devfest 2021

Managing the **ML Lifecycle** without a headache

Hampus Londögård
Team Lead Future Technologies, AFRY
lundez
@hlondogard

Official Partner

AFRY ÅF PÖYRY

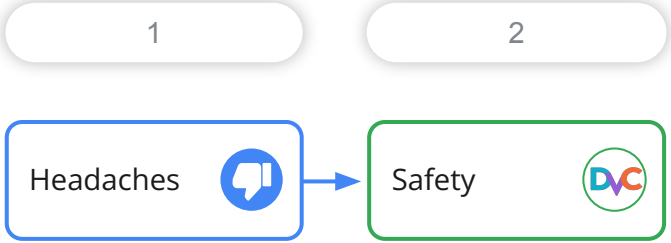
Google Developer Groups
West Sweden

Content

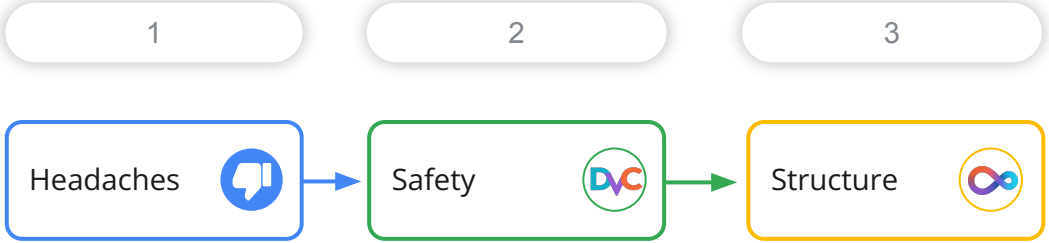
1

Headaches

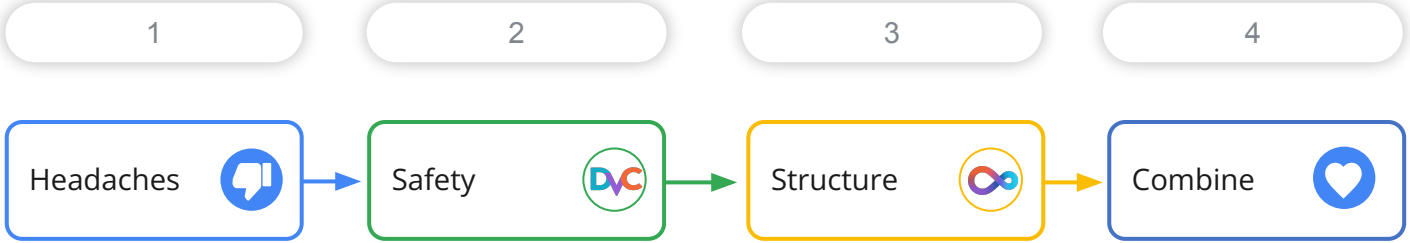
Content



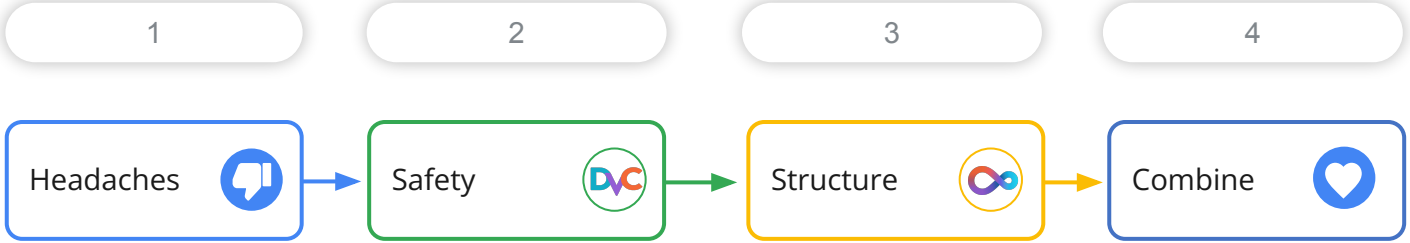
Content



Content

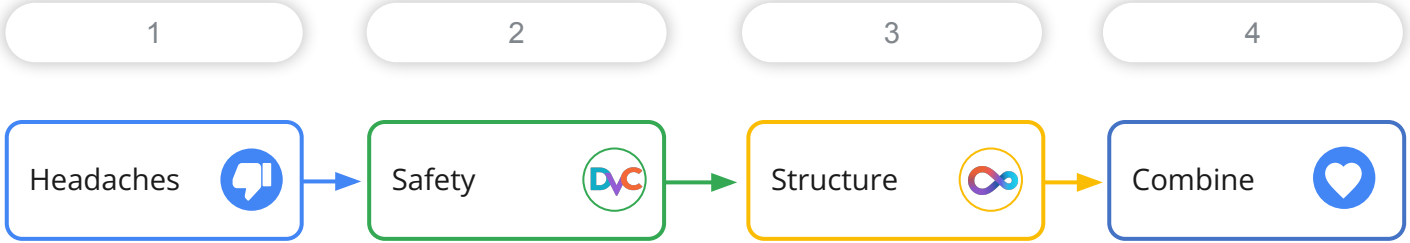


Content



Open Source

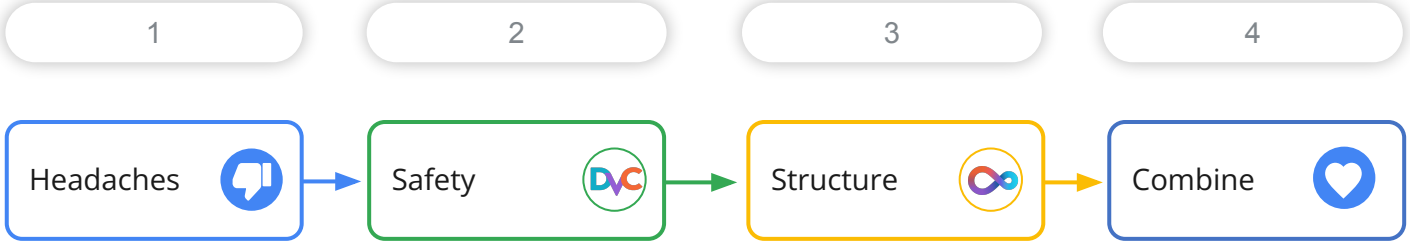
Content



Open Source

Agnostic

Content



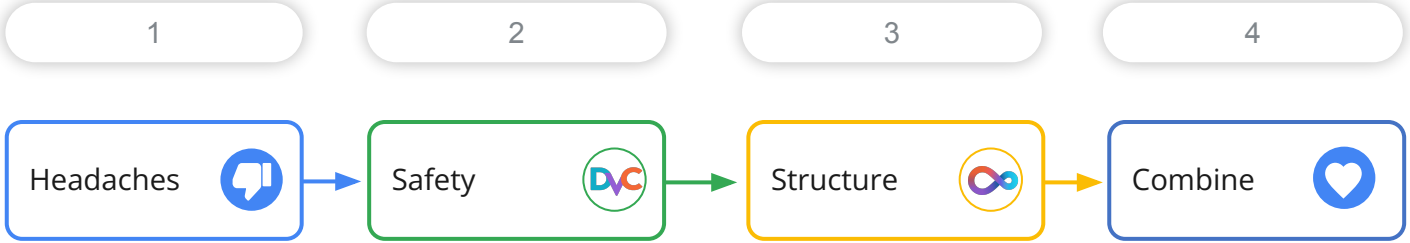
Open Source

Agnostic

Google Cloud

aws

Content



Open Source

Agnostic

+ ssh, local, HDFS, & more!

Google Cloud

aws

Headaches

- Experiments based on folders?

- Experiments based on folders?

📁

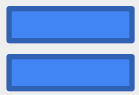
- Experiments in a notepad? 📝

- Experiments based on folders? 🗂️
- Experiments in a notepad? 📝
- Weights on G-Drive?

- Experiments based on folders? 🗂️
- Experiments in a notepad? 📝
- Weights on G-Drive?
- Sharing code through email? 📧

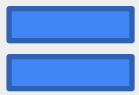
- Experiments based on folders? 🗂️
- Experiments in a notepad? 📝
- Weights on G-Drive?
- Sharing code through email? 📧
- Difficulties reproducing the best model? ⟲

- Experiments based on folders? 🗂️
- Experiments in a notepad? 📝
- Weights on G-Drive?
- Sharing code through email? 📧
- Difficulties reproducing the best model? 🔄



Chaos when something breaks

- Experiments based on folders? 🗂️
- Experiments in a notepad? 📝
- Weights on G-Drive?
- Sharing code through email? 📧
- Difficulties reproducing the best model? ⟲



Chaos ***when*** something breaks

Can we fix it?

Can we fix it?

Yes we can

Safety

Version Control

Version Control

Version Control

Data
Version
Control

git

commit 8d7aa3d

commit 8d7aa3d

data

code

model

Jan 2021	——	V0.0.1	→	V0.0.1
Mar 2021	——	V0.0.2	→	V0.0.2
Apr 2021	——	V0.0.3	→	V0.0.3

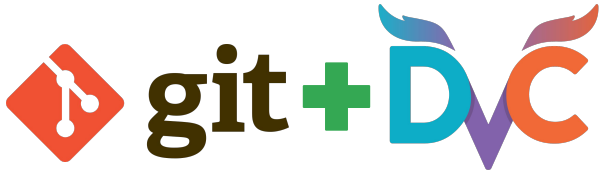
commit 8d7aa3d

data

code

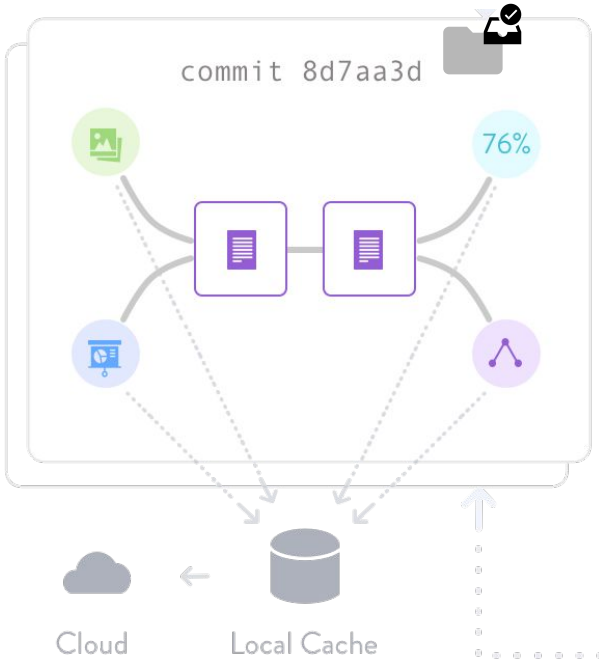
model

Jan 2021		V0.0.1 →	V0.0.1?
Mar 2021		V0.0.2 →	V0.0.2?
Apr 2021		V0.0.3 →	V0.0.3?



