TensorFlow in the Real World: Deploying Models

Gusti Triandi Winata

Community Researcher of RL & LLM
Cohere for AI

bangkit

About Me

Gusti Triandi Winata

Latest Work Experiences:

- Researcher, Cohere for AI **2024 - present**
- MLE Consultant, Freeport Indonesia **2023 - 2024**
- Mid. MLE, eFishery **2022 - 2023**
- OSS Fellowship with Adobe, Major League Hacking **2021**

Education:

- Bandung Institute of Technology Graduated at **2021**
  Bachelor of Electrical and Computer Engineering
  Was a Bangkit Graduate of the first batch (2020)!

bangkit

Ground Rules

Observe the following rules to ensure a supportive, inclusive, and engaging classes

Give full attention in class

Mute your microphone when you're not talking

Keep your camera on

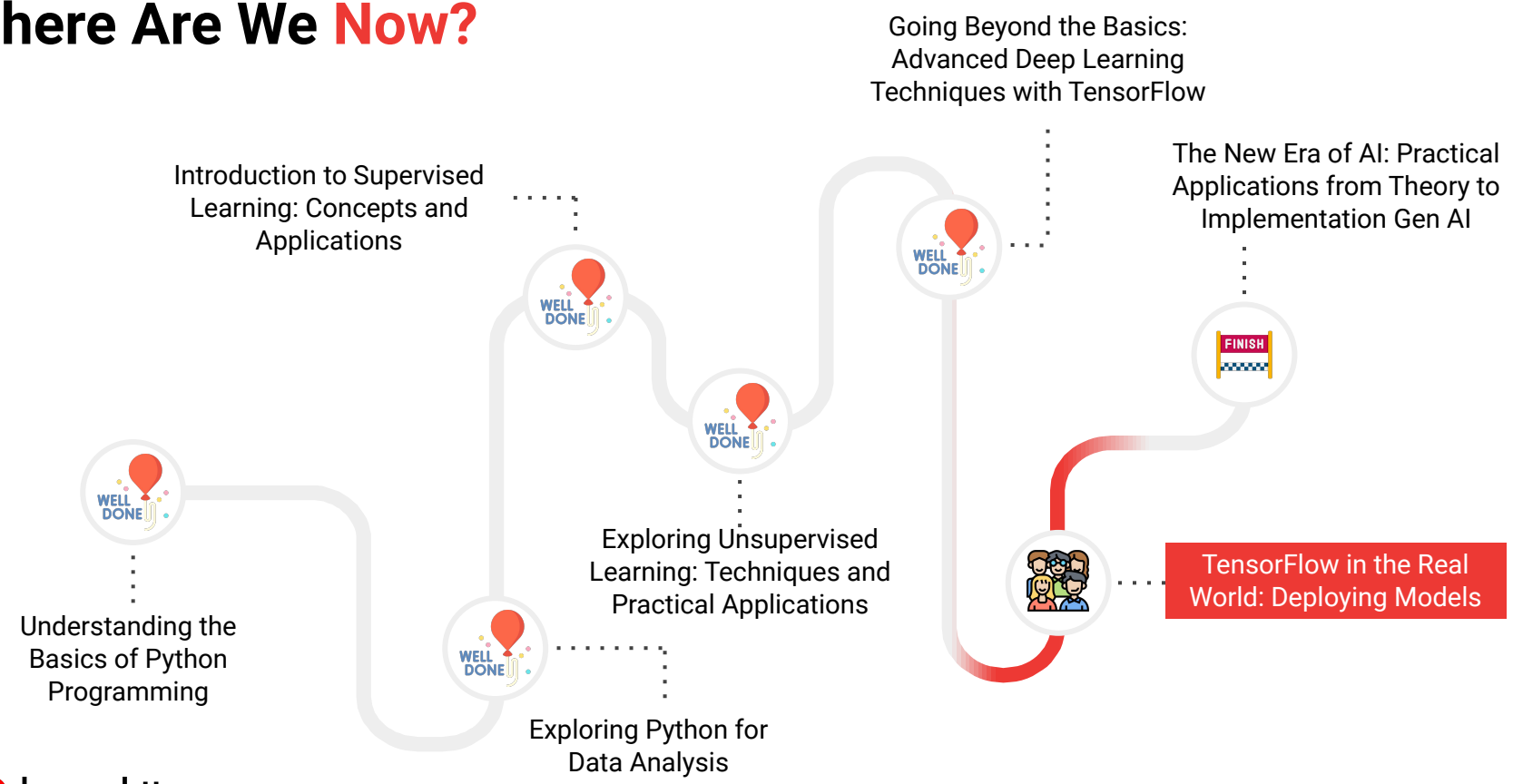
Turn on the CC Feature on Meet

Use raise hand or chat to ask questions

Make this room a safe place to learn and share

bangkit

Where Are We Now?



