Beyond Inference: Bringing ML into Production

DevConf.CZ 2021

Isabel Zimmerman
Data Science Intern
Forward Deployed AI Engineering

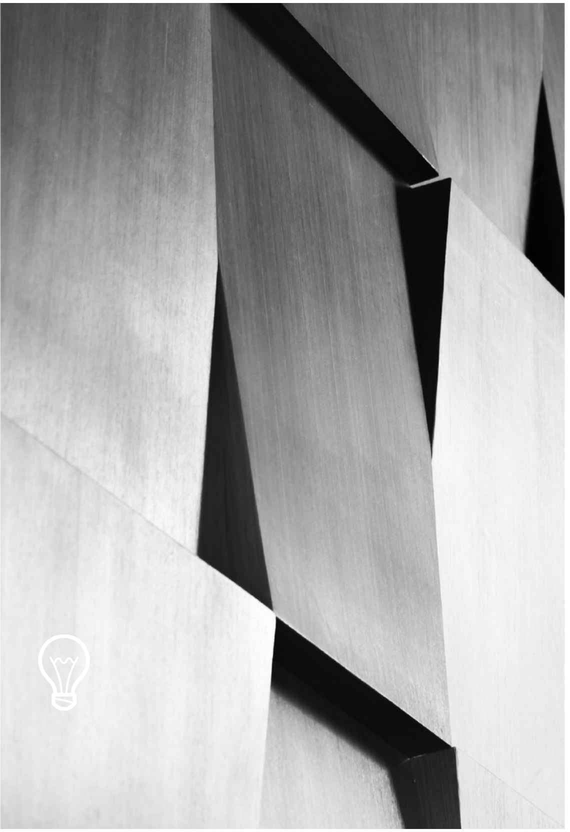
Red Hat

Things to learn

What is model serving?

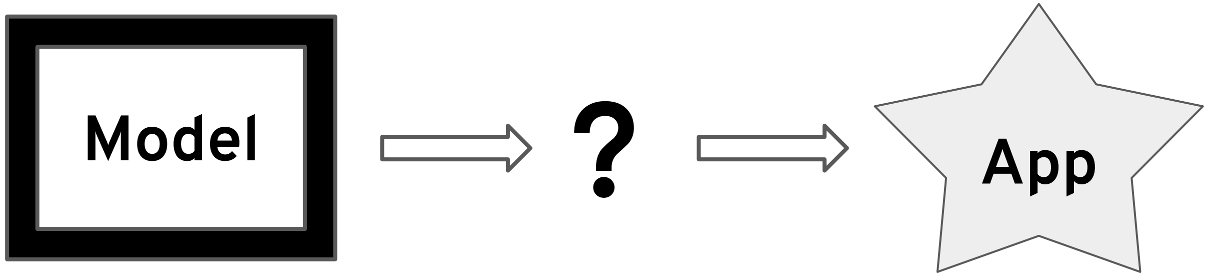
What features of model serving can I utilize?

How can I deploy a model?

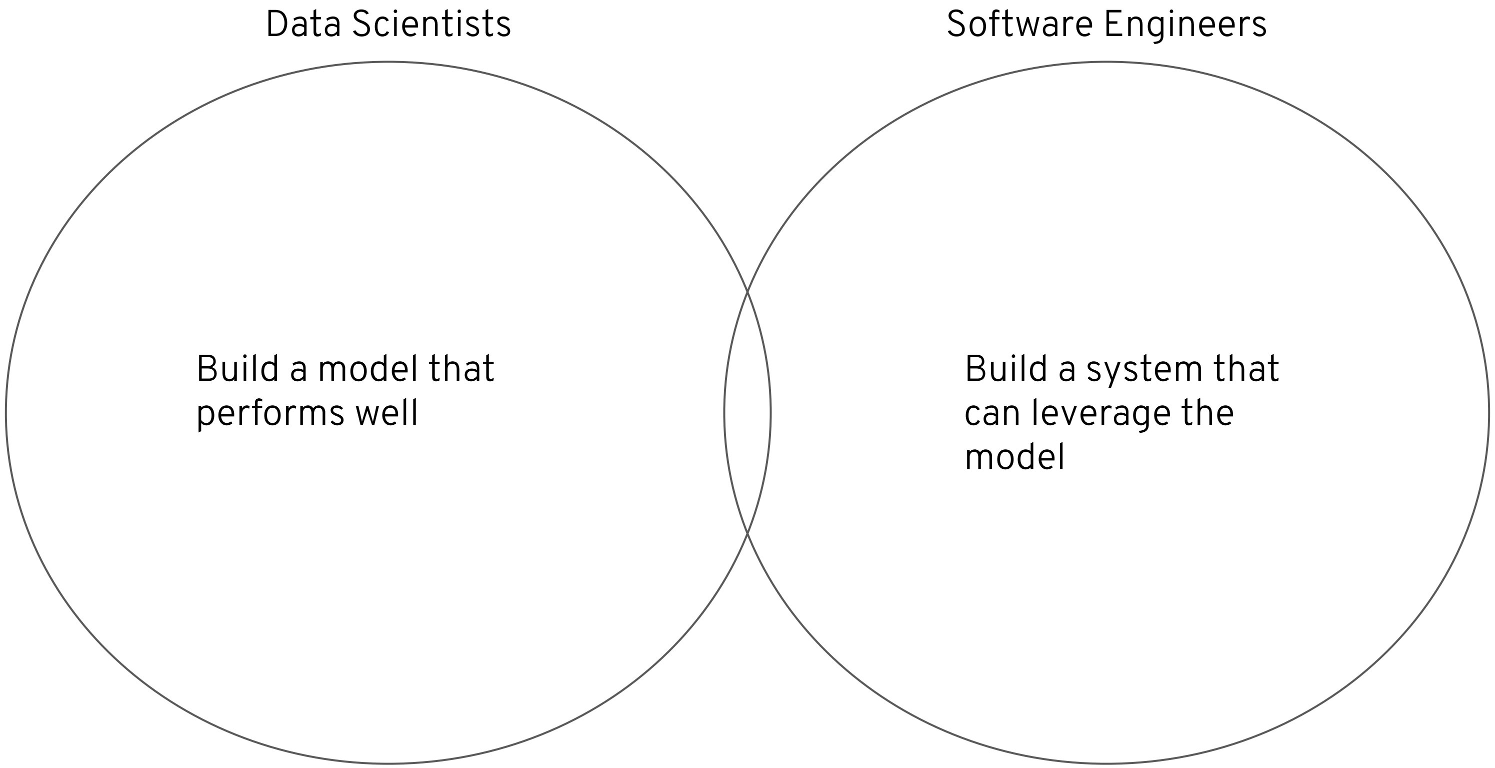


Why model serving?

Why do we need model serving?



Do we need model serving?



