MLOps: Deploying Models

Ed Shee - Seldon

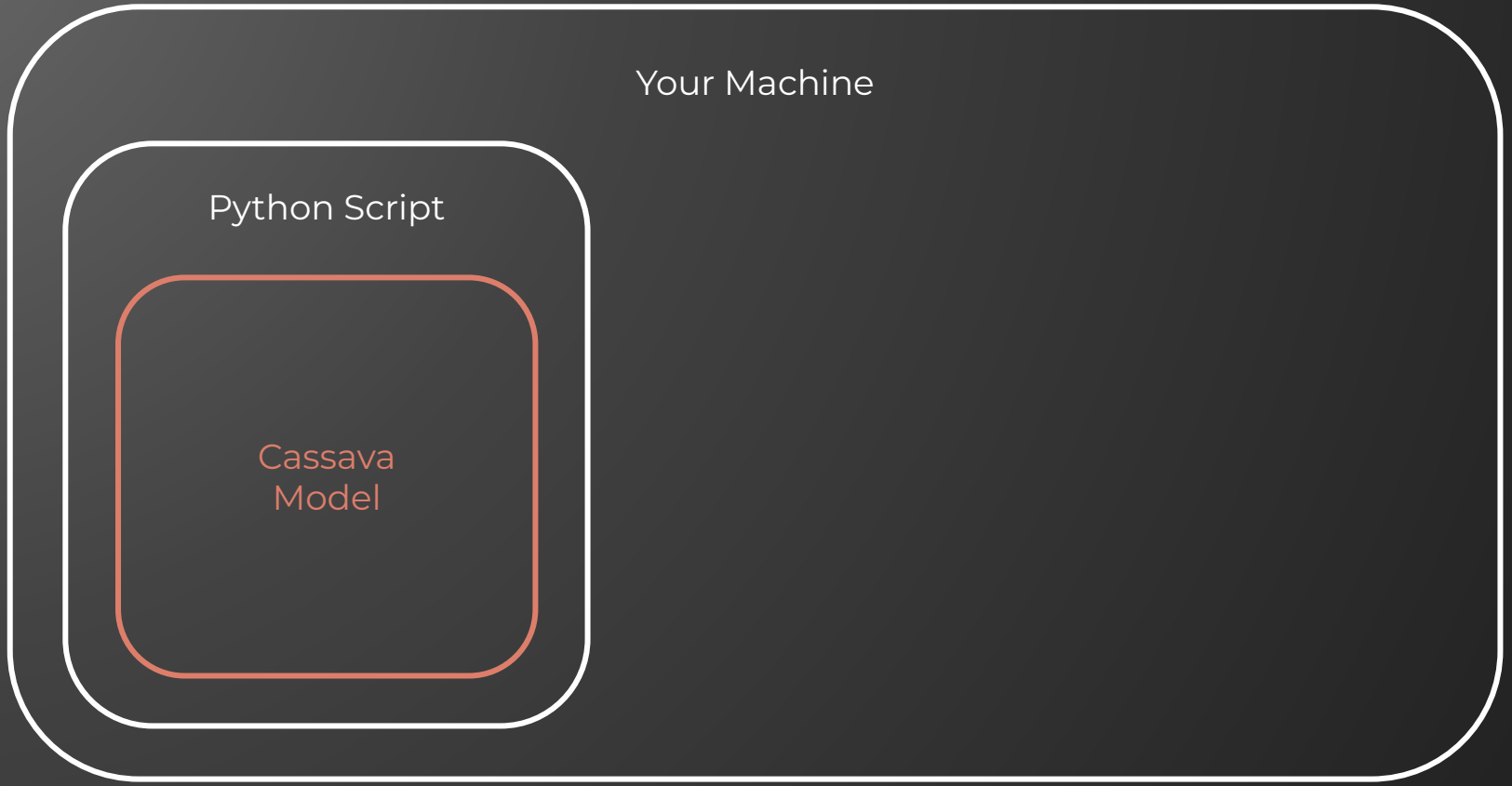
What we'll cover

- How to turn your model into an API
- Containerize your model
- Store your container in a registry
- Deploy your model to a cloud compute cluster
- Scale up your model

SELDON

Demo

The ML Model



SELDON

The model can’t be accessed outside of the python process...

```python
import numpy as np
from flask import Flask, request
import pickle

app = Flask(__name__)
model = pickle.load(open('model.pkl', 'rb'))

@app.route('/predict', methods=['POST'])
def predict():

  int_features = [int(x) for x in request.form.values()]
  final_features = [np.array(int_features)]
  prediction = model.predict(final_features)

  output = round(prediction[0], 2)

  return output

if __name__ == "__main__":
  app.run(debug=True)
```

Why not use Flask or FastAPI?

