Produccionamiento de Sistemas de ML

Pedro Ferrari

Juan Martín Pampliega

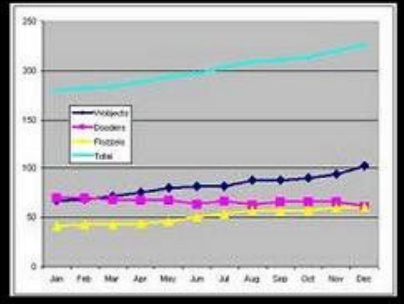
01 | Problemática

Data Scientist

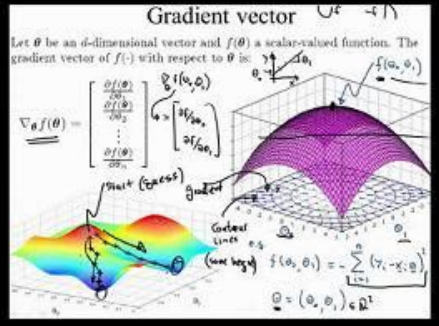
What my friends think I do

What my mom thinks I do

What society thinks I do

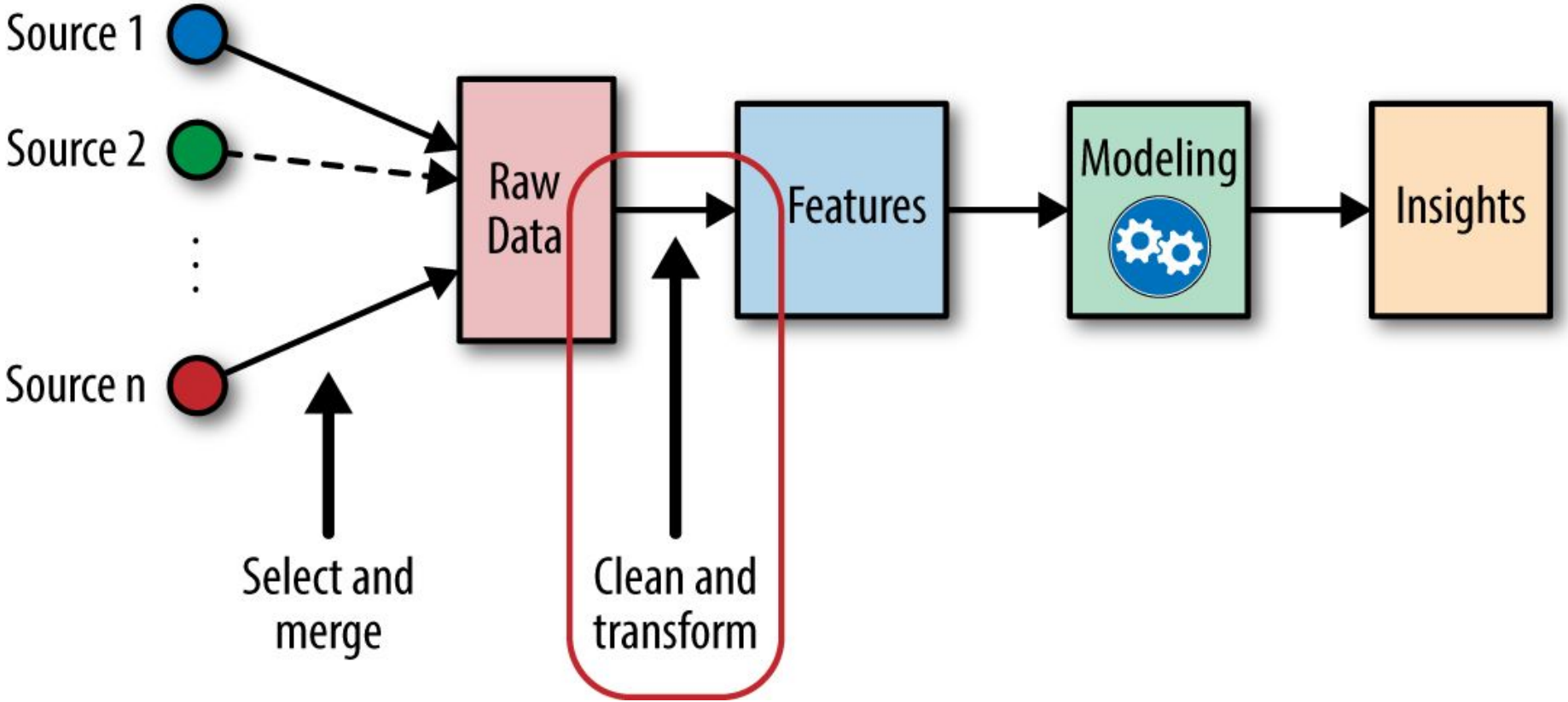


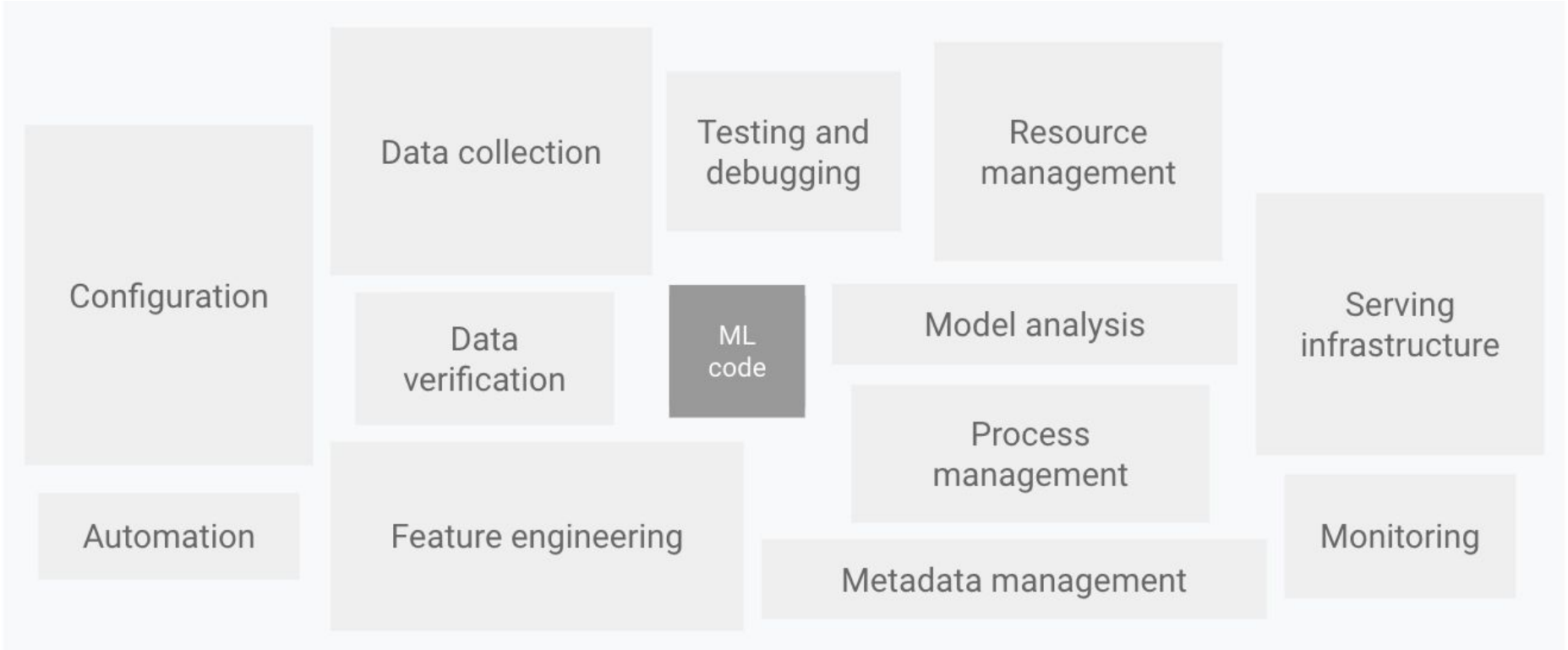
What my boss thinks I do

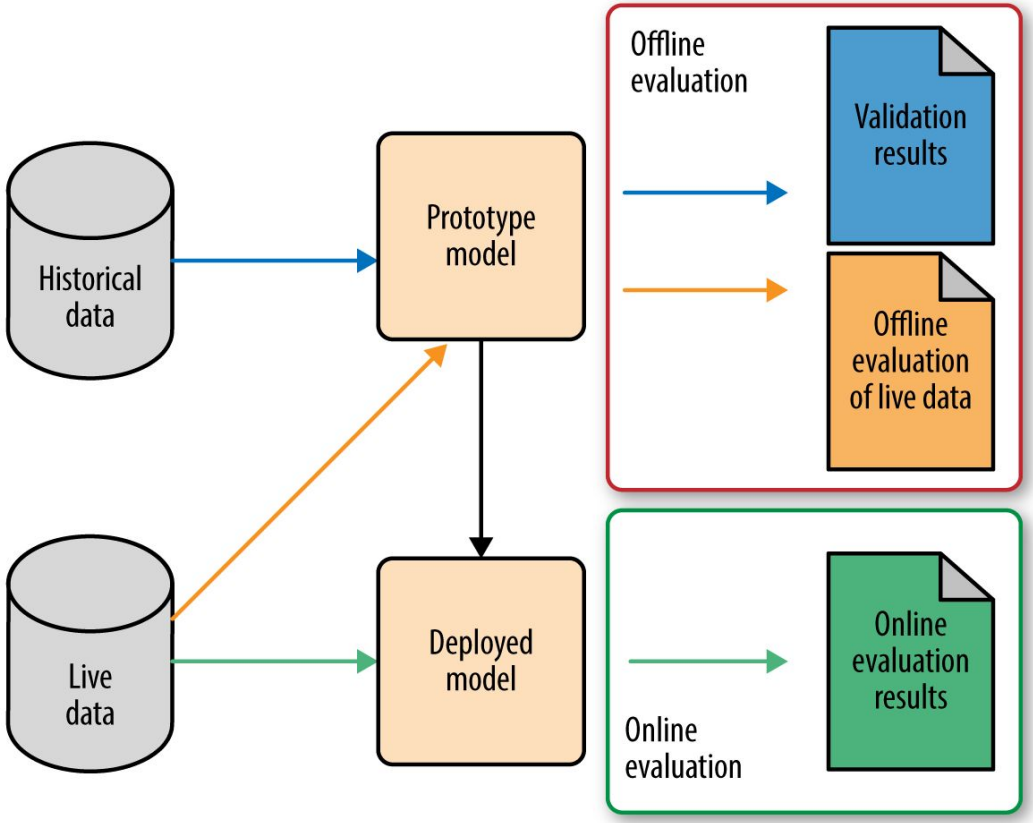


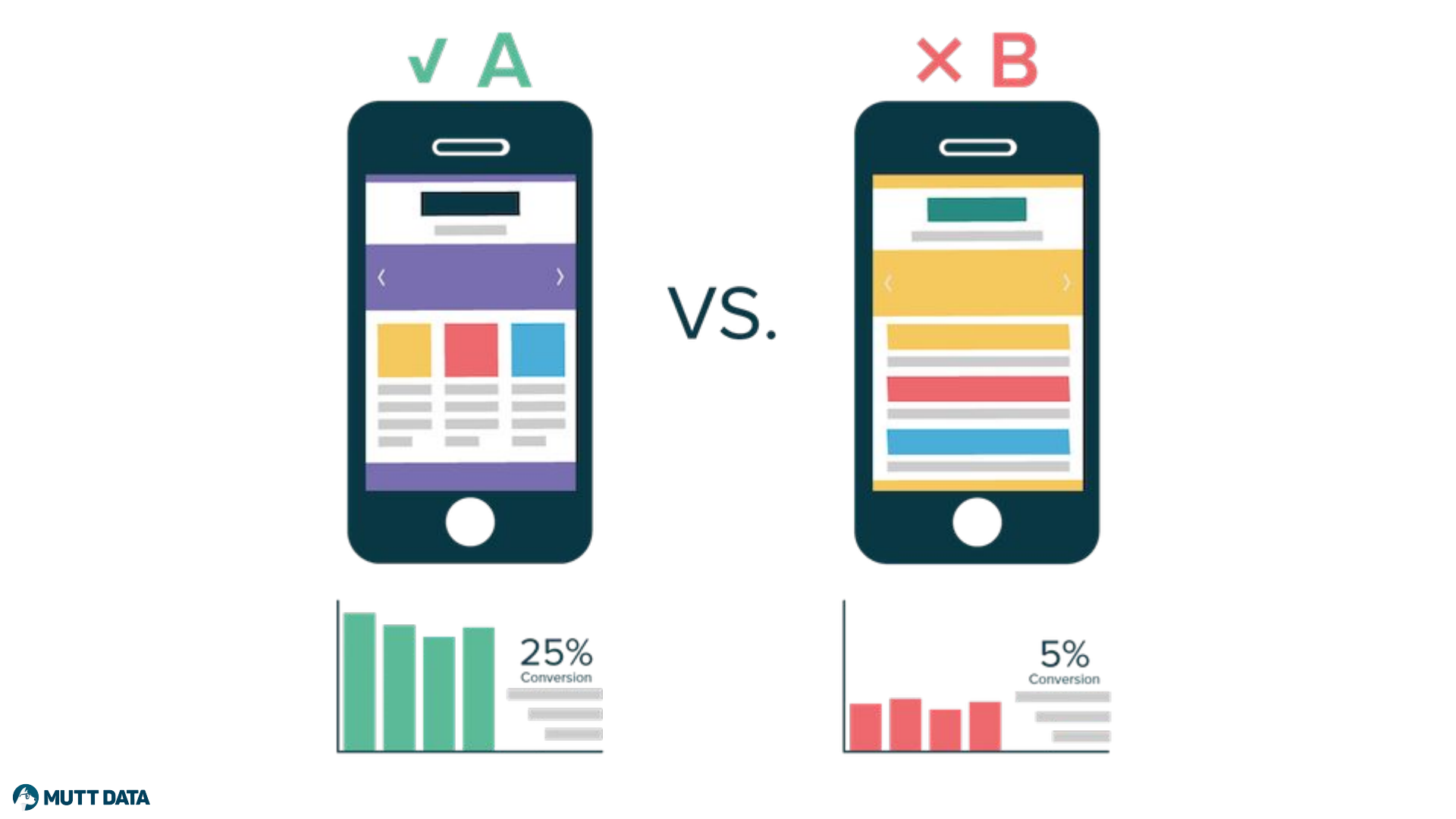
What I think I do

What I actually do


Source 1
Source 2
Source n
Raw Data
Features
Modeling
Insights
Select and merge
Clean and transform


Configuration
Data collection
Testing and debugging
Resource management
Data verification
ML code
Model analysis
Serving infrastructure
Process management
Automation
Feature engineering
Metadata management
Monitoring


Offline evaluation
Validation results
Offline evaluation of live data
Historical data
Prototype model
Live data
Deployed model
Online evaluation
Online evaluation results


