Seamless MLOps with Seldon and MLflow

Adaptive. Intelligent.
Agile. Intuitive.
Alive. Inspiring.

SELDON

About me

Adrian Gonzalez-Martin

Machine Learning Engineer

agm@seldon.io

@kaseyo23

github.com/adriangonz

About Seldon

We are hiring!

seldon.io/careers/

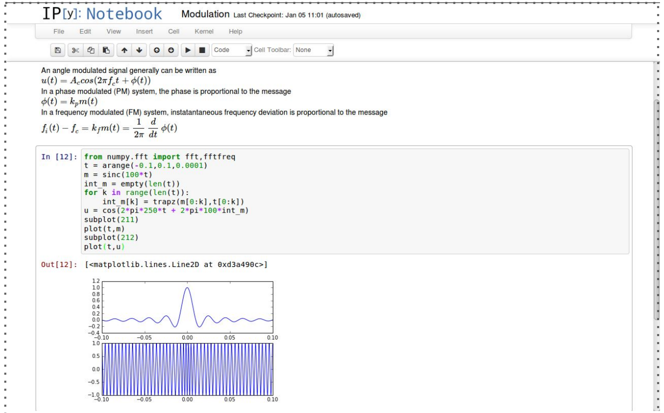
Outline

- ➔ Why is MLOps hard?
- ➔ Training with MLflow
- ➔ Serving with Seldon
- ➔ **Demo!**

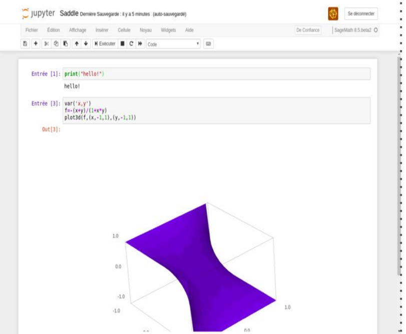
Why is MLOps hard?

What do we mean by MLOps?

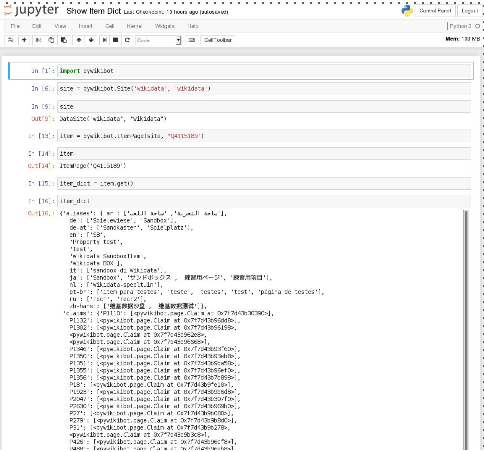
Notebooks (by themselves) **don't scale!**



Untitled.ipynb



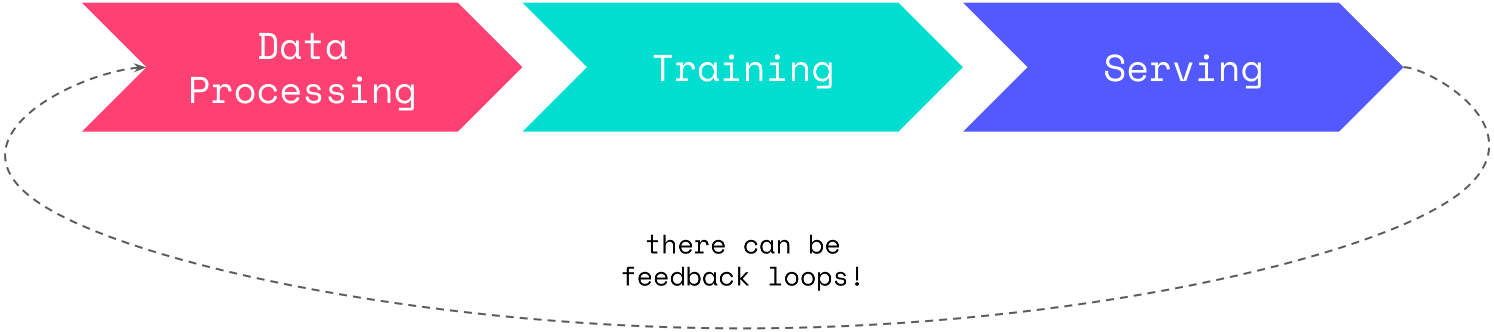
Untitled_2.ipynb



Untitled_3.ipynb

What do we mean by MLOps?

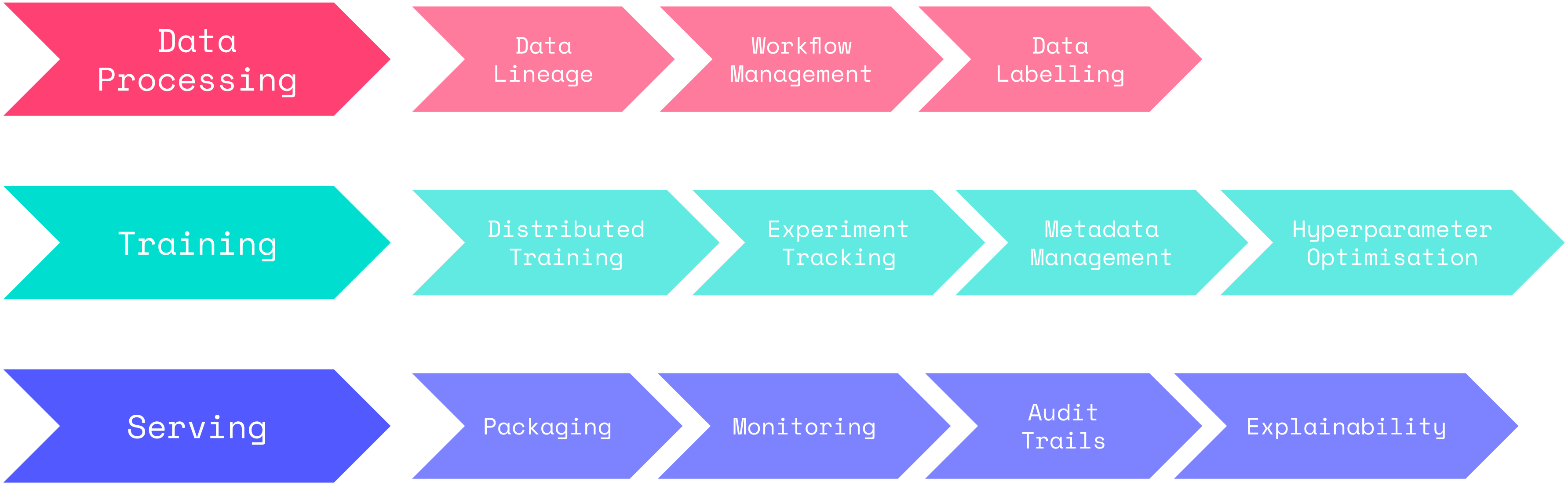
And training is **not the end goal**!



DISCLAIMER: This is just a high-level overview!

What do we mean by MLOps?

There is a **larger ML lifecycle**



DISCLAIMER: This is just a high-level overview!

What do we mean by MLOps?

*"**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 lifecycle**."* [1]

[1] https://en.wikipedia.org/wiki/MLOps

Why is MLOps hard?

Managing the ML lifecycle is **hard**!

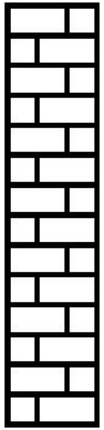
- ➔ Wide heterogeneous requirements
- ➔ Technically challenging (e.g. monitoring)
- ➔ Process needs to **scale up** across **every ML model**!
- ➔ **Organizational challenge**
    - The lifecycle needs to jump across **many walls**

Why is MLOps hard?

Organizational challenge

This is what **DevOps** tried to solve

SW Engineering

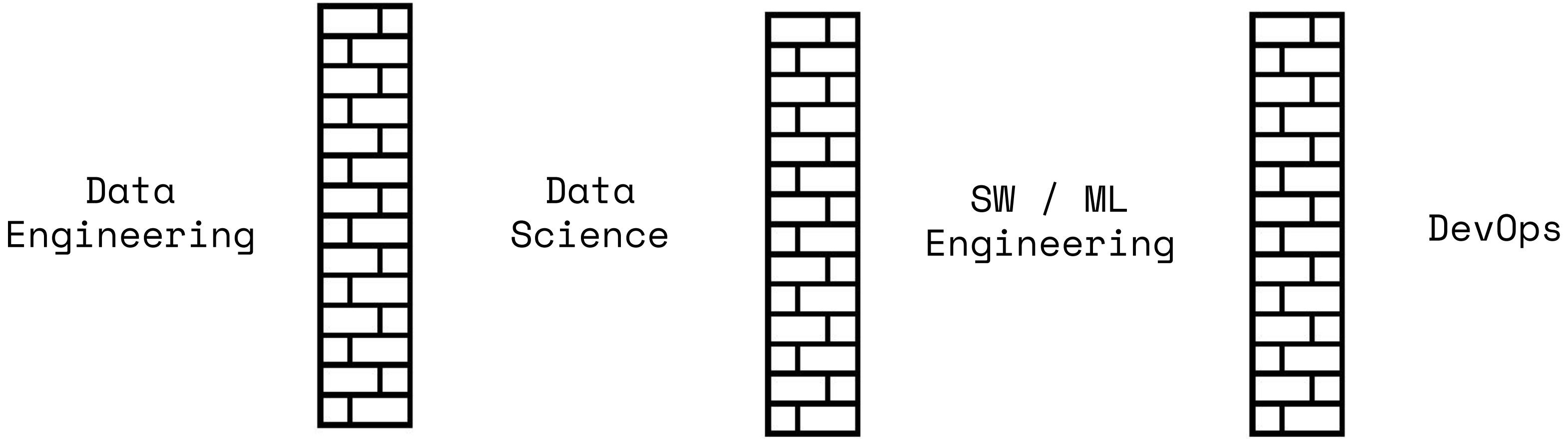


DevOps

Why is MLOps hard?

