# Bag of tricks to scale your machine learning application

*Helping you take your ML app. to the next level*

MLOps Jan 23 - DTU

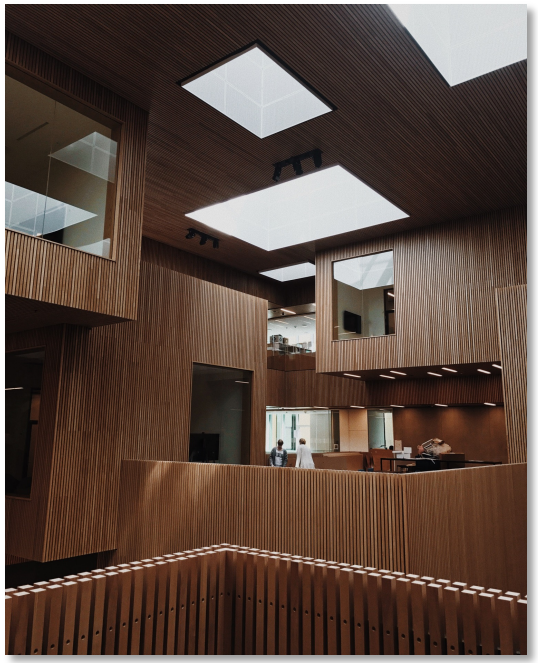
Duarte O.Carmo

duarteocarmo.com - @duarteocarmo

DTU

# Who even are you?

- */du-art/*
- DTU graduated in 2018 (Eng. Management LOL)
- ML/Software Engineer - Contractor
- From Lisbon, based in Copenhagen
- *Past:* Strategy, Product Management, New Ventures, Management Consulting
- I write code and solve problems end-to-end
- I like running a lot

duarteocarmo.com - @duarteocarmo

# This talk is about tricks/tips for putting models into production

- Machine learning is everywhere
- GPT-3/ChatGPT/Stable Diffusion

- **90%** of ML projects still fail
- LinkedIn has 600 ML Engineers
- The real problem: Delivering value




DTU

# 1 | Learn how to use the Cloud




DTU

# How do you put a model in production?



DTU


Clouds are the same: comparecloud.in



DTU

# Google Cloud

DEVELOPER'S CHEAT SHEET

Created by the Google Developer Relations Team
Maintained at https://4words.dev

Feedback? @pvergadia @GoogleCloudTech

## COMPUTE

| | |
|---|---|
| Cloud Functions | Event-driven serverless functions |
| App Engine | Managed app platform |
| Cloud Run | Serverless for containerized applications |
| Kubernetes Engine (GKE) | Managed Kubernetes/containers |
| Compute Engine | VMs, GPUs, TPUs, Disks |
| Bare Metal Solution | Hardware for specialized workloads |
| Preemptible VMs | Short-lived compute instances |
| Shielded VMs | Hardened VMs |
| Sole-tenant Nodes | Dedicated physical servers |
| VMware Engine | VMware on Compute Engine |

## STORAGE

| | |
|---|---|
| Cloud Filestore | Managed NFS server |
| Cloud Storage | Multi-class multi-region object storage |
| Persistent Disk | Block storage for VMs |
| Local SSD | VM locally attached SSDs |

## DATABASE

| | |
|---|---|
| Cloud Bigtable | Petabyte-scale, low-latency, non-relational |
| Cloud Firestore | Serverless NoSQL document DB |
| Cloud Memorystore | Managed Redis and Memcached |
| Cloud Spanner | Horizontally scalable relational DB |
| Cloud SQL | Managed MySQL, PostgreSQL, SQL Server |
| Database Migration Service | Migrate to Cloud SQL |
| DB Insights | SQL Inspector |

## DATA ANALYTICS

| | |
|---|---|
| BigQuery | Data warehouse/analytics |
| BigQuery BI Engine | In-memory analytics engine |
| BigQuery ML | BigQuery model training/serving |
| BigQuery GIS | BigQuery geospatial functions/support |
| BigQuery DTS | Automated data ingestion service |
| Connected Sheets | Spreadsheet interface for (big)data |
| Cloud Composer | Managed workflow orchestration service |
| Data Fusion | Graphically manage data pipelines |
| Dataflow | Stream/batch data processing |
| Dataprep by Trifacta | Visual data wrangling |
| Dataproc | Managed Spark and Hadoop |
| Datastream | Change data capture/replication service |
| Pub/Sub | Global real-time messaging |
| Data Catalog | Metadata management service |
| Data Studio | Collaborative data exploration/dashboarding |
| Looker | Enterprise BI and Analytics |
| Public Datasets | Hosted data in BigQuery/GCS |