commit 8d7aa3d

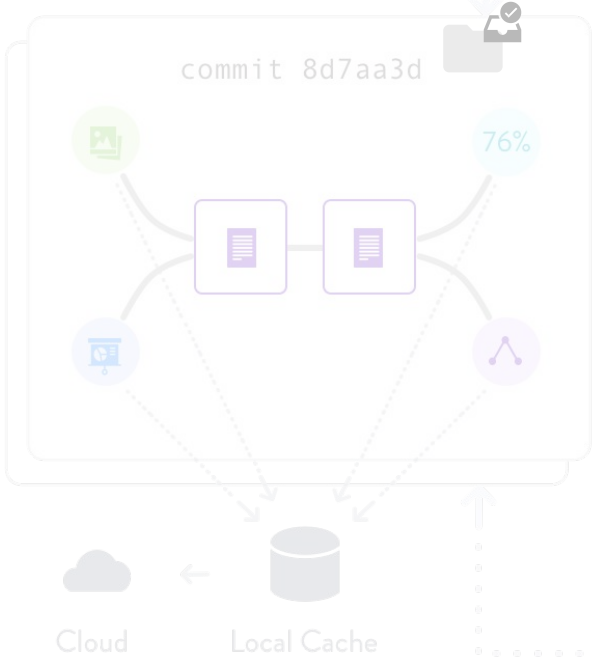
git + DVC

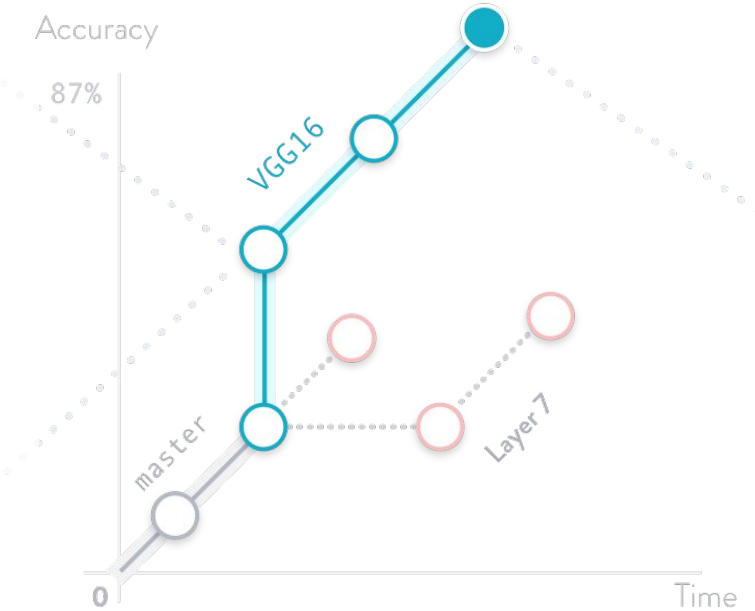


dvc.org

No more complicated folder-structure
🥳

git + DVC





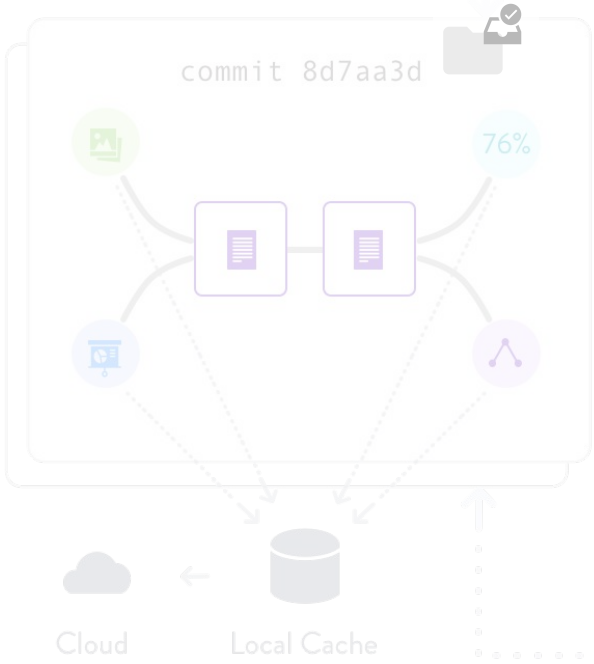
Rollback

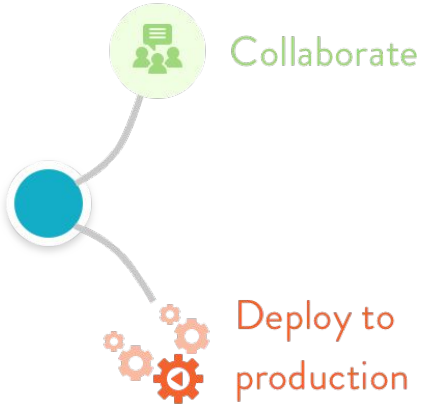
dvc.org

No more complicated folder-structure 🥳

Rollback hours, days or even years! ⟲

git + DVC





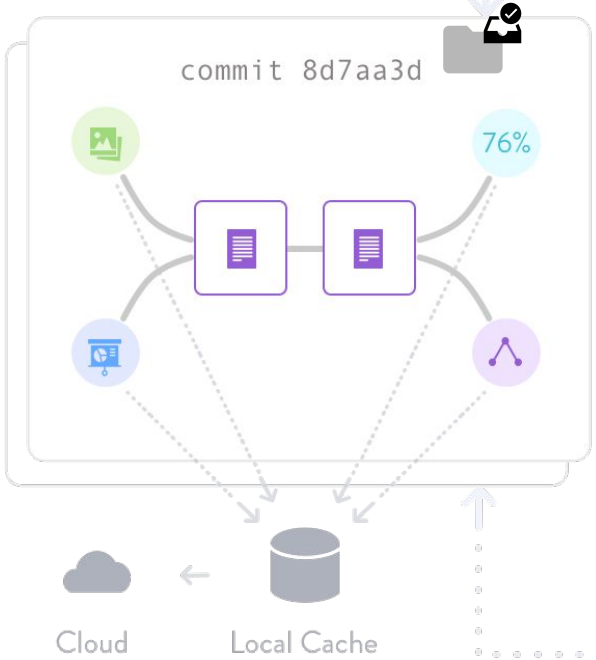
dvc.org

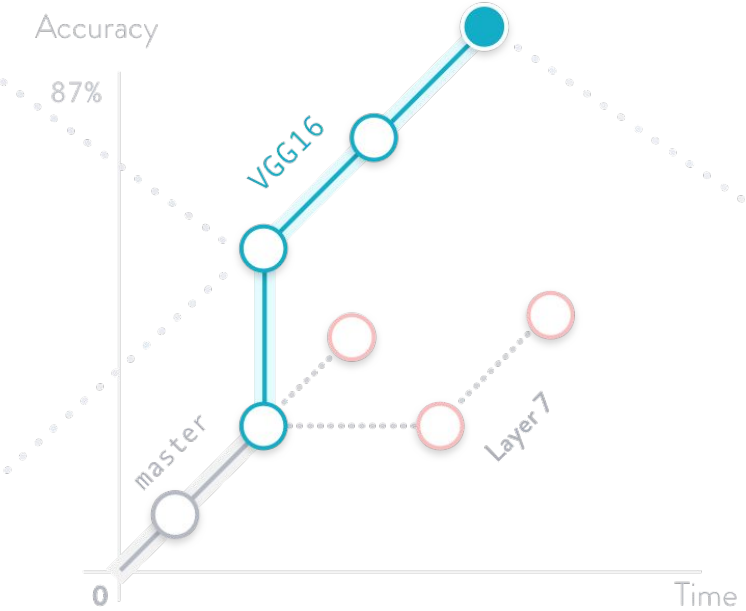
No more complicated folder-structure 🥳

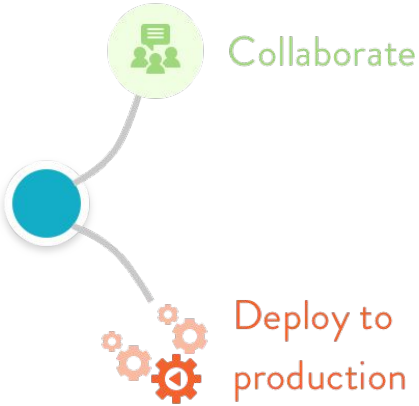
Rollback hours, days or even years! ⏲

Simplify selection & creation of release ☕

git + DVC







Rollback

dvc.org

No more complicated folder-structure 🥳

Rollback hours, days or even years! ⏲

Simplify selection & creation of release ☕

Structure

CI/CD

CI/CD

Continuous **I**ntegration

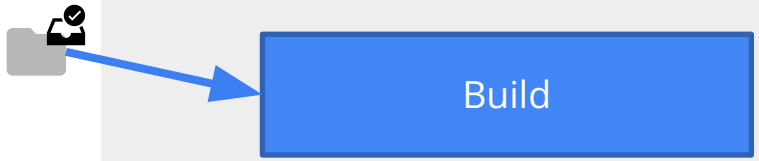