Organizational challenge

This is what MLOps
has to solve



How can we make MLOps better?

Breaking up **siloes**

- ➔ Automate, automate, automate!
- ➔ Measure and monitor everything
- ➔ “Shift-left” on responsibilities, e.g.
    - Data scientist to “own” training pipelines
    - Data scientists “own” production models
- ➔ We need **tooling** to allow this while **keeping the infrastructure “hidden”**

How can we make MLOps better?

What **tools** do we have available?

awesome | Maintained? YES | Release PROD | Languages MULTI | License MIT | Follow 2.1k

Awesome production machine learning

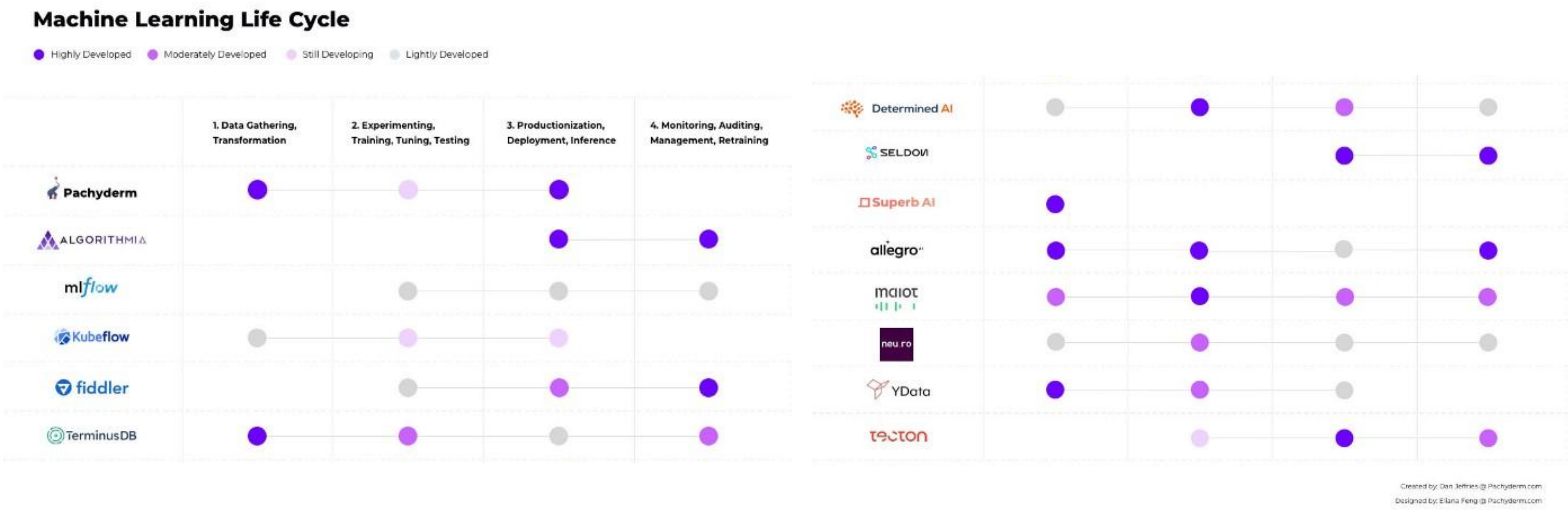
This repository contains a curated list of awesome open source libraries that will help you deploy, monitor, version, scale, and secure your production machine learning.

Quick links to sections in this page

Explaining predictions & models	Privacy preserving ML	Model & data versioning
Model Training Orchestration	Model Serving and Monitoring	Neural Architecture Search
Reproducible Notebooks	Visualisation frameworks	Industry-strength NLP
Data pipelines & ETL	Data Labelling	Data storage
Functions as a service	Computation distribution	Model serialisation
Optimized calculation frameworks	Data Stream Processing	Outlier and Anomaly Detection
Feature engineering	Feature Stores	Adversarial Robustness
Commercial Platforms		

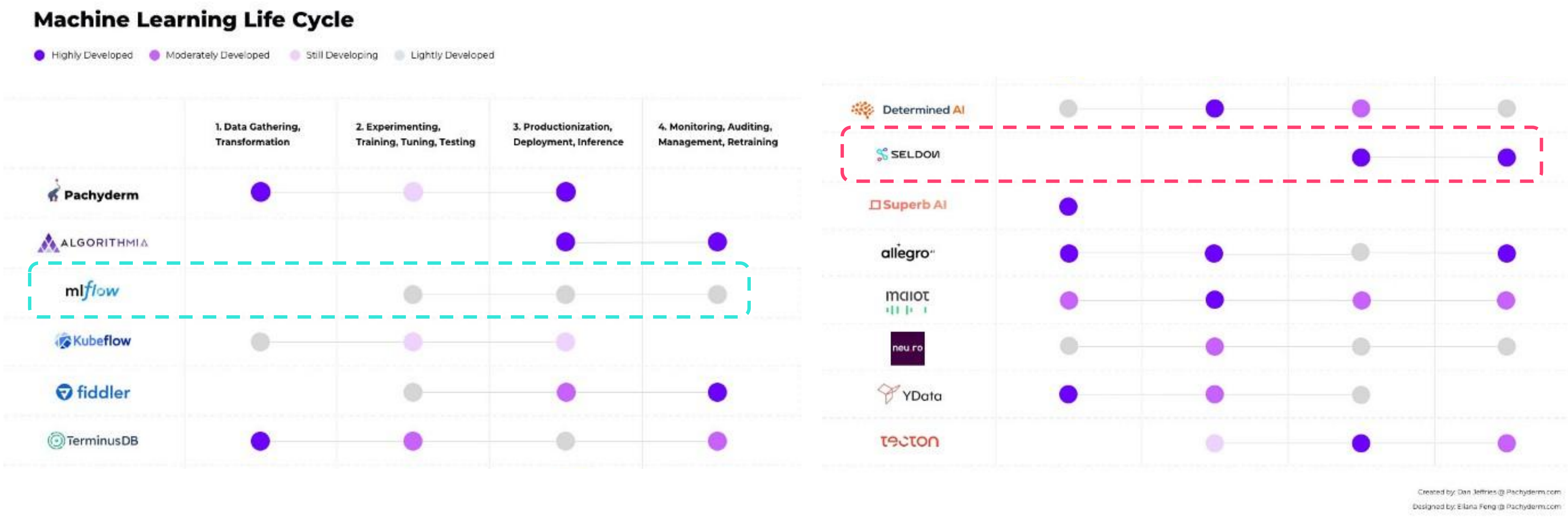
[2] https://github.com/EthicalML/awesome-production-machine-learning

How can we make MLOps better?



[3] https://towardsdatascience.com/rise-of-the-canonical-stack-in-machine-learning-724e7d2faa75

How can we make MLOps better?



[3] https://towardsdatascience.com/rise-of-the-canonical-stack-in-machine-learning-724e7d2faa75

How can we make MLOps better?

