Superb AI

Introducing MLOps

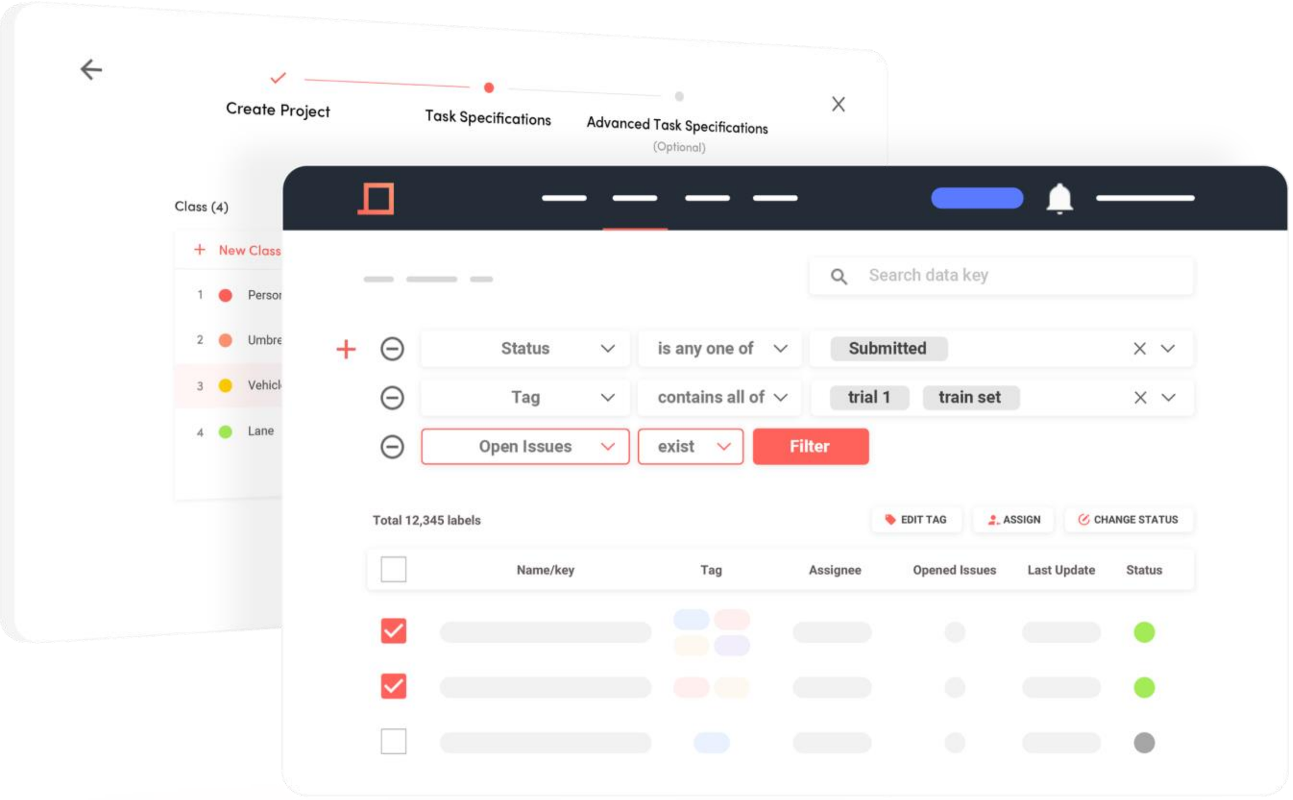
Date 10 08, 2020

Speaker Moonsu Cha

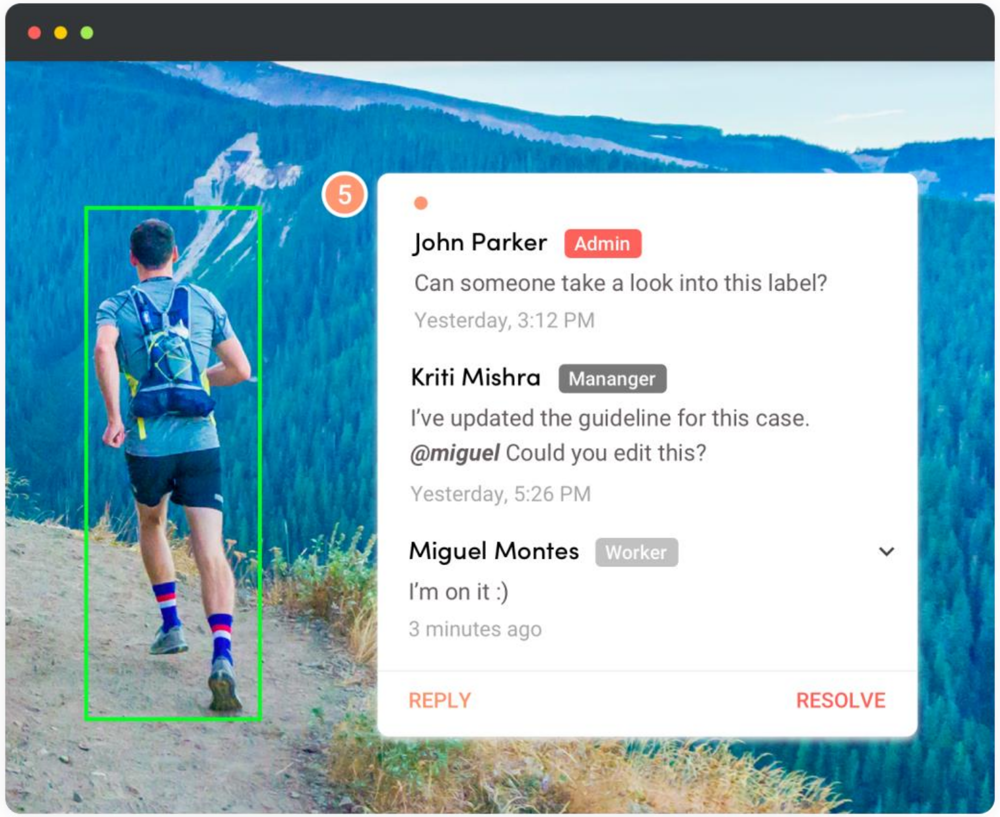
Co-founder, Machine Learning Engineer, Superb AI

CONTENTS

Superb AI



Create Project
Task Specifications
Advanced Task Specifications
(Optional)
Class (4)
+ New Class
1 Person
2 Umbre
3 Vehicl
4 Lane
Search data key
Status
is any one of
Submitted
Tag
contains all of
trial 1
train set
Open Issues
exist
Filter
Total 12,345 labels
EDIT TAG
ASSIGN
CHANGE STATUS
Name/key
Tag
Assignee
Opened Issues
Last Update
Status



5
John Parker
Admin
Can someone take a look into this label?
Yesterday, 3:12 PM
Kriti Mishra
Mananger
I've updated the guideline for this case.
@miguel Could you edit this?
Yesterday, 5:26 PM
Miguel Montes
Worker
I'm on it :)
3 minutes ago
REPLY
RESOLVE

Machine Learning Life Cycle

https://blog.superb-ai.com/machine-learning-training-data-workflows/

Machine Learning Life Cycle

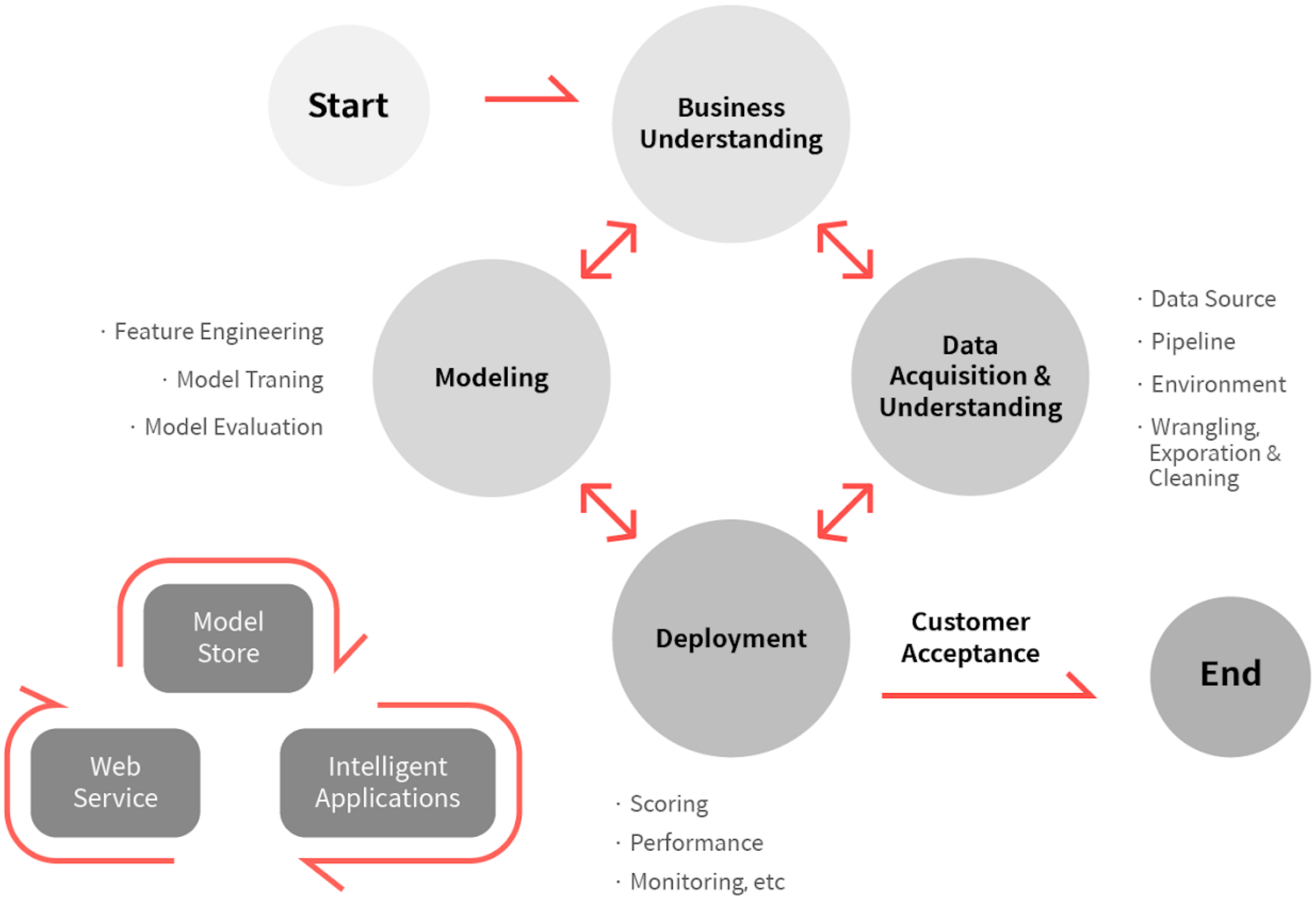


Image Credit: http://programmersought.com/article/250152726/

https://blog.superb-ai.com/machine-learning-training-data-workflows/

17 Key Questions to Ask Before Embarking on Your Next ML Project

Business Understanding

- What do you expect to gain from the machine learning system?
- What are the exact application scenarios and the expected business impact?
- Do you understand the expected performance and limitations of using an ML?
- How will you monitor and measure the performance of ML models?