CI/CD

Continuous **I**ntegration

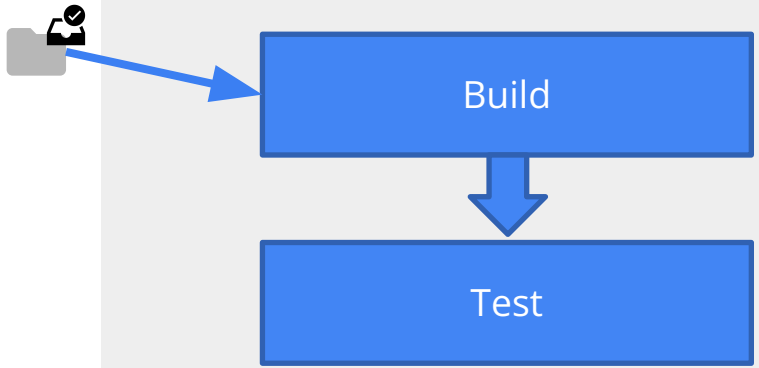
Continuous **D**elivery

CI/CD

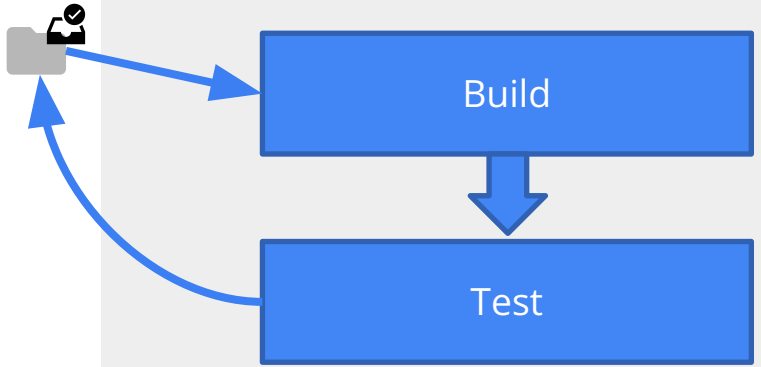
CI/CD



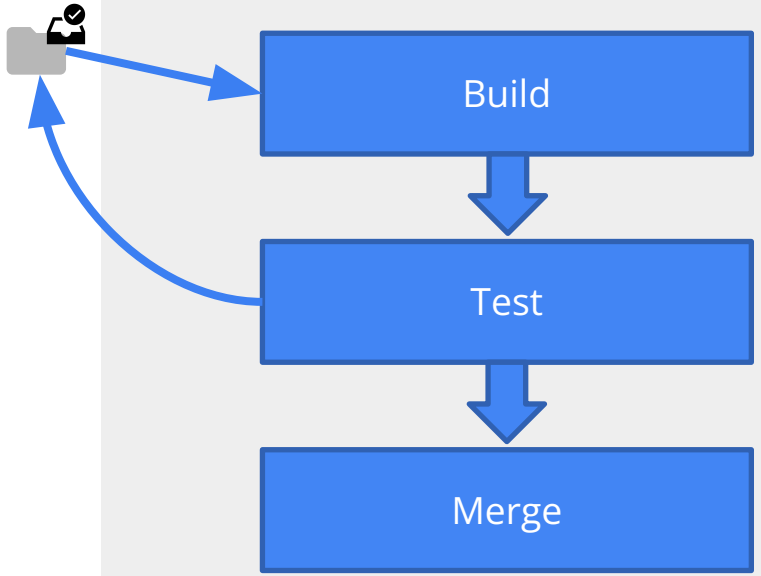
CI/CD



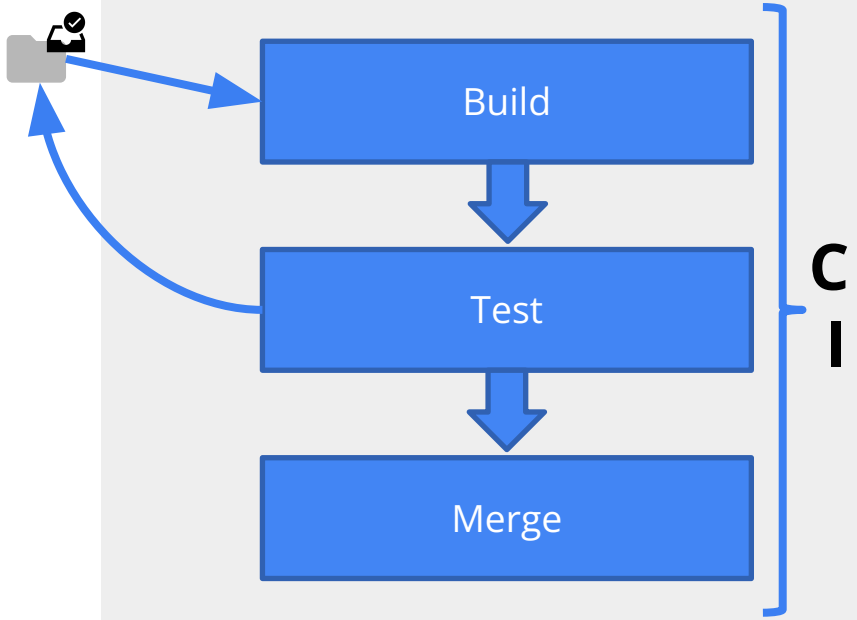
CI/CD



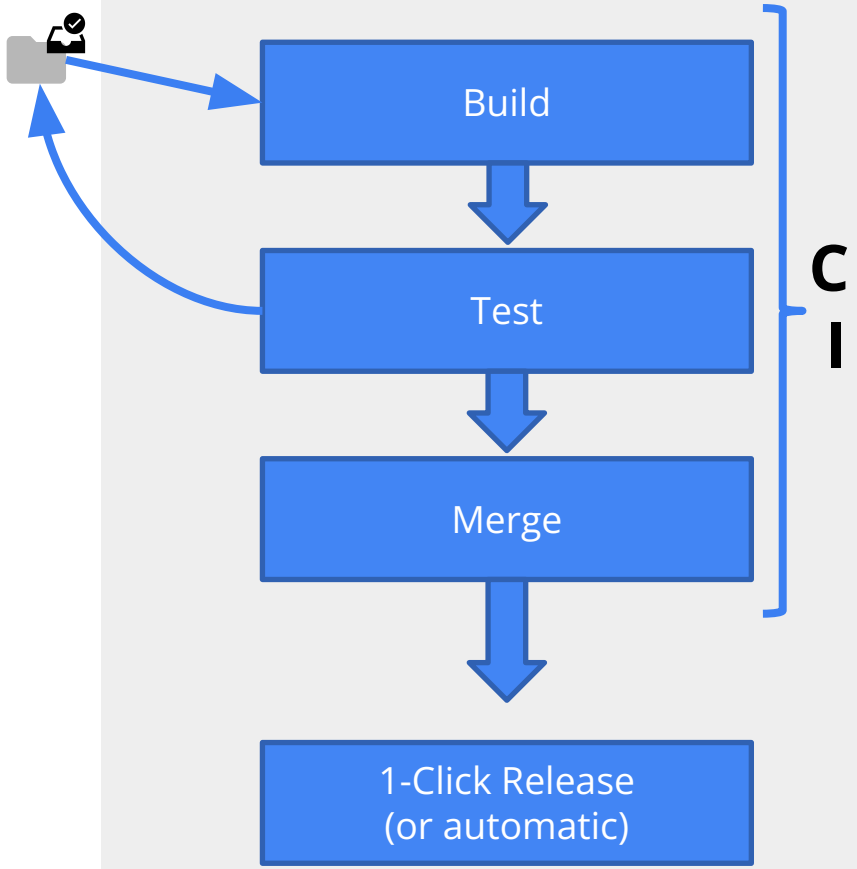
CI/CD



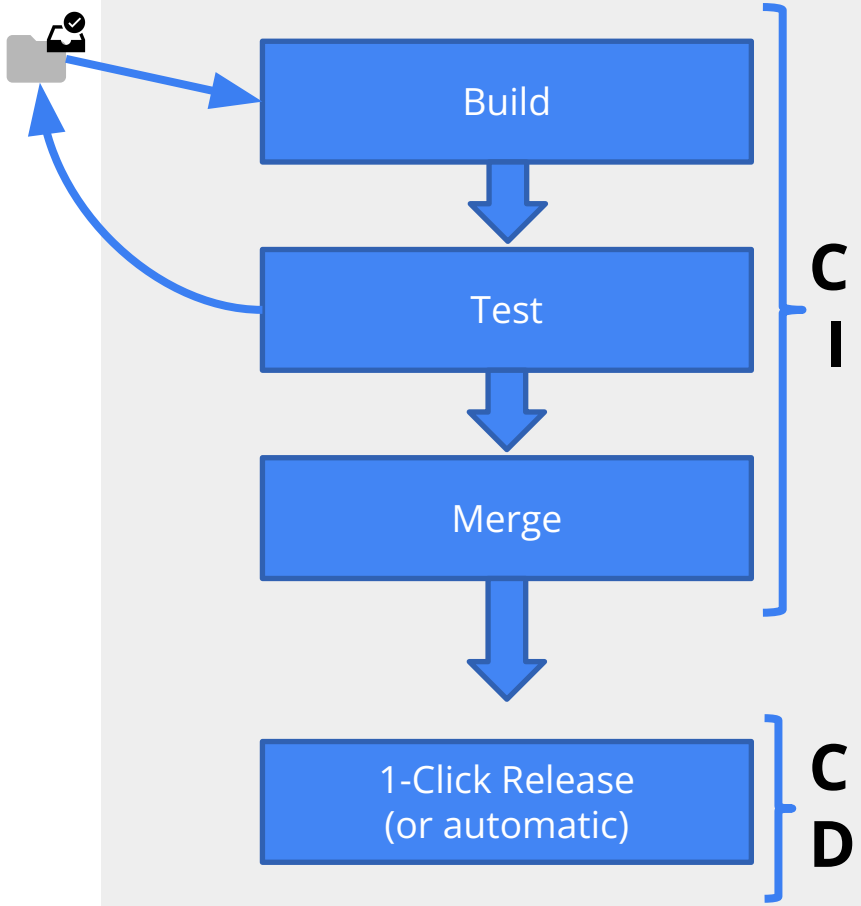
CI/CD



CI/CD

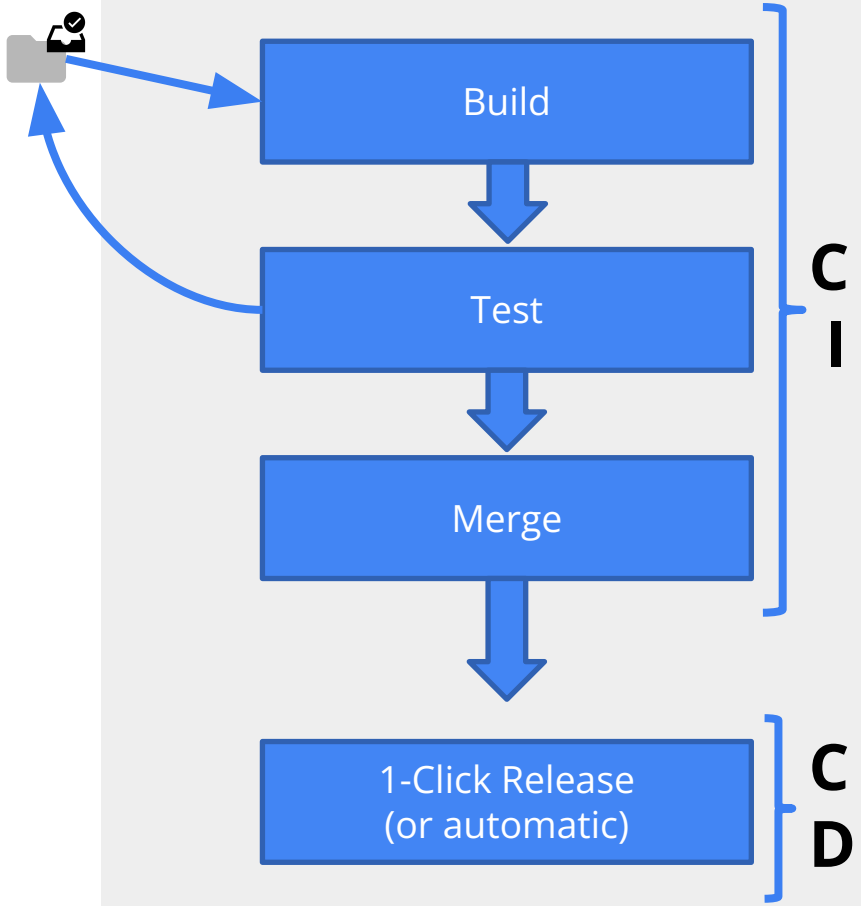


CI/CD



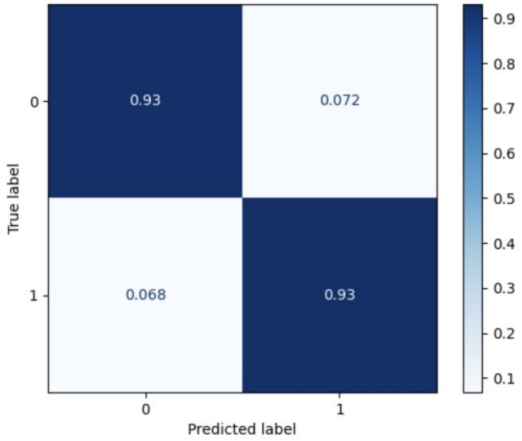
CI/CD

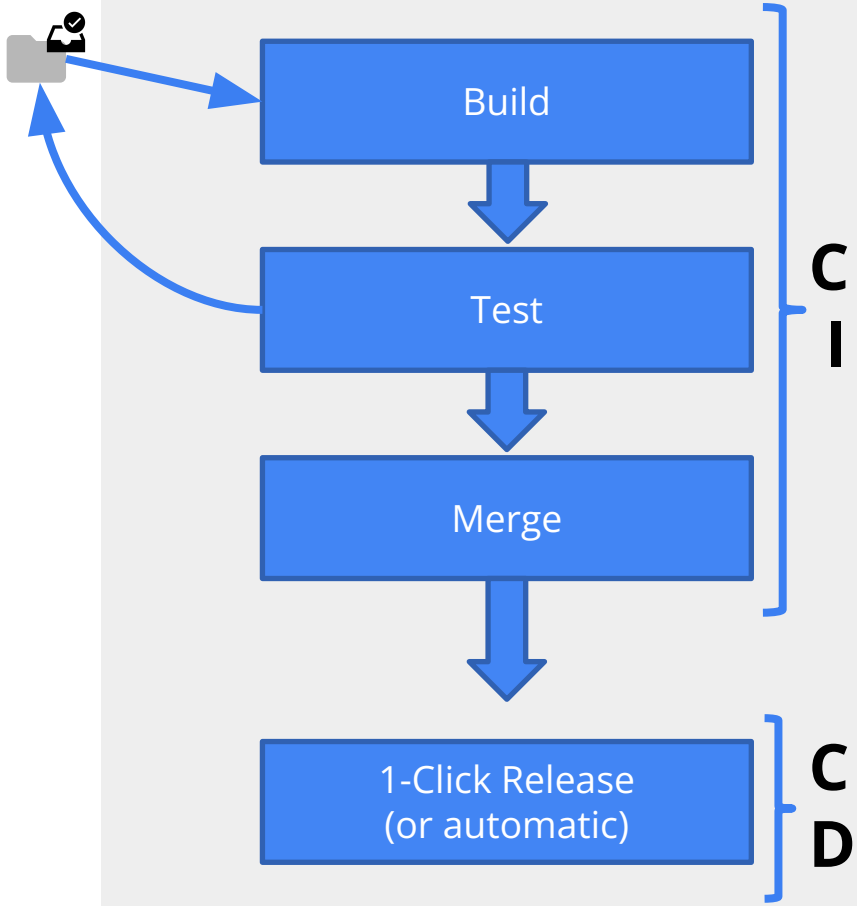
✅ predict_same.py
✅ predict_human.py
✅ predict_dog.py
✅ predict_nothing.py
...



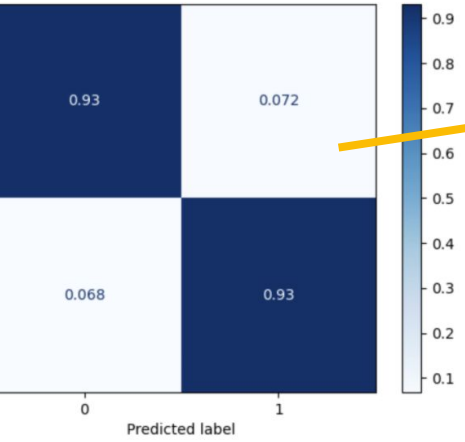
CI/CD

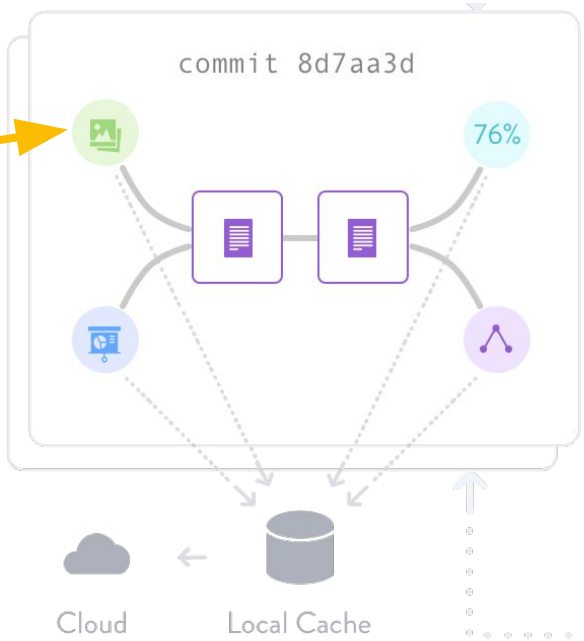
Accuracy: 0.93

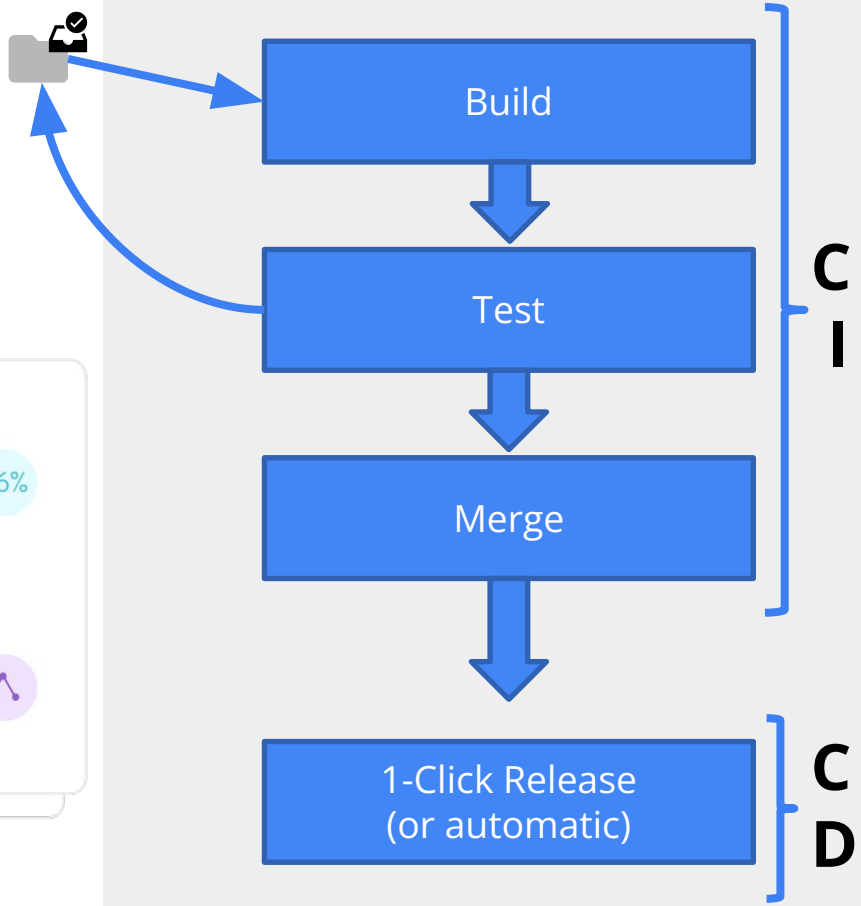




CI/CD


0.93
0.072
0.068
0.93
0
1
Predicted label
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1