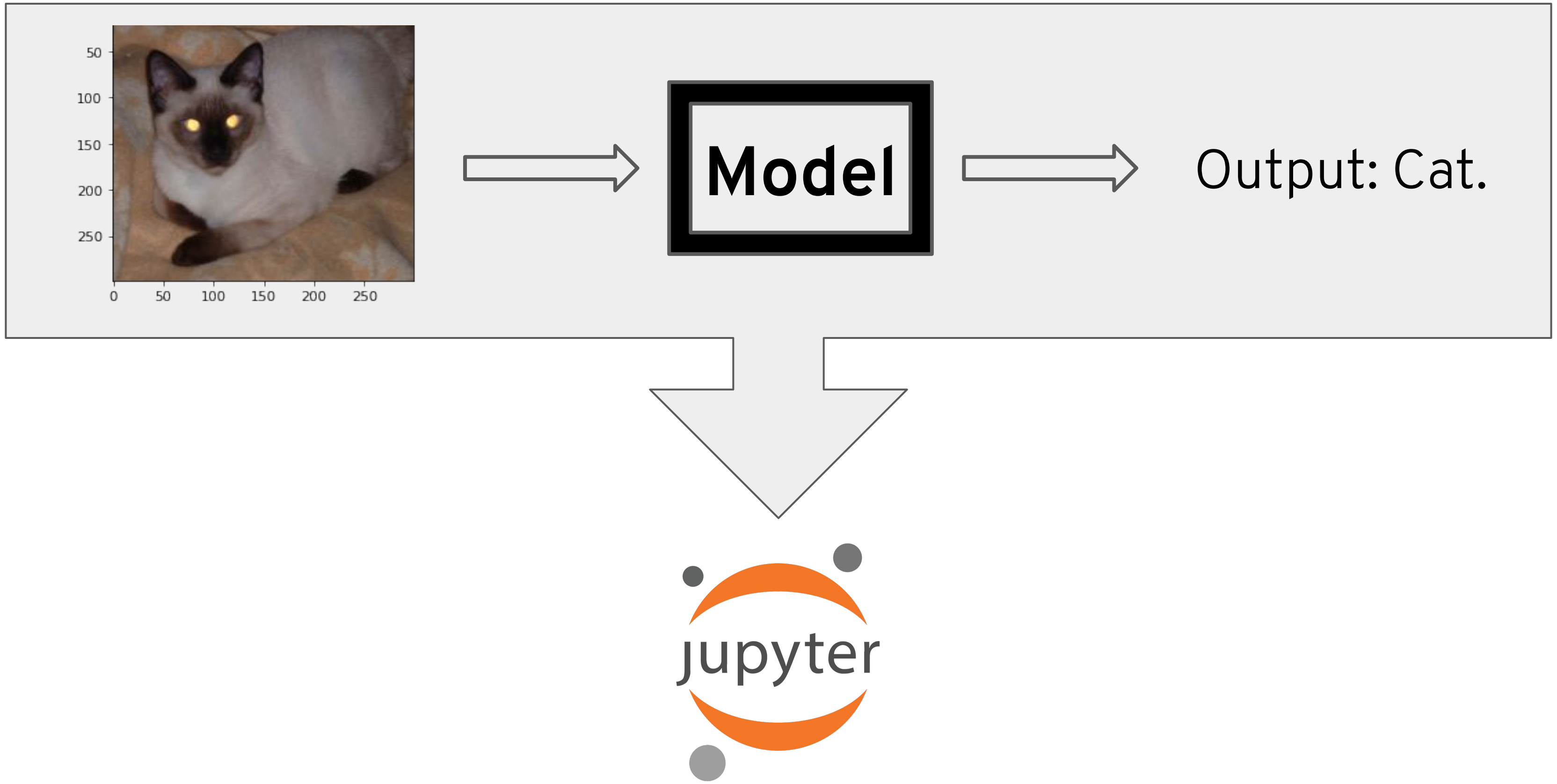
Data Scientists

Build a model that performs well

Software Engineers

Build a system that can leverage the model

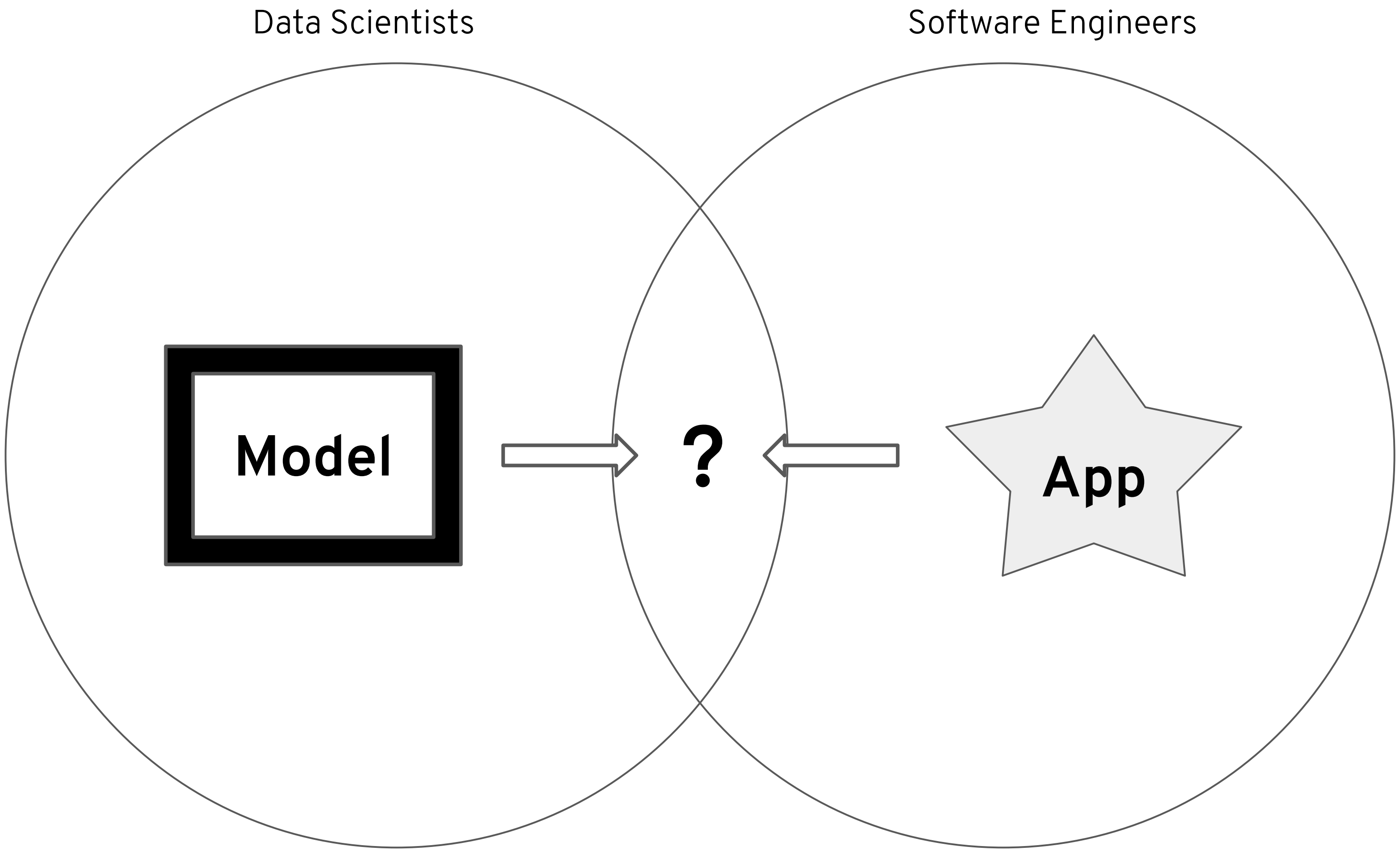
How do models act?