Modeling

- What machine learning model will you use?
- What are your requirements for performance, in terms of computing speed (inference speed), accuracy, precision, and recall?
- What are your requirements for training? Will you rely on a cloud computing server? Will you continuously update your model?

Model Training

- After you build the dataset, what infrastructure will you use to train the models? In-house GPU servers or cloud servers?
- Do you need access to advanced hardware, i.e. TPUs?
- Will you train the models internally or will you use third-party model training services? Will you automate the hyper-parameter tuning and architecture search (Auto-ML)?

Data Acquisition & Understanding

- Do you have enough data to train your said model? If not, how will you collect additional data? Will you crowdsource, web-crawl, or purchase pre-made datasets?
- What are the legal implications of your data source? Are they copyrighted?
- Will you implement data augmentation techniques?
- Can your use-case resort to synthetically generated data?

Data Labeling

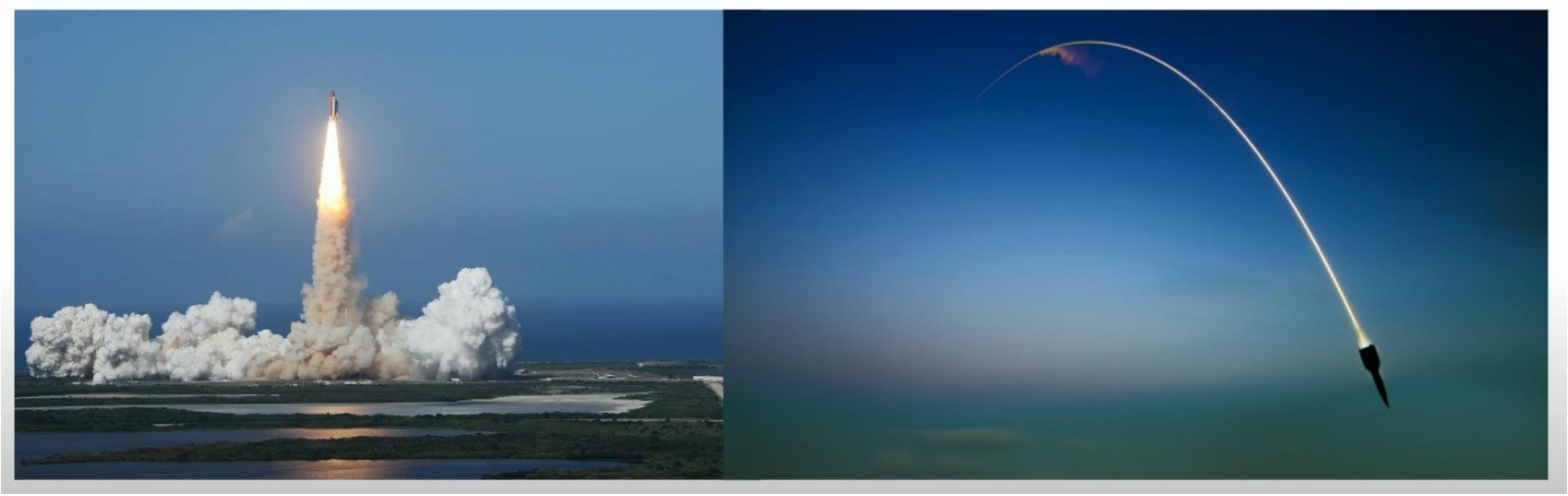
- Who will label your data? Do you have your own in-house data labeling team? Will you outsource to a labeling agency?
- Which data labeling tool will you use? Does it support whatever functions you need, such as visualization, statistics, version control, and multi-person collaboration?
- Will you use pre-trained machine learning models to speed-up the labeling process? If so, do you have access to training the said model? How often will you be re-training this model?

https://blog.superb-ai.com/machine-learning-training-data-workflows/

Machine Learning Component

- Configuration
- Data Collection
- Testing And Debugging
- Experiment Management
- Resource Management
- Model Versioning
- Data Versioning
- Data Verification
- ML Code
- Model Analysis
- Serving & Infrastructure
- Data Labeling
- Feature Engineering
- Process Management
- Monitoring
- Automation

Launching is easy, Operation is hard



What is ML Ops? Best Practices for DevOps for ML (Cloud Next '18)

Hidden Technical Debt in Machine Learning Systems

Complex Models Erode Boundaries

Data Dependencies Cost More than Code Dependencies

Feedback Loops

https://papers.nips.cc/paper/5656-hidden-technical-debt-in-machine-learning-systems.pdf, google, NIPS 2015

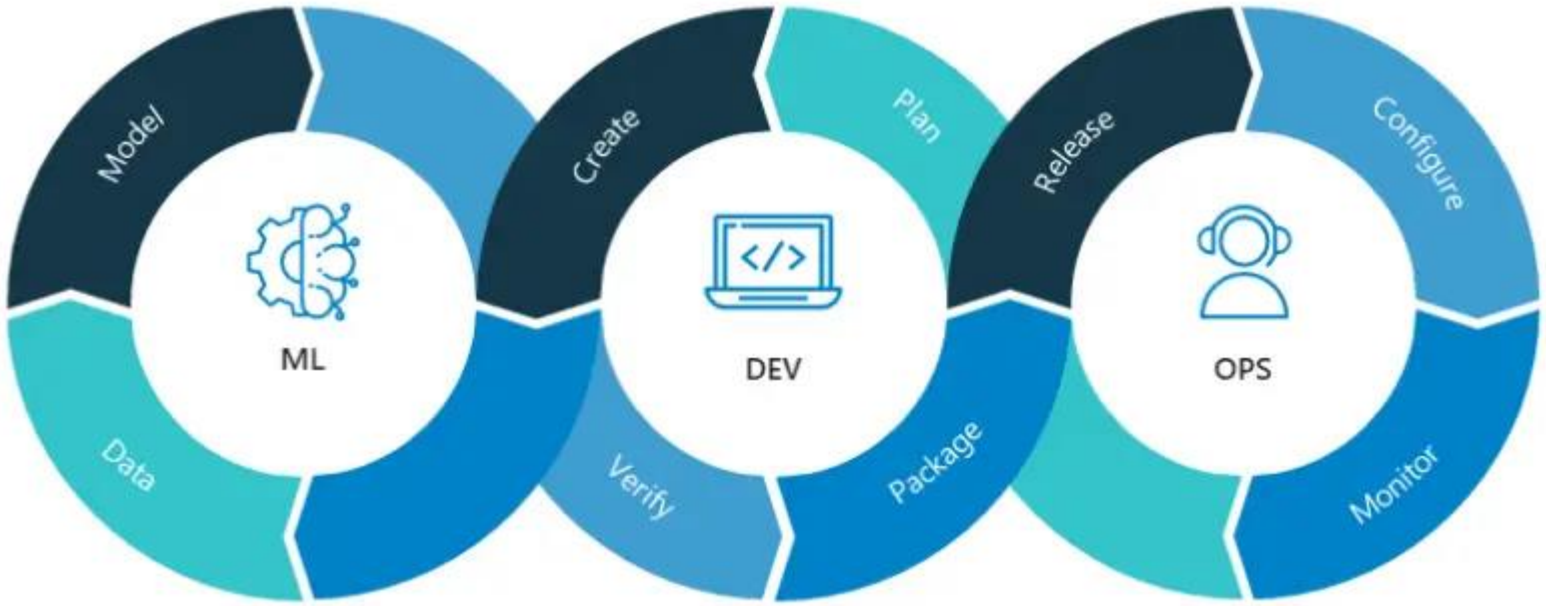
Hidden Technical Debt in Machine Learning Systems

ML-System Anti-Patterns

Glue Code	Pipeline Jungles	Dead Experimental Codepaths
Abstraction Debt	Multiple-Language Smell	Prototype Smell

https://papers.nips.cc/paper/5656-hidden-technical-debt-in-machine-learning-systems.pdf, google, NIPS 2015

MLOps



Wiki

MLOps (a compound of “machine learning” and “operations”) is a practice for **collaboration** and communication between **data scientists** and **operations professionals** to help manage production ML (or deep learning) lifecycle.

Google

MLOps is an ML engineering culture and practice that aims at **unifying** ML system development (Dev) and ML system operation (Ops). Practicing MLOps means that you advocate for **automation and monitoring** at all steps of ML system construction, including integration, testing, releasing, deployment and infrastructure management.

