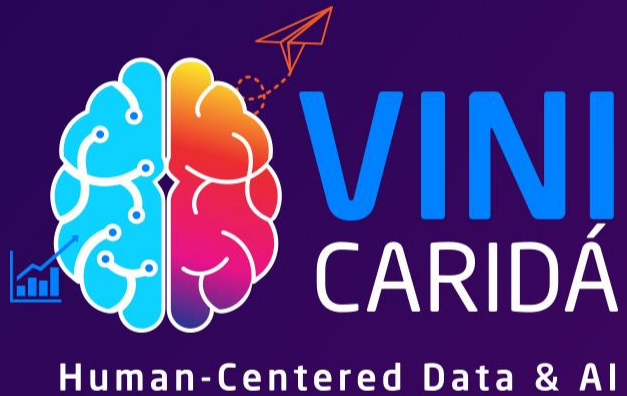

VINI
CARIDÁ
Human-Centered Data & AI

VINICARIDÁ
Human-Centered Data & AI

Vinicius Caridá, Ph.D.

- Executive Specialist, Artificial Intelligence and Data - Itaú
- MBA Professor – FIAP and ESPM

@vinicius caridá @vfcarida @vinicius caridá @vfcarida @vinicius caridá @vfcarida

Inteligência de Dados

- ORGANIZAÇÃO DATA DRIVEN
- VISUALIZAÇÃO DE DADOS
- MODELOS PREDITIVOS PARA A TOMADA DE DECISÃO
- DATA STORYTELLING E DATA ART
- CIÊNCIA DE DADOS APLICADA AOS NEGÓCIOS
- INTELIGÊNCIA ARTIFICIAL
- JORNADA ANALÍTICA

Inteligência de Dados

- ORGANIZAÇÃO DATA DRIVEN
- VISUALIZAÇÃO DE DADOS
- MODELOS PREDITIVOS PARA A TOMADA DE DECISÃO
- DATA STORYTELLING E DATA ART
- CIÊNCIA DE DADOS APLICADA AOS NEGÓCIOS
- INTELIGÊNCIA ARTIFICIAL
- JORNADA ANALÍTICA

Inteligência de Dados

- ORGANIZAÇÃO DATA DRIVEN
- VISUALIZAÇÃO DE DADOS
- MODELOS PREDITIVOS PARA A TOMADA DE DECISÃO
- DATA STORYTELLING E DATA ART
- CIÊNCIA DE DADOS APLICADA AOS NEGÓCIOS
- INTELIGÊNCIA ARTIFICIAL
- JORNADA ANALÍTICA

Três caminhos de AI na Google Cloud

seus dados + seu modelo


Cloud TPUs


Compute Engine


Cloud Dataproc

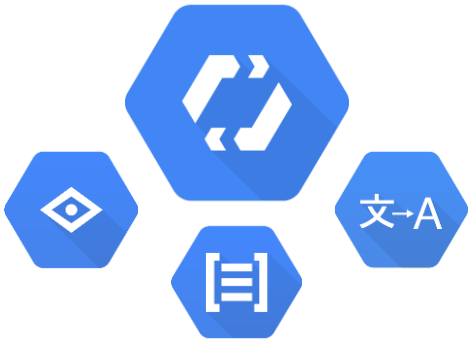

Kubernetes Engine


Cloud ML Engine


BigQuery ML

seus dados + nossos modelos

AutoML



nossos dados + nossos modelos


Cloud Translation API


Cloud Vision API


Cloud Speech API


Cloud Video Intelligence API


Data Loss Prevention API


Cloud Speech Synthesis API


Cloud Natural Language API


Dialogflow

Customização

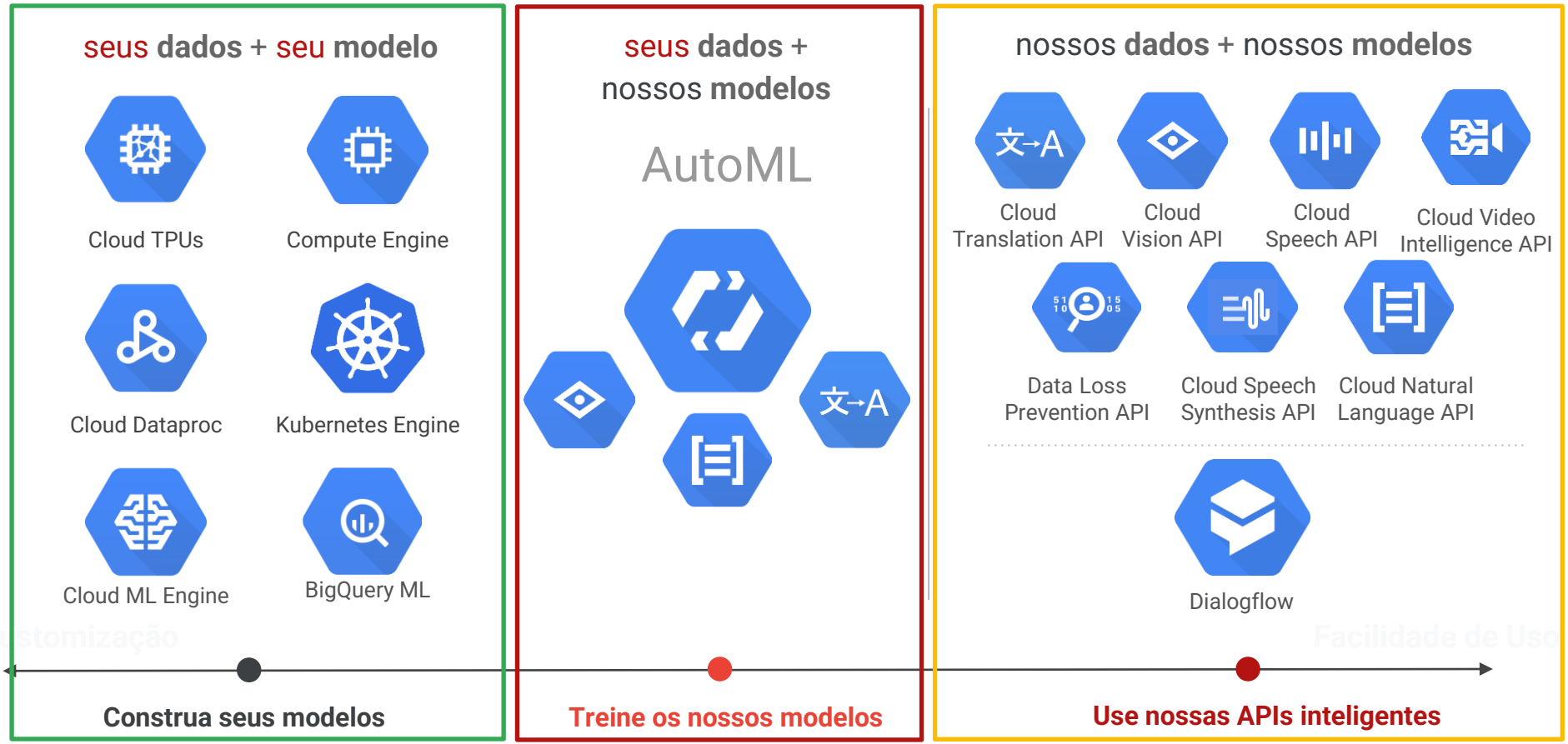
Facilidade de Uso

Construa seus modelos

Treine os nossos modelos

Use nossas APIs inteligentes

Três caminhos de AI na Google Cloud

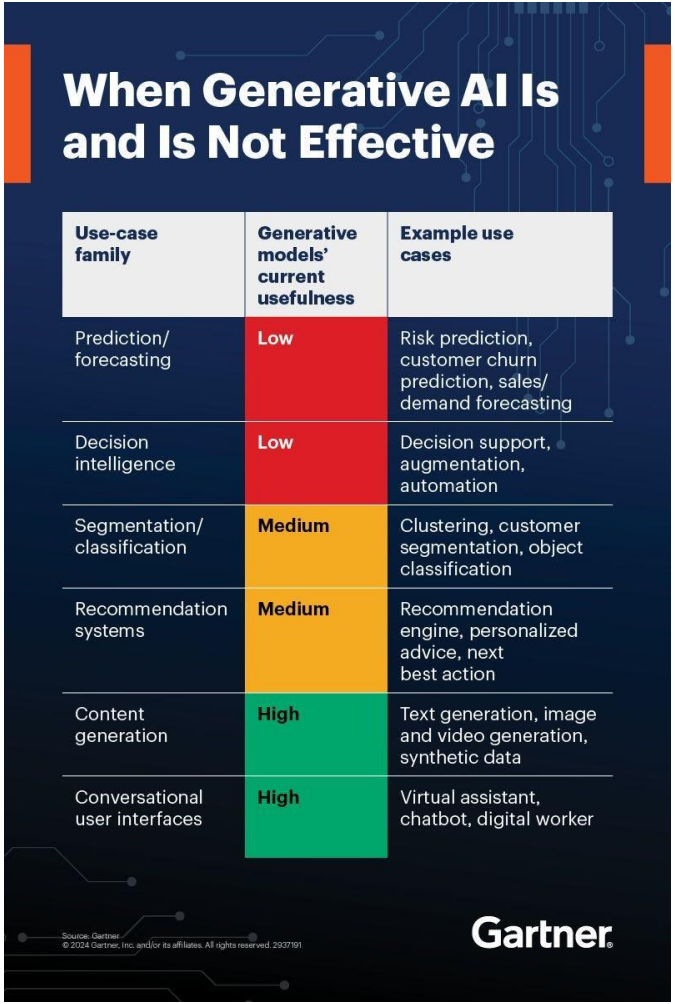


When Generative AI Is and Is Not Effective

Use-case family	Generative models' current usefulness	Example use cases
Prediction/ forecasting	Low	Risk prediction, customer churn prediction, sales/ demand forecasting
Decision intelligence	Low	Decision support, augmentation, automation
Segmentation/ classification	Medium	Clustering, customer segmentation, object classification
Recommendation systems	Medium	Recommendation engine, personalized advice, next best action
Content generation	High	Text generation, image and video generation, synthetic data
Conversational user interfaces	High	Virtual assistant, chatbot, digital worker

Source: Gartner

Gartner

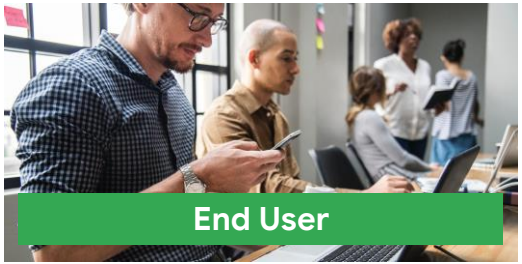


1. Ambiente Cloud
2. Algoritmos e técnicas
3. Fuzzy e Algoritmos Genéticos
4. Redes Neurais avançadas
5. Inteligência Artificial Generativa