✓ A
✕ B
VS.
25%
Conversion
5%
Conversion

Data

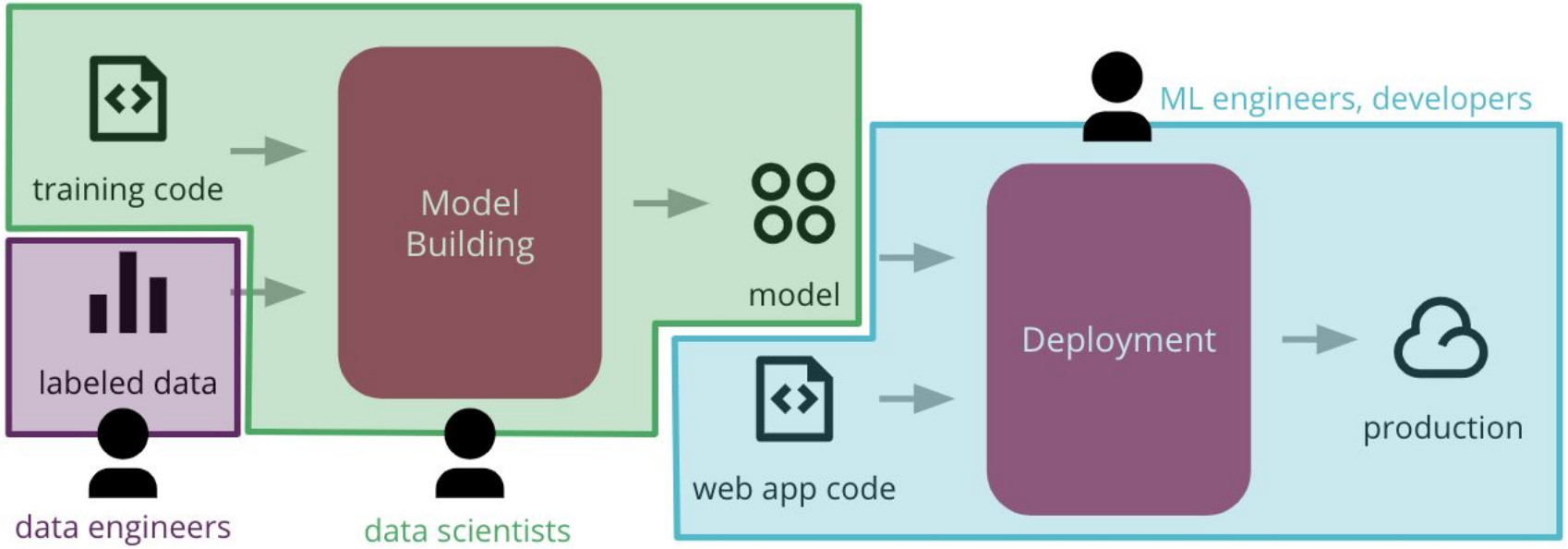
- Schema
- Sampling over Time
- Volume

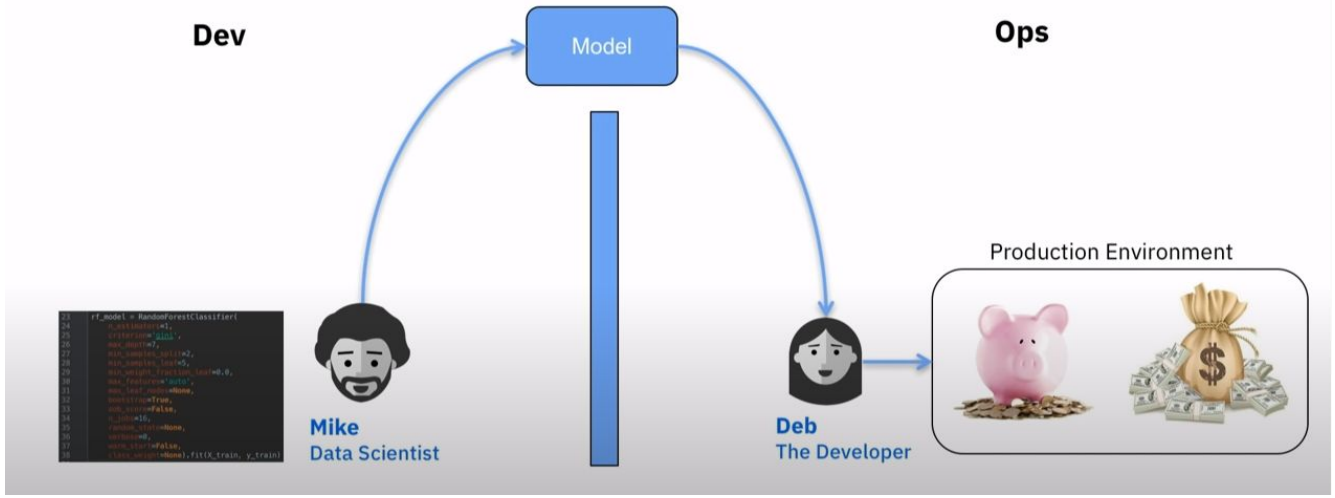
Model

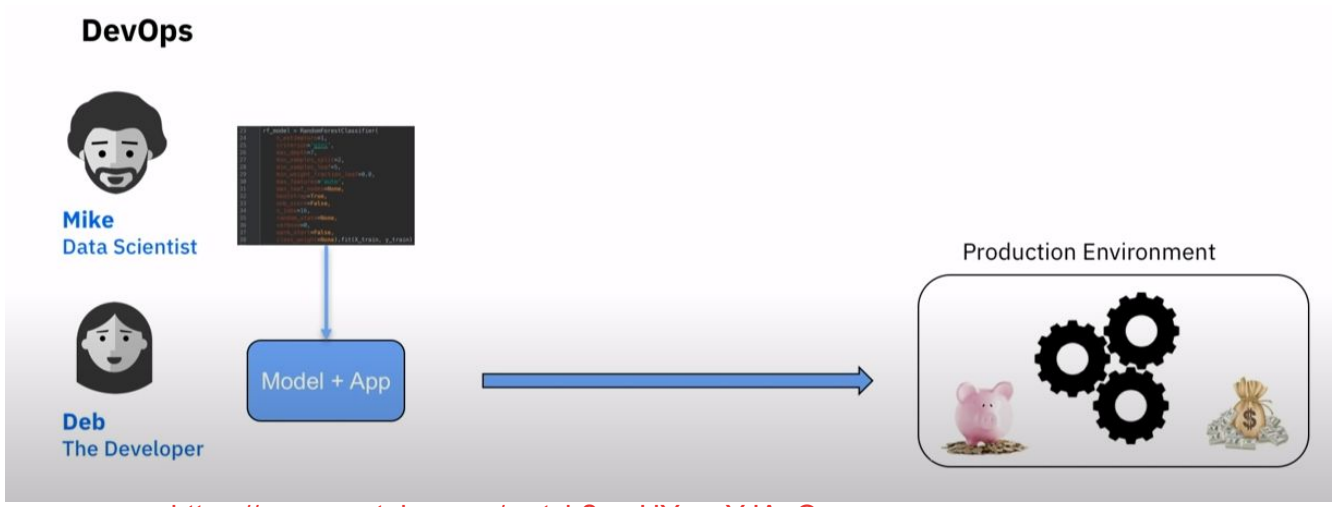
- Algorithms
- More Training
- Experiments

Code

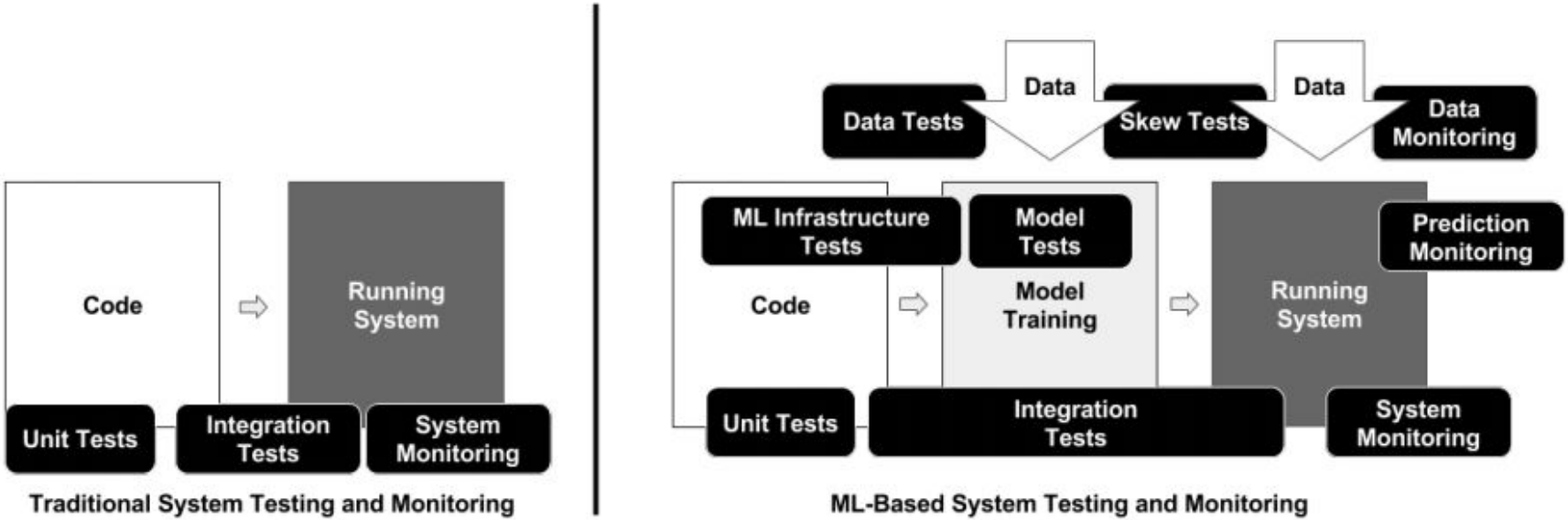
- Business Needs
- Bug Fixes
- Configuration





