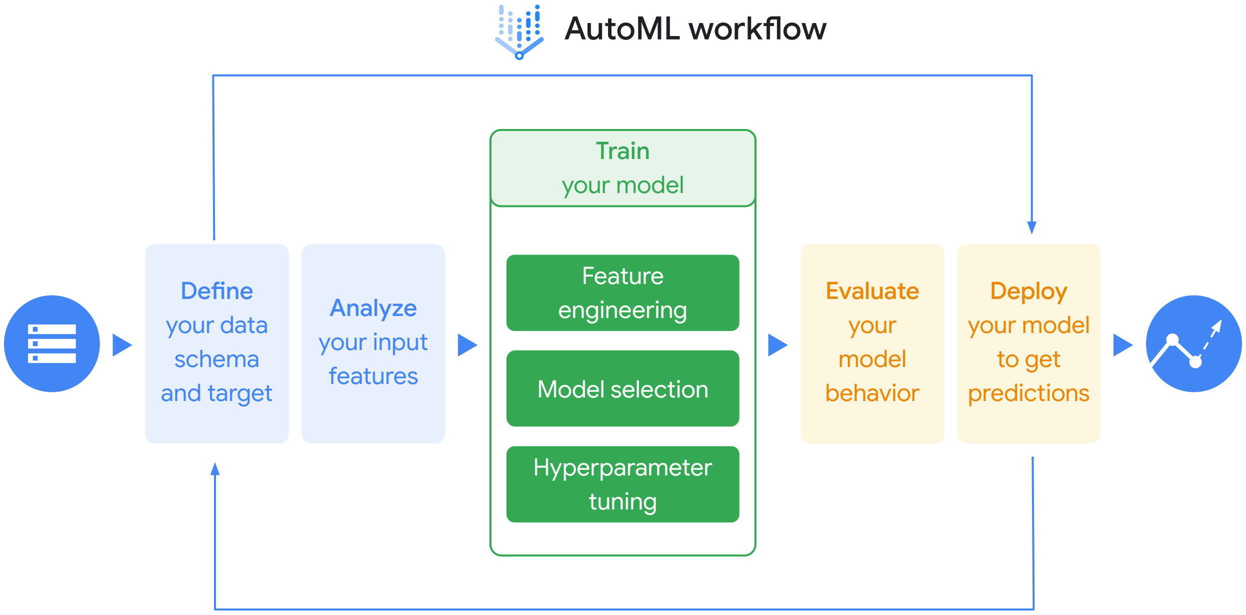
AutoML - Fastest path from data to value

AutoML Workflow



Low/No code

Point and click to build custom, high-quality models using the AutoML workflow in Vertex AI



Automatically search through Google's whole model zoo...

Linear, logistic

Feedforward DNN

Wide and Deep NN

Gradient Boosted Decision Tree (GBDT)

DNN + GBDT Hybrid

Adanet ensemble

Neural + Tree Architecture Search

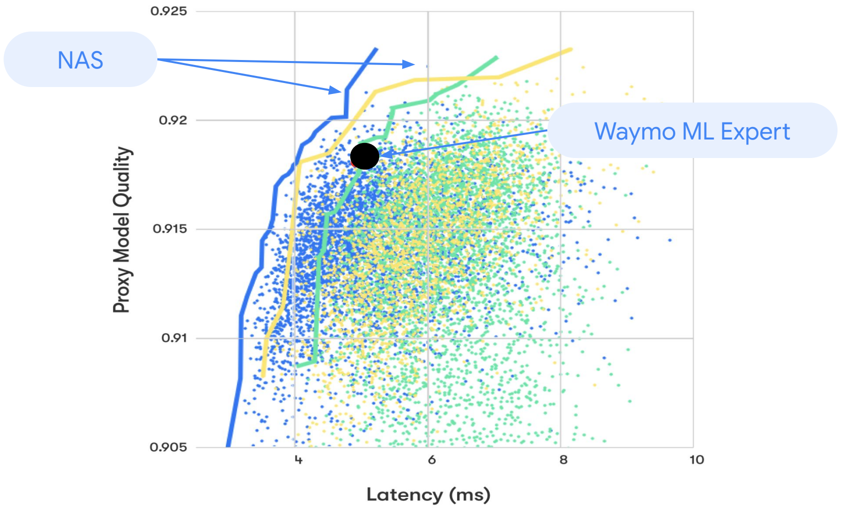
...and more!