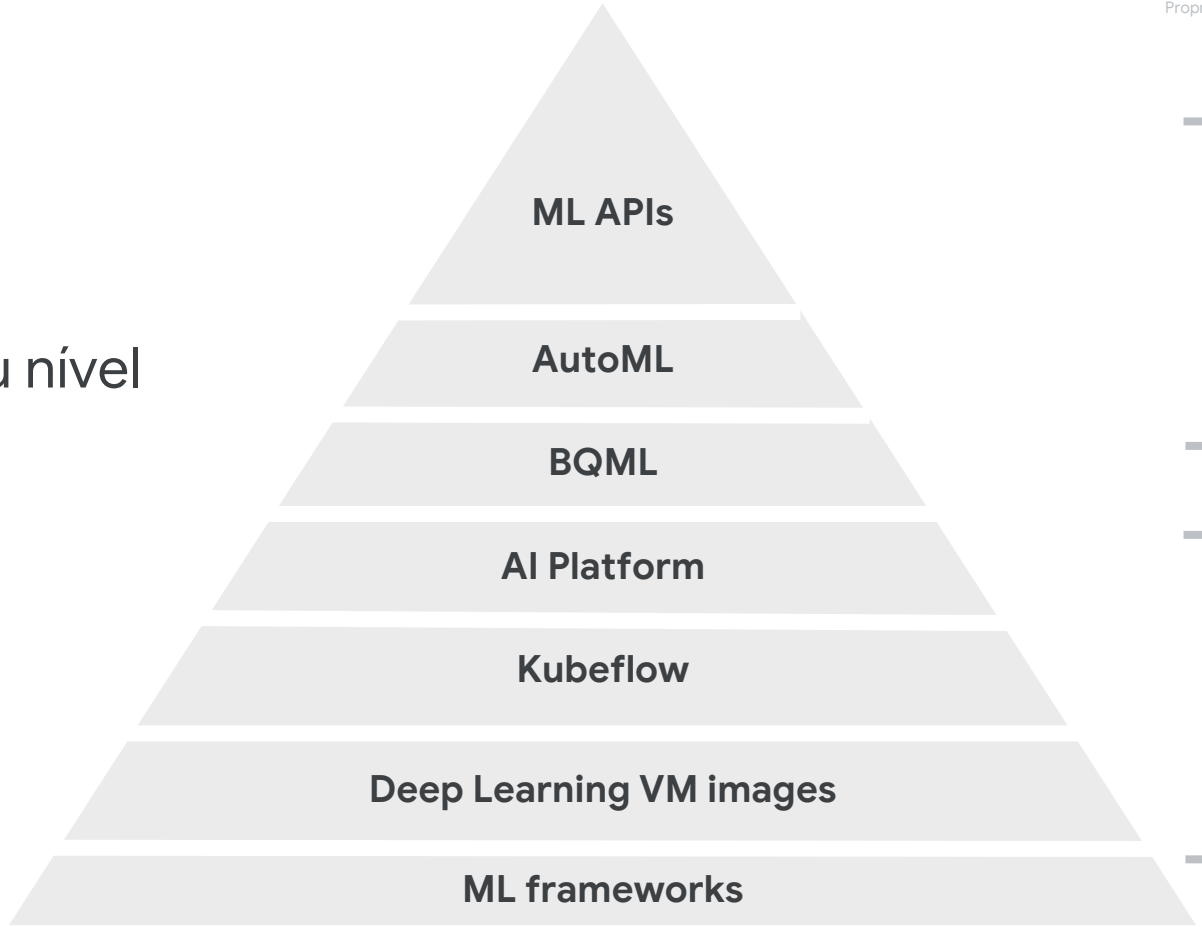
AI Platform Overview

Google Cloud

Como conectar usuários para produzir resultados 10x mais impactantes?

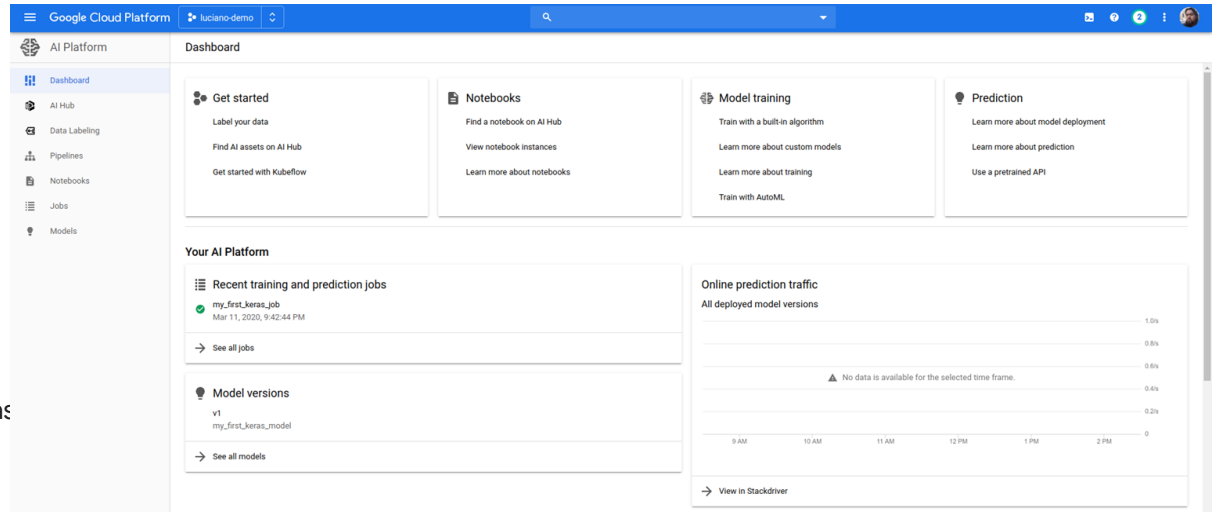


Escolha o seu nível de abstração



Introdução ao Cloud AI Platform

- Ambiente de desenvolvimento fim-a-fim para IA dentro da console GCP
- Construído com Kubeflow, oferece uma cadeia de ferramentas integrada - de engenharia de dados à implementação de modelos. Sem “lock-in”
- Permite implementar suas soluções on-premises ou na Google Cloud sem mudanças significativas de código
- Acesso à tecnologias Google AI de ponta como Tensorflow, Tensorflow Extended (TFX), TPUs quando implementando sua solução em produção



O que está incluso no Cloud AI Platform

AI Hub

Kubeflow
(On premises)

Deep learning VM image (DLVM)

Machine Learning mais rápido e fácil no GCE

- **Prototipação Acelerada**
  Prototipe seu projeto rapidamente utilizando VMs pré-configuradas para Deep Learning

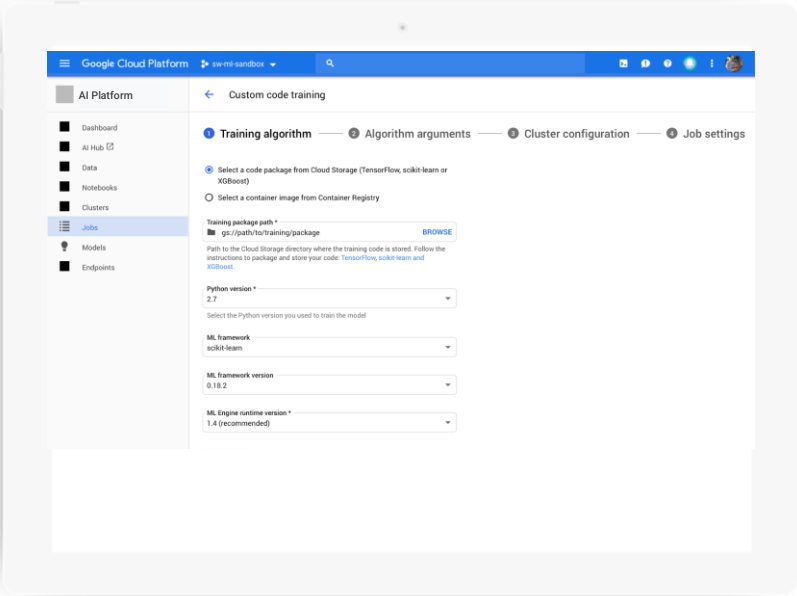
- **Suporte a CPU, GPU e TPU**
  Utilize aceleradores para seus modelos, como GPU ou TPU, com poucos cliques

- **Performance otimizada para Google Cloud**
  Bibliotecas e configurações otimizadas para a melhor performance para a sua infraestrutura - assim você não se preocupa com isso

- **Flexibilidade**
  Escolha entre diferentes ML frameworks como TensorFlow, PyTorch e scikit-learn ou instale o framework que você preferir sob a imagem base

Gerenciamento serverless de modelos

- Treine modelos abstraindo a infraestrutura
- Suporta os frameworks de machine learning mais populares. Suporta, também, containers Docker customizados
- Realiza treinamentos distribuídos e utiliza GPUs e TPUs para finalizar jobs mais rápido
- Melhora modelos através da otimização de hiperparâmetros automatizada

