



KubeCon



CloudNativeCon

China 2018

Introduction to Modern Data Science

Sam Kreter



Code and Slides



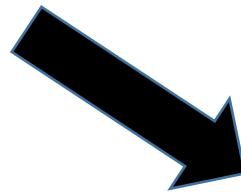
github.com/samkreter/KubeconAsia2018

The Process

Business Need /
Problem Discovery

The Process

Business Need /
Problem Discovery



Development



KubeCon



CloudNativeCon

China 2018

The Process



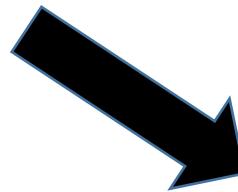
KubeCon



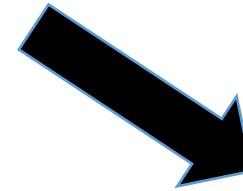
CloudNativeCon

China 2018

Business Need /
Problem Discovery



Development



Production /
Actual User Impact

The Process



KubeCon



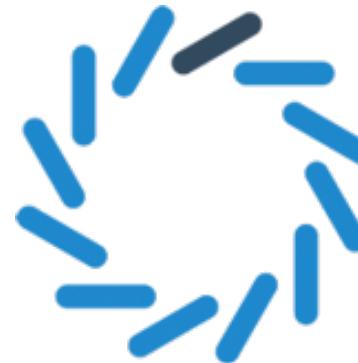
CloudNativeCon

China 2018

Business Need /
Problem Discovery



Production /
Actual User Impact



DOMINO
Pandas



The Process



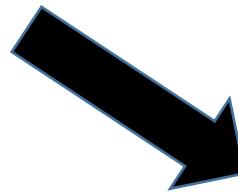
KubeCon



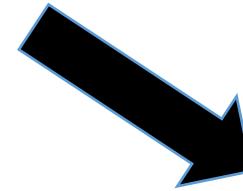
CloudNativeCon

China 2018

Business Need /
Problem Discovery

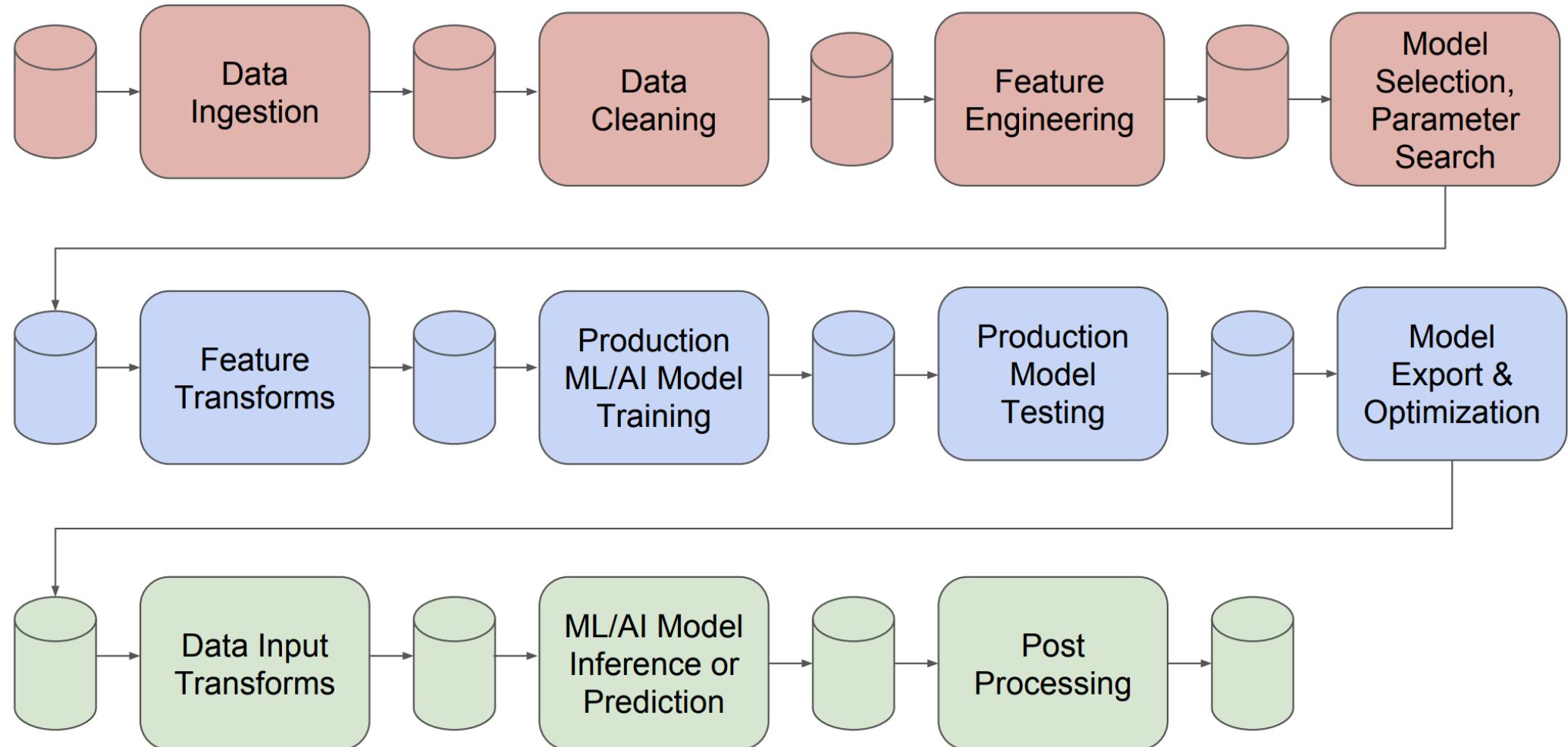


Development



Production /
Actual User Impact

The Data Science Pipeline





KubeCon

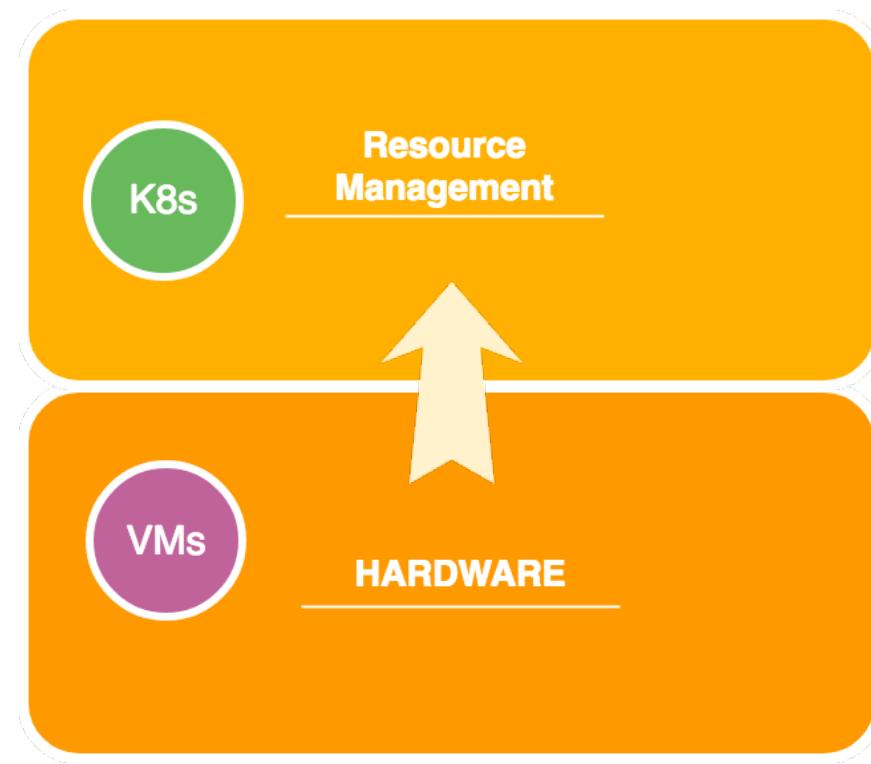


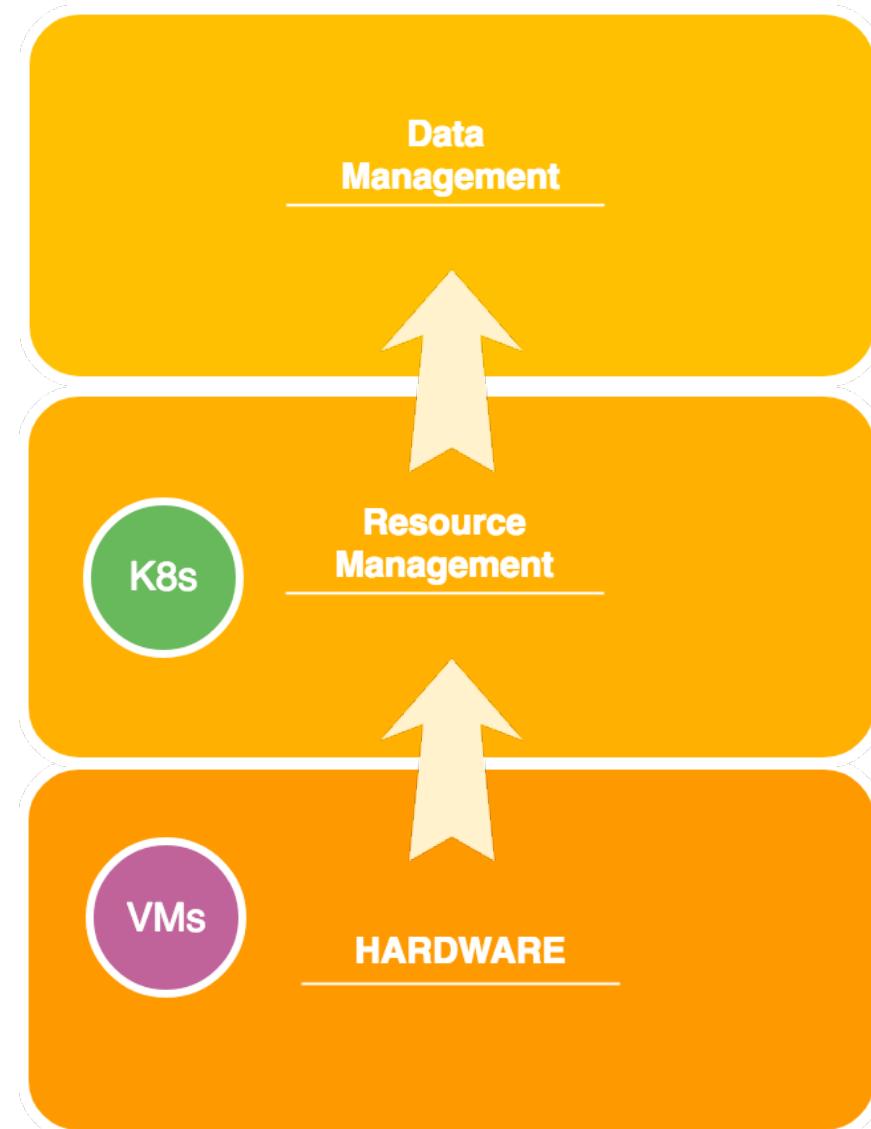
