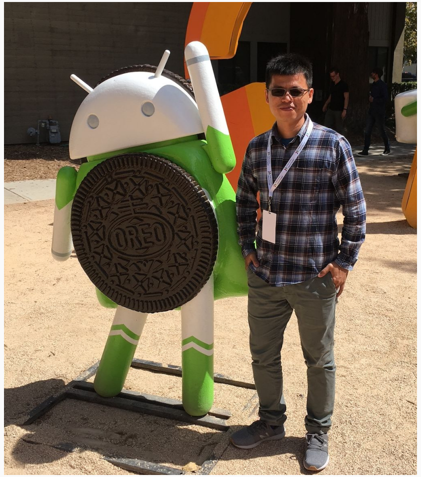
東海大學專題演講
題目: AI自動化部署架構

講者: 葉信和 先生
日期: 2021/05/12
Email: hsinho.yeh@footprint-ai.com

信誠金融科技股份有限公司
XINCHEN FINTECH CO., LTD.

About me

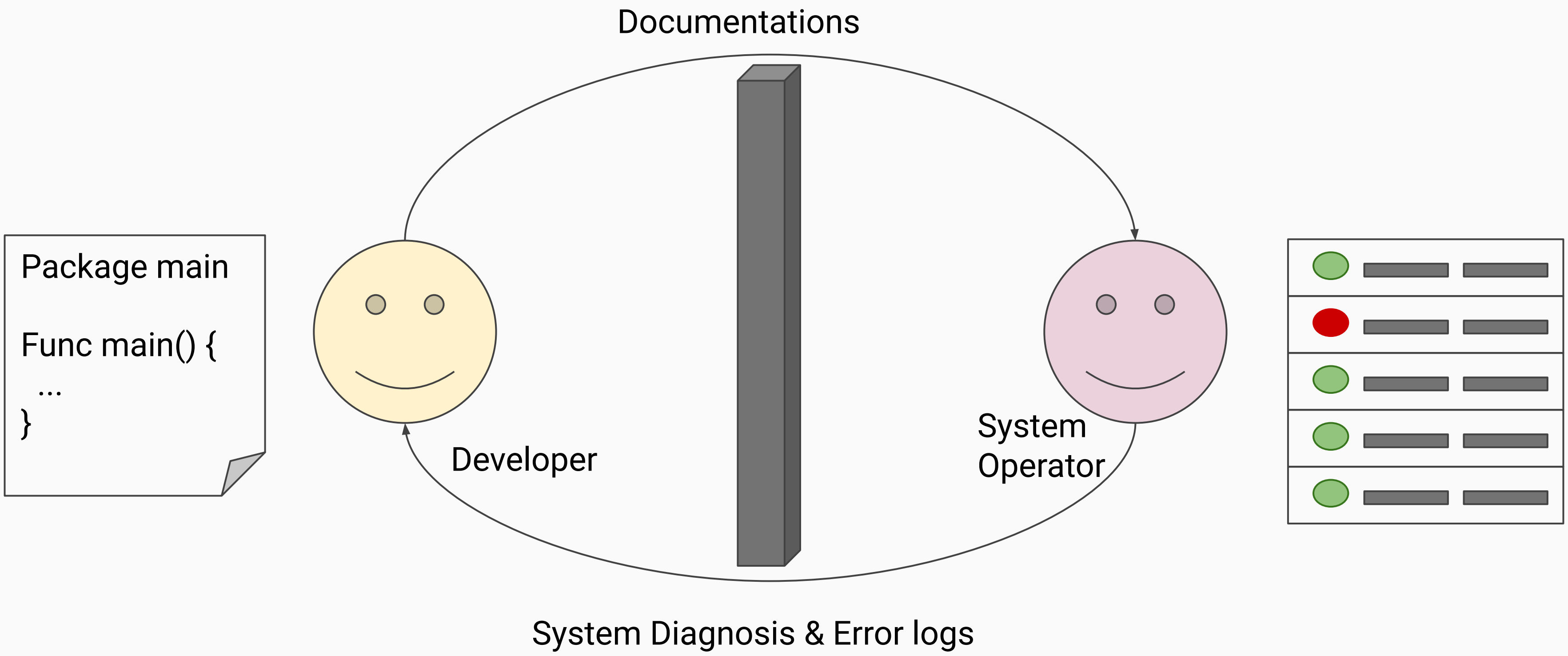
- 2020 - Present at 信誠金融科技
  - Tech solution provider for financial sectors
    - Deepselling: A deep analytics platform for ecommerce
    - Tintin: Everyone-can-use machine learning platform
- 2016 - 2020 at IglooInsure (16M+ in series A+ 2020)
  - Provide digital insurance for e-conomic world
  - Funded in KUL, Headquartered in Singapore
  - First employee/ Engineering Lead / Regional Head/ Chief Engineer
- 2013 - 2016 at Studio Engineering @ hTC
  - Principal Engineer on Cloud Infrastructure Team
- 2009 - 2012 at IIS @ Academia Sinica
  - Computer vision, pattern recognition, and data mining
- CS@CCU, CS@NCKU alumni



Agenda

- Why we need Deployment Automation?
- What is DevOps?
- What is MLOps?
- MLOps Architecture
- Q&A

Before DevOps: developers and system operators were in different teams...