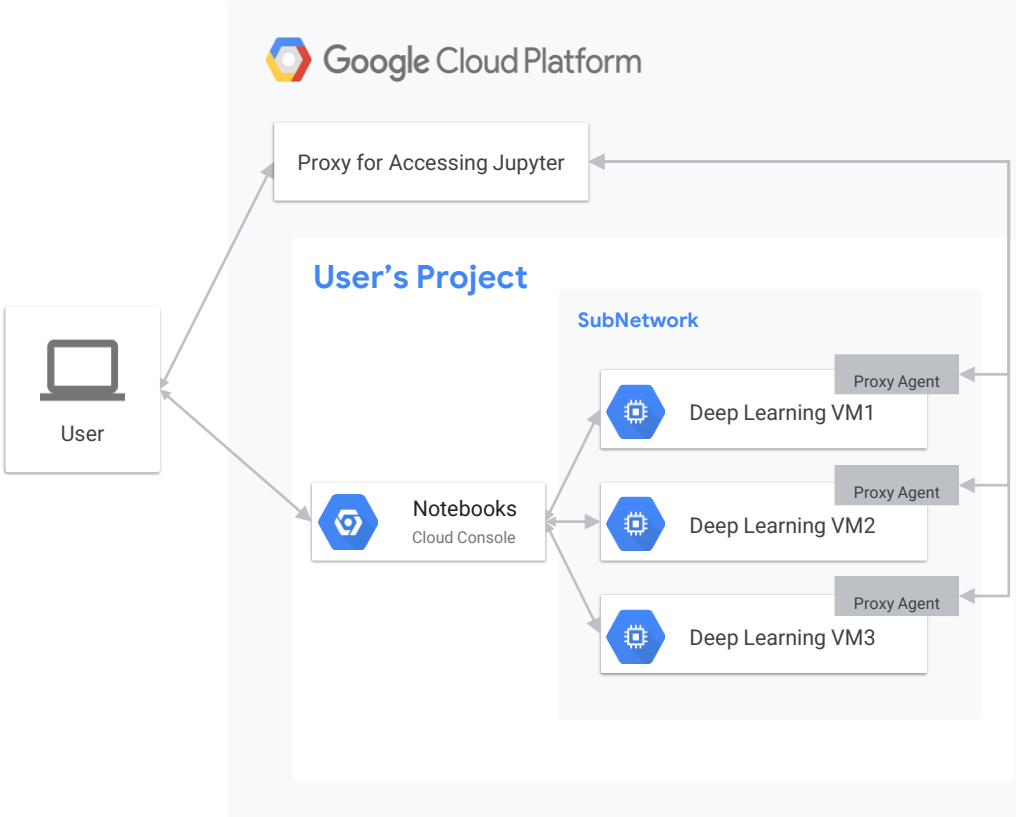


Notebooks gerenciados e ambientes pré-configurados

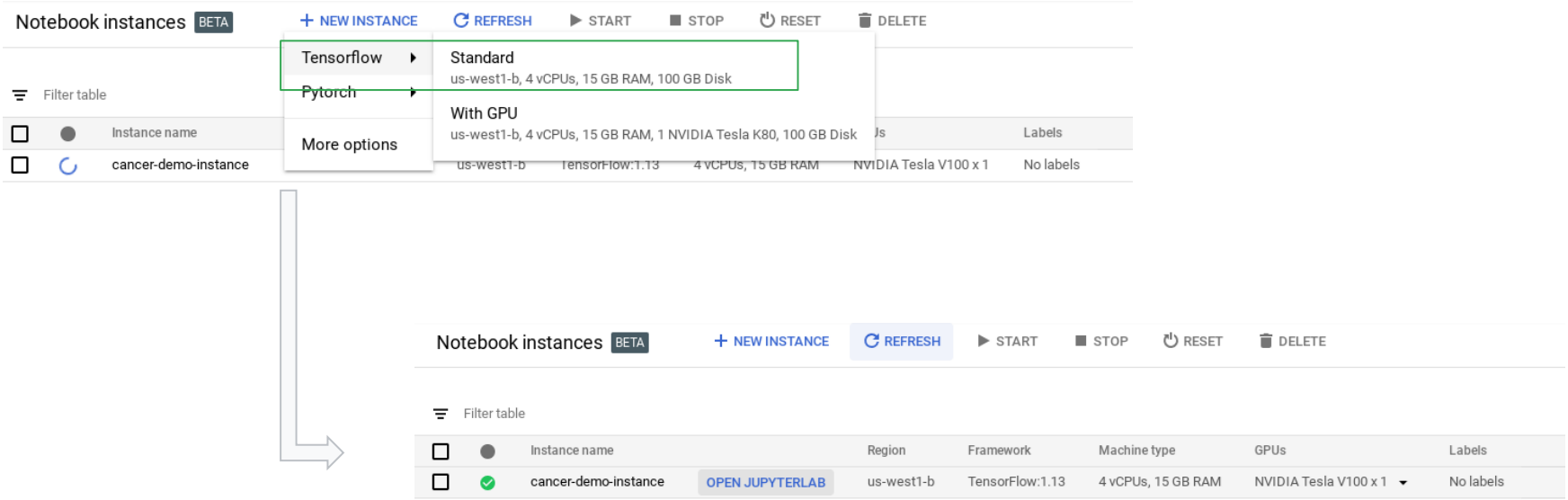
Bibliotecas GCP pré-instaladas

Ambientes pré-configurados para:

NumPy SciPy scikit-learn matplotlib TensorFlow Keras dmlc XGBoost PyTorch



Crie uma instância de Notebook



Importe seus dados do Cloud Storage

[ ]:
```
!pip3 install tensorflow_hub
```

[1]:
```
import sys
# Workaround: pip installs to a wrong location on these notebooks
sys.path.insert(0,'/home/jupyter/.local/lib/python3.5/site-packages')
import os
import json
import random
import base64
import tensorflow as tf
import tensorflow_hub as hub
import tensorflow.keras as ks
import IPython.display
from tensorflow.python.lib.io import file_io

storage_bucket = 'gs://cancer-demo-data/ht-kg-histopathologic-cancer-detection/'
train_files = [storage_bucket+f for f in ['train_0.tfrecords', 'train_1.tfrecords', 'train_2.tfrecords']]
eval_files = [storage_bucket+f for f in ['train_3.tfrecords']]
xval_files = [storage_bucket+f for f in ['train_4.tfrecords']]

row_count_file = storage_bucket + 'record_counts.json'
examples_file = storage_bucket + 'examples.zip'
```

Treine seu modelo onde quiser

On-Premises

```
from fairing.config.config_kubeflow import KubeflowConfig
from fairing.jobs.train_job import TrainJob

job = TrainJob(train,
               base_docker_image="gcr.io/caip-dexter-dev/jaas:py3.5.3",
               destination_dir='gs://cancer_detection_demo/',
               runtime_config=KubeflowConfig())

job.submit()
job.get_logs()
```

Google Cloud

```
from fairing.config.config_cmle import CMLEConfig
from fairing.jobs.train_job import TrainJob

job = TrainJob(train,
               base_docker_image="gcr.io/caip-dexter-dev/jaas:py3.5.3",
               destination_dir='gs://cancer_detection_demo/',
               runtime_config=CMLEConfig())

job.submit()
```

Simples modificações para treinar seu modelo on-premises com Kubeflow ou na Google Cloud using AI Platform

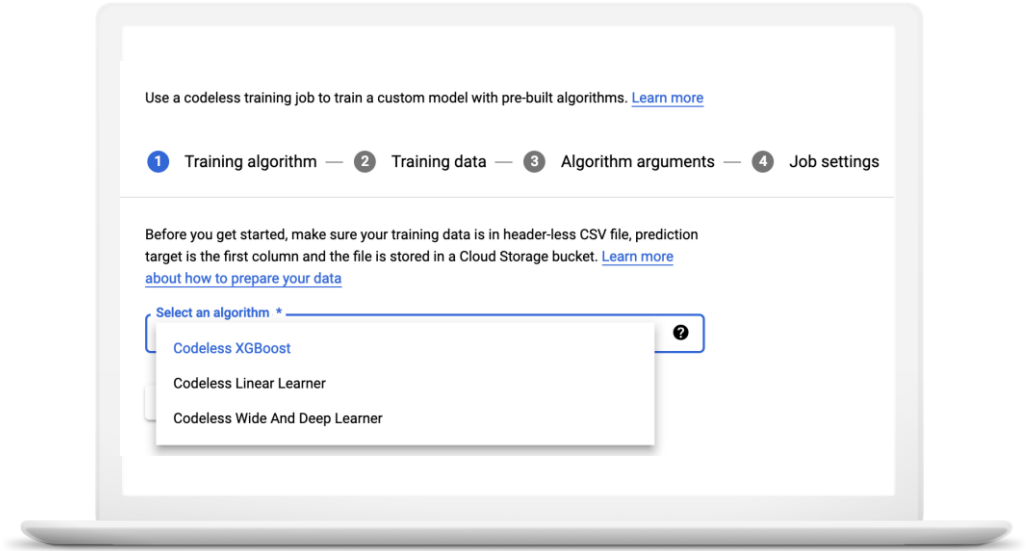
Implemente seu modelo com facilidade

```
%%bash -s "$SAVED_MODEL_DIR"
sm_dir=$1
model_name=kg_cancer_detection
version_name=versions


# Create the model and base version if it doesn't exist
gcloud ml-engine models describe $model_name > /dev/null
if [ $? -ne 0 ]; then gcloud ml-engine models create $model_name; fi
gcloud ml-engine versions describe $version_name --model $model_name 2> /dev/null
if [ $? -ne 0 ]; then
    gcloud ml-engine versions create $version_name \
      --model $model_name \
      --origin=$sm_dir \
fi
```

Modelos Built-in no Cloud AI Platform

- Otimização de hiperparâmetros embutida
- Suporte à algorítimos populares
- Busque soluções no AI Hub
- Facilidade para incluir novos algorítimos como containers



Três caminhos de AI na Google Cloud

seus dados + seu modelo


Cloud TPUs


Compute Engine


Cloud Dataproc

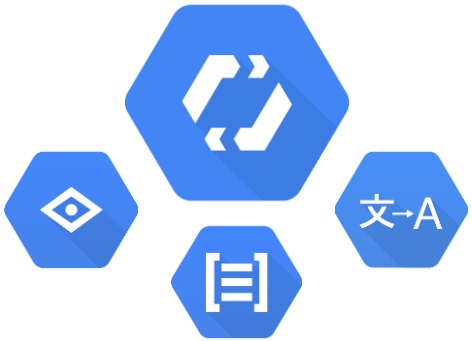

Kubernetes Engine


Cloud ML Engine


BigQuery ML

seus dados + nossos modelos

AutoML



nossos dados + nossos modelos


Cloud Translation API


Cloud Vision API


Cloud Speech API


Cloud Video Intelligence API


Data Loss Prevention API


Cloud Speech Synthesis API


Cloud Natural Language API


Dialogflow

Customização

Facilidade de Uso

Construa seus modelos

Treine os nossos modelos

Use nossas APIs inteligentes

Inteligência de Dados

- ORGANIZAÇÃO DATA DRIVEN
- VISUALIZAÇÃO DE DADOS
- MODELOS PREDITIVOS PARA A TOMADA DE DECISÃO
- DATA STORYTELLING E DATA ART
- CIÊNCIA DE DADOS APLICADA AOS NEGÓCIOS
- INTELIGÊNCIA ARTIFICIAL
- JORNADA ANALÍTICA

