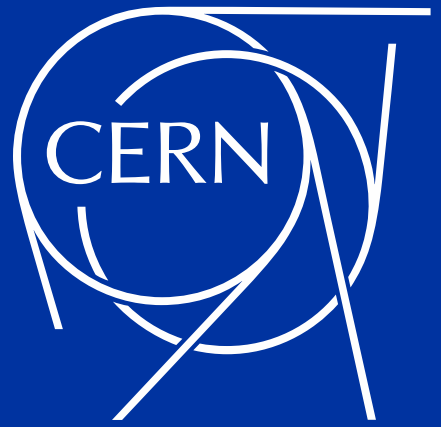


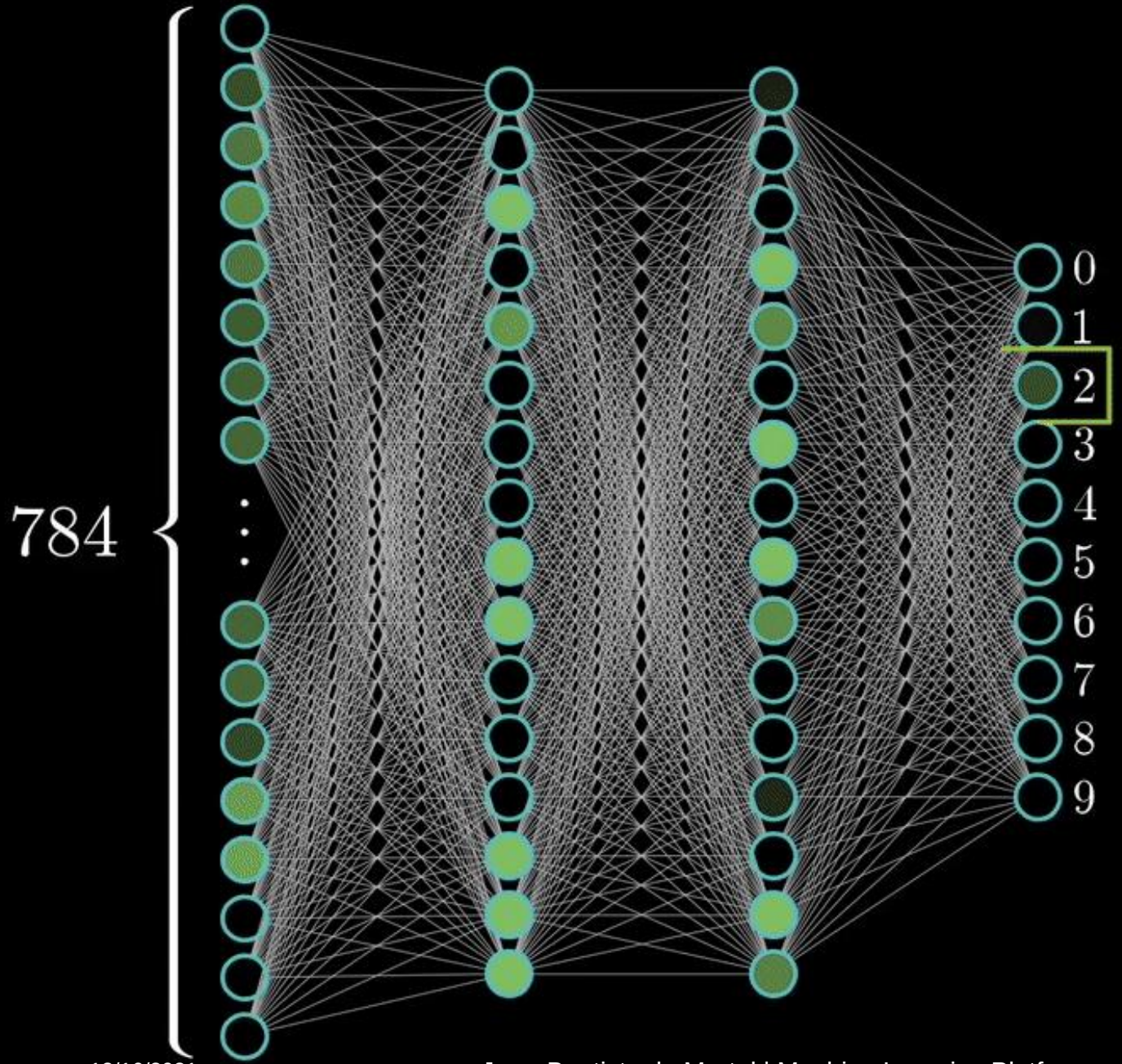
Machine Learning Platform: Deploying and Managing Models in the CERN Control System

Jean-Baptiste de Martel - Nico Madysa, Roman Gorbonosov, Verena Kain

12/16/2021 – CERN meets SLAC

Objectives

- **Provide a common approach to storage, versioning, deployment and usage of models**
- **Accelerate and simplify the model lifecycle by abstracting infrastructural concerns**
- **Fulfill the specific needs of the accelerator control system**
  - Reliability
  - Traceability
  - Security
  - Standardization
- **Stay out of the user's way**
  - Minimize constraints on model developer's workflow
  - Avoid constraints on choice of tools