DevOps vs MLOps

A set of practices intended to reduce the time between committing a change to a system and the change being placed into normal production, while ensuring high quality

Team Skills

- 데이터 과학자 또는 ML 연구원을 포함
- 프로덕션 수준의 서비스를 빌드 가능한 소프트웨어 엔지니어 없을 수 있음

Testing

- 소프트웨어 시스템 테스트보다 더 복잡
- 데이터 검증, 학습된 모델 품질 평가, 모델 검증이 필요

Developments

- ML은 기본적으로 실험적임.
- 다양한 실험을 바탕으로 문제에 가장 적합한 것을 최대한 빨리 찾아야 함.

Deployment

- ML 시스템을 사용하면 모델을 자동으로 재학습시키고 배포하기 위해 다단계 파이프라인을 구성 필요

Production

- 지속적으로 진화하는 데이터 프로필로 인해 성능이 저하 가능
- 데이터의 요약 통계를 추적하고 모델의 온라인 성능을 모니터링 필수

CI (Continuous Integration)

- 더 이상 코드 및 구성요소만 테스트하고 검증하는 것이 아니라 데이터, 데이터 스키마, 모델도 테스트하고 검증 필요

CD (Continuous Delivery)

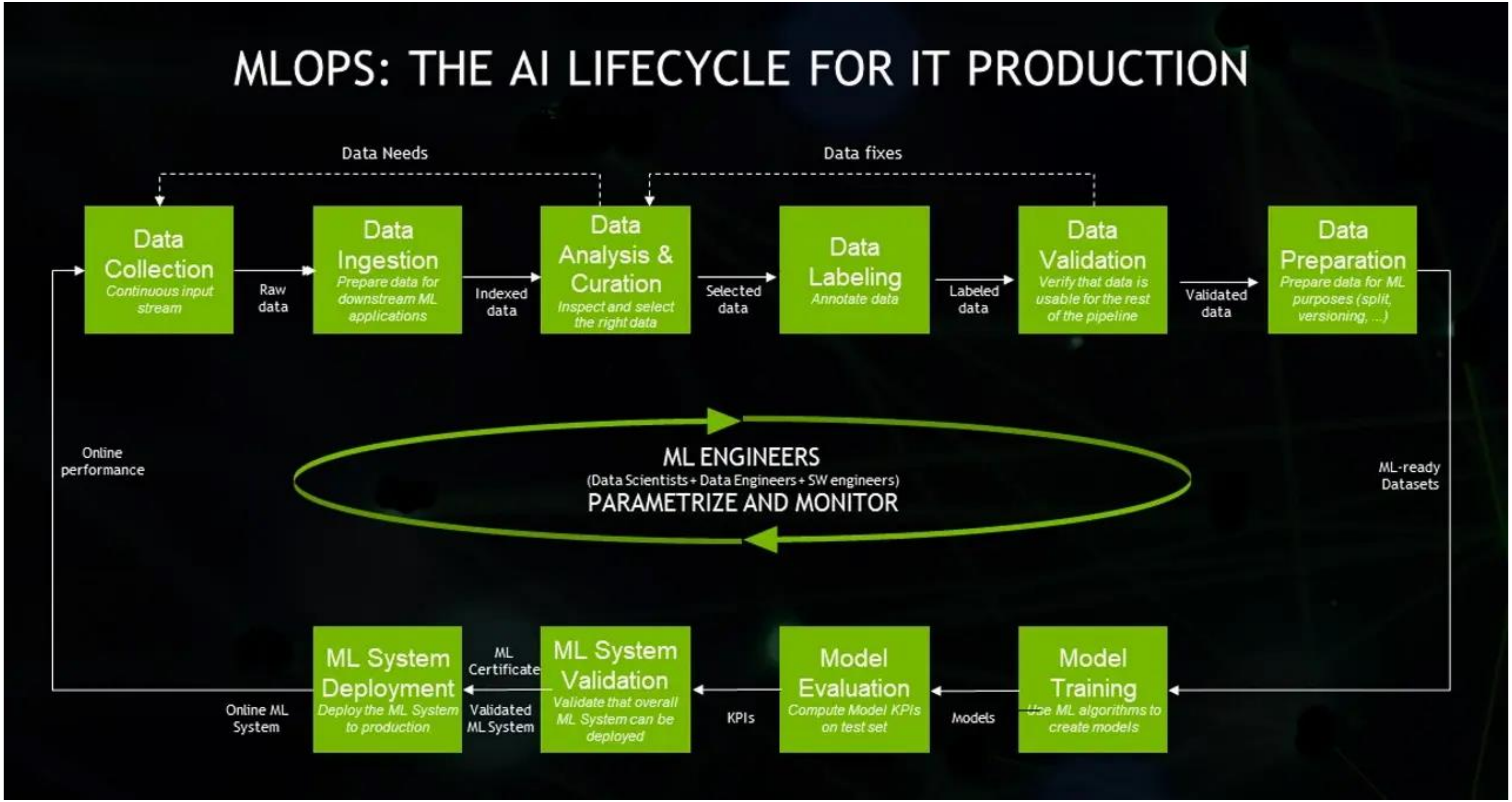
- 더 이상 단일 소프트웨어 패키지 또는 서비스만이 아니라 다른 서비스(모델 예측 서비스)를 자동으로 배포해야 하는 시스템(ML 학습 파이프라인)

CT (Continuous Training)

- ML 시스템에 고유한 새 속성으로, 모델을 자동으로 재학습 필요

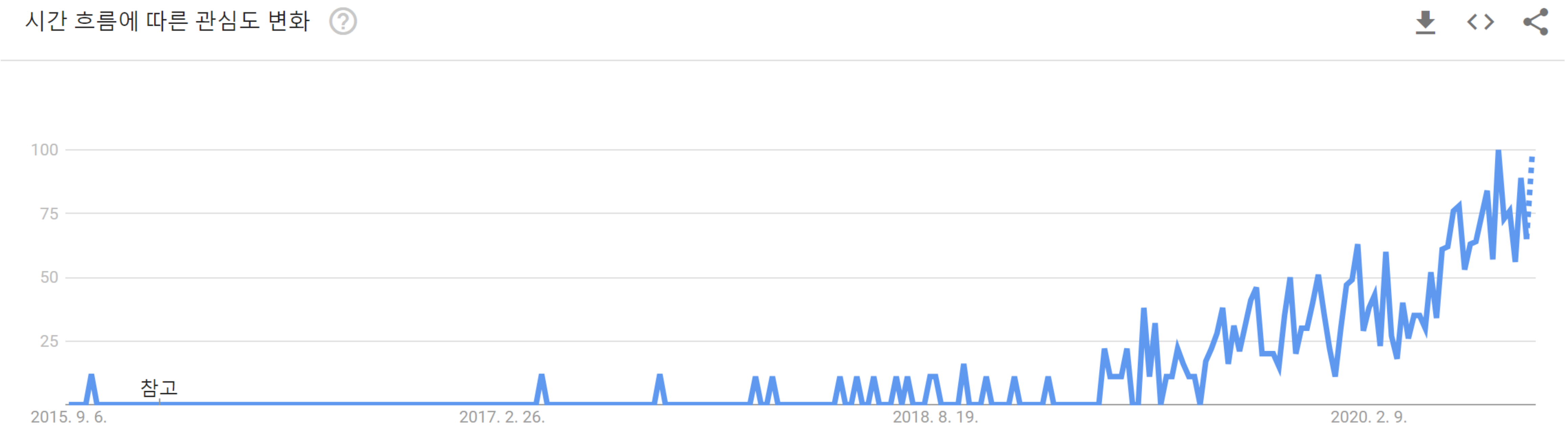
https://cloud.google.com/solutions/machine-learning/mlops-continuous-delivery-and-automation-pipelines-in-machine-learning