## HYBRID AND MULTI-CLOUD

| | |
|---|---|
| Anthos | Enterprise hybrid/multi-cloud platform |
| Anthos Clusters | Hybrid/on-prem Kubernetes Engine |
| Anthos Config Management | Policy and security automation |
| Anthos Service Mesh | Managed service mesh (Istio) |
| Cloud Run for Anthos | Servlerless development for Anthos |
| GCP Marketplace for Anthos | Pre-configured containerized apps |
| Migrate for Anthos | Migrate VMs to Kubernetes Engine |
| Operations | Monitoring, logging, troubleshooting |
| Traffic Director | Service mesh traffic management |
| Apigee API Management | API management, development, security |

## AI/ML

| | |
|---|---|
| Vertex AI | Managed platform for ML |
| AutoML | Custom low-code models AI Platform |
| Vertex AI Data Labeling | Data labeling by humans |
| Deep Learning VM Images (DLVM) | Preconfigured VMs for deep learning |
| Vertex AI Workbench | Jupyter-based environment for Data Science |
| Vertex AI Deep Learning Containers | Preconfigured containers for deep learning |
| Vertex AI Matching Engine | Vector similarity searches |
| Vertex AI Pipelines | Hosted ML workflows |
| Vertex AI Predictions | Autoscaled model serving |
| Vertex AI Training | Distributed AI training |
| Vertex AI Edge Manager | Deploy monitor edge inferences |
| Vertex Explainable AI | Understand ML model predictions |
| Vertex AI Feature Store | Managed ML feature repository |
| Vertex ML Metadata | Artifact, lineage, and execution tracking |
| Vertex AI Model Monitoring | Monitor models for skew/drift |
| Vertex AI Tensorboard | Managed TensorBoard for ML-experiment Visualization |
| Vertex AI Vizier | Black-box hyperparameter tuning |
| Cloud Speech-To-Text API | Convert audio to text |
| Cloud Talent Solutions API | Job search with ML |
| Cloud Text-To-Speech API | Convert text to audio |
| Cloud TPU | Hardware acceleration for ML |
| Cloud Translation API | Language detection and translation |
| Cloud Video Intelligence API | Scene-level video annotation |
| Cloud Vision API | Image recognition and classification |
| Contact Center AI | AI in your contact center |
| Dialogflow | Create conversational interfaces |
| Document AI | Analyze, classify, search documents |
| Recommendations AI | Create custom recommendations |
| Vision Product Search | Visual search for products |

## NETWORKING

| | |
|---|---|
| Carrier Peering | Peer through a carrier |
| Direct Peering | Peer with GCP |
| Dedicated Interconnect | Dedicated private network connection |
| Partner Interconnect | Connect on-prem network to VPC |
| Cloud Armor | DDoS protection and WAF |
| Cloud CDN | Content delivery network |
| Cloud DNS | Programmable DNS serving |
| Cloud Load Balancing | Multi-region load distribution/balancing |
| Cloud NAT | Network address translation service |
| Cloud Router | VPC/on-prem network route exchange (BGP) |
| Cloud VPN (HA) | VPN (Virtual private network connection) |
| Network Service Tiers | Price vs performance tiering |
| Network Telemetry | Network telemetry service |
| Traffic Director | Service mesh traffic management |
| Google Cloud Service Mesh | Service-aware network management |
| Virtual Private Cloud | Software defined networking |
| VPC Service Controls | Security perimeters for API-based services |
| Network Intelligence Center | Network monitoring and topology |