commit 8d7aa3d
76%
Cloud
Local Cache


Build
Test
Merge
1-Click Release
(or automatic)
C
I
C
D

CI/CD

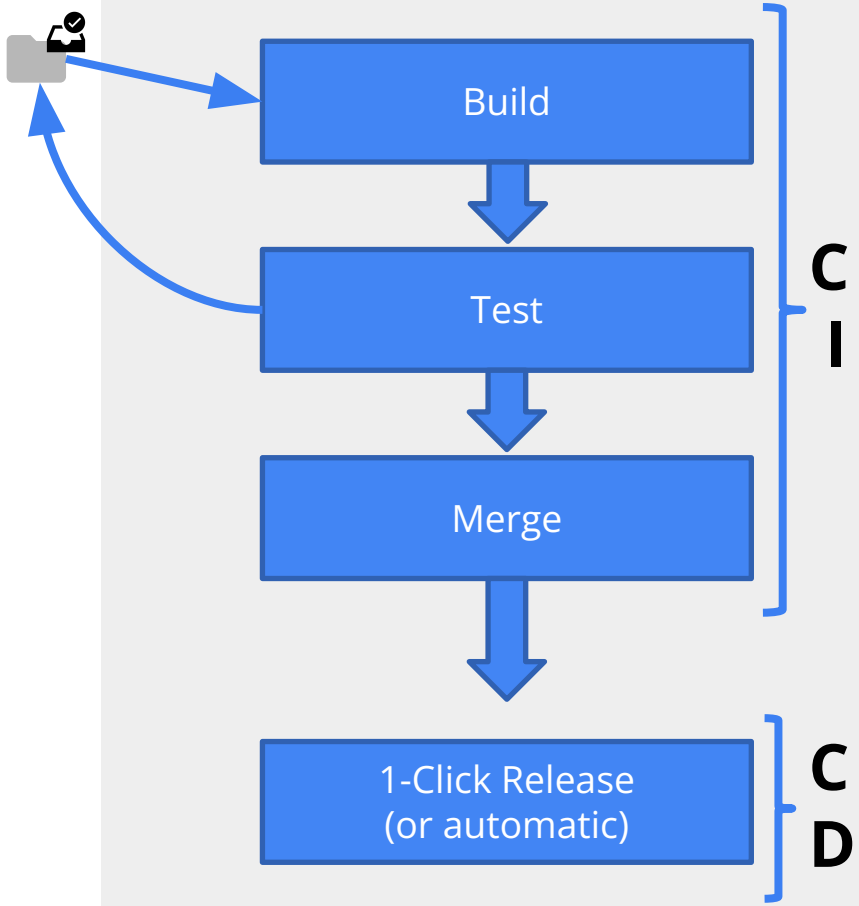
Workspace vs. Master

Training metrics

Path	Param	Old	New
params.yaml	style_weight	100.0	50.0

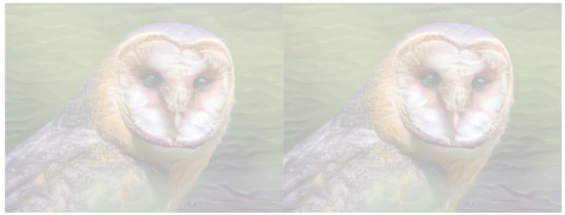
GPU info

Metric	Value
Iterations	2
Run time (s)	779.615
Final loss	5.16476e+10



CI/CD

Workspace vs. Master

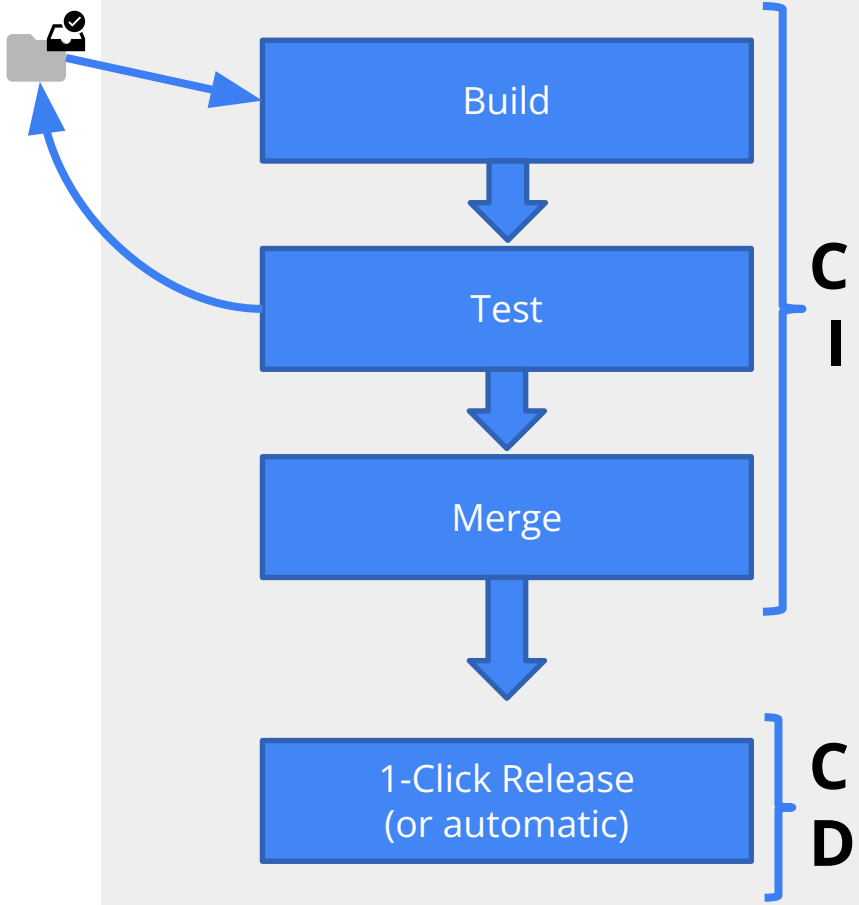


Training metrics

Path	Param	Old	New
params.yaml	style_weight	100.0	50.0

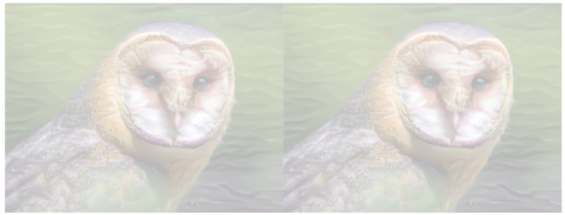
GPU info

Metric	Value
Iterations	2
Run time (s)	779.615
Final loss	5.16476e+10



CI/CD

Workspace vs. Master

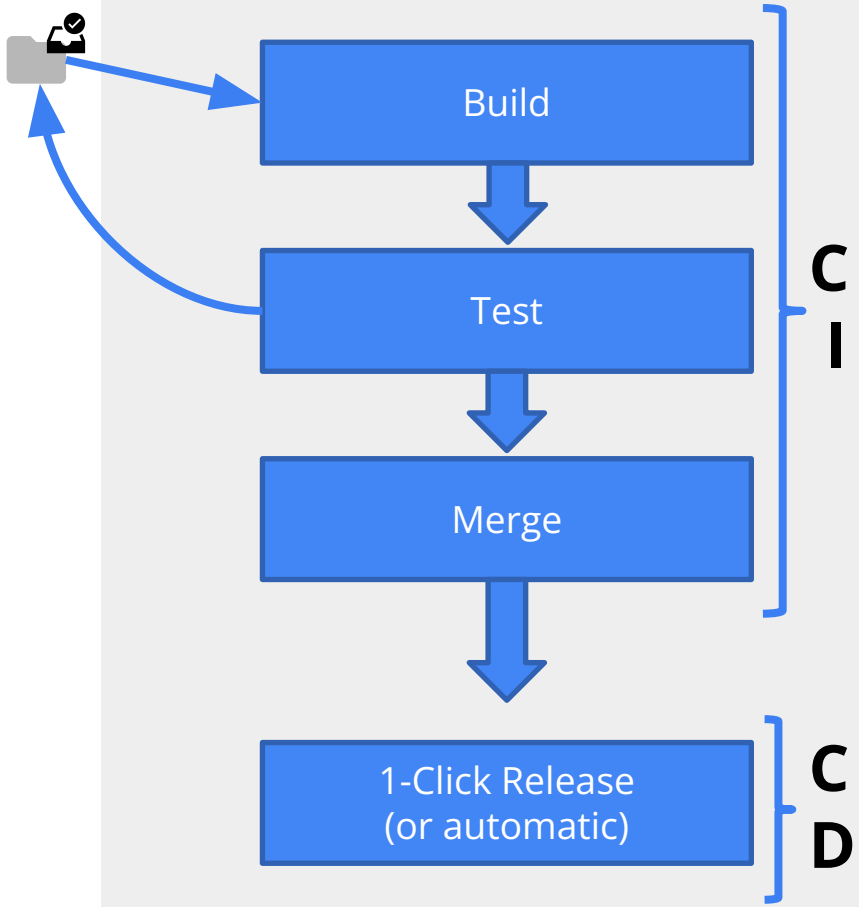


Training metrics

Path	Param	Old	New
params.yaml	style_weight	100.0	50.0

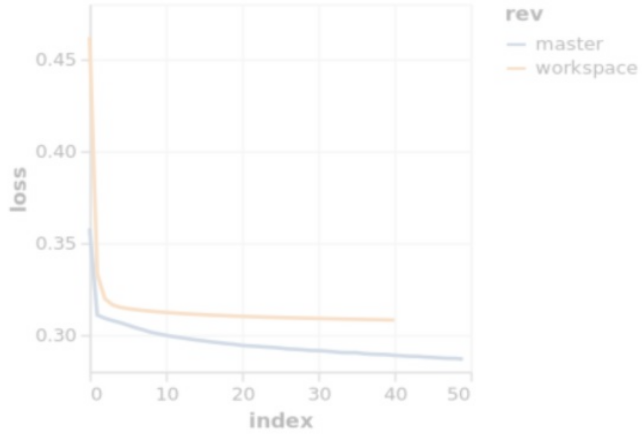
GPU info

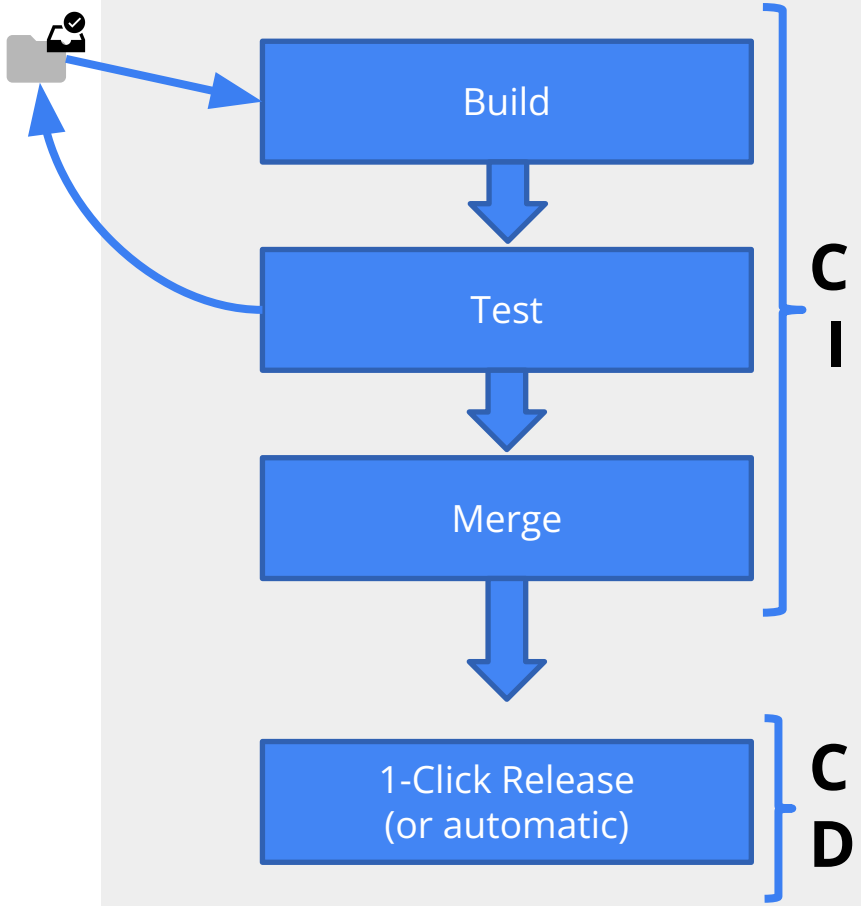
Metric	Value
Iterations	2
Run time (s)	779.615
Final loss	5.16476e+10



CI/CD

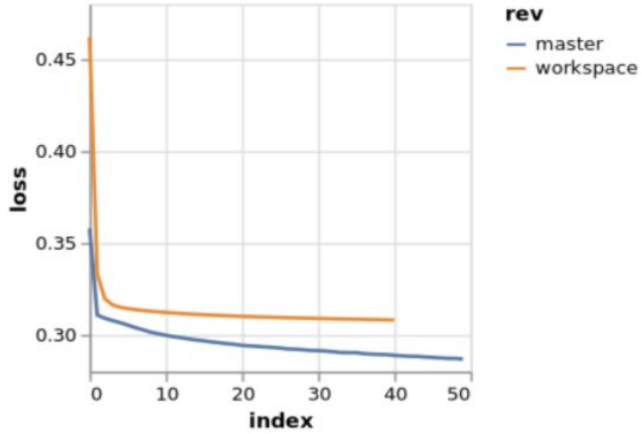
Path	Metric	Value	Change
metrics.json	accuracy	0.8694	-0.00296
metrics.json	precision	0.86793	0.00375
metrics.json	recall	0.87352	-0.01216

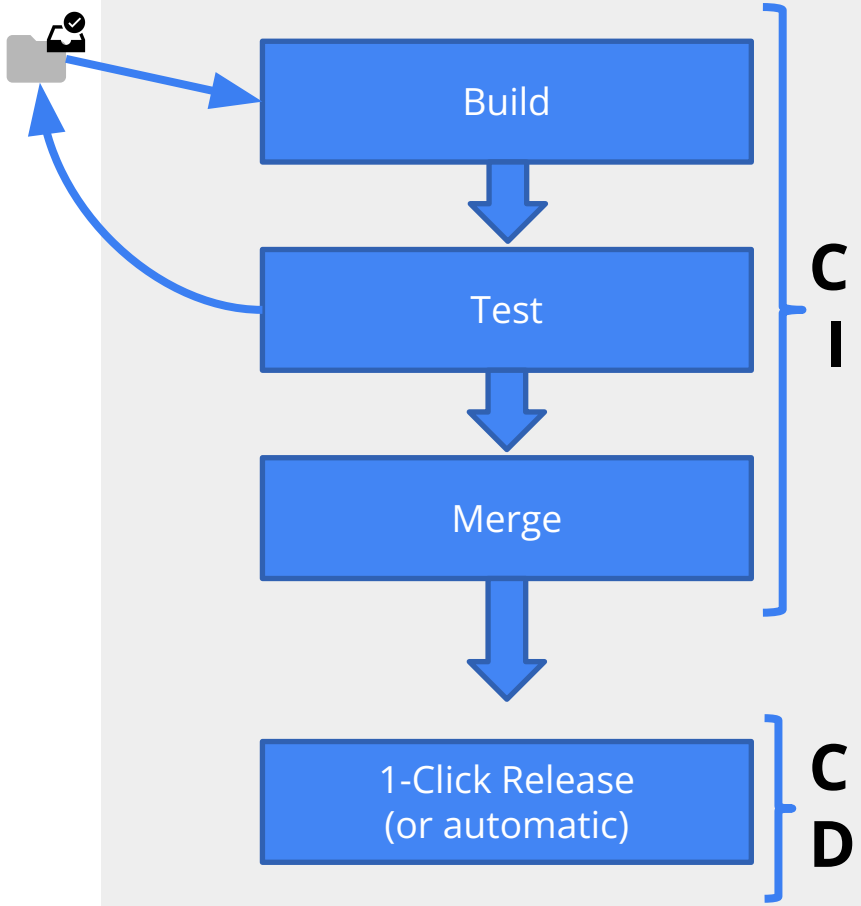




CI/CD

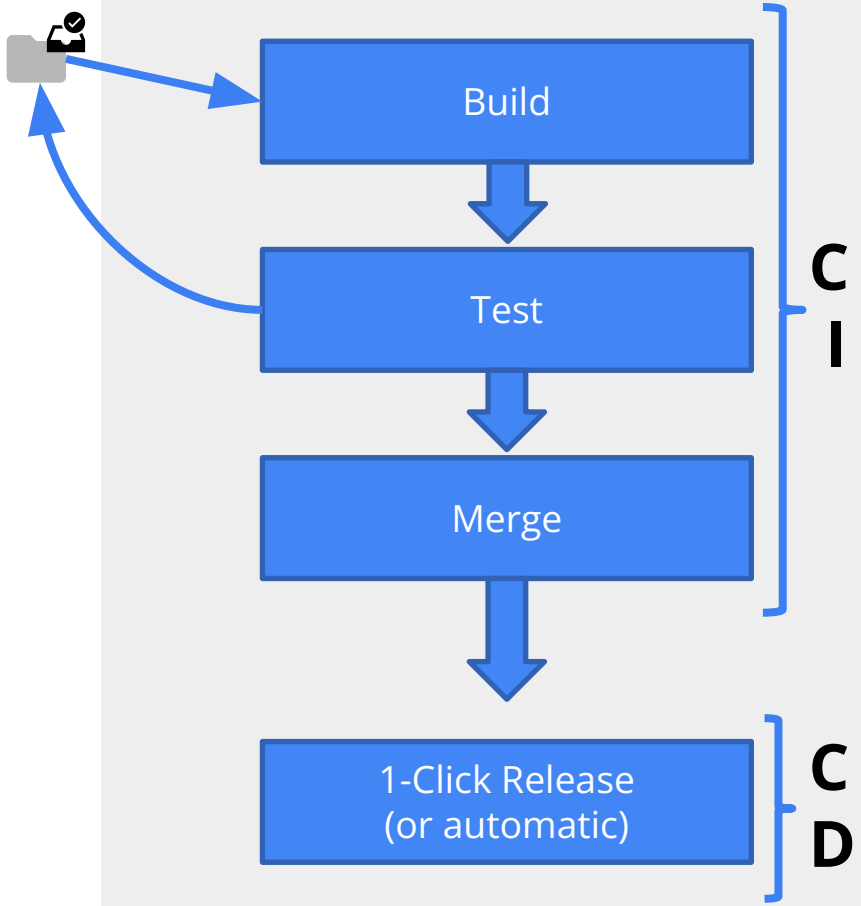
Path	Metric	Value	Change
metrics.json	accuracy	0.8694	-0.00296
metrics.json	precision	0.86793	0.00375
metrics.json	recall	0.87352	-0.01216