bangkit

Learning Objectives

4 **Operate how to use TensorFlow Dataset**

3 **Illustrate basic concepts in Federated Learning**

2 Conceptualize how to use TensorFlow.js, TensorFlow Lite, & TensorFlow Serving

1 Explain basic concepts in Machine Learning Deployment

bangkit

Today's Agenda

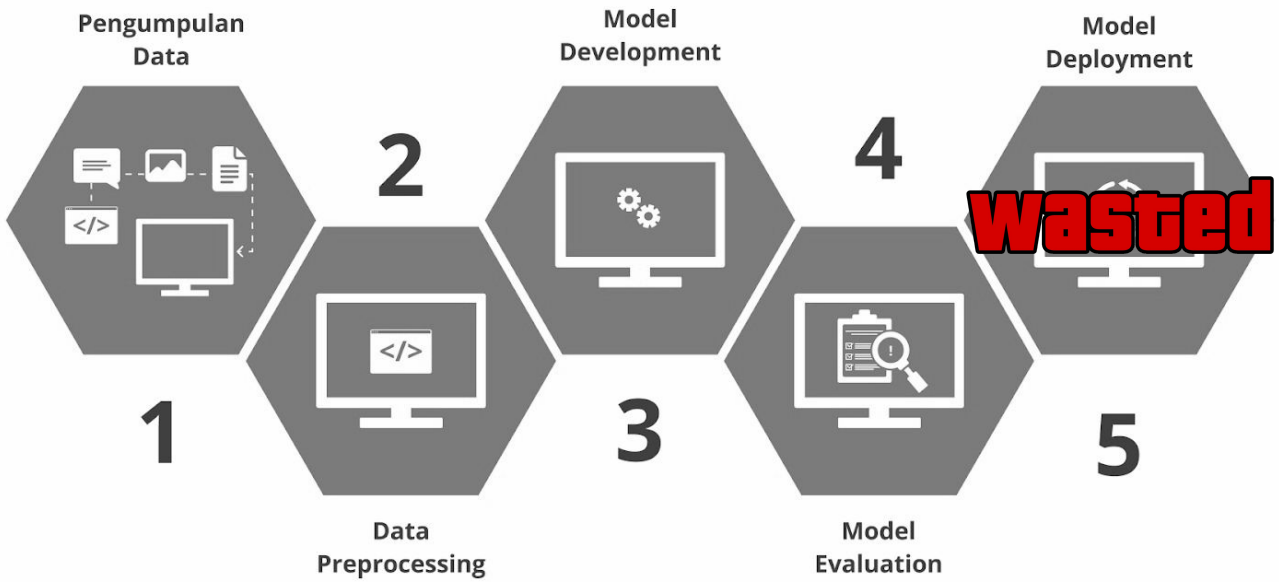
Embarking on the Journey of Machine Learning Model Deployment

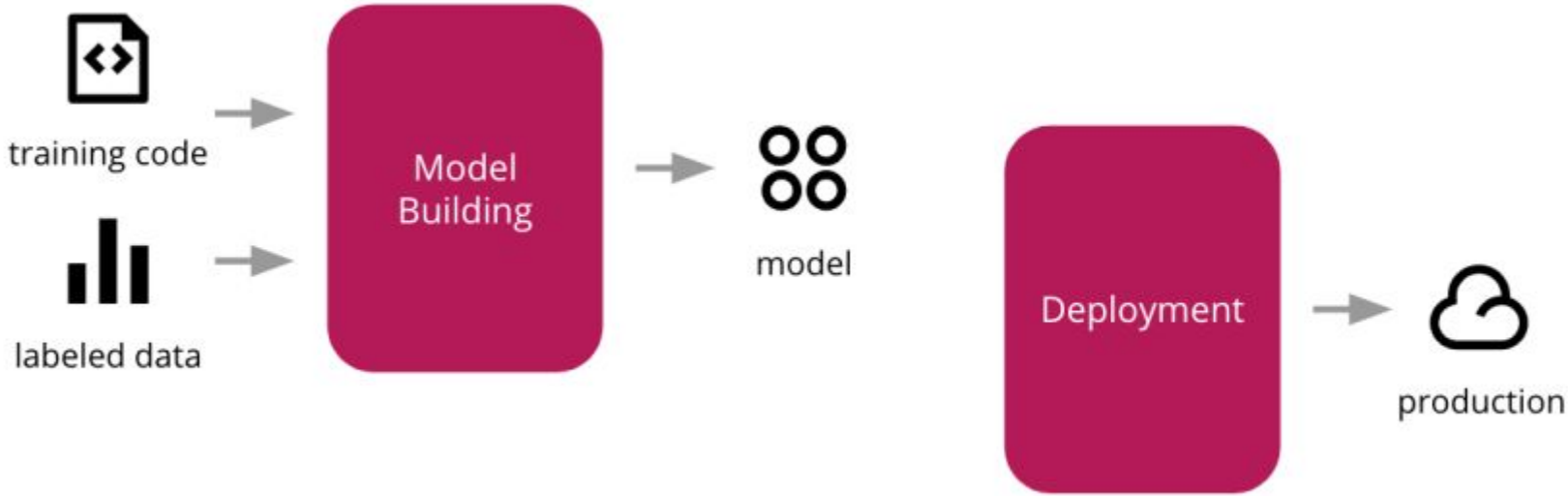
Utilizing TensorFlow Tools for Deploying Models in Real-World Applications