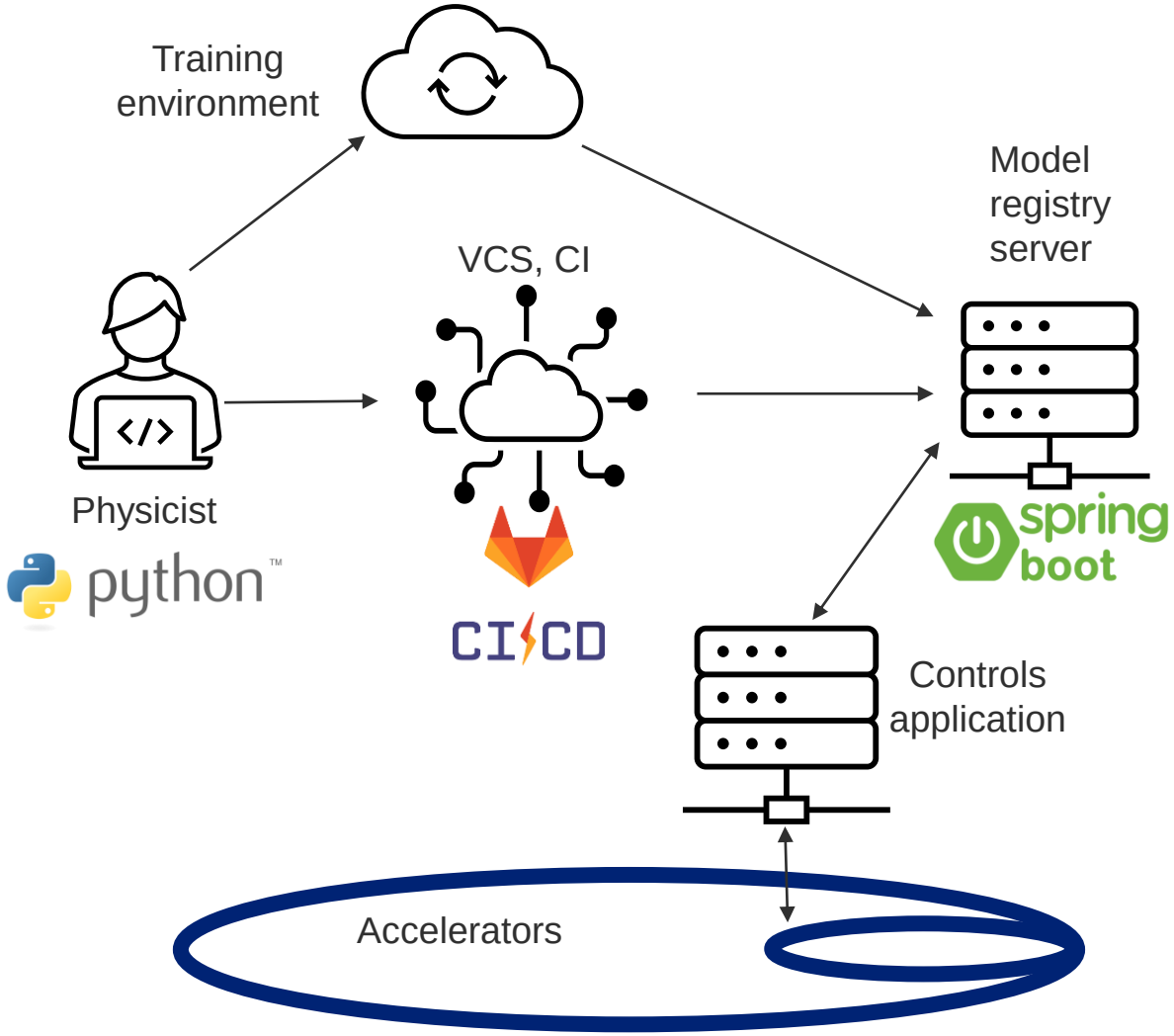
Model parameters:
"Trained weights"
- values assigned to the neurons and connections after training

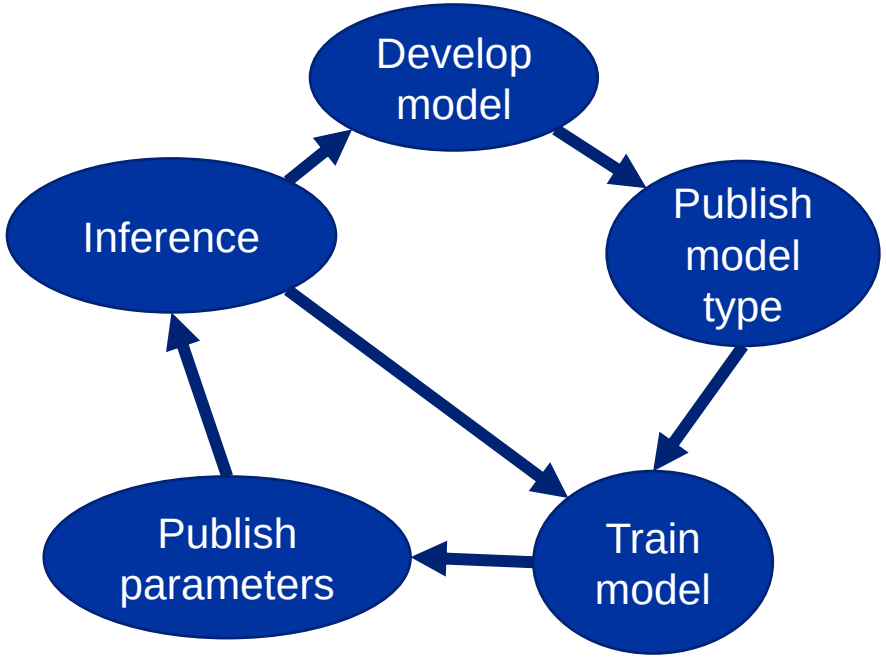
Model type:
Layout/architecture of the neural network –
i.e., number of neurons, how they are connected, etc…