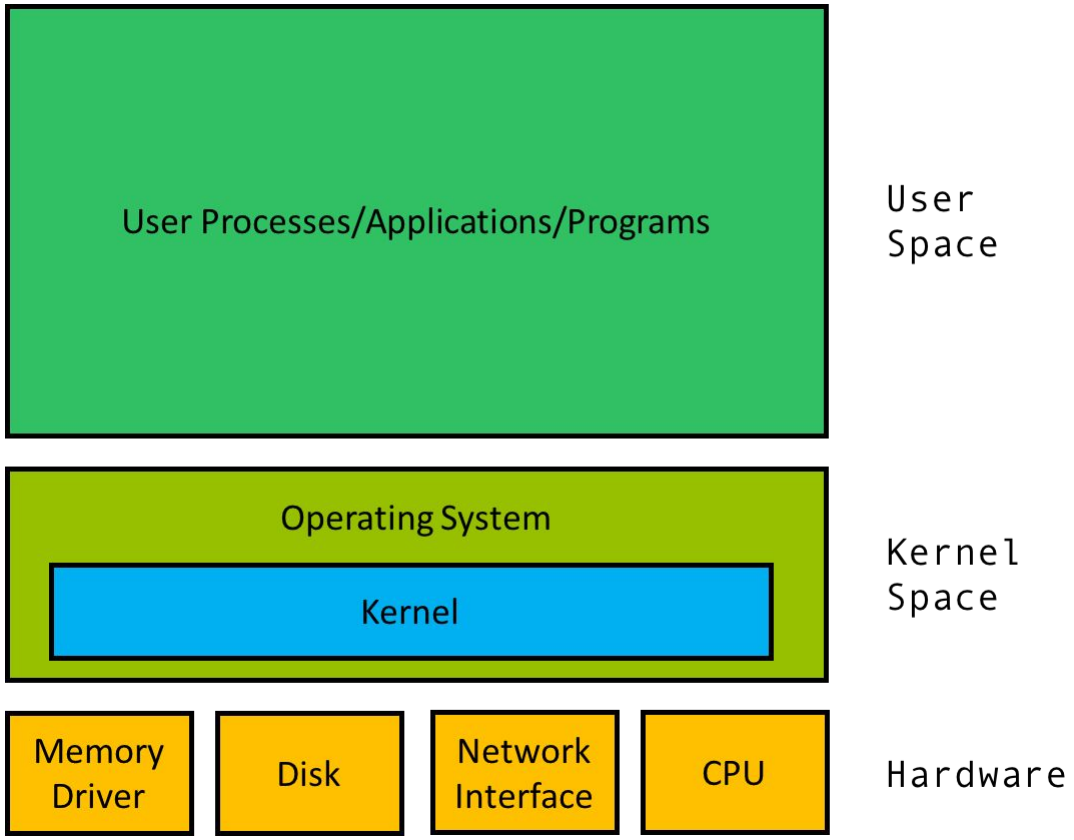


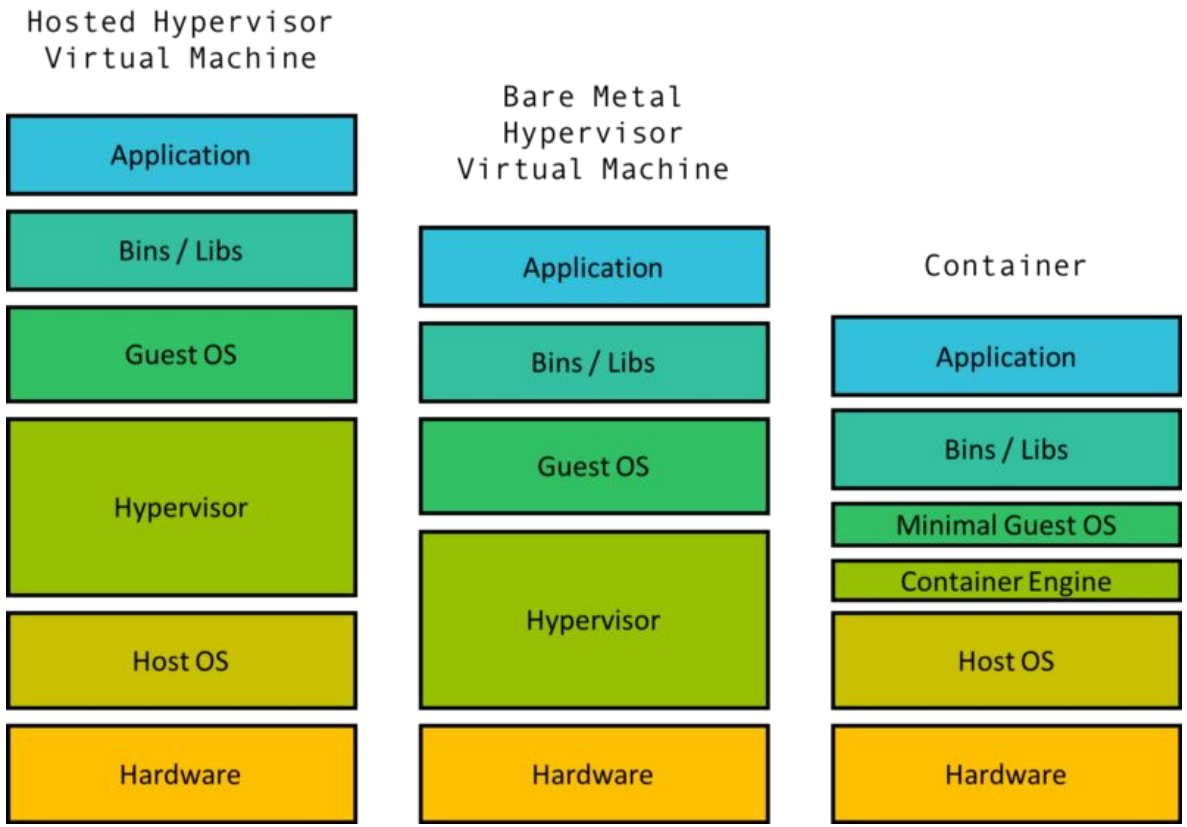
https://www.youtube.com/watch?v=-UYyyeYJAoQ

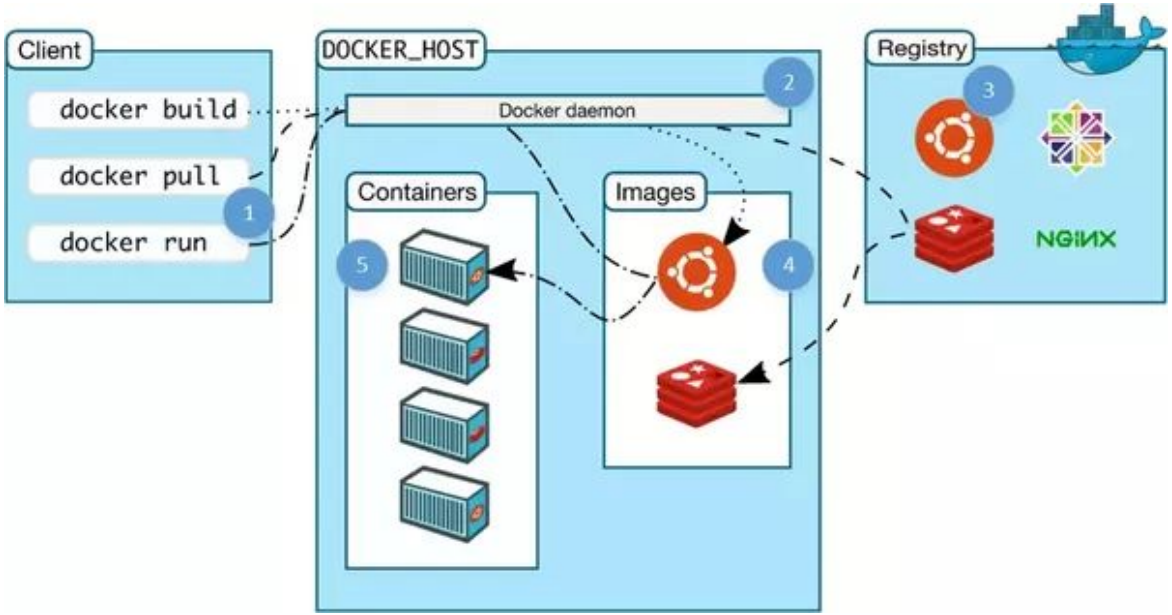


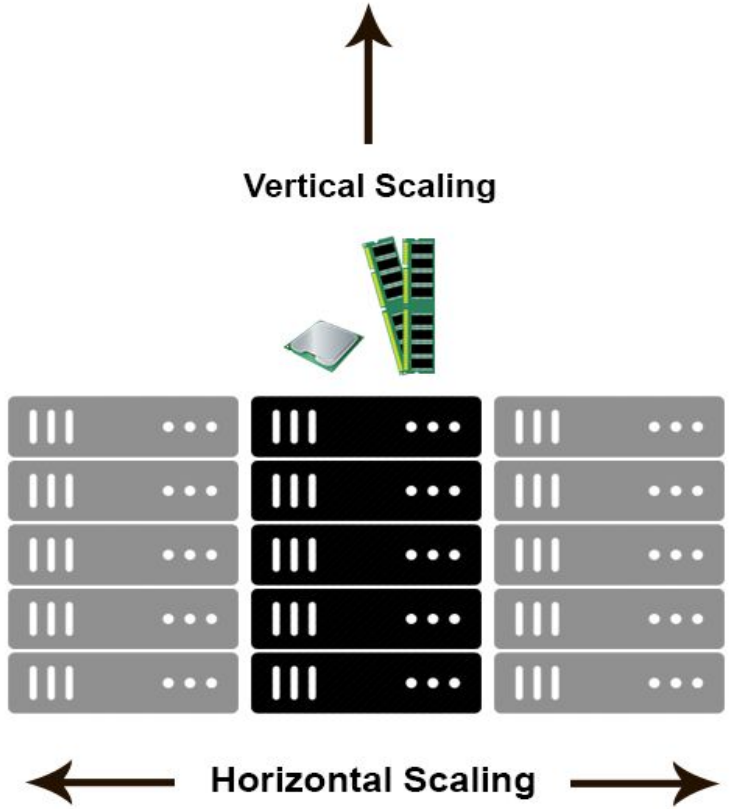
A Rubric for ML Production Readiness and Technical Debt Reduction
https://storage.googleapis.com/pub-tools-public-publication-data/pdf/aad9f93b86b7addfea4c419b9100c6cdd26cacea.pdf

02 | Cloud Computing


User Processes/Applications/Programs
User Space
Operating System
Kernel
Kernel Space
Memory Driver
Disk
Network Interface
CPU
Hardware


Hosted Hypervisor
Virtual Machine
Application
Bins / Libs
Guest OS
Hypervisor
Host OS
Hardware
Bare Metal
Hypervisor
Virtual Machine
Application
Bins / Libs
Guest OS
Hypervisor
Hardware
Container
Application
Bins / Libs
Minimal Guest OS
Container Engine
Host OS
Hardware

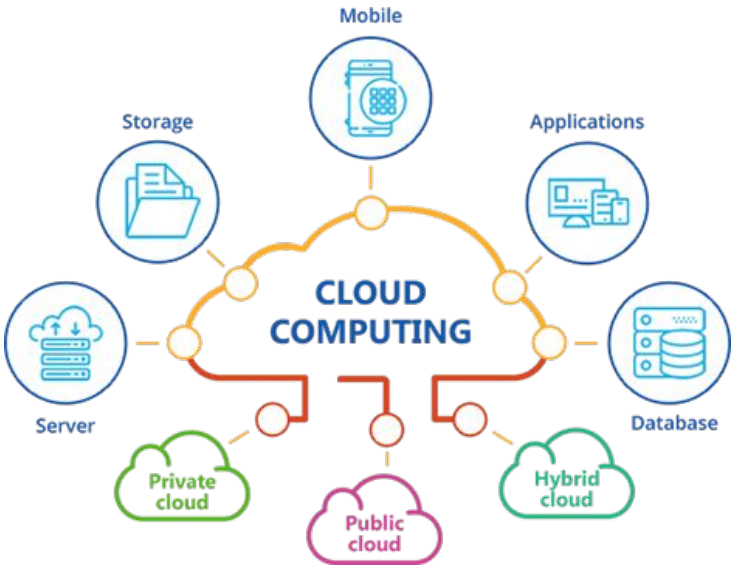

Client
docker build
docker pull
docker run
1
DOCKER_HOST
Docker daemon
2
Containers
5
Images
4
Registry
3
NGINX


Vertical Scaling
Horizontal Scaling

Cloud Computing

Es la prestación de servicios informáticos, incluidos servidores, almacenamiento, bases de datos, redes, software, análisis e inteligencia, a través de Internet ("la nube").

Por lo general, uno sólo paga por los servicios que usa, lo que le ayuda a reducir sus costos operativos, ejecutar su infraestructura de manera más eficiente y escalar a medida que cambian las necesidades de su negocio.