CloudNativeCon

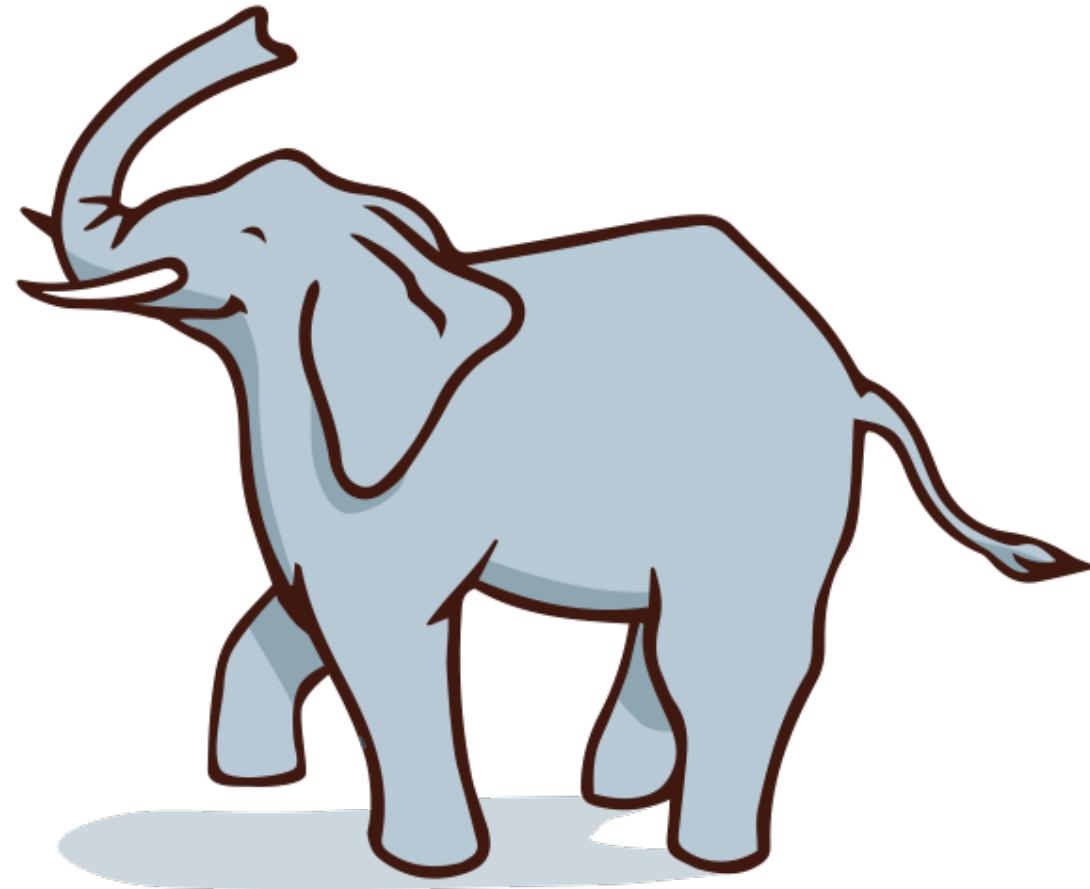
China 2018

VMs

HARDWARE





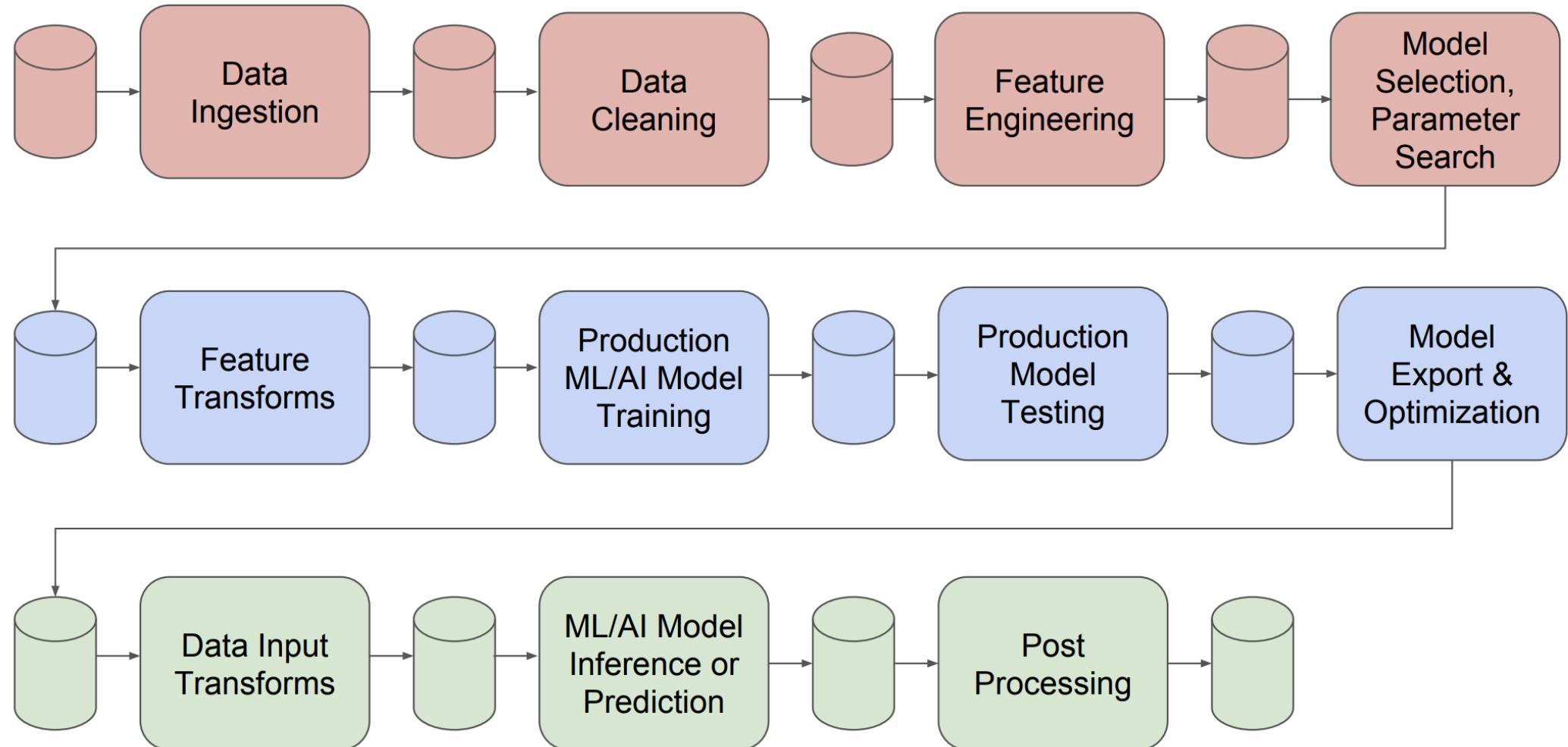


Pachyderm

Pipelines

```
1  {
2      "pipeline": {
3          "name": "wordcount"
4      },
5      "transform": {
6          "image": "wordcount-image",
7          "cmd": ["/binary", "/pfs/data", "/pfs/out"]
8      },
9      "parallelism_spec": {
10         "coefficient": 2
11     },
12     "input": {
13         "atom": {
14             "repo": "data",
15             "glob": "*"
16         }
17     }
18 }
```