SELDON

Production deployment challenges

- Maximizing infrastructure usage
- Dependency management
- Working with multiple ML frameworks
- Standardizing API definitions
- Capturing payload structures
- Handling multiple versions of models
- Collecting metrics

SELDON

SELDON®
ML SERVER

SELDON

MLServer

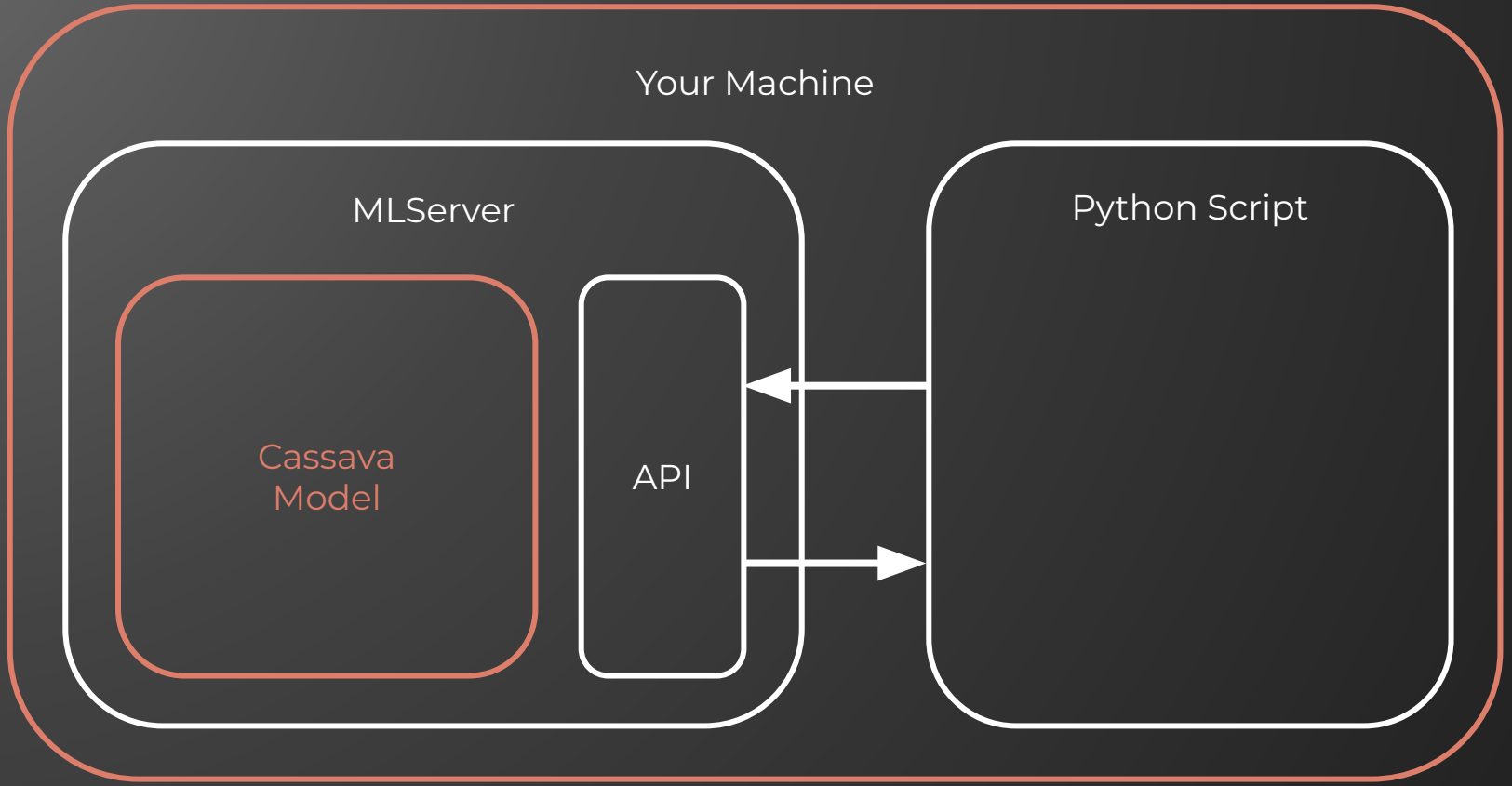
- An Open Source inference server for machine learning models
- Serves models over standardised REST and gRPC interfaces
- Supports popular python based ML frameworks