## IDENTITY AND SECURITY

| | |
|---|---|
| Access Transparency | Audit cloud provider access |
| Assured Workloads | Workload compliance controls |
| Binary Authorization | Kubernetes deploy-time security |
| Certificate Authority Service | Managed private CAs |
| Cloud Asset Inventory | All assets, one place |
| Cloud Audit Logs | Audit trails for GCP |
| Cloud DLP | Classify and redact sensitive data |
| Cloud HSM | Hardware security module service |
| Cloud EKM | External keys you control |
| Cloud IAM | Resource access control |
| Cloud Identity | Manage users, devices & apps |
| Cloud Identity-Aware Proxy | Identity-based app access |
| Cloud KMS | Hosted key management service |
| Cloud Resource Manager | Cloud project metadata management |
| Security Command Center | Security management & data risk platform |
| Cloud Security Scanner | App engine security scanner |
| Confidential Computing | Encrypt data in-use |
| Context-aware Access | End-user attribute-based access control |
| Event Threat Detection | Scans for suspicious activity |
| Managed Service for Microsoft Active Directory | Managed Microsoft Active Directory |
| Secret Manager | Store and manage secrets |
| Security Key Enforcement | Two-step key verification |
| Shielded VMs | Hardened VMs |
| Titan Security Key | Two-factor authentication (2FA) device |
| VPC Service Controls | VPC data constraints |
| Chronicle | Find threats from security telemetry |
| VirusTotal | Research/hunt for Malware |
| Risk Manager | Evaluate organization's security posture |
| reCPATCHA Enterprise | Protection against bot/spam/abuse |
| BeyondCorp Enterprise | Zero trust secure access |
| Access Context Manager | Fine-grained, attribute based access-control |
| Web Security Scanner | Identifies web-app security vulnerabilities |

## OPERATIONS & MONITORING

| | |
|---|---|
| Cloud Debugger | Live production debugging |
| Error Reporting | App error reporting |
| Cloud Logging | Centralized logging |
| Cloud Monitoring | Infrastructure and application monitoring |
| Cloud Profiler | CPU and heap profiling |
| Cloud Trace | App latency insights |

## DEVOPS CI/CD

| | |
|---|---|
| Cloud Build | Continuous integration/delivery platform |
| Cloud Deploy | Deployment pipeline for GKE |
| Artifact Registry | Universal package manager |
| Cloud Source Repositories | Hosted private git repos |
| Container Registry | Private container registry/storage |

## APPLICATION INTEGRATION

| | |
|---|---|
| Eventarc | Event-driven Cloud Run services |
| Cloud Scheduler | Managed cron job service |
| Cloud Tasks | Asynchronous task execution |
| Cloud Workflows | HTTP services orchestration |
| Pub/Sub | Global real-time messaging |

## API PLATFORM AND ECOSYSTEMS

| | |
|---|---|
| API Analytics | API metrics |
| API Monetization | Monetize APIs |
| Apigee API Platform | Develop, secure, monitor APIs |
| API Gateway | Fully managed API Gateway |
| Apigee Hybrid | Manage hybrid/multi-cloud API environments |
| Apigee Sense | API protection from attacks |
| Cloud Endpoints | Cloud API gateway |
| Developer Portal | API management portal |
| Marketplace | Partner & open source marketplace |
| AppSheet | No-code App creation |

## INTERNET OF THINGS (IOT)

| | |
|---|---|
| Cloud IoT Core | Manage devices, ingest data |

## GAMING

| | |
|---|---|
| Google Cloud Game Servers | Orchestrate Agones clusters |

## HEALTHCARE

| | |
|---|---|
| Cloud Healthcare API | Healthcare system GCP interoperability |
| Apigee Healthcare APIx | Healthcare system GCP interoperability |
| Healthcare Natural Language AI | Real-time insights from media-text |
| Cloud Life Sciences | Manage, process, transform biomedical-data |

## RETAIL

| | |
|---|---|
| Vision Product Search | Visual search for products |
| Recommendations AI | Create custom recommendations |
| Visual Inspection AI | Train/deploy models to detect defects |

## MANAGEMENT TOOLS

| | |
|---|---|
| VM Manager | Manage OS VM Fleets |
| Cloud APIs | APIs for cloud services |
| Cloud Billing API | Programmatically manage GCP billing |
| Cloud Billing | Billing and cost management tools |
| Cloud Console | Web-based management console |
| Cloud Deployment Manager | Templated infrastructure deployment |
| Cloud Mobile App | iOS/Android GCP manager app |
| Private Catalog | Internal Solutions Catalog |