Documentations

```
Package main

Func main() {
  ...
}
```

Developer

System Operator

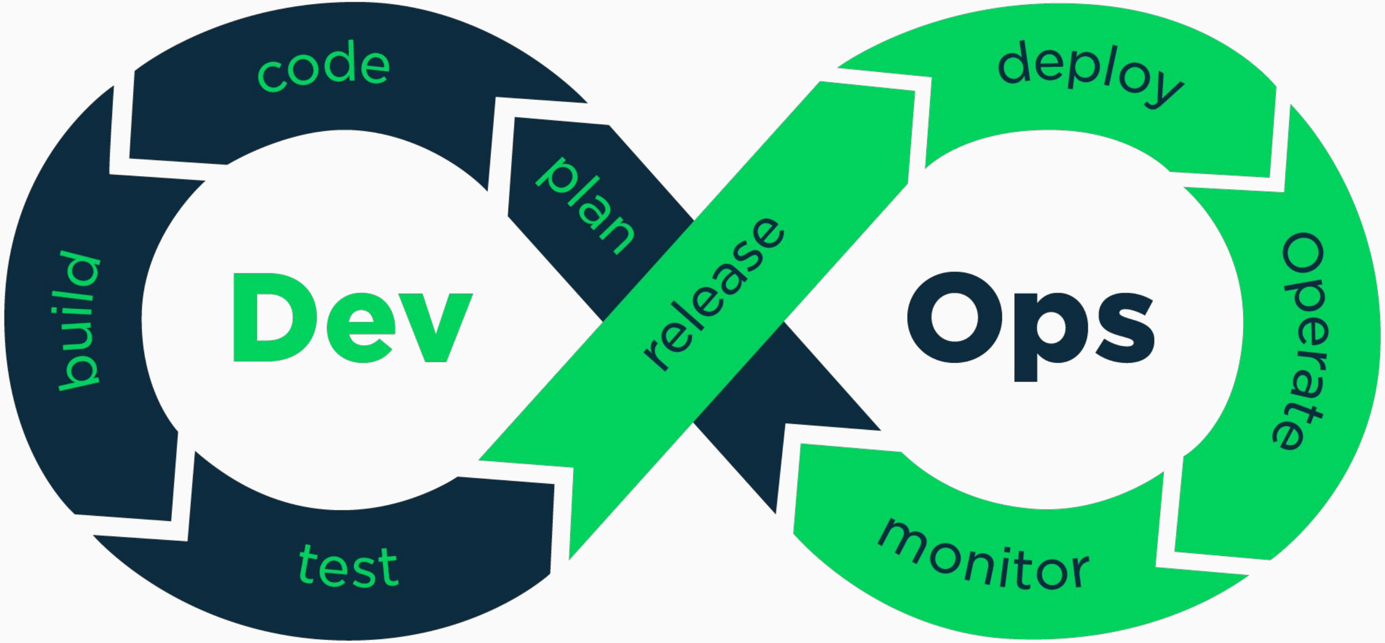
System Diagnosis & Error logs

Before DevOps: developers and system operators were in different teams...

- Pitfall
  - Environment
    - Developer: those codes works on MY machine
    - Operator: those codes are not working in production environment.
  - Existing Silos
    - Operation team keep complaining documentary is out-of-date but development team are too busy to update it…
  - Slow release cycle
    - Operator needs to take time to verify and carefully deploy into production

Inefficient coworking model, a finger pointing culture and blaming.

DevOps: 一個撒尿牛丸的概念



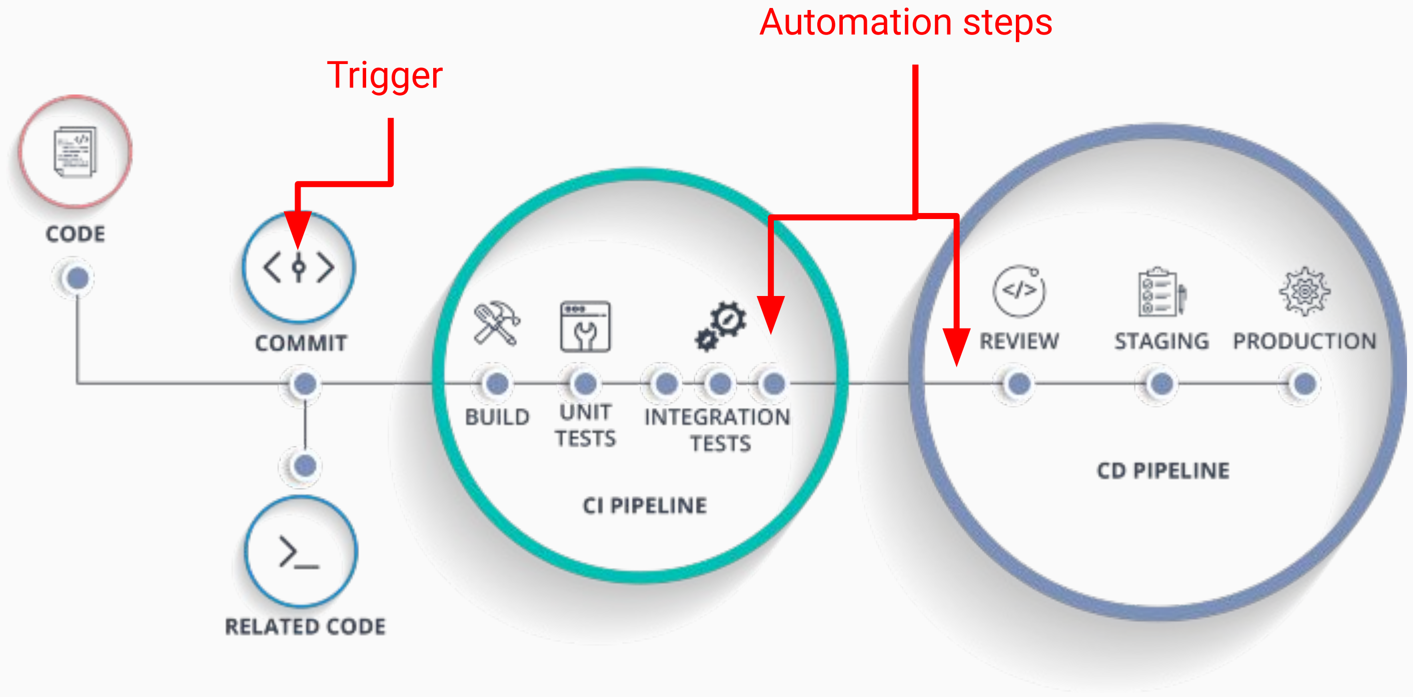
Ref: https://devopedia.org/devops

What is the definition of DevOps?

- Many definitions...
  - DevOps is development and operations collaboration.
  - DevOps is ops who think like devs and devs who think like ops
    - Developers need to learn how to create high-quality, production-ready software, and ops needs to learn that Agile techniques are actually powerful tools to enable effective, low-risk change management [1].
  - DevOps integrates developers and operations teams to improve collaboration and productivity by aiming **automation infrastructure**, workflows and **continuous improvement** product performance [2].

Ref:
[1] https://devopedia.org/devops
[2] https://www.youtube.com/watch?v=_I94-tJlovg&t=284s

What is Continuous Integration(CI)/Continuous deployment(CD) Pipelines



Ref: https://nanduribalajee.medium.com/what-is-ci-cd-pipeline-e2f25db99bbe

What is DA in Machine learning?