SELDON

Demo

Turning your model into an API

SELDON



SELDON

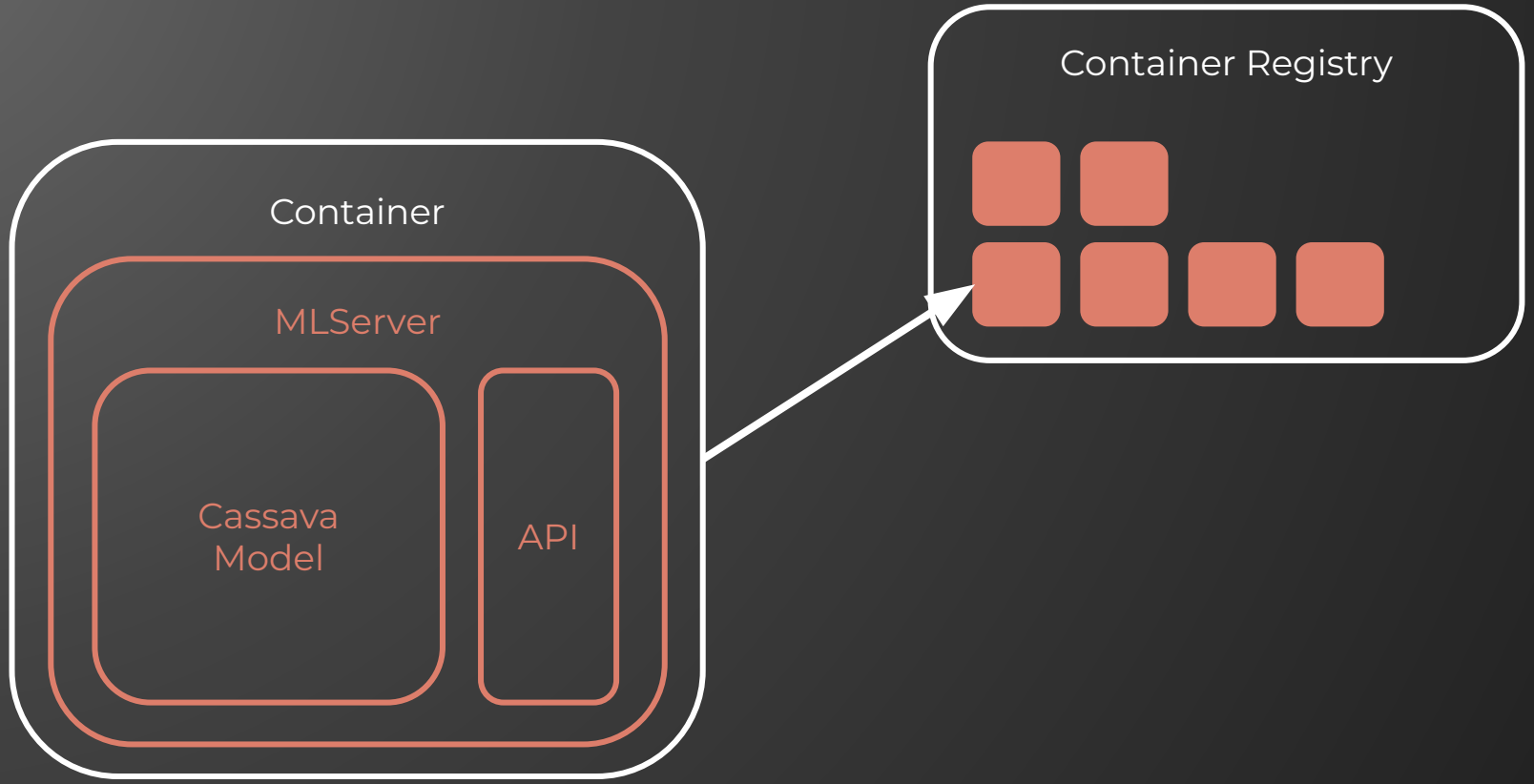
Your model is useless on your machine...