Access to Google's best-in-class algorithms like NAS

Use of Neural Architecture Search (NAS) at Waymo

"Going from months of engineering time to generate and fine tune a architecture manually to "automatically generating" neural nets with NAS"

- 20–30% lower latency/same quality
- 8–10% lower error rate/same latency
- NAS model in 2 weeks vs months (1 year of GPU time) searching over 10k architectures



Waymo blog

MLOps

A set of **standardized** processes and technology capabilities for building, deploying, and operationalizing ML systems **rapidly** and **reliably**

Vertex AI

Applications

- Vision and Video
- Conversation
- Language
- Structured Data

Custom machine learning

Workbench	AutoML	NAS	Prediction	ML Metadata
Data Labeling	Training	Explainable AI	Feature Store	Model Monitoring
Experiments	Vizier (Optimization)	Pipelines	Matching Engine	AI Accelerators

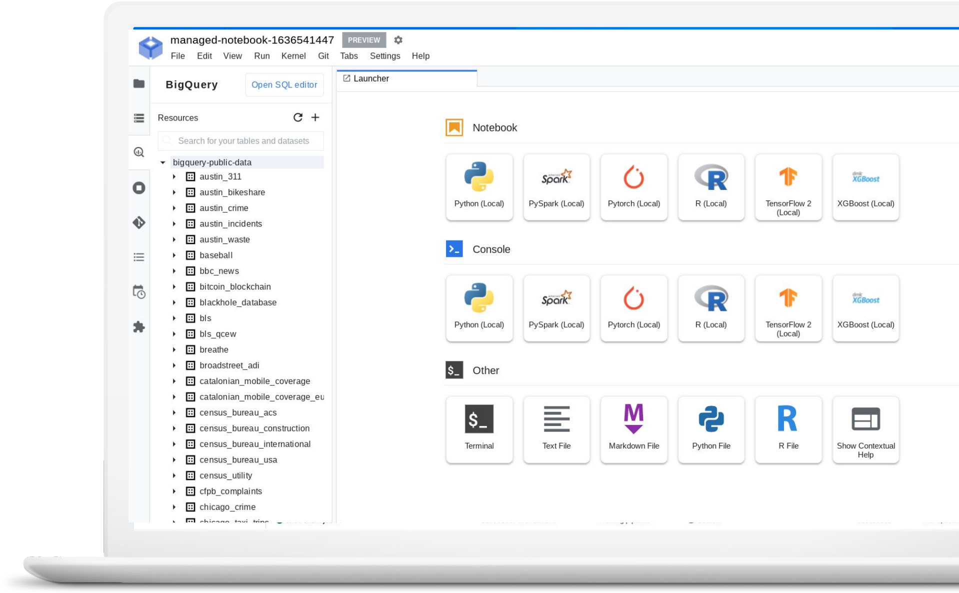
Vertex AI Workbench with Managed Notebooks

A one-stop interface for Data Science

Vertex AI Workbench

- Notebooks
- Multi-Kernel in UI
- Managed Pool
  - Instant Provisioning and Switching
  - Idle Shutdown
  - Direct access to data in BQ & Cloud Storage

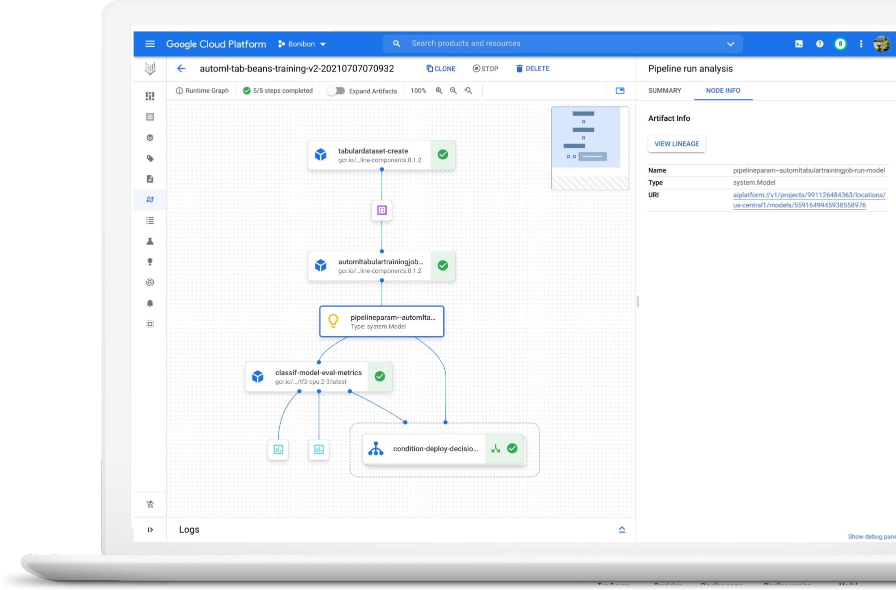
Analyse	Train	Deploy
BigQuery	Vertex Training (Custom and AutoML)	Vertex Pipelines
Spark		Vertex Prediction