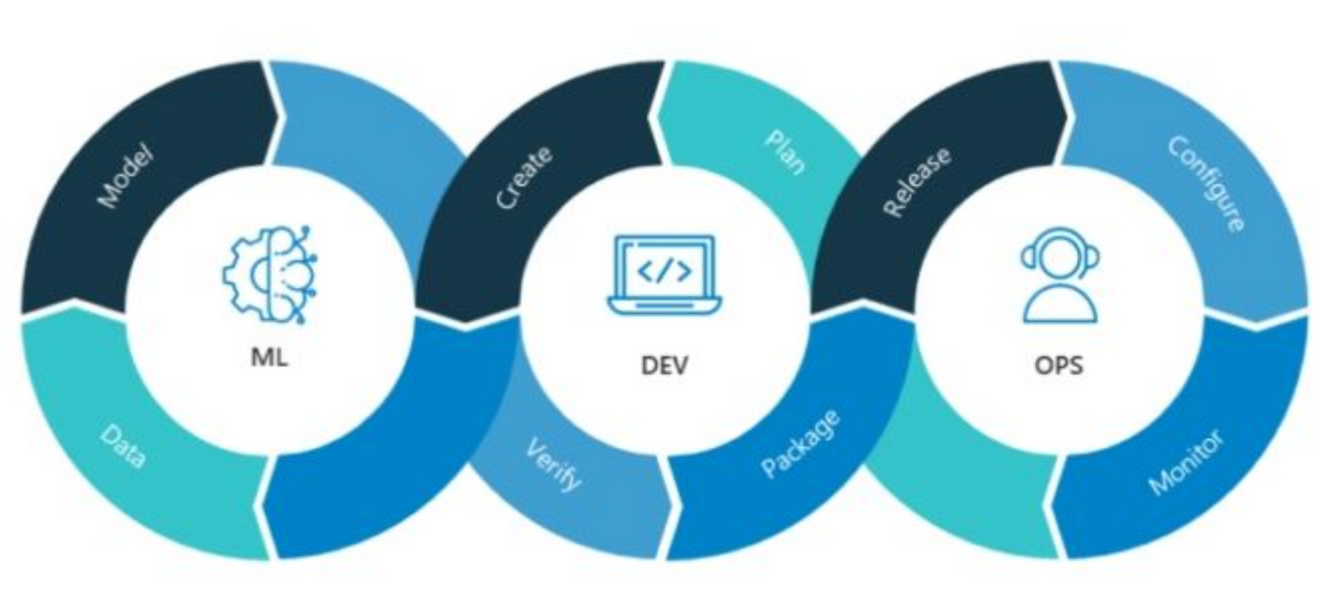
DevOps + ML = MLOps

MLOps is the process of taking an experimental Machine Learning model into a production system by including continuous development practice of DevOps in the software field.

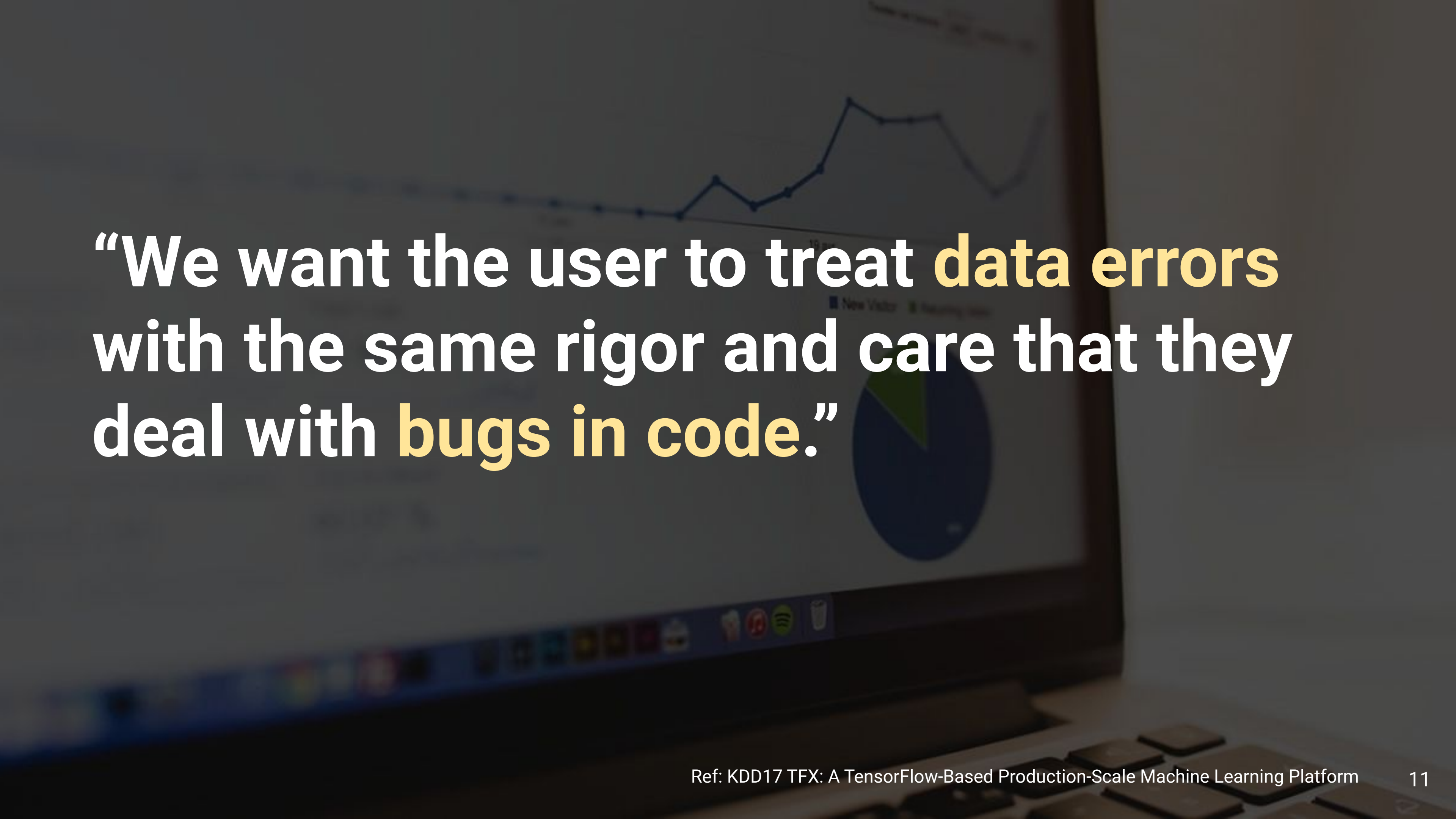
Ref: https://en.wikipedia.org/wiki/MLOps