MLOps Life Cycle



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

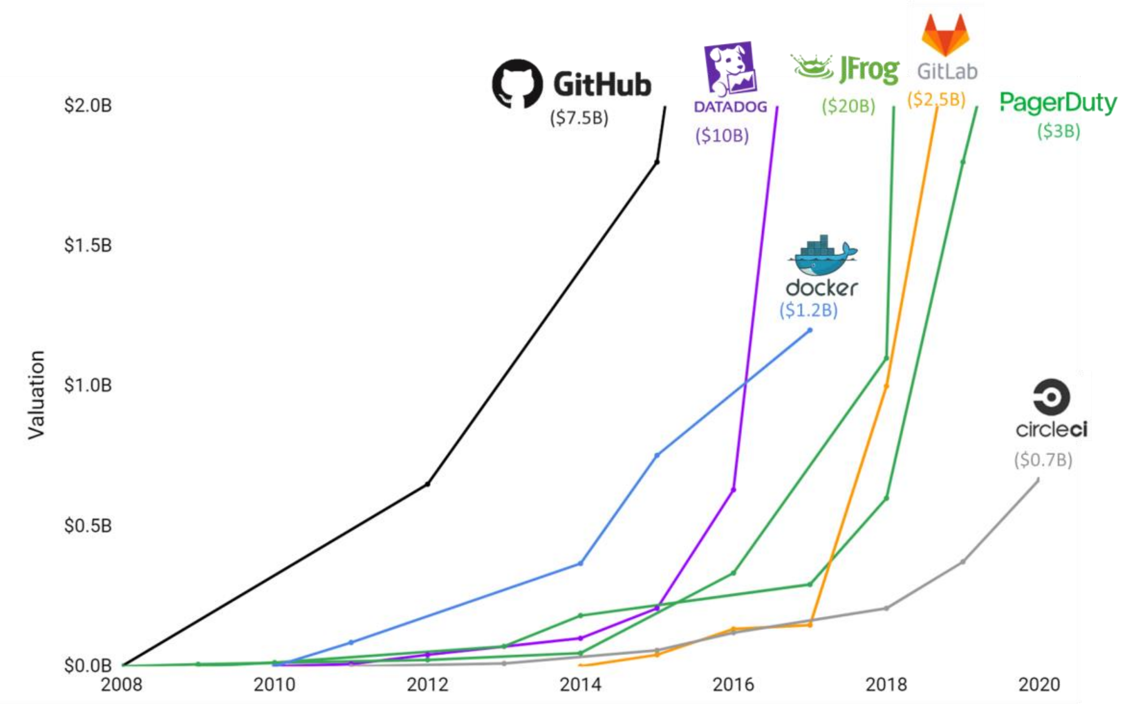
Market– Google Trends



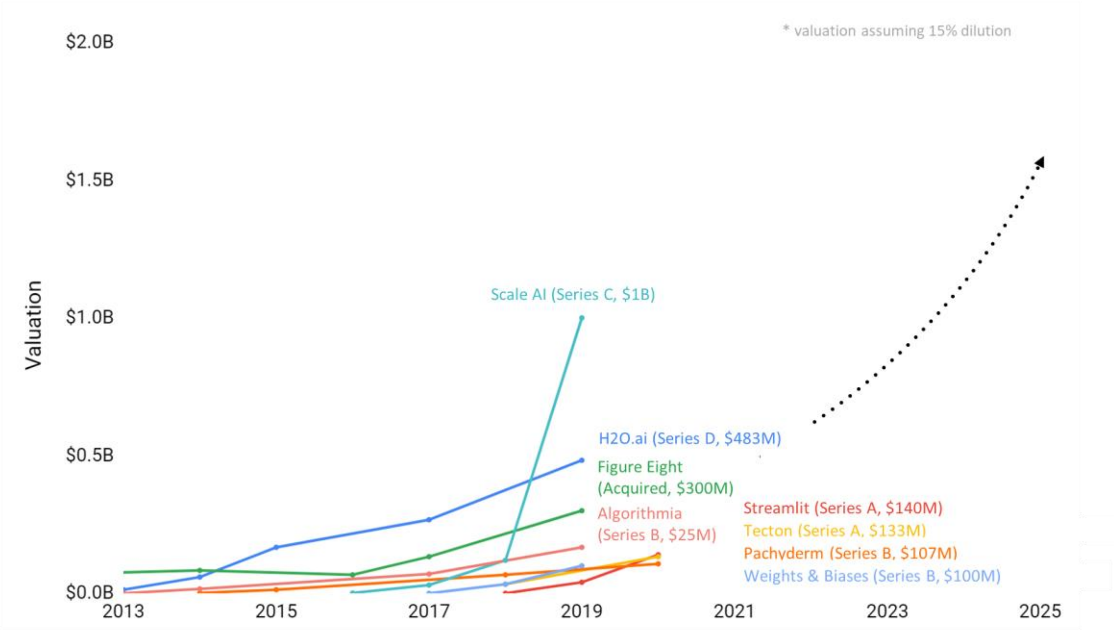
https://trends.google.com/trends/explore?date=today%205-y&q=MLOps

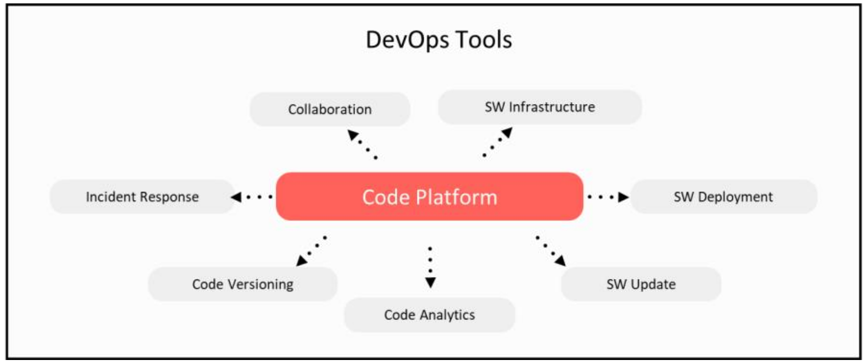
Market

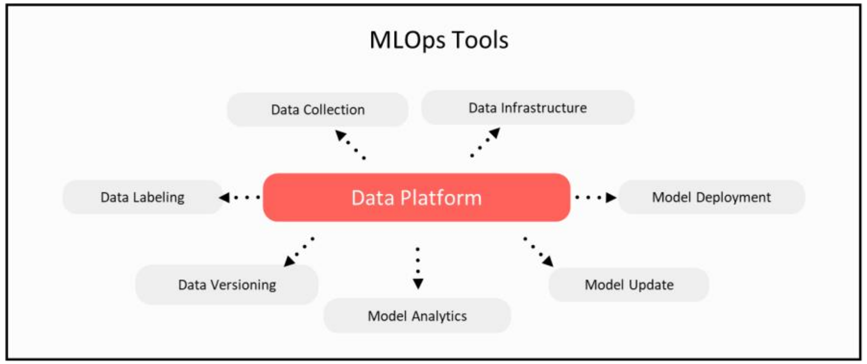
(* SW engineering tools that enable faster SW development and delivery)



(** ML engineering tools that enable faster ML development and delivery)