SELDON

Demo

Packaging your model as a container

SELDON

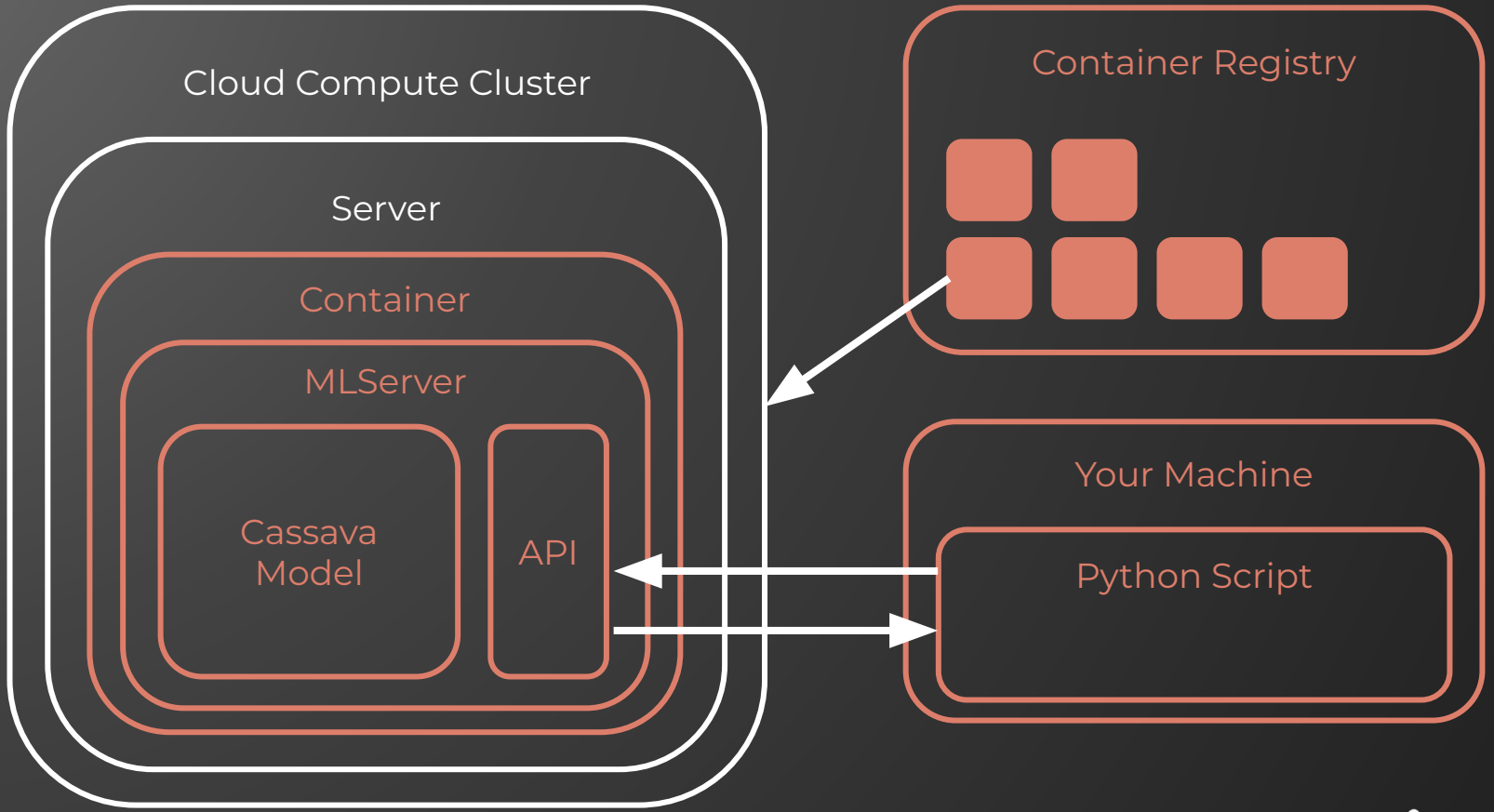


SELDON

Demo

Deploying your model to the cloud

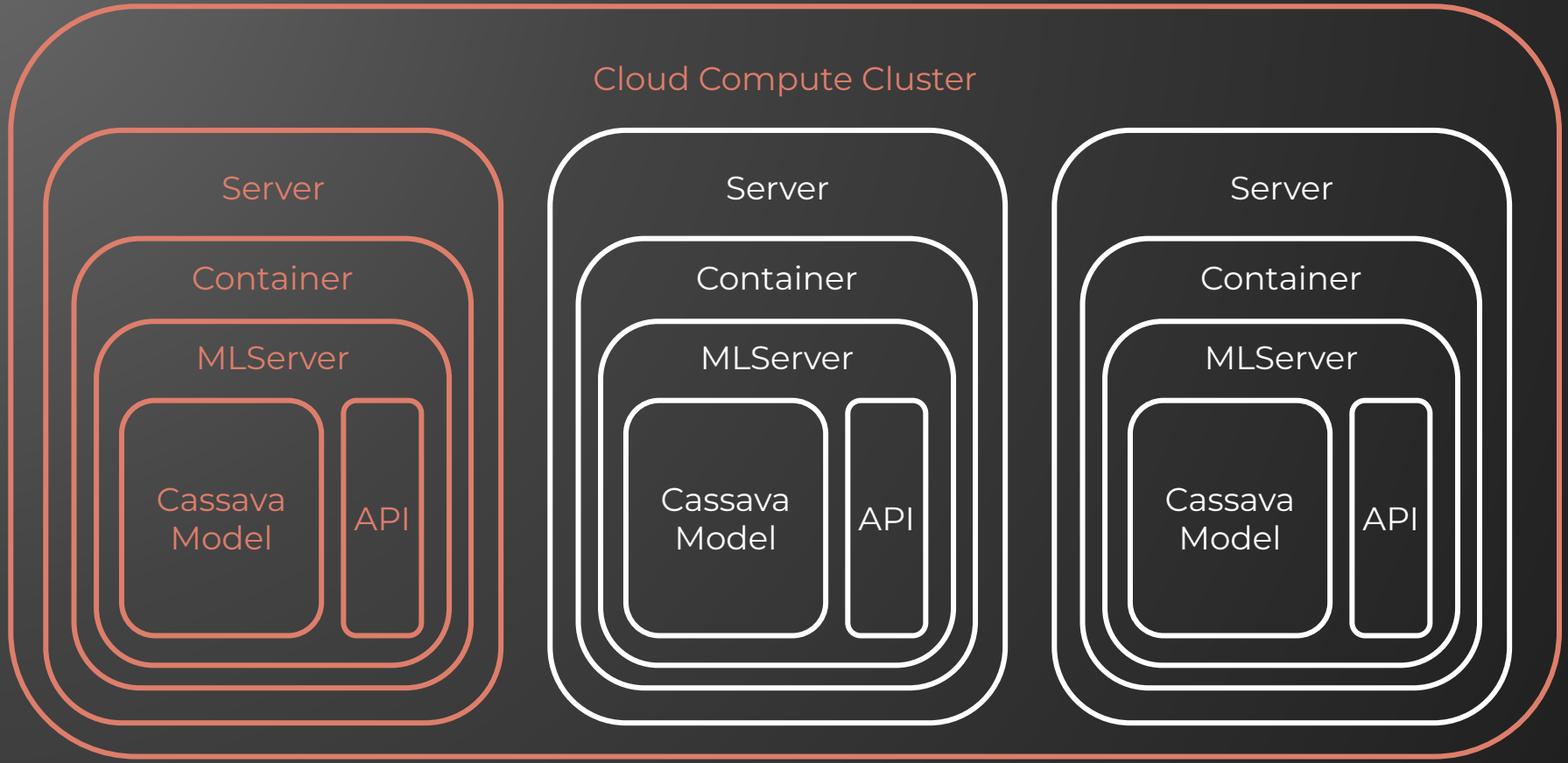
SELDON


Cloud Compute Cluster
Server
Container
MLServer
Cassava Model
API
Container Registry
Your Machine
Python Script
SELDON

Demo

Scaling your model

SELDON


Cloud Compute Cluster
Server
Container
MLServer
Cassava Model
API
Server
Container
MLServer
Cassava Model
API
Server
Container
MLServer
Cassava Model
API

SELDON

Conclusion

- Your model is useless on your machine
- Creating an API makes your model accessible
- Containerization is an easy way to package models
- Deploying to cloud keeps your API running
- Scaling helps you avoid downtime

SELDON

Thank You

Ed Shee


SeldonIO


es@seldon.io


/in/edshee

SELDON