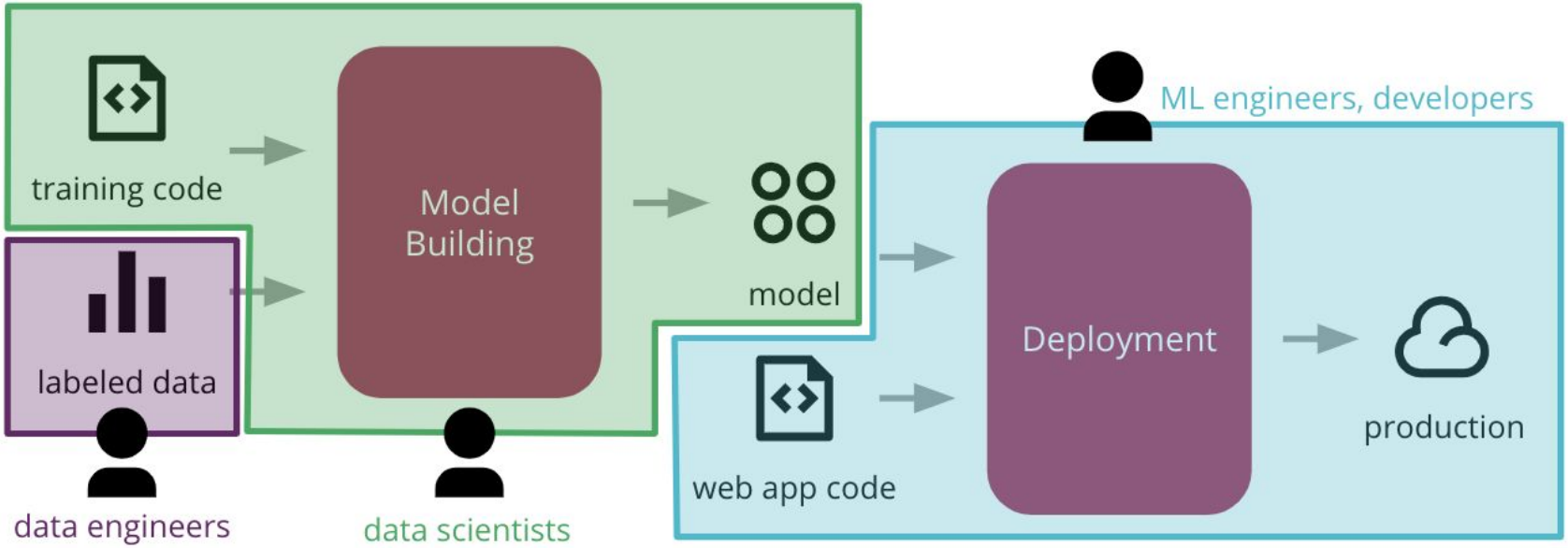
Highlighting Basic Principles of Federated Learning