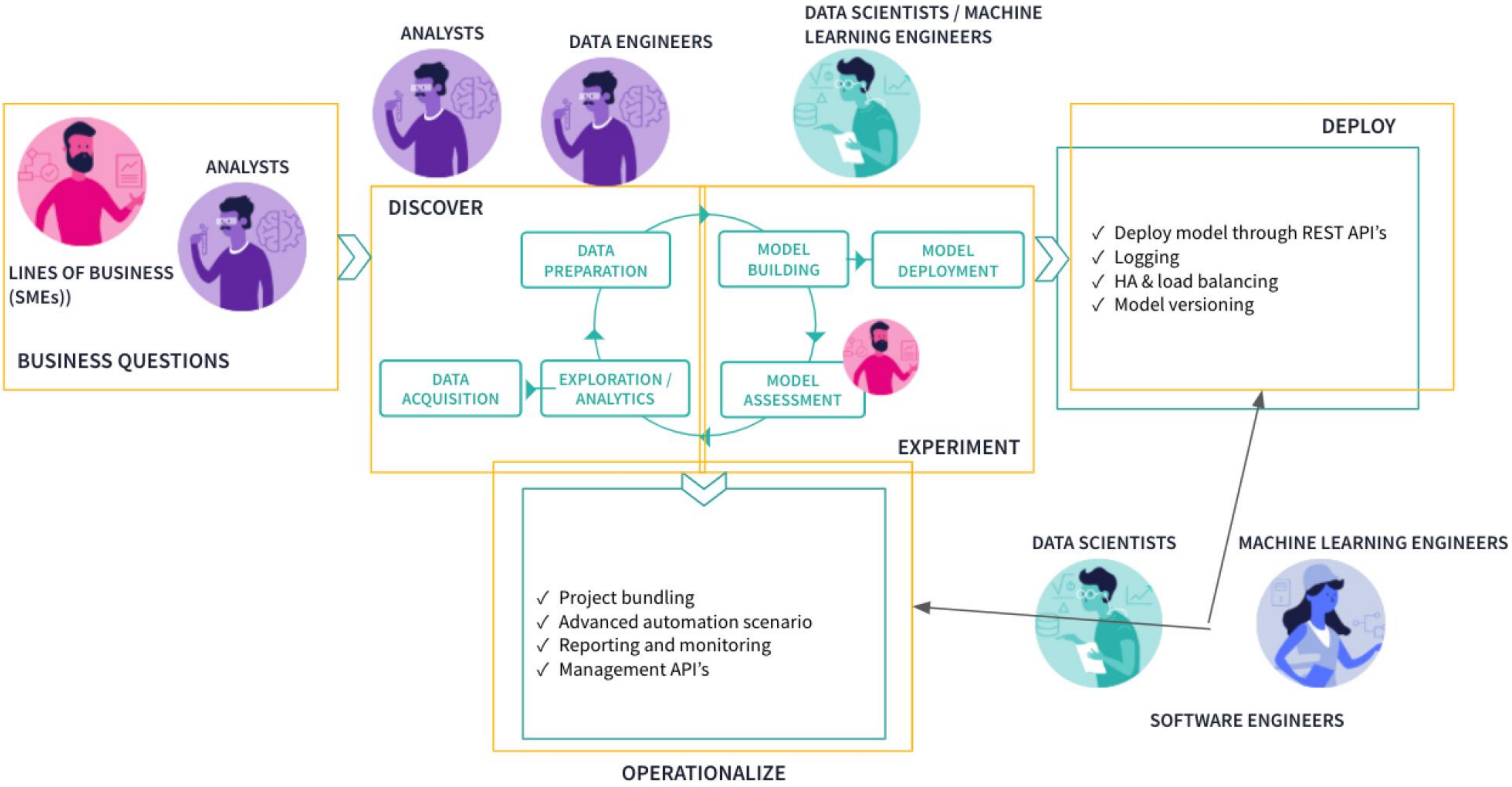




Market

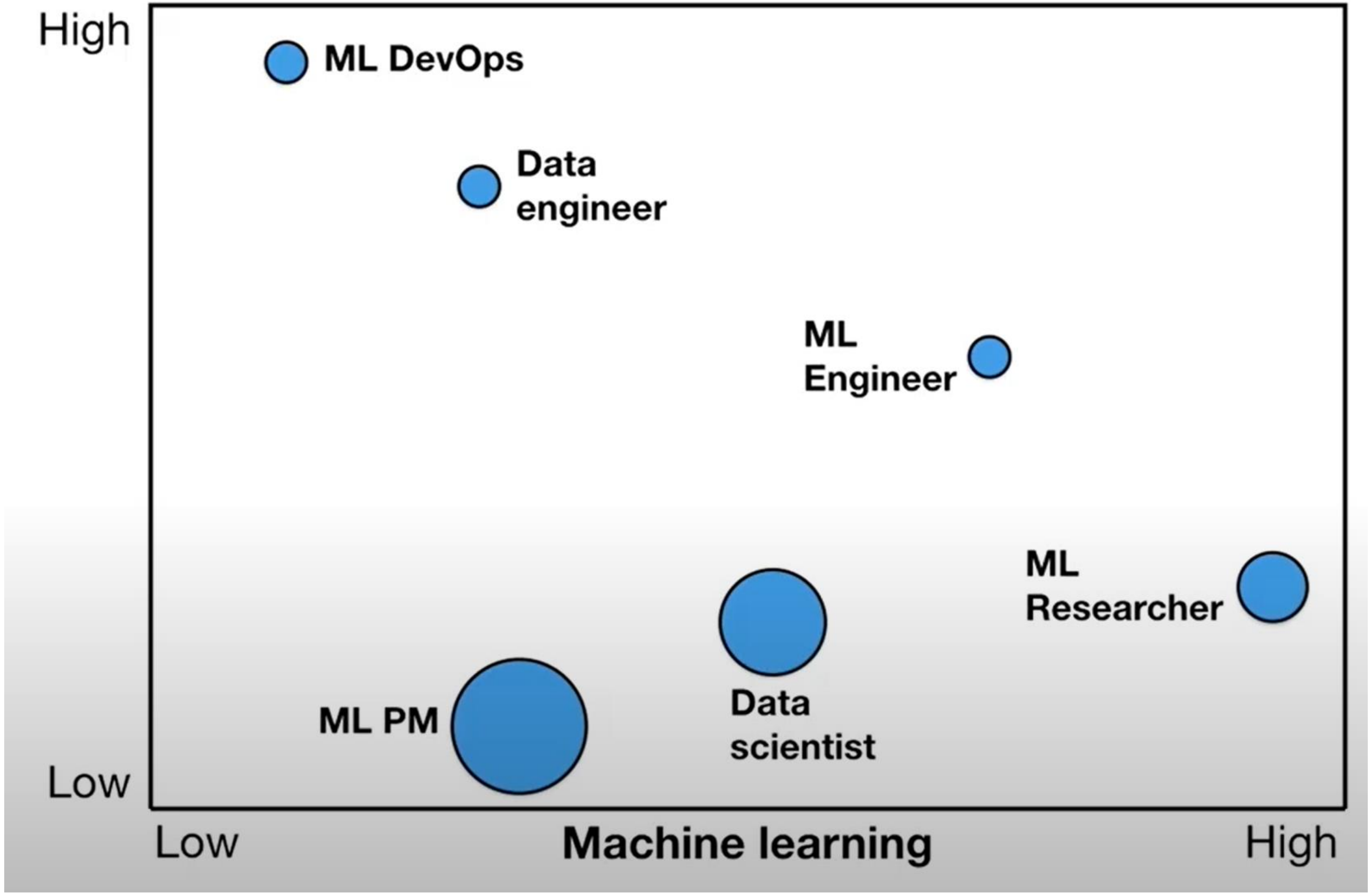
	Data Labeling	Data Platform	Model Management
Market Size Growth Rate	$1B (2020) 15% CAGR	$1B (2020) 30% CAGR	$1.6B (2020) 19% CAGR
Market Maturity	High	Low	Medium
Notable Players	Figure-Eight (Acquired by Appen, 2019) Mighty AI (Acquired by Uber, 2019) Scale AI (1B+ Value, 2019)	SuperAnnotate ($3M Seed, 2020) Platform AI ($2M Seed, 2018) Labelbox ($25M Series B, 2020) Scale-Nucleus (launched in 2020 Q3)	Model Train: Comet AI ($4.5M Series A, 2020) Model Deploy: Algorithmia ($25M Series B, 2019) Monitor: Weights & Biases ($15M Series B, 2019)

People



https://pages.dataiku.com/oreilly-introducing-mlops

People



https://course.fullstackdeeplearning.com/course-content/ml-teams/roles

Key Features

Business Objective

Governance

Data Collection

Data Labeling

Data Versioning

Data Analysis

Model Training

Model Versioning

Experiment Management

Model Evaluation

Model Serving

Model Quantization

CI/CD

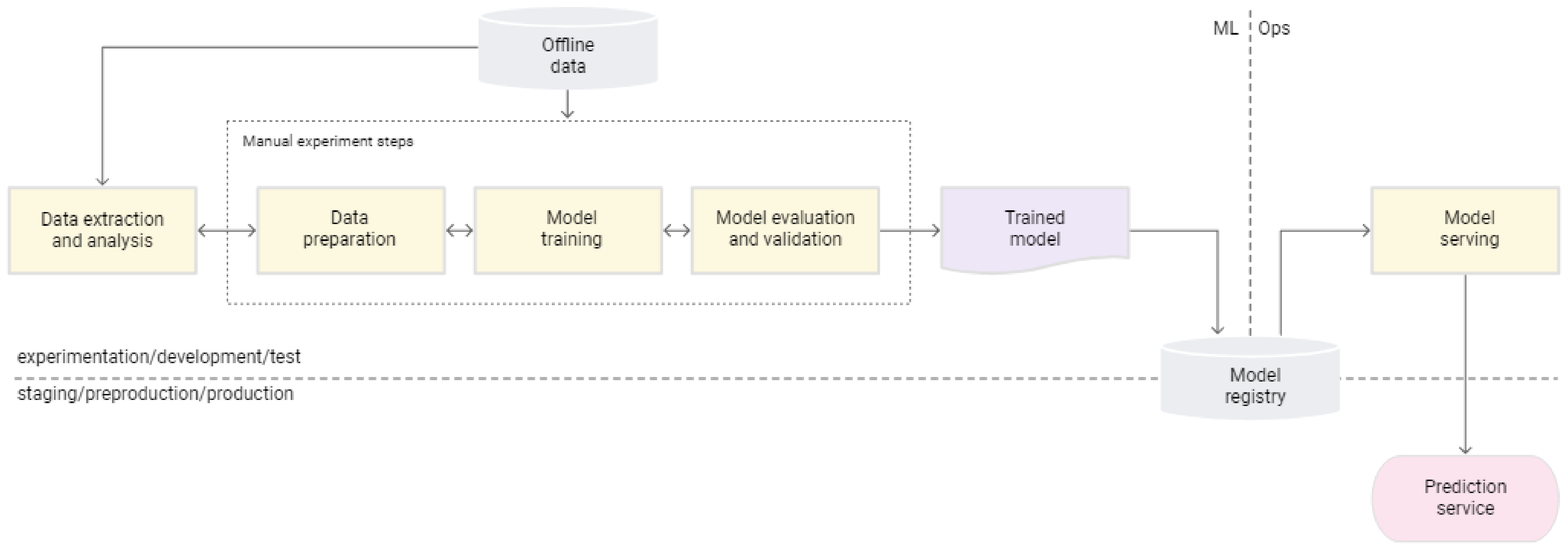
Monitoring (Performance Measure)

Iteration – Continuous Training (CT)