Source: https://www.kubeflow.org/

MLOps



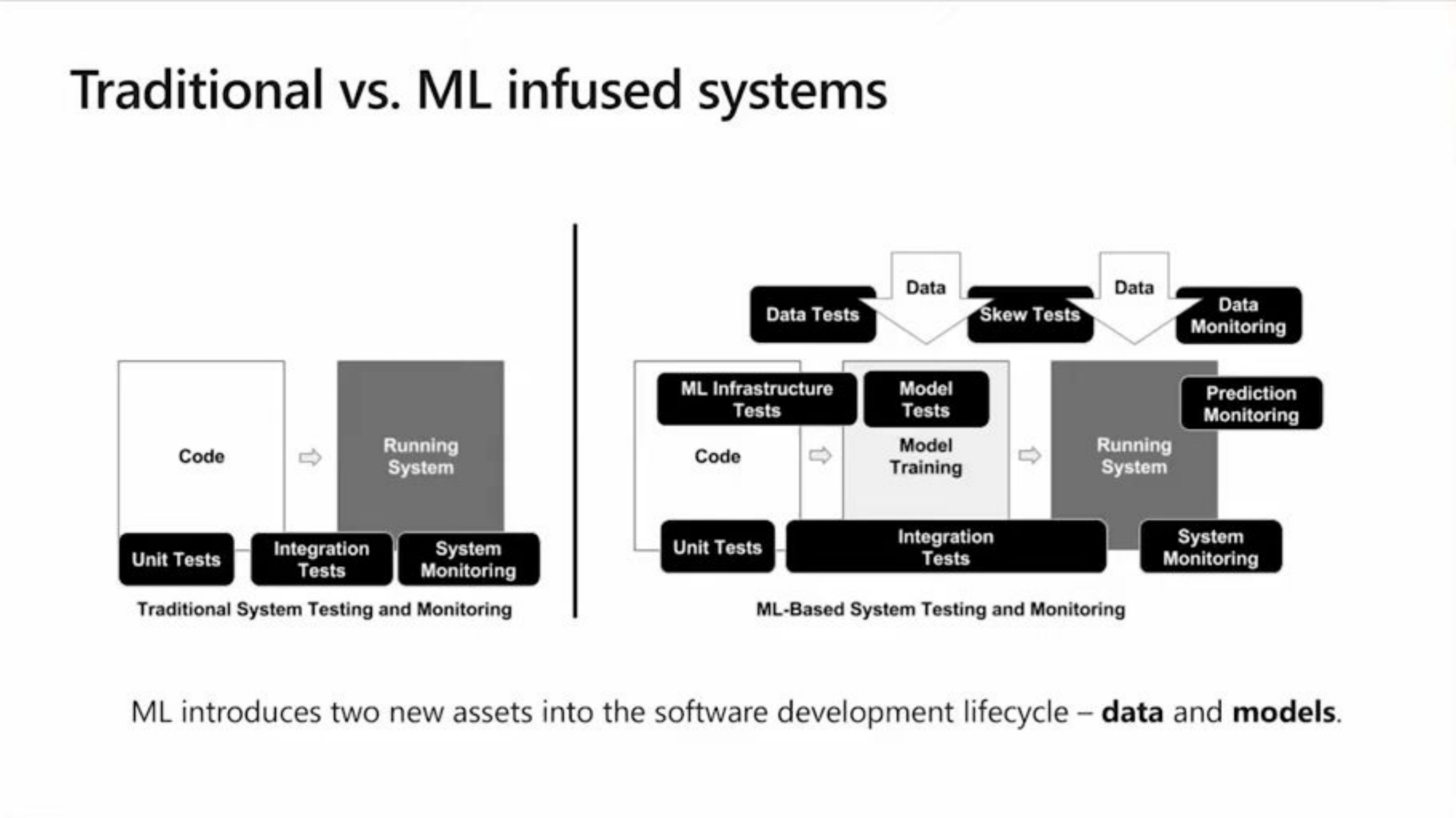
Ref: https://blogs.nvidia.com/blog/2020/09/03/what-is-mlops/



“We want the user to treat **data errors** with the same rigor and care that they deal with **bugs in code**.”

Ref: KDD17 TFX: A TensorFlow-Based Production-Scale Machine Learning Platform

How MLOps is different from DevOps (1/2)



Traditional vs. ML infused systems

Traditional System Testing and Monitoring

ML-Based System Testing and Monitoring

ML introduces two new assets into the software development lifecycle – **data** and **models**.

Ref: DevOps Enterprise Summit’ 19 https://www.youtube.com/watch?v=pqppGvTJm-A

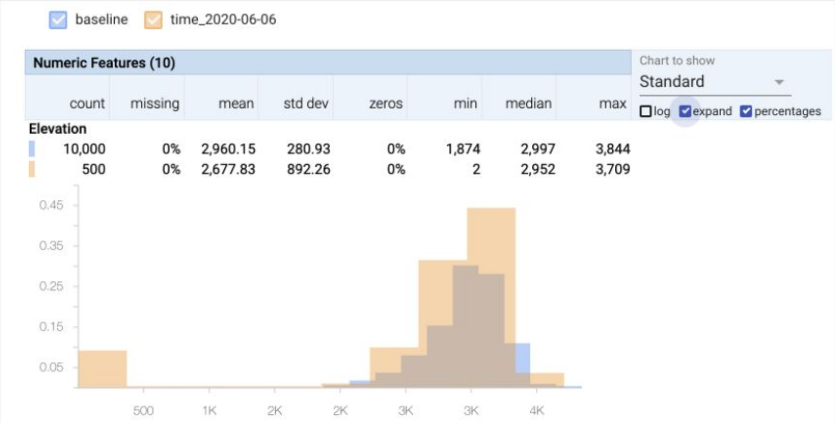
How MLOps is different from DevOps (2/2)

- Team Skills
  - DS(data scientists) and DR(data researcher) usually focused on data analysis, model deployment, and experimentation. May not have production-class experiences like SE(software engineer) do.
- Development
  - ML is experimental in nature, the challenge is tracking what worked (features/algorithms/model frameworks/parameters) and what didn't, and maintaining reproducibility while maximizing code reusability.
- Testing
  - Additional to software testing, data/model validation and model quality evaluation.
- Deployment
  - Not just deploy an offline-trained model to production, but requires a multi-step pipelines to retrain/deploy models as well as steps that are manually done by data scientists to train and validate new models.
- Production
  - Model could decay as the distribution of data could be drifting. You need to track summary statistics of your data and monitor the online performance and retrain if necessary.