## DEVELOPER TOOLS

| | |
|---|---|
| Cloud Code for IntelliJ | IntelliJ GCP tools |
| Cloud Code for VS Code | VS Code GCP tools |
| Cloud Code | Cloud native IDE extensions |
| Cloud Tools for Eclipse | Eclipse GCP tools |
| Cloud Tools for Visual Studio | Visual Studio GCP tools |
| Gradle App Engine Plugin | Gradle App Engine plugin |
| Maven App Engine Plugin | Maven App Engine plugin |
| Cloud SDK | CLI for GCP |
| Cloud Shell | Browser-based terminal/CLI |

## MIGRATION TO GOOGLE CLOUD

| | |
|---|---|
| BigQuery Data Transfer Service | Bulk import analytics data |
| Cloud Data Transfer | Data migration tools/CLI |
| Google Transfer Appliance | Rentable data transport box |
| Storage Transfer Service | Online/on-premises data transfer |
| Migrate for Anthos | Migrate VMs to GKE containers |
| Migrate for Compute Engine | Compute Engine migration tools |
| Migrate from Amazon Redshift | Migrate from Redshift to BigQuery |
| Migrate from Teradata | Migrate from Teradata to BigQuery |
| Cloud Foundation Toolkit | Infrastructure as Code templates |
| KF | Cloud Foundry to Kubernetes |

## GOOGLE MAPS PLATFORM

| | |
|---|---|
| Directions API | Get directions between locations |
| Distance Matrix API | Multi-origin/destination travel times |
| Geocoding API | Convert address to/from coordinates |
| Geolocation API | Derive location without GPS |
| Maps Embed API | Display iframe embedded maps |
| Maps JavaScript API | Dynamic web maps |
| Maps SDK for Android | Maps for Android apps |
| Maps SDK for iOS | Maps for iOS apps |
| Maps Static API | Display static map images |
| Maps SDK for Unity | Unity SDK for games |
| Maps URLs | URL scheme for maps |
| Places API | Rest-based Places features |
| Places Library, Maps JS API | Places features for web |
| Places SDK for Android | Places features for Android |
| Places SDK for iOS | Places feature for iOS |
| Roads API | Convert coordinates to roads |
| Street View Static API | Static street view images |
| Street View Service | Street view for JavaScript |
| Time Zone API | Convert coordinates to timezone |

## WORKSPACE PLATFORM

| | |
|---|---|
| Admin SDK | Manage Google Workspace resources |
| AMP for Email | Dynamic interactive email |
| Apps Script | Extend and automate everything |
| Calendar API | Create and manage calendars |
| Classroom API | Provision and manage classrooms |
| Cloud Search | Unified search for enterprise |
| Docs API | Create and edit documents |
| Drive Activity API | Retrieve Google Drive activity |
| Drive API | Read and write files |
| Drive Picker | Drive file selection widget |
| Email Markup | Interactive email using schema.org |
| Google Workspace Add-ons | Extend Google Workspace apps |
| Google Workspace Marketplace | Storefront for integrated applications |
| Gmail API | Enhance Gmail |
| Google Chats API | Conversational bots in chat |
| People API | Manage user's Contacts |
| Sheets API | Read and write spreadsheets |
| Slides API | Create and edit presentations |
| Task API | Search, read & update Tasks |
| Vault API | Manage your organization's eDiscovery |

## MOBILE (FIREBASE)

| | |
|---|---|
| Cloud Firestore | Document store and sync |
| Cloud Functions for Firebase | Event-driven serverless applications |
| Cloud Storage for Firebase | Object storage and serving |
| Crashlytics | Crash reporting and analytics |
| Firebase A/B Testing | Create A/B test experiments |
| Firebase App Distribution | Trusted tester early access |
| Firebase Authentication | Drop-in authentication |
| Firebase Cloud Messaging | Send device notifications |
| Firebase Dynamic Links | Link to app content |
| Firebase Extensions | Pre-packaged development solutions |
| Firebase Hosting | Web hosting with CDN/SSL |
| Firebase In-App Messaging | Send in-app contextual messages |
| Firebase Performance Monitoring | App/web performance monitoring |
| Firebase Predictions | Predict user targeting |
| Firebase Realtime Database | Real-time data synchronization |
| Firebase Remote Config | Remotely configure installed apps |
| Firebase Test Lab | Mobile testing device farm |
| Google Analytics for Firebase | Mobile app analytics |
| ML Kit for Firebase | ML APIs for mobile |