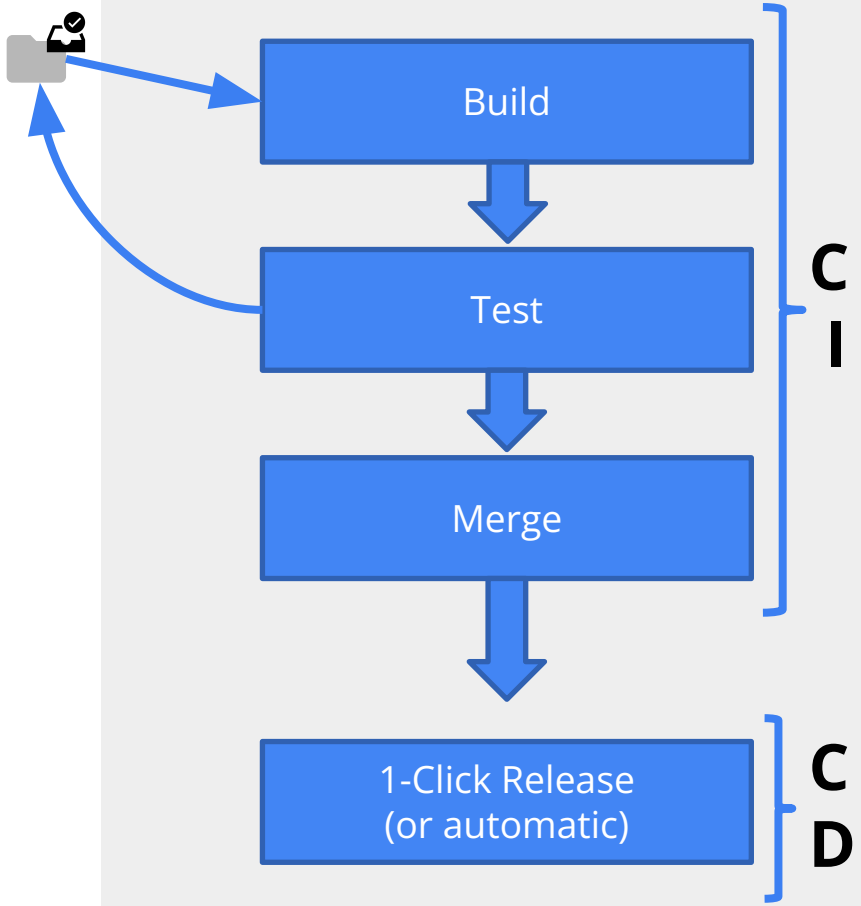
CI/CD

- Nightly Jobs



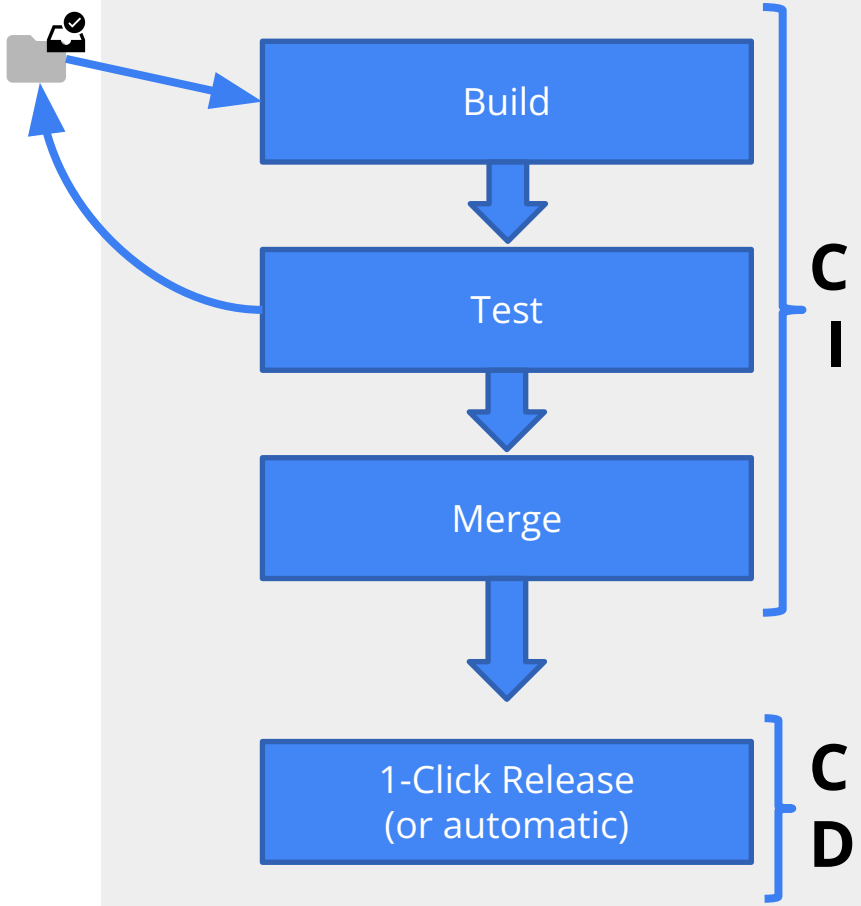
CI/CD

- Nightly Jobs
- New Data



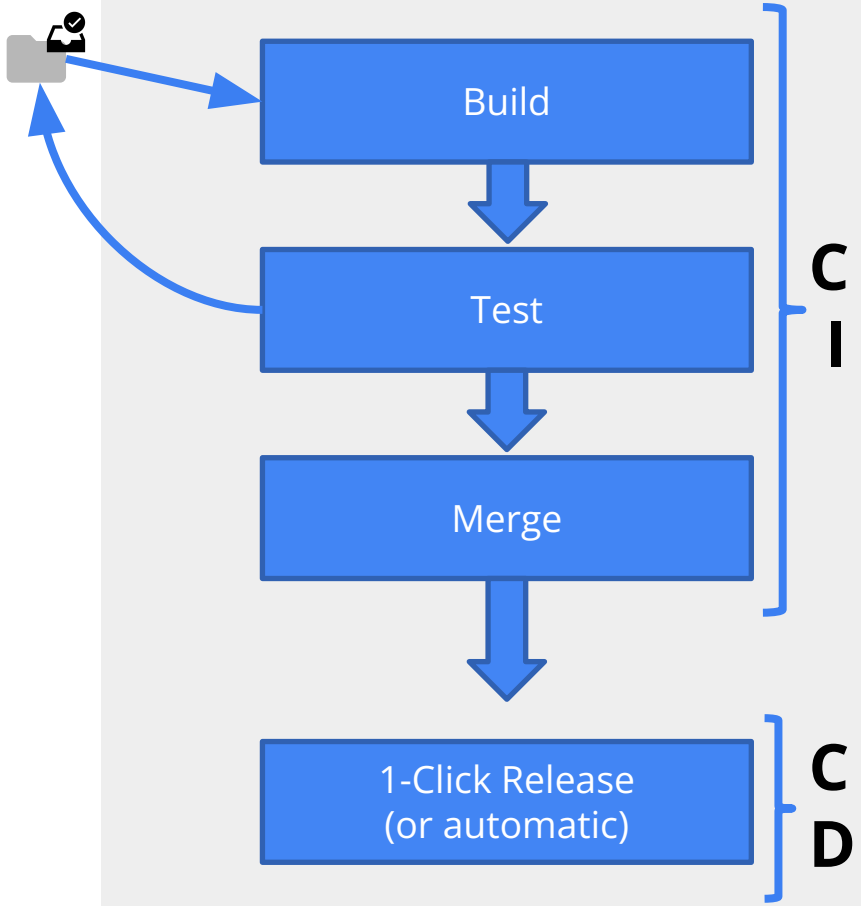
CI/CD

- Nightly Jobs
- New Data
- Pipeline Updates



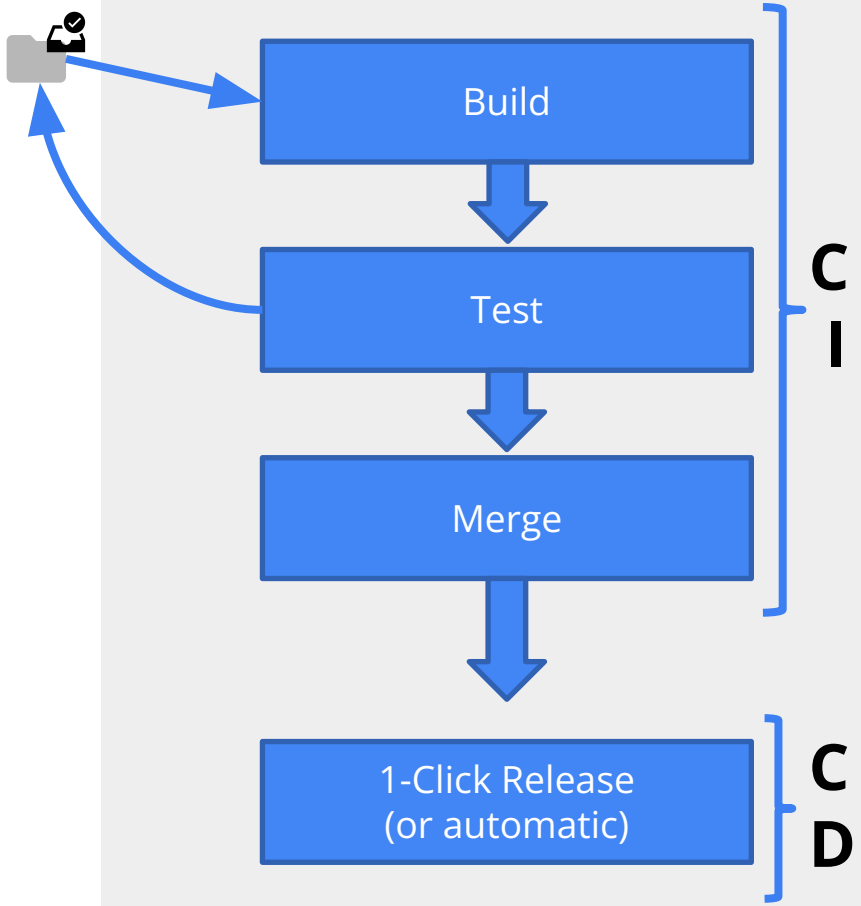
CI/CD

- Nightly Jobs
- New Data
- Pipeline Updates
- Model/Param Updates



CI/CD

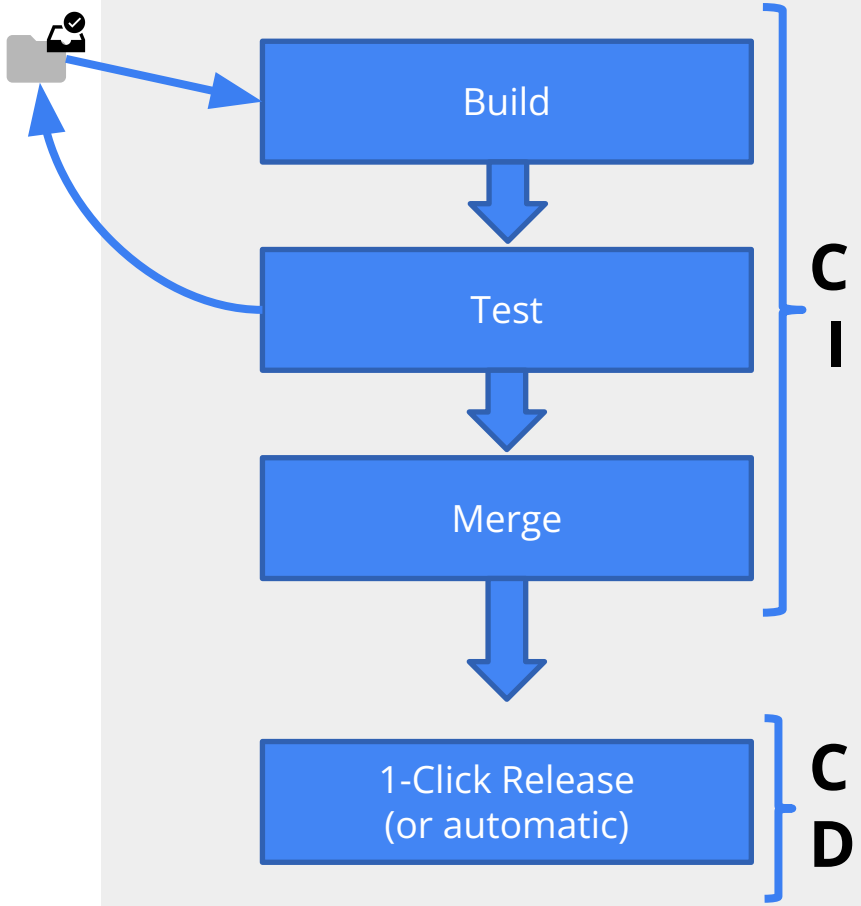
- Nightly Jobs
- New Data
- Pipeline Updates
- Model/Param Updates
- Hard Regressions (Tests)