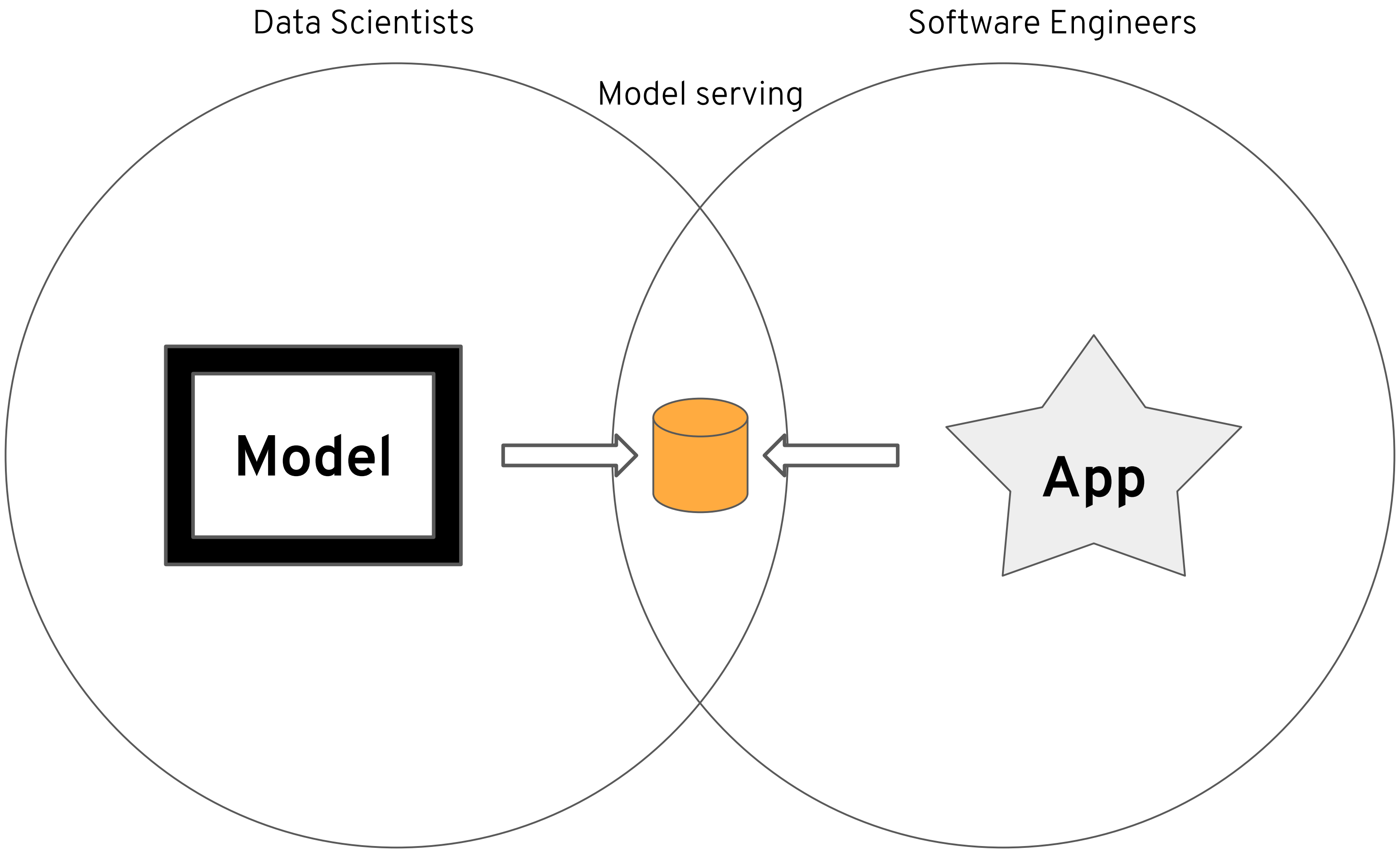
Model

Output: Cat.

What is model serving?



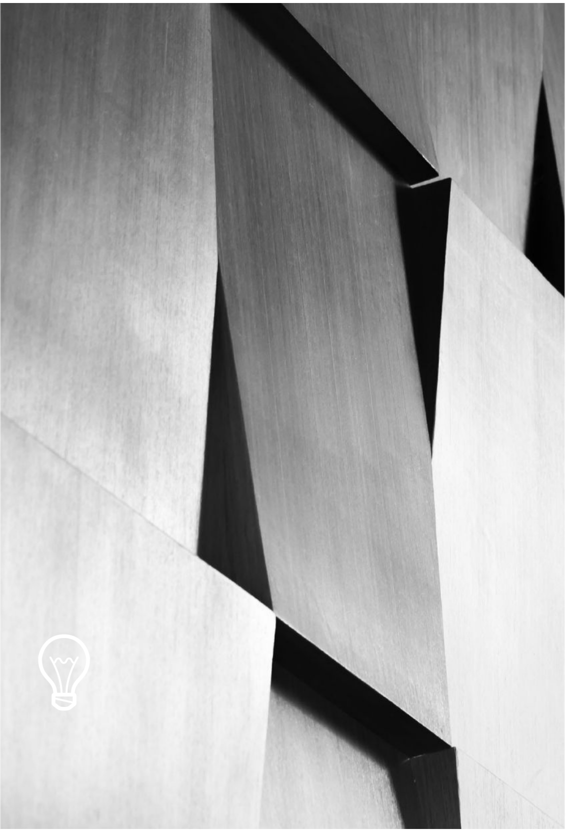
What is model serving?