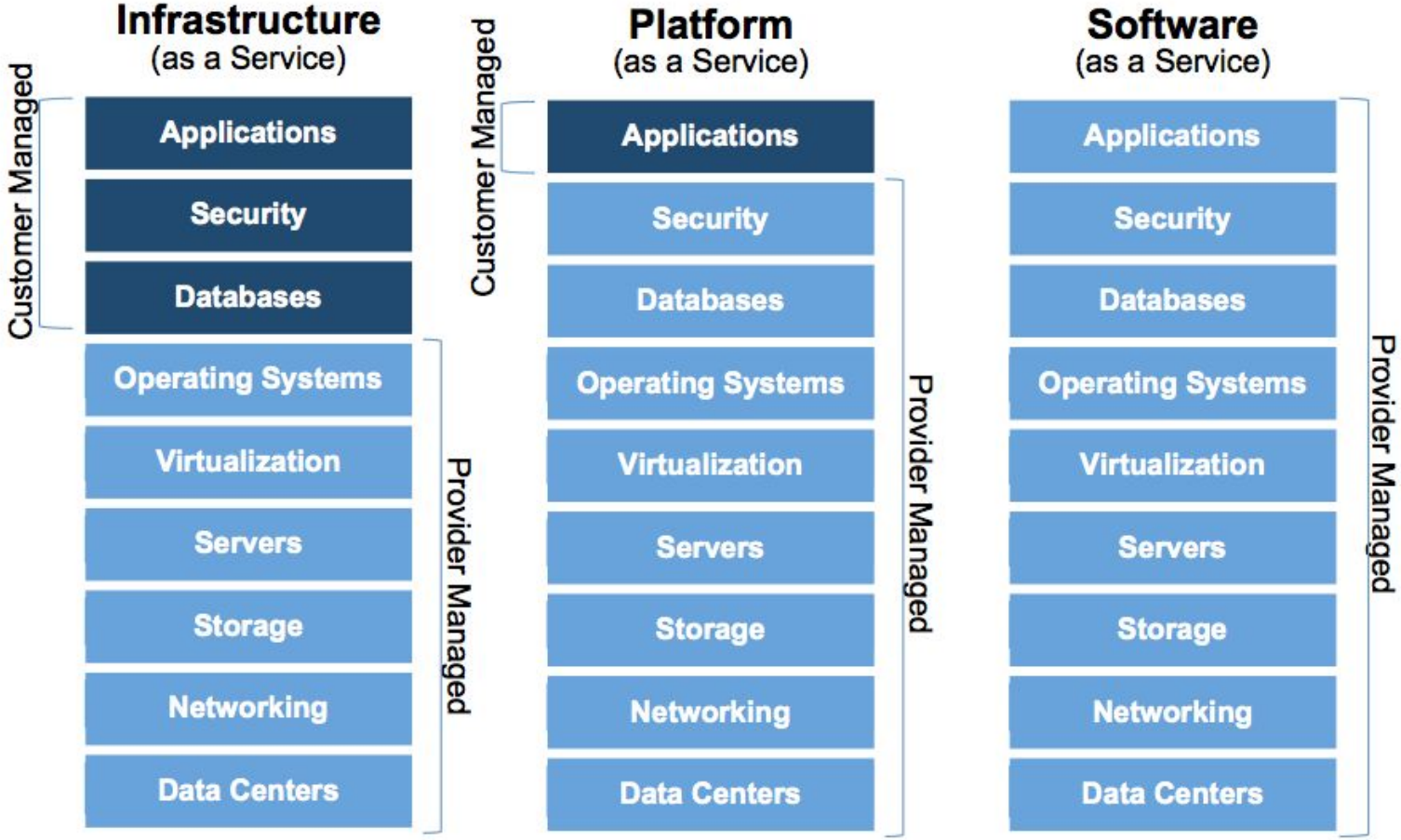
Public Cloud Computing

Es la manera más usual de utilizar servicios Cloud.

Se refiere a contratar servicios Cloud a una empresa que los provee públicamente a cualquier usuario potencial.

Los proveedores ofrecen interfaces web y APIs para poder aprovisionar y administrar los recursos que se necesiten y se paga mensualmente por lo utilizado.

El mercado actualmente se encuentra fuertemente concentrado en tres competidores donde Amazon tiene una ventaja significativa sobre el resto.

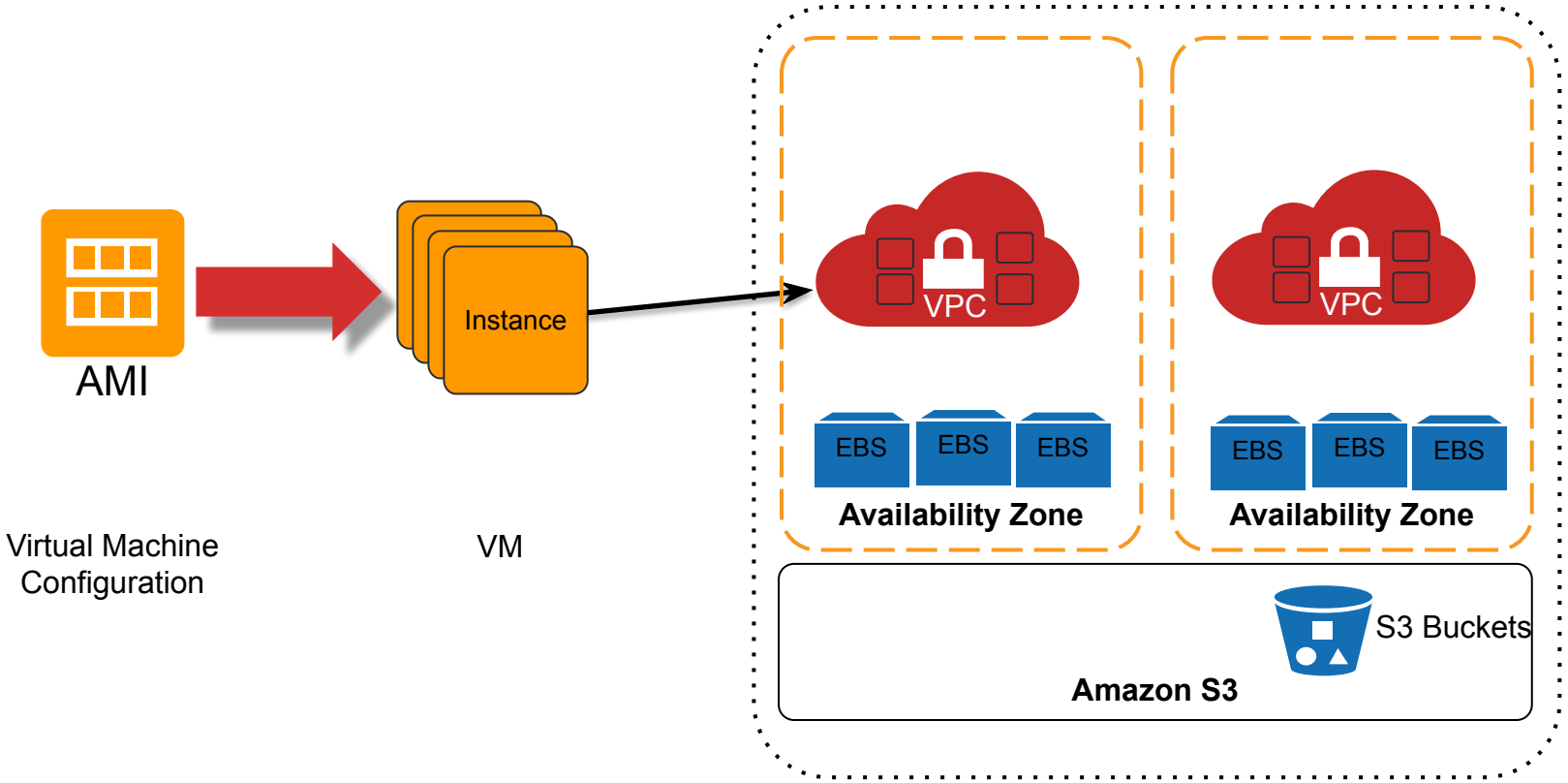

Infrastructure
(as a Service)
Customer Managed
Applications
Security
Databases
Operating Systems
Virtualization
Servers
Storage
Networking
Data Centers
Provider Managed
Platform
(as a Service)
Customer Managed
Applications
Security
Databases
Operating Systems
Virtualization
Servers
Storage
Networking
Data Centers
Provider Managed
Software
(as a Service)
Applications
Security
Databases
Operating Systems
Virtualization
Servers
Storage
Networking
Data Centers
Provider Managed

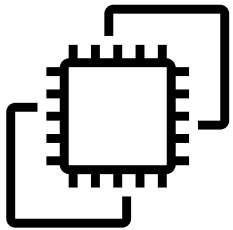
Ventajas

1. Gasto de capital comercial por gasto variable: no se requiere inversión inicial en hardware o software.
2. Dejar de adivinar la capacidad
3. Aumentar la velocidad y la agilidad reduciendo el tiempo necesario para que los recursos estén disponibles.
4. Dejar de gastar dinero en administrar un datacenter.
5. Alcance global con poco esfuerzo.
6. Mayor seguridad.