We'll focus on **training** and **serving**

Data Processing → Training → Serving

Pachyderm mlflow CORE

Training with MLFlow

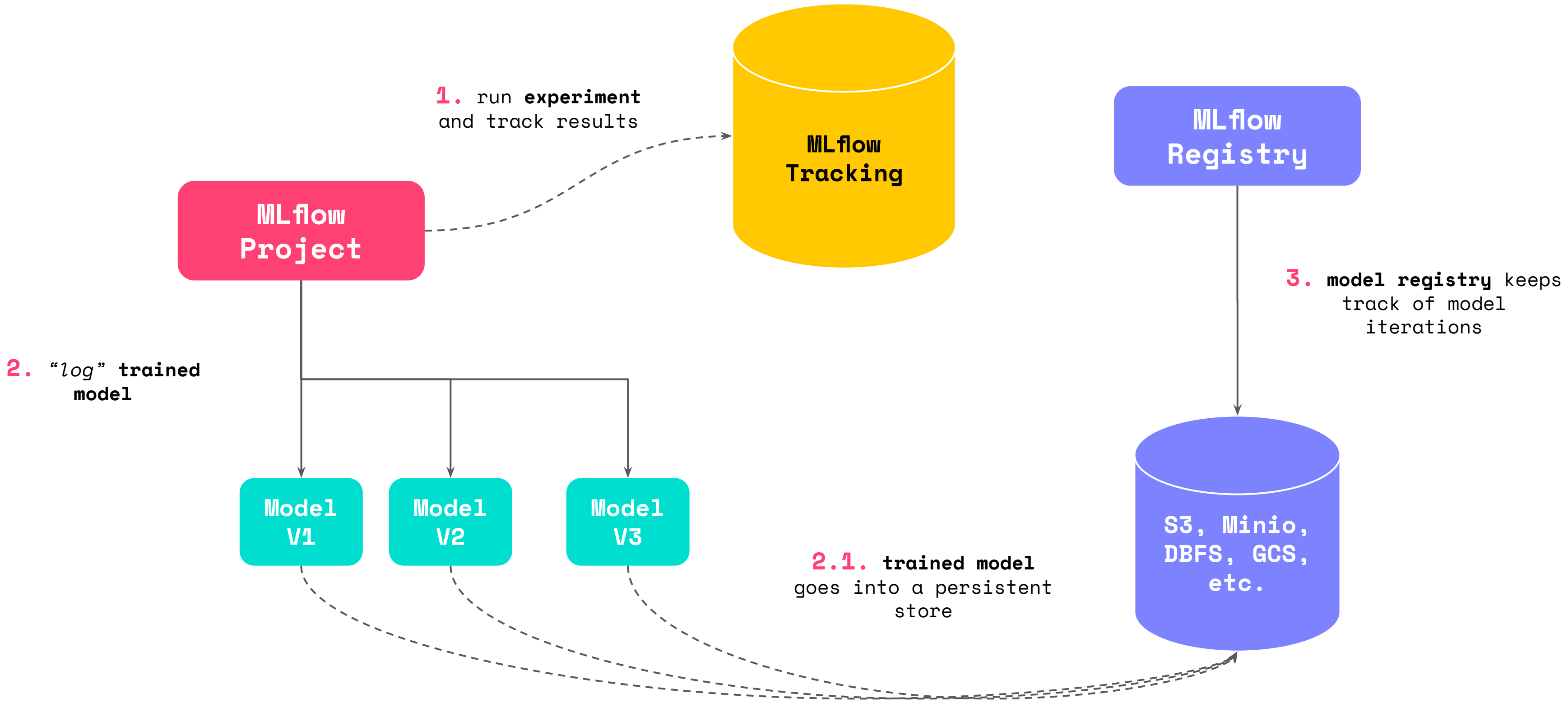
What is MLflow?

- Open Source project initially started by Databricks
- Now part of the **LFAI**

> *"**MLflow** is an open source platform to manage the ML lifecycle, including experimentation, reproducibility, deployment, and a central model registry."* [4]

[4] https://mlflow.org/

What is MLflow?



What is MLflow?

- MLflow **Project**
  - Defines **environment**, **parameters** and **model's interface**.
- MLflow **Tracking**
  - API to track **experiment results** and **hyperparameters**.
- MLflow **Model**
  - Snapshot / version of the model.
- MLflow **Registry**
  - Keeps track of **model metadata**

How does MLflow work?

MLproject file and training

```
MLproject

name: mlflow-talk

conda_env: conda.yaml

entry_points:
  main:
    parameters:
      alpha: float
      l1_ratio: {type: float, default: 0.1}
    command: "python train.py {alpha} {l1_ratio}"
```

```
$ mlflow run ./training -P alpha=0.5
$ mlflow run ./training -P alpha=1.0
```

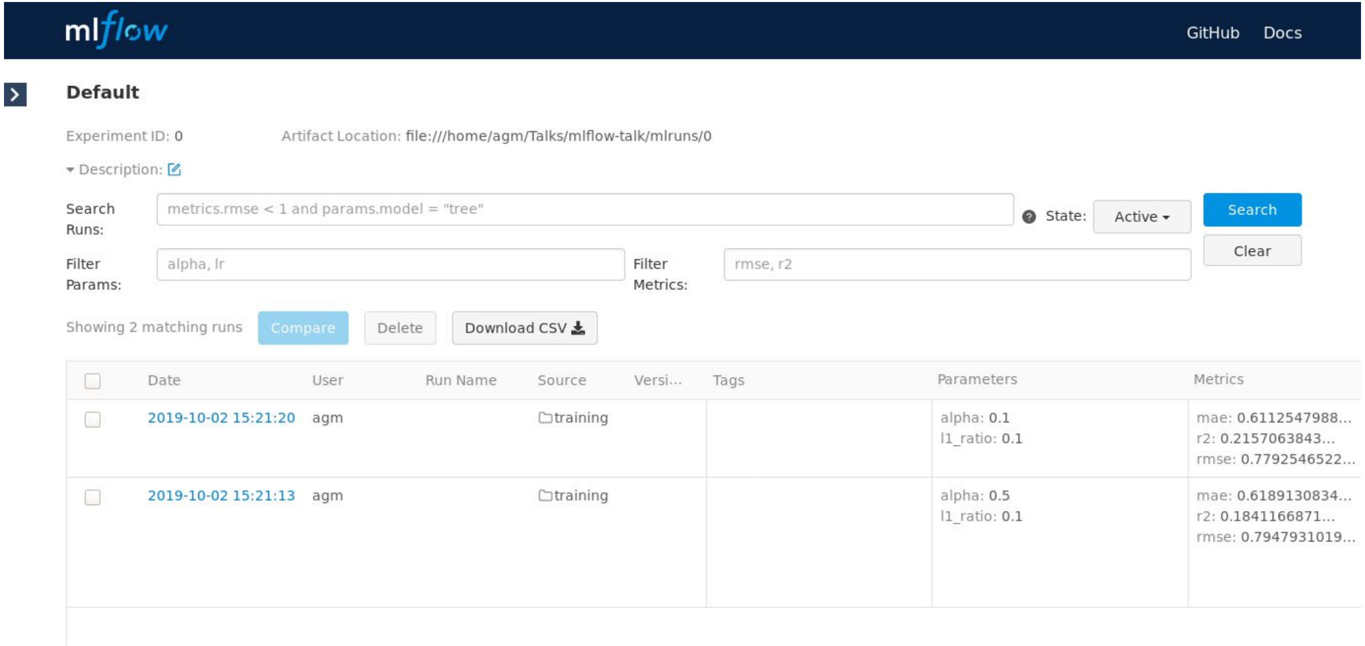
How does MLflow work?

MLmodel snapshot

MLmodel

```yaml
artifact_path: model
flavors:
  python_function:
     data: model.pkl
     env: conda.yaml
     loader_module: mlflow.sklearn
     python_version: 3.6.9
  sklearn:
     pickled_model: model.pkl
     serialization_format: cloudpickle
     sklearn_version: 0.19.1
run_id: 5a6be5a1ef844783a50a6577745dbdc3
utc_time_created: '2019-10-02 14:21:15.783806'
```

How does MLflow work?



Serving with Seldon

What is Seldon Core?

➔ Open Source project created by **Seldon**

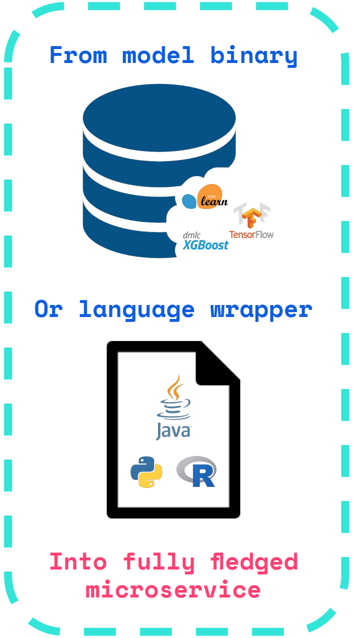
*"An **MLOps framework** to package, deploy, monitor and manage thousands of production machine learning models"* [6]