Model:
Combination of a model type and model parameters

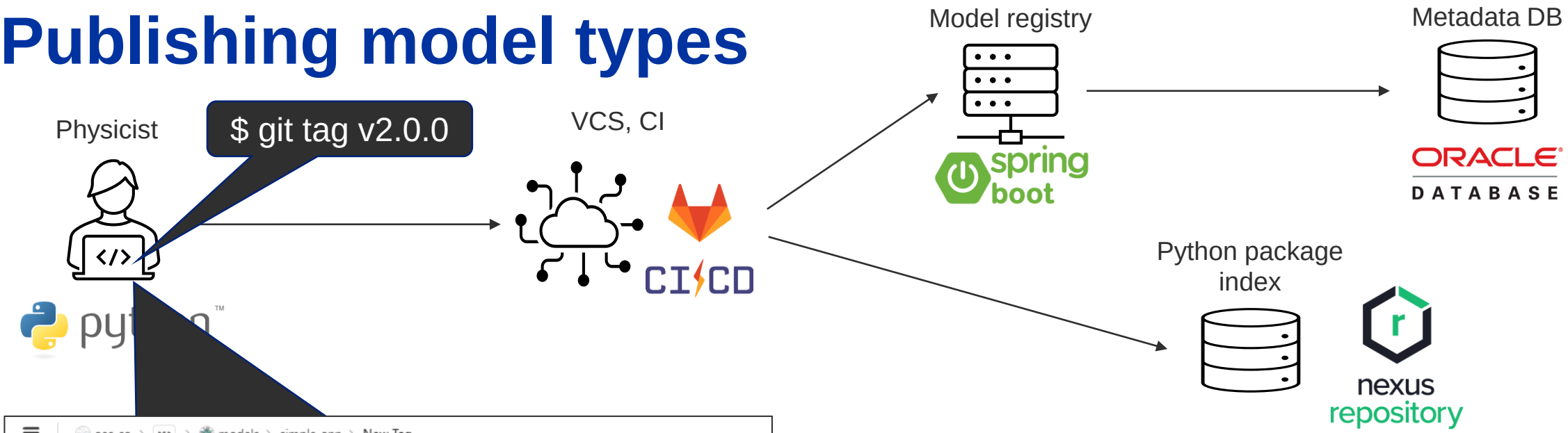
Source: https://www.youtube.com/watch?v=aircAruvnKk