Data Scientists

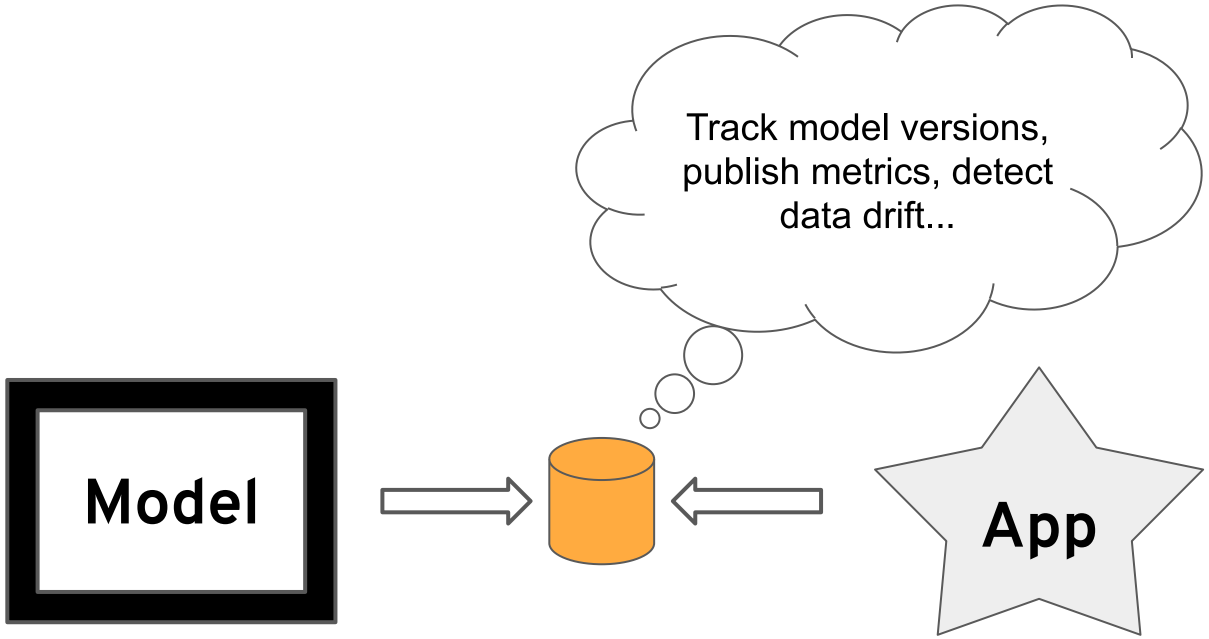
Software Engineers

Model serving



What features should we look for?

How can we make model servers more useful?



Track model versions, publish metrics, detect data drift...

How are you deploying?

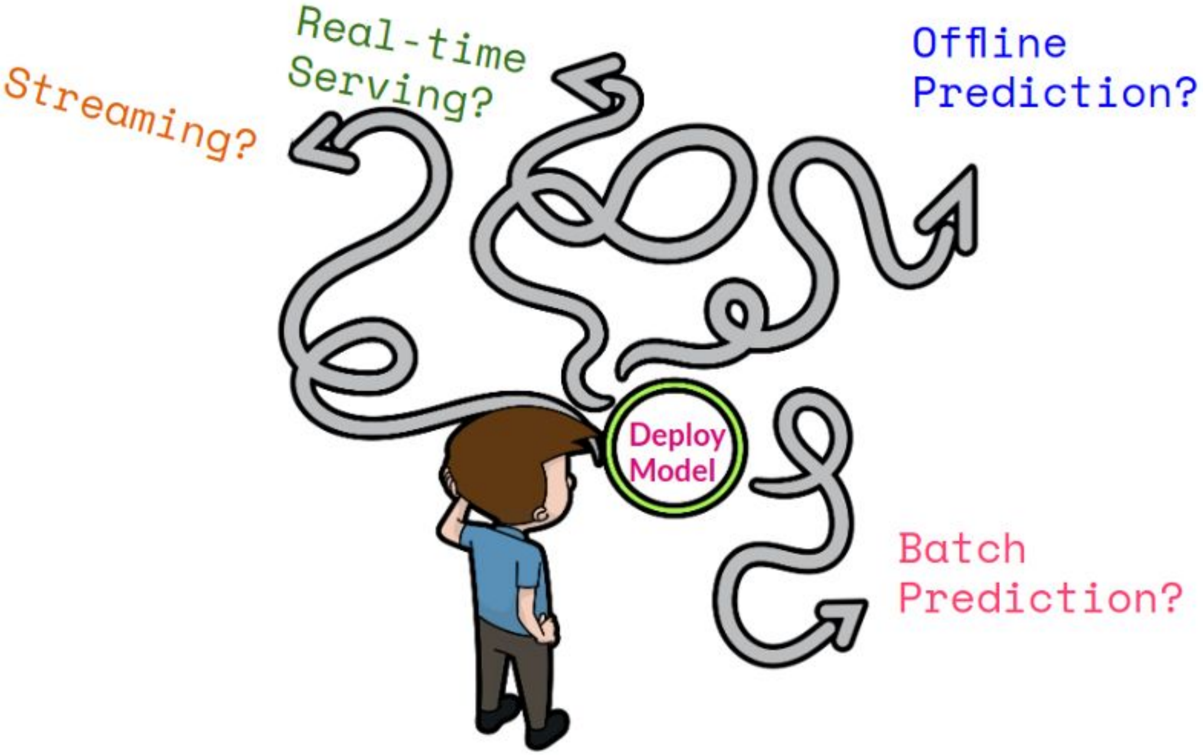
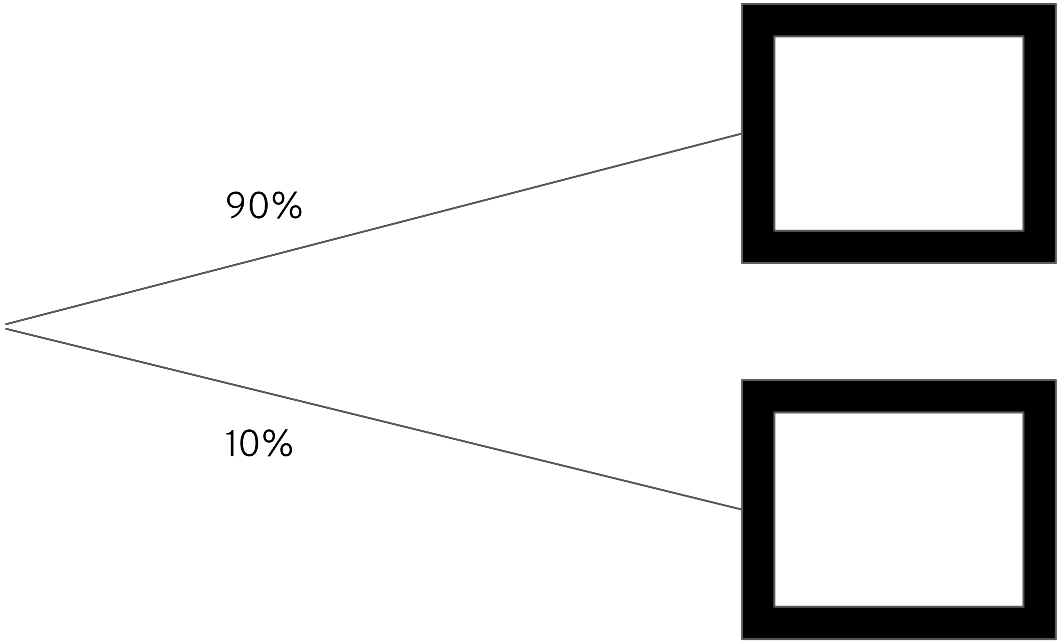
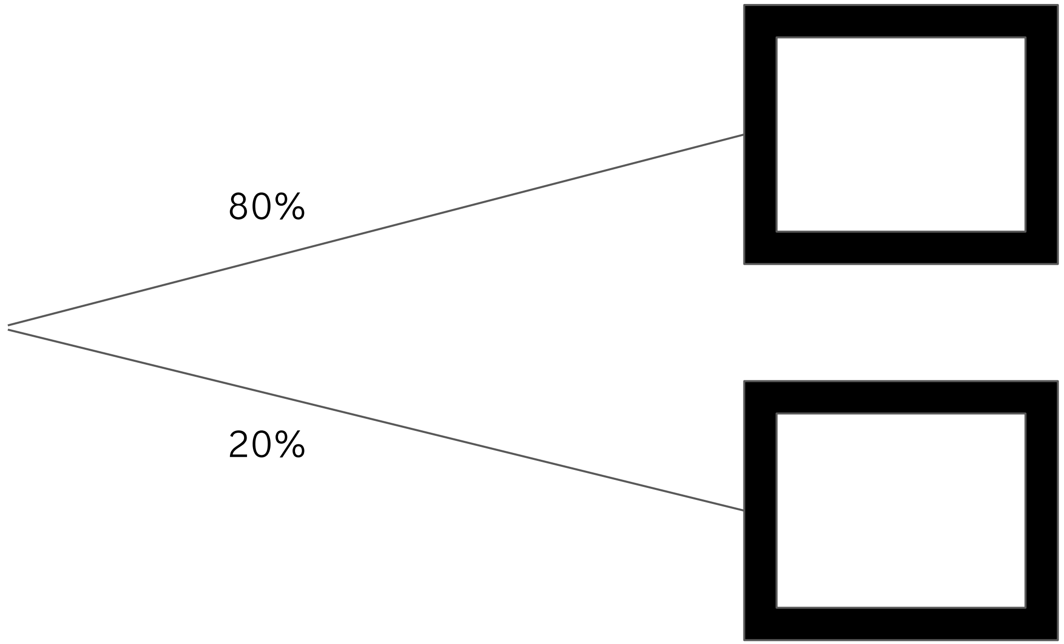