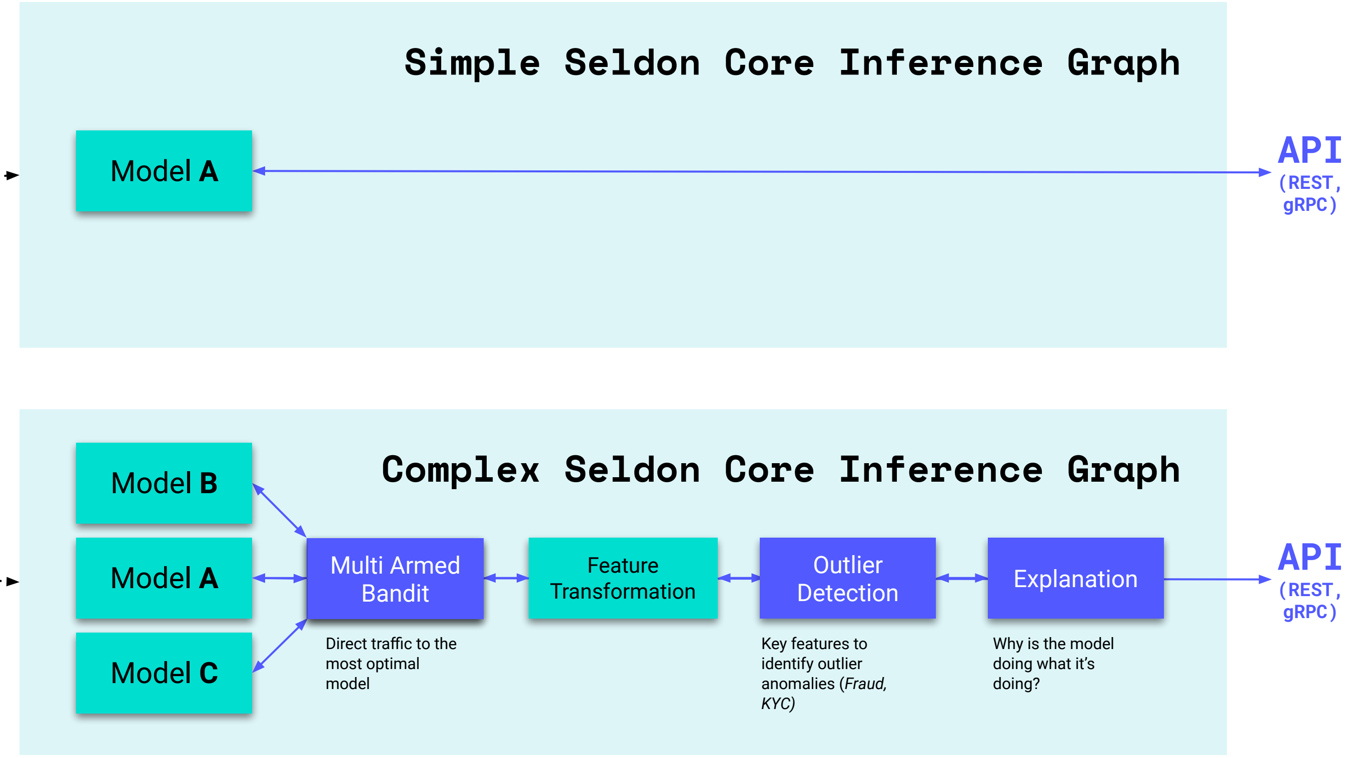
CORE

[6] https://github.com/SeldonIO/seldon-core

What is Seldon Core?

1. **Containerise**
2. **Deploy**
3. **Monitor**

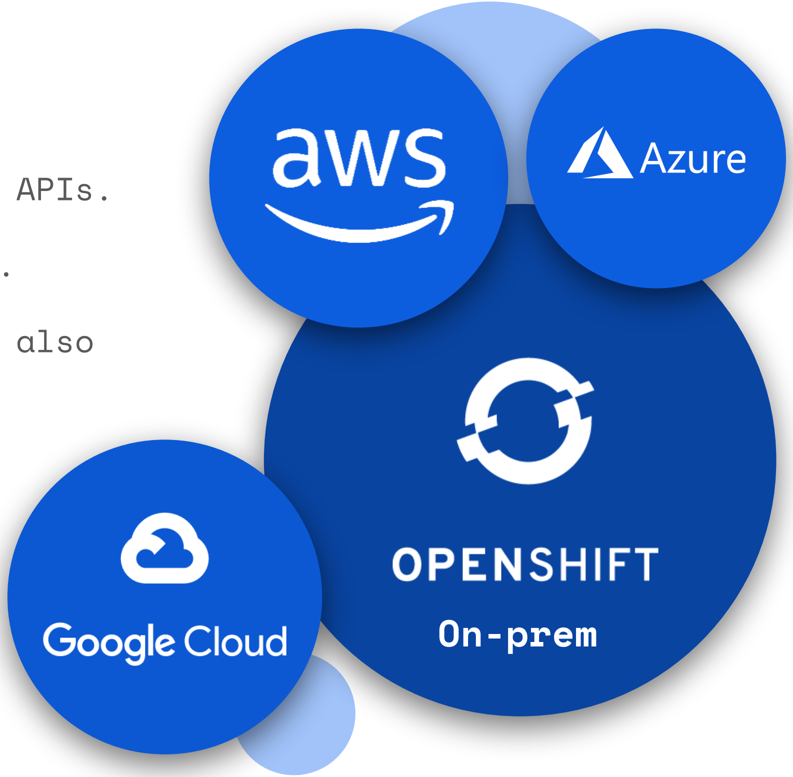




What is Seldon Core?

Cloud Native

- ➔ Built on top of Kubernetes Cloud Native APIs.
- ➔ All major cloud providers are supported.
- ➔ On-prem providers such as OpenShift are also supported.

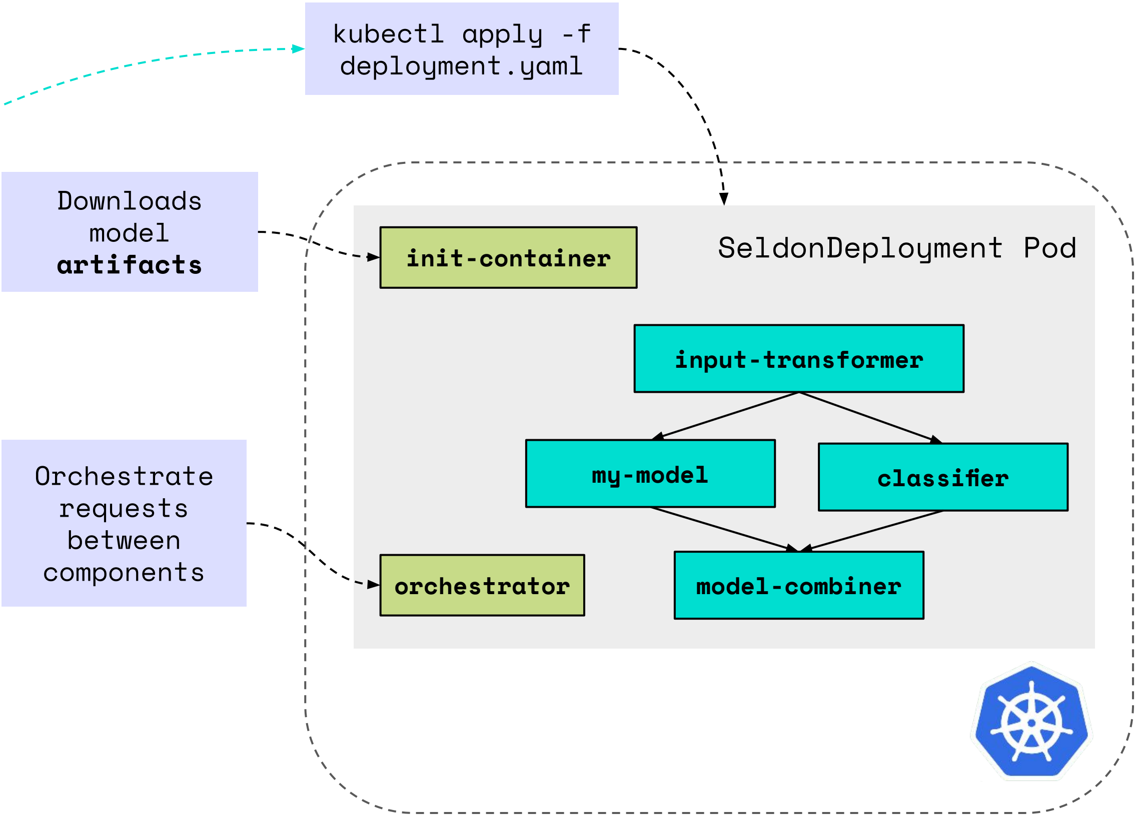


https://docs.seldon.io/projects/seldon-core/en/latest/examples/notebooks.html

How does Seldon Core work?

deployment.yaml

```yaml
apiVersion: machinelearning.seldon.io/v1
kind: SeldonDeployment
metadata:
  name: example-model
spec:
  name: example
  predictors:
  - componentSpecs:
    - spec:
        containers:
        - image: model:0.1
          name: my-model
        - image: transformer:0.1
          name: input-transformer
        - image: combiner:0.1
          name: model-combiner
    graph:
      name: input-transformer
      type: TRANSFORMER
      children:
        - name: model-combiner
          type: COMBINER
          children:
          - name: my-model
            type: MODEL
          - name: classifier
            implementation: MLFLOW_SERVER
            modelUri: gs://seldon-models/mlflow/model-a
    name: default
    replicas: 1
```



What is Seldon Core?