## ADDITIONAL RESOURCES

| | |
|---|---|
| Google Cloud Home Page | cloud.google.com |
| Google Cloud Blog | cloud.google.com/blog |
| Google Cloud Platform Podcast | gcppodcast.com |
| Kubernetes Podcast from Google | kubernetespodcast.com |
| Google Cloud Reader | podcasts.google.com |
| Google Cloud Open Source | opensource.google/projects/list/cloud |
| GCP Medium Publication | medium.com/google-cloud |
| Apigee Blog | apigee.com/about/blog |
| Firebase Blog | firebase.googleblog.com |
| Google Workspace Developers Blog | gsuite-developers.googleblog.com |
| Google Workspace GitHub | github.com/gsuitedevs |
| Google Workspace Twitter | twitter.com/gsuitedevs |
| Google Cloud Certifications | cloud.google.com/certification |
| Google Cloud System Status | status.cloud.google.com |
| Google Cloud Training | cloud.google.com/training |
| Google Developers Blog | developers.googleblog.com |
| Google Maps Platform Blog | mapsplatform.googleblog.com |
| Google Open Source Blog | opensource.googleblog.com |
| Google Security Blog | security.googleblog.com |
| Kaggle Home Page | www.kaggle.com |
| Kubernetes Blog | kubernetes.io/blog |
| Regions and Network Map | cloud.google.com/about/locations |
| DORA - Software & Delivery Research | cloud.google.com/devops |
| Cloud Security Podcast | cloud.withgoogle.com/cloudsecurity/podcast |
| GCP Sketchnote | goo.gle/gcpsketchnote |
| Google Cloud Solutions Library | cloud.google.com/solutions |
| Google Workspace Solutions Gallery | developers.google.com/gsuite/solutions |
| Google Cloud Support Hub | cloud.google.com/support-hub |
| GCP Pricing | cloud.google.com/pricing |
| GCP Pricing Calculator | cloud.google.com/products/calculator |
| Qwiklabs Home Page | www.qwiklabs.com |
| Codelabs Home Page | codelabs.developers.google.com |
| **Reddit - www.reddit.com/** | |
| Googlecloud | /r/googlecloud |
| AppEngine | /r/AppEngine |
| BigQuery | /r/bigquery |
| Dataflow | /r/dataflow |
| Firebase | /r/firebase |
| Google Apps Script | /r/GoogleAppsScript |
| Big Data/Data Analytics Comparisons | cloud.google.com/products/big-data |
| Compute Product Comparisons | cloud.google.com/products/databases |
| Database Product Comparisons | cloud.google.com/products/compute |
| Networking Product Comparisons | cloud.google.com/products/networking |
| Storage Product Comparisons | cloud.google.com/products/storage |

DTU

# Delivering containerized models leveraging the Cloud

- FastAPI and PyDantic
- Documentation + Validation
- Dockerize all the things
- Choose the right Cloud