BigQuery ML Overview

Google Cloud

BigQuery ML

BigQuery ML empowers both **data analysts** and **data scientists**

Execute ML initiatives without moving data from BigQuery

Iterate on models in SQL in BigQuery to increase development speed

Automate model selection, and hypertuning

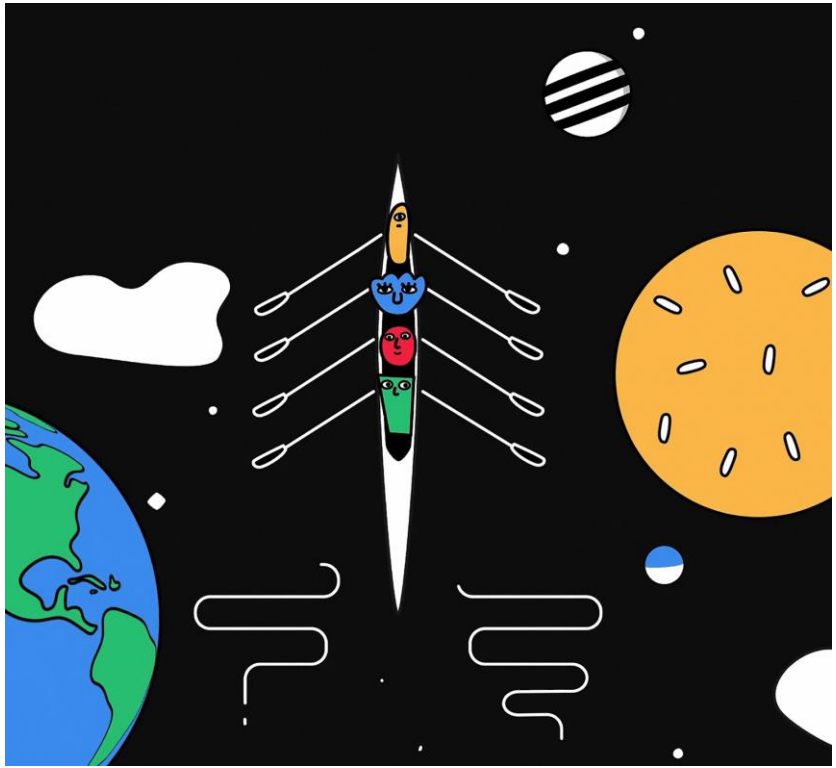
BigQuery ML

1. **Execute** ML initiatives without moving data from BigQuery
2. **Iterate** on models in SQL in BigQuery to increase development speed
3. **Automate** common ML tasks and hyperparameter tuning

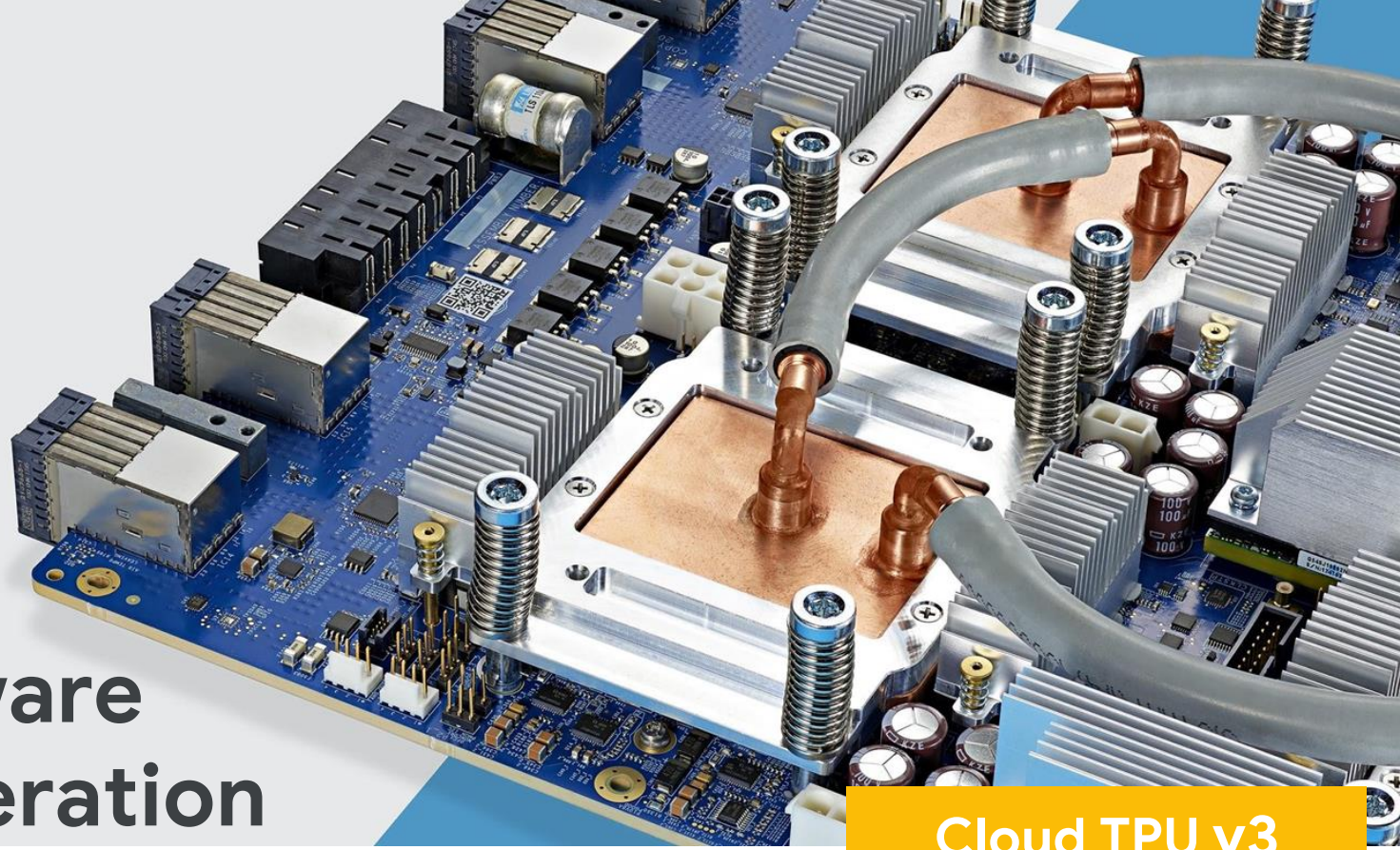
Behind the scenes

Through two lines of SQL

- Leverage BigQuery's processing power to build a model
- Auto-tuned learning rate
- Auto-split of data into training and test
- Null imputation
- Standardization of numeric features
- One-hot encoding of strings
- Class imbalance handling



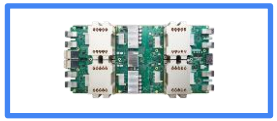
Hardware Acceleration



Cloud TPU v3

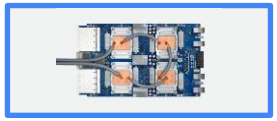
Accelerator families comparison

TPU v1
int8 only - inference only required model quantization



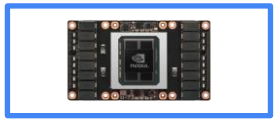
TPU v2
bfloat16 - inference and training no model changes necessary

180 TFLOPS (bfloat16)
/ board



TPU v3
bfloat16 - inference and training no model changes necessary

420 TFLOPS
(bfloat16)
/ board



NVIDIA P100
3584 CUDA cores - float16
but float16 x float16 => float32 not available

18 TFLOPS (float16)
/ chip

NVIDIA V100
5120 CUDA cores + 640 TensorCores - float16
model tweaking necessary for float16

112 TFLOPS (float16)
/ chip

Google Cloud

Reference models for Cloud TPUs

Image Recognition
Object Detection

Image Recognition
AmoebaNet-D
ResNet-50/101/152/200
Inception v2/v3/v4
DenseNet

Object Detection
RetinaNet

Low-Resource Models
MobileNet
SqueezeNet

Machine Translation
& Language Modeling

Models
Machine translation
Language modeling
Sentiment analysis
Question-answering
(all Transformer-based)

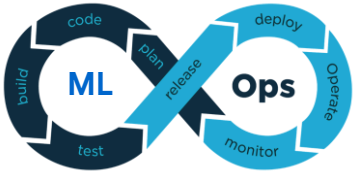
Speech
Recognition

Models
ASR Transformer
(LibriSpeech)

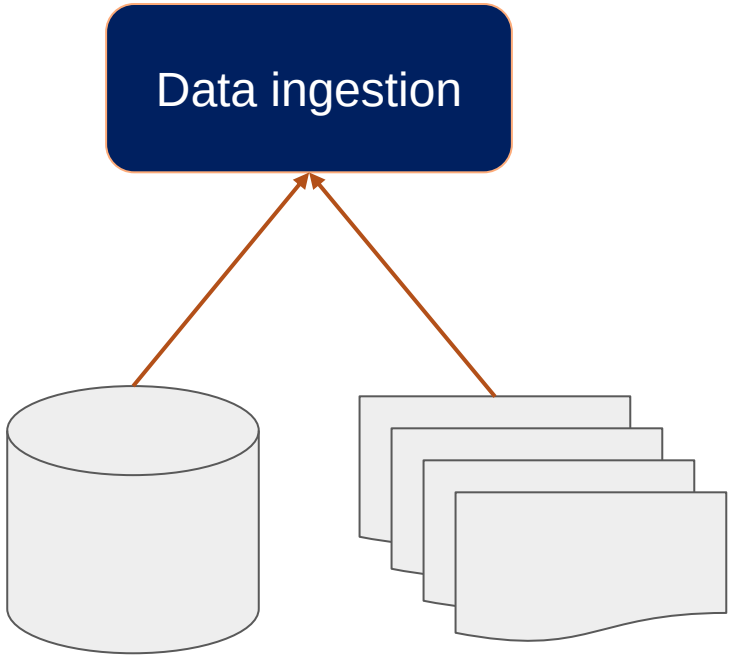
Image
Generation

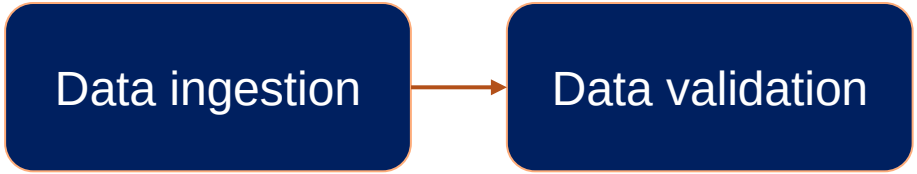
Models
Image Transformer
DCGAN

github.com/tensorflow/tpu



MLOps (ML Operationalization)


Data ingestion

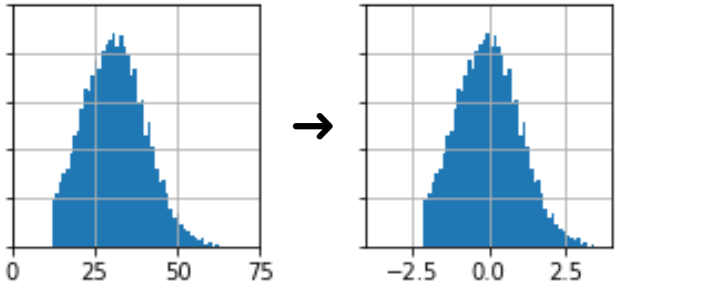


`age` is missing

`country` not in:

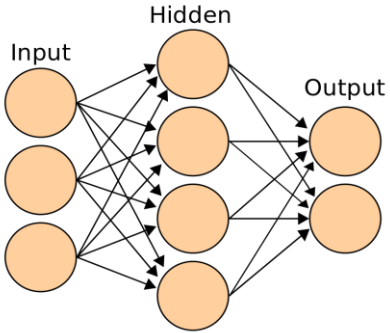
- China
- India
- USA

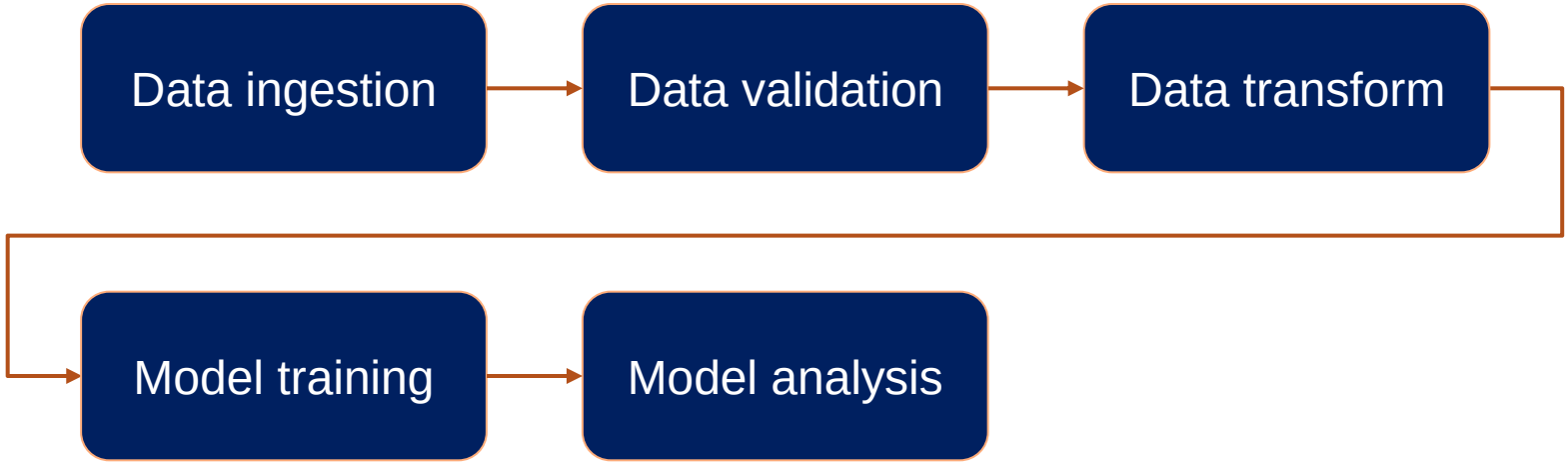
Data ingestion → Data validation → Data transform

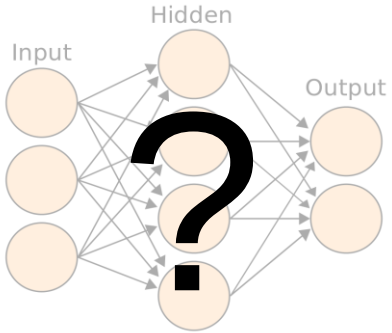


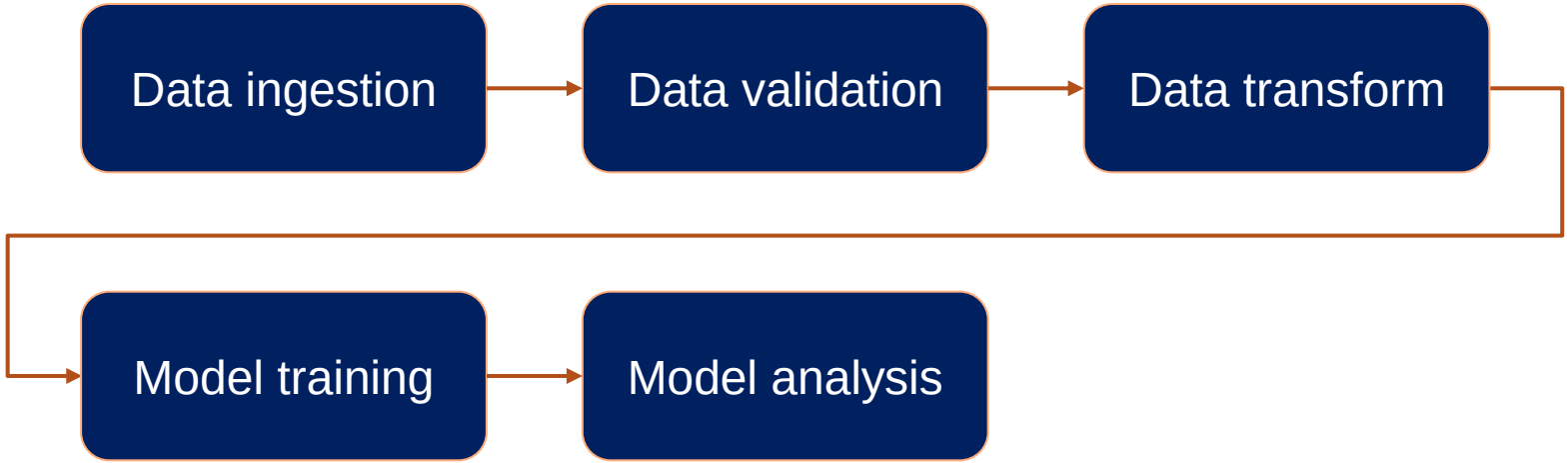
China → [1, 0, 0]
India → [0, 1, 0]
USA → [0, 0, 1]

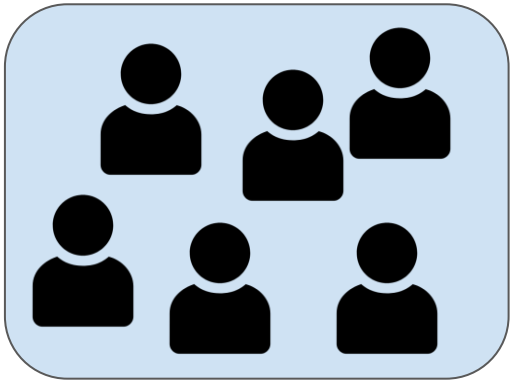
Data ingestion → Data validation → Data transform → Model training

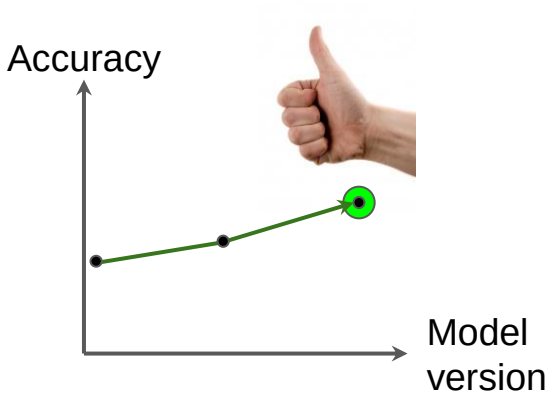


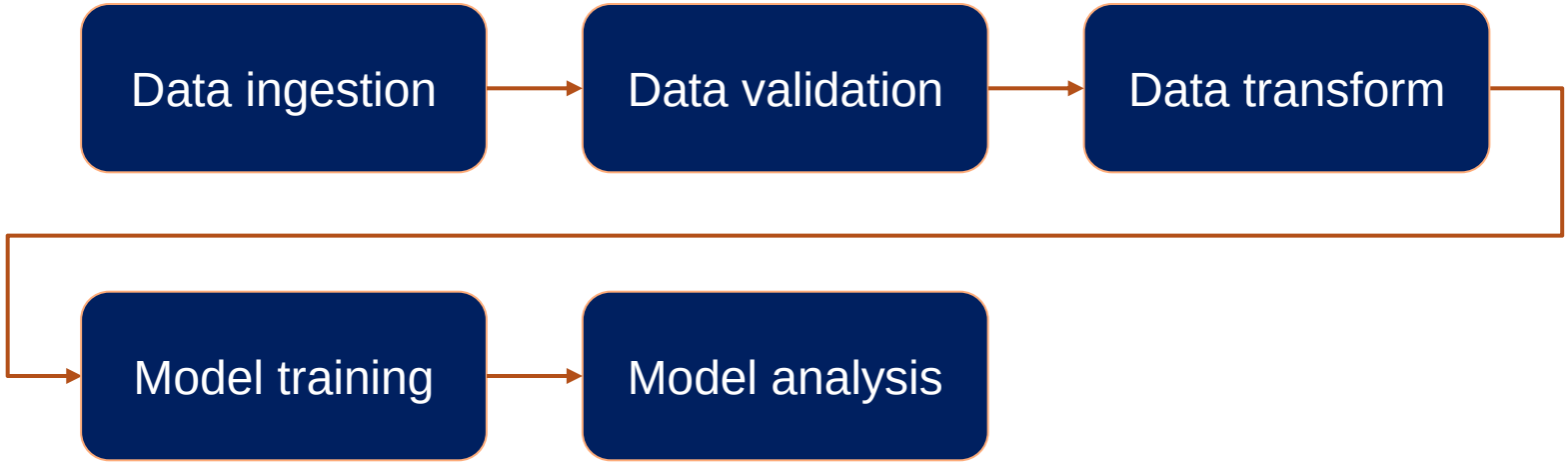

Data ingestion
Data validation
Data transform
Model training
Model analysis

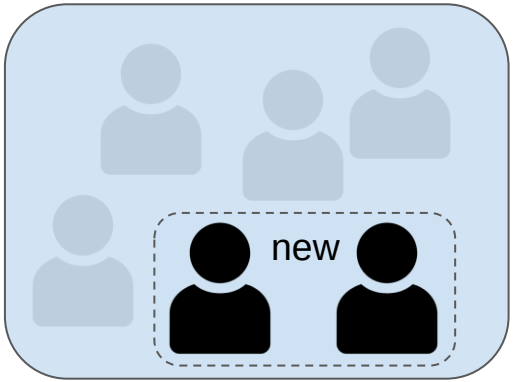

Hidden
Input
Output
?