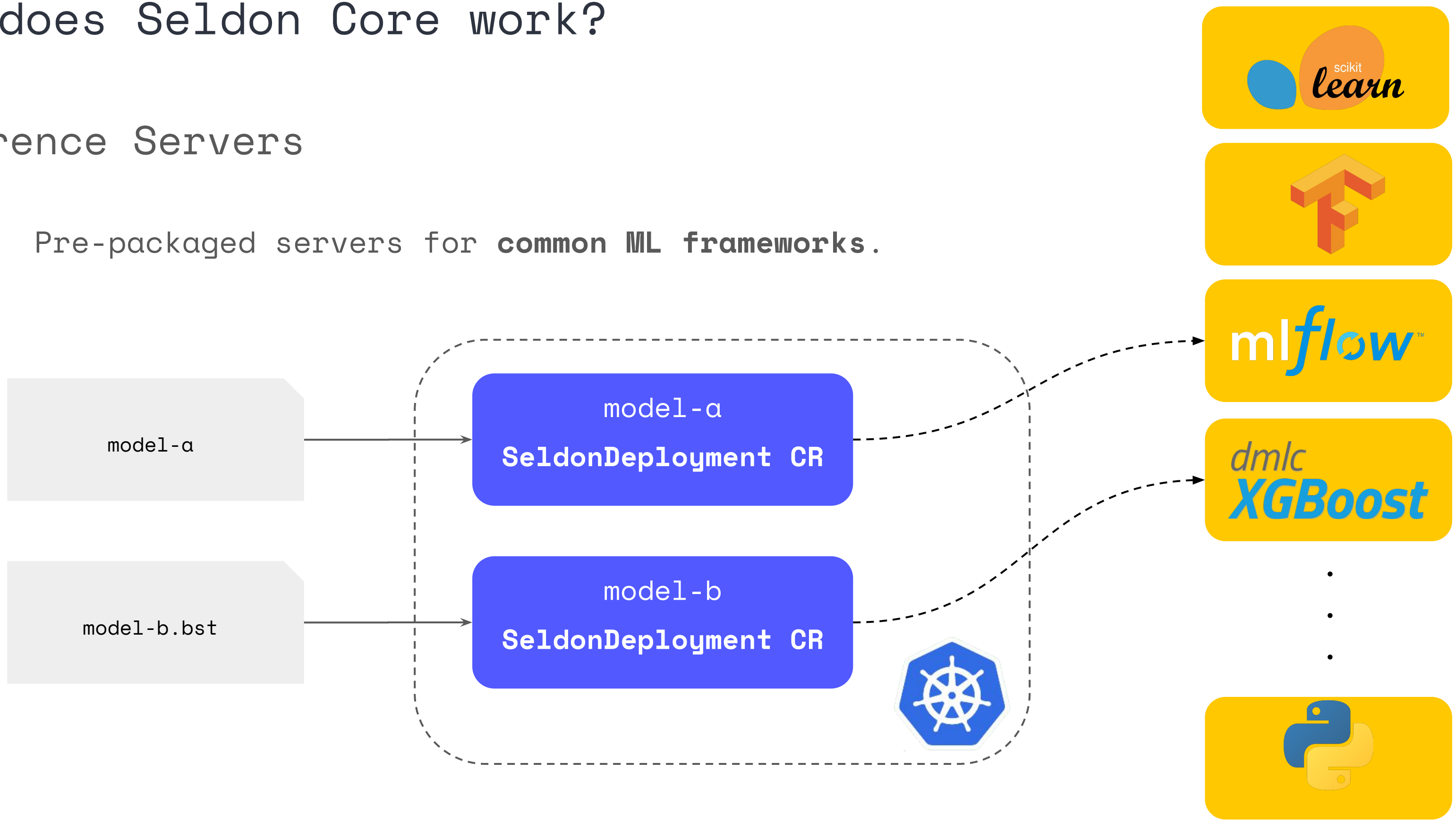
SeldonDeployment CRD to manage ML deployments

- ➔ Abstraction for Machine Learning deployments: **SeldonDeployment CRD**
  - ◆ Simple enough to deploy a model only pointing to stored weights
  - ◆ Powerful enough to keep full control over the created resources
- ➔ Pre-built **inference servers** for a subset of ML frameworks
  - ◆ Ability to write custom ones
- ➔ **A/B tests**, **shadow deployments**, etc.
- ➔ Integrations with **Alibi explainers**, **outlier detectors**, etc.
- ➔ Tools and integrations for **monitoring**, **logging**, **scaling**, etc.

How does Seldon Core work?

Inference Servers

➔ Pre-packaged servers for **common ML frameworks**.

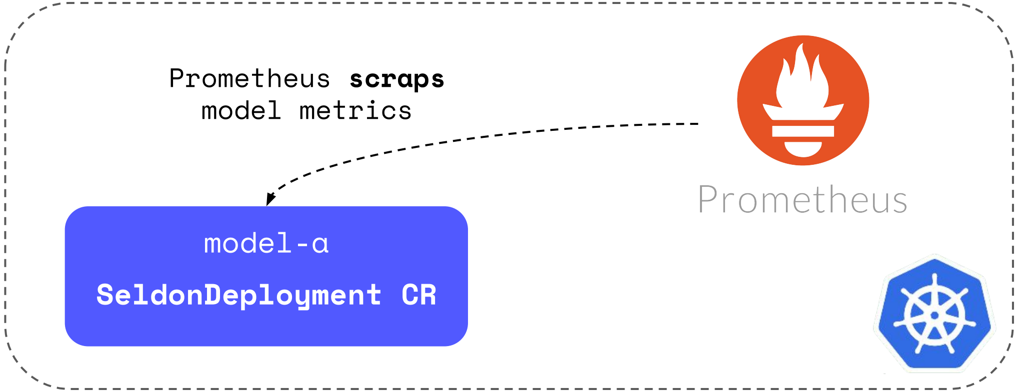


https://docs.seldon.io/projects/seldon-core/en/latest/servers/overview.html

How does Seldon Core work?

Monitoring

- ➔ Seldon integrates with **Prometheus** for metrics
- ➔ Out of the box: **memory**, **CPU**, **latency**, etc.
- ➔ **Custom metrics** are also supported



https://docs.seldon.io/projects/seldon-core/en/latest/examples/notebooks.html

How does Seldon Core work?

Monitoring

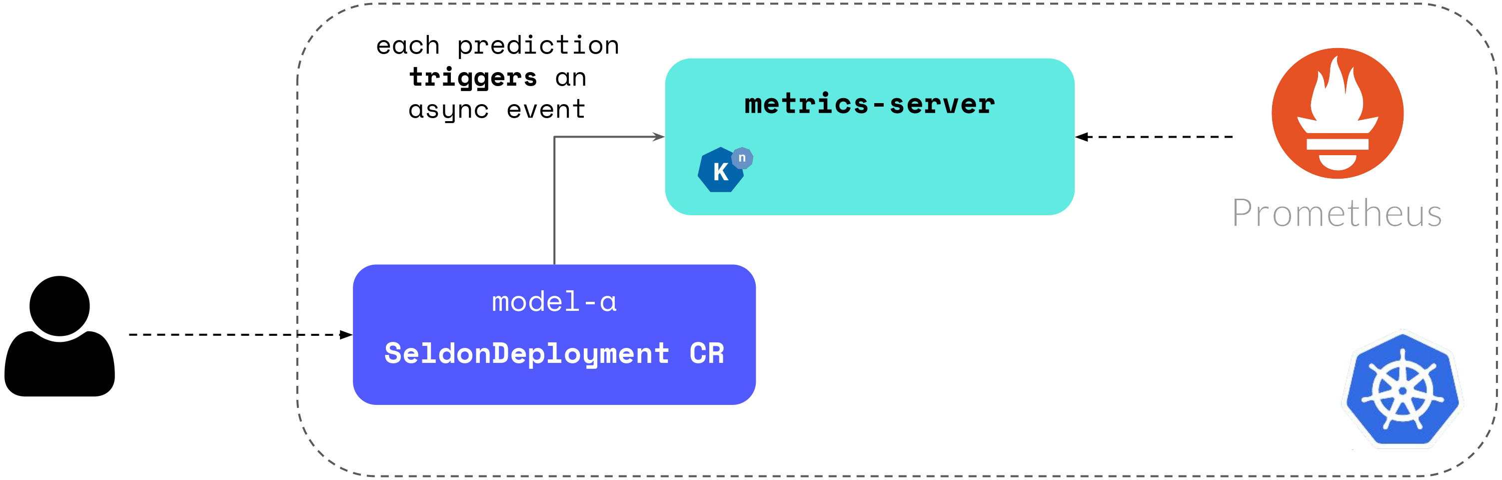
- Seldon leverages KNative for (more advanced) **async monitoring pipelines**
    - **Outlier detection** (through Alibi Detect)
    - **Drift detection** (through Alibi Detect)
    - **Custom metrics**

https://docs.seldon.io/projects/seldon-core/en/latest/examples/notebooks.html

How does Seldon Core work?

Monitoring

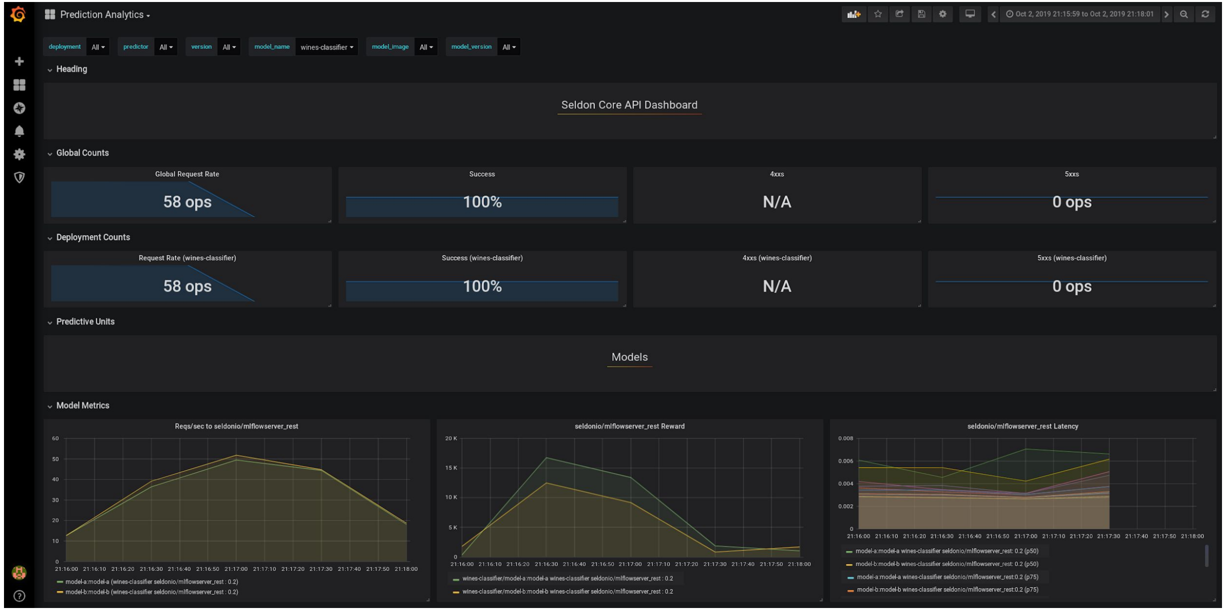
- ➔ Seldon leverages KNative for (more advanced) **async monitoring pipelines**



https://docs.seldon.io/projects/seldon-core/en/latest/examples/notebooks.html

How does Seldon Core work?

