Cloud Infrastructure

ULiège | INFO9023 - Machine Learning Systems Design

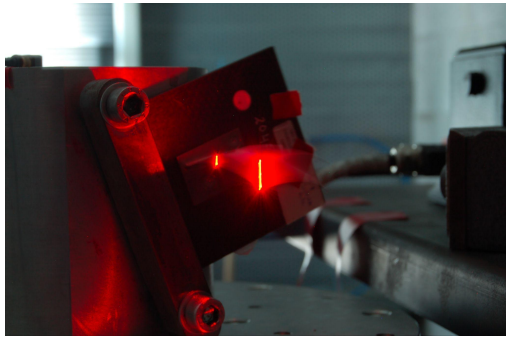
March 18th 2024

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

kaggle

TABLE OF CONTENTS

About me – Philippe Modard



Aerospace Engineering University Liège

Universidad Politécnica Madrid

Master Management HEC

V2i

Modygo

Storify

livefyre

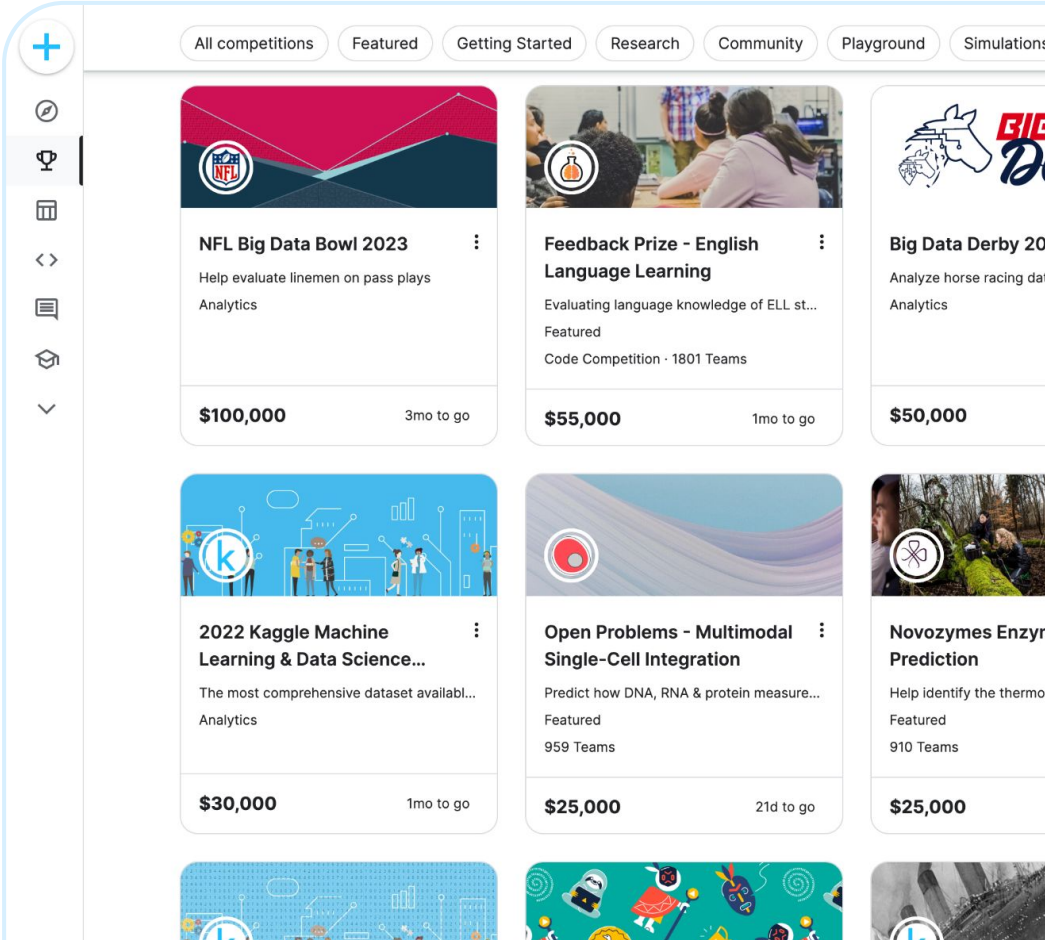
GCB

Software Engineer
Tech Lead

Kaggle platform

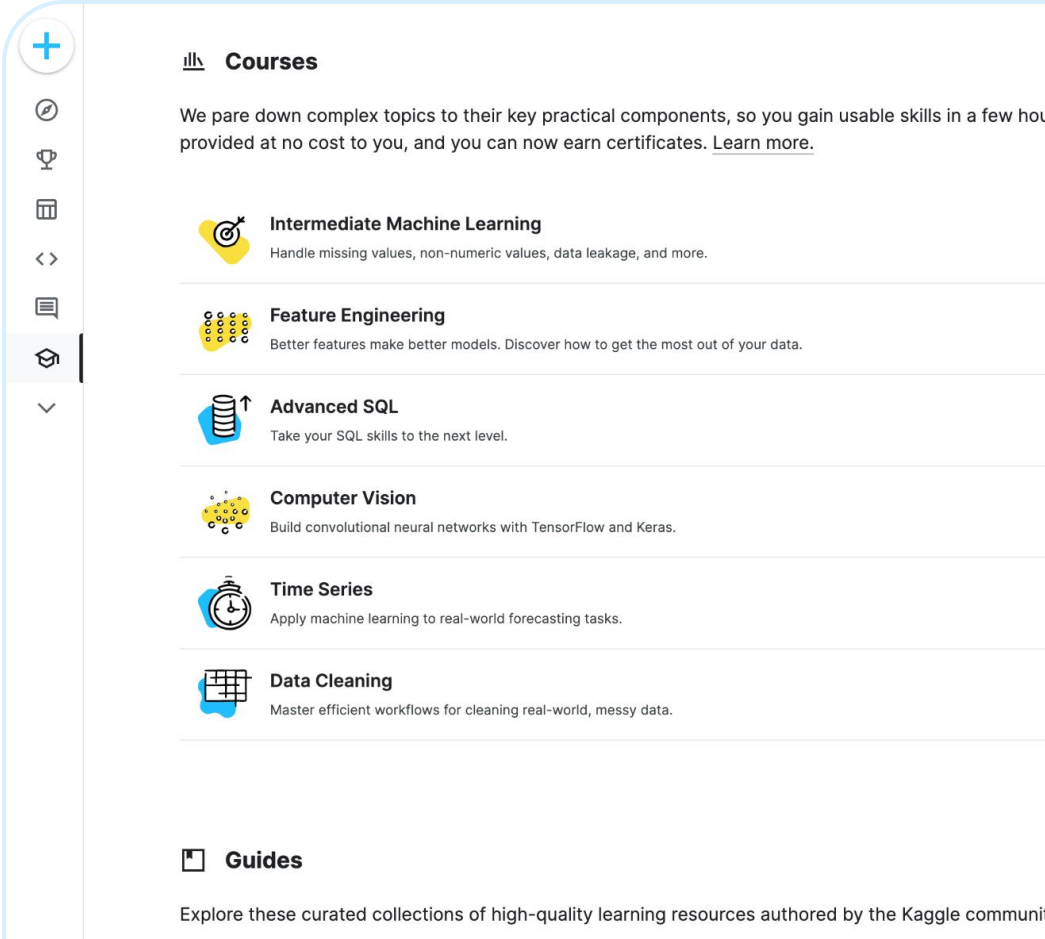
Competitions