Image based on Clip Art by Vector Toons, CC BY-SA 4.0, via Wikimedia Commons

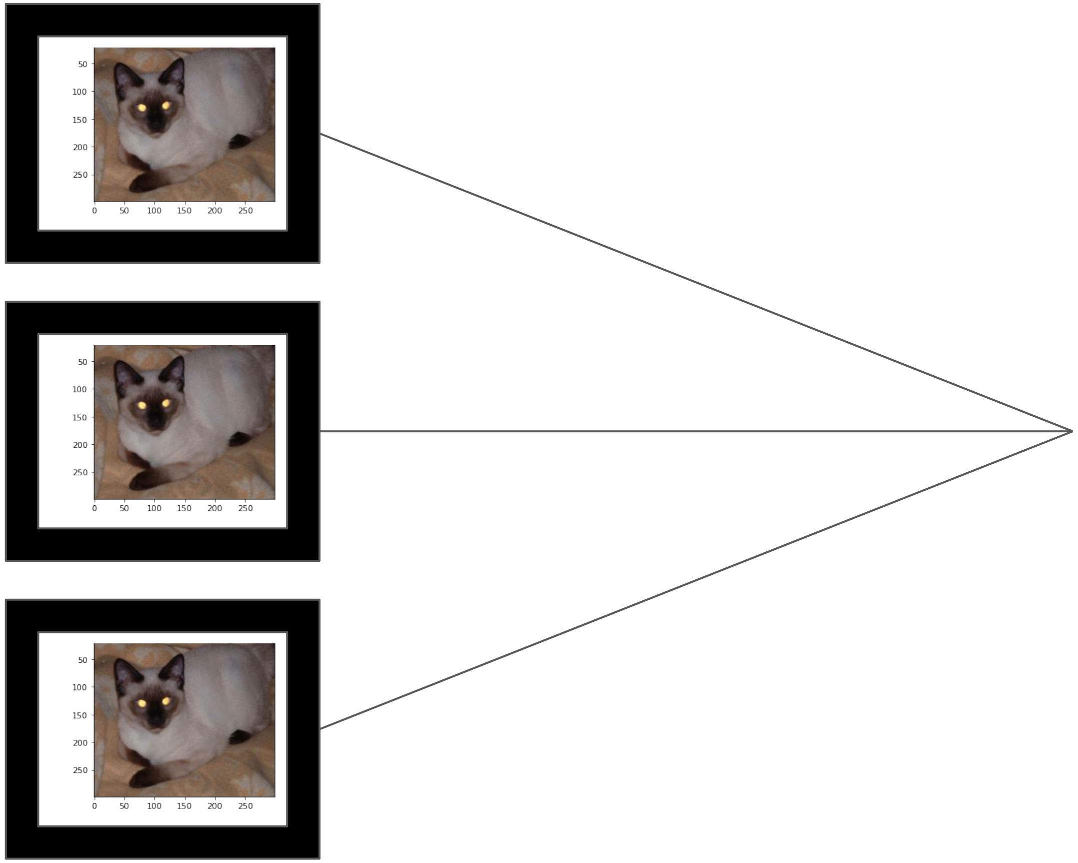
A/B Testing



A/B Testing

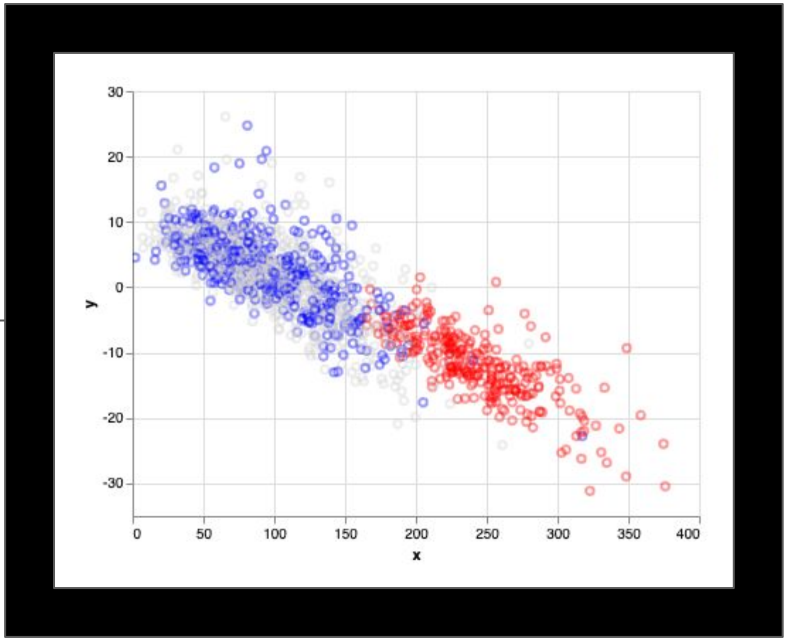


Ensemble



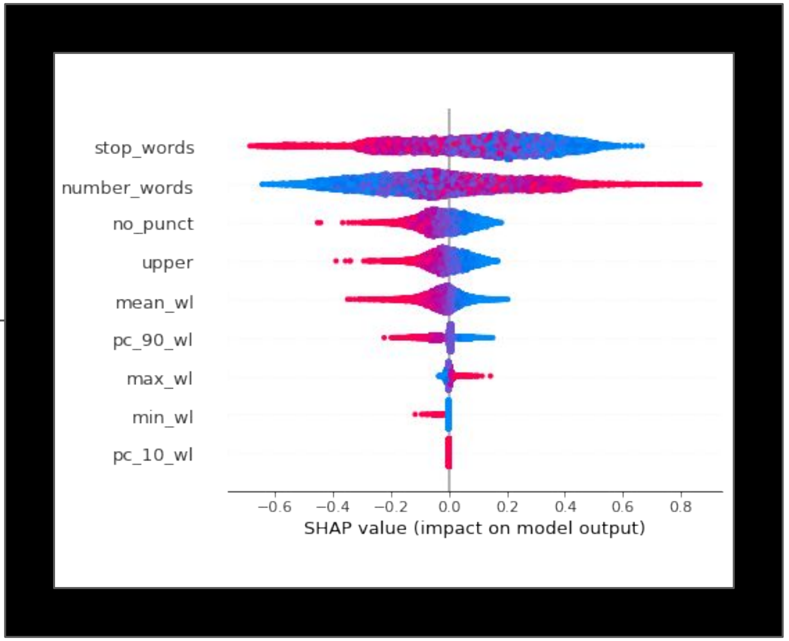
Output: Cat.

Data Drift

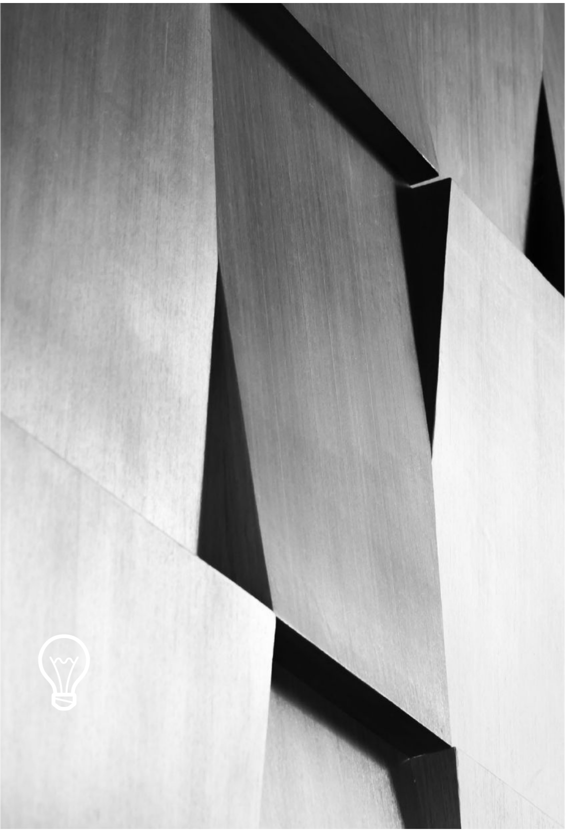