The Data Science Pipeline





KubeCon



CloudNativeCon

China 2018

Principles





KubeCon



CloudNativeCon

China 2018

1. Autonomy



I HAVE A VERY PARTICULAR SET OF SKILLS





KubeCon



CloudNativeCon

China 2018

1. Autonomy





KubeCon

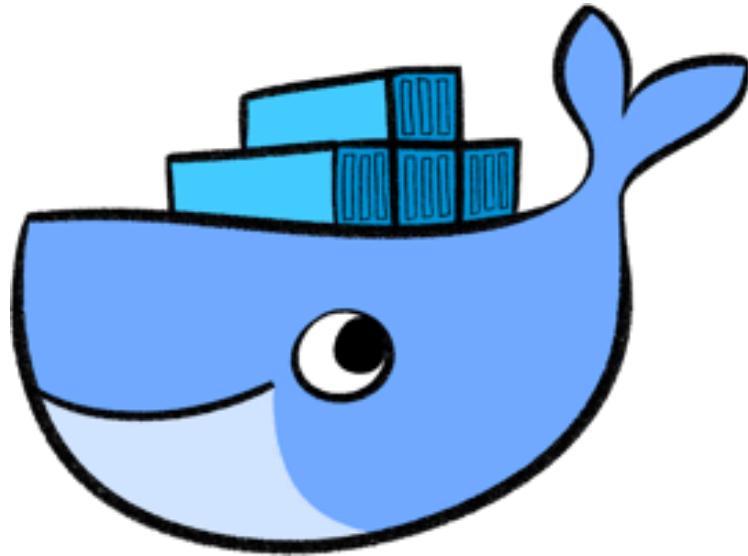


CloudNativeCon

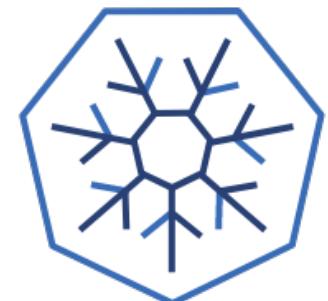
China 2018



Autonomy



 Rocket



cri-o

Containerization

1. Single Operations per Container

Containerization

1. Single Operations per Container
2. Use Parameterize Data Flow
 - Data Inputs
 - Data Outputs

Distributing Workloads

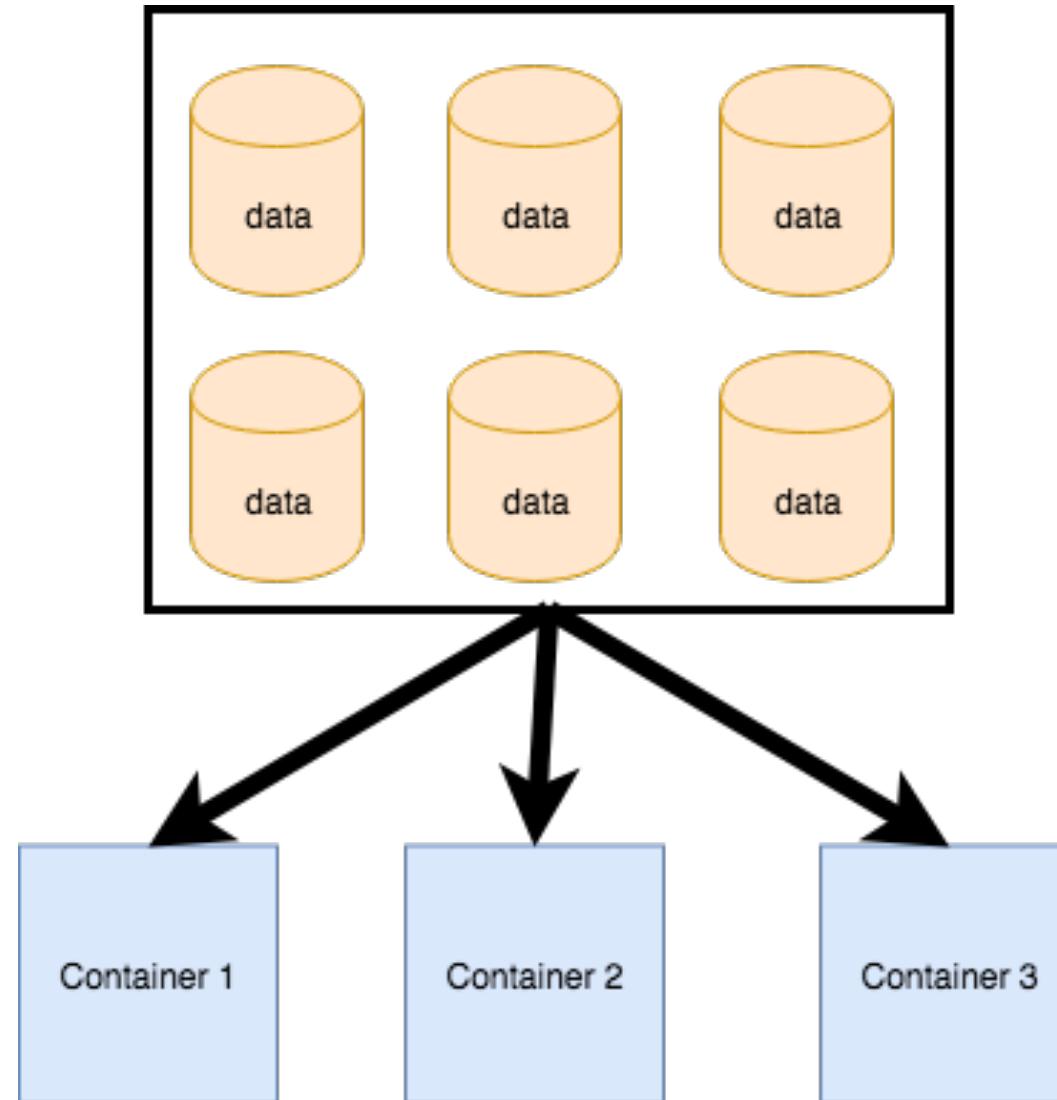


KubeCon



CloudNativeCon

China 2018





KubeCon



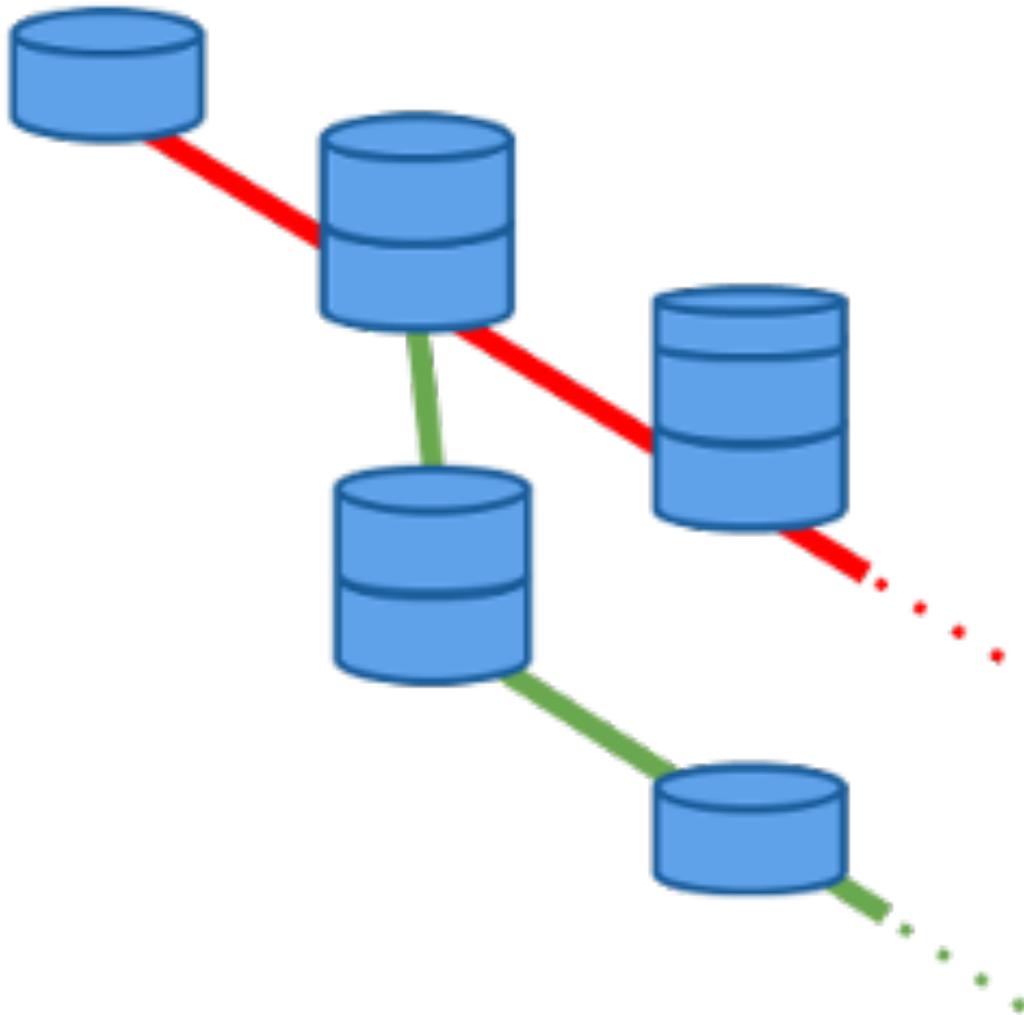
CloudNativeCon

China 2018

2. Reproducibility



Data Versioning



Reproducibility

For Developers

Reproducibility