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



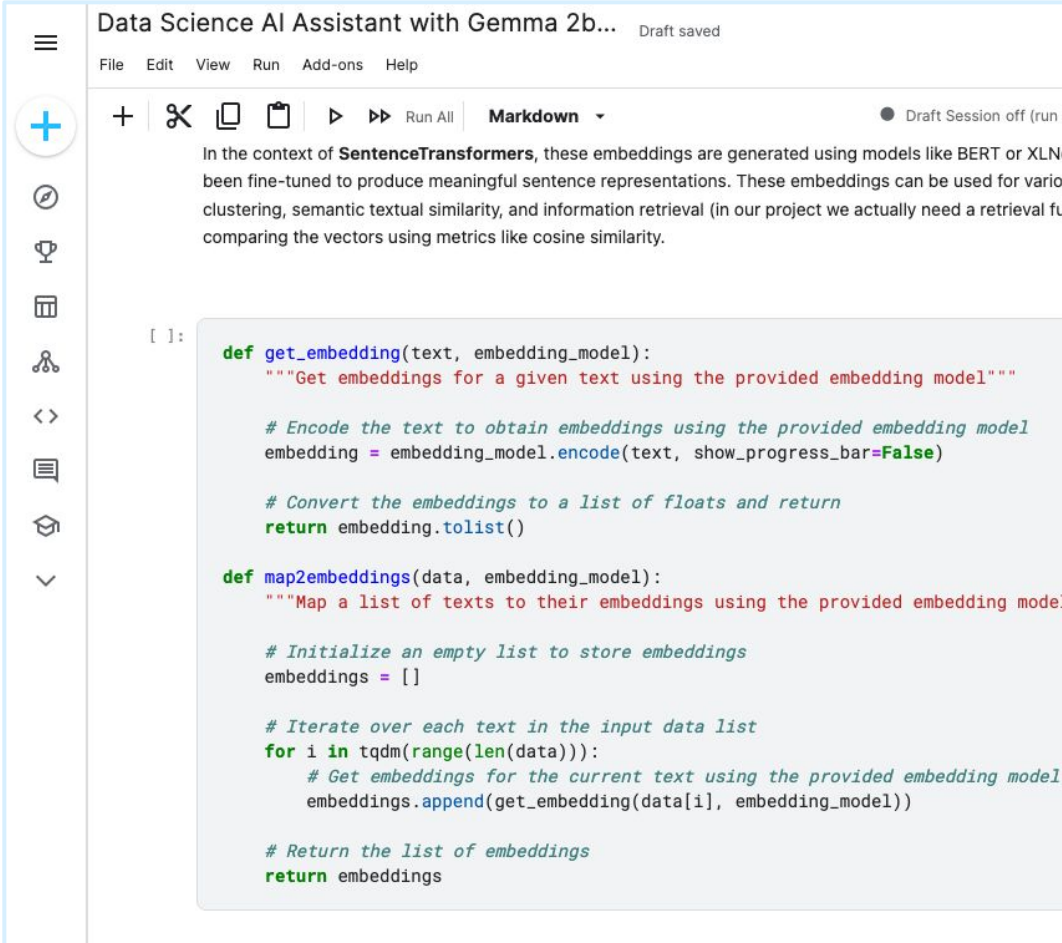
Learn

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



Code

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



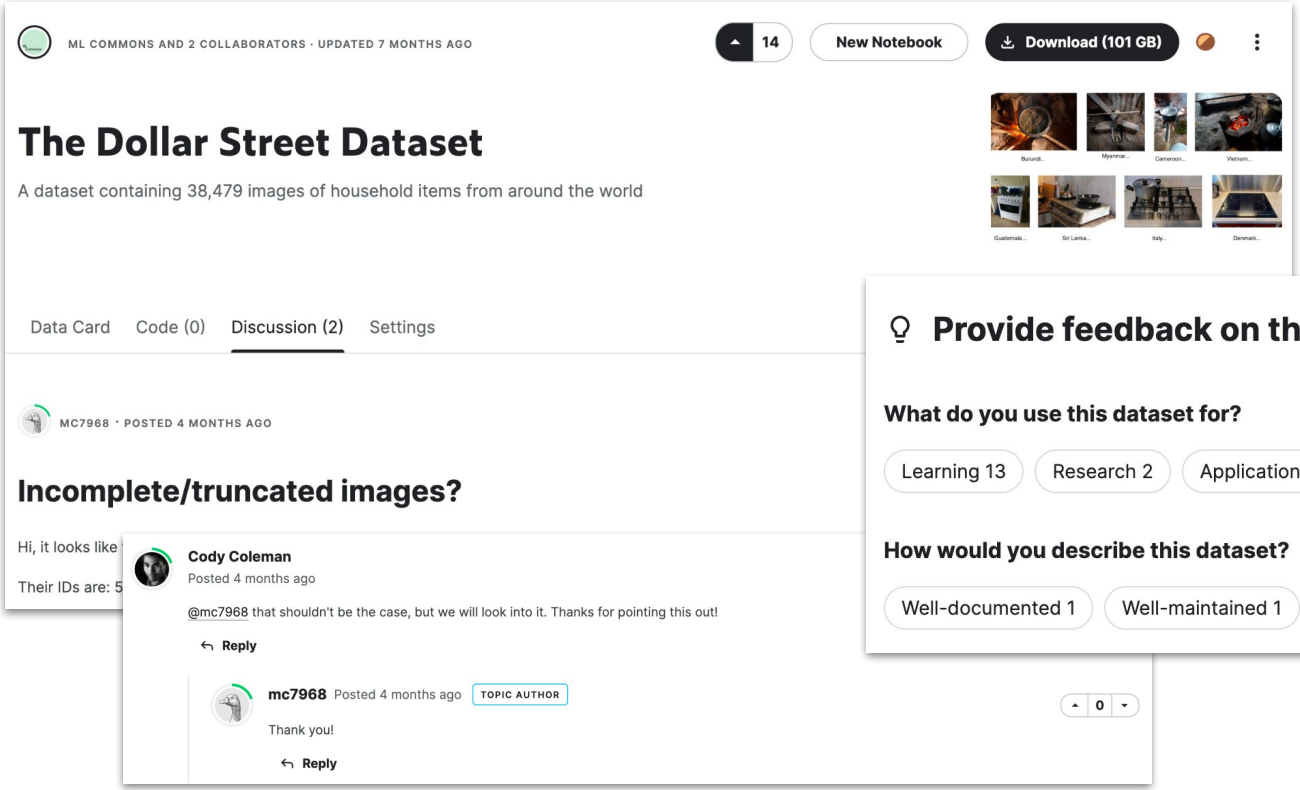
Largest ML Hub

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

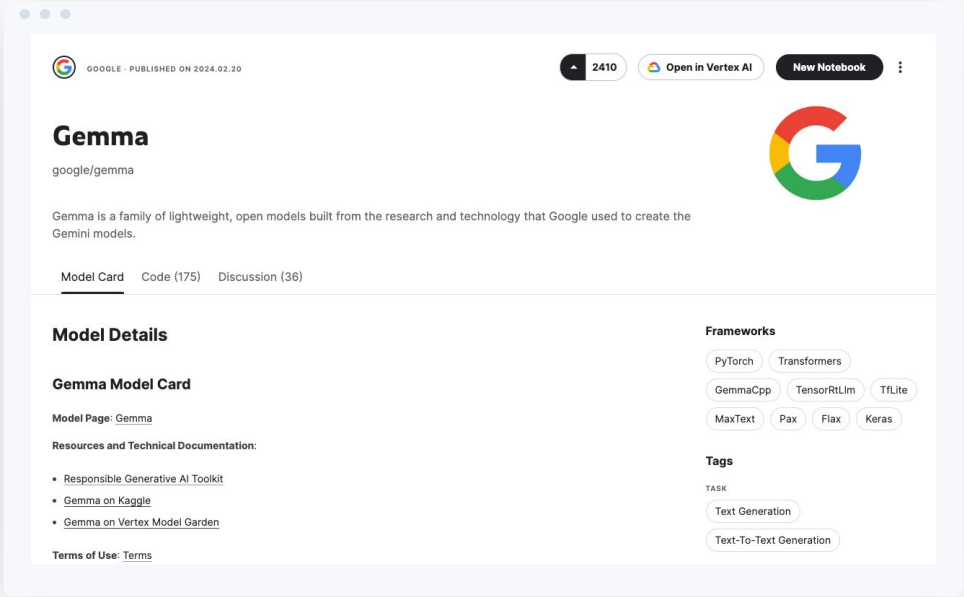
200K
Public Datasets

3K
Public Models

A community makes datasets into living, vibrant resources



Gemma



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

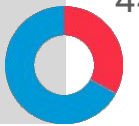