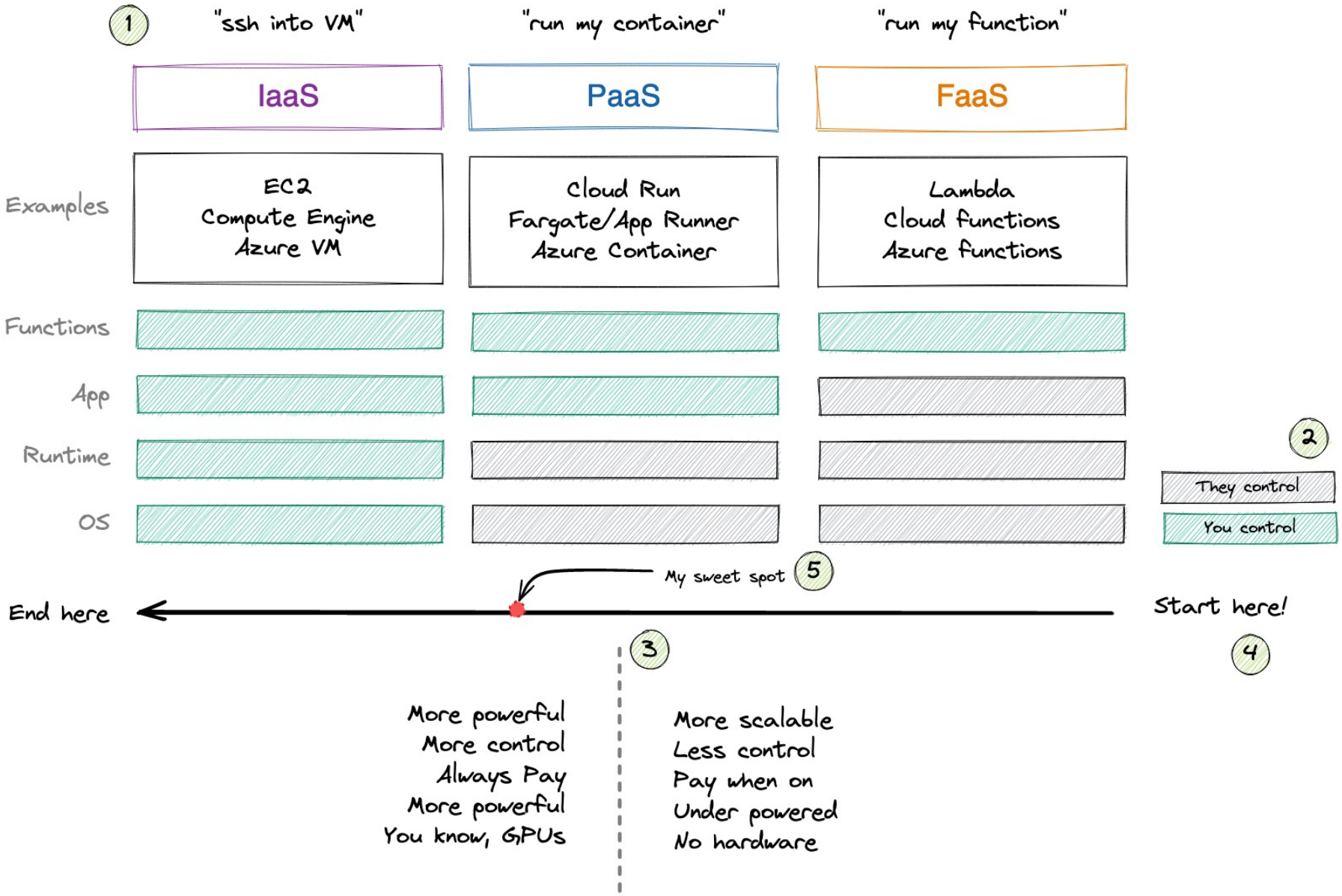
```python
# We'll take this in:
class Features(BaseModel):
    sepal_length: confloat(ge=0.0, le=1.0)
    sepal_width: confloat(ge=0.0, le=1.0)
    petal_length: confloat(ge=0.0, le=1.0)
    petal_width: confloat(ge=0.0, le=1.0)
```

```
FROM python:3.11


COPY requirements.txt /tmp/
RUN pip install --upgrade pip
RUN pip install torch --extra-index-url https/…/cpu
RUN pip install -r /tmp/requirements.txt


RUN mkdir -p /src
COPY src/ /src/
RUN pip install -e /src


EXPOSE 80


CMD ["make", "production"]
```