Vertex Pipelines

Automate, monitor, and govern your ML systems by orchestrating your ML workflow in a serverless manner, and storing your workflow's artifacts using Vertex ML Metadata

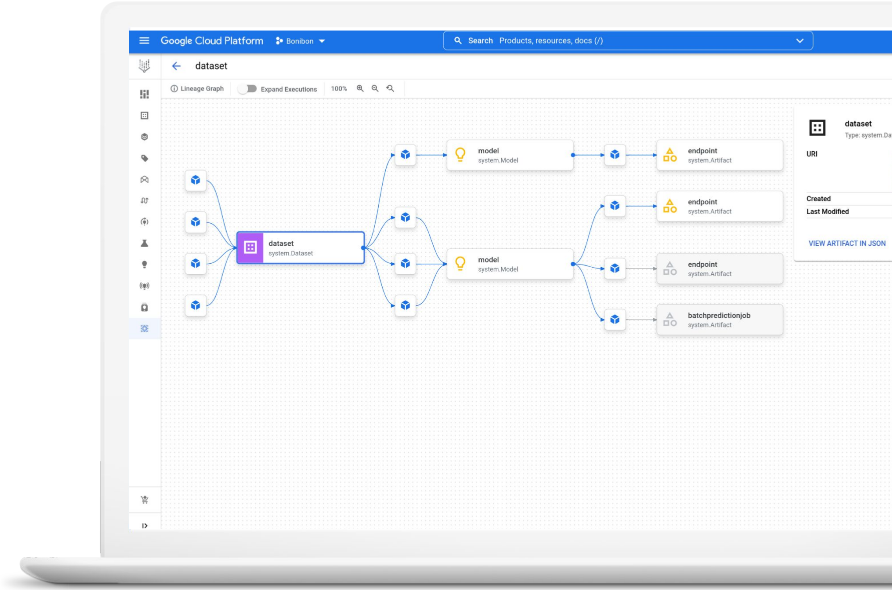
- **Easy to use Python SDKs:** Build your Pipelines using the battle-tested and easy-to-use KFP SDK and TFX SDK
- **Scalable:** Run as many pipelines on as much data as you want without having to worry about compute resources
- **Cost-effective:** Pay for the pipelines you run and the resources they use.
- **Secure:** Integrated with GCP security features like IAM, VPC-SC, and CMEK.
- **Metadata Tracking and Lineage:** Automatically store metadata about every artifact produced by the Pipelines.