Monitoring

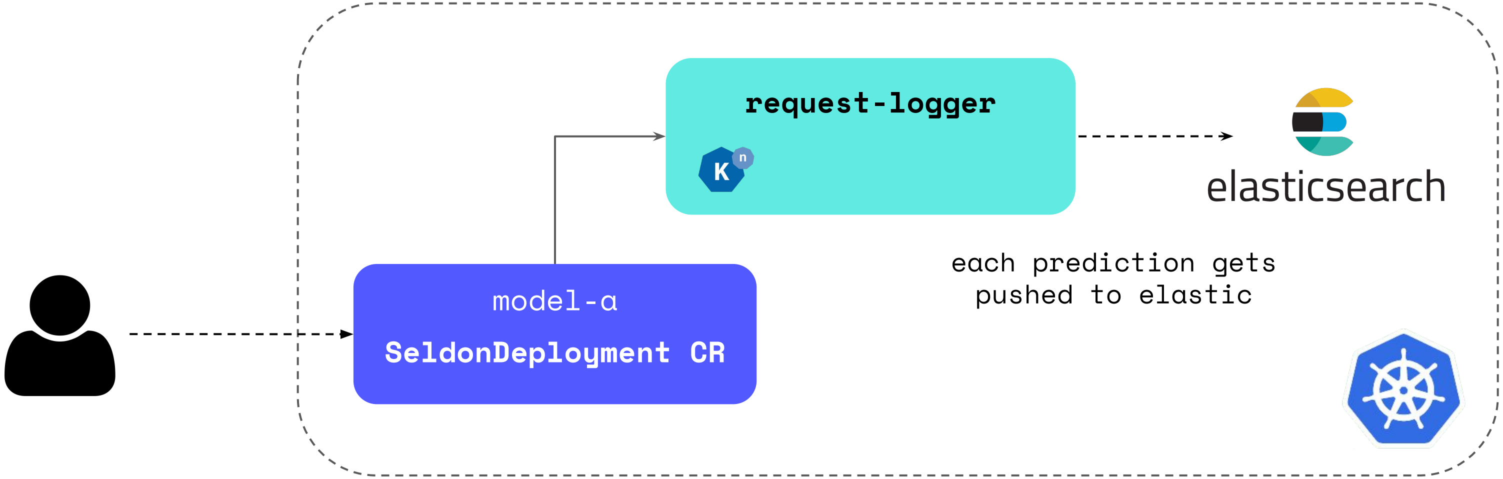


https://docs.seldon.io/projects/seldon-core/en/latest/examples/notebooks.html

How does Seldon Core work?

Auditability

➔ We can leverage a similar pattern to **log each prediction request**

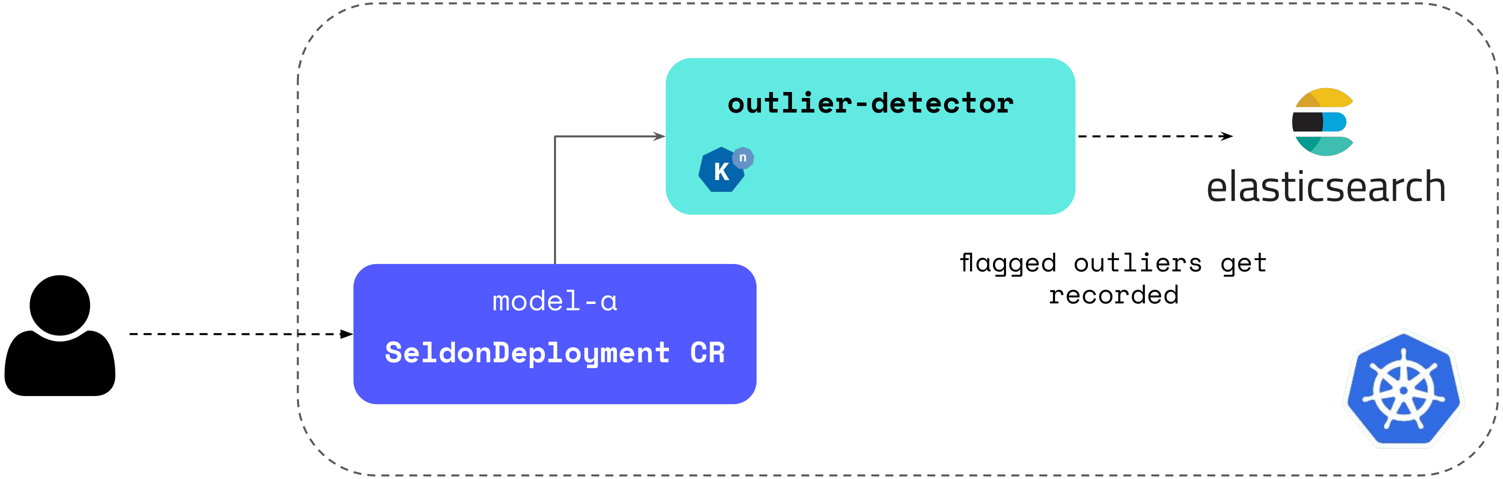


https://docs.seldon.io/projects/seldon-core/en/latest/examples/notebooks.html

How does Seldon Core work?

Auditability

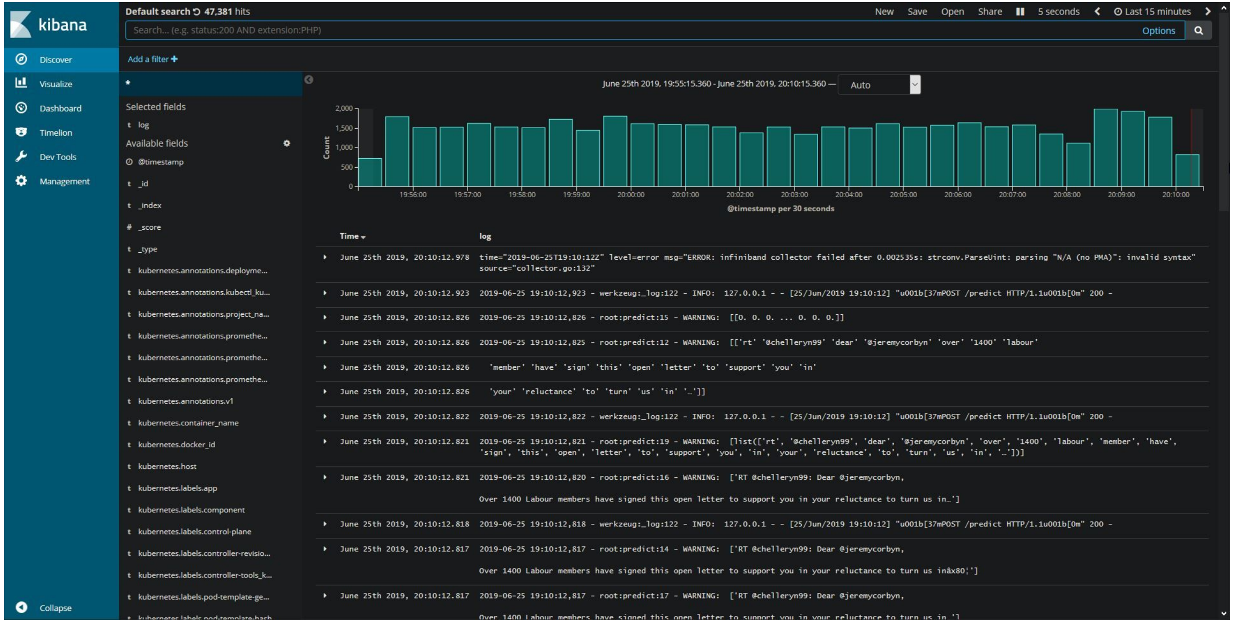
➔ And / or some attributes of each instance (e.g. **outliers**)



https://docs.seldon.io/projects/seldon-core/en/latest/examples/notebooks.html

How does Seldon Core work?

Auditability

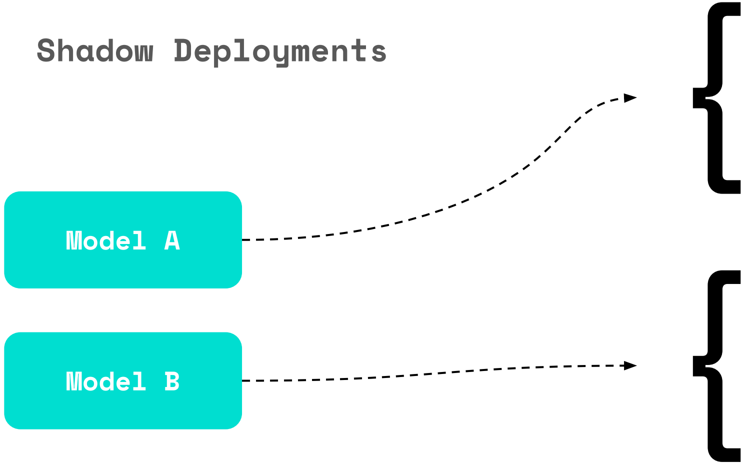


https://docs.seldon.io/projects/seldon-core/en/latest/examples/notebooks.html

How does Seldon Core work?