Introduction to Data Handling with TensorFlow Datasets

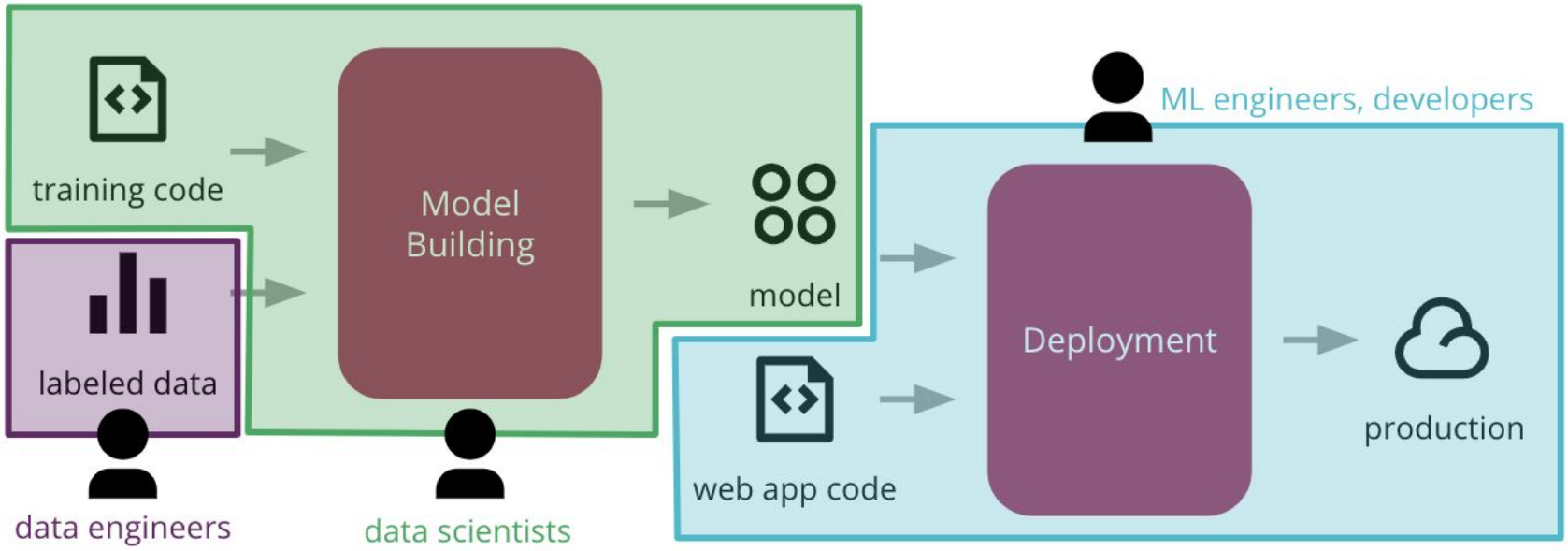
bangkit


Pengumpulan Data
1
2
Data Preprocessing
Model Development
3
4
Model Evaluation
Model Deployment
5
wasted


training code
labeled data
Model Building
model
Deployment
production


training code
Model Building
model
labeled data
data engineers
data scientists
ML engineers, developers
Deployment
production
web app code


Bangkit ML Students


training code
labeled data
data engineers
Model Building
data scientists
model
ML engineers, developers
web app code
Deployment
production

Introduction to Machine Learning Deployment

bangkit

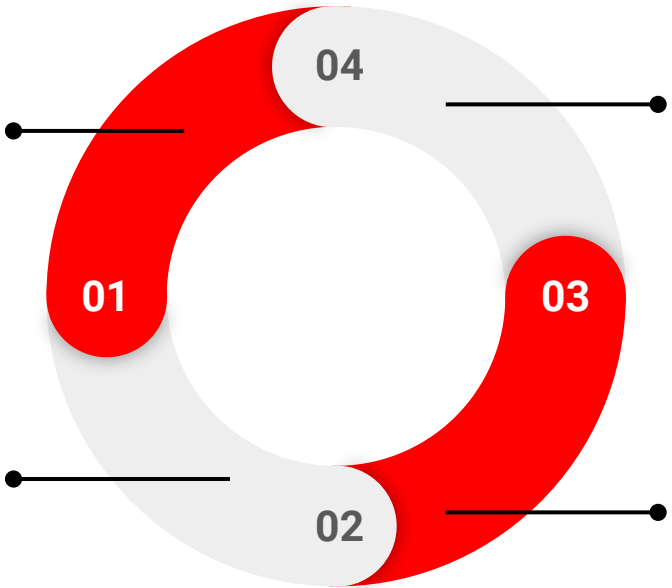
Lifecycle of an ML Project

Project Planning & Setup

Decide the problem to work on, determine the requirements and goals, as well as figure out how to allocate resources properly.

Data Collection & Labeling

Collect & organize data (images, text, tabular, etc.) & potentially annotate them with ground truth, depending on the specific sources where they come from.

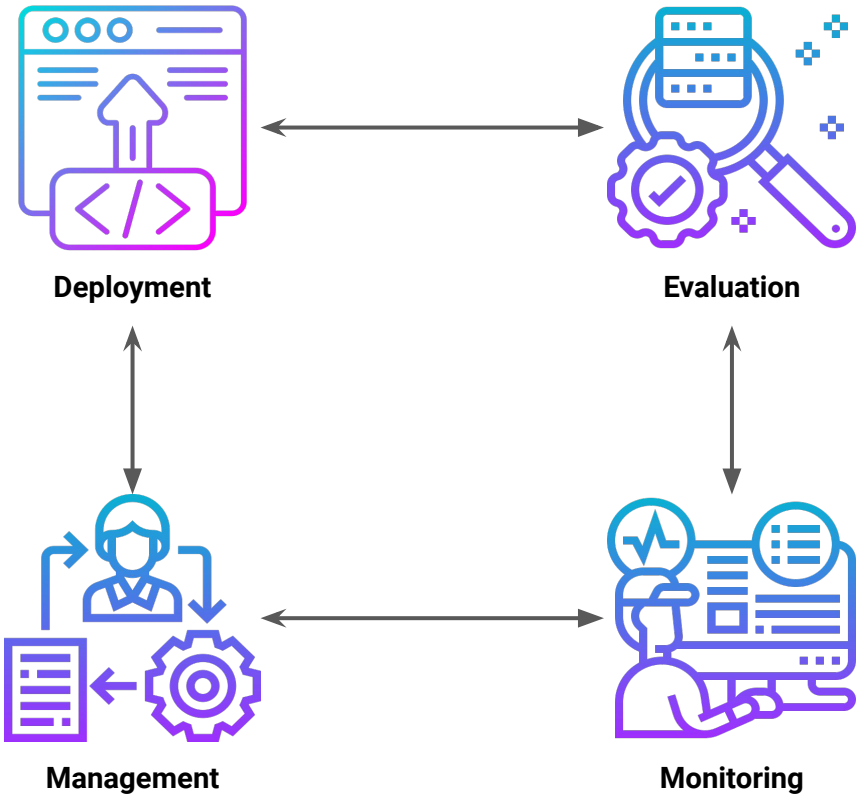


Deployment & Monitoring

Put the model into production, write the software needed to make the model run, and make predictions.