Source:
https://github.com/isabelizimm/openshift-ml-workflows-workshop/blob/develop/source/05-data-drift.ipynb

Explainability



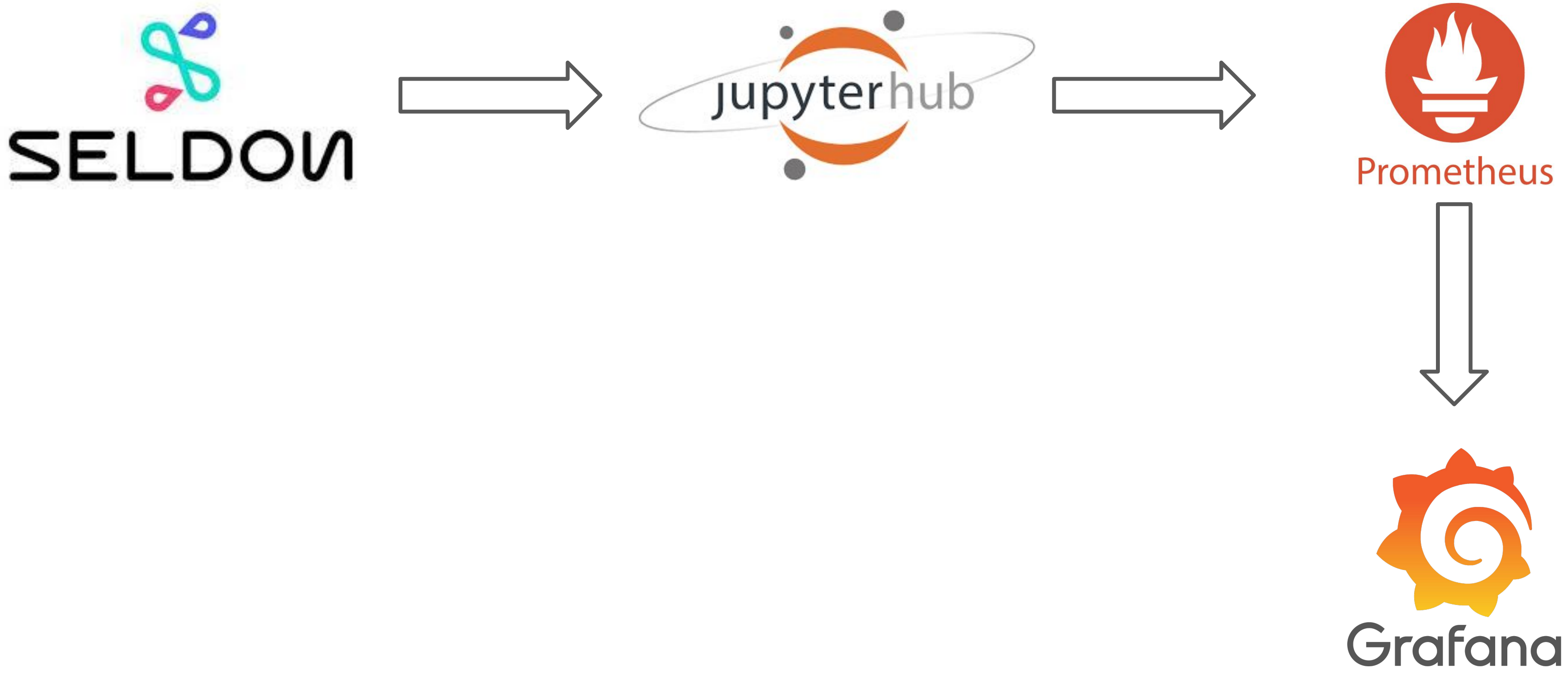
Source:
https://github.com/isabelizimm/openshift-ml-workflows-workshop/blob/develop/source/05-data-drift.ipynb