Test data is used to evaluate prediction accuracy

Training data

Age	Income	Claim Filed
58	$95,824	True
73	$20,708	False
59	$82,152	False
66	$25,334	True

Test data

Age	Income	Claim Filed
73	$53,445	
61	$36,679	
47	$90,422	
44	$79,040	



Submissions are ranked on a live leaderboard

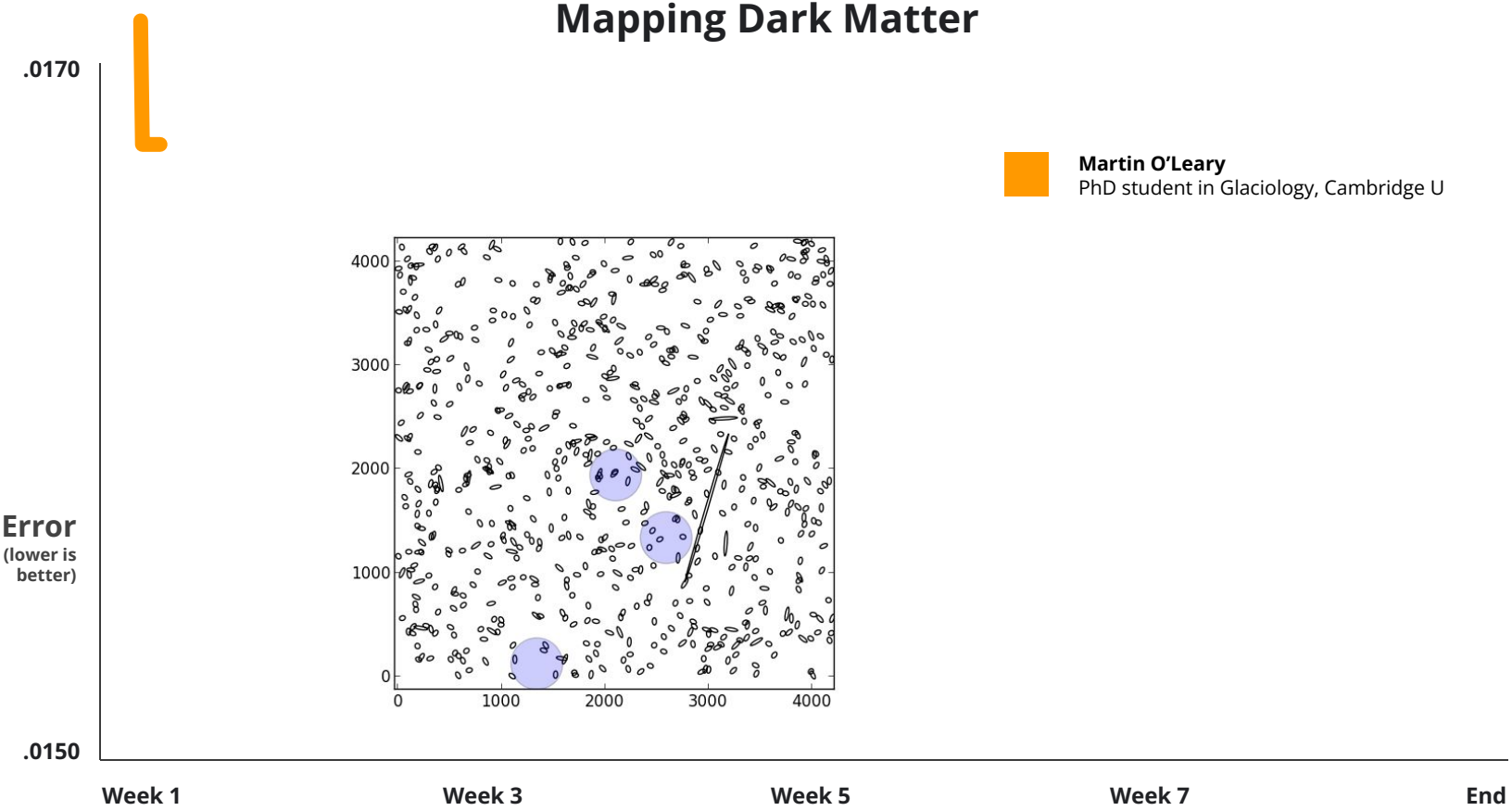
Public Private

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



"The world's brightest physicists have been working for decades on solving one of the great unifying problems of our universe"