Metadata and Lineage on Vertex AI

Artifact, lineage, and execution tracking for your ML workflow

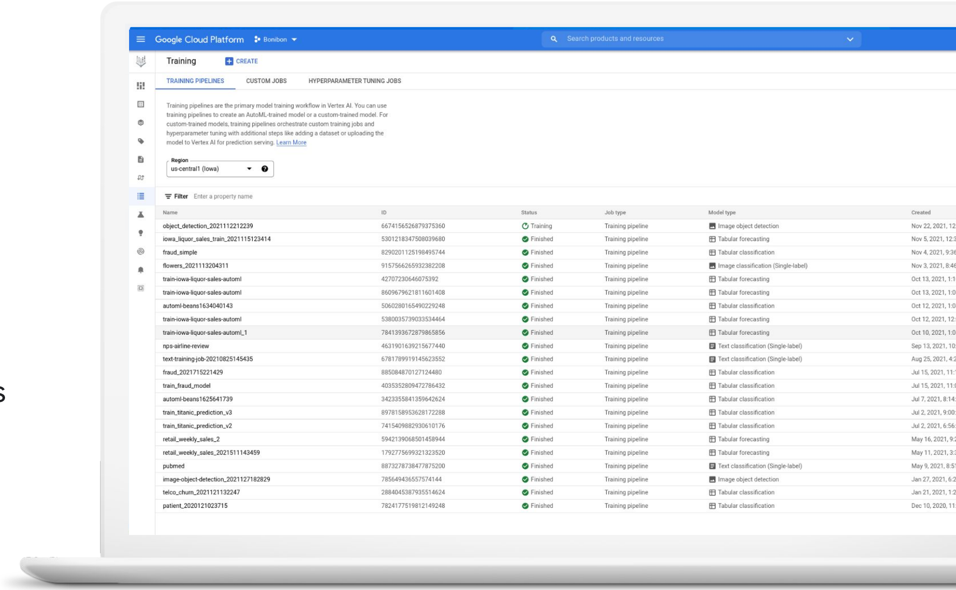
- Automatically track inputs and outputs to all components in an ML pipeline, and their lineage.
- Visualize the workflow for faster debugging with a DAG of all related executions.
- Manage artifacts by projects, group by experiments, and track the usage of datasets and models in your organization



Vertex Training

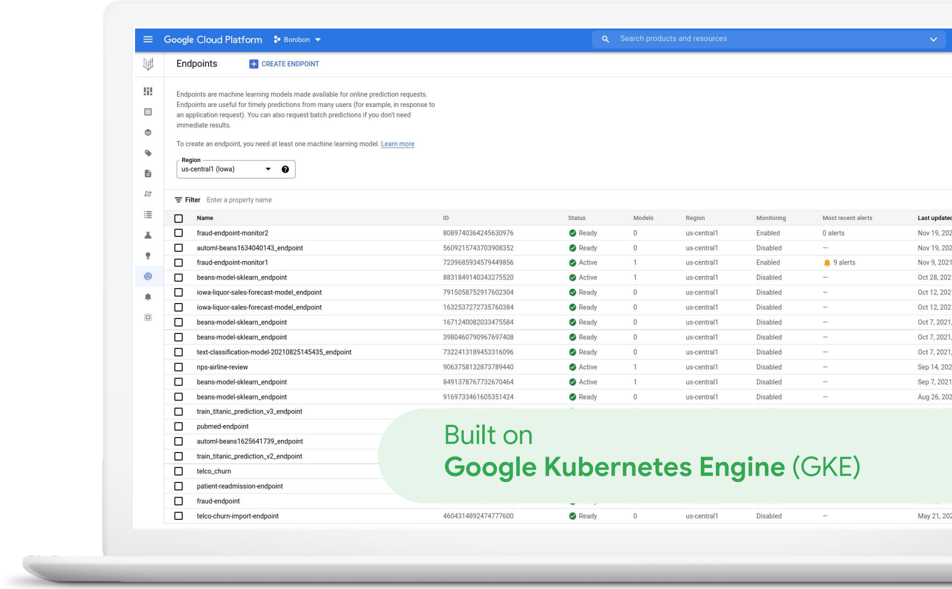
Artifact, lineage, and execution tracking for ML workflows, with an easy-to-use Python SDK.

- **Fully Managed:** Train without provisioning or managing servers. Pay only for the compute you consume. Zero administration.
- **High Performance:** Optimized for Machine Learning. Scalable distributed orchestration with the most advanced cloud Accelerators (GPUs and TPUs)
- **HyperParameter Optimization:** Automatically tune models with Google's Vizier optimizer
- **Customizable:** Supports predefined (Tensorflow, Sklearn, XGB, Keras) and custom containers with flexible machines
- **Built-in logging and monitoring:** review your execution jobs and monitor resources utilization for your jobs



Vertex Prediction

- Serve **online** endpoints for low-latency predictions, or predictions on massive **batches** of data.
- **Built-in security and compliance:** VPC peering and security perimeter. Custom managed encryption keys. Fine-tuned access control.
- **Low TCO:** Scale **automatically** based on your traffic, and alleviate operational overhead.
- **Intelligent and assistive:** Built-in Model Explainability and proactive model monitoring.
- Log prediction requests and responses to **BigQuery** for monitoring and debugging
- **Fast inference on GPUs:** Support for a broad range of machine types specialized for ML, such as GPUs.