Ref: https://cloud.google.com/architecture/mlops-continuous-delivery-and-automation-pipelines-in-machine-learning#devops_versus_mlops

Why we should care about drifting?

- Data drifting
  - A skew grows between training data and serving data.
  - The discrepancies between training data and serving data can usually be classified as schema skews or distribution skews
- Concept drifting
  - The interpretation of the relationship between the input predictors and the target feature evolves



☑ baseline ☑ time_2020-06-06

Numeric Features (10)

Chart to show: Standard

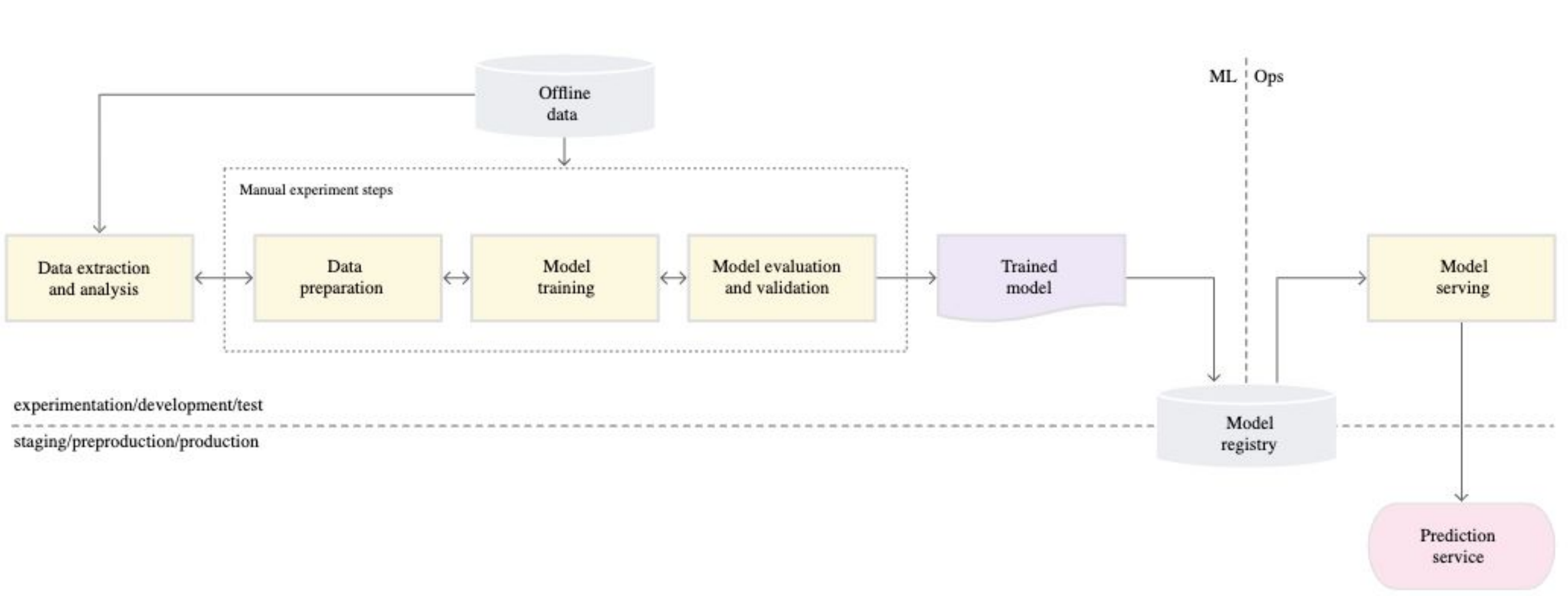
□ log ☑ expand ☑ percentages

	count	missing	mean	std dev	zeros	min	median	max
Elevation								
🟦	10,000	0%	2,960.15	280.93	0%	1,874	2,997	3,844
🟧	500	0%	2,677.83	892.26	0%	2	2,952	3,709

Ref: https://cloud.google.com/architecture/ml-modeling-monitoring-analyzing-training-server-skew-in-ai-platform-prediction-with-tfdv

Evolution of MLOps Architecture

MLOps Architecture: Manual process

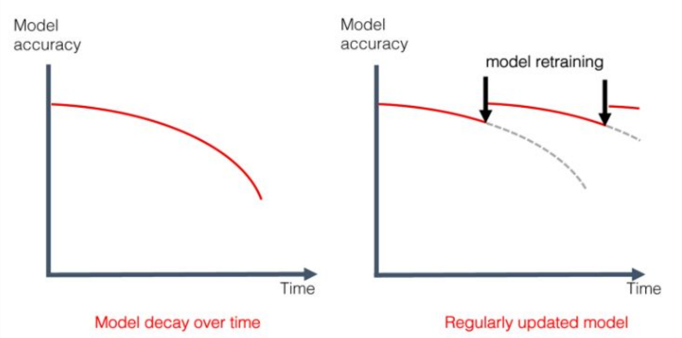


Ref: https://cloud.google.com/architecture/mlops-continuous-delivery-and-automation-pipelines-in-machine-learning#devops_versus_mlops