For Developers

For the Team

Reproducibility



For Developers

For the Team

For Production



KubeCon

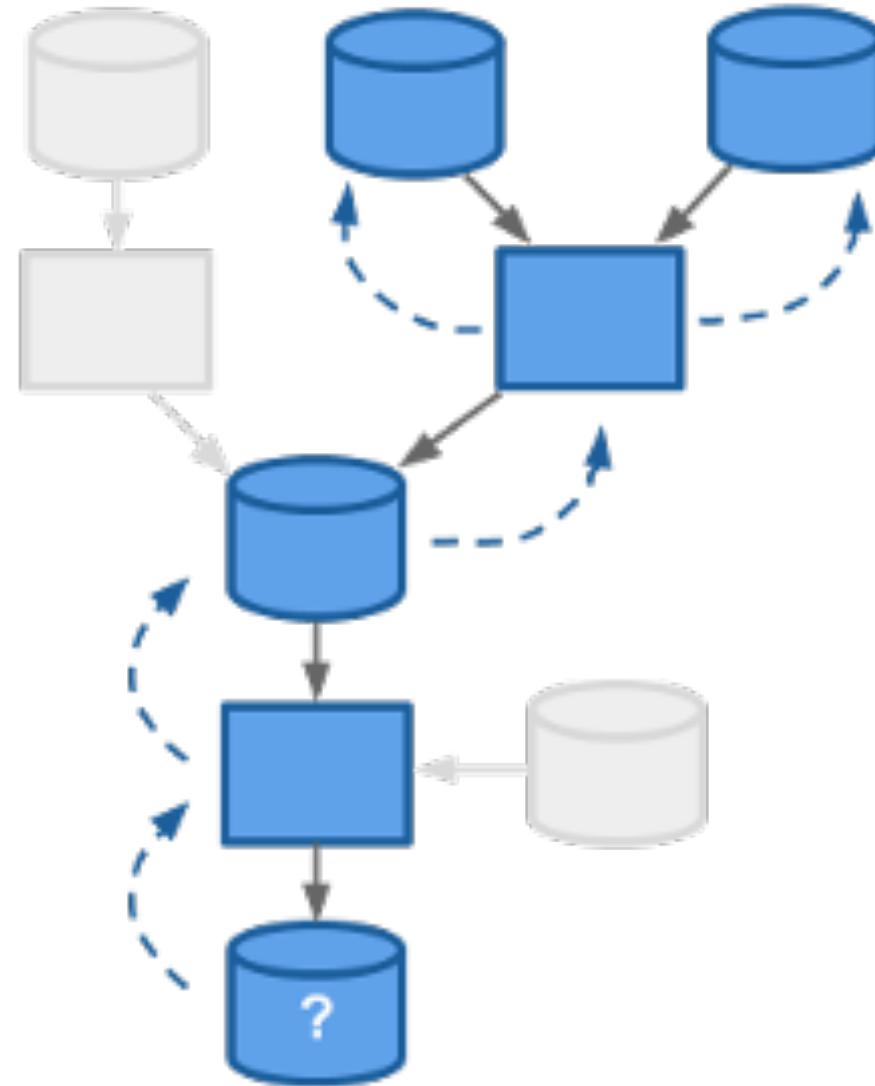


CloudNativeCon

China 2018

3. Data Provenance

Data Provenance



Data Provenance

1. Retry after failure

Data Provenance

1. Retry after failure
2. Rolling Back Models

Data Provenance

1. Retry after failure
2. Rolling Back Models
3. Clarity / Organizational Trust



KubeCon



CloudNativeCon

China 2018

4. Automation (CI/CD)





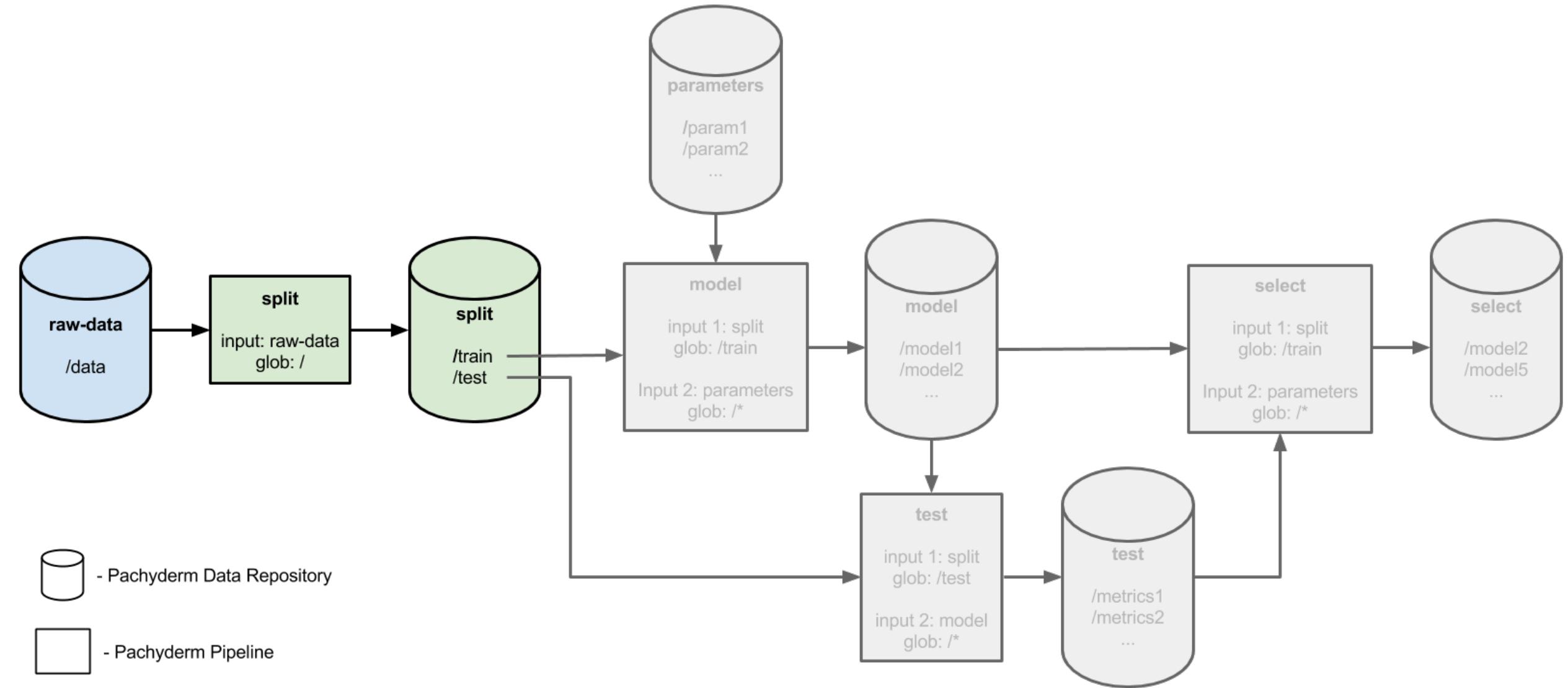
Azure DevOps



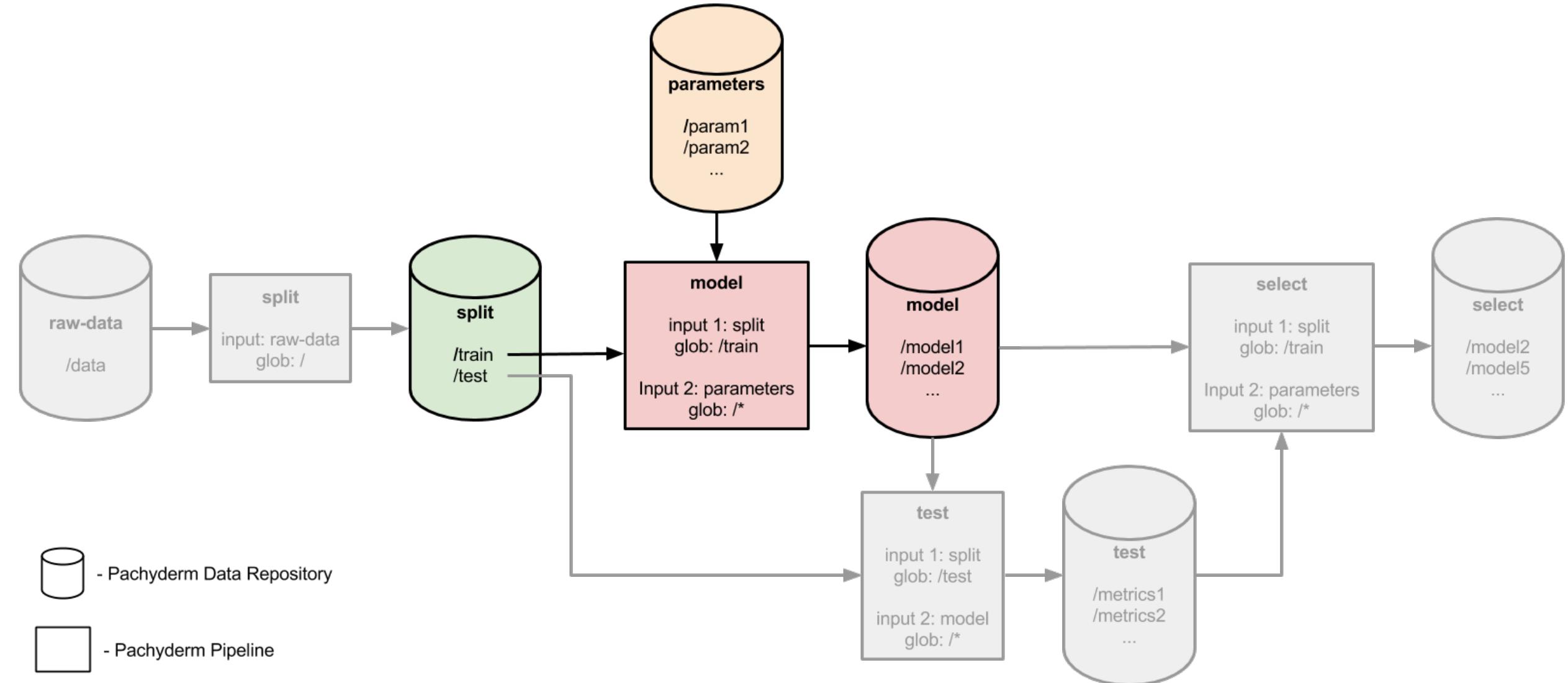
Summary

1. Autonomy
2. Reproducibility
3. Data Provenance
4. Automation

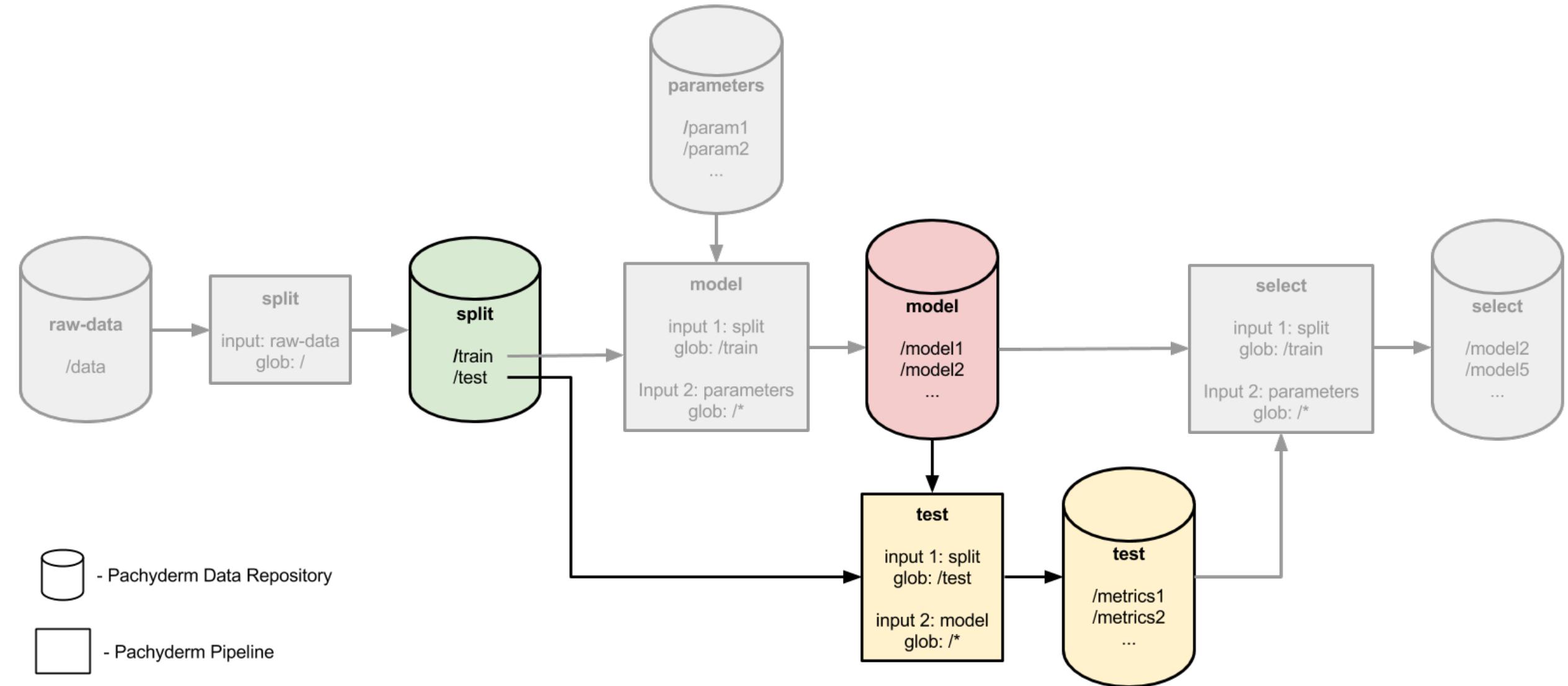
Demo



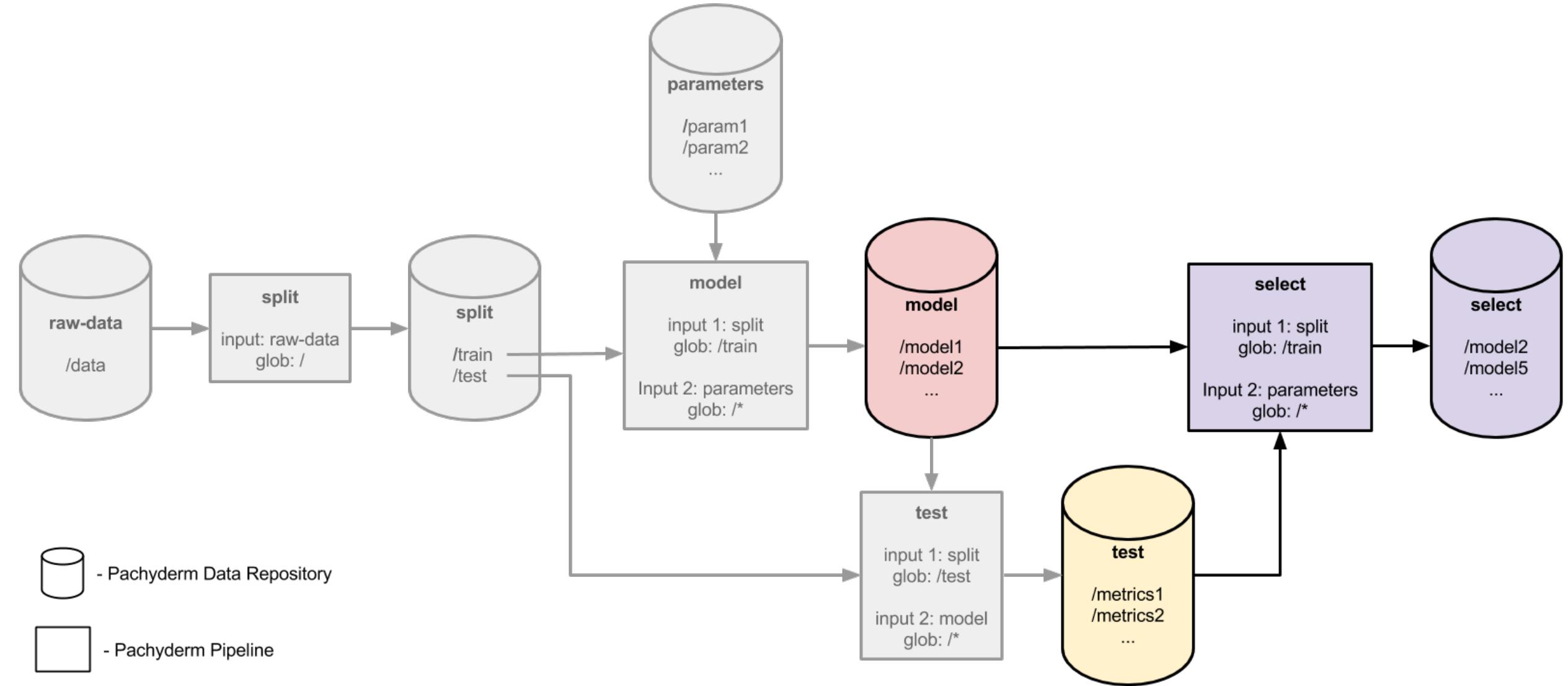
Demo



Demo



Demo



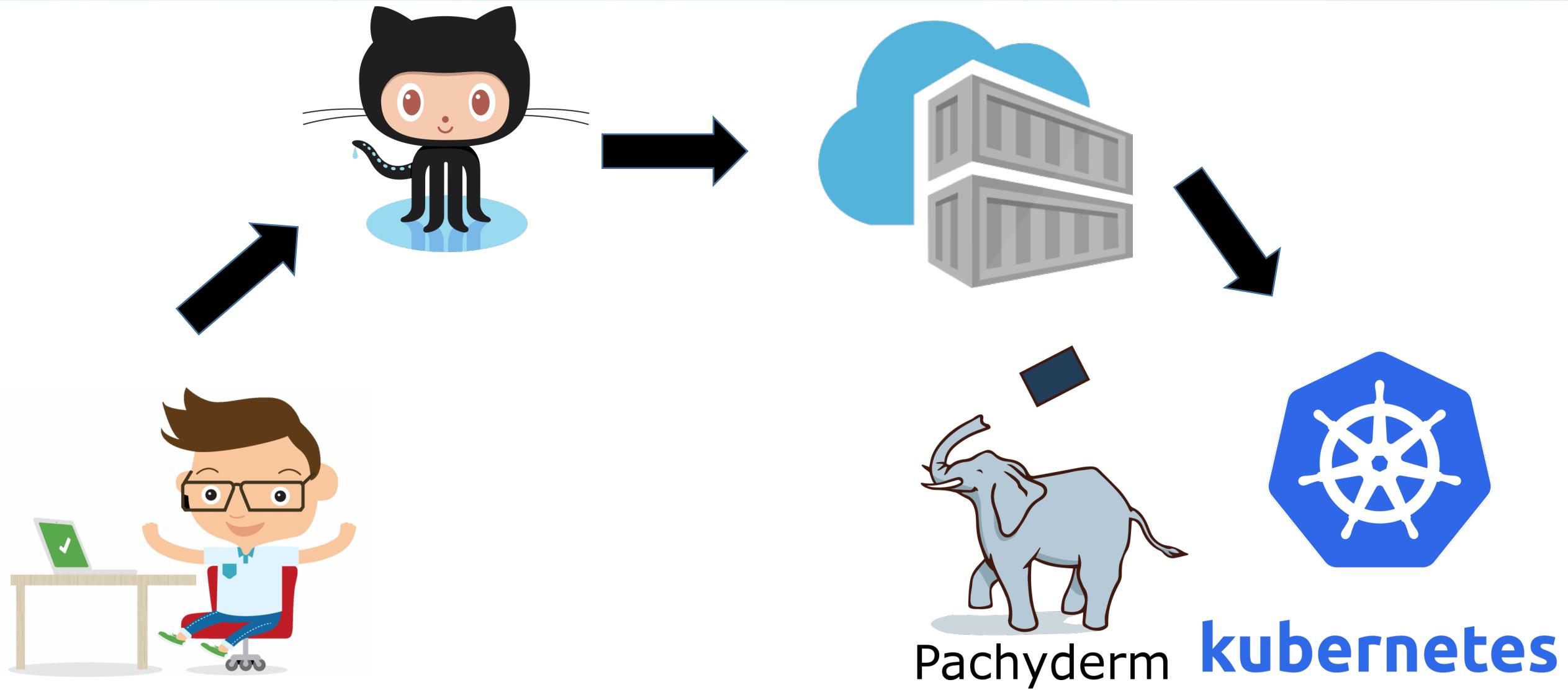


1. bash

```