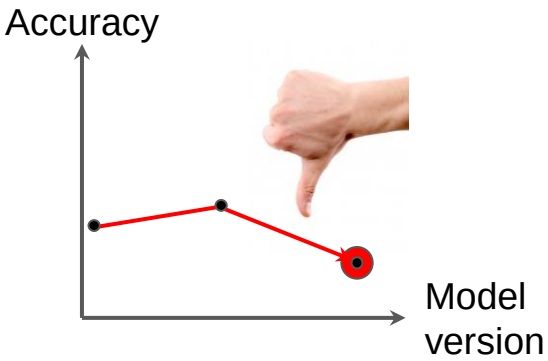

Data ingestion
Data validation
Data transform
Model training
Model analysis

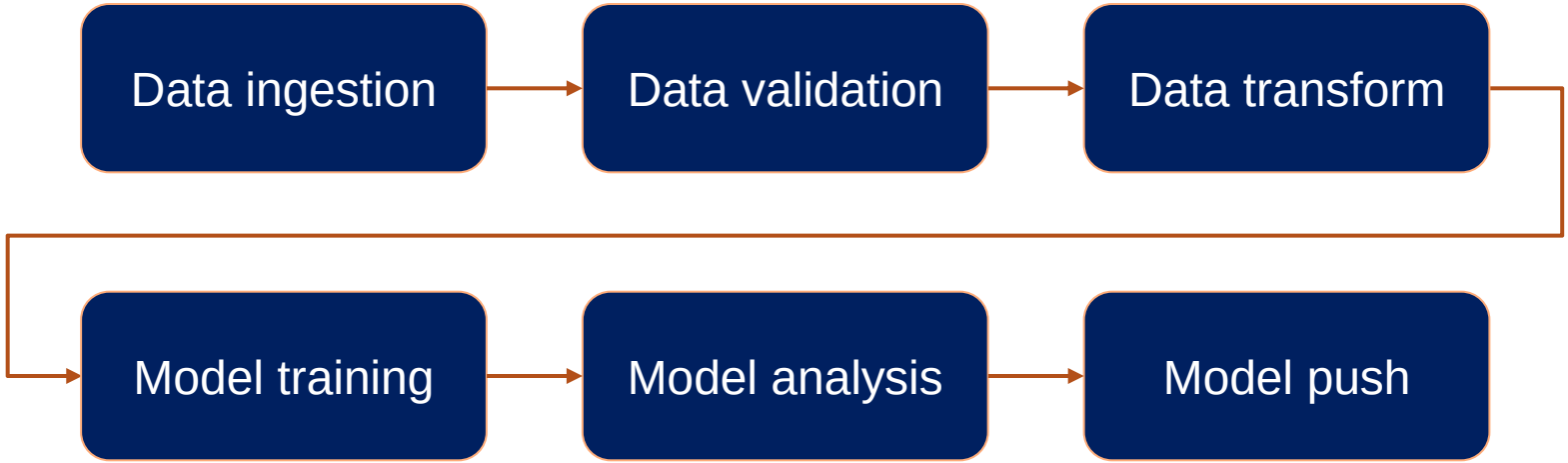


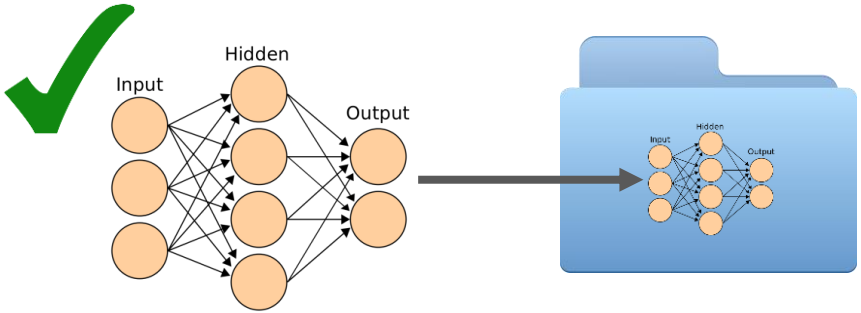

Accuracy
Model version

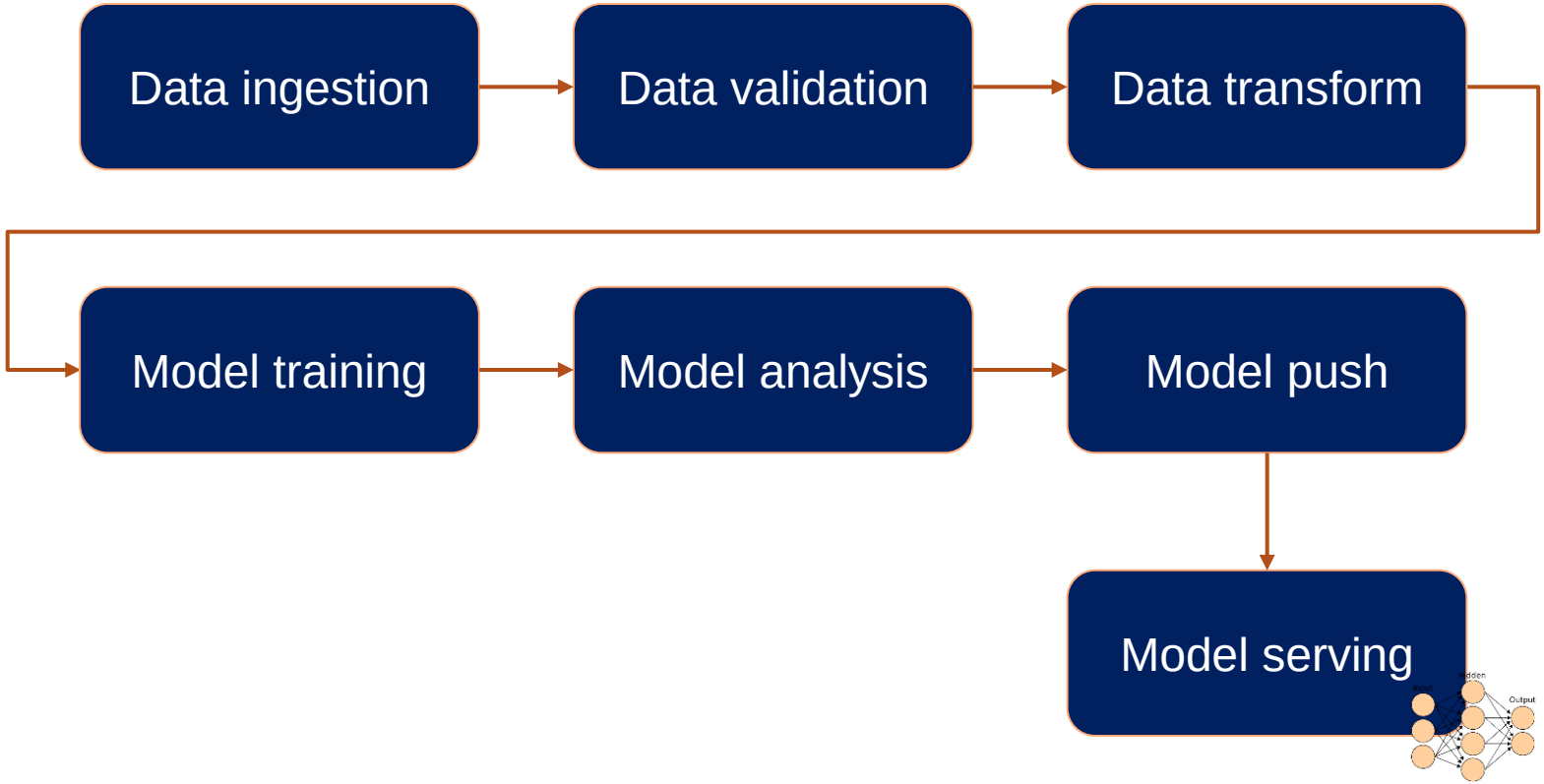

Data ingestion
Data validation
Data transform
Model training
Model analysis

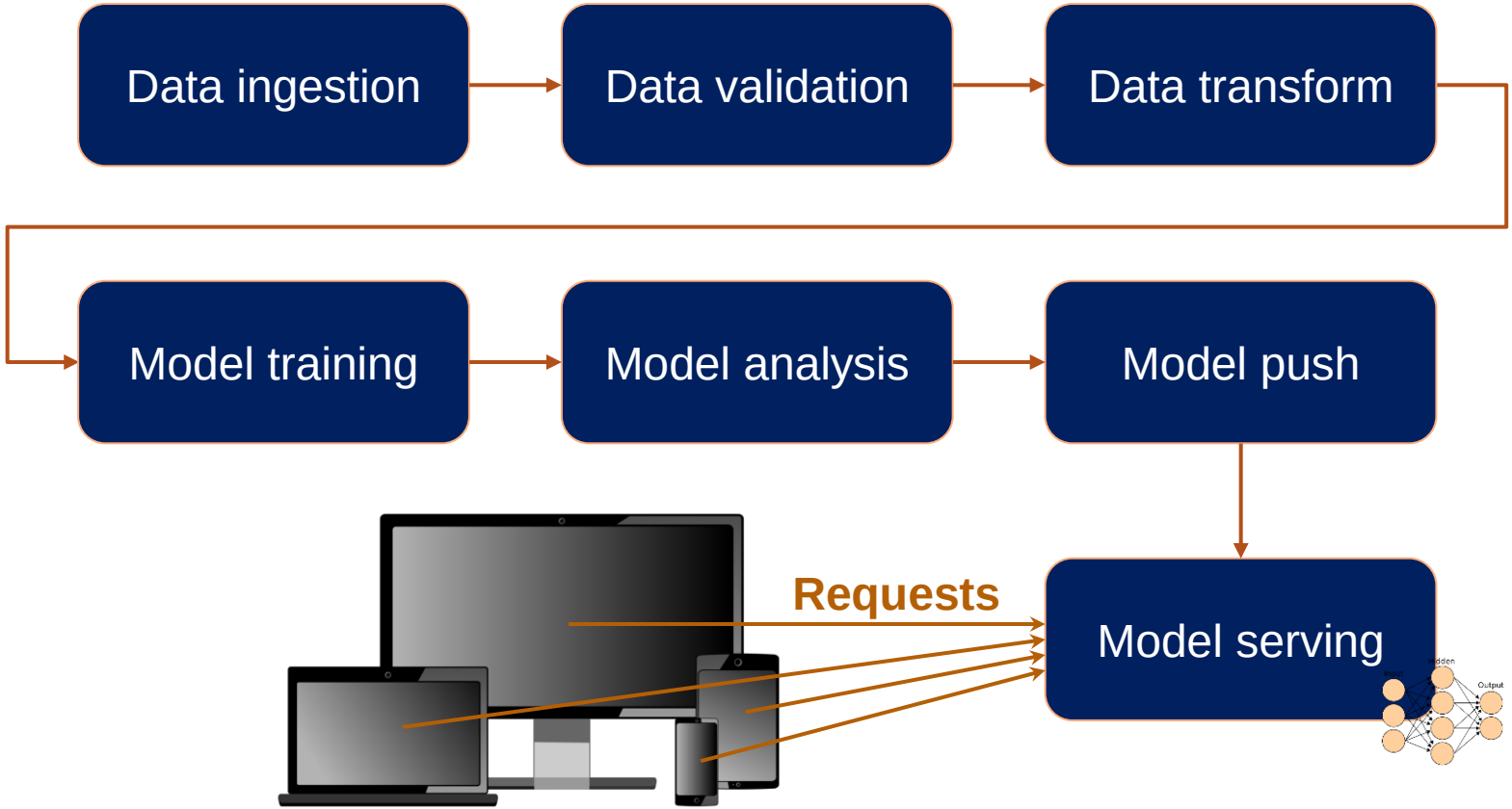

new

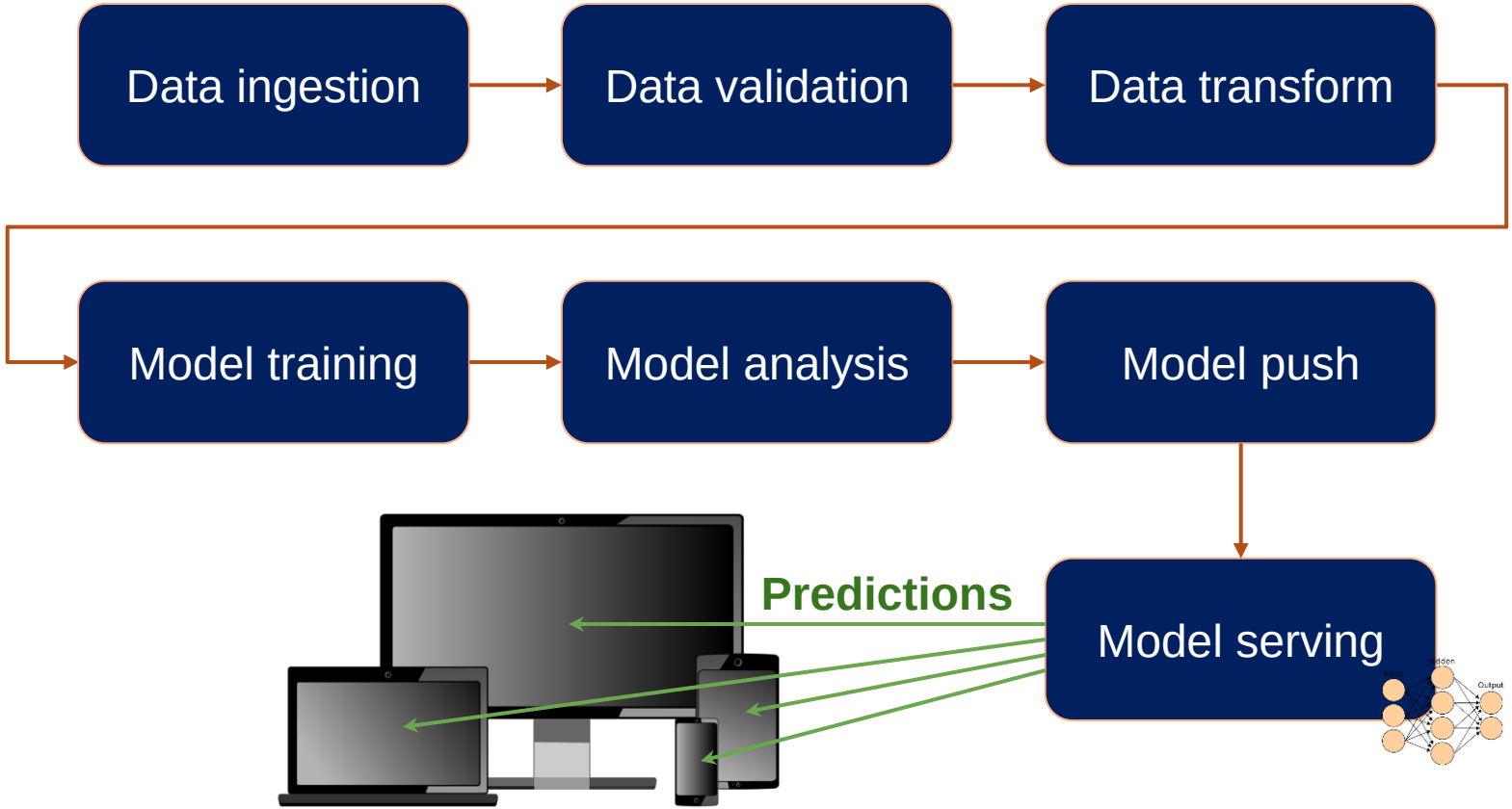

Accuracy
Model version

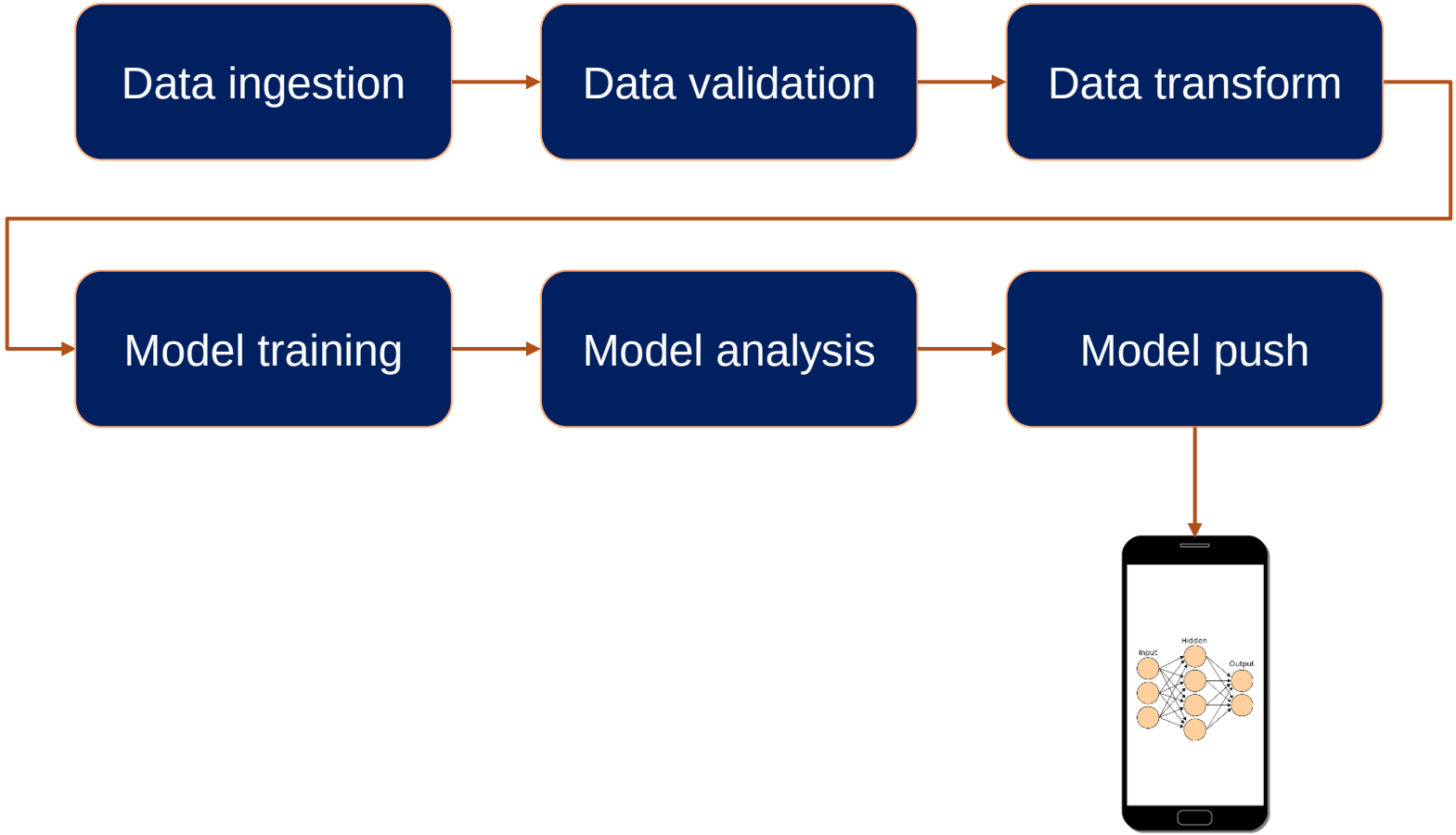

Data ingestion
Data validation
Data transform
Model training
Model analysis
Model push


Input
Hidden
Output
Input
Hidden
Output


Data ingestion
Data validation
Data transform
Model training
Model analysis
Model push