Manual process

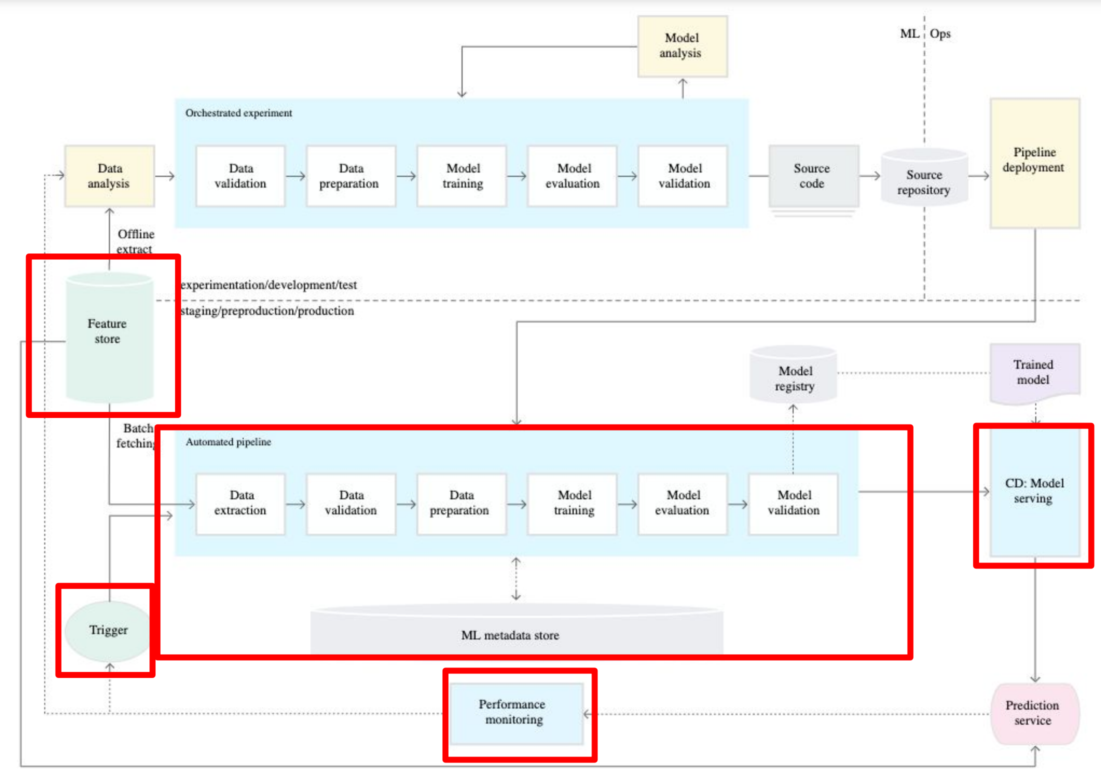
- In reality
  - The most common architecture to many businesses when they are beginning to apply ML.
  - The models fail to adapt to the drifting scenarios and the customers are the first person to spot the issue.

- Begging for improvement?
  - Monitor model quality
  - Frequently retrain your model
  - Continuous experiment



Ref: https://cloud.google.com/architecture/mlops-continuous-delivery-and-automation-pipelines-in-machine-learning
https://evidentlyai.com/blog/machine-learning-monitoring-data-and-concept-drift

MLOps Architecture: ML pipeline automation



Feature store provides an unify interface for accessing data from training/inference phase.

Automated pipeline is constructed for continuously experimenting new code with fresh data and delivery latest model.

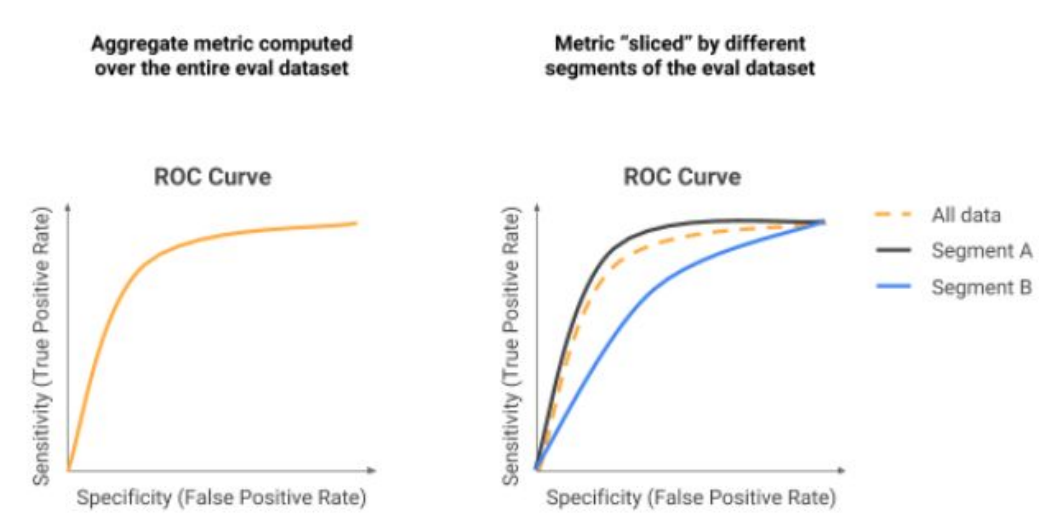
Performance monitoring keeps detecting performance degradation.

Metadata store keep tracks of code version and arguments for reproducibility.

Ref: https://cloud.google.com/architecture/mlops-continuous-delivery-and-automation-pipelines-in-machine-learning

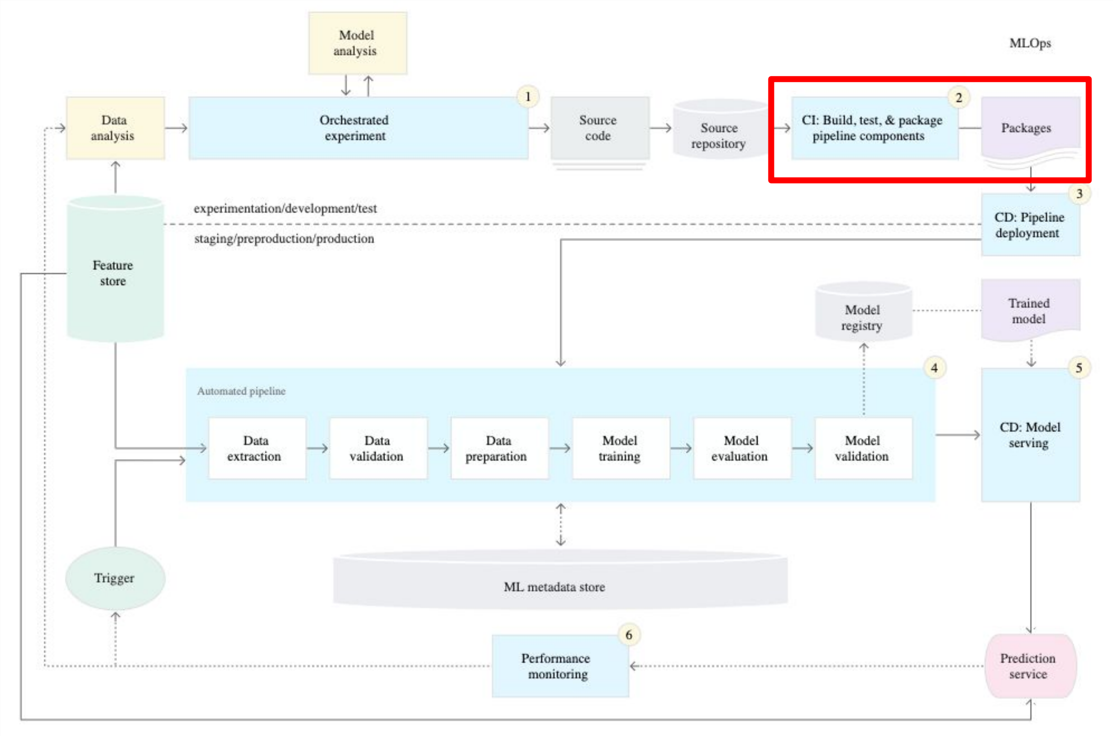
How to evaluate model quality? A TFMA viewpoint.