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

Competition Shines Light on Dark Matter

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

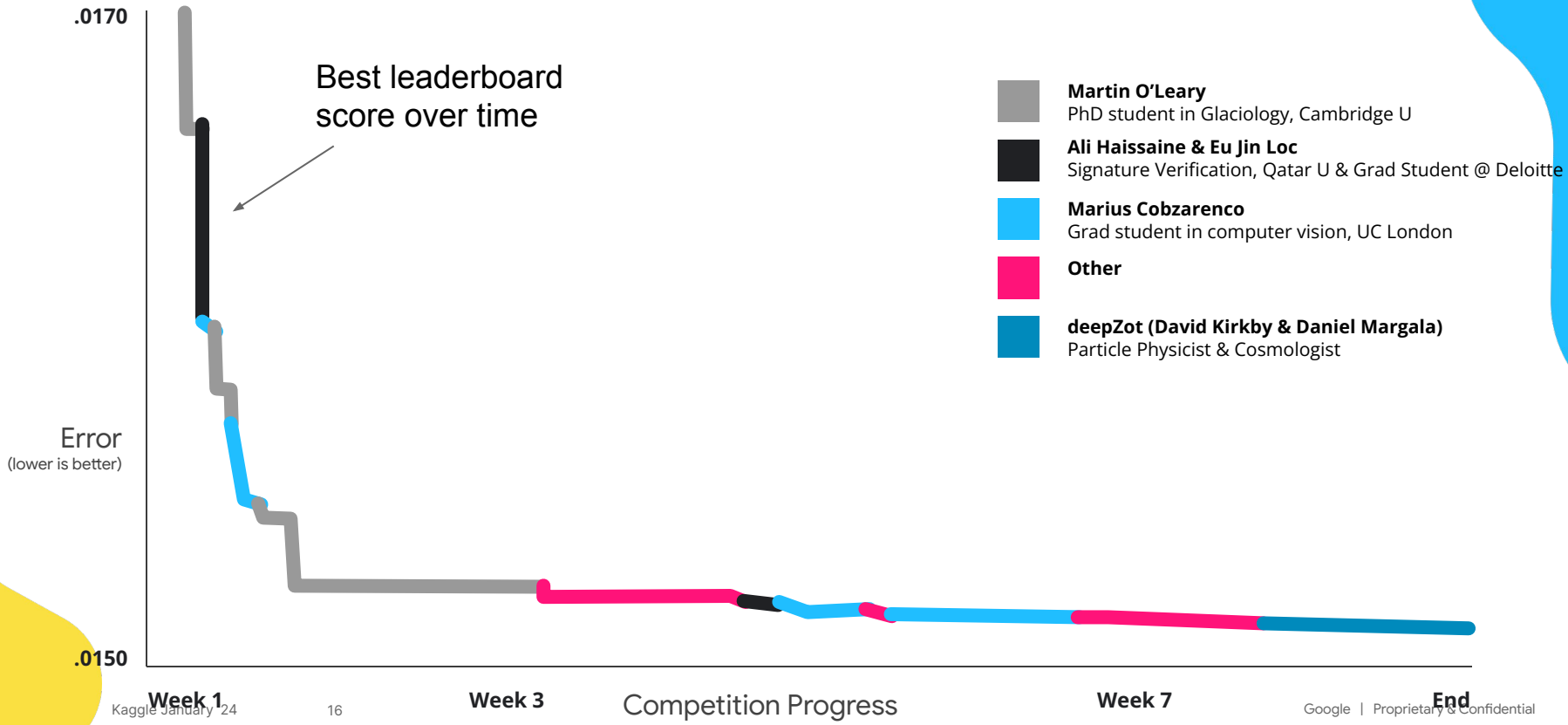
The world's brightest physicists have been working for decades on solving one of the great unifying problems of our universe. It is a problem that explores our place in the cosmos and, as was the case with Newton's law of gravitation and Einstein's theory of relativity, would change our understanding of the Universe if solved. Recently, top experts challenged anyone to find a breakthrough from an unexpected place.

On May 23, a consortium of the very best from the data-mining, computer science and astronomy communities posted the problem on the data science challenge site Kaggle.com for all the world to weigh in. In less than a week, Martin O'Leary, a PhD student in glaciology, 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 crowdsoucing, and this Admnistration'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,