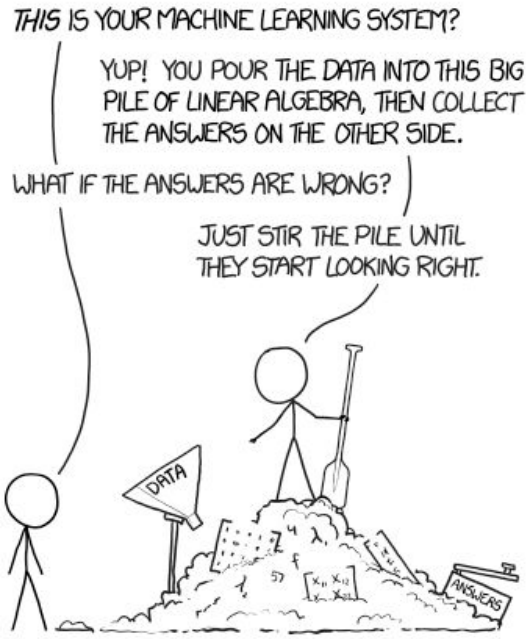
Model Training & Debugging

Implement baseline models quickly, find and reproduce state-of-the-art methods for the problem domain, debug our implementation, and improve the model performance for specific tasks.


Deployment
Evaluation
Management
Monitoring

The Challenges of Model Deployment

- ML models are sensitive to incoming data's **semantics, quantity, and completeness**.
- The performance of ML models in production **degrades over time** due to changes in actual data.
- ML models only work with data from a **specific demographic**.