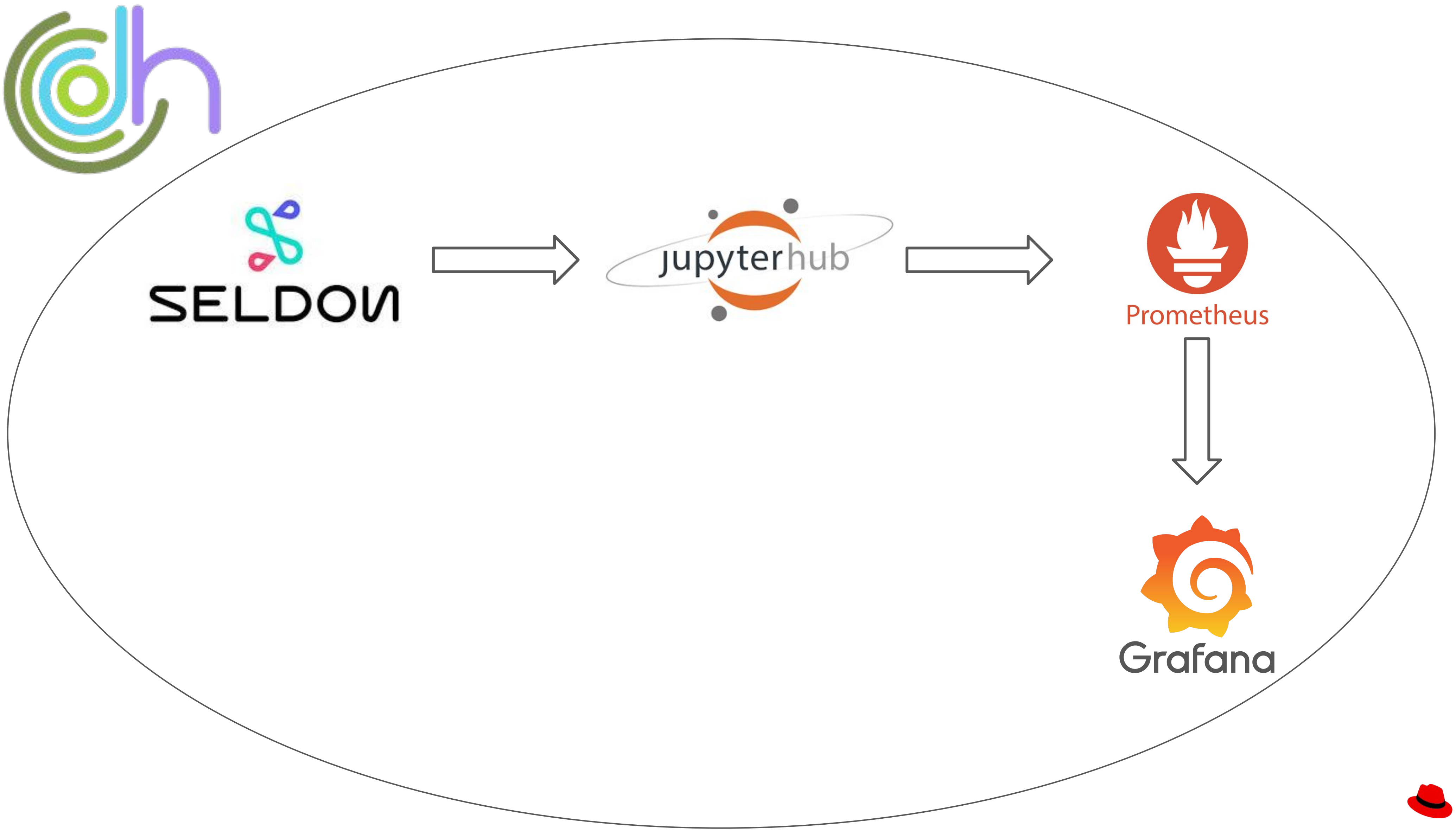


How do I deploy a model?

```yaml
apiVersion: machinelearning.seldon.io/v1
kind: SeldonDeployment
metadata:
  name: image-default
  namespace: seldon
spec:
  name: income
  predictors:
  - graph:
      children: []
      implementation: SKLEARN_SERVER
      modelUri: gs://seldon-models/sklearn/image/model-0.23.2
      name: classifier
    explainer:
      type: AnchorTabular
      modelUri: gs://seldon-models/sklearn/image/explainer-py36-0.5.2
    name: default
     replicas: 1
```

