

Cloud Infrastructure

ULiège | INFO9023 - Machine Learning Systems Design

March 18th 2024

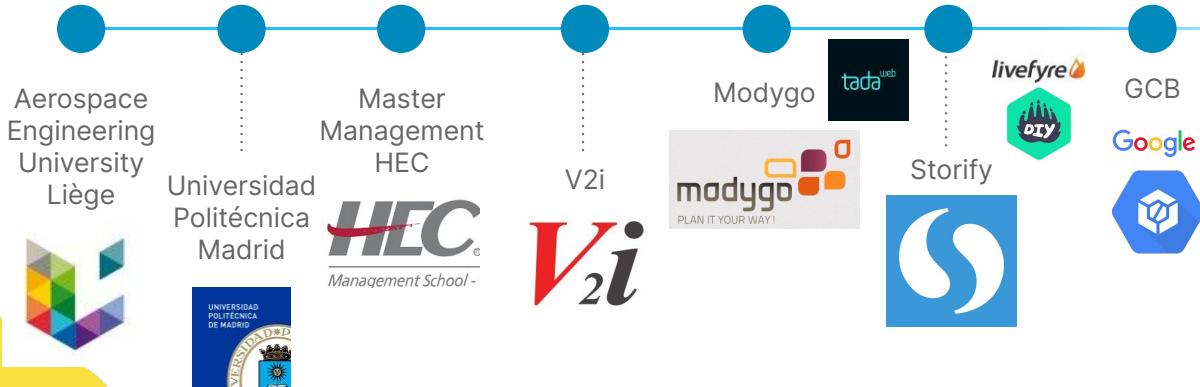
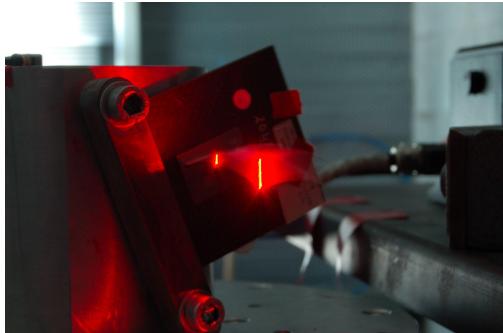
philmod@google.com | philippe.modard@gmail.com

TABLE OF CONTENTS

- **About me**
- **Kaggle Platform**
- **Cloud Computing**
 - Introduction
 - GCP Services
 - Vertex AI



About me – Philippe Modard



Software Engineer
Tech Lead

kaggle
Google

Kaggle platform



Competitions

Each year, we host thousands of ML competitions that attract over 300K participants who make nearly 5 million submissions.

The image shows a screenshot of the Kaggle competition dashboard. On the left, there is a vertical sidebar with icons for creating a new competition, editing, deleting, and more. At the top, there are tabs for 'All competitions', 'Featured', 'Getting Started', 'Research', 'Community', 'Playground', and 'Simulations'. Below these tabs, there are several competition cards displayed in a grid. Each card includes a thumbnail image, the competition name, a brief description, the prize amount, and a time remaining.

Competition	Description	Prize	Time Remaining
NFL Big Data Bowl 2023	Help evaluate linemen on pass plays Analytics	\$100,000	3mo to go
Feedback Prize - English Language Learning	Evaluating language knowledge of ELL st... Featured Code Competition - 1801 Teams	\$55,000	1mo to go
Big Data Derby 2023	Analyze horse racing data Analytics	\$50,000	
2022 Kaggle Machine Learning & Data Science Competition	The most comprehensive dataset available... Analytics	\$30,000	1mo to go
Open Problems - Multimodal Single-Cell Integration	Predict how DNA, RNA & protein measure... Featured 959 Teams	\$25,000	21d to go
Novozymes Enzyme Prediction	Help identify the thermophilic enzymes Featured 910 Teams	\$25,000	



Learn

Every month, over 50K people learn from our 17 hands-on, no cost ML courses.



Courses

We pare down complex topics to their key practical components, so you gain usable skills in a few hours provided at no cost to you, and you can now earn certificates. [Learn more.](#)



Intermediate Machine Learning

Handle missing values, non-numeric values, data leakage, and more.



Feature Engineering

Better features make better models. Discover how to get the most out of your data.



Advanced SQL

Take your SQL skills to the next level.



Computer Vision

Build convolutional neural networks with TensorFlow and Keras.



Time Series

Apply machine learning to real-world forecasting tasks.



Data Cleaning

Master efficient workflows for cleaning real-world, messy data.

Guides

Explore these curated collections of high-quality learning resources authored by the Kaggle community.

Code

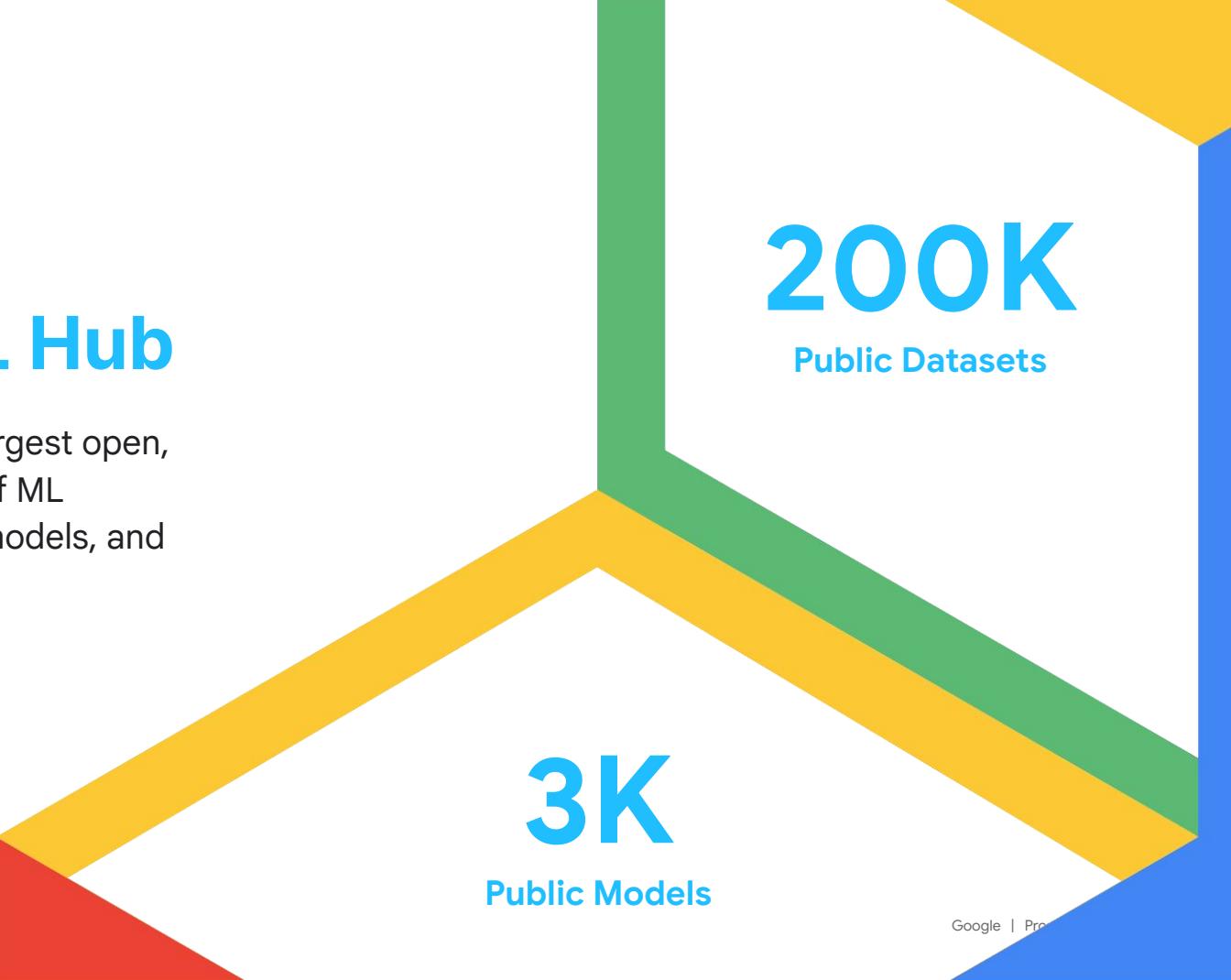
One-click to start coding on a fully managed Python/R environment