Figure 1. Magic Quadrant for Cloud Infrastructure as a Service, Worldwide

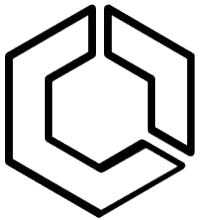
Source: Gartner (July 2019)


AMI
Instance
VPC
VPC
EBS
EBS
EBS
EBS
EBS
EBS
Availability Zone
Availability Zone
S3 Buckets
Amazon S3
Virtual Machine Configuration
VM



Amazon EC2

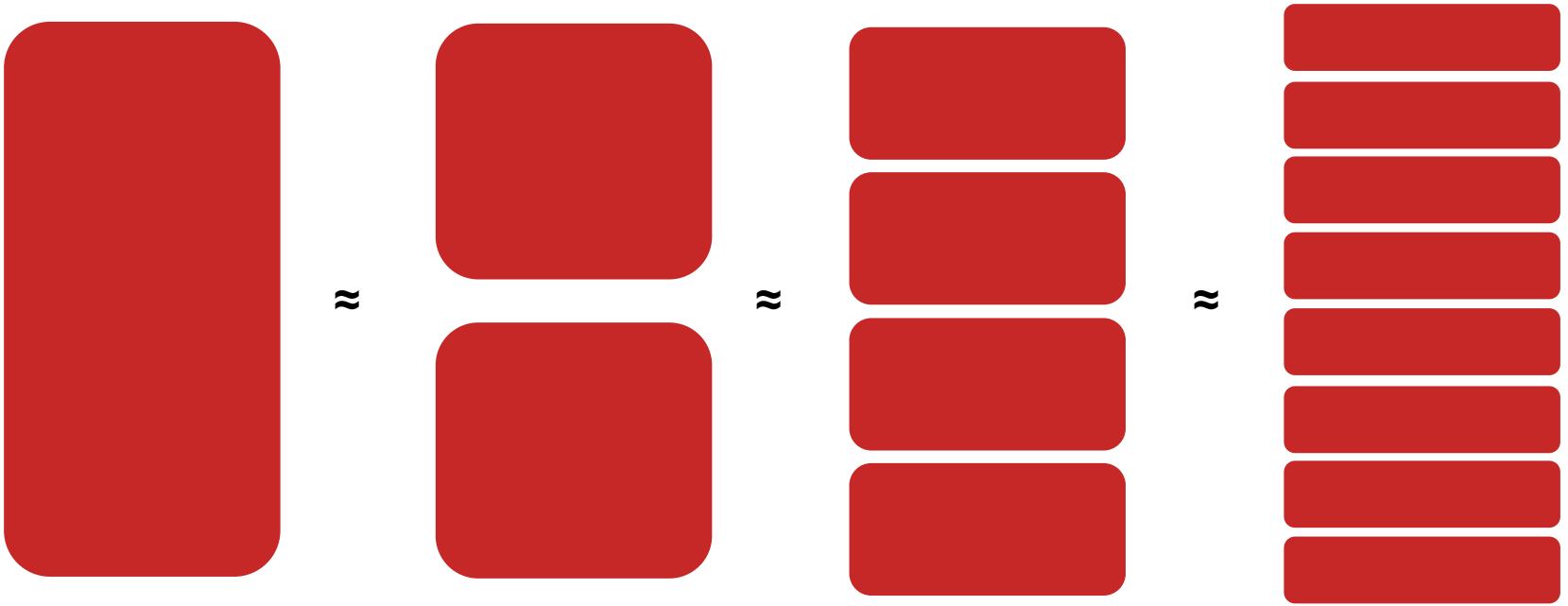
Virtual Machines



Amazon ECS

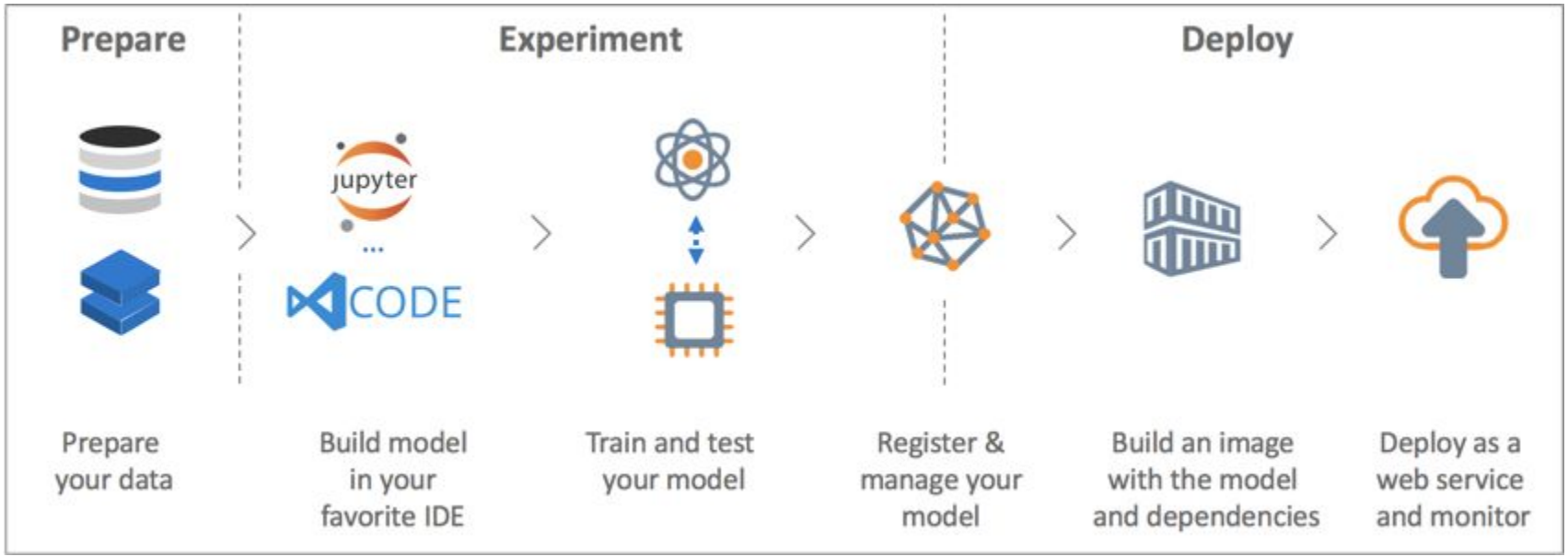
Container management services over EC2

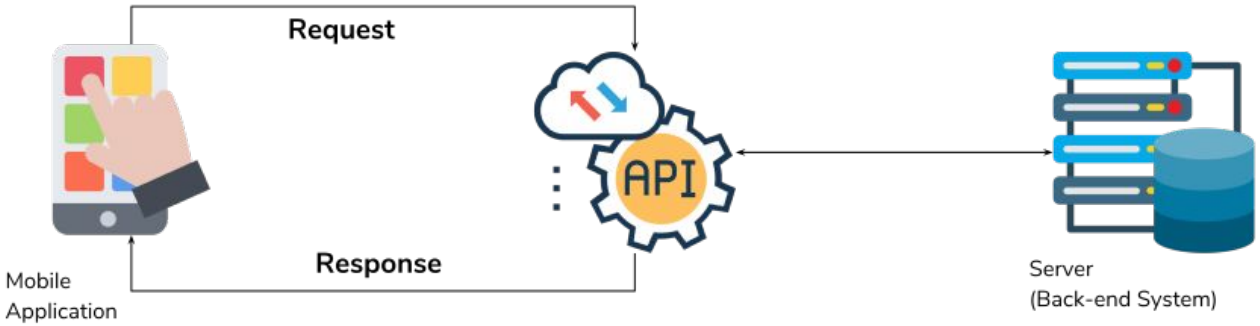
Amazon EC2: Instance Sizes

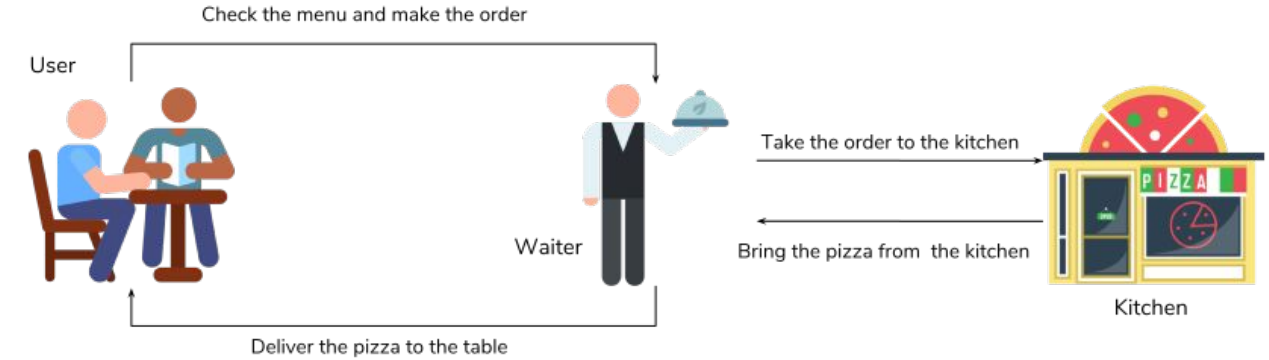


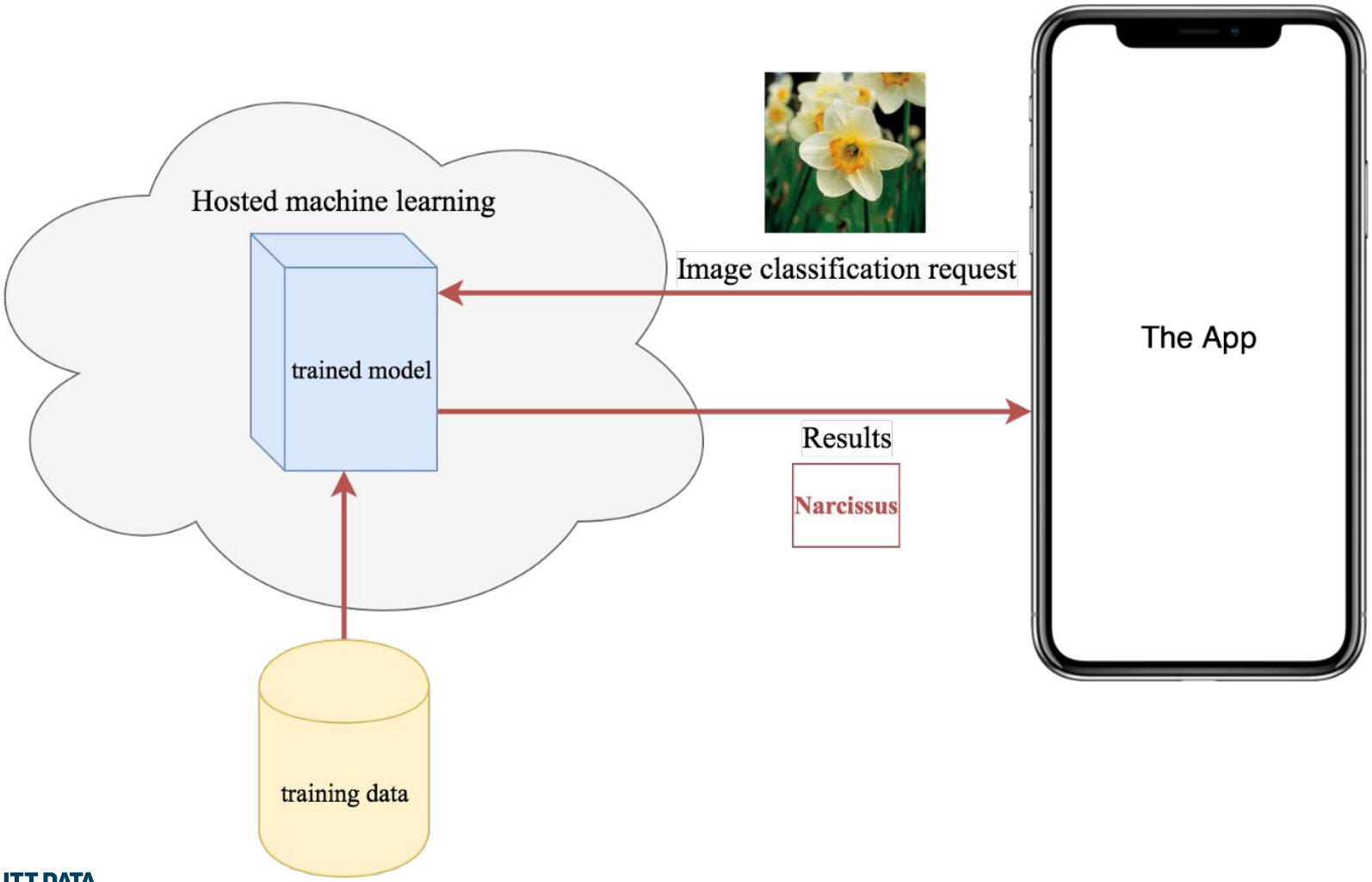
c4.8xlarge ≈ 2 - c4.4xlarge ≈ 4 - c4.2xlarge ≈ 8 - c4.xlarge

03 | ML en Producción


Prepare
Experiment
Deploy
jupyter
CODE
Prepare your data
Build model in your favorite IDE
Train and test your model
Register & manage your model
Build an image with the model and dependencies
Deploy as a web service and monitor


Request
API
Response
Mobile
Application
Server
(Back-end System)


Check the menu and make the order
User
Waiter
Take the order to the kitchen
Bring the pizza from the kitchen
PIZZA
Kitchen
Deliver the pizza to the table


Hosted machine learning
trained model
training data
Image classification request
Results
Narcissus
The App


The App
trained model
Narcissus

mlflow Github Docs

Listing Price Prediction

Experiment ID: 0 Artifact Location: /Users/matei/mlflow/demo/mlruns/0

Search Runs: metrics.R2 > 0.24 Search

Filter Params: alpha, lr Filter Metrics: rmse, r2 Clear