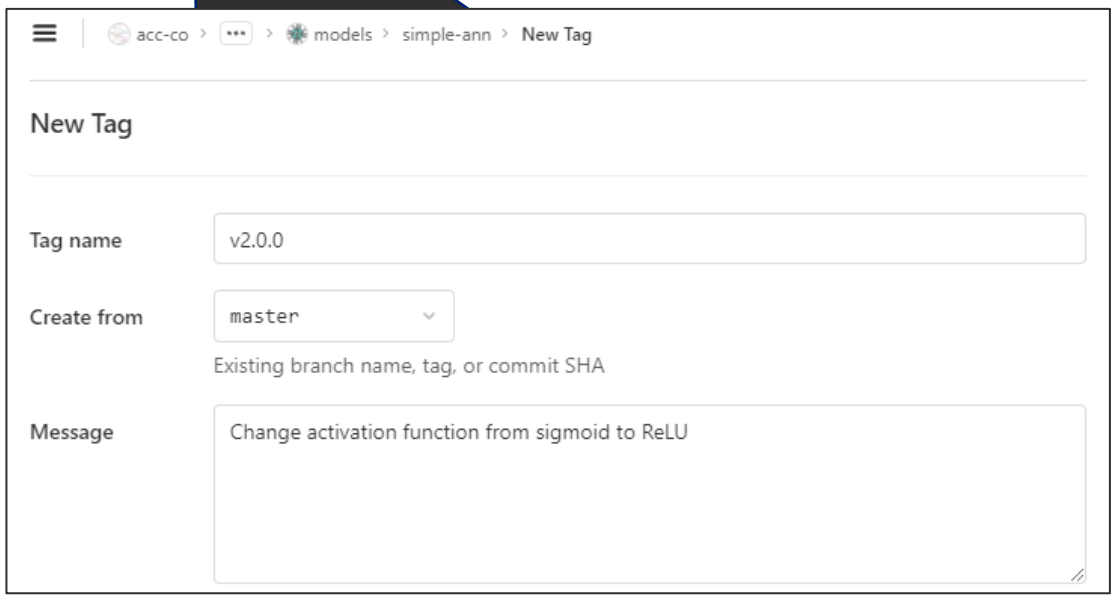
Development workflow





Publishing model types

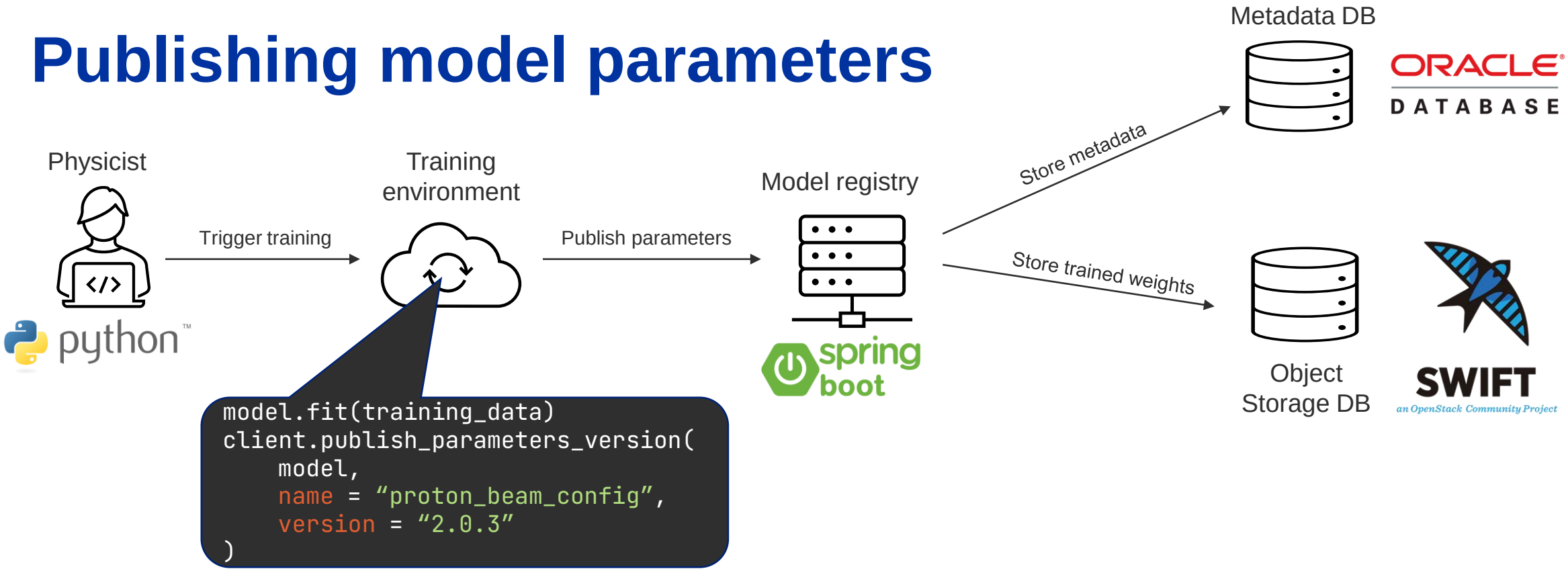




Advantages

- **Access control and traceability for model types**
- **Quick & easy, no need to learn new tools, complexity is hidden**

Publishing model parameters



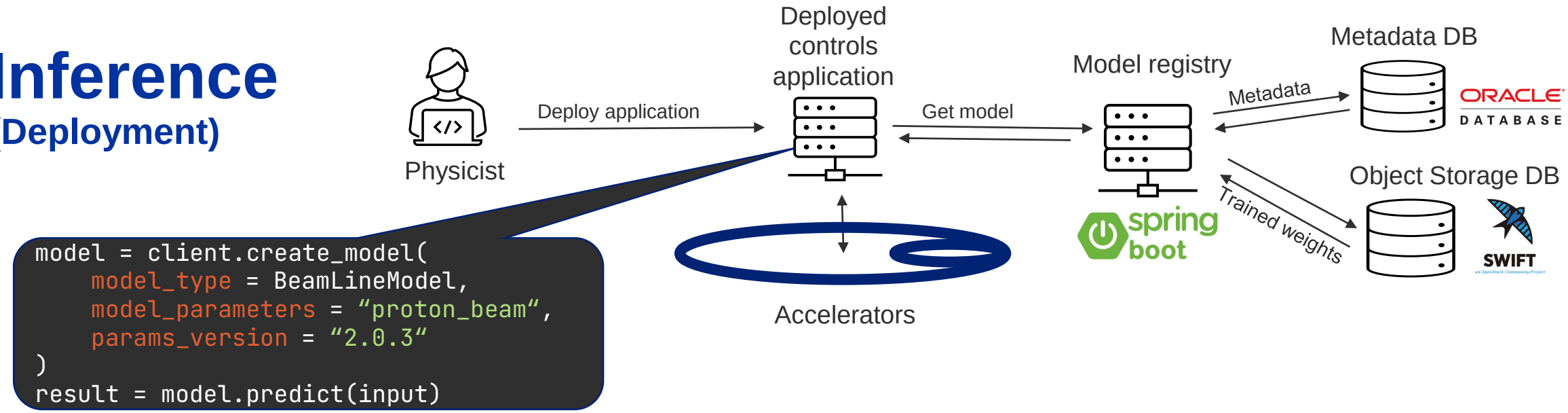
Usage

- **Choose parameters name and version**
- **Use the client library to publish**

Advantages

- **All parameters stored centrally and reliably**
- **Compatibility is fully managed**

Inference
(Deployment)



Usage

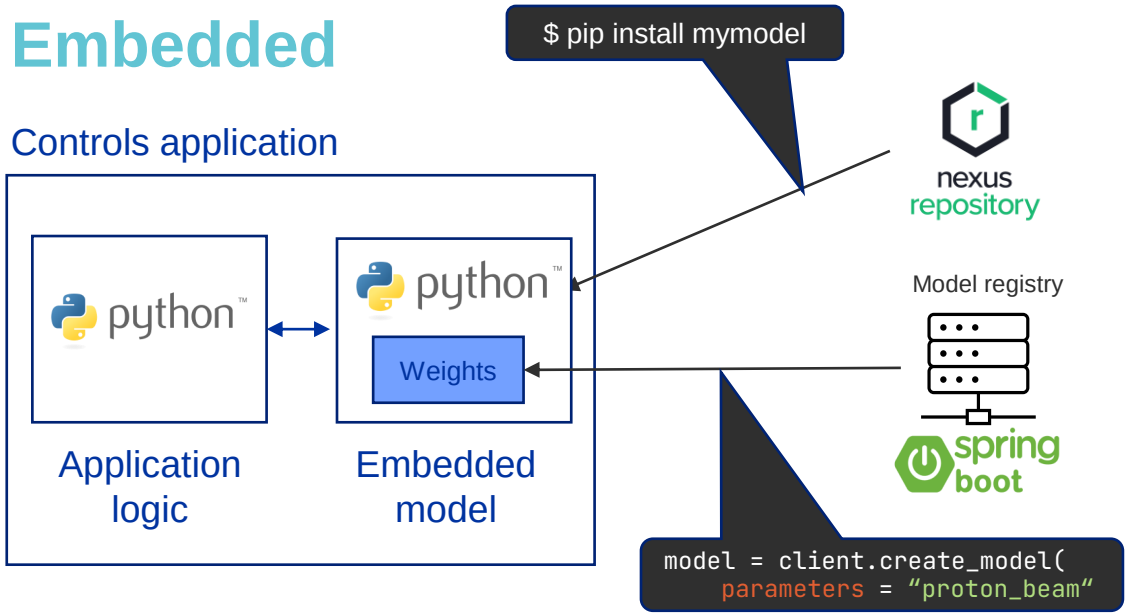
- **Use the MLP client library to instantiate the model**
- **Provide model type, parameters name and version**