Model serving
Hidden
Output

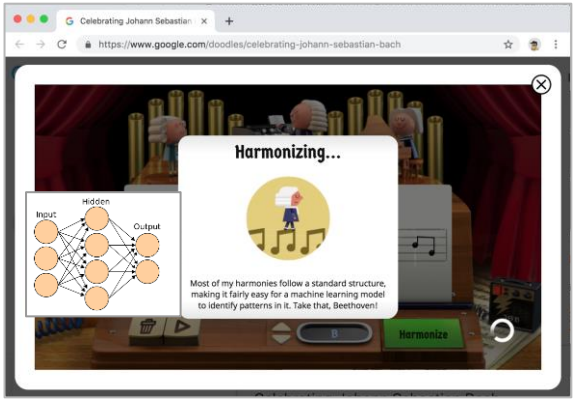

Data ingestion
Data validation
Data transform
Model training
Model analysis
Model push
Requests
Model serving
Hidden
Output

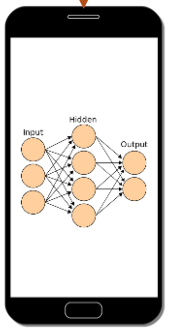

Data ingestion
Data validation
Data transform
Model training
Model analysis
Model push
Predictions
Model serving
Hidden
Output


Data ingestion
Data validation
Data transform
Model training
Model analysis
Model push
Input
Hidden
Output

VINICARIDÁ
Human-Centered Data & AI

Data ingestion → Data validation → Data transform → Model training → Model analysis → Model push



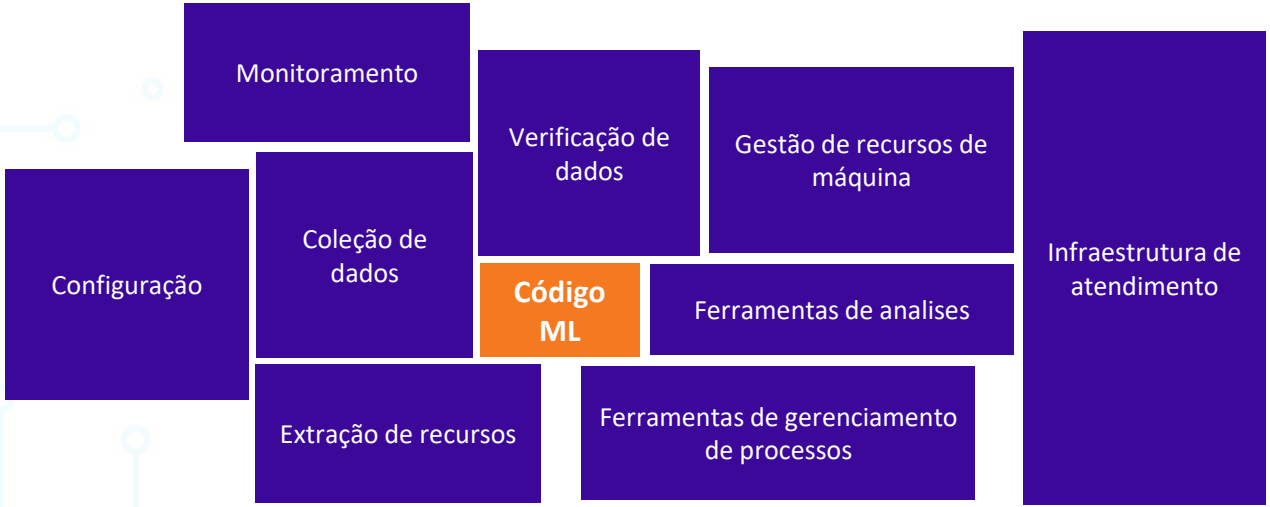


Machine Learning | Produção

Além de treinar um modelo incrível...