The screenshot shows a user interface for a Data Science AI Assistant. The top bar includes the title "Data Science AI Assistant with Gemma 2b...", a "Draft saved" indicator, and a "File" menu. Below the title, there's a toolbar with icons for "New", "Delete", "Copy", "Run", "Add-ons", and "Help". A "Markdown" dropdown is also present. On the far right, a status bar shows "Draft Session off (run)".
The main area contains a text editor with the following Python code:

```
[ ]:  
def get_embedding(text, embedding_model):  
    """Get embeddings for a given text using the provided embedding model"""  
  
    # Encode the text to obtain embeddings using the provided embedding model  
    embedding = embedding_model.encode(text, show_progress_bar=False)  
  
    # Convert the embeddings to a list of floats and return  
    return embedding.tolist()  
  
def map2embeddings(data, embedding_model):  
    """Map a list of texts to their embeddings using the provided embedding model"""  
  
    # Initialize an empty list to store embeddings  
    embeddings = []  
  
    # Iterate over each text in the input data list  
    for i in tqdm(range(len(data))):  
        # Get embeddings for the current text using the provided embedding model  
        embeddings.append(get_embedding(data[i], embedding_model))  
  
    # Return the list of embeddings  
    return embeddings
```

Largest ML Hub

But, Kaggle is also the largest open, community-driven hub of ML resources: public data, models, and code.



200K
Public Datasets

3K
Public Models



A community makes datasets into living, vibrant resources

ML COMMONS AND 2 COLLABORATORS · UPDATED 7 MONTHS AGO

14 New Notebook Download (101 GB) ...

The Dollar Street Dataset
A dataset containing 38,479 images of household items from around the world

Data Card Code (0) Discussion (2) Settings

MC7968 · POSTED 4 MONTHS AGO

Incomplete/truncated images?

Hi, it looks like
Their IDs are: 5

Cody Coleman Posted 4 months ago
@mc7968 that shouldn't be the case, but we will look into it. Thanks for pointing this out!

Reply

mc7968 Posted 4 months ago TOPIC AUTHOR
Thank you!

Reply

Provide feedback on this dataset

What do you use this dataset for?

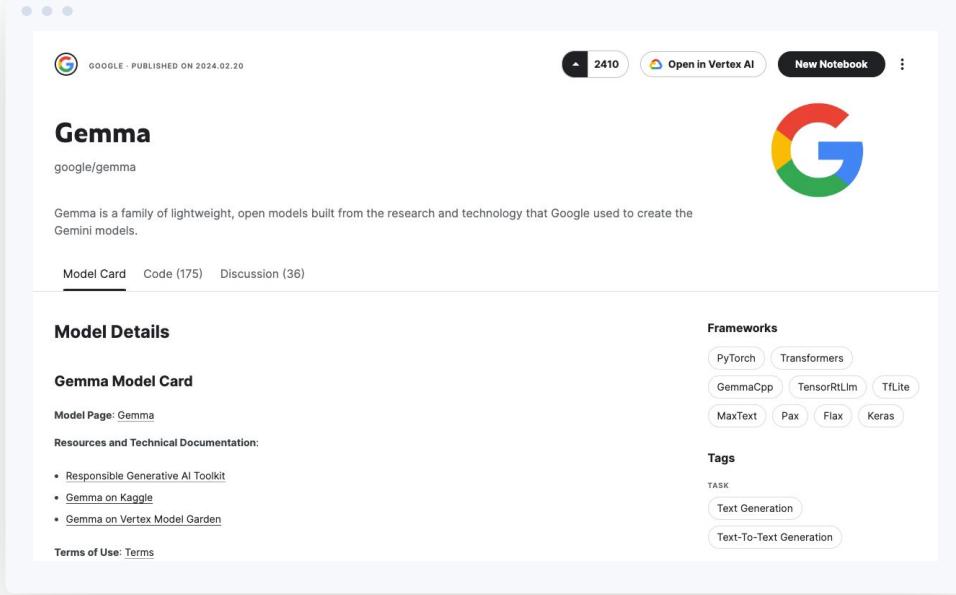
Learning 13 Research 2 Application 0

How would you describe this dataset?

Well-documented 1 Well-maintained 1 Clean data 1 Original 2 High-quality notebooks 1

0

Gemma



The screenshot shows the Gemma model card page on Kaggle. At the top, there's a navigation bar with three dots, a Google logo, and the text "GOOGLE - PUBLISHED ON 2024.02.20". To the right are buttons for "2410" (upvotes), "Open in Vertex AI", "New Notebook", and a more options menu. Below the navigation is a large Google "G" logo.

Gemma
google/gemma

Gemma is a family of lightweight, open models built from the research and technology that Google used to create the Gemini models.

[Model Card](#) [Code \(175\)](#) [Discussion \(36\)](#)

Model Details

Gemma Model Card

[Model Page: Gemma](#)

Resources and Technical Documentation:

- [Responsible Generative AI Toolkit](#)
- [Gemma on Kaggle](#)
- [Gemma on Vertex Model Garden](#)

[Terms of Use: Terms](#)

Frameworks

PyTorch, Transformers, GemmaCpp, TensorRTLM, TfLite, MaxText, Pax, Flax, Keras

Tags

TASK
Text Generation, Text-To-Text Generation

In partnership with Google Deep Mind, we released the latest open sourced LLM, Gemma, on Kaggle.





Competitions

Most Kaggle competitions have a simple setup

Training data is used to build algorithms & generate predictions

Test data is used to evaluate prediction accuracy

Training data			Test data			Claim Filed
Age	Income	Claim Filed	Age	Income		
58	\$95,824	True	73	\$53,445		
73	\$20,708	False	61	\$36,679		
59	\$82,152	False	47	\$90,422		
66	\$25,334	True	44	\$79,040		