Competitions extract all the signal from a dataset

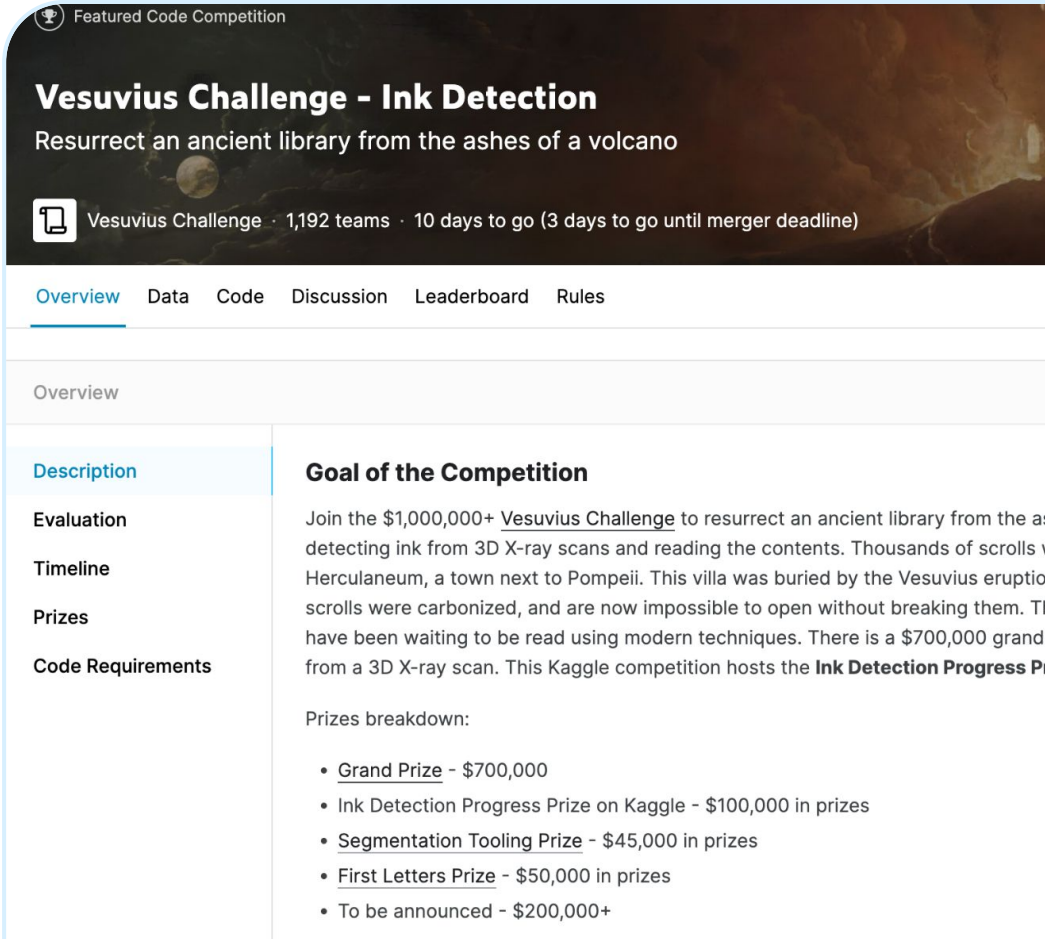


Some recent unique competitions

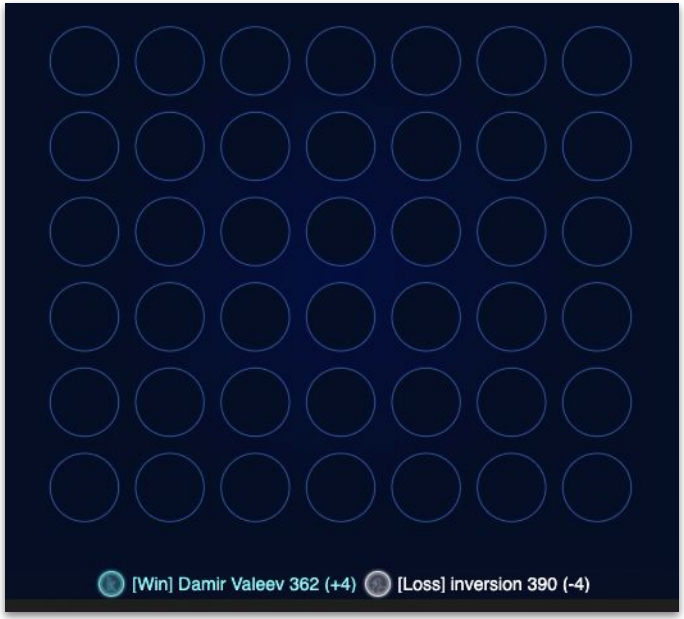
Vesuvius Ink Detection

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

$1,000,000 in prizes



Simulations - Reinforcement Learning Competitions



Intro to Game AI and Reinforcement Learning Home Page

Introduction

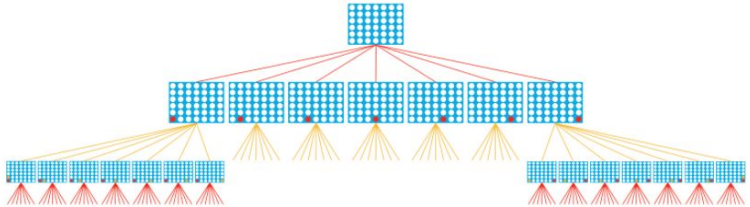
Even if you're new to Connect Four, you've likely developed several game-playing strategies. In this tutorial, you'll learn to use a **heuristic** to share your knowledge with the agent.

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.

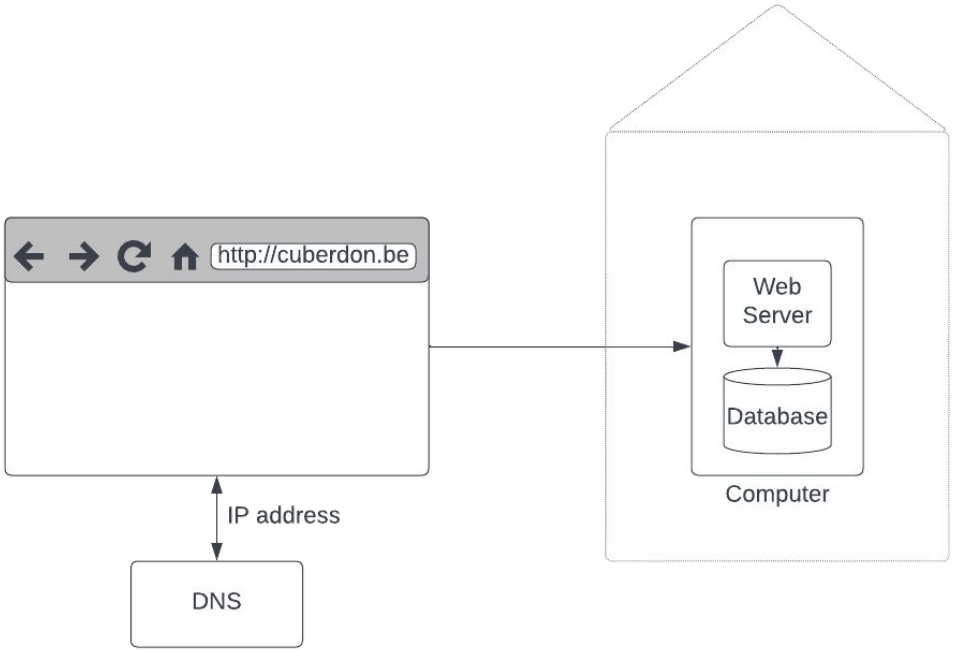
We can formalize this idea and represent all possible outcomes in a **(complete) game tree**.



Cloud Computing

Introduction

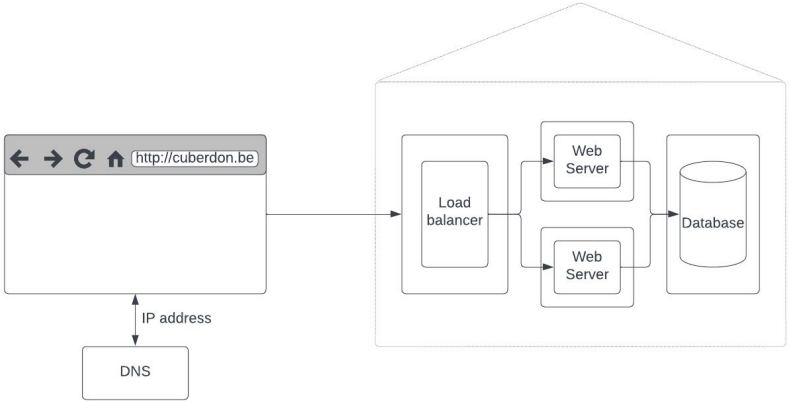
Imagine you are building a new online store website to sell ... cuberdons.



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

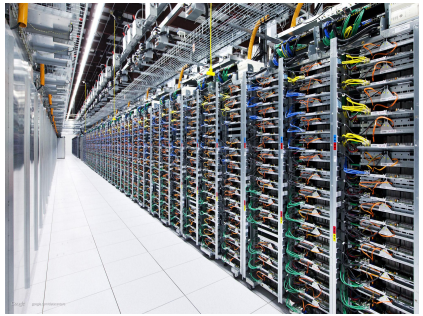
- Successful services needs to scale
- Difficult to provision hardware

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.