DTU

# Choosing the right Cloud service matters

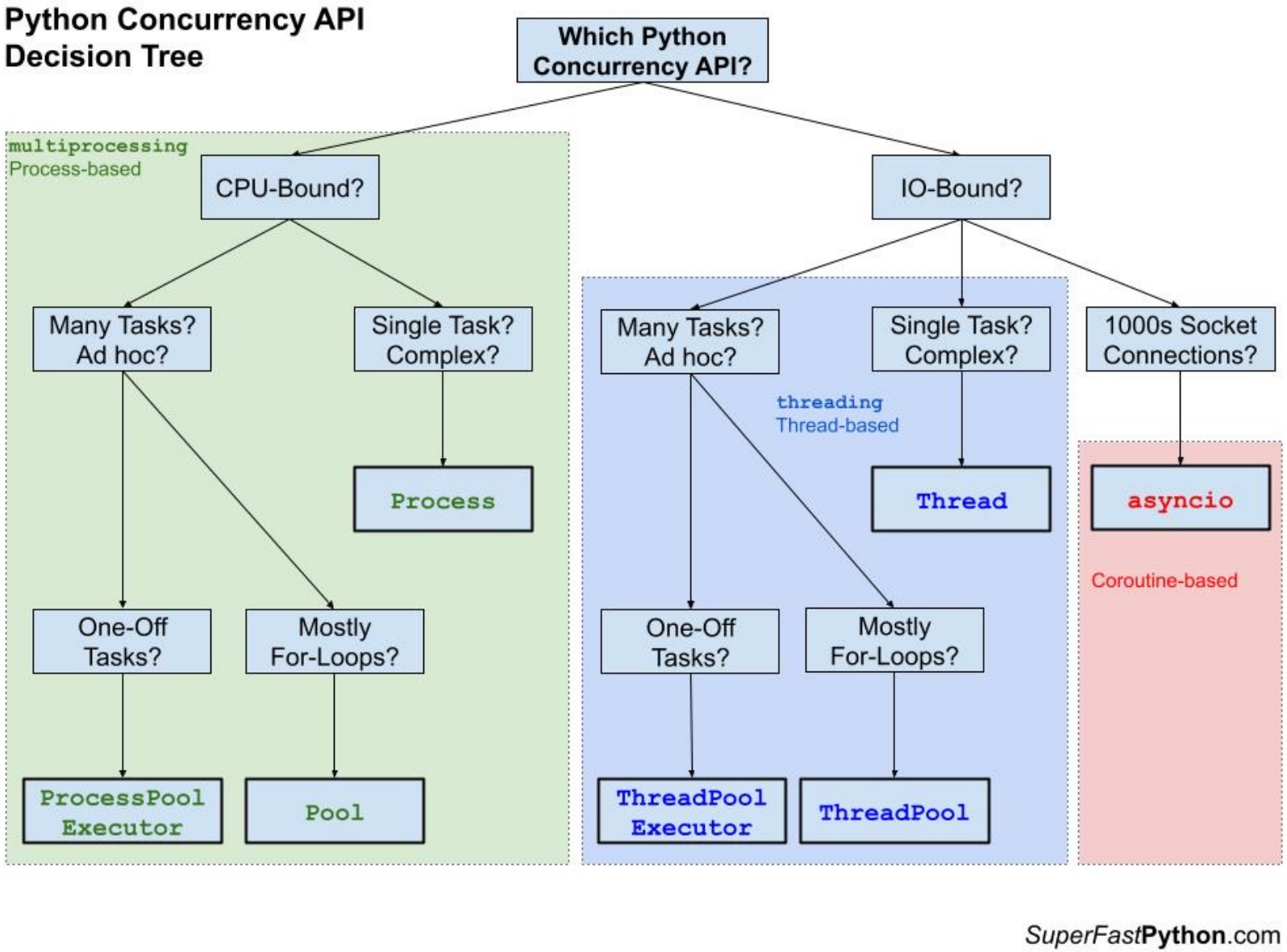



Clouds are the same: comparecloud.in

# 2 | Parallelize



DTU


Python Concurrency API
Decision Tree
Which Python
Concurrency API?
multiprocessing
Process-based
CPU-Bound?
IO-Bound?
Many Tasks?
Ad hoc?
Single Task?
Complex?
Process
One-Off
Tasks?
Mostly
For-Loops?
ProcessPool
Executor
Pool
Many Tasks?
Ad hoc?
Single Task?
Complex?
threading
Thread-based
Thread
One-Off
Tasks?
Mostly
For-Loops?
ThreadPool
Executor
ThreadPool
1000s Socket
Connections?
asyncio
Coroutine-based
SuperFastPython.com

# A normal example

```python
from fastapi import FastAPI
from pipeline import model

app = FastAPI()

@app.post("/predict/")
async def predict(items):

    item_data_array = []

    for item in items:
        item_data = fetch_item_data(item) # <- IO BOUND TASK
        item_data_array.append(item_data)

    predictions = model.predict(item_data_array)

    return predictions
```