Submissions are ranked on a live leaderboard

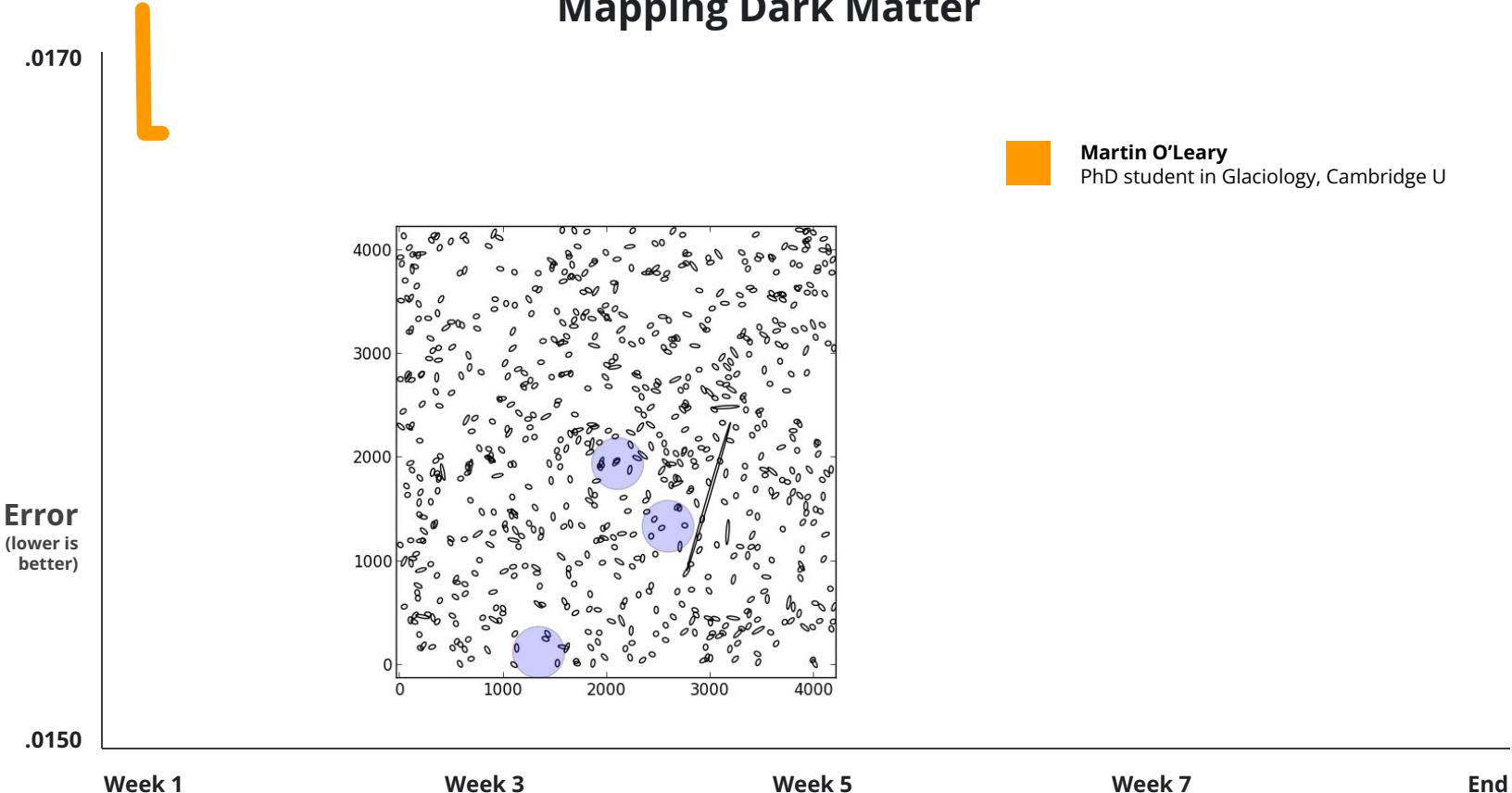
Public Private

This leaderboard is calculated with approximately 50% of the test data. The final results will be based on the other 50%, so the final standings may be different.

Prize Contenders

#	Team	Members	Score	Entries	Last	Join
1	🏆🏆	🟡	0.756	50	11h	
2	George Megre	🟡	0.736	7	4d	
3	Roman Yakunin	🟡	0.695	8	5d	
4	forest	🟡	0.682	21	2h	
5	The Tinnitus Man	🟡	0.678	43	2h	
6	Just A game on your lips	🟡	0.621	5	8d	
7	ohkawa3	🟡	0.619	15	19h	
8	Wondering Alice	🟡	0.553	43	4h	
9	shun39	🟡	0.536	10	2d	
10	Amirreza Ghasemi	🟡	0.520	17	2d	
11	bert vdb	🟡	0.483	11	3d	

Mapping Dark Matter



[Home](#) • [The Administration](#) • [Office of Science and Technology Policy](#)[Office of Science and Technology Policy](#)

Competition Shines Light on Dark Matter

Posted by Jason Rhodes on June 27, 2011 at 04:32 PM EDT

The world's brightest physicists have been working for decades on solving one of the great unifying problems of our universe. It is a problem that explores our place in the cosmos and, as was the case with Newton's law of gravitation and Einstein's theory of relativity, "If solved, it would change everything we know about the Universe if solved. Recently, top experts came together to break through from a dead end place,

"In less than a week, Martin O'Leary, a PhD student in glaciology, outperformed the state-of-the-art algorithms"

On May 23, a consortium of the very best from the National Science Foundation and the American Physical Society [posted the problem](#) on the data-mining website Kaggle and Challenge.gov for all the world to weigh in. In less than a week, Martin O'Leary, a PhD student at the University of Colorado Boulder, identified an algorithm that outperformed the state-of-the-art algorithms most commonly used in astronomy for mapping dark matter.

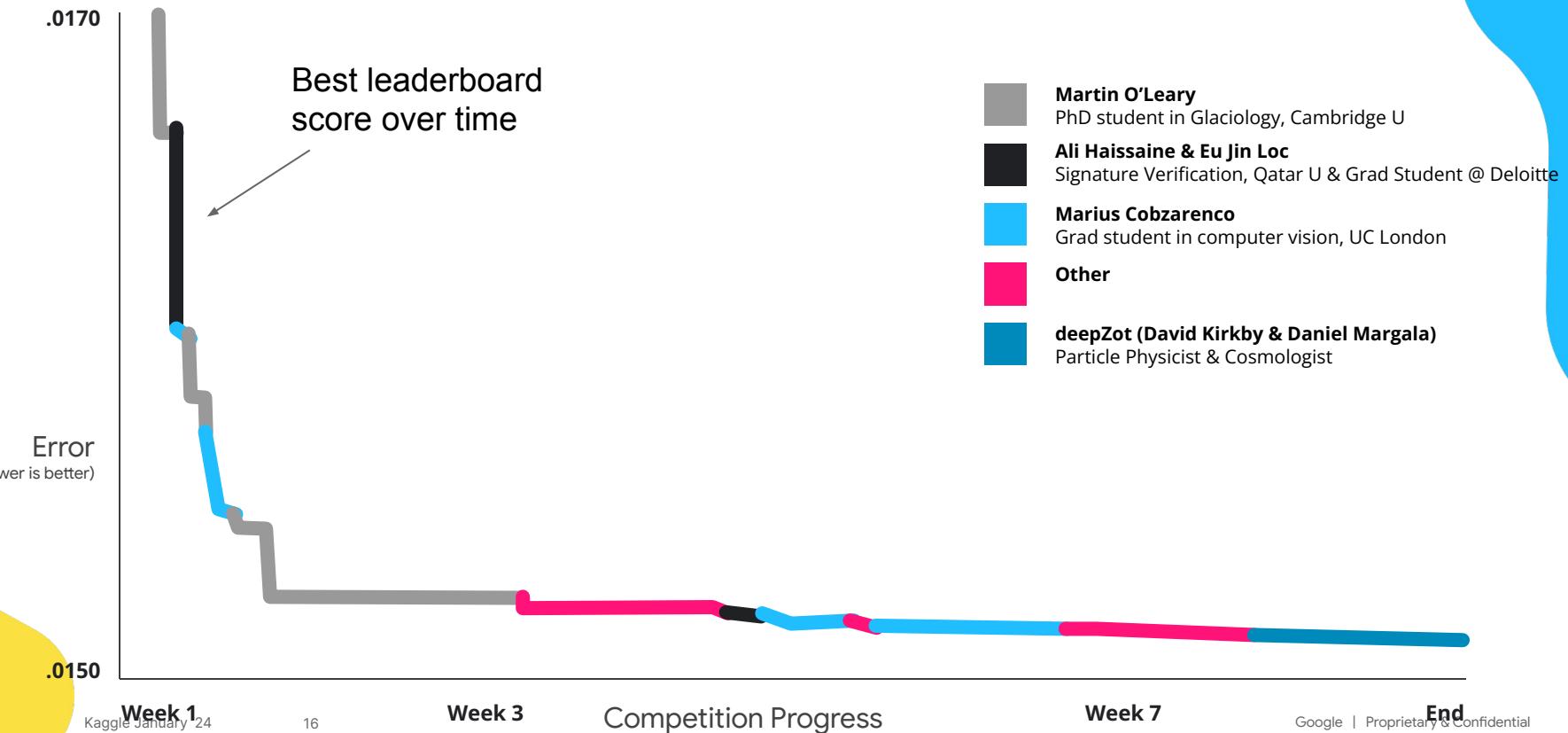
Chalk another one up for the power of crowdsourcing, and this Administration's commitment to using prizes and challenges to find solutions to some of our most pressing problems—here on Earth as well as in the furthest reaches of space!