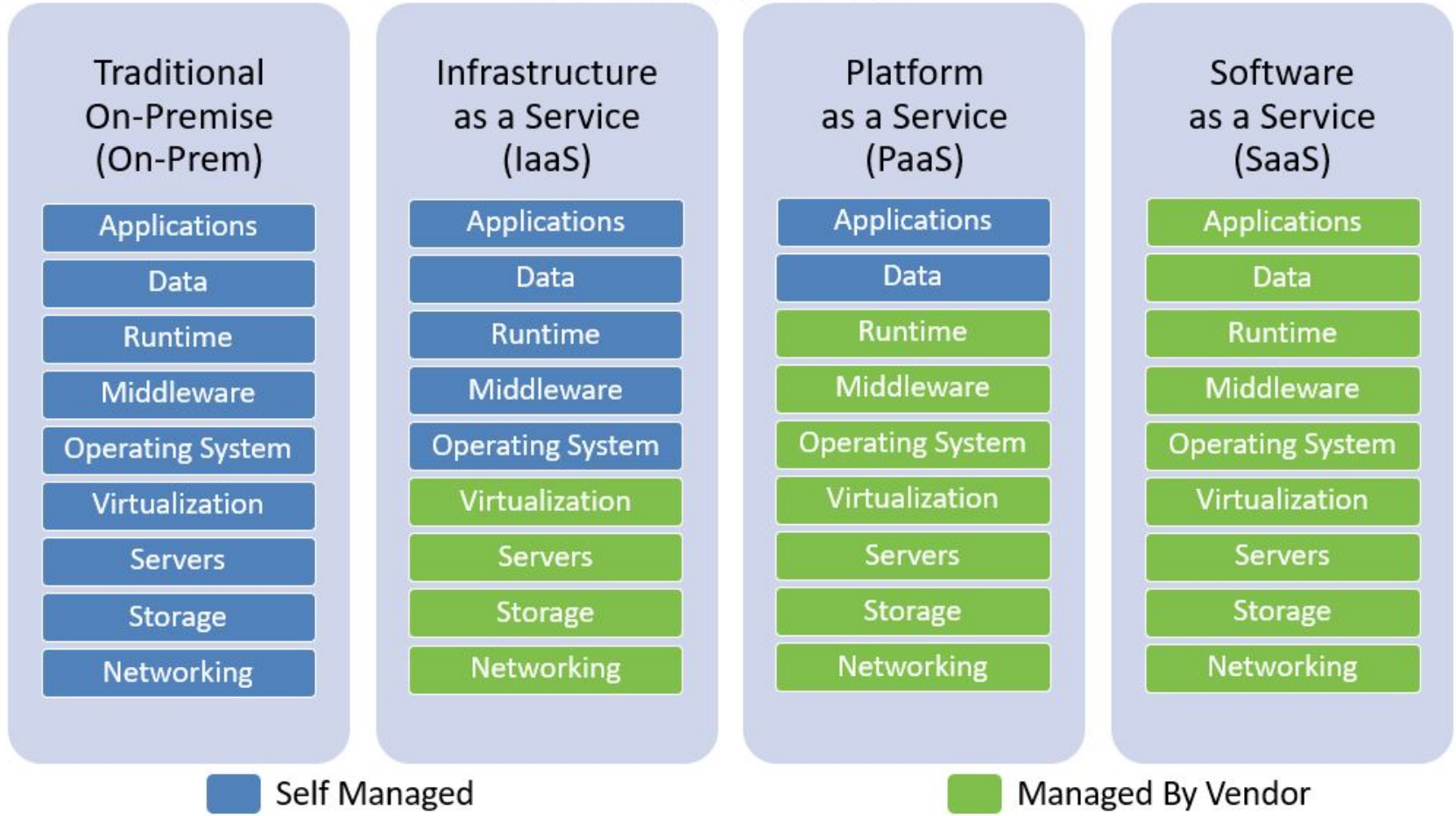
amazon web services™

Microsoft Azure

Google Cloud Platform

Multi-cloud. For best of redundancy.

	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
• 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	• 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**
• 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	• 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**
• 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	• 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**
• 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	• 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

Google Cloud Platform

Powered by 20+ years of Google Innovation

Google
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

The term **"Cloud Computing"** enters the common vernacular.

Process Containers initiative to bring containers to the Linux kernel

2008

April 2008
Google Cloud is launched

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

2013

Work begins to open source Google's Borg as Kubernetes

2013 Docker launched

2015

Nov 2014
GKE Alpha

June 2014
First Github commit for Kubernetes

Aug 2015
GKE GA

July 2015
Kubernetes 1.0

Envoy 1.0

2018

Knative announced

GKE on Prem Announced

Istio 1.0

gVisor

Istio announced

2019

Anthos

Cloud Run

2021

GKE Autopilot

Google Cloud Platform

Current region with 3 zones

Future region with 3 zones

Edge point of presence

Regions and PoPs

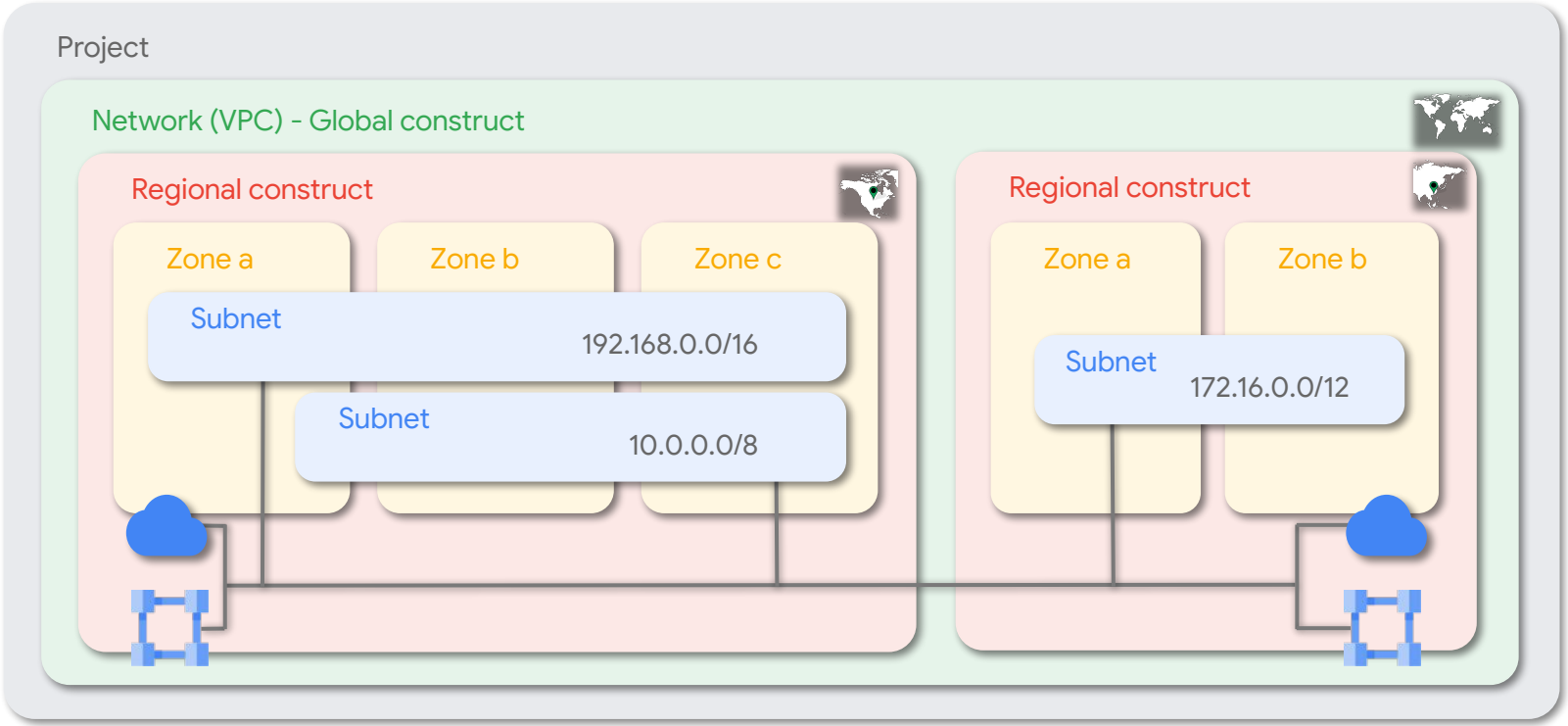
*Exception: region has 4 zones.