4 matching runs Compare Selected Download CSV

					Parameters		Metrics		
	Time	User	Source	Version	alpha	l1_ratio	MAE	R2	RMSE
☐	17:37	matei	linear.py	3a1995	0.5	0.2	84.27	0.277	158.1
☐	17:37	matei	linear.py	3a1995	0.2	0.5	84.08	0.264	159.6
☐	17:37	matei	linear.py	3a1995	0.5	0.5	84.12	0.272	158.6
☐	17:37	matei	linear.py	3a1995	0	0	84.49	0.249	161.2

04 | Sagemaker

ML Stack AWS

APPLICATION SERVICES

Vision
Rekognition Image
Rekognition Video
Amazon Textract

Speech
Amazon Polly
Amazon Transcribe

Language
Amazon Lex
Amazon Translate
Amazon Comprehend

Forecasting
Amazon Forecast

Personalization
Amazon Personalize

PLATFORMS

Amazon SageMaker

FRAMEWORKS ET INFRASTRUCTURE

DL AMI

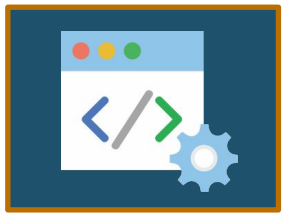
EC2 P3, C5, F1

Elastic Inference

Greengrass

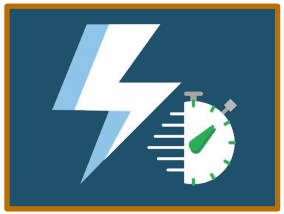
Sagemaker

1



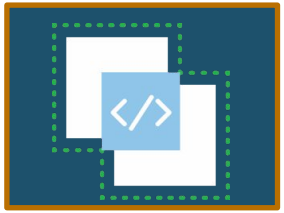
Notebook Instances con 200+ Ejemplos

2



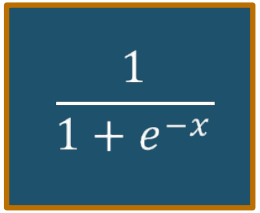
17 Algoritmos built-in

3



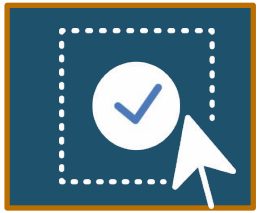
ML Training Service

4



Tuning de hiperparámetros

5



Hosting de Servicios

6

Batch Transform
- Inferencia
- Transformación de datos

7

SDK's:
- Python
- Spark

8

Documentación & White Papers & Blog Posts

Clasificación
- Linear Learner *
- XGBoost
- KNN

Trabajar con texto
- Blazing Text
  - Supervised
  - Unsupervised *

Traducción de secuencias
- Seq2Seq *

Computer Vision
- Image Classification <>
- Object Detection <>
- Semantic Segmentation