Advantages

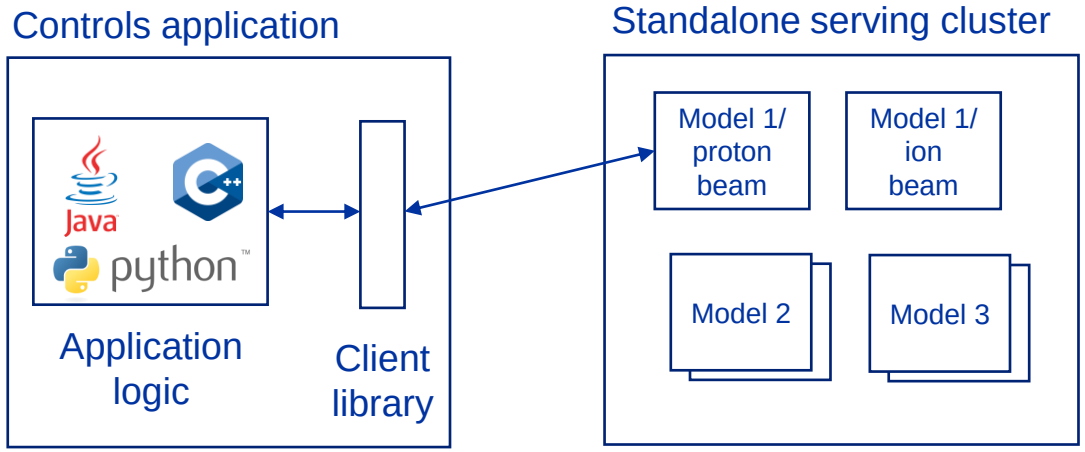
- **Parameters retrieved and loaded transparently**
- **Parameter traceability**

Embedded vs standalone deployment

Embedded



Standalone



- **Python only**
- **Model type must be installed**
- **Parameters retrieved then stored locally**

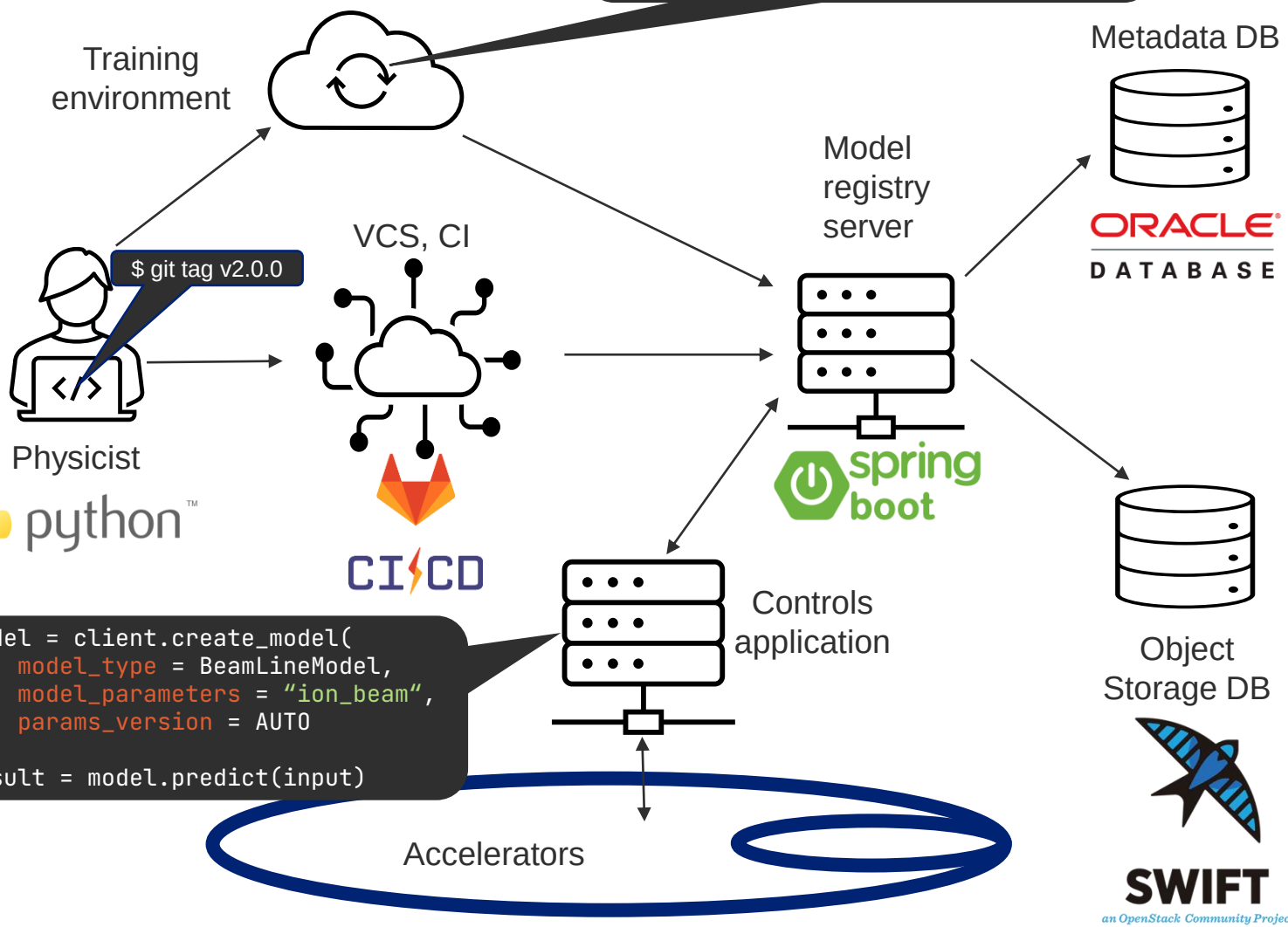
- **Call models from any language**
- **Install mlp-client only**
- **Everything happens remotely**