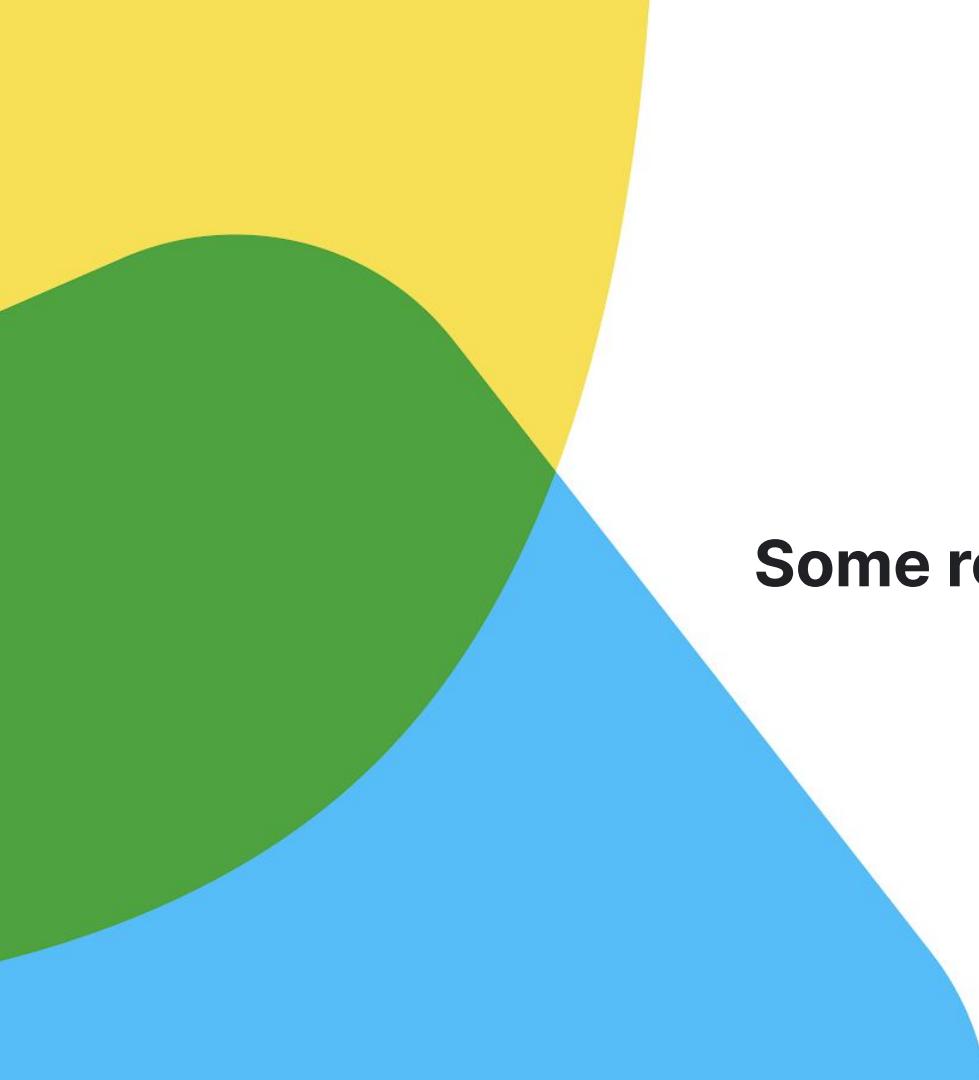
The posted problem had to do with how scientists can go about mapping "dark matter." Our Universe, it turns out,

The sidebar on the right contains a green button labeled "Launch the Receipt".

WHITE HOUSE BLOG

Competitions extract all the signal from a dataset





Some recent unique competitions

Vesuvius Ink Detection

An ambitious challenge to “resurrect” destroyed papyrus scrolls from the Vesuvius volcano which erupted nearly 2,000 years ago

\$1,000,000 in prizes

The screenshot shows the Kaggle competition page for the Vesuvius Challenge - Ink Detection. At the top, there's a banner with the title "Vesuvius Challenge - Ink Detection" and the subtitle "Resurrect an ancient library from the ashes of a volcano". Below the banner, there's a navigation bar with tabs: Overview (which is active), Data, Code, Discussion, Leaderboard, and Rules. The main content area has a sidebar on the left with links to Description, Evaluation, Timeline, Prizes, and Code Requirements. The main content area features a section titled "Goal of the Competition" with a detailed description of the challenge. It mentions the \$1,000,000+ prize pool and the specific task of detecting ink from 3D X-ray scans. The "Prizes" section lists several categories with their respective monetary values.

Featured Code Competition

Vesuvius Challenge - Ink Detection

Resurrect an ancient library from the ashes of a volcano

Vesuvius Challenge · 1,192 teams · 10 days to go (3 days to go until merger deadline)

[Overview](#) [Data](#) [Code](#) [Discussion](#) [Leaderboard](#) [Rules](#)

Overview

Description

Evaluation

Timeline

Prizes

Code Requirements

Goal of the Competition

Join the \$1,000,000+ [Vesuvius Challenge](#) to resurrect an ancient library from the ashes of a volcano. The competition involves detecting ink from 3D X-ray scans and reading the contents. Thousands of scrolls were found in Herculaneum, a town next to Pompeii. This villa was buried by the Vesuvius eruption in 79 AD. The scrolls were carbonized, and are now impossible to open without breaking them. These scrolls have been waiting to be read using modern techniques. There is a \$700,000 grand prize from a 3D X-ray scan. This Kaggle competition hosts the [Ink Detection Progress Prize](#).

Prizes breakdown:

- [Grand Prize](#) - \$700,000
- Ink Detection Progress Prize on Kaggle - \$100,000 in prizes
- [Segmentation Tooling Prize](#) - \$45,000 in prizes
- [First Letters Prize](#) - \$50,000 in prizes
- To be announced - \$200,000+



Simulations - Reinforcement Learning Competitions



[Intro to Game AI and Reinforcement Learning Home Page](#)

Introduction

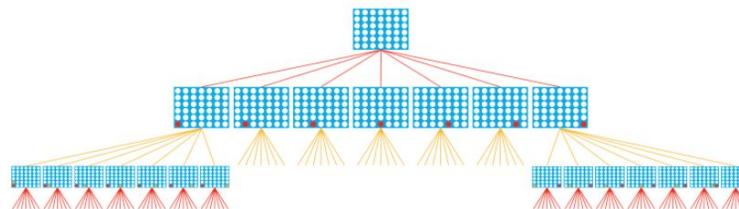
Even if you're new to Connect Four, you've likely developed several game-playing strategies. In this tutorial, you'll learn to use a **heuristic** to share your knowledge with the agent.

Game trees

As a human player, how do you think about how to play the game? How do you weigh alternative moves?

You likely do a bit of forecasting. For each potential move, you predict what your opponent is likely to do in response, along with how you'd then respond, and what the opponent is likely to do then, and so on. Then, you choose the move that you think is most likely to result in a win.

We can formalize this idea and represent all possible outcomes in a **(complete) game tree**.

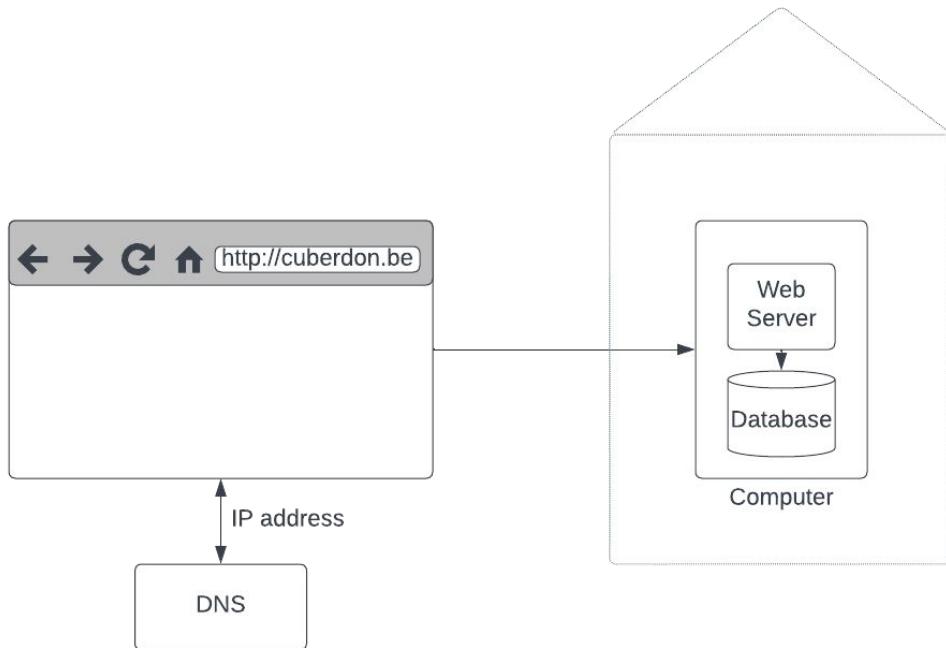


Cloud Computing



Introduction

Imagine you are building a new online store website to sell ... cuberdons.