Recomendación
- Factorization Machines *

Detección de anomalías
- Random Cut Forests *
- IP Insights *

Regresión
- Linear Learner *
- XGBoost
- KNN

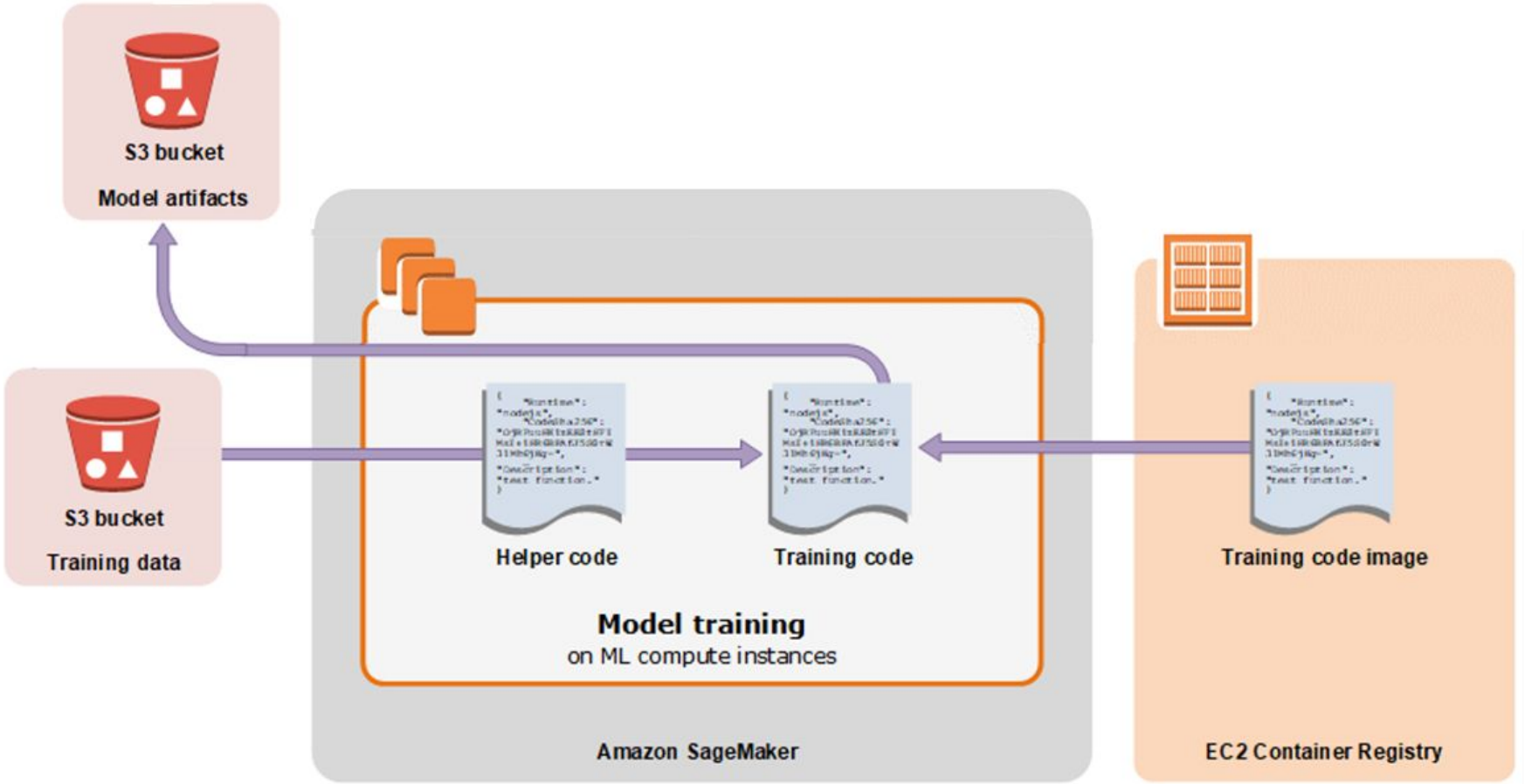
Topic Modeling
- LDA
- NTM

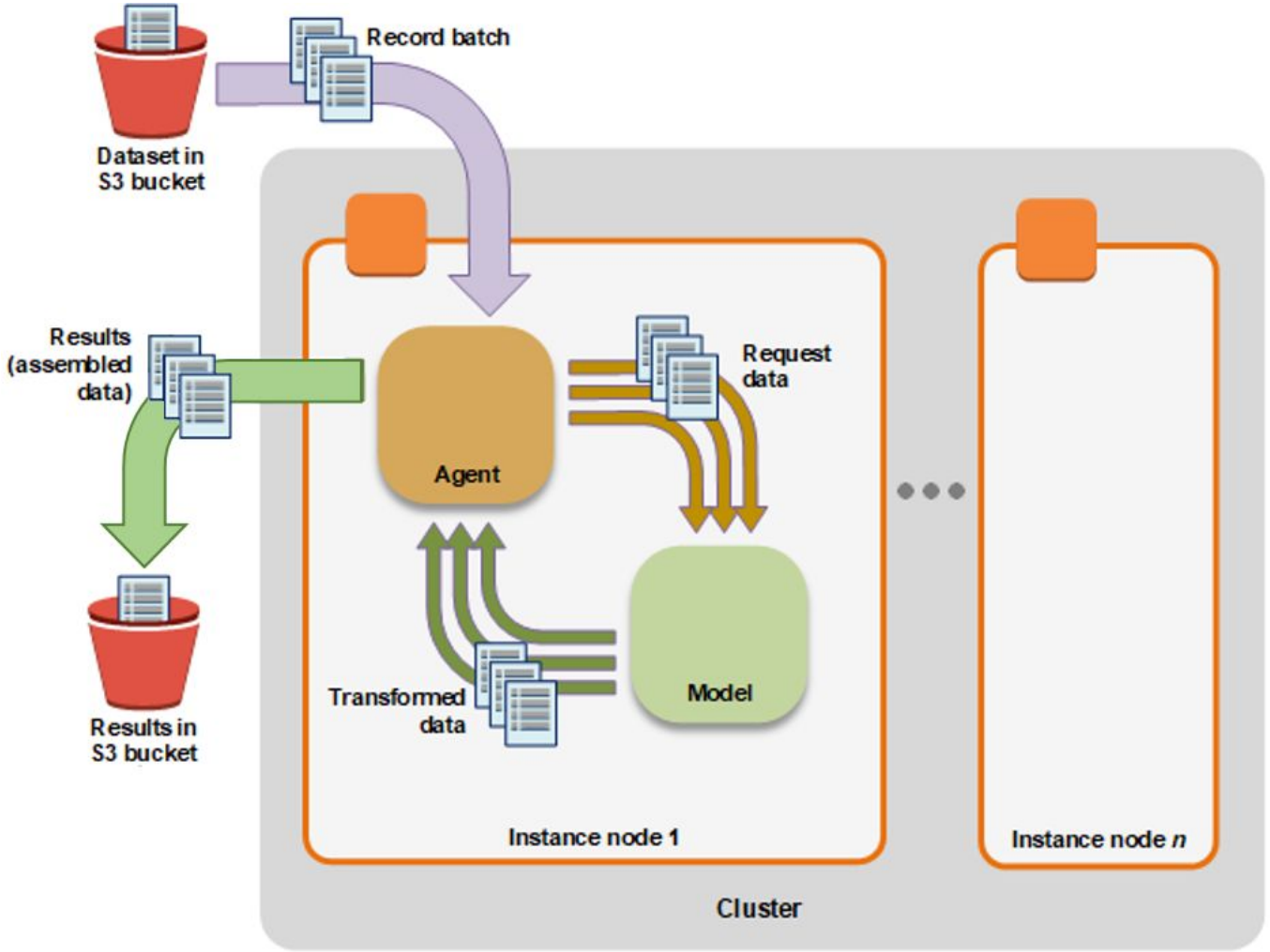
Forecasting
- DeepAR *

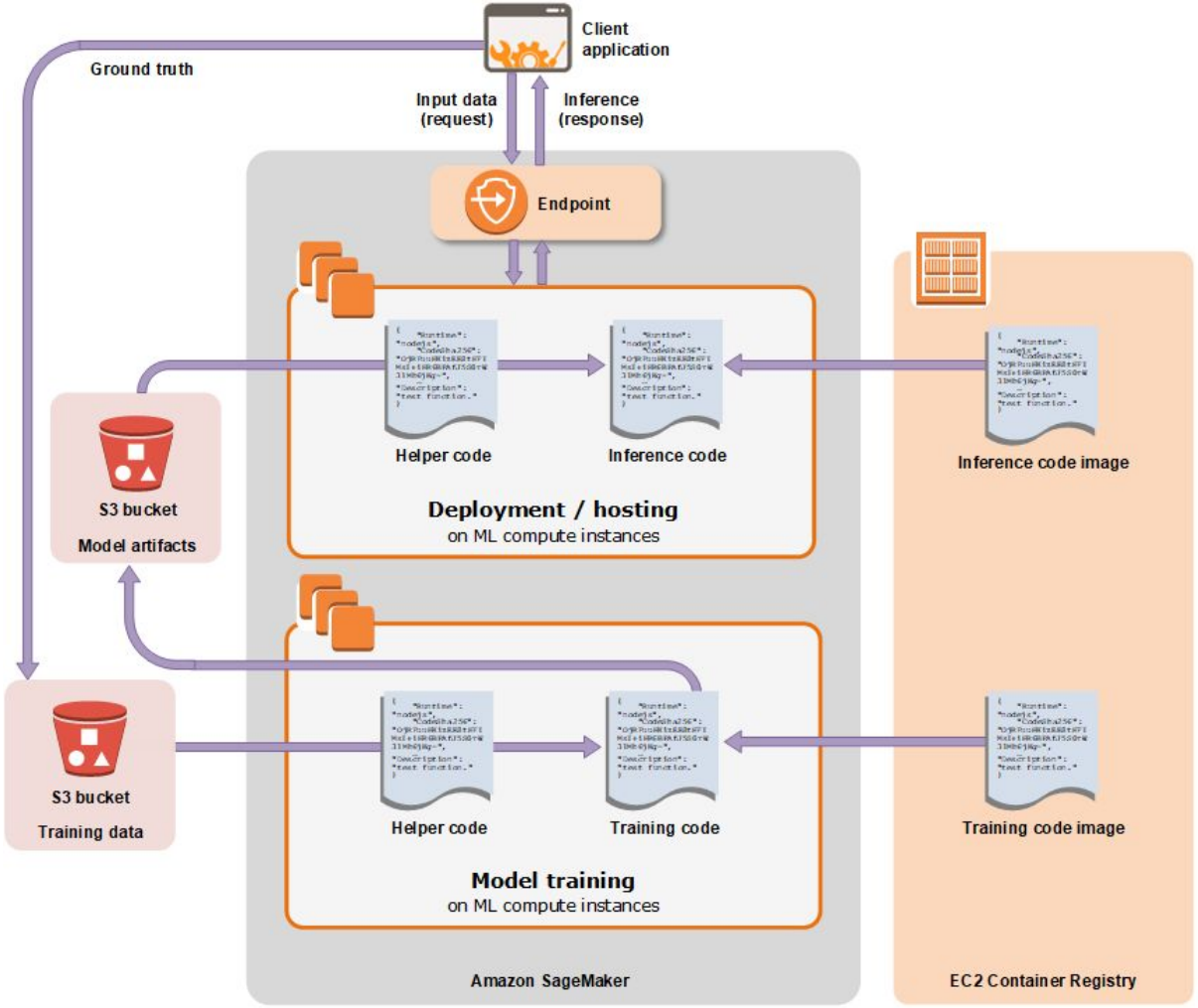
Clustering
- Kmeans *

Reducción de dimensiones
- PCA
- Object2Vec

https://aws.amazon.com/marketplace/solutions/machine-learning


S3 bucket
Model artifacts
S3 bucket
Training data
Helper code
Training code
Model training
on ML compute instances
Amazon SageMaker
Training code image
EC2 Container Registry


Record batch
Dataset in S3 bucket
Results (assembled data)
Agent
Request data
Transformed data
Model
Results in S3 bucket
Instance node 1
Instance node n
Cluster


Client application
Ground truth
Input data (request)
Inference (response)
Endpoint
Helper code
Inference code
Inference code image
Deployment / hosting
on ML compute instances
S3 bucket
Model artifacts
S3 bucket
Training data
Helper code
Training code
Training code image
Model training
on ML compute instances
Amazon SageMaker
EC2 Container Registry


Muchas Gracias!
MUTT DATA