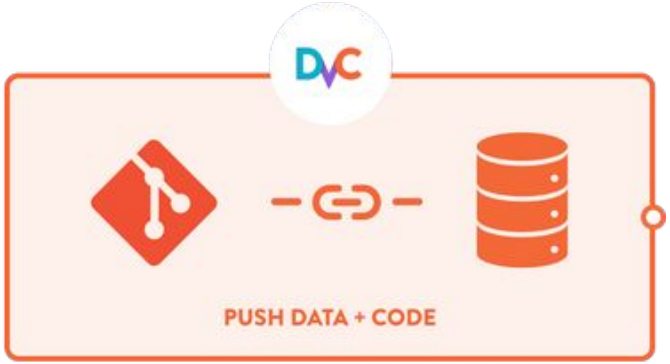
CI/CD

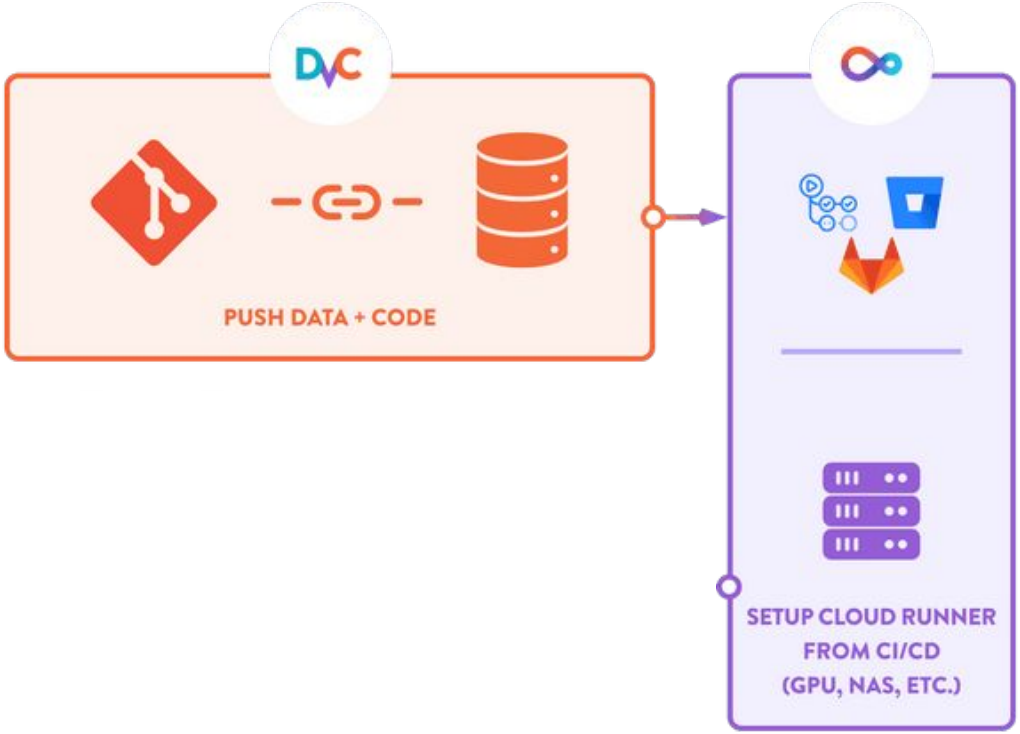
- Nightly Jobs
- New Data
- Pipeline Updates
- Model/Param Updates
- Hard Regressions (Tests)

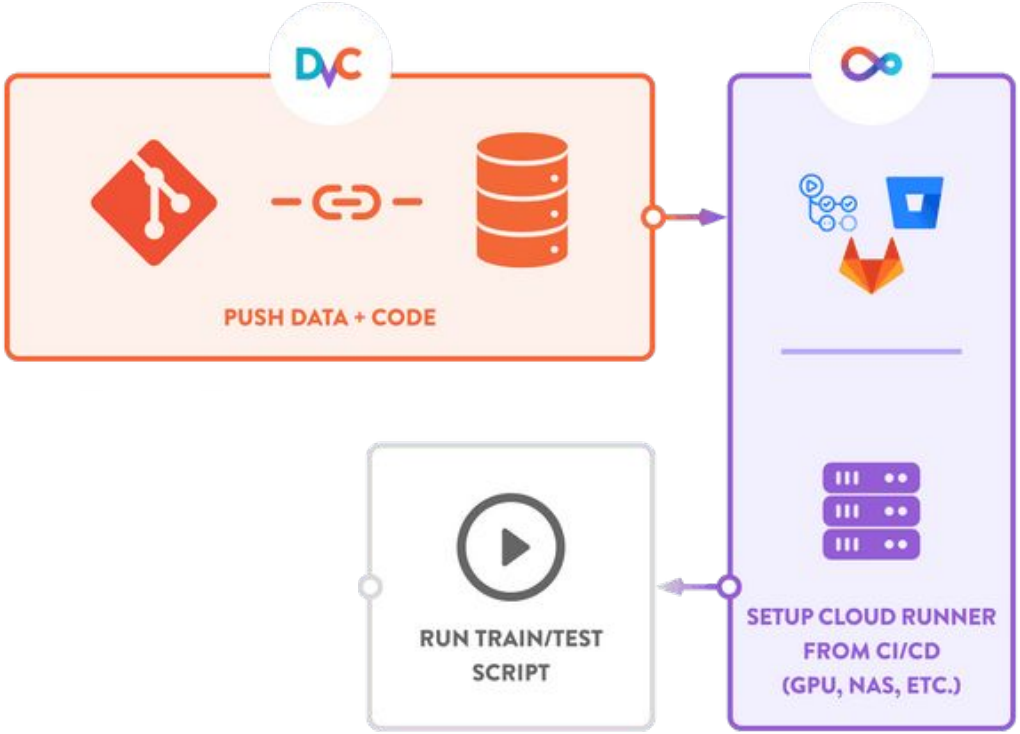
Everything is **tracked** and **saved**

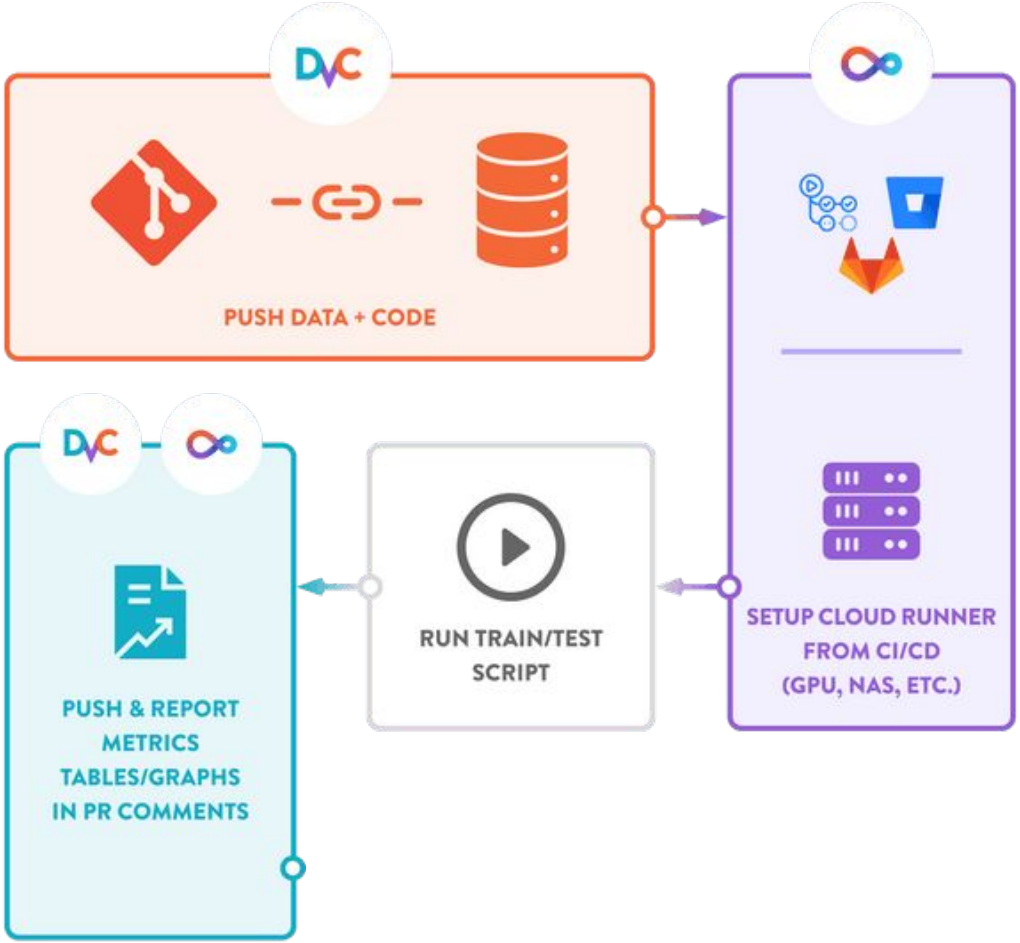


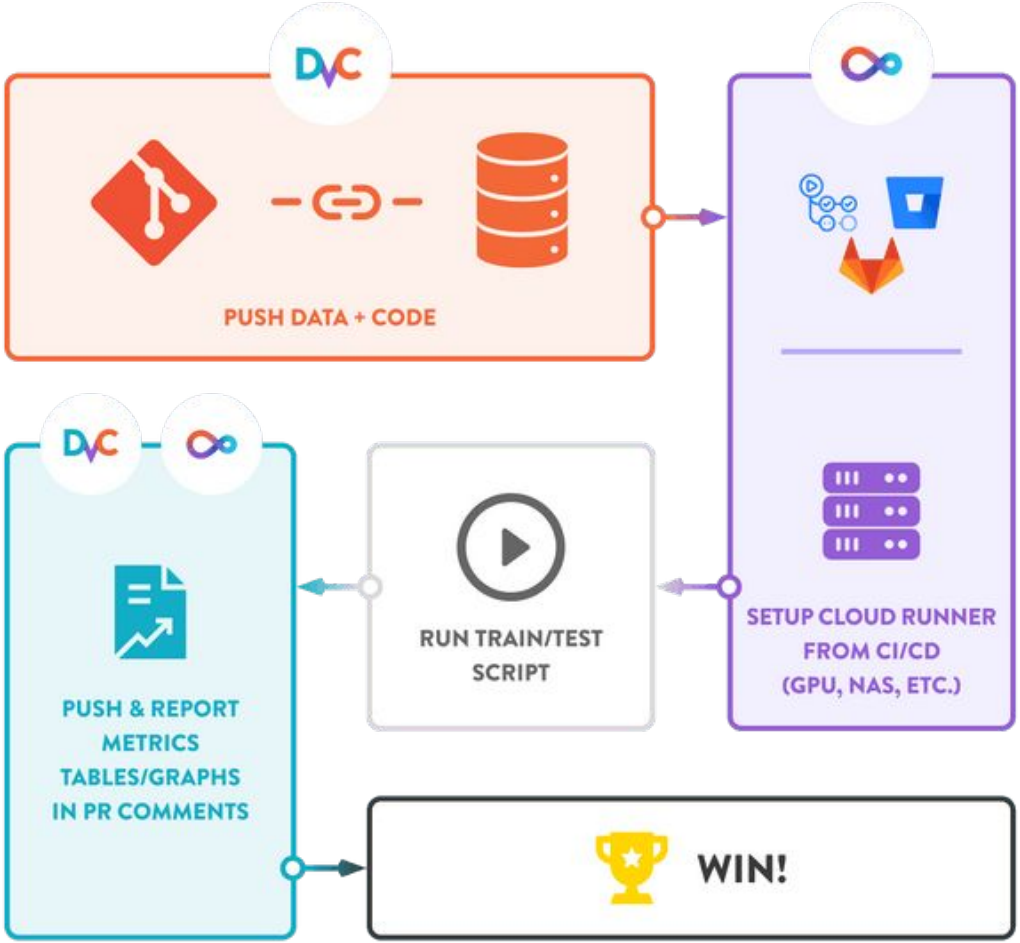
Combine


DVC
PUSH DATA + CODE

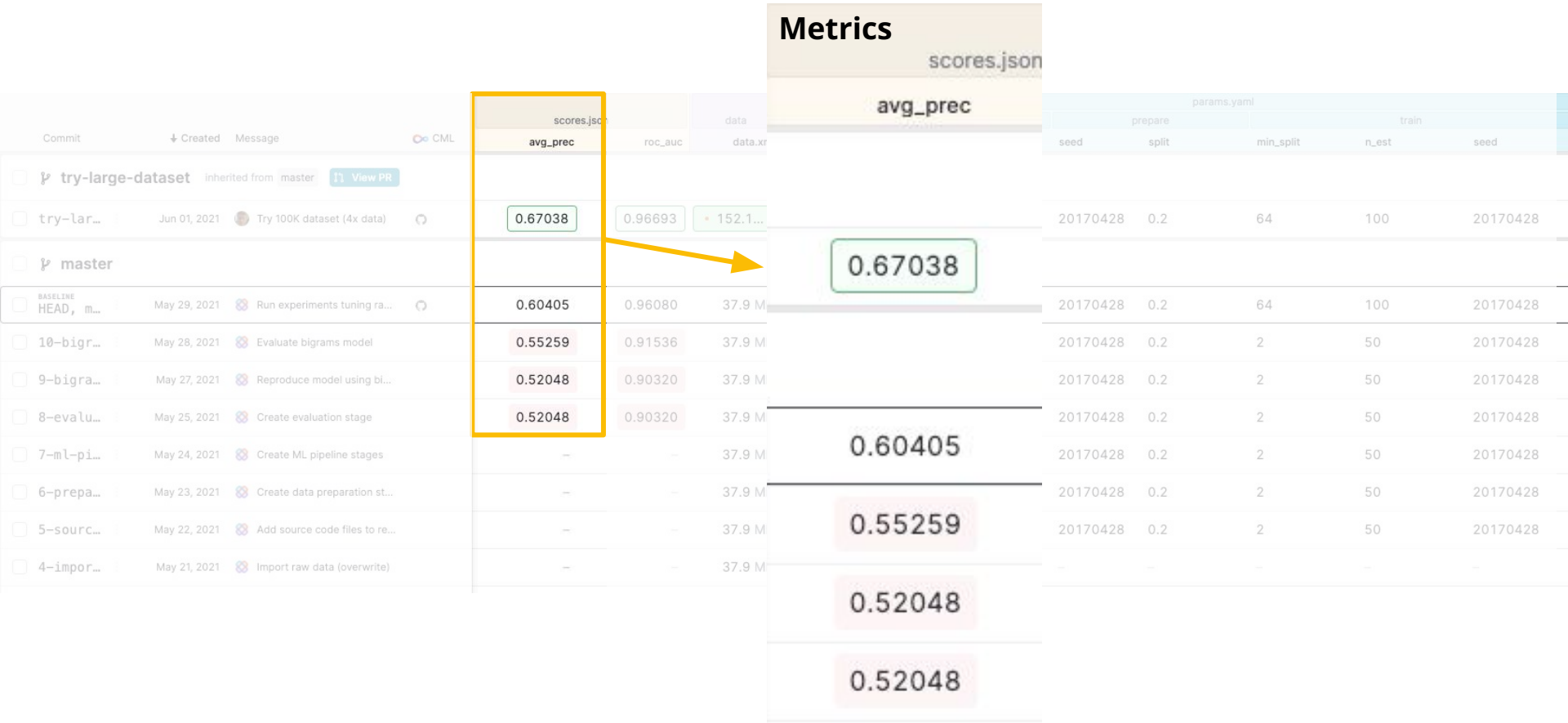

DVC
PUSH DATA + CODE
SETUP CLOUD RUNNER
FROM CI/CD
(GPU, NAS, ETC.)