Solutions

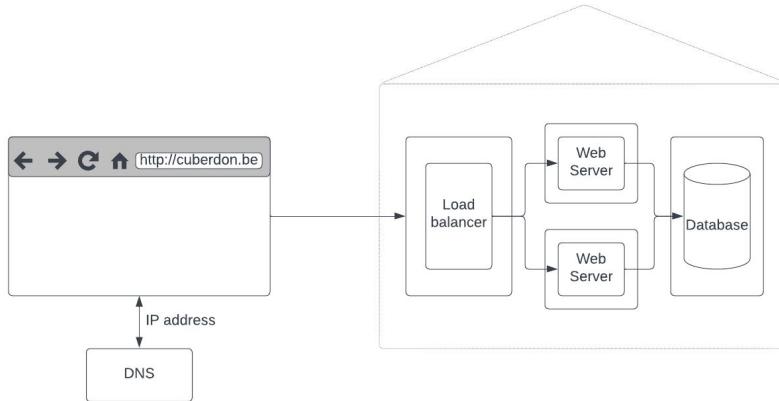
Vertical Scaling

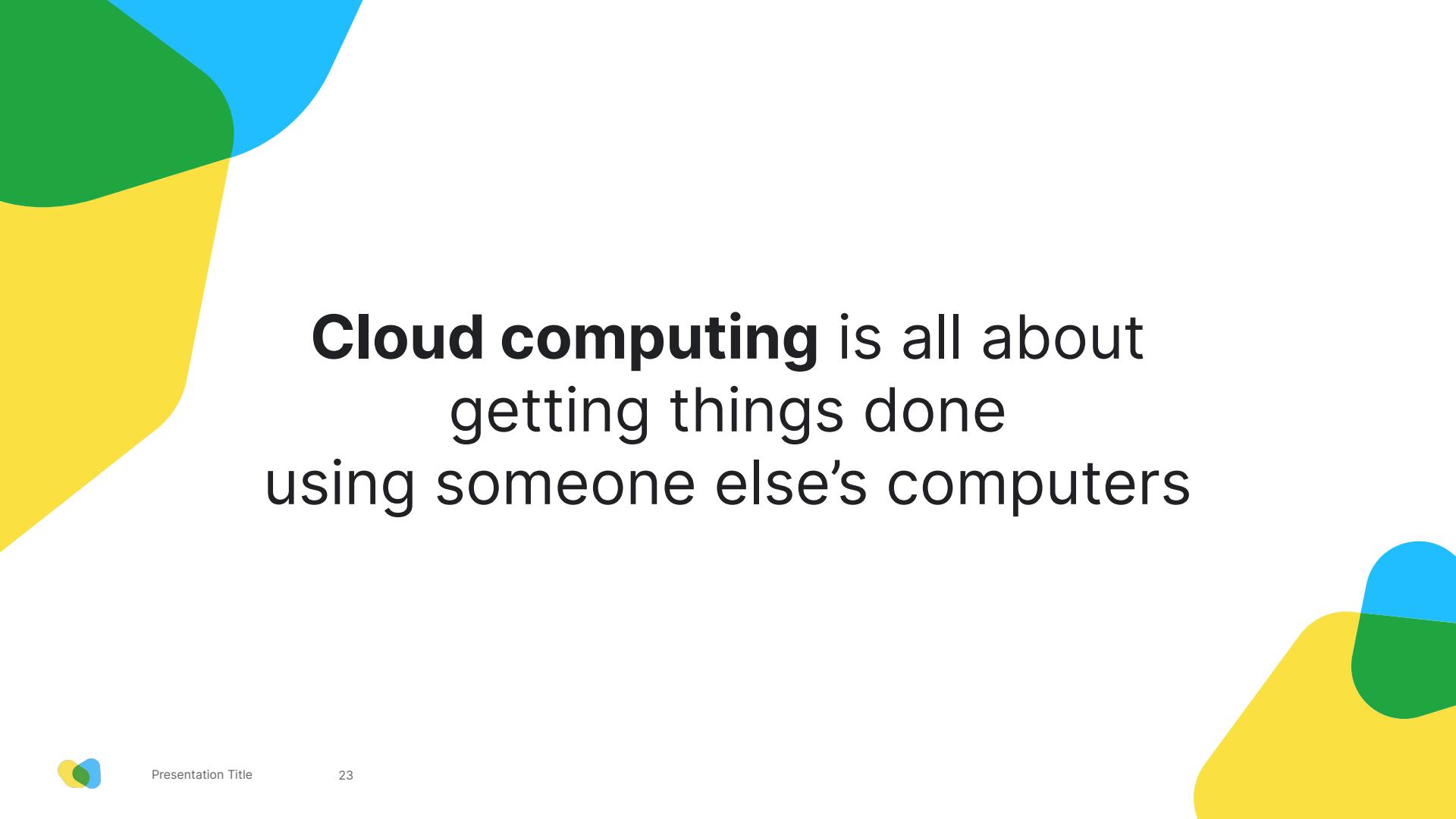
Pros

Easy

Cons

Limit
Only solving
scaling issue





Cloud computing is all about
getting things done
using someone else's computers

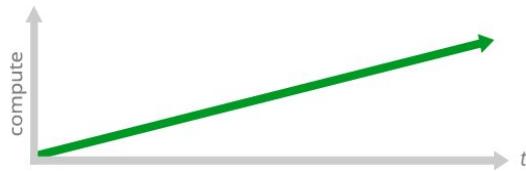


On and Off



- On & off workloads (e.g. batch job)
- Over provisioned capacity is wasted

Growth



- Successful services needs to scale
- Difficult to provision hardware

Bursting



- Unexpected/unplanned peak in demand
- Sudden spike impacts performance
- Can't over-provision for extreme cases





Private Cloud. Runs in local datacenter (either customers or a hoster like Rackspace).



Hybrid. Customers who have some elements from private and public clouds.



Public Cloud. Massive centralized compute, IaaS, PaaS and SaaS. In essence this is GCP, AWS and Azure.



	On-Premises	Cloud
Cost	Upfront, Calculated for peak usage	Ondemand, Ongoing
Performance		Best internet connection, servers closer to users
Operations	Responsible for all infrastructure setup and maintenance, needs knowledge	Only responsible for data management and service configuration; monitoring
Security	Fully control the data; responsible for all security setup and patching	Cloud provider is responsible for underlying security
Reliability, availability	Responsible for designing redundant system	Deploy service across multiple regions for reliability; easy backup
Scalability	Responsible for planning growth	Easily scale up and down, pay as you go



Traditional On-Premise (On-Prem)

Applications

Data

Runtime

Middleware

Operating System

Virtualization

Servers

Storage

Networking

Infrastructure as a Service (IaaS)

Applications

Data

Runtime

Middleware

Operating System

Virtualization

Servers

Storage

Networking

Platform as a Service (PaaS)

Applications

Data

Runtime

Middleware

Operating System

Virtualization

Servers

Storage

Networking

Software as a Service (SaaS)

Applications

Data

Runtime

Middleware

Operating System

Virtualization

Servers

Storage

Networking

Self Managed

Managed By Vendor



IaaS pros	IaaS cons
<ul style="list-style-type: none"> • Highest level of control over infrastructure • On-demand scalability • No single point of failure for higher reliability • Reduced upfront capital expenditures (for example, pay-as-you-go pricing) • Fewer provisioning delays and wasted resources • Accelerated development and time to market 	<ul style="list-style-type: none"> • Responsible for your own data security and recovery • Requires hands-on configuration and maintenance • Difficulties securing legacy applications on cloud-based infrastructure
CaaS pros	CaaS cons
<ul style="list-style-type: none"> • Ideal for running, managing, and scaling microservices • Streamlined development speeds up time to market • More control and configuration of networks and application components • Increases workload portability between environments, such as hybrid cloud and multicloud • Built-in performance monitoring and container orchestration 	<ul style="list-style-type: none"> • Some CaaS solutions have limited language support available depending on the cloud service provider • Container security risks may increase when using CaaS as they share the same kernel with the OS (although they are considered safer than VMs)
PaaS pros	PaaS cons
<ul style="list-style-type: none"> • Instant access to a complete, easy-to-use development platform • Cloud service provider is responsible for maintenance and securing infrastructure • Available over any internet connection on any device • On-demand scalability 	<ul style="list-style-type: none"> • Application stack can be limited to the most relevant components • Vendor lock-in may be an issue depending on the cloud service provider • Less control over operations and the overall infrastructure • More limited customizations
SaaS pros	SaaS cons
<ul style="list-style-type: none"> • Easy to set up and start using • The provider manages and maintains everything, from hardware to software • Software is accessible over any internet connection on any device 	<ul style="list-style-type: none"> • No control over any of the infrastructure or security controls • Integration issues with your existing tools and applications • Vendor lock-in may be an issue depending on the cloud service provider • Little to no customization



The background features a large, abstract graphic composed of three overlapping curved shapes in yellow, green, and blue, creating a dynamic, modern feel.