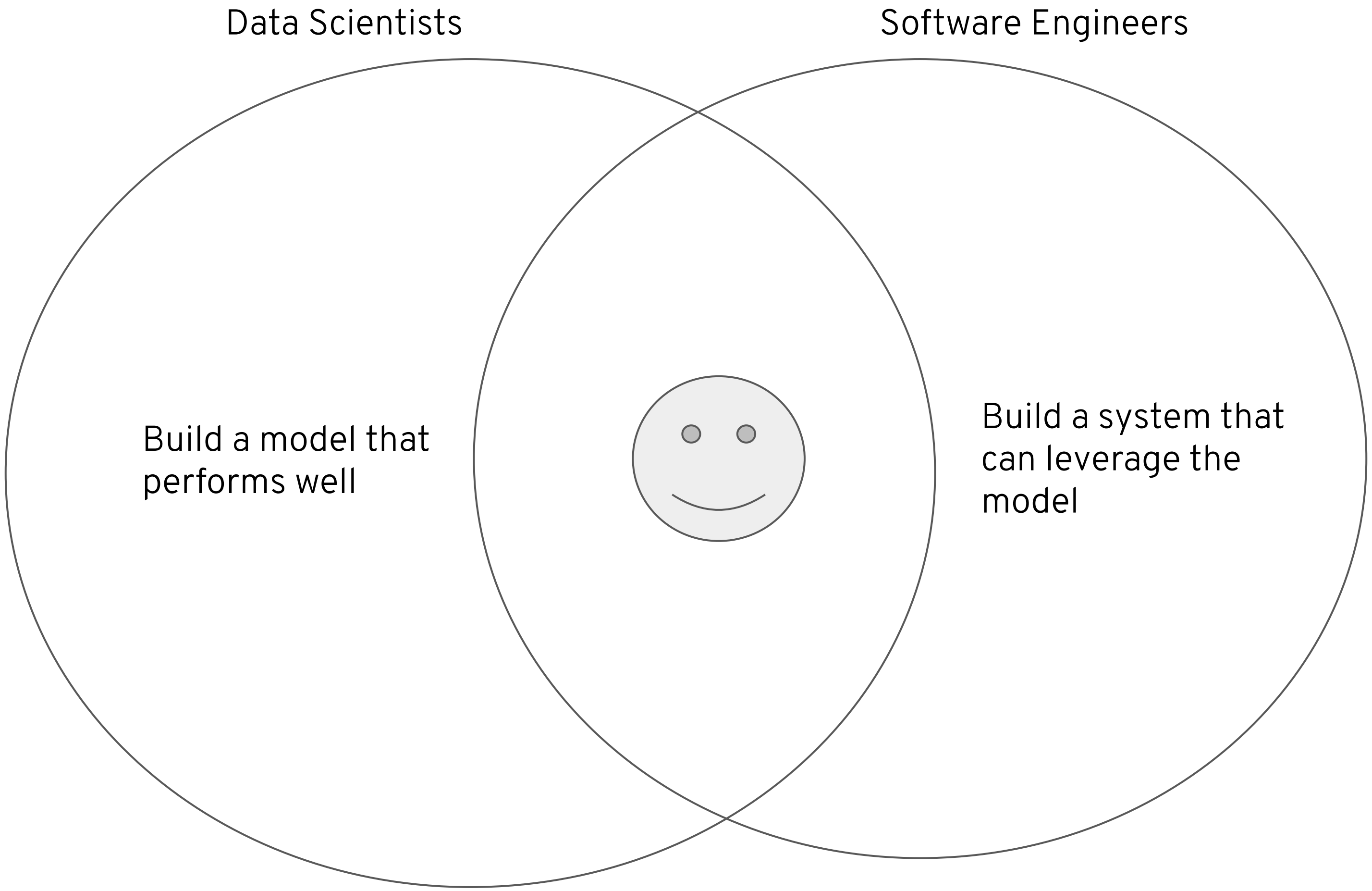
https://docs.seldon.io/projects/seldon-core/en/latest/analytics/explainers.html


SELDON
jupyterhub
Prometheus
Grafana


SELDON
jupyterhub
Prometheus
Grafana

Demo

Do we need model serving?



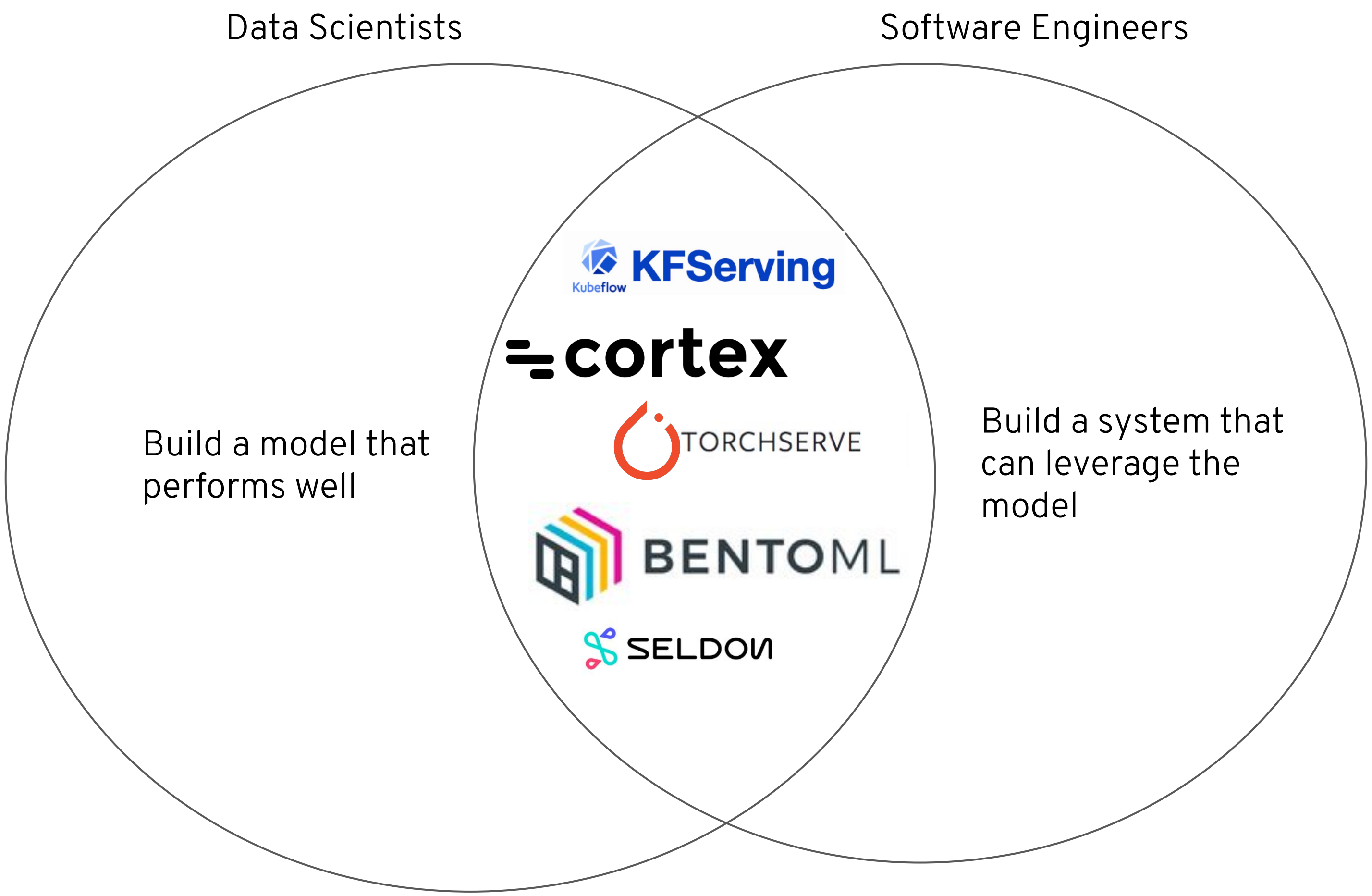
Data Scientists

Software Engineers

Build a model that performs well

Build a system that can leverage the model

Do we need model serving?



Data Scientists

Software Engineers

Build a model that performs well

Build a system that can leverage the model

https://github.com/EthicalML/awesome-production-machine-learning#model-serving-and-monitoring

Stay Connected.

linkedin.com/isabel-zimmerman

twitter.com/isabelizimm