Google Cloud

Location Considerations

	Regional	Dual-Region	Multi-Region
Availability	• Data r**edundancy across availability zones (synchronous)** • **RTO=0:** automated failover and failback on zonal failure (no need to change storage paths)	• **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)	• **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	• **200 Gbps** (per region, per project) • Scalable to many Tbps by requesting higher bandwidth quota	• **200 Gbps** (per region, per project) • Scalable to many Tbps by requesting higher bandwidth quota	• **50 Gbps** (per region, per project) • Limited performance scaling, variable performance for reads
Pricing	• **Lowest storage price** • No replication charges • No egress charges when reading data inside the same region	• **Highest storage price** • Replication charges apply on write • No egress charges when reading data within either region	• **Higher storage price than regional, but lower than dual-region** • Replication charges apply on write • Egress charges always apply when reading data

Network Concepts



Running Code

Compute Engine

App Engine

Cloud Run

Complexity
Customizability

Storing Data

Storage

SQL

Firestore

Big Data / ML

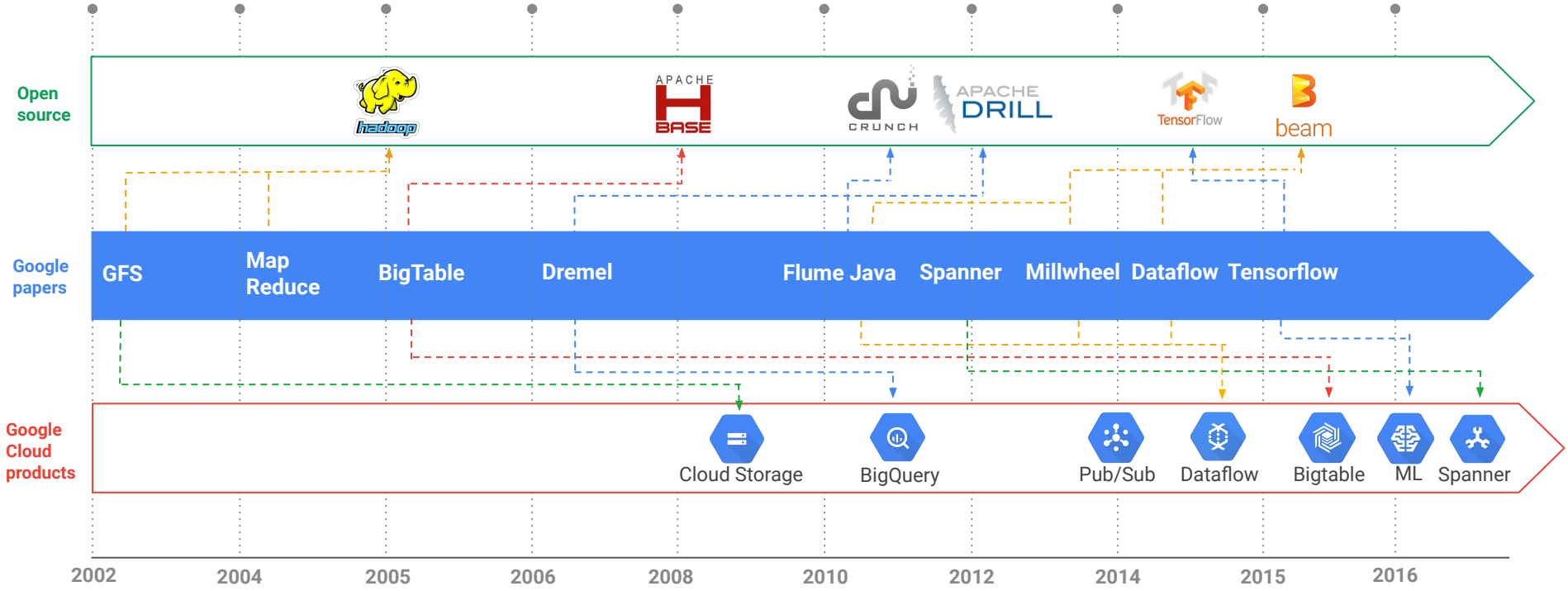
BigQuery

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

DEMO

Google Vertex AI

A deep history of research and innovation at Google

2017
Transformer

Google invents Transformer kickstarting LLM revolution

2018
BERT

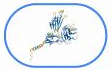
Google's groundbreaking large language model, BERT