Feedback Loop

Pipelines

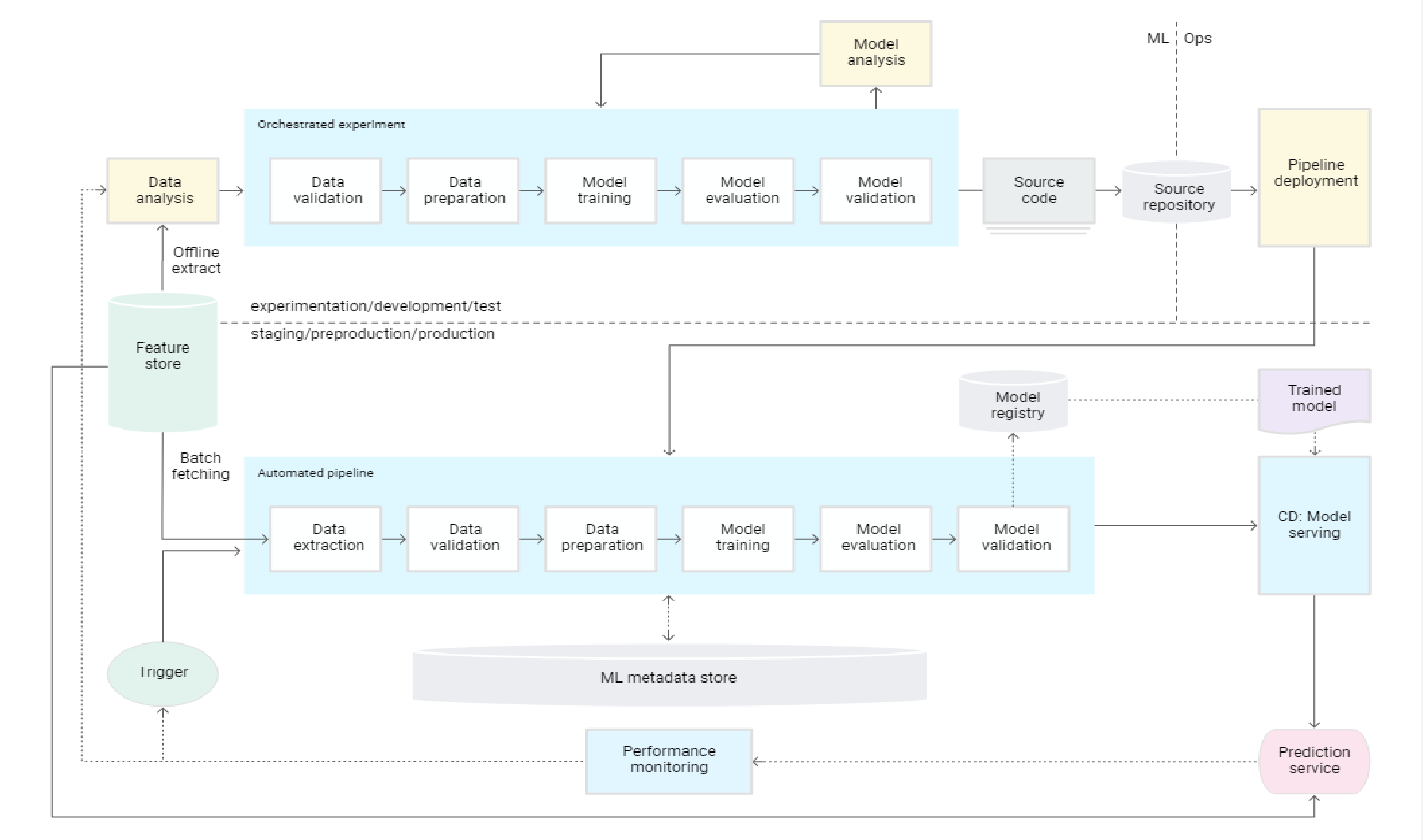
MLOps level 0: Manual process



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

Pipelines

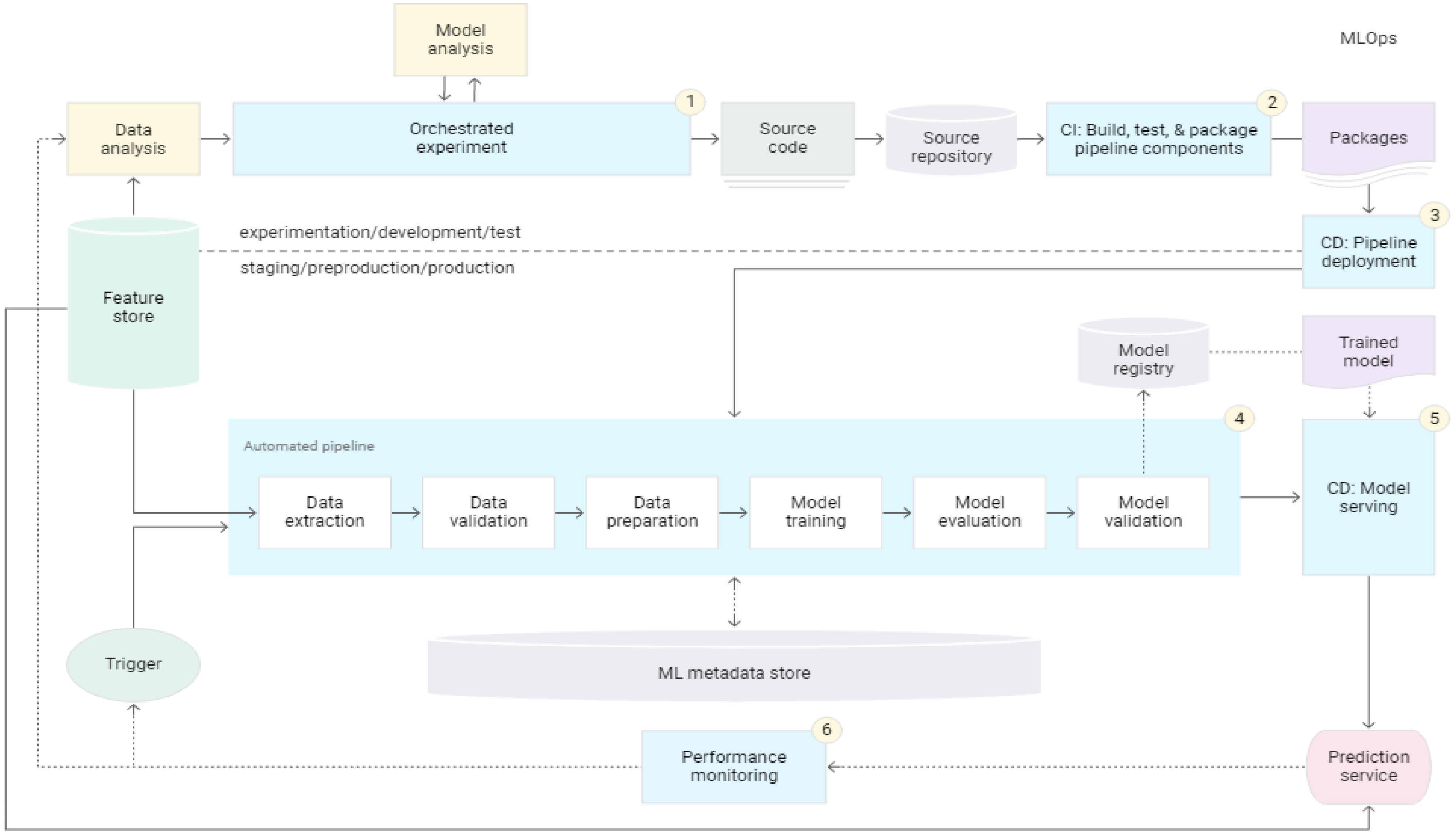
MLOps level 1: ML pipeline automation



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

Pipelines

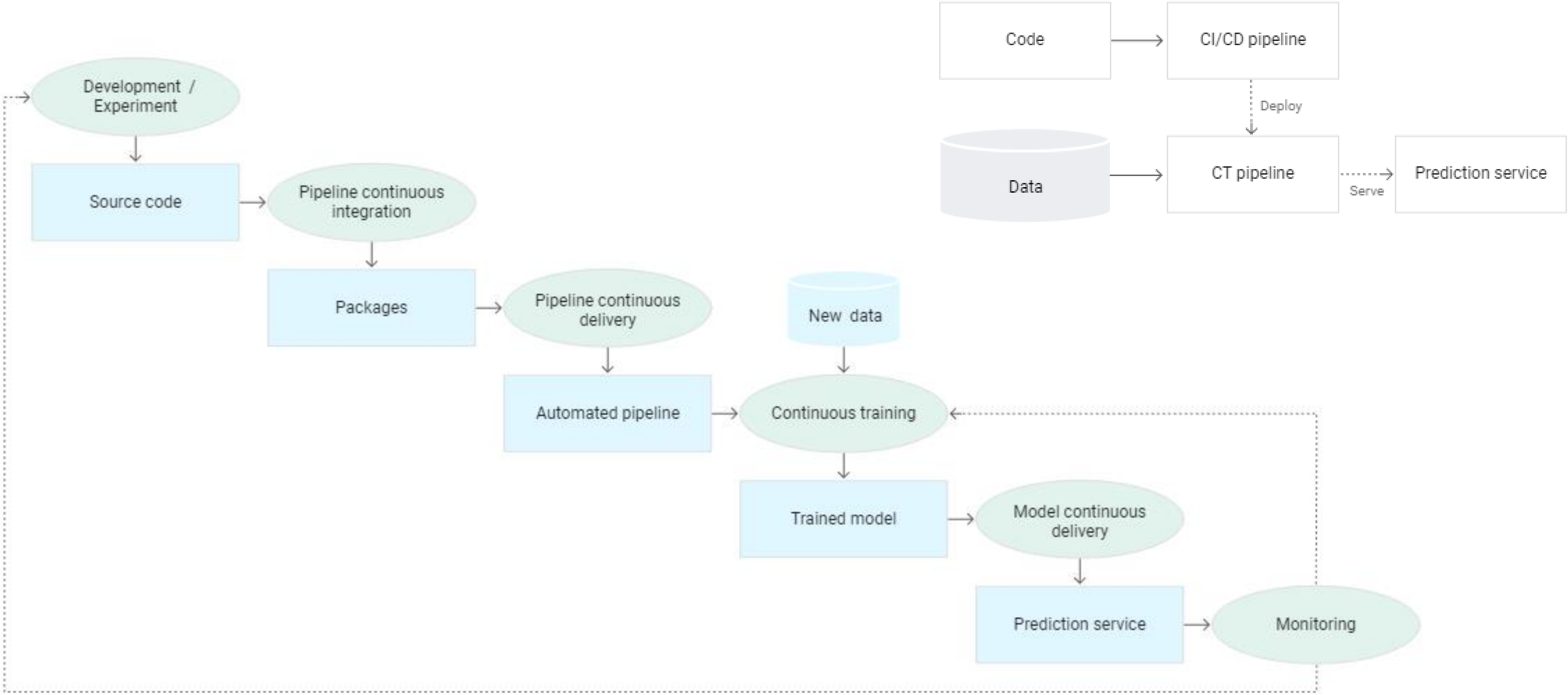
MLOps level 2: CI/CD/CT pipeline automation