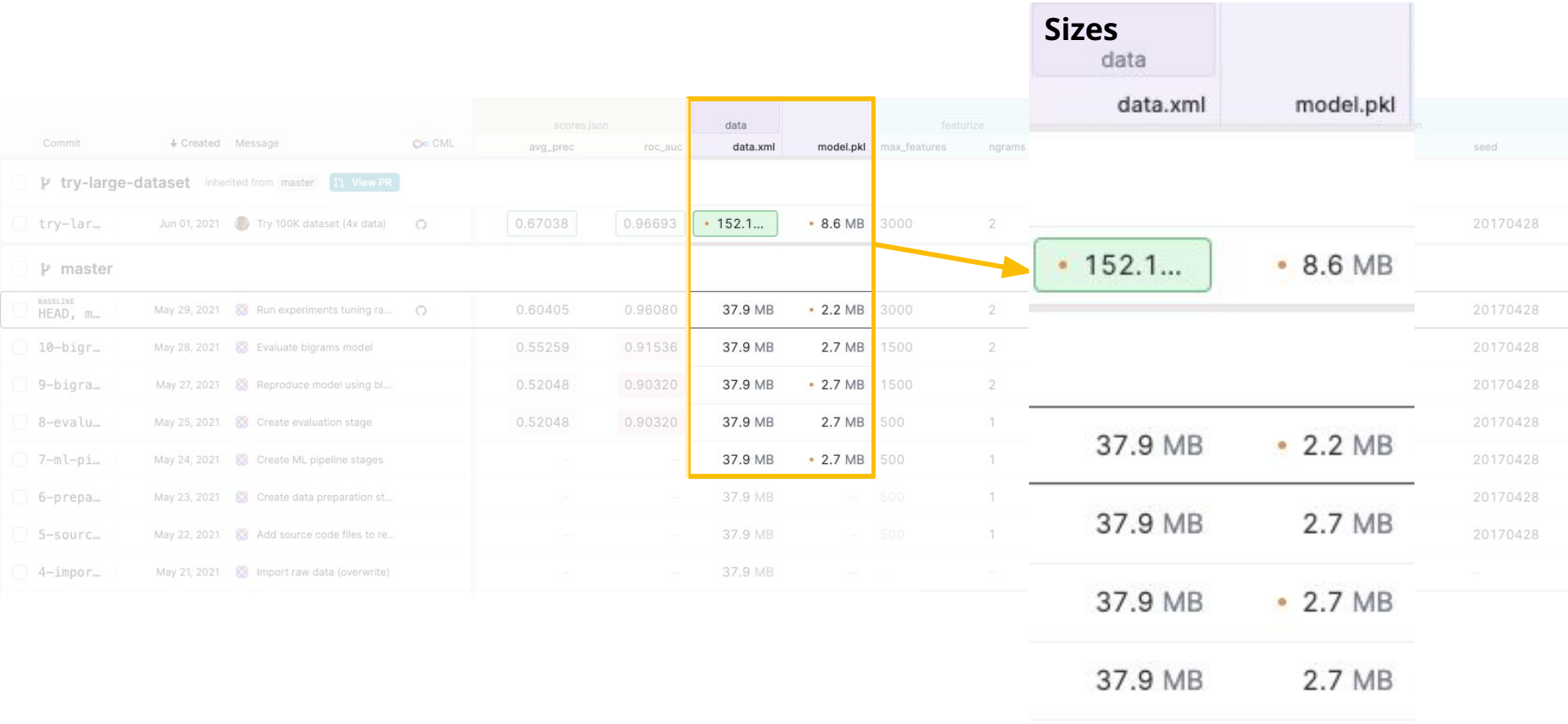

DVC
PUSH DATA + CODE
RUN TRAIN/TEST
SCRIPT
SETUP CLOUD RUNNER
FROM CI/CD
(GPU, NAS, ETC.)

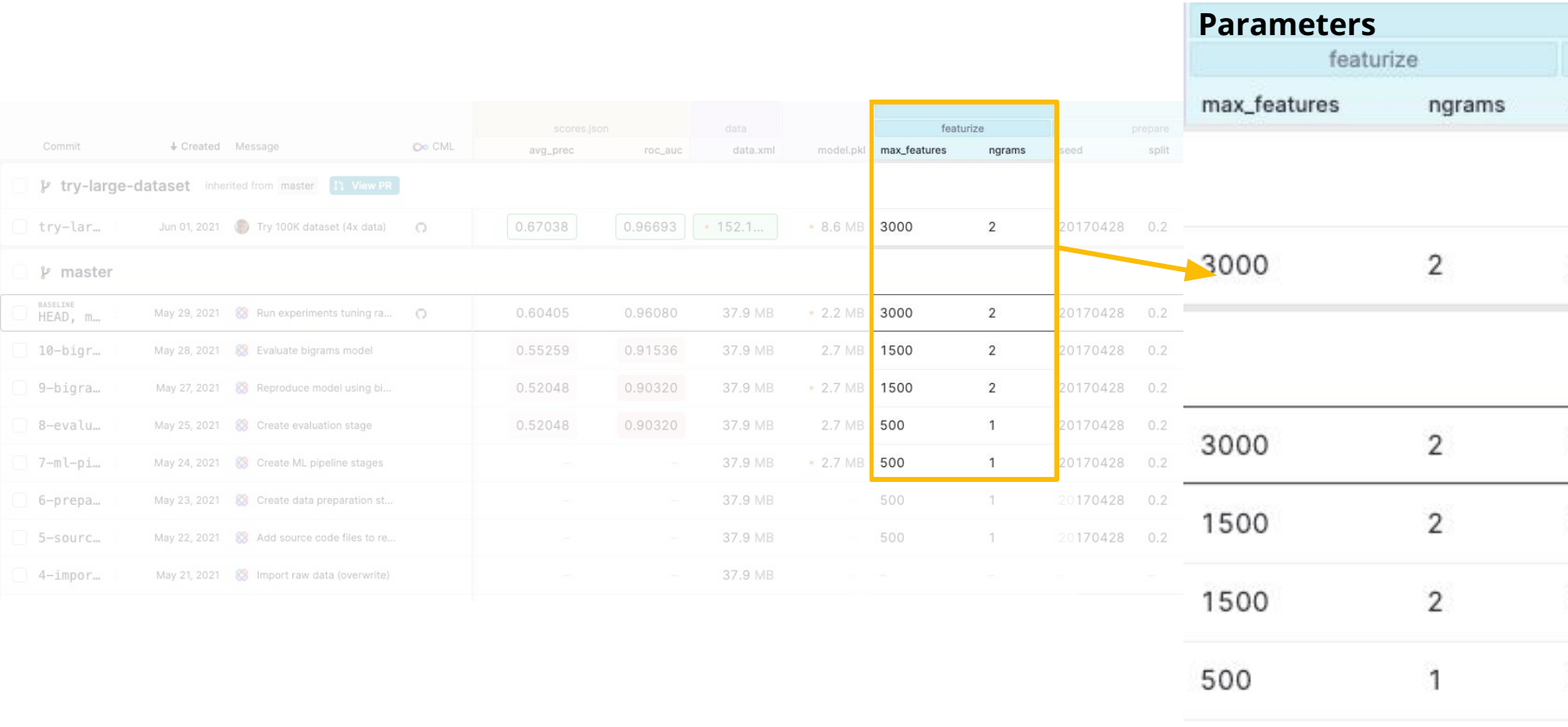

PUSH DATA + CODE
SETUP CLOUD RUNNER
FROM CI/CD
(GPU, NAS, ETC.)
RUN TRAIN/TEST
SCRIPT
PUSH & REPORT
METRICS
TABLES/GRAPHS
IN PR COMMENTS

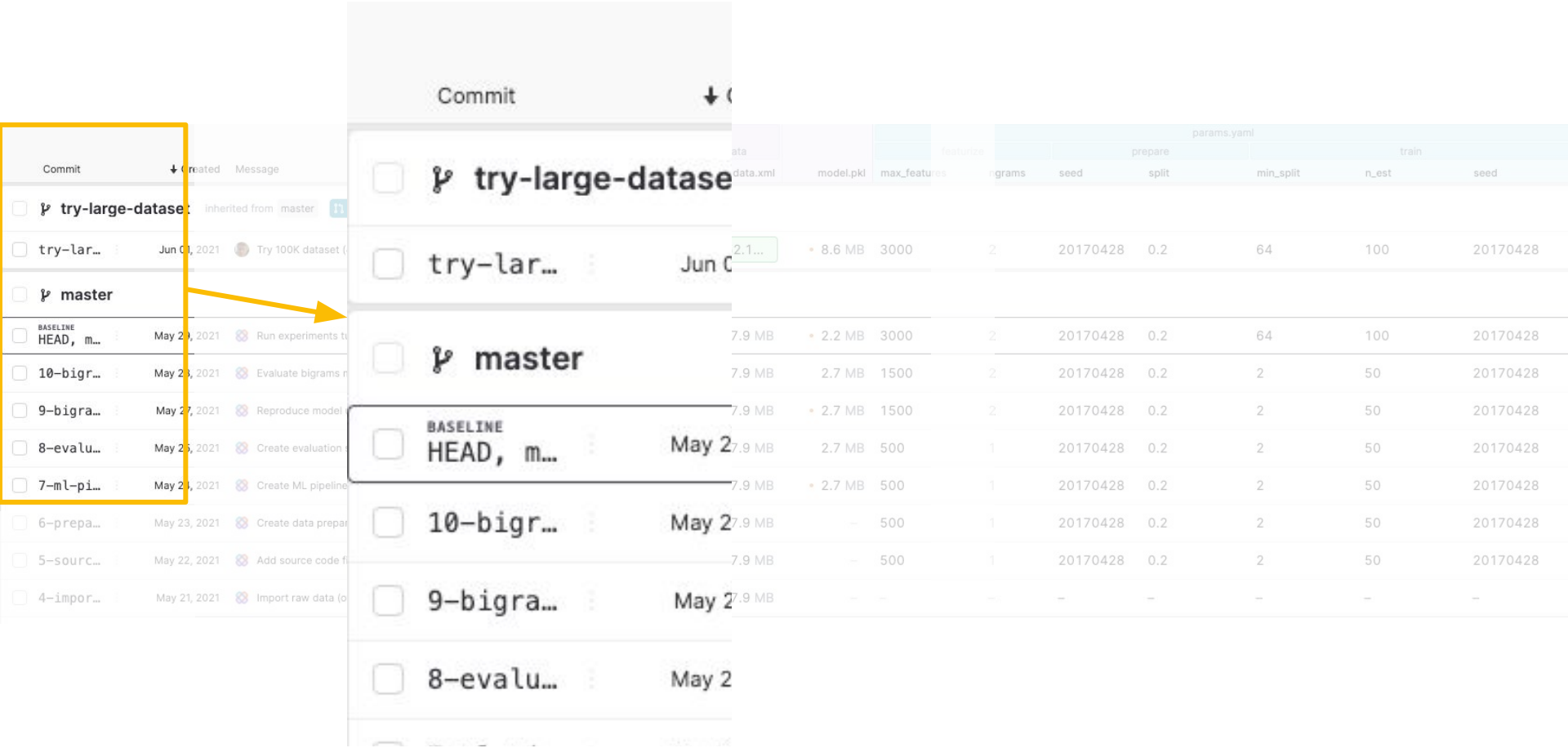

PUSH DATA + CODE
SETUP CLOUD RUNNER
FROM CI/CD
(GPU, NAS, ETC.)
RUN TRAIN/TEST
SCRIPT
PUSH & REPORT
METRICS
TABLES/GRAPHS
IN PR COMMENTS
WIN!

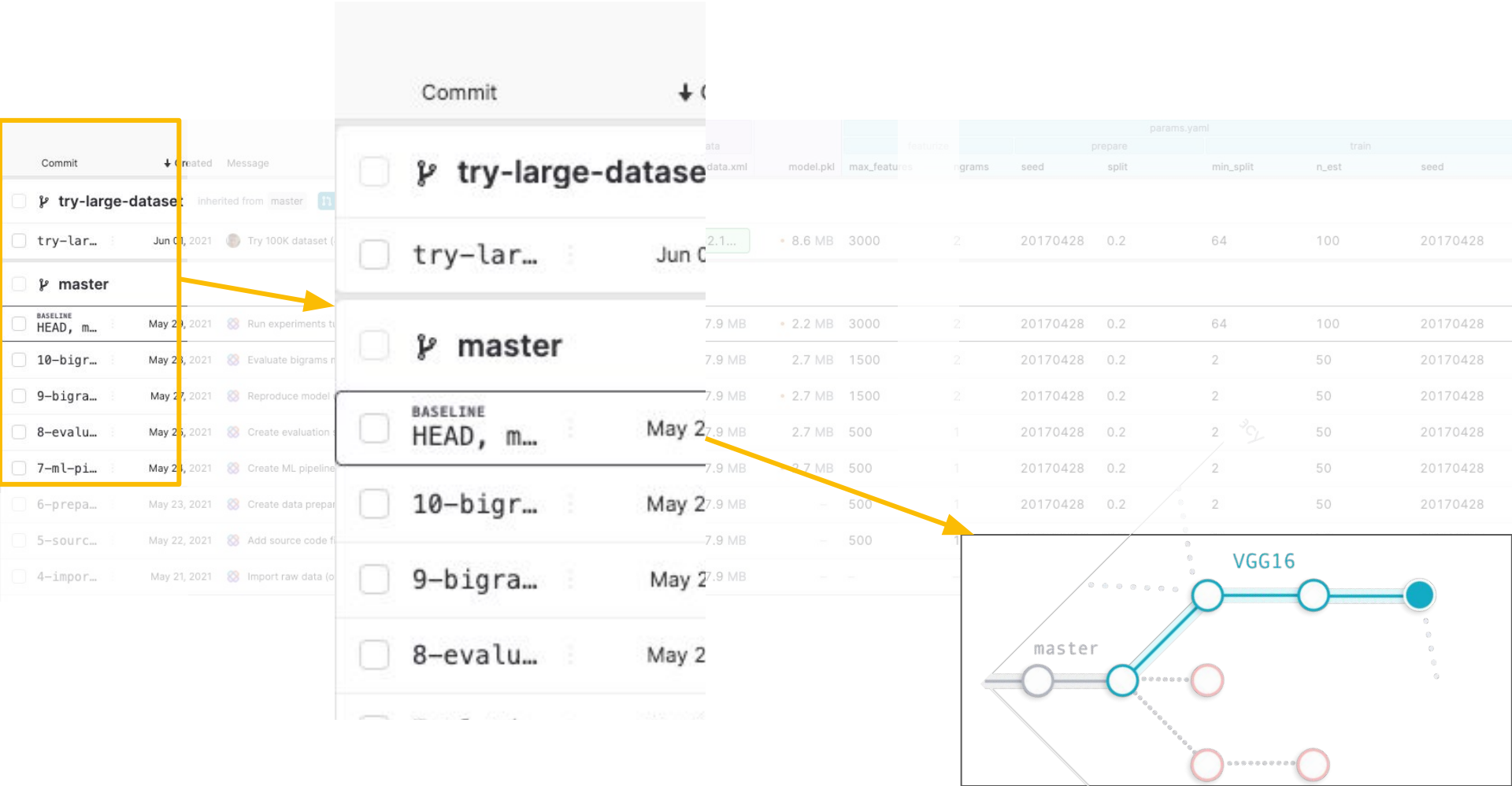
Commit	Created	Message	CML	scores.json		data		params.yaml						
								featurize		prepare		train		
				avg_prec	roc_auc	data.xml	model.pkl	max_features	ngrams	seed	split	min_split	n_est	seed
try-large-dataset inherited from master View PR														
try-lar…	Jun 01, 2021	Try 100K dataset (4x data)		0.67038	0.96693	152.1…	8.6 MB	3000	2	20170428	0.2	64	100	20170428
master														
BASELINE HEAD, m…	May 29, 2021	Run experiments tuning ra…		0.60405	0.96080	37.9 MB	2.2 MB	3000	2	20170428	0.2	64	100	20170428
10-bigr…	May 28, 2021	Evaluate bigrams model		0.55259	0.91536	37.9 MB	2.7 MB	1500	2	20170428	0.2	2	50	20170428
9-bigra…	May 27, 2021	Reproduce model using bi…		0.52048	0.90320	37.9 MB	2.7 MB	1500	2	20170428	0.2	2	50	20170428
8-evalu…	May 25, 2021	Create evaluation stage		0.52048	0.90320	37.9 MB	2.7 MB	500	1	20170428	0.2	2	50	20170428
7-ml-pi…	May 24, 2021	Create ML pipeline stages		–	–	37.9 MB	2.7 MB	500	1	20170428	0.2	2	50	20170428
6-prepa…	May 23, 2021	Create data preparation st…		–	–	37.9 MB	–	500	1	20170428	0.2	2	50	20170428
5-sourc…	May 22, 2021	Add source code files to re…		–	–	37.9 MB	–	500	1	20170428	0.2	2	50	20170428
4-impor…	May 21, 2021	Import raw data (overwrite)		–	–	37.9 MB	–	–	–	–	–	–	–	–