samkreter@Sams-MacBook-Pro-3:~/go/src/github.com/samkreter/KubeconAsia2018/scripts$  
samkreter@Sams-MacBook-Pro-3:~/go/src/github.com/samkreter/KubeconAsia2018/scripts$  
samkreter@Sams-MacBook-Pro-3:~/go/src/github.com/samkreter/KubeconAsia2018/scripts$  
samkreter@Sams-MacBook-Pro-3:~/go/src/github.com/samkreter/KubeconAsia2018/scripts$  
samkreter@Sams-MacBook-Pro-3:~/go/src/github.com/samkreter/KubeconAsia2018/scripts$  
samkreter@Sams-MacBook-Pro-3:~/go/src/github.com/samkreter/KubeconAsia2018/scripts$  
samkreter@Sams-MacBook-Pro-3:~/go/src/github.com/samkreter/KubeconAsia2018/scripts$ ls  
data/      pipelines/ setup.sh*  update.sh  
samkreter@Sams-MacBook-Pro-3:~/go/src/github.com/samkreter/KubeconAsia2018/scripts$ █
```

}

Demo



Small model changes by samkreter



GitHub, Inc. [US] | https://github.com/samkreter/KubeconAsia2018/pull/18



Apps

Todo

When Your Bored

customer relation...

SubtlePatterns Bo...

HSK 4 Vocabulary...

CSS-Only Respon...

chinese topics

Hacking a Google ...



Small model changes #18

Edit

Open

samkreter wants to merge 1 commit into master from modelchanges8

Conversation 0

Commits 1

Checks 0

Files changed 1

+1 -1



samkreter commented 2 minutes ago

Owner



...

No description provided.

small model changes

58cd57a

Add more commits by pushing to the modelchanges8 branch on samkreter/KubeconAsia2018.



This branch has no conflicts with the base branch

Merging can be performed automatically.

Merge pull request

You can also open this in GitHub Desktop or view command line instructions.



Write

Preview

AA

B

i

“

<>

↶

≡

≡

≡

@

★

↶

Leave a comment

Unsubscribe

You're receiving notifications because you authored the thread.