Google Cloud Platform

Powered by 20+ years of Google Innovation

Google Innovation Timeline

Innovation Timeline

2003



Birth of Borg
3-4 Google Engineers working to automate cluster management inside Google.

Scheduling ~ Several **BILLION** containers per week in 2020 across the entire Google environment.

2006



Process Containers
initiative to bring containers to the Linux kernel

2008



The term “**Cloud Computing**” enters the common vernacular.

April 2008

Google Cloud is launched

2013



LxC launched, complete Linux container manager merged into the Linux Kernel

2015



2013 Docker launched

Nov 2014

GKE Alpha



Work begins to open source Google’s Borg as Kubernetes

July 2015

Kubernetes 1.0



Aug 2015

GKE GA



Envoy 1.0

2018



Istio 1.0



Istio announced



Knative announced



GKE on Prem Announced

2019



Cloud Run

2021





Google Cloud Platform

Regions and PoPs

*Exception: region has 4 zones.

Current region
with 3 zones

Future region
with 3 zones

Edge point
of presence

Location Considerations

	Regional	Dual-Region	Multi-Region
Availability	<ul style="list-style-type: none"> • Data redundancy across availability zones (synchronous) • RTO=0: automated failover and failback on zonal failure (no need to change storage paths) 	<ul style="list-style-type: none"> • Higher availability than regional • Data redundancy across regions (asynchronous) • Turbo replication option for replication within 15 minutes • RTO=0: automated failover and failback on regional failure (no need to change storage paths) 	<ul style="list-style-type: none"> • Higher availability than regional • Data redundancy across regions (asynchronous) • RTO=0: automated failover and failback on regional failure (no need to change storage paths)
Performance	<ul style="list-style-type: none"> • 200 Gbps (per region, per project) • Scalable to many Tbps by requesting higher bandwidth quota 	<ul style="list-style-type: none"> • 200 Gbps (per region, per project) • Scalable to many Tbps by requesting higher bandwidth quota 	<ul style="list-style-type: none"> • 50 Gbps (per region, per project) • Limited performance scaling, variable performance for reads
Pricing	<ul style="list-style-type: none"> • Lowest storage price • No replication charges • No egress charges when reading data inside the same region 	<ul style="list-style-type: none"> • Highest storage price • Replication charges apply on write • No egress charges when reading data within either region 	<ul style="list-style-type: none"> • Higher storage price than regional, but lower than dual-region • Replication charges apply on write • Egress charges always apply when reading data

Network Concepts

Project

Network (VPC) - Global construct

Regional construct

Zone a

Zone b

Zone c

Subnet

192.168.0.0/16

Subnet

10.0.0.0/8



Regional construct

Zone a

Zone b

Subnet

172.16.0.0/12



Running Code



Compute Engine



App Engine



Cloud Run

Complexity
Customizability



Storing Data



Storage



SQL



Firebase



Big Data / ML



BigQuery



Dataflow



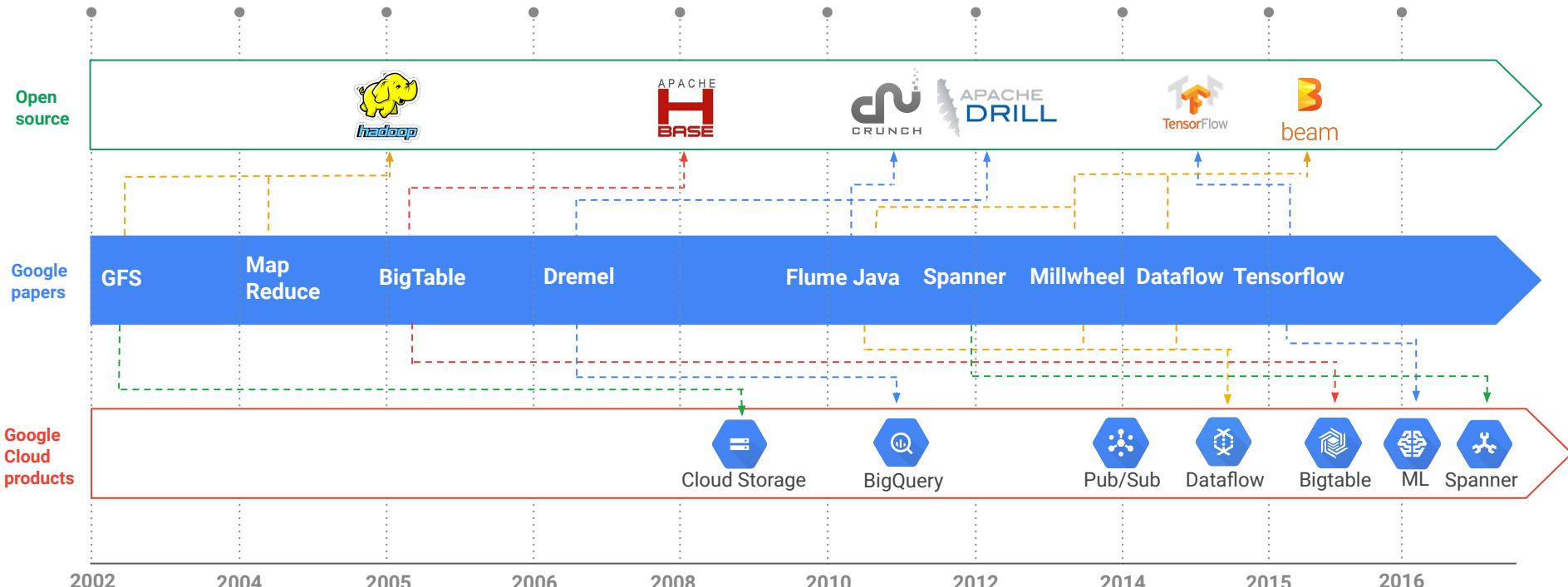
Vision API, ...



Vertex AI



Twenty years of tackling big data problems



All Google Cloud Services: <https://googlecloudcheatsheet.withgoogle.com>



The most comprehensive portfolio of purpose-built infra TPU & GPU optionality for all AI workloads

N1+T4
(T4)

A2
(A100)

G2
(L4)

A3
(H100)

A3+
(H100)