- Aggregate metric vs sliced metric
  - Slicing metrics allows us to analyze the performance of a model on a more granular level.
  - This enables us to identify slices where examples may be mislabeled, or where the model over- or under-predicts.



Ref: https://blog.tensorflow.org/2018/03/introducing-tensorflow-model-analysis.html

MLOps Architecture: CI/CD automation

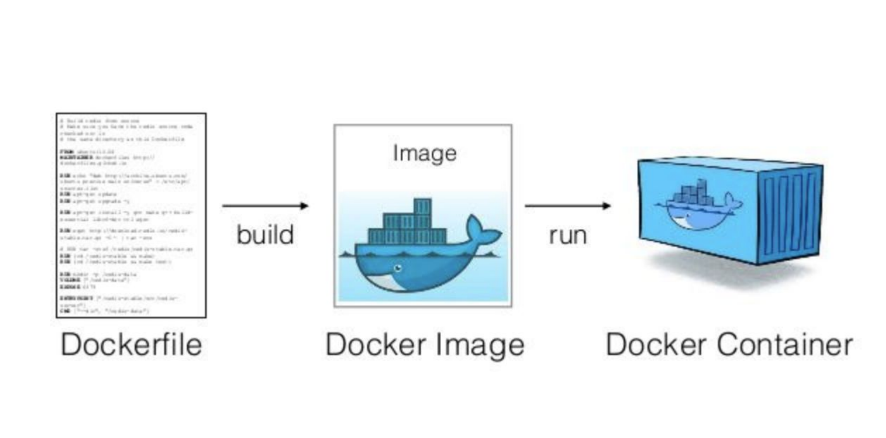


Continuous integration(CI) keeps build the latest source code, run various test cases (over/under fitting testing, model analysis), package pipeline components into deployable container.

Ref: https://cloud.google.com/architecture/mlops-continuous-delivery-and-automation-pipelines-in-machine-learning#devops_versus_mlops

What is Container?

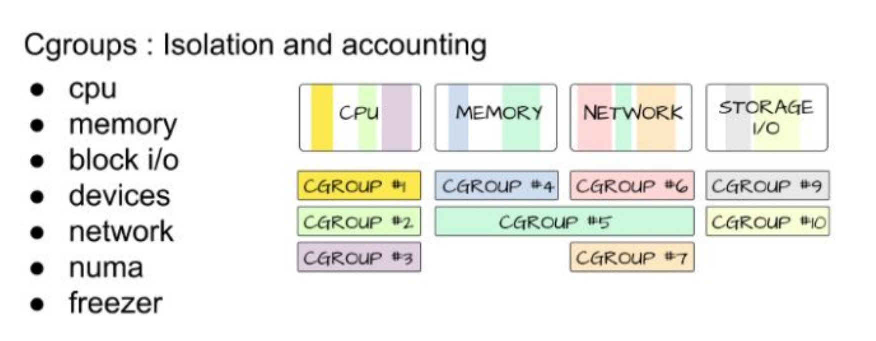
- Container Image
  - = Application code + dependencies
  - Runtime environment (cgroups, namespaces, env vars)
- Container Registry
  - Container repository



Ref: https://medium.com/platformer-blog/practical-guide-on-writing-a-dockerfile-for-your-application-89376f88b3b5

How container works?

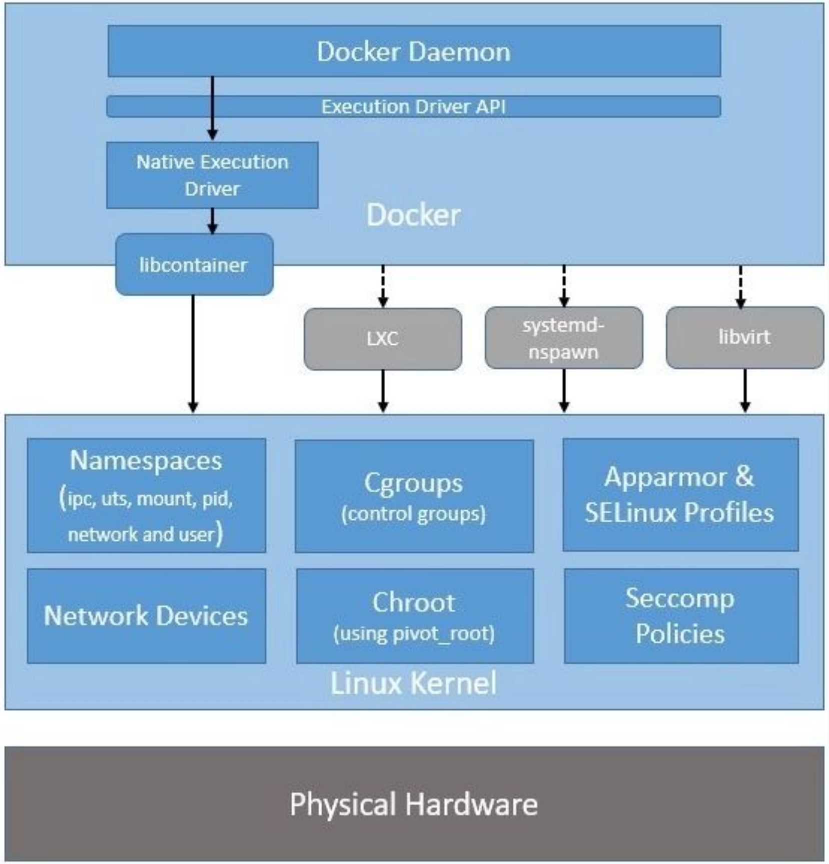
- Namespace for isolation
- Cgroups for resource limiting