Reviewers



No reviews

Assignees



No one—assign yourself

Labels



None yet

Projects



None yet

Milestone

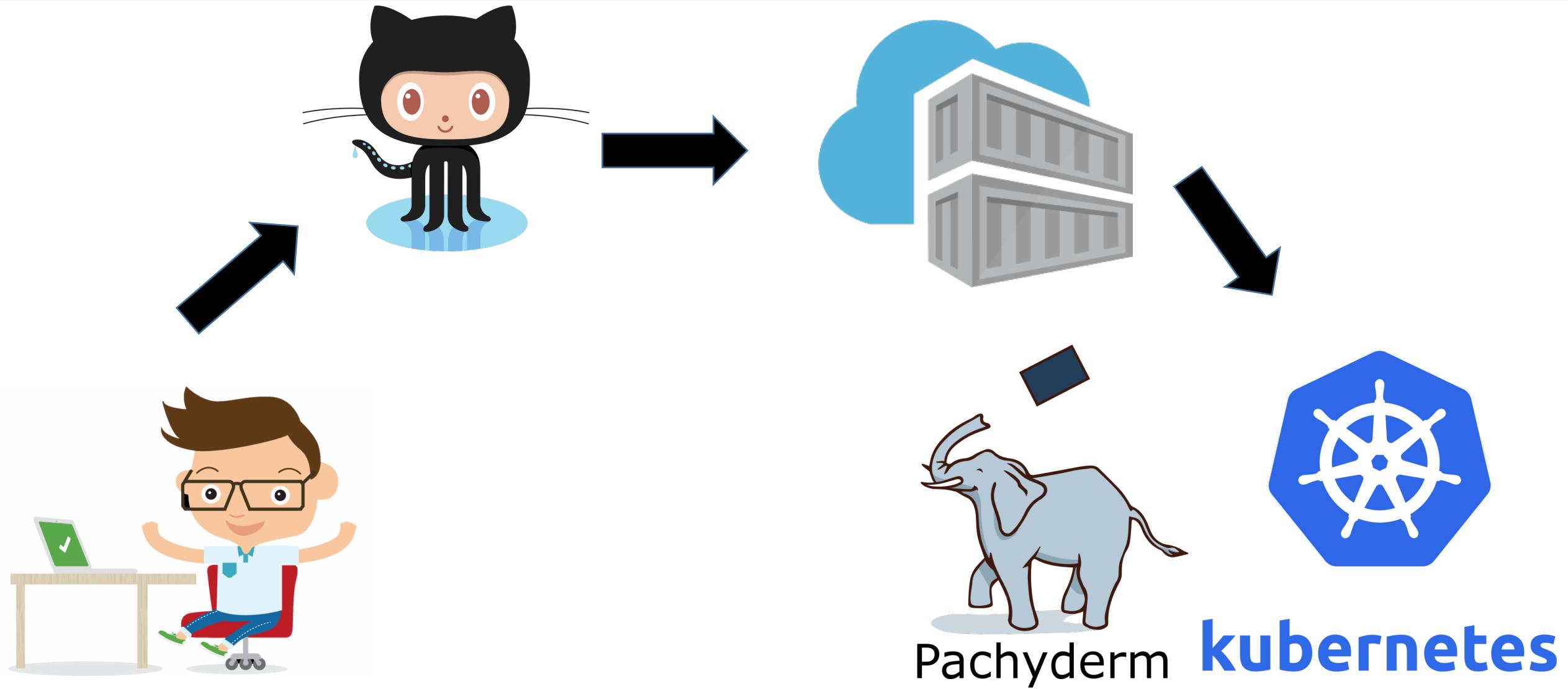


No milestone

Notifications

Unsubscribe

Demo



Contact Me



Twitter: @samkreter

Github: samkreter

Medium: samkreter

Sources / Resources

Pachyderm: <http://www.pachyderm.io/>

Data Science Bill of Rights: <http://www.pachyderm.io/dsbor.html>

Azure Container Registry Build: <https://docs.microsoft.com/en-us/azure/container-registry/container-registry-tasks-overview>

Azure Kubernetes Service: <https://azure.microsoft.com/en-us/services/kubernetes-service/>

Pipeline Images:

<https://github.com/pachyderm/pachyderm/tree/master/doc/examples/ml/hyperparameter>



KubeCon



CloudNativeCon

China 2018

谢谢！





KubeCon



CloudNativeCon

China 2018