Built on
Google Kubernetes Engine (GKE)

Feature Store on Vertex AI

A rich feature repository to serve, share and re-use ML features.

Share and reuse ML features across use cases

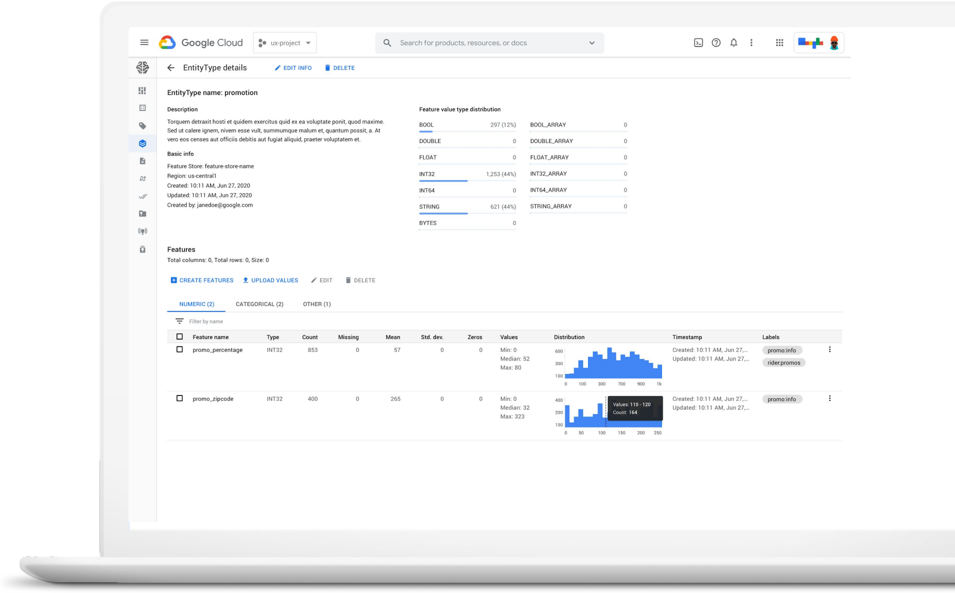
Centralized feature repository with easy APIs to search & discover features, fetch them for training/serving and manage permissions.

Serve ML Features at scale with low latency

Offload the operational overhead of handling infrastructure for low latency scalable feature serving.

Alleviate training serving skew

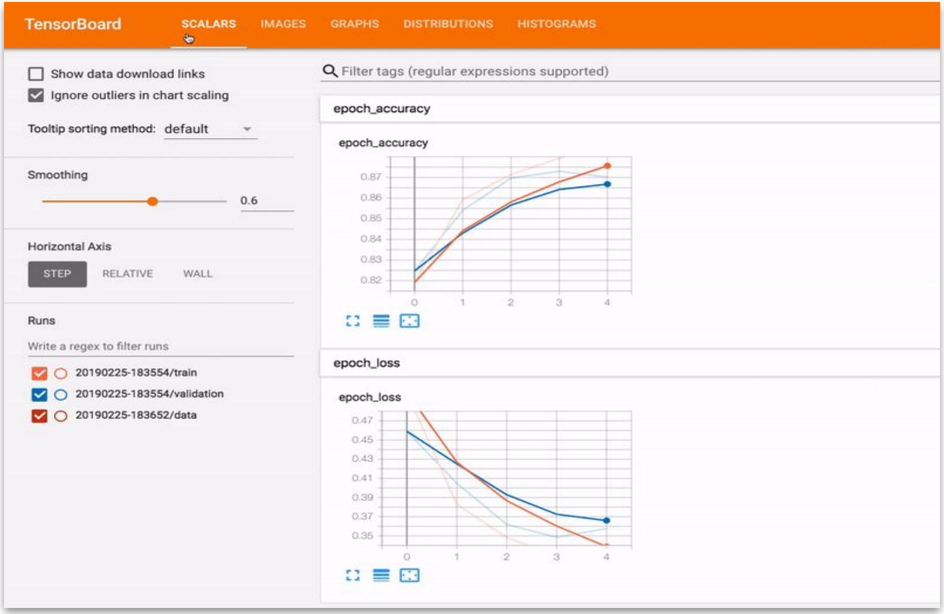
- Compute feature values once, re-use for training and serving
- Track & monitor for drift and other quality issues