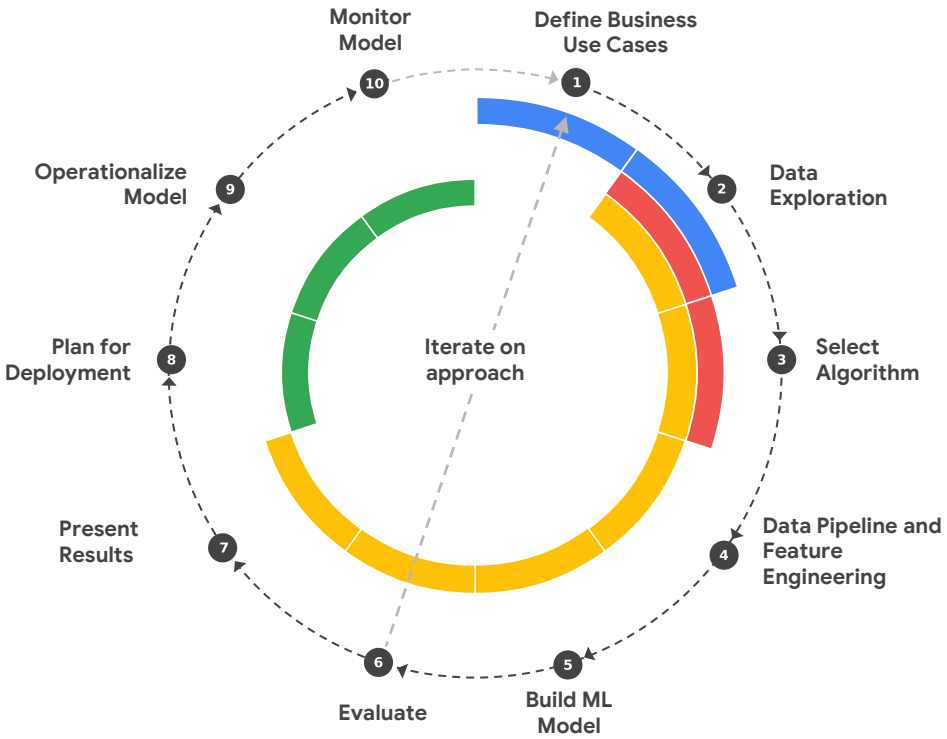
Machine Learning | Produção

Realidade: ML requer DevOps

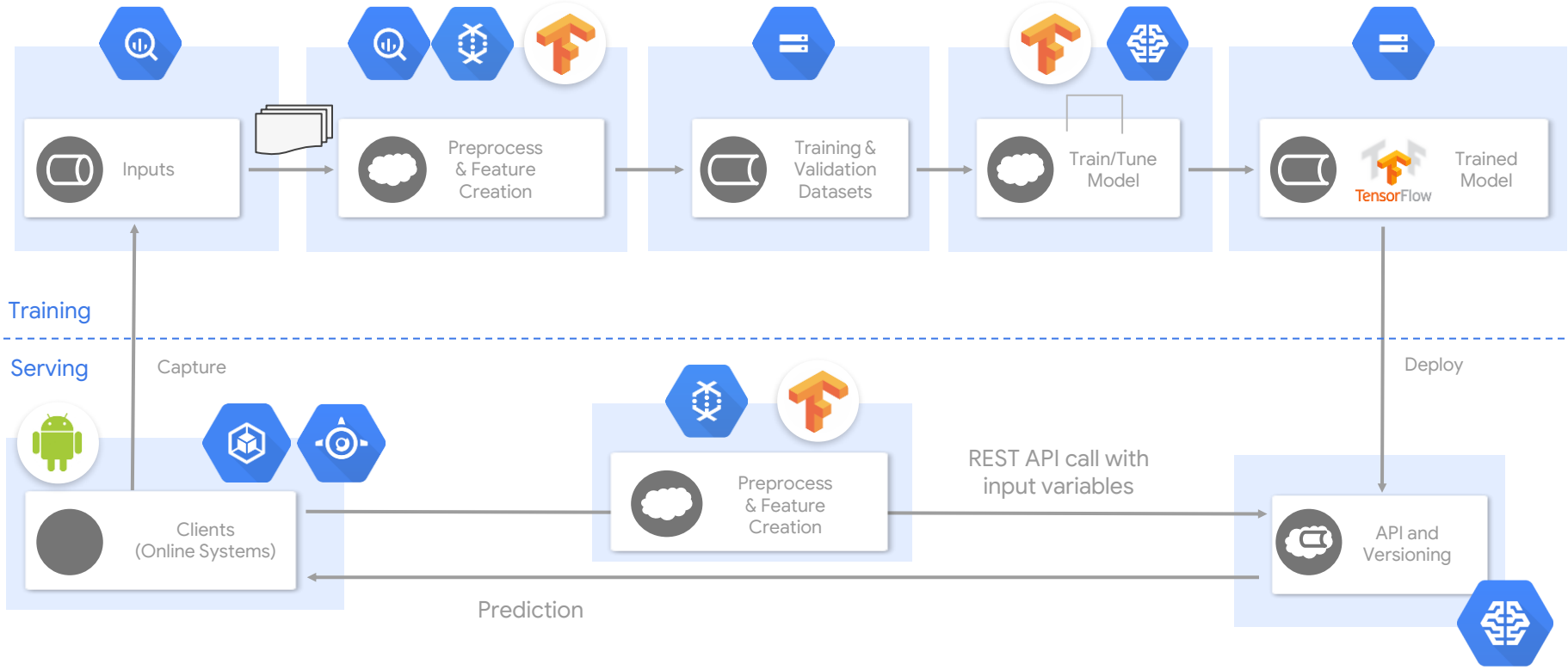


Typical machine learning lifecycle

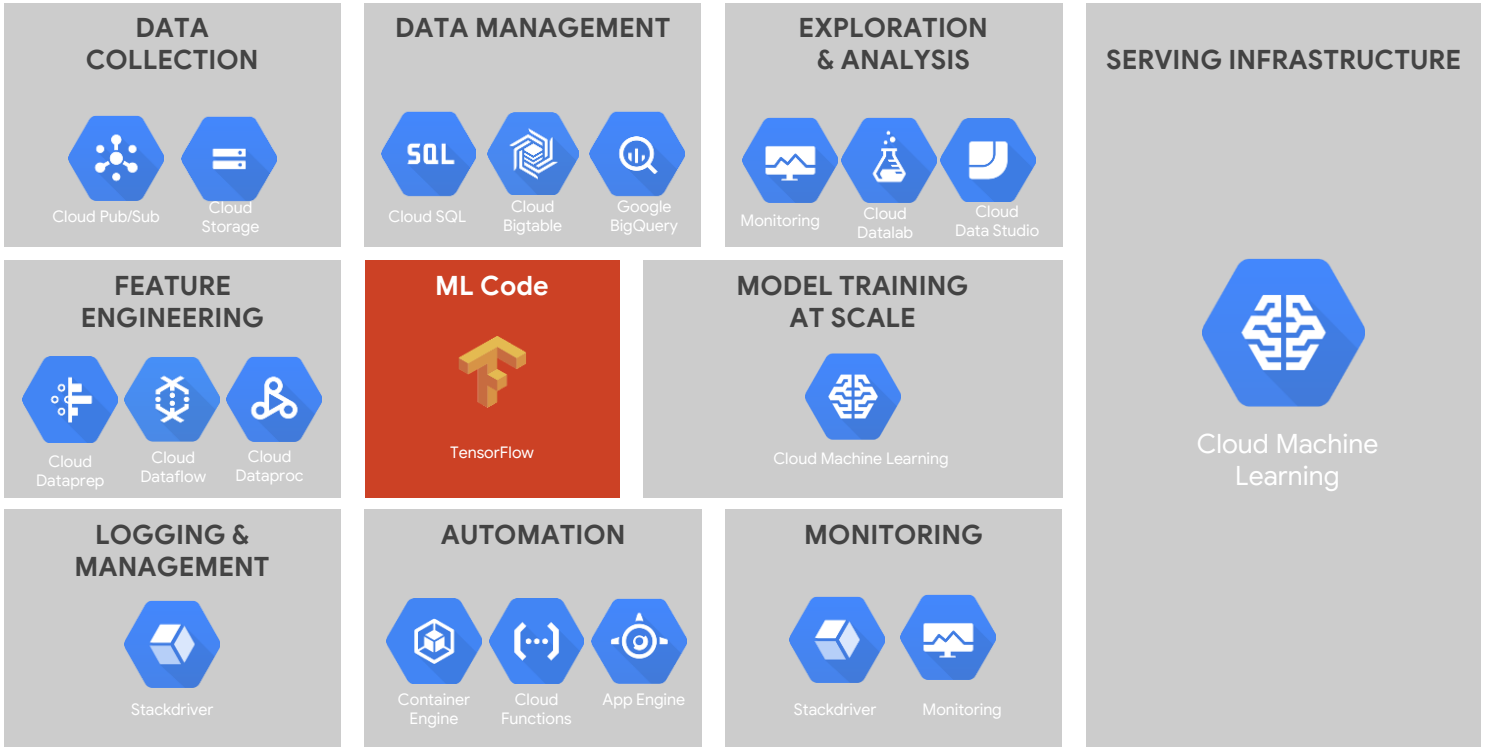
Step-by-step solution of ML problem



Machine learning pipeline @ GCP



Operational ML - end-to-end ML solution on GCP



Implementação de Exemplo

VINICARIDÁ
Human-Centered Data & AI

Thank you!

@vinicius caridá

@vfcarida

vfcarida@gmail.com

https://linktr.ee/vfcarida