Conclusion

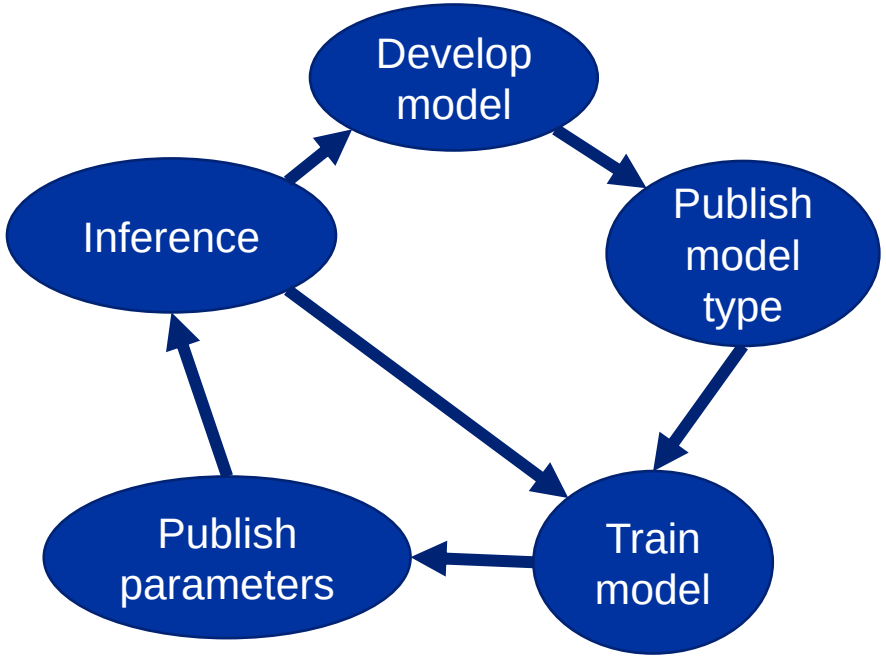
- **We want to help physicists develop models faster and unburden them from infrastructural concerns while minimizing constraints**
- **We also want to apply software engineering best practices to ensure reliability and maintainability of the control system**
- **MLP provides a basis to achieve these goals and is now being adopted**
- **Could not cover everything, simplified a lot – please contact us offline!**
  - jean-baptiste.de.martel@cern.ch
  - nico.madysa@cern.ch
  - roman.gorbonosov@cern.ch

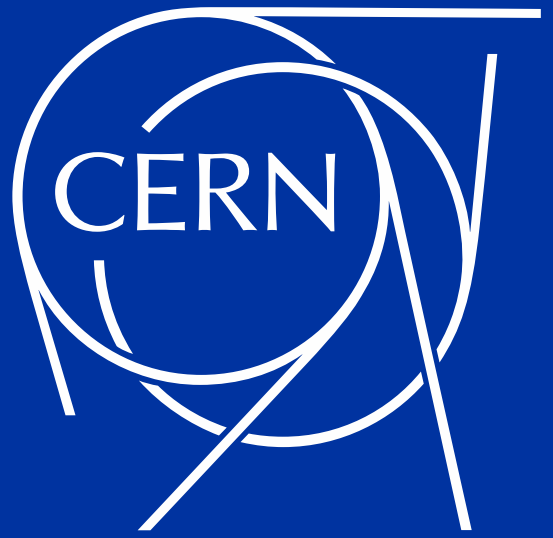
Thank you !

```
model.fit(training_data)
client.publish_parameters_version(
    model,
    name = “ion_beam”,
    version = AUTO # generated
)
```




```
model = client.create_model(
    model_type = BeamLineModel,
    model_parameters = “ion_beam“,
    params_version = AUTO
)
result = model.predict(input)
```