TensorBoard: ML visualization toolkit

TensorBoard provides the visualization and tooling needed for ML experimentation

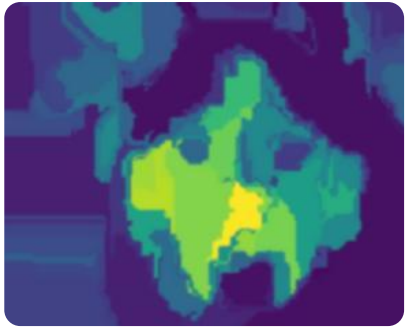
- Tracking and visualizing metrics such as loss and accuracy
- Visualizing the model graph (ops and layers)
- Viewing histograms of weights, biases, or other tensors as they change over time
- Projecting embeddings to a lower dimensional space
- Displaying images, text, and audio data
- Profiling TensorFlow programs



[Get started with TensorBoard Docs]

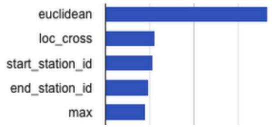
Explainable AI tells you how important each input feature is

Images



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
prcp	0.03	-0.00134132



Text

The cake tastes delicious!

Sentiment score: 0.9

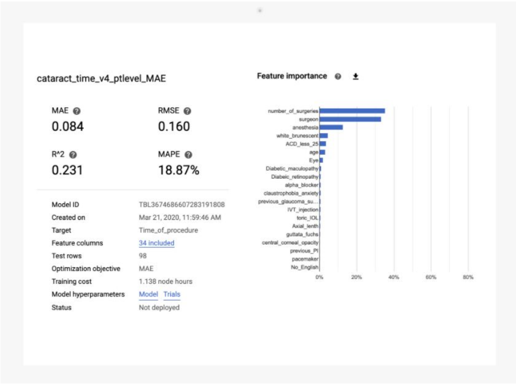
Explanations tell you:

What **image pixels or regions** most contributed to the model's classification?

How much did each **feature column** contribute to a single prediction or the model overall?

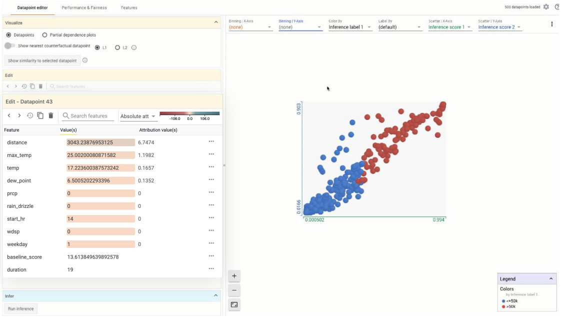
How much did each **word** or **token** contribute to the text classification?

Explainable AI in AutoML, BQML and Vertex Prediction



Integrated in AutoML, BQML, Vertex Prediction and Custom Containers, enabling explanations and “feature attributions” for any prediction, both local and global.

Pre-installed in Vertex Workbench TensorFlow instances for rapid, local prototyping and analysis



Use What-if Tool in Vertex AI Notebooks to visualize explanations on tabular data and LIT (Language Interpretability Tool) on text and image data.

Enables interactive model evaluation, counterfactual explainability, and fairness analysis