Advanced Deployment Models

- ➔ **A/B Tests**
  - ◆ We'll see this in the demo!
- ➔ **Shadow Deployments**



a-b-deployment.yaml

```yaml
apiVersion: machinelearning.seldon.io/v1alpha2
kind: SeldonDeployment
metadata:
  name: wines-classifier
spec:
  name: wines-classifier
  predictors:
  - graph:
        children: []
        implementation: MLFLOW_SERVER
        modelUri: gs://seldon-models/mlflow/model-a
        name: wines-classifier
        name: model-a
        replicas: 1
        traffic: 50
  - graph:
        children: []
        implementation: MLFLOW_SERVER
        modelUri: gs://seldon-models/mlflow/model-b
        name: wines-classifier
        name: model-b
        replicas: 1
        traffic: 50
```

https://docs.seldon.io/projects/seldon-core/en/latest/examples/mlflow_server_ab_test_ambassador.html

Demo!

https://github.com/adriangonz/mlflow-talk

Demo

Wine e-commerce

- ➔ We want to predict wine quality for new wines
- ➔ We want to listen to feedback from customers

Demo

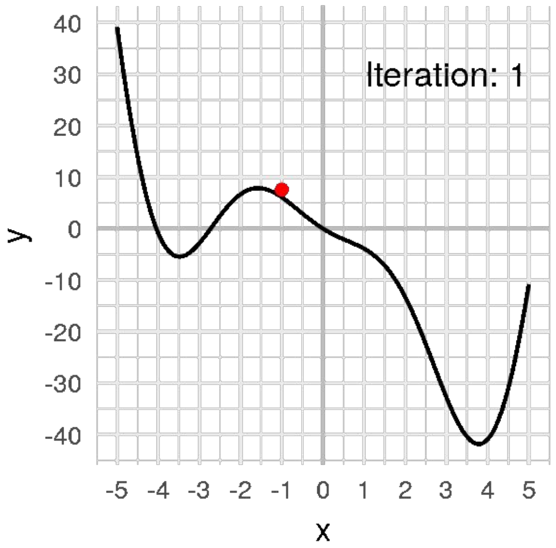
Wine quality dataset

Fixed Acidity	Volatile Acidity	Citric Acid	...	Sulphates	Alcohol	Quality
7	0.27	0.36		0.45	8.8	6
6.3	0.3	0.34		0.49	9.5	7
8.1	0.28	0.4		0.44	10.1	1
7.2	0.23	0.32		0.4	9.9	2
7.2	0.23	0.32		0.4	9.9	5
...						

Demo

ElasticNet

- ➔ Linear regression with L1 and L2 regularisers.
- ➔ Two hyperparameters: $\{a, b\}$

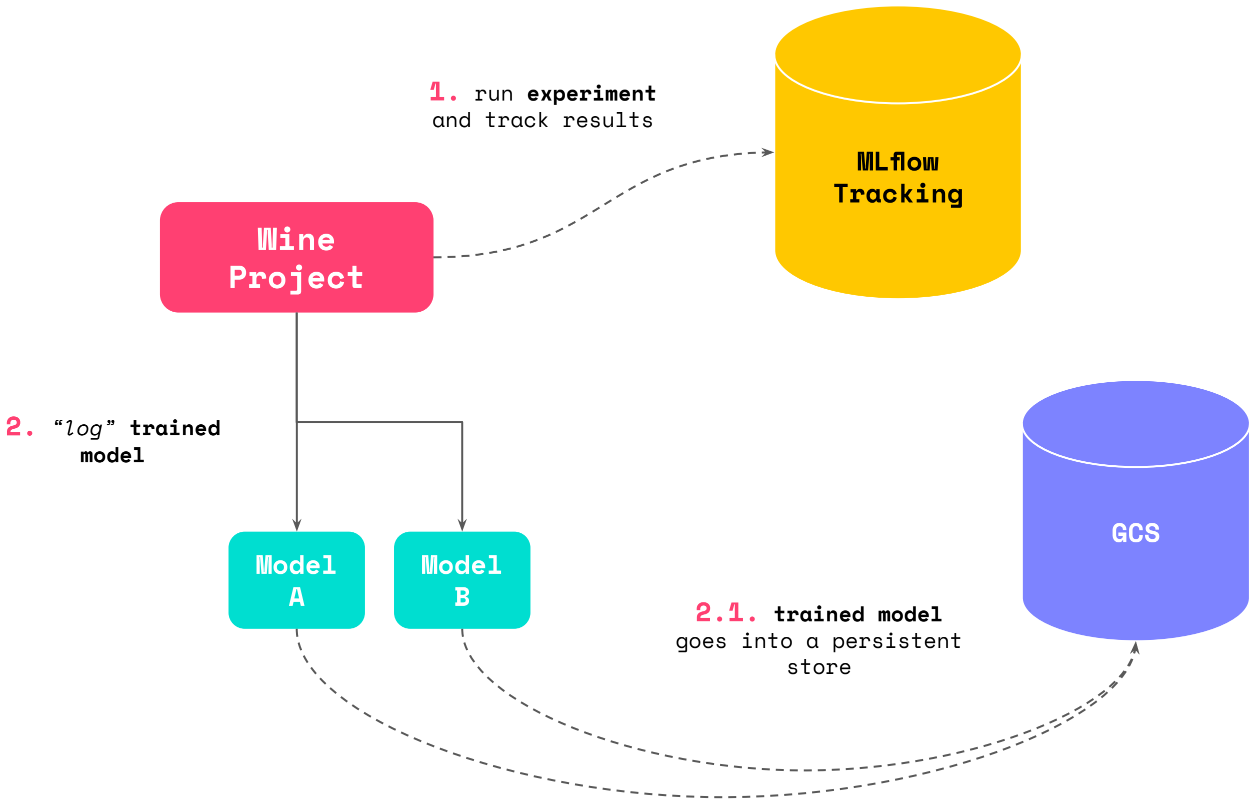


$$y = \boldsymbol{x}^{\mathsf{T}} \hat{\boldsymbol{\beta}}$$

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\| + a\,\|\boldsymbol{\beta}\|_1 + b\,\|\boldsymbol{\beta}\|_2^2$$

Demo

Training



Demo

Hyperparameter setting?

- ➔ Train two versions of ElasticNet
- ➔ It’s not clear which one would be **better in production**
- ➔ Deploy both and do **A/B test** based on **user’s feedback**

Model A

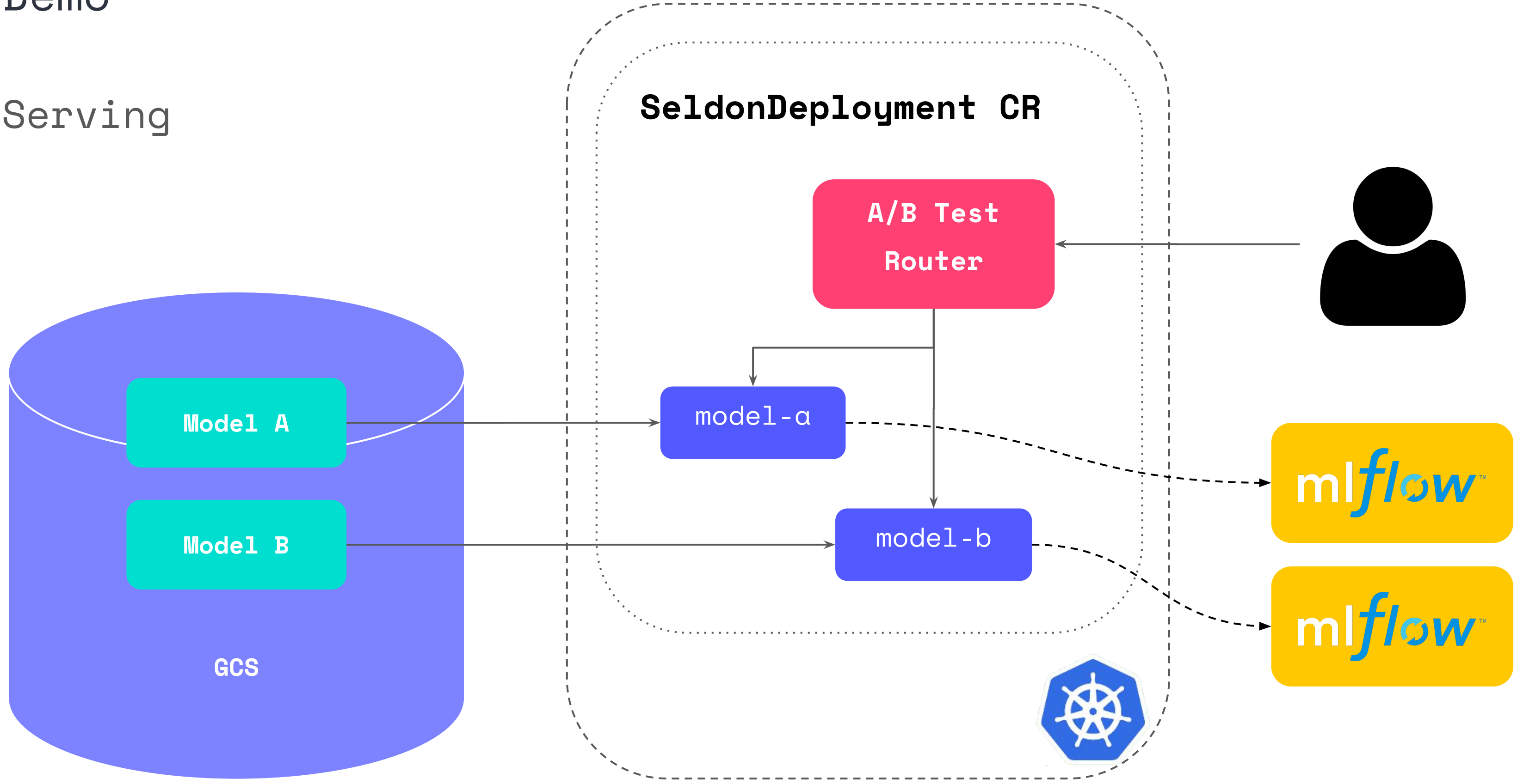
Model B

?