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

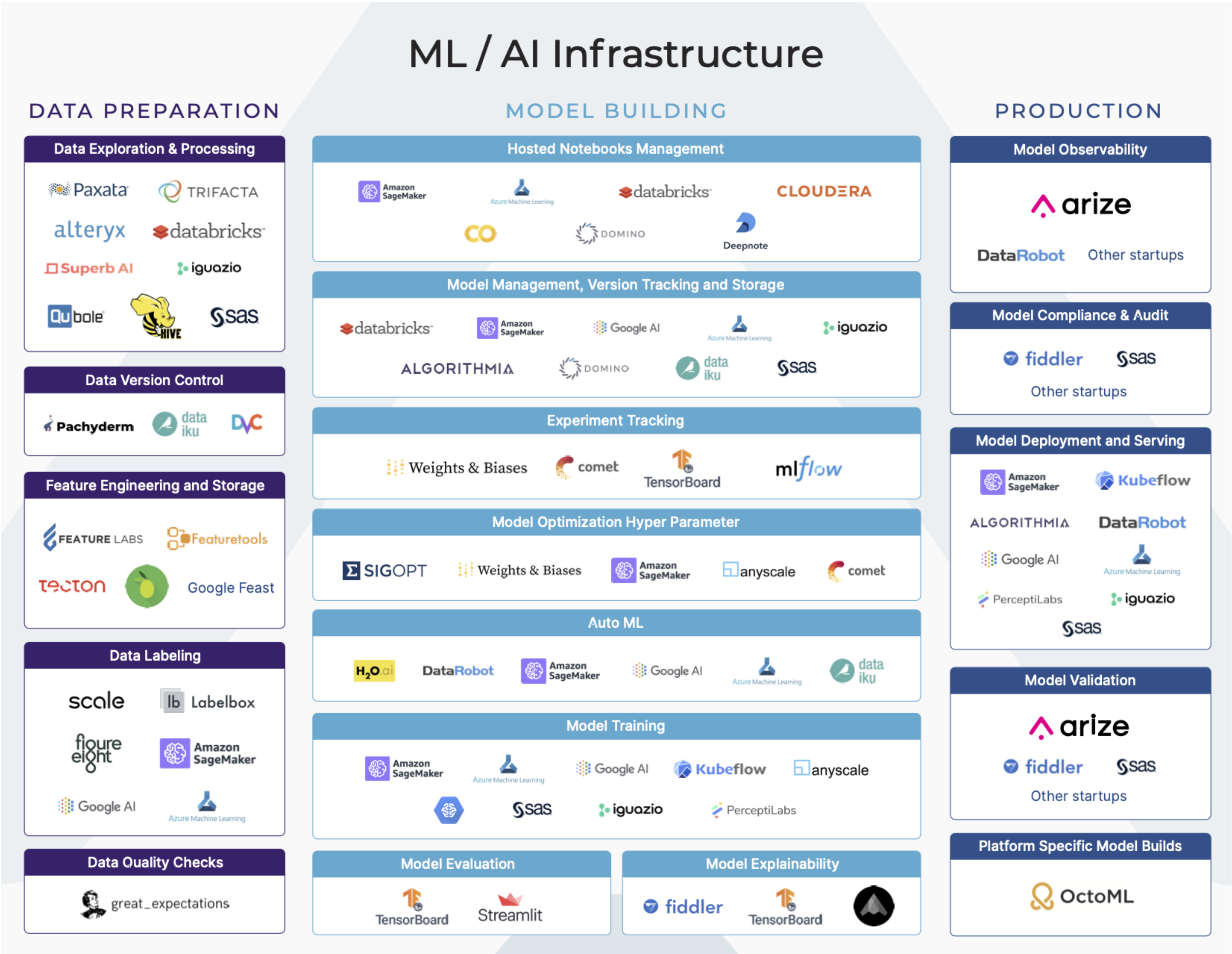
Pipelines

MLOps level 2: CI/CD/CT pipeline automation



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

Tools



ML / AI Infrastructure
DATA PREPARATION
MODEL BUILDING
PRODUCTION
Data Exploration & Processing
Paxata
TRIFACTA
alteryx
databricks
Superb AI
iguazio
Qubole
HIVE
sas
Data Version Control
Pachyderm
data iku
DVC
Feature Engineering and Storage
FEATURE LABS
Featuretools
tecton
Google Feast
Data Labeling
scale
Labelbox
figure eight
Amazon SageMaker
Google AI
Azure Machine Learning
Data Quality Checks
great_expectations
Hosted Notebooks Management
Amazon SageMaker
Azure Machine Learning
databricks
CLOUDERA
CO
DOMINO
Deepnote
Model Management, Version Tracking and Storage
databricks
Amazon SageMaker
Google AI
Azure Machine Learning
iguazio
ALGORITHMIA
DOMINO
data iku
sas
Experiment Tracking
Weights & Biases
comet
TensorBoard
mlflow
Model Optimization Hyper Parameter
SIGOPT
Weights & Biases
Amazon SageMaker
anyscale
comet
Auto ML
H2O.ai
DataRobot
Amazon SageMaker
Google AI
Azure Machine Learning
data iku
Model Training
Amazon SageMaker
Azure Machine Learning
Google AI
Kubeflow
anyscale
sas
iguazio
PerceptiLabs
Model Evaluation
TensorBoard
Streamlit
Model Explainability
fiddler
TensorBoard
Model Observability
arize
DataRobot
Other startups
Model Compliance & Audit
fiddler
sas
Other startups
Model Deployment and Serving
Amazon SageMaker
Kubeflow
ALGORITHMIA
DataRobot
Google AI
Azure Machine Learning
PerceptiLabs
iguazio
sas
Model Validation
arize
fiddler
sas
Other startups
Platform Specific Model Builds
OctoML

Labeling & Data Management – Superb AI Suite

- Open-source Version Control System for Machine Learning Projects

Data Management	Data Analytics & Processing	Labeling	Auto Labeling

Labeling & Data Management – Superb AI Suite

- Open-source Version Control System for Machine Learning Projects

```
$ pip install spb-cli
$ spb --version
0.0.xx
```

```
$ spb configure
Suite Account Name: foo
Access Key: bar
```

- Describe Projects

```
$ spb describe projects
┌──────────────────────────────────────────────┬────────┬──────────┐
│ NAME                                         │ LABELS │ PROGRESS │
├──────────────────────────────────────────────┼────────┼──────────┤
│ my-project                                   │   5837 │    13.7% │
...
Press any button to continue to the next page (1/10). Otherwise press ‘Q’ to quit.
```

Labeling & Data Management – Superb AI Suite

- Open-source Version Control System for Machine Learning Projects

  - Upload Data

```
$ cd some-folder
$ spb upload dataset
Project Name: my-project
Dataset Name: my-dataset

Uploading 3 data and 0 labels to dataset 'my-dataset' under project 'my-project'. P
Uploading data:
100%|██████████████████████████████████████████████| 3/3 [00:03<00:00,  1.06s/it]
```

  - Upload Data with Label

```
$ cd some-folder
$ spb upload dataset --include-label
Project Name: my-project
Dataset Name: my-dataset

Uploading 3 data and 0 labels to dataset 'my-dataset' under project 'my-project'. P
Uploading data:
100%|██████████████████████████████████████████████| 3/3 [00:03<00:00,  1.06s/it]
Uploading labels:
100%|██████████████████████████████████████████████| 3/3 [00:03<00:00,  3.40s/it]
```

Labeling & Data Management – Superb AI Suite

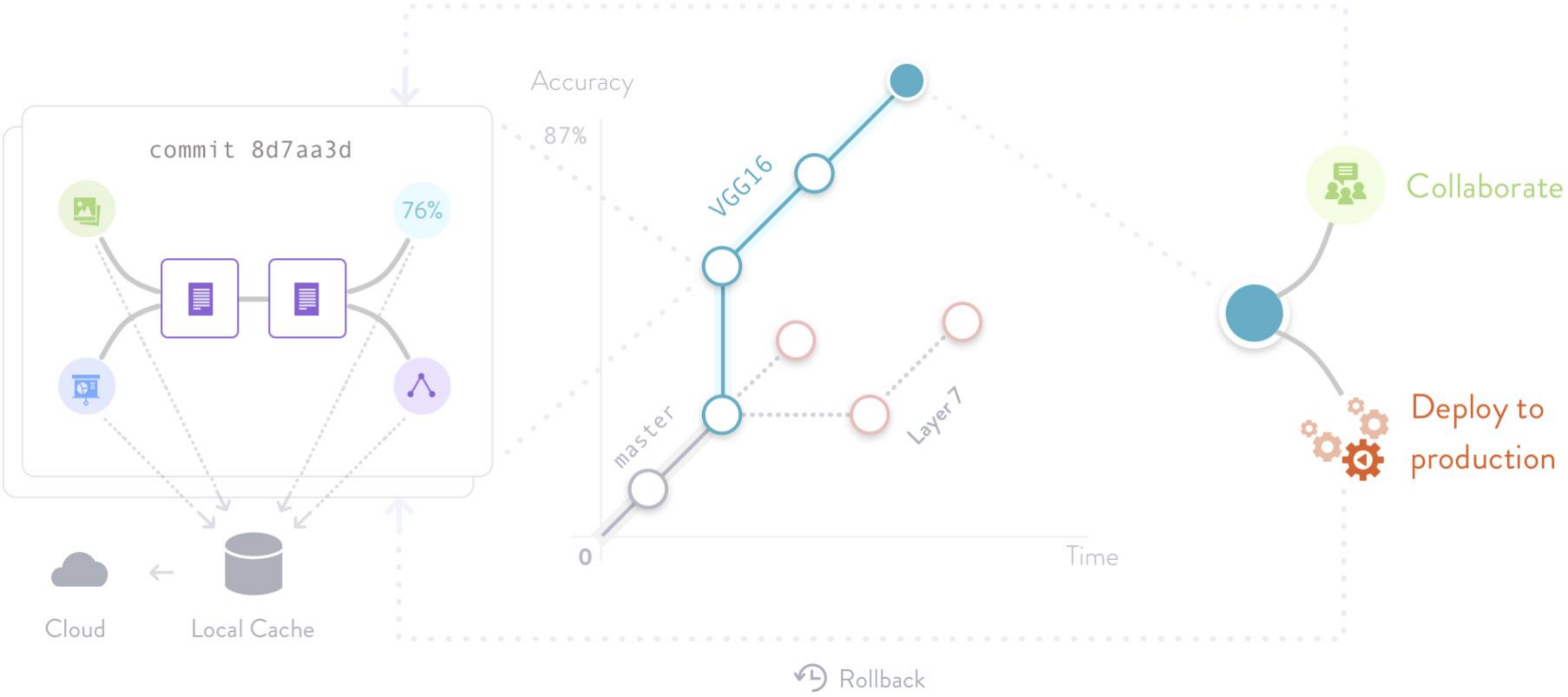
- Open-source Version Control System for Machine Learning Projects
  - Download Data

```
$ cd some-folder
$ spb download
Project Name: my-project

Downloading 3 data and 3 labels from project 'my-project' to '.'. Proceed? [y/N]: y
100%|██████████████████████████████████████████████| 1/1 [00:00<00:00,  1.11it/s]

** Result Summary **
Successful download of 3 out of 3 labels. (100.0%)
Successful download of 3 out of 3 data. (100.0%)
```

Versioning - DVC

- Open-source Version Control System for Machine Learning Projects

Data Versioning	Data Access	Data Pipeline	Data Experiments



Data, Labeling, Model Versioning - DVC

- Open-source Version Control System for Machine Learning Projects

```
$ pip install dvc
```

```
$ dvc init
```

```
$ git status
Changes to be committed:
        new file:   .dvc/.gitignore
        new file:   .dvc/config
        ...
$ git commit -m "Initialize DVC"
```