Cgroups : Isolation and accounting

- cpu
- memory
- block i/o
- devices
- network
- numa
- freezer

Ref: https://www.baeldung.com/linux/docker-containers-evolution
https://medium.com/@BeNitinAgarwal/understanding-the-docker-internals-7ccb052ce9fe



Docker Daemon

Execution Driver API

Native Execution Driver

Docker

libcontainer

LXC

systemd-nspawn

libvirt

Namespaces (ipc, uts, mount, pid, network and user)

Cgroups (control groups)

Apparmor & SELinux Profiles

Network Devices

Chroot (using pivot_root)

Seccomp Policies

Linux Kernel

Physical Hardware

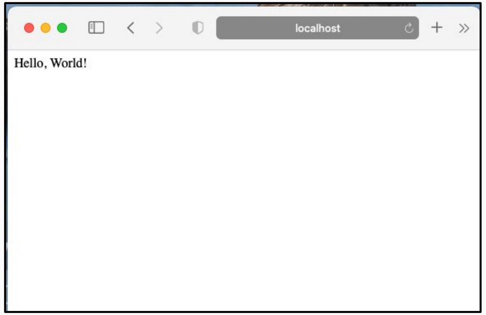
Minimize container example

Workshop link: k8s-workshop

```
FROM php:7.0-apache

COPY src/index.php /var/www/html/index.php

EXPOSE 80
```

Dockerfile

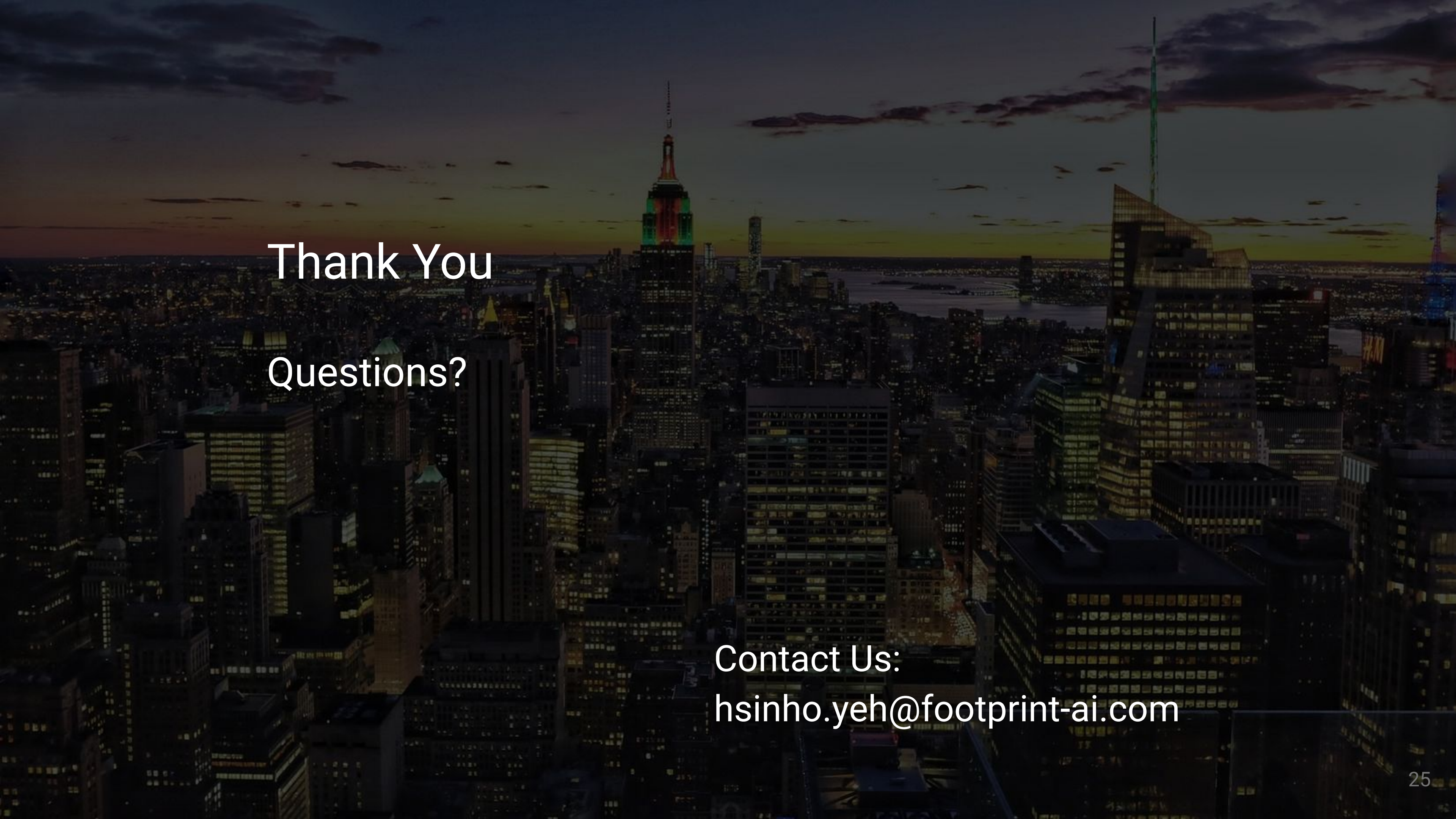
```
// build container image …
docker build -t phg-helloworld . -f Dockerfile

// run container image
docker run -p 80:80 phg-helloworld

// test php service
curl -vvv http://localhost
```



Hello, World!

Conclusion

- Deploy and serve machine learning models in production environments is easily to go wrong, you need an automated tool to simplify the flow.
- MLOps introduces highly automation tools/concepts to minimize errors from code, manual process, and data drifting, but comes with a cost (time/skill sets/...).
- A suitable solution is far better than a comprehensive solution.



Thank You

Questions?

Contact Us:
hsinho.yeh@footprint-ai.com