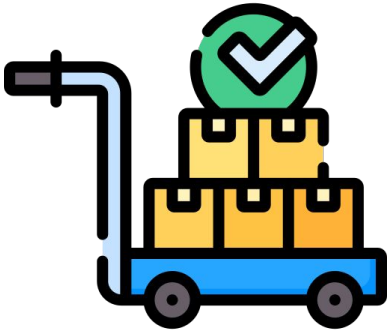


Model Deployment Options

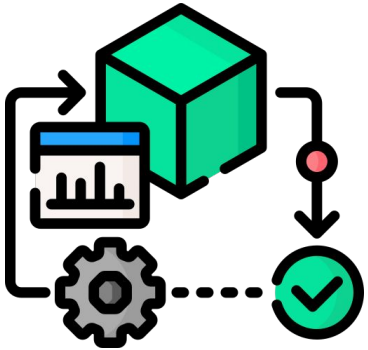
Centralize model in server

Distribute model on user device

Model Deployment Options



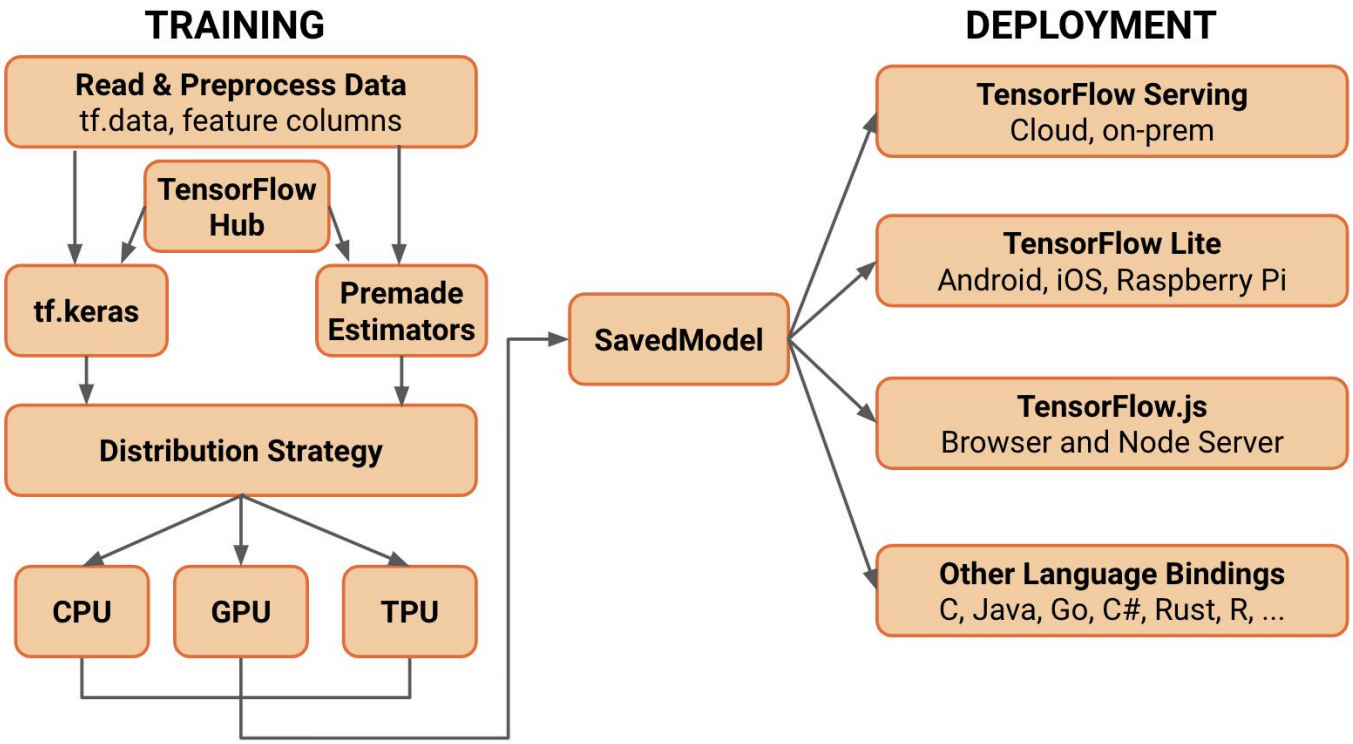
Batch Prediction



Real Time Prediction

How does the user use the model?

Model Deployment Options



Model Deployment: TensorFlow.js

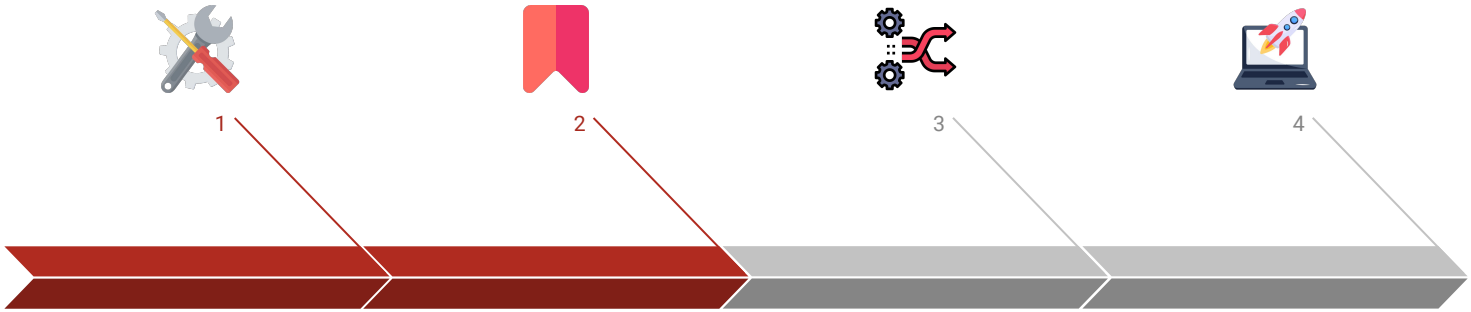
What is TensorFlow.js?

An **open-source JavaScript library** for training and deploying machine learning models in the **client's browser** or **Node.js server.**

- Run existing models
- Retrain existing models
- Develop ML with JavaScript

TensorFlow.js
Disadvantages

General Steps in TensorFlow.js



Build a Model

Build & train a ML model with TensorFlow

Save Model

Save model in SavedModel or .h5 format.

Convert

Convert it to JSON format

Deploy

Deploy the Converted model on a webpage

[Hands-on] Bridging Machine Learning with Web Development

bangkit

Model Deployment: TensorFlow Lite

What is TensorFlow Lite?

An open-source deep learning framework to **run TensorFlow models on-device**

- Optimized for on-device machine learning
- Multiple platform support
- Diverse language support
- Hardware acceleration & model optimization


We'll create
a TensorFlow Lite model
It will be easy right?
Easy right?
imgflip.com