2019
T5

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

2020
LaMDA

Google LaMDA model trained to converse



2021
AlphaFold

AlphaFold predicts structures of all known proteins

2022
PaLM

Industry leading large language model

2023
Gemini

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

- Ingest
- Analyze
- Transform

Datasets
Labeling

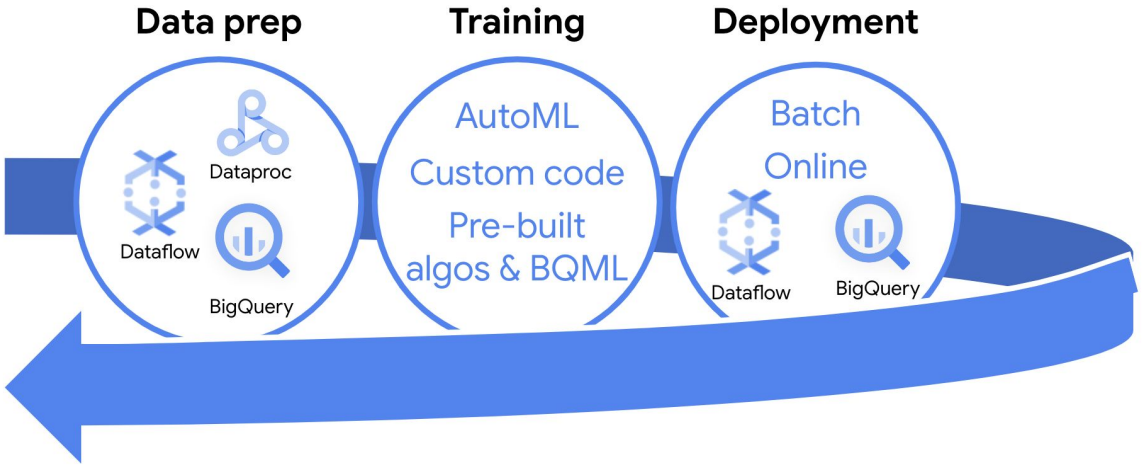
- Train
- Model

AutoML
Custom

- Evaluate
- Deploy
- Predict

Registry
Predictions
CLI, UI, SDK, API

- Monitoring



Vertex AI

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

No code / low code workflow

Data Science tool kit

Custom workflow

MLOps

Model Garden

Pre-trained/ Up-trainable Models

Vision | Video | Language | Speech | Translation | Tables | Forecast | TabNet

BigQuery

BigQuery ML

Vertex AI Studio

Workbench

Integration with Data Cloud

Cloud Storage | BigQuery | Spark | Spanner | AlloyDB | BI

Experiment

Datasets

SDK

Experiments

Train

Training

NAS

Vizier

Deploy

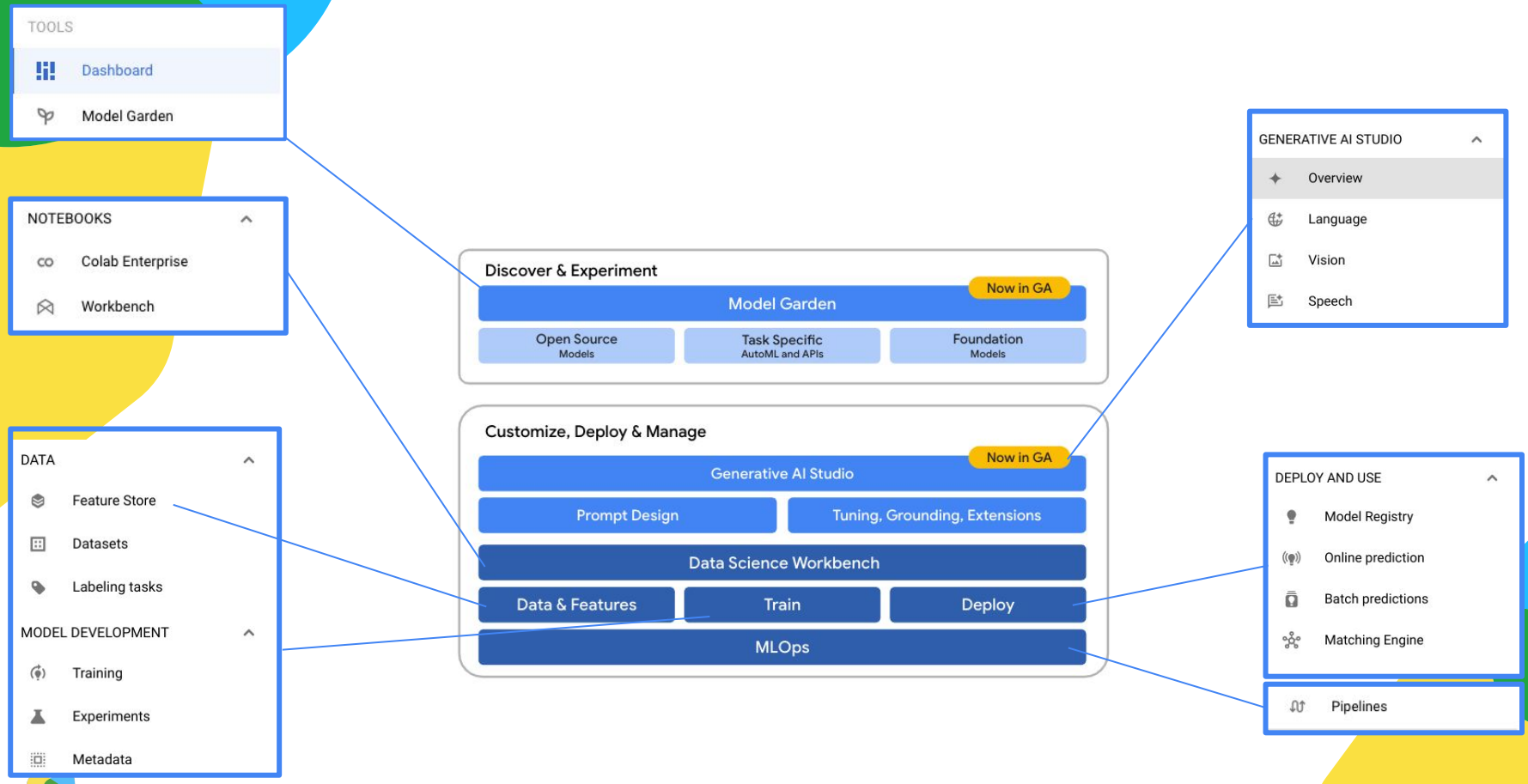
Prediction

Vector Store

Streaming Ingestion

Model Monitoring | Explainability | ML Metadata | Feature Store | Model Registry | Model Evaluation

Pipelines


TOOLS
Dashboard
Model Garden
NOTEBOOKS
Colab Enterprise
Workbench
DATA
Feature Store
Datasets
Labeling tasks
MODEL DEVELOPMENT
Training
Experiments
Metadata
Discover & Experiment
Model Garden
Now in GA
Open Source
Models
Task Specific
AutoML and APIs
Foundation
Models
Customize, Deploy & Manage
Generative AI Studio
Now in GA
Prompt Design
Tuning, Grounding, Extensions
Data Science Workbench
Data & Features
Train
Deploy
MLOps
GENERATIVE AI STUDIO
Overview
Language
Vision
Speech
DEPLOY AND USE
Model Registry
Online prediction
Batch predictions
Matching Engine
Pipelines

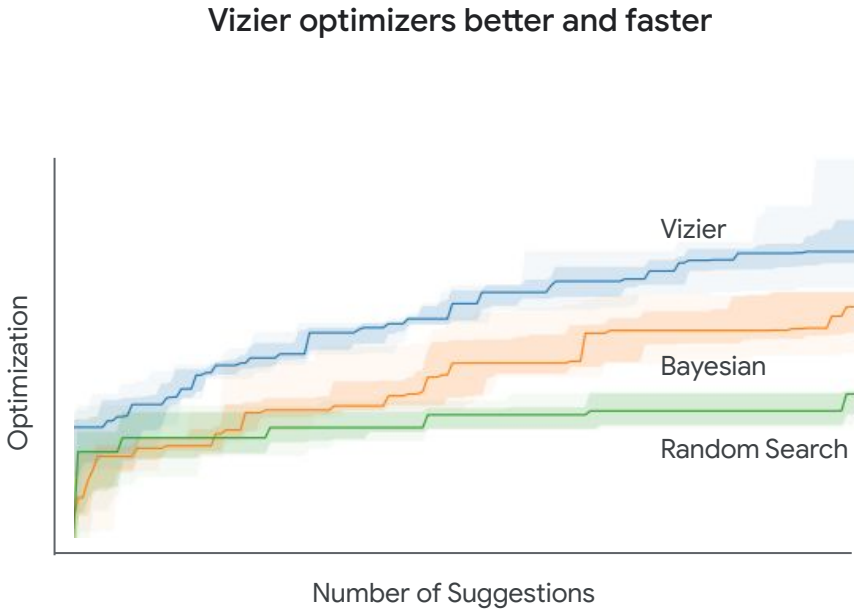
Hyper Parameter Optimization
Vizier

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



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
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 · 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