TPU
v3 & v4

TPU
v5e

TPU
v5p

Giant Model
Training

ML
Training

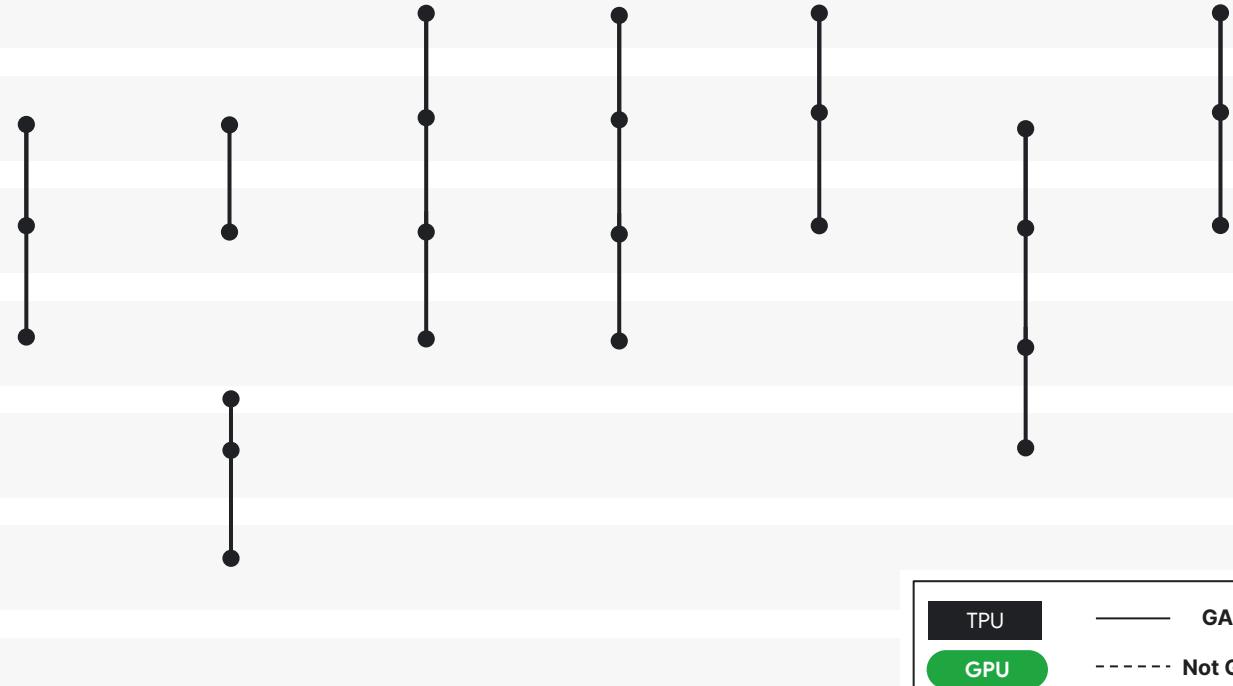
Model Fine
Tuning

Large Model
Inference

Medium Model
Inference

Small Model
Inference

Low-Cost
Presentation Title
Inference



DEMO





Google Vertex AI

A deep history of research and innovation at Google



2017
Transformer



2018
BERT



2019
T5



2020
LaMDA



2021
AlphaFold



2022
PaLM



2023
Gemini

Google invents
Transformer
kickstarting LLM
revolution

Google's
groundbreaking
large language
model, BERT

Text-to-Text
Transfer Transformer
LLM 10B P model open
sourced

Google LaMDA
model trained to
converse

AlphaFold predicts
structures of all
known proteins

Industry leading
large language
model

A conversational AI
Service powered by
LaMDA.