- Add Data

```
$ dvc add data/data.xml
```

```
$ git add data/data.xml.dvc data/.gitignore
$ git commit -m "Add raw data"
```

Versioning - DVC

- Open-source Version Control System for Machine Learning Projects

  - Add Remote

```
$ dvc remote add -d storage s3://my-bucket/dvc-storage
$ git commit .dvc/config -m "Configure remote storage"
```

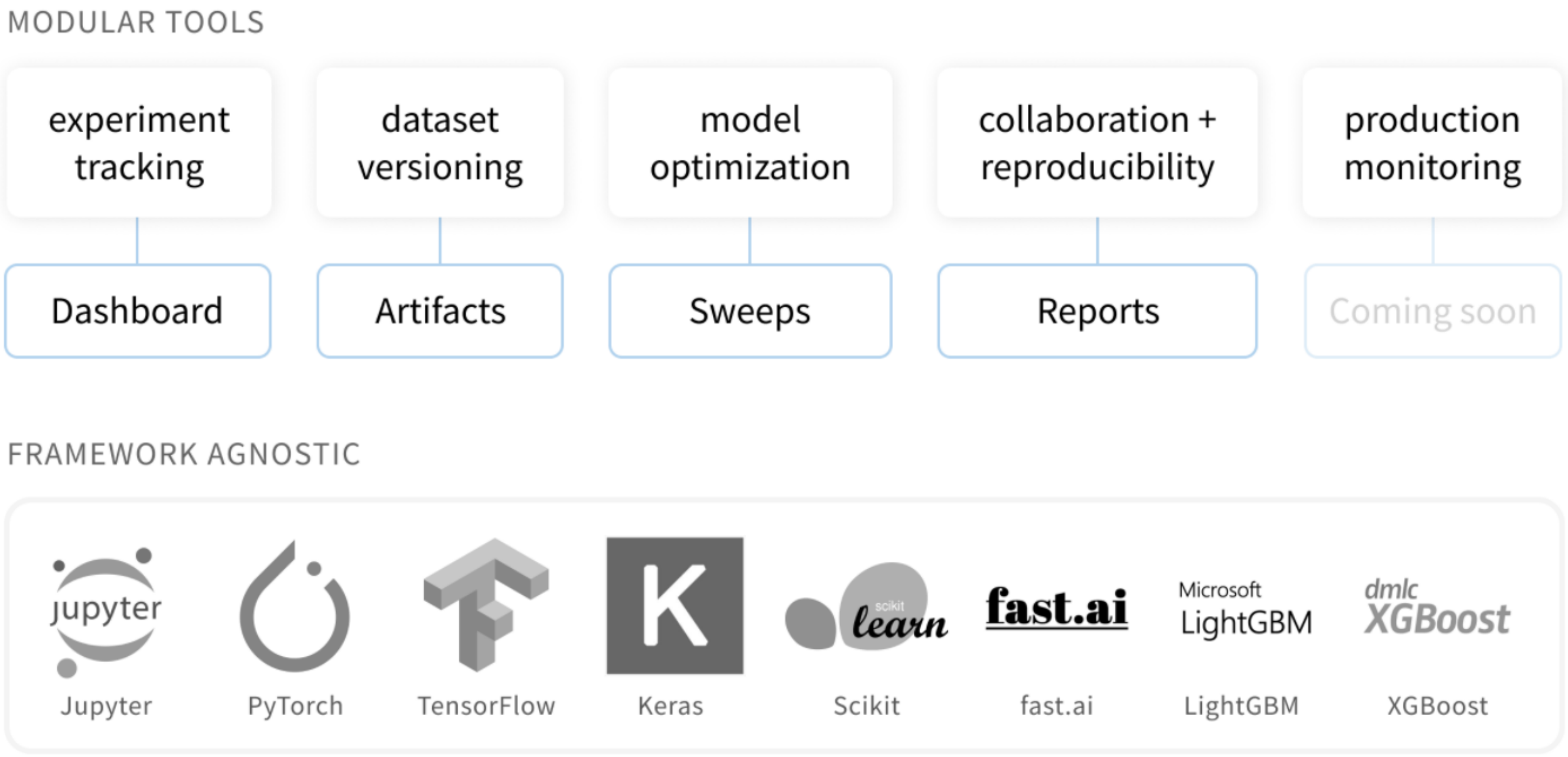
```
$ dvc push
```

  - Retrieve
    - git clone & git pull

```
$ dvc pull
```

Monitoring & Tracking – Weights & Biases

- Developer tools for machine learning
  - Experiment tracking, model optimization, and dataset versioning



Monitoring & Tracking – Weights & Biases

```
pip install wandb
```

```
wandb login
```

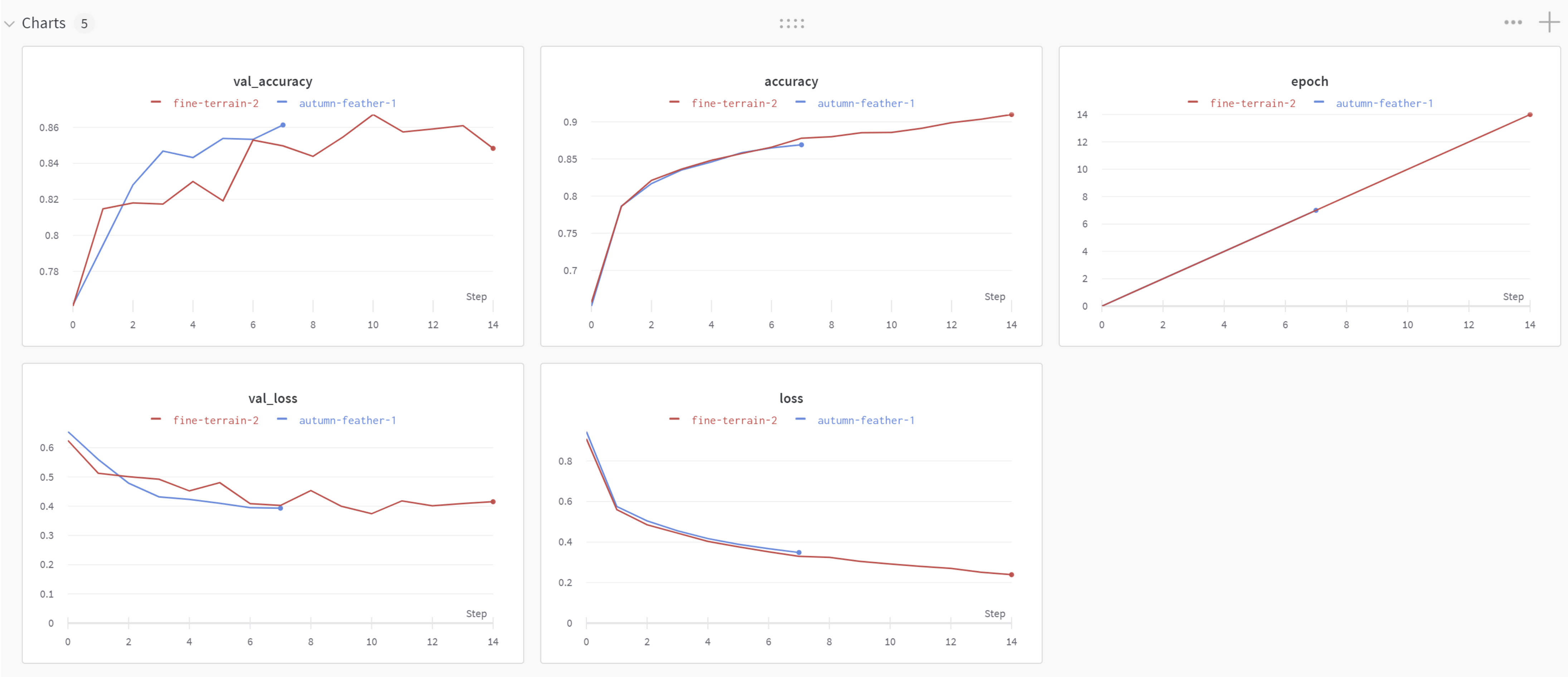
```python
# Inside my model training code
import wandb
wandb.init(project="my-project")
```

```python
wandb.config.dropout = 0.2
wandb.config.hidden_layer_size = 128
```

```python
def my_train_loop():
    for epoch in range(10):
        loss = 0 # change as appropriate :)
        wandb.log({'epoch': epoch, 'loss': loss})
```

https://github.com/wandb/tutorial

Monitoring & Tracking – Weights & Biases



End

Superb AI

Let's talk with Superb AI!

References

- https://cloud.google.com/solutions/machine-learning/mlops-continuous-delivery-and-automation-pipelines-in-machine-learning
  https://cloud.google.com/solutions/machine-learning/architecture-for-mlops-using-tfx-kubeflow-pipelines-and-cloud-build?hl=ko
- https://en.wikipedia.org/wiki/MLOps
- https://blogs.nvidia.com/blog/2020/09/03/what-is-mlops/
- https://dvc.org/doc/start/data-access
- https://docs.wandb.com/quickstart
- https://docs.superb-ai.com/developers/command-line-interface/downloading-data-and-labels
- https://blog.superb-ai.com/machine-learning-training-data-workflows/
- https://norman3.github.io/papers/docs/hidden_technical_debt
- https://papers.nips.cc/paper/5656-hidden-technical-debt-in-machine-learning-systems.pdf
- https://storage.googleapis.com/pub-tools-public-publication-data/pdf/43146.pdf
- https://pages.dataiku.com/oreilly-introducing-mlops
- https://towardsdatascience.com/ml-infrastructure-tools-for-model-building-464770ac4fec
- https://trends.google.com/trends/explore?date=today%205-y&q=MLOps
- https://course.fullstackdeeplearning.com/course-content/ml-teams/roles