Which one is **best**?

Demo

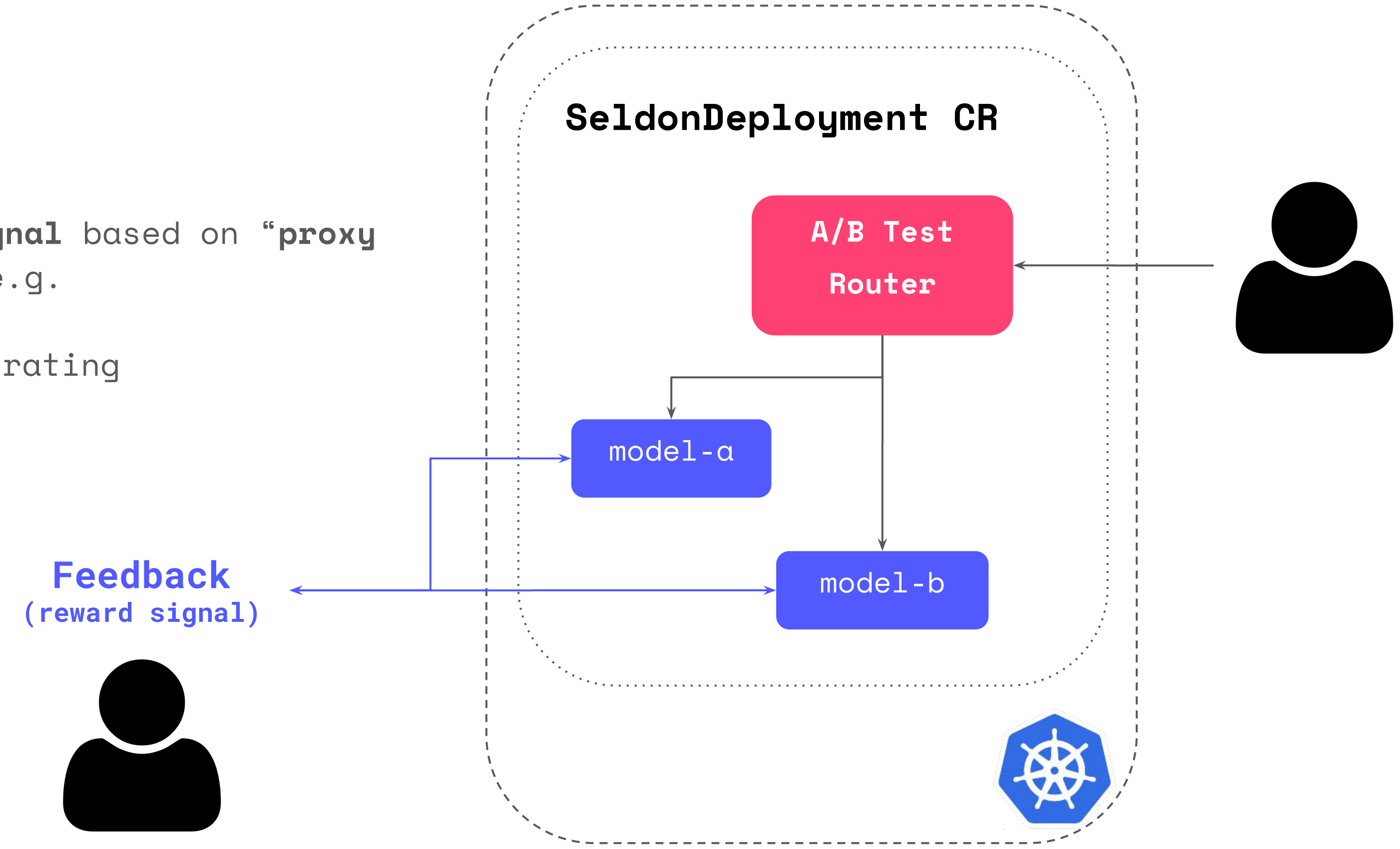
Serving



Demo

Feedback

➔ **Reward signal** based on “**proxy metric**”, e.g.
  - Sales
  - User rating



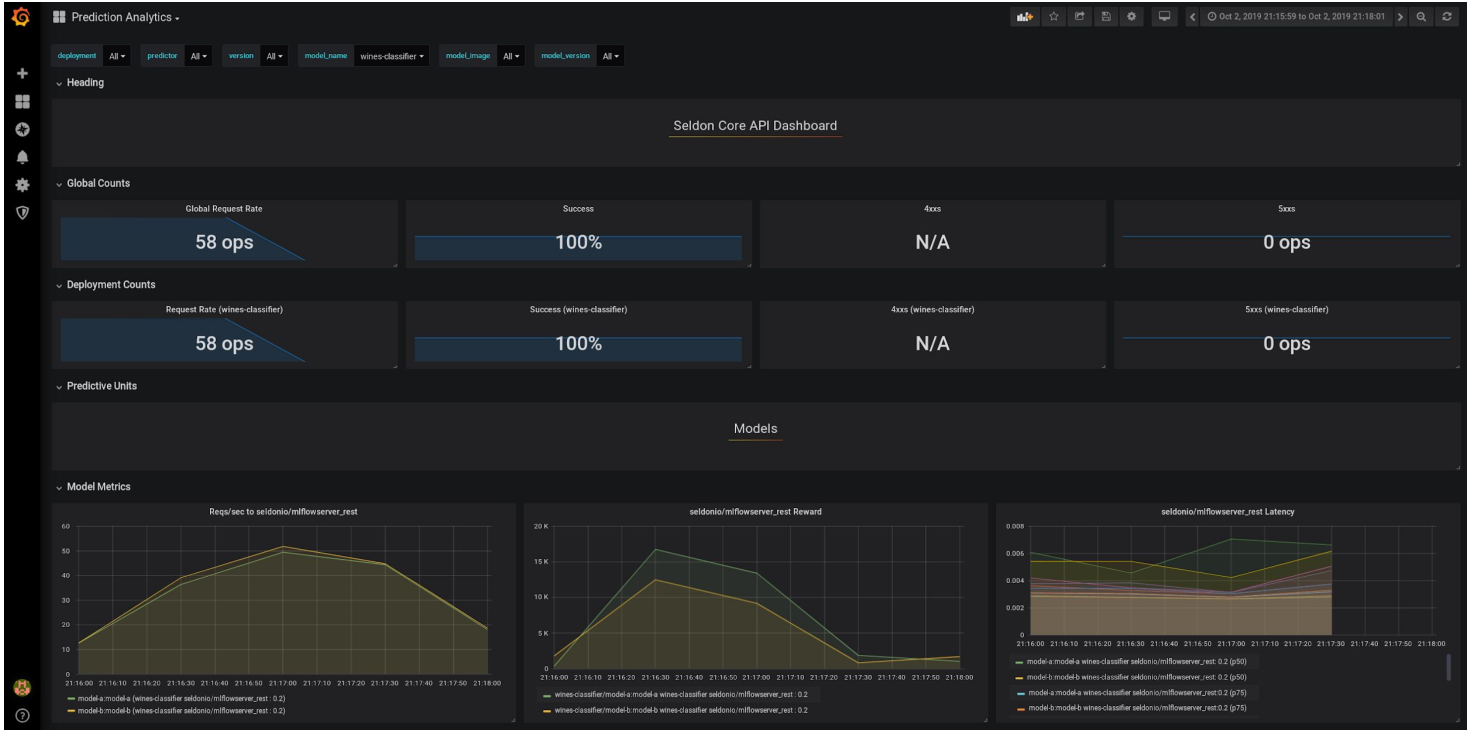
Demo

Feedback

➔ We can build a **rough reward signal** using the **squared error**

$$R(x_n) = \begin{cases} \frac{1}{(y_n - f(x_n))^2} & , y_n \neq f(x_n) \\ 500 & , y_n = f(x_n) \end{cases}$$

Seldon Analytics



Demo

https://github.com/adriangonz/mlflow-talk

Thanks!

agm@seldon.io

@kaseyo23

github.com/adriangonz

We are hiring!

seldon.io/careers/