Monitoring of model performance

For models deployed in the Vertex Prediction service

Monitor and alert

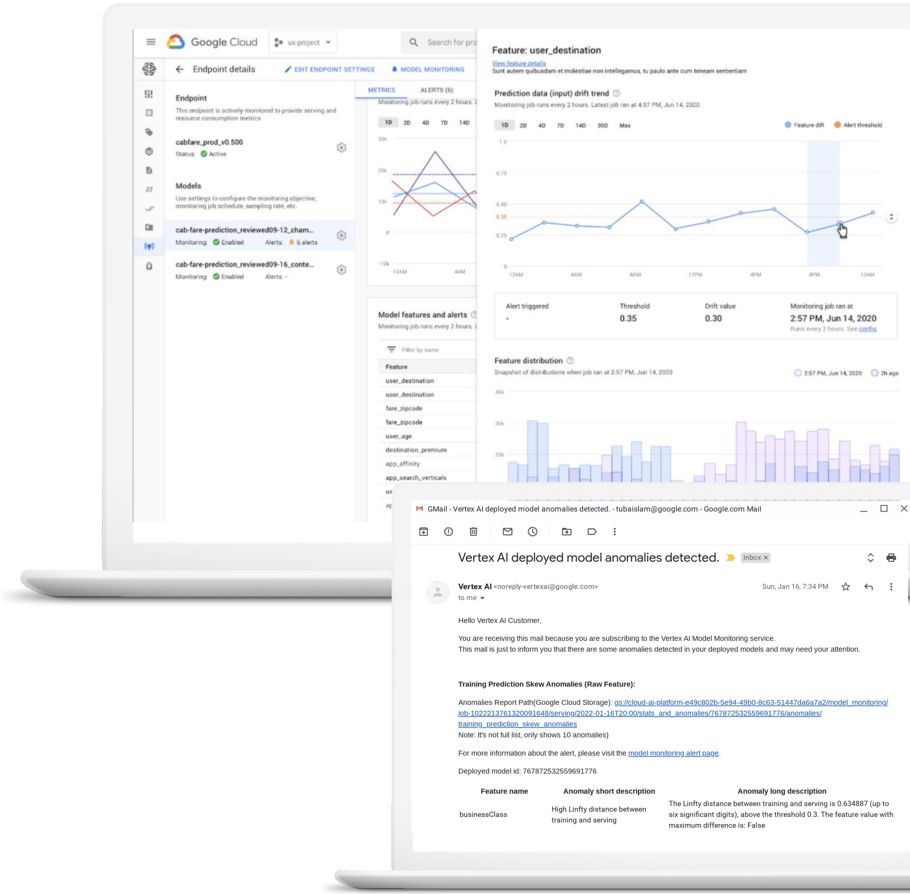
Monitor signals for model's predictive performance, and alert when those signals deviate.

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.



New ML Tools on Vertex AI: Matching Engine

30-50% cheaper than alternatives, while delivering higher scale and lower latencies.