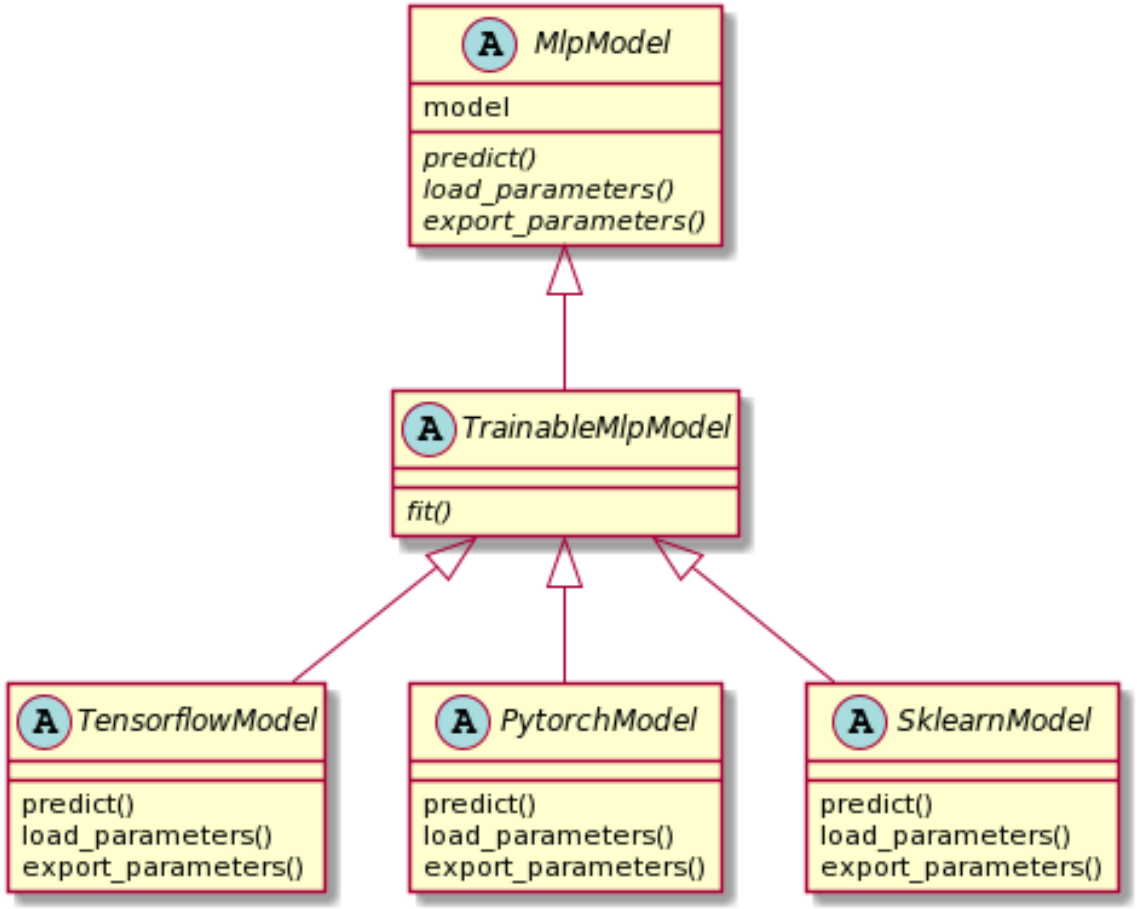



CERN

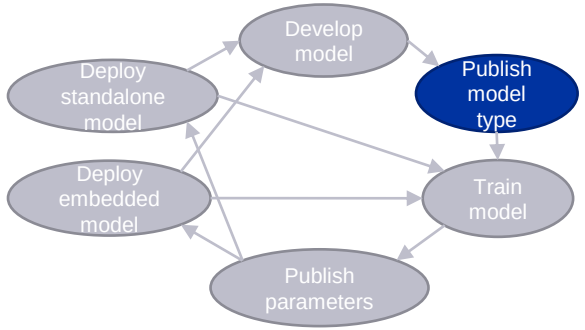
Reserve slides

Implementing a model

- **We define a common API for all controls models**
  - shared abstraction layer
- **Interface defines 4 methods:**
  - *Fit* – train the model on the provided data
  - *Export parameters* – extract current values of all model parameters
  - *Load parameters* – configure the model using the provided parameters
  - *Predict* – return a prediction from the input data
- **Default extensible implementations are available for common frameworks**



Publishing model types



Custom jobs

...