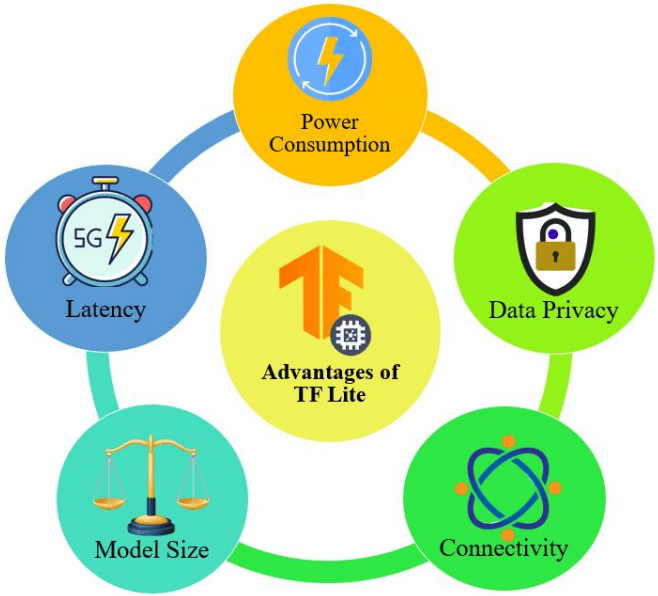
bangkit

The Challenges of on-Device Models

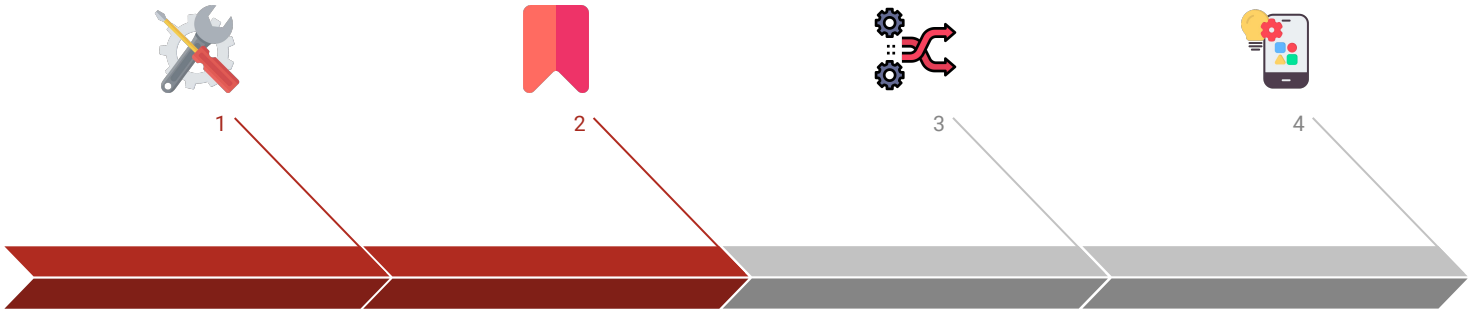
Running on-device models means we need to handle **diverse devices** & we do **not have access** to actual data.

Need to **balance** between:

- Power efficiency
- Inference latency
- Model accuracy & complexity



General Step in TensorFlow Lite



Build a Model

Build & train a ML model with TensorFlow

Save Model

Save model in SavedModel or .h5 format.

Convert

Convert it to TensorFlow Lite model

Deploy

Deploy the Converted model on a mobile/edge device

[Hands-on] Streamlining TensorFlow Lite Deployment for Mobile Applications

bangkit

Drive Video Demonstration

Break Time

Model Deployment: TensorFlow Serving

bangkit

What is TensorFlow Serving?

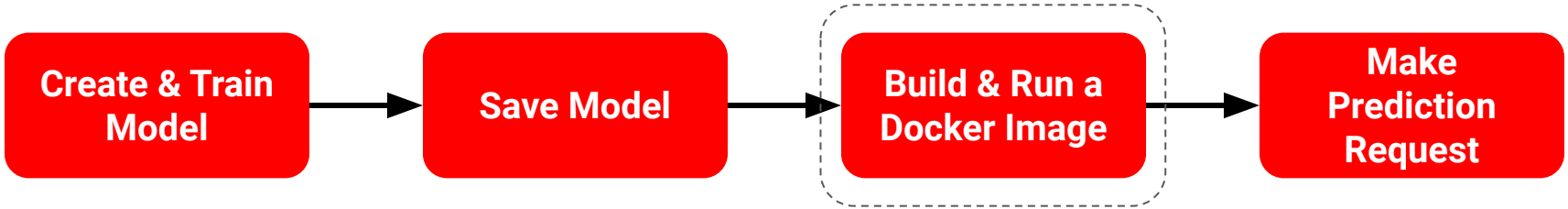
TensorFlow Serving is a **flexible** & **high-performance** serving system for machine learning models, designed for **production environments**

TensorFlow Serving

TensorFlow Serving allows us to have a **centralized model**

- Easy to manage model version
- Easy to manage hardware resources based on demand
- Can have multiple serving processes

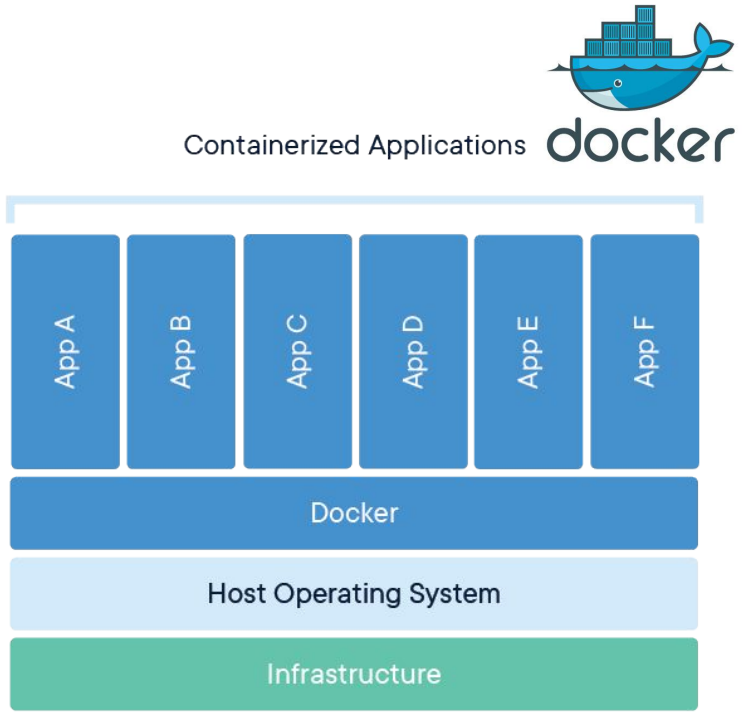
Deploy Model with TF Serving Workflow



What is Container? Docker?