Faster i.e. low latency

Find nearest neighbors in a few millisecond

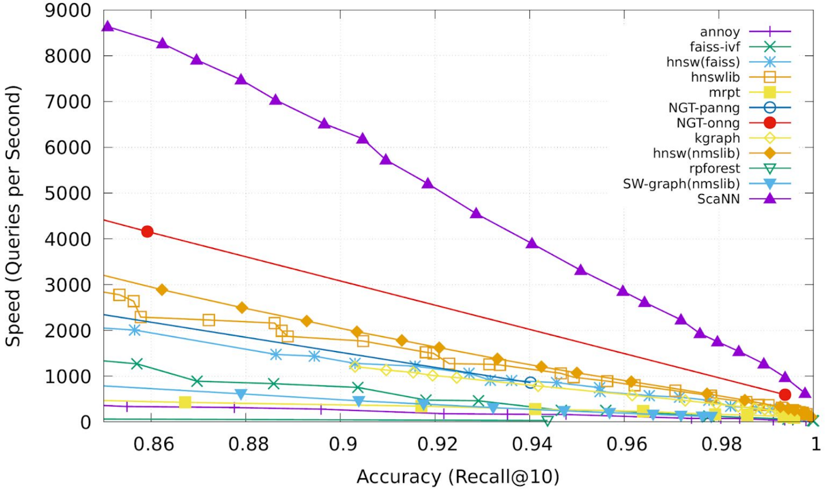
The most scalable

Scales to billions of vectors

Cheaper

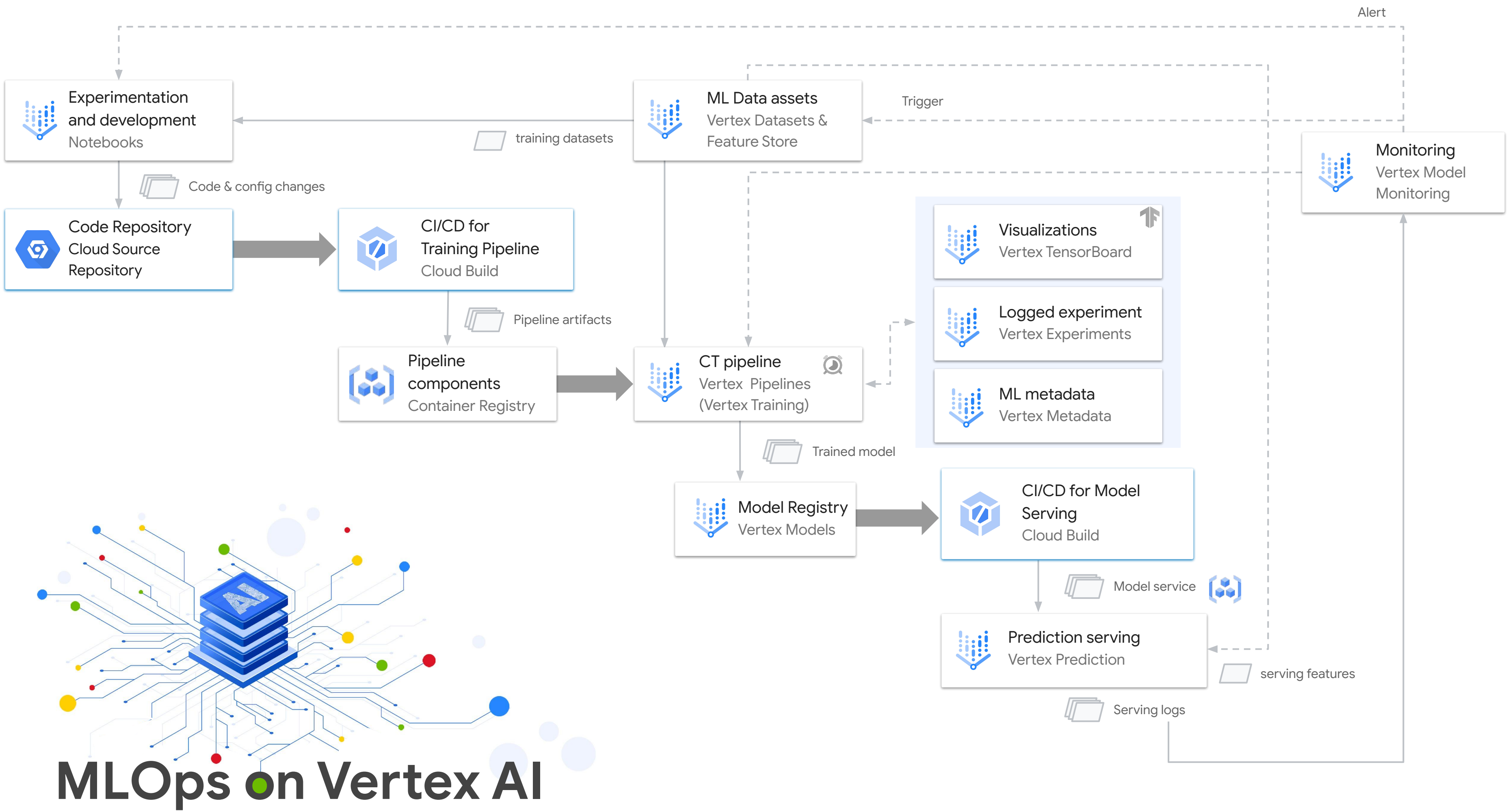
Requires fewer VMs to serve the same workload

- 1/4th the CPU consumption of **faiss**
- 1/3rd the memory consumption of **nmslib**



Google's technology (labelled **ScaNN**) compared with popular ANN services

[Demo link]



Alert
Experimentation and development
Notebooks
ML Data assets
Vertex Datasets & Feature Store
Trigger
training datasets
Monitoring
Vertex Model Monitoring
Code & config changes
Code Repository
Cloud Source Repository
CI/CD for Training Pipeline
Cloud Build
Visualizations
Vertex TensorBoard
Logged experiment
Vertex Experiments
Pipeline artifacts
Pipeline components
Container Registry
CT pipeline
Vertex Pipelines
(Vertex Training)
ML metadata
Vertex Metadata
Trained model
Model Registry
Vertex Models
CI/CD for Model Serving
Cloud Build
Model service
Prediction serving
Vertex Prediction
serving features
Serving logs
MLOps on Vertex AI

Demo's Marijn

Data & AI

Advanced analytics

Anomaly detection

Forecasting

Document processing

Image recognition

Conversational AI