Responsible AI at the foundation



Machine learning APIs



AutoML



Vertex AI Custom training



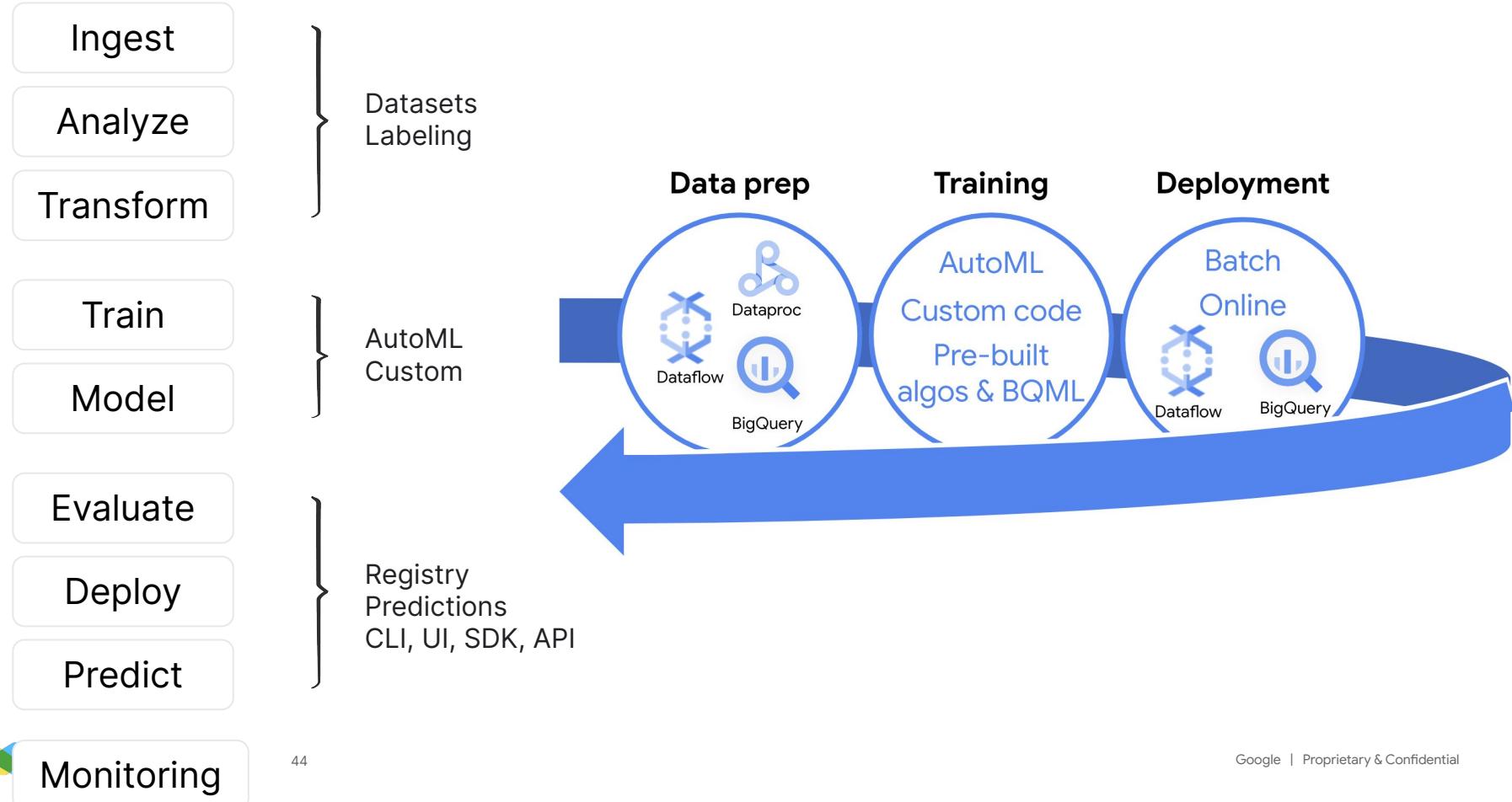
AI Infrastructure tools



Require zero machine learning knowledge

Require more machine learning knowledge

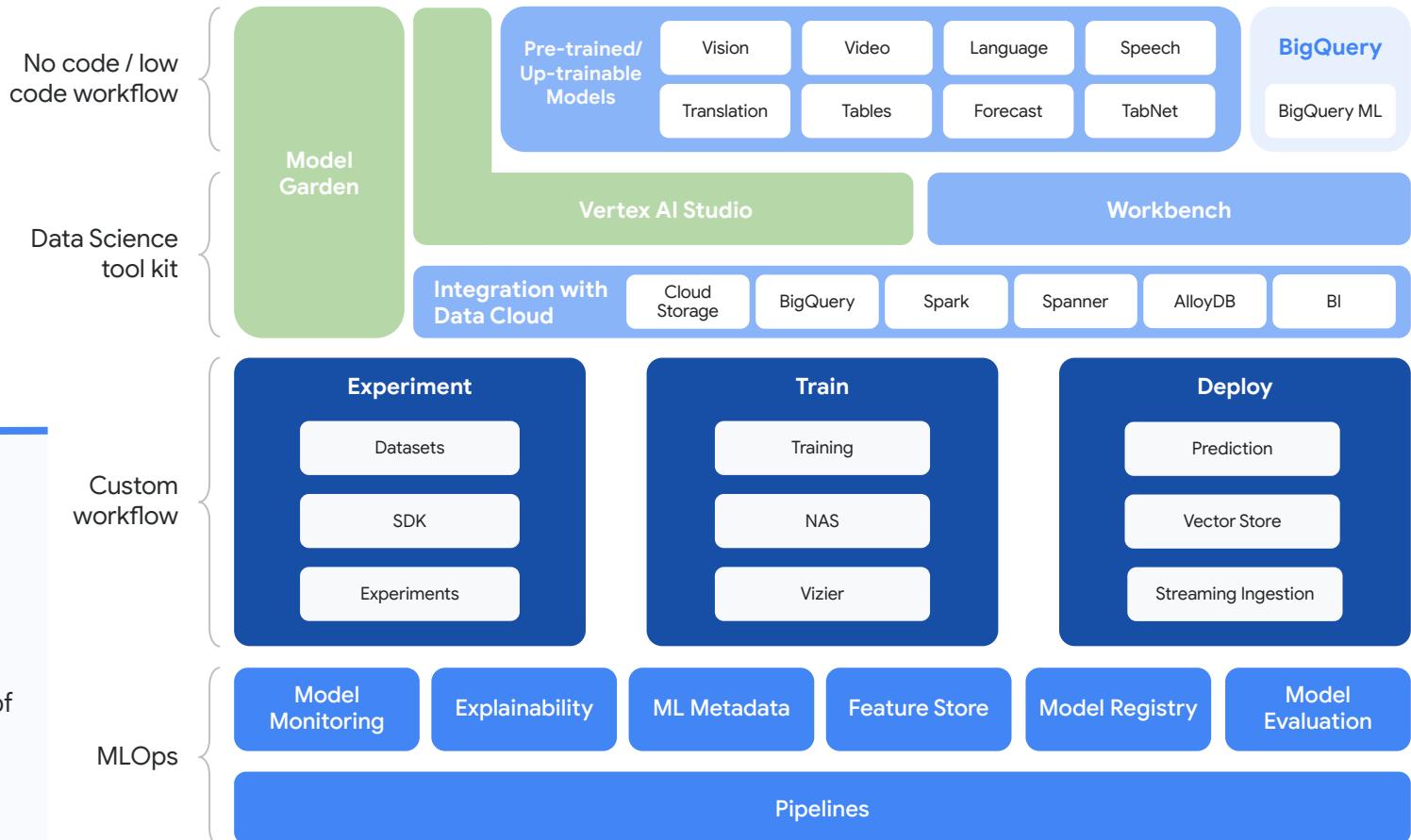
Typical Machine Learning Workflow

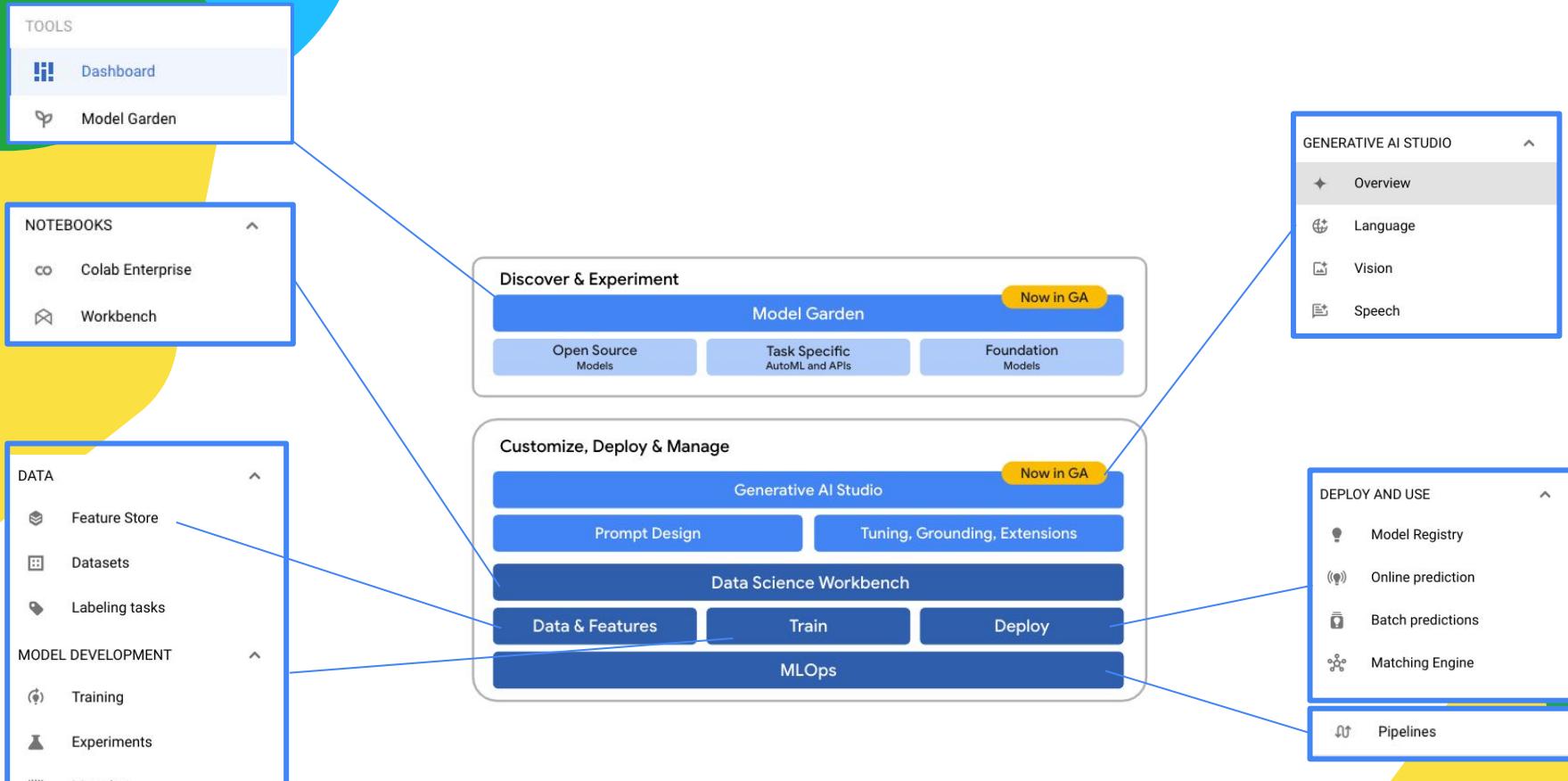




Vertex AI

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





Hyper Parameter Optimization

Vizier



Automatic Hyperparameter tuning

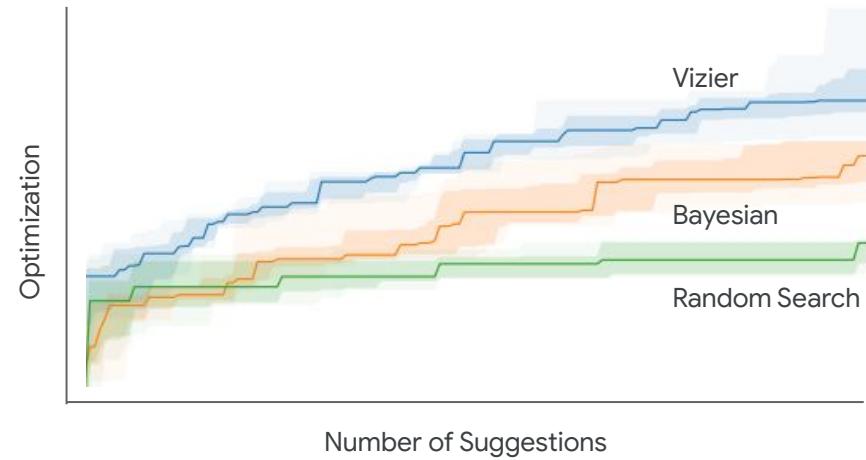
Easily sweep through different configurations to optimize your deep ML models using Vertex AI's HPTuning Jobs



Cloud Vizier Optimizer

Google's generalized optimization engine, used by major Google apps for better and more efficient optimization optimization

Vizier optimizers better and faster



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

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 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 **best model for prod deployment**
- Assess the performance of **models** on different slices and evaluated annotations



Model Garden



Gemini
Foundation
Models

Gemini 1.0
Pro

Gemini 1.5
Pro

Gemini 1.0
Ultra

Google
Foundation
Models

PaLM 2

Imagen

Chirp

Codey

Embeddings API
Embeddings

Google Task
Specific
Models

Speech-to-Text

Text-to-Speech

Natural Language

Translation

Doc AI OCR

Occupancy analytics

Vision

Video Intelligence

Google
Domain
Specific
Models



MedLM
Life Science and
Healthcare



Sec-PaLM
Cybersecurity

Partner &
Open
Ecosystem

Llama 2
Code Llama

TII
Technology
Innovation
Institute

Falcon

Claude 2
Pre-announce

MISTRAL
AI_

Gemma

- **Choice and flexibility** with Google, open source, and third-party foundation models
- **Multiple modalities** to match every use case
- **Multiple model sizes** to match cost and efficacy needs
- **Domain-specific models** for specialized industries
- Enterprise ready with **safety, security, and responsibility**
- Decrease time to value with **fully integrated platform**



So why Vertex AI?

- No need for managing infra, focus on your job
- Unlimited amount of parallel training power on demand
- Best tools, like Vizier for hyperparameters tuning
- All your models, datasets, etc... in one place
- Access to the best Generative and Foundation AI models, that you could fine tune to meet your needs



DEMO



Q & A