Container is a standard unit of software that **packages up code** and **all its dependencies** so the application runs portably.

- Portable
- Lightweight
- Isolation



bangkit

Install Docker Desktop on Windows: https://docs.docker.com/desktop/windows/install/

How to Create Container?

[Hands-on] TensorFlow Serving Deployment in Cloud-native Environments

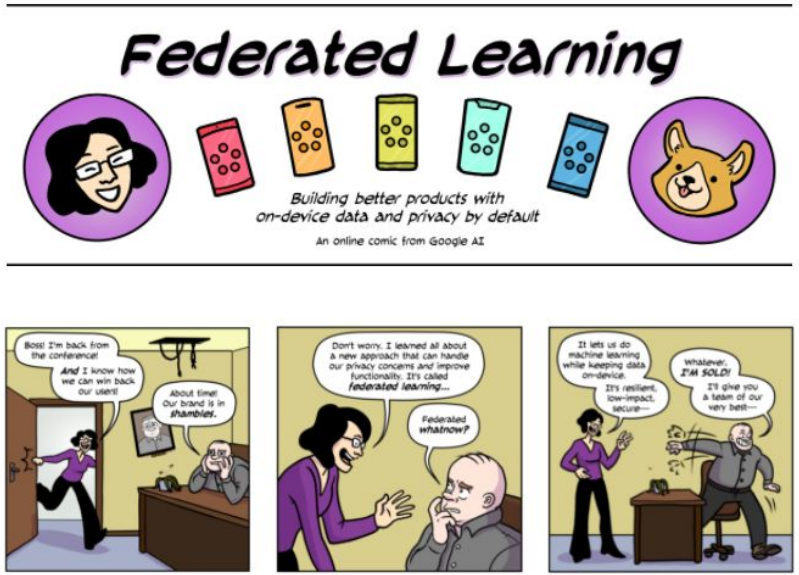
bangkit

Federated Learning

What is Federated Learning?

Federated learning allows each client **independently train** its own model using its own data right on the device.

- **Lower latency**
- **Less power consumption**
- **Ensuring privacy**



[Fun comic that explains the concept]

bangkit

Deployment Option Summary

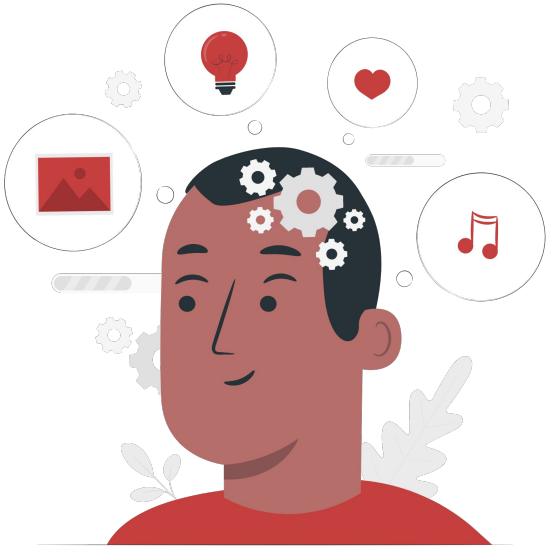
	TF JS	TF Lite	TF Serving	Federated Learning
Model runtime	Node.JS server / client's browser	On-device	Server	On-device
Computing power	Depends on usage	Low	High	Low
Latency	Depends on the model complexity	Low	Depends on the infrastructure and model complexity	Low
Model complexity	Depends on usage	Lighter	Heavier	Lighter
Need server connection	Anytime	One time only / once in a while update	Anytime	One time only / once in a while update
Privacy	Depends on model runtime	No need to send data to the server	Need to send data to the server	Ensuring user privacy

Introduction to TensorFlow Datasets

What is TensorFlow Datasets?

TensorFlow Datasets is a **collection of datasets ready to use**, with TensorFlow or other Python ML frameworks.

- **Simplicity & Performance**
- **Determinism/Reproducibility**
- **Customizability**



[Hands-on] Integrating TensorFlow Datasets into Your Machine Learning Workflow

bangkit

Quiz

Discussions

Conclusion

You have learned **all the options to deploy a machine learning model** using TensorFlow.

In the end, you also learn about **TFDS** and how to use it to build an **efficient data pipeline.**

bangkit

Thank You

bangkit