test

lint

triggers

push tag v1.0.2

Model developer

CI/CD

MLP jobs

build Python wheel

Publish wheel to artifact registry

Packaging & publication

scan entry points

modelA

register MTV **modelA v1.0.2**

modelB

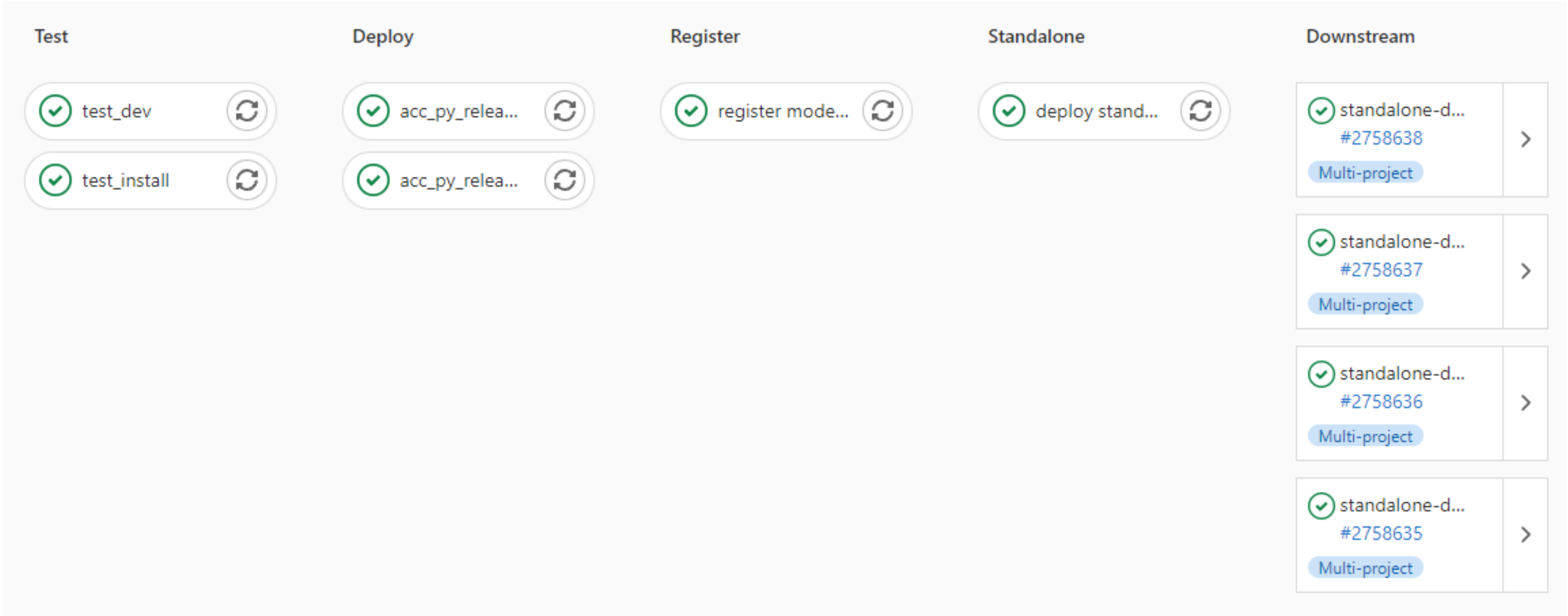
register MTV **modelB v1.0.2**

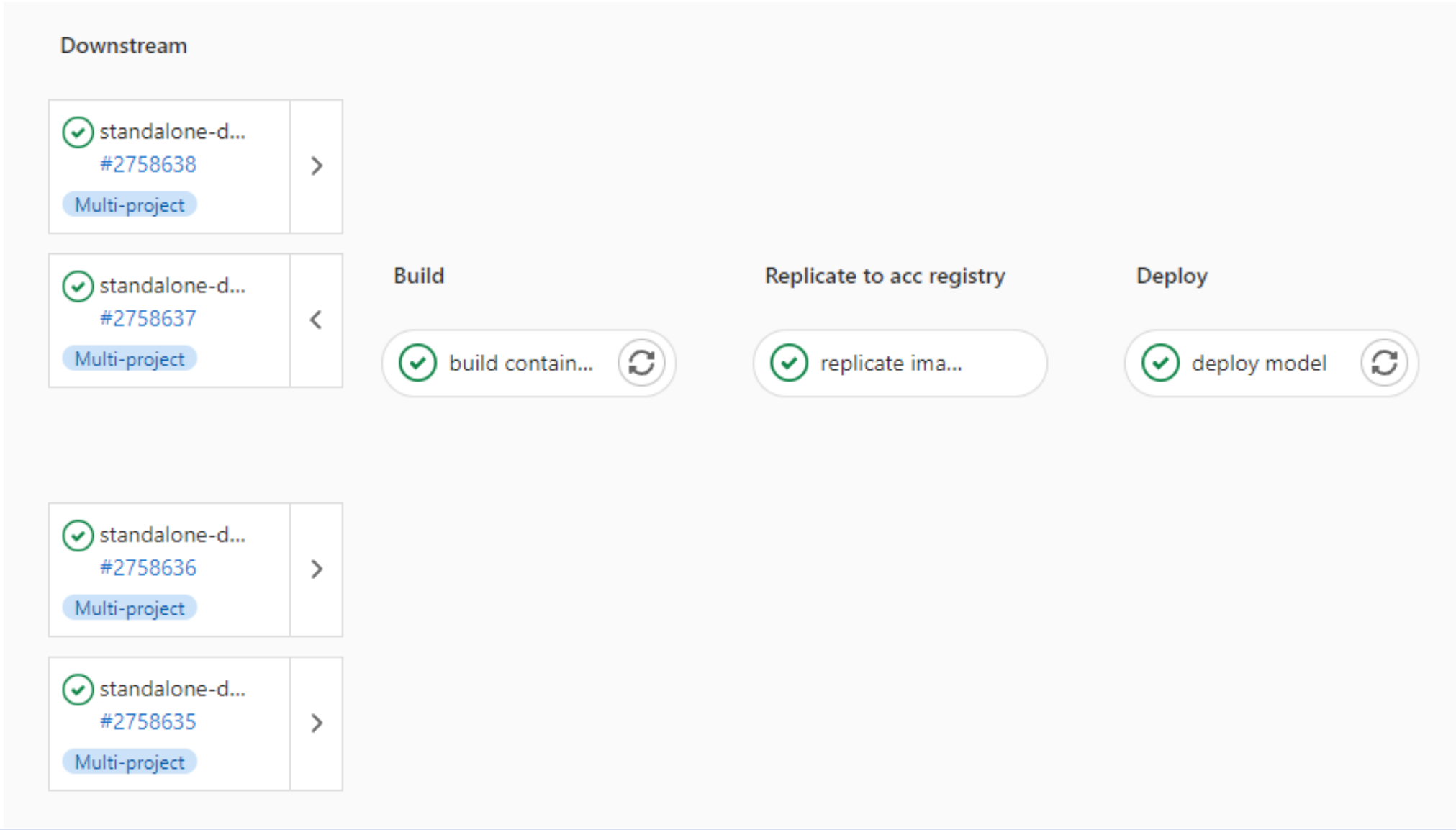
Model registration

Model parameters version number generation

Model type version	Highest existing parameters version	->	Generated parameters version
1.0.0	None exist yet	->	1.0
1.0.0	1.0	->	1.1
1.6.0	1.1	->	1.2
2.0.0	1.2	->	2.0
3.3.0	4.0 (no 3.x)	->	ambiguity
3.3.0	4.0 (3.3 exists)	->	3.4

Standalone deployment CI




Downstream
standalone-d...
#2758638
Multi-project
standalone-d...
#2758637
Multi-project
Build
Replicate to acc registry
Deploy
build contain...
replicate ima...
deploy model
standalone-d...
#2758636
Multi-project
standalone-d...
#2758635
Multi-project

Compatibility