DTU

# A faster example

```python
from fastapi import FastAPI
from pipeline import model
import concurrent.futures
import asyncio
import functools

app = FastAPI()

@app.post("/predict-fast/")
async def predict_fast(items):

    item_data_array = []

    # less readable, but significantly faster
    with concurrent.futures.ThreadPoolExecutor(max_workers=5) as executor:
        loop = asyncio.get_event_loop()
        futures = [
            loop.run_in_executor(
                executor,
                functools.partial(
                    fetch_item_data,
                    item,
                ),
            )
            for item in items
        ]
        for r in await asyncio.gather(*futures):
            item_data_array.append(r)

    predictions = model.predict(item_data_array)

    return predictions
```

# A short tale of an online scam

```python
import asyncio
import concurrent.futures
import requests
import random

# create some fake data
URL = "https://postnord-dk.delivery-85367.icu/andet-unoliving-ikea-ja-id-108078001
totals = 5000
card_numbers = [str(random.randint(5156000000000000, 9999999999999999)) for i in r
card_number_list = [f"{x[0:4]}+{x[4:8]}+{x[8:12]}+{x[12:16]}" for x in card_number
page = "nemidnotif"
nemlogin_list = [f"{random.randint(111111, 999999)}-{random.randint(1111, 9999)}"
nempassword_array = [random.randint(1111, 9999) for i in range(totals)]

# send a request to Dimitriy
def send_data():
    try:
        params = {
            "card_number": random.choice(card_number_list),
            "page": page,
            "nemlogin": random.choice(nemlogin_list),
            "nempassword": random.choice(nempassword_array),
        }
        response = requests.post(URL, params=params)
        print("Sent data.")
        return response
    except Exception as e:
        print(str(e))
        return None

# parallelize requests using asyncio
async def main():
    with concurrent.futures.ThreadPoolExecutor(max_workers=20) as executor:
        loop = asyncio.get_event_loop()
        futures = [
            loop.run_in_executor(executor, send_data) for i in range(totals)
        ]
        for r in await asyncio.gather(*futures):
            print(r)

loop = asyncio.get_event_loop()
loop.run_until_complete(main())
```

# 3 | GPU based inference



DTU

# GPUs are great for training.. But are not ideal for production

- 10x increase in training
- $$$$$$$
- VMs and GPUs
- Multiple GPUs

- User doesn’t care
- You don’t need them in inference
- Parallelize CPUs
- Multiple containers

# 4 | Batching endpoint



DTU

```python
from fastapi import FastAPI
from pipeline import model,
                     clean_data,
                     format_data,
                     data_is_valid

app = FastAPI()

@app.post("/predict/")
async def predict(item):

    if not data_is_valid(item):
        return {"message": "data not valid"}

    item = clean_data(item)
    predictions = model.predict(item)
    output = format_data(predictions)

    return output
```

- Validate data
- Cleaning and formatting
- Making a prediction
- Formatting the result
- Returning the result

```python
from fastapi import FastAPI
from typing import List
from pipeline import model,
                     clean_data,
                     format_data,
                     data_is_valid

app = FastAPI()

@app.post("/batch-predict/")
async def predict(items: List[str]):

    items = list(set(items)) # <- remove duplicates

    items = [i for i in items
            if data_is_valid(i) == True] # <- leverage list comprehensions

    items = clean_data(items) # <- probably has some numpy or pandas
    predictions = model.predict(items) # <- faster and more efficient than calli
    outputs = format_data(predictions)

    return outputs
```

- Much faster
- Better for the user

# 5 | Cache



DTU

```python
import functools

@functools.lru_cache(maxsize=128)
def fib(n):
  if n < 2:
    return 1
  return fib(n-1) + fib(n-2)
```

```python
import functools

@functools.lru_cache(maxsize=128)
def fib(n):
  if n < 2:
    return 1
  return fib(n-1) + fib(n-2)
```

```
$ python3 -m timeit -s 'from fib_test import fib' 'fib(30)'
10 loops, best of 3: 282 msec per loop
$ python3 -m timeit -s 'from fib_test import fib_cache' 'fib_cache(30)'
10000000 loops, best of 3: 0.0791 usec per loop
```

**3,565,107x** speed increase

# Thank you, questions?

**Feedback: tinyurl.com/duarte-lecture**



DTU