Metrics
scores.json
avg_prec
0.67038
0.60405
0.55259
0.52048
0.52048
Commit
Created
Message
CML
scores.jsc
avg_prec
roc_auc
data
data.xr
params.yaml
prepare
train
seed
split
min_split
n_est
seed
try-large-dataset
inherited from master
View PR
try-lar…
Jun 01, 2021
Try 100K dataset (4x data)
0.67038
0.96693
152.1…
20170428
0.2
64
100
20170428
master
BASELINE
HEAD, m…
May 29, 2021
Run experiments tuning ra…
0.60405
0.96080
37.9 M
20170428
0.2
64
100
20170428
10-bigr…
May 28, 2021
Evaluate bigrams model
0.55259
0.91536
37.9 M
20170428
0.2
2
50
20170428
9-bigra…
May 27, 2021
Reproduce model using bi…
0.52048
0.90320
37.9 M
20170428
0.2
2
50
20170428
8-evalu…
May 25, 2021
Create evaluation stage
0.52048
0.90320
37.9 M
20170428
0.2
2
50
20170428
7-ml-pi…
May 24, 2021
Create ML pipeline stages
–
–
37.9 M
20170428
0.2
2
50
20170428
6-prepa…
May 23, 2021
Create data preparation st…
–
–
37.9 M
20170428
0.2
2
50
20170428
5-sourc…
May 22, 2021
Add source code files to re…
–
–
37.9 M
20170428
0.2
2
50
20170428
4-impor…
May 21, 2021
Import raw data (overwrite)
–
–
37.9 M
– – – –


Sizes
data
data.xml
model.pkl
Commit
Created
Message
CML
scores.json
avg_prec
roc_auc
data
data.xml
model.pkl
featurize
max_features
ngrams
seed
try-large-dataset
inherited from master
View PR
try-lar…
Jun 01, 2021
Try 100K dataset (4x data)
0.67038
0.96693
152.1…
8.6 MB
3000
2
20170428
master
BASELINE
HEAD, m…
May 29, 2021
Run experiments tuning ra…
0.60405
0.96080
37.9 MB
2.2 MB
3000
2
20170428
10-bigr…
May 28, 2021
Evaluate bigrams model
0.55259
0.91536
37.9 MB
2.7 MB
1500
2
20170428
9-bigra…
May 27, 2021
Reproduce model using bi…
0.52048
0.90320
37.9 MB
2.7 MB
1500
2
20170428
8-evalu…
May 25, 2021
Create evaluation stage
0.52048
0.90320
37.9 MB
2.7 MB
500
1
20170428
7-ml-pi…
May 24, 2021
Create ML pipeline stages
37.9 MB
2.7 MB
500
1
20170428
6-prepa…
May 23, 2021
Create data preparation st…
37.9 MB
500
1
20170428
5-sourc…
May 22, 2021
Add source code files to re…
37.9 MB
500
1
20170428
4-impor…
May 21, 2021
Import raw data (overwrite)
37.9 MB
152.1…
8.6 MB
37.9 MB
2.2 MB
37.9 MB
2.7 MB
37.9 MB
2.7 MB
37.9 MB
2.7 MB


Parameters
featurize
max_features
ngrams
3000
2
3000
2
1500
2
1500
2
500
1


Commit
try-large-datase
try-lar…
Jun
master
BASELINE
HEAD, m…
May 2
10-bigr…
May 2
9-bigra…
May 2
8-evalu…
May 2


Commit
try-large-datase
try-lar…
master
BASELINE
HEAD, m…
May 2
10-bigr…
May 2
9-bigra…
May 2
8-evalu…
May 2
VGG16
master

Commit	Created	Message	CML	scores.json		data		params.yaml						
								featurize		prepare		train		
				avg_prec	roc_auc	data.xml	model.pkl	max_features	ngrams	seed	split	min_split	n_est	seed
try-large-dataset inherited from master View PR														
try-lar…	Jun 01, 2021	Try 100K dataset (4x data)		0.67038	0.96693	152.1…	8.6 MB	3000	2	20170428	0.2	64	100	20170428
master														
BASELINE HEAD, m…	May 29, 2021	Run experiments tuning ra…		0.60405	0.96080	37.9 MB	2.2 MB	3000	2	20170428	0.2	64	100	20170428
10-bigr…	May 28, 2021	Evaluate bigrams model		0.55259	0.91536	37.9 MB	2.7 MB	1500	2	20170428	0.2	2	50	20170428
9-bigra…	May 27, 2021	Reproduce model using bi…		0.52048	0.90320	37.9 MB	2.7 MB	1500	2	20170428	0.2	2	50	20170428
8-evalu…	May 25, 2021	Create evaluation stage		0.52048	0.90320	37.9 MB	2.7 MB	500	1	20170428	0.2	2	50	20170428
7-ml-pi…	May 24, 2021	Create ML pipeline stages		–	–	37.9 MB	2.7 MB	500	1	20170428	0.2	2	50	20170428
6-prepa…	May 23, 2021	Create data preparation st…		–	–	37.9 MB	–	500	1	20170428	0.2	2	50	20170428
5-sourc…	May 22, 2021	Add source code files to re…		–	–	37.9 MB	–	500	1	20170428	0.2	2	50	20170428
4-impor…	May 21, 2021	Import raw data (overwrite)		–	–	37.9 MB	–	–	–	–	–	–	–	–


Views
example-get-started
Demo
Private
Help
Filters
Columns
Show plots
Compare
Selected only
Trends
Delta mode
Commits
try-large...
try-lar…
master
BASELINE
HEAD, m…
Directories
Search
Clear all
Selected 2 of 2
prc.json
roc.json
Plots
HEAD
try-large-dataset
prc.json
precision
recall
roc.json
tpr
fpr

Why care?

Safe

Safe
Clean

Safe
Clean
Standardized

Quick Google of companies using...

DVC / CML

aporia

codecentric

BILLIE

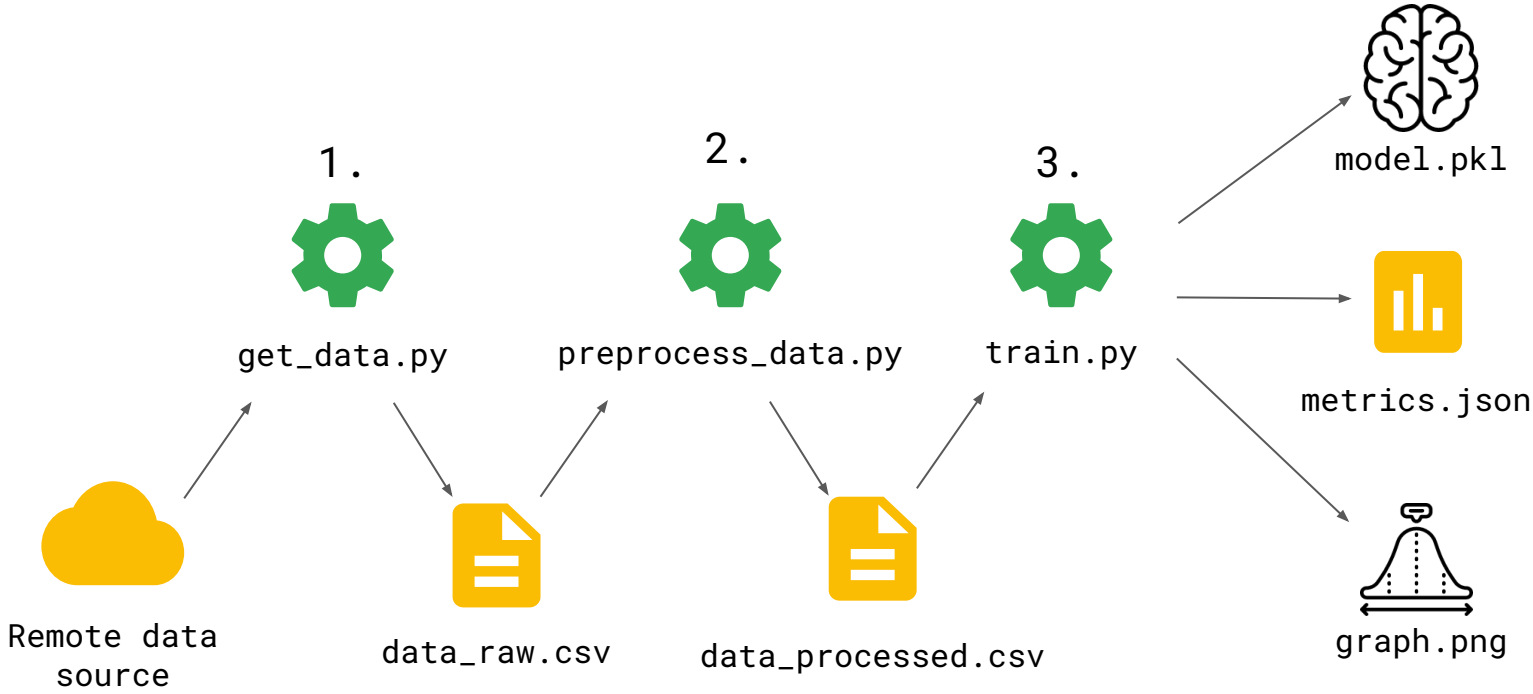
Any type of CI

flair

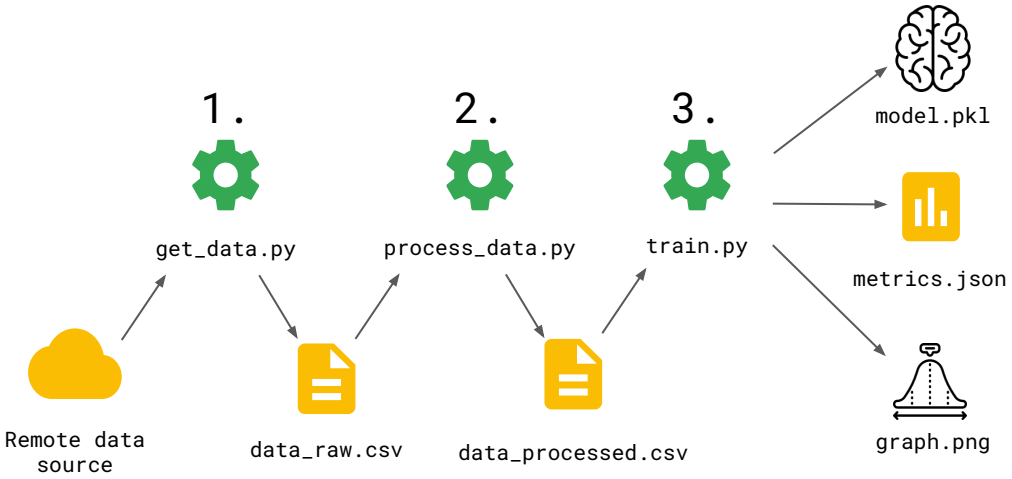
spaCy

Quick walk-through

DVC **101**-style



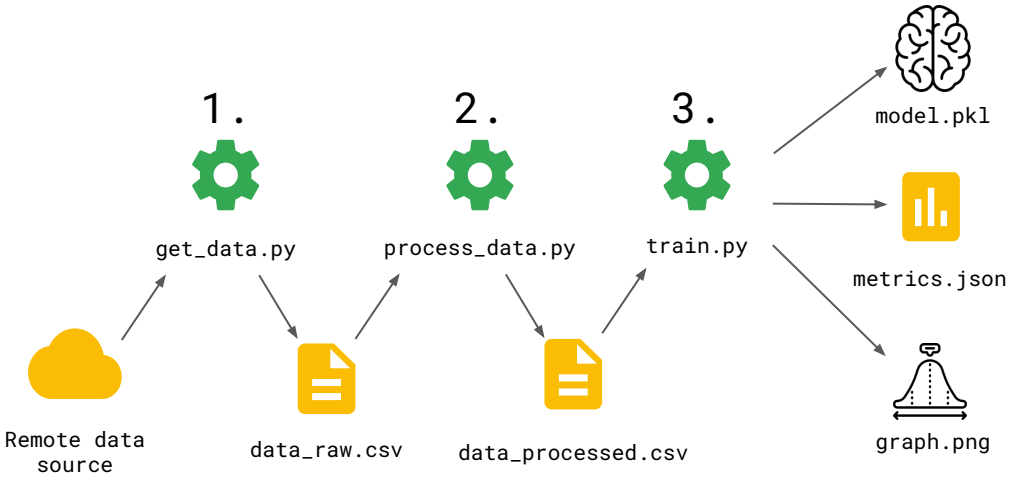
DVC **101**-style



dvc.yml

```
stages:
 get_data:
   cmd: python get_data.py
   deps:
   - get_data.py
   outs:
   - data_raw.csv
 process:
   cmd: python process_data.py
...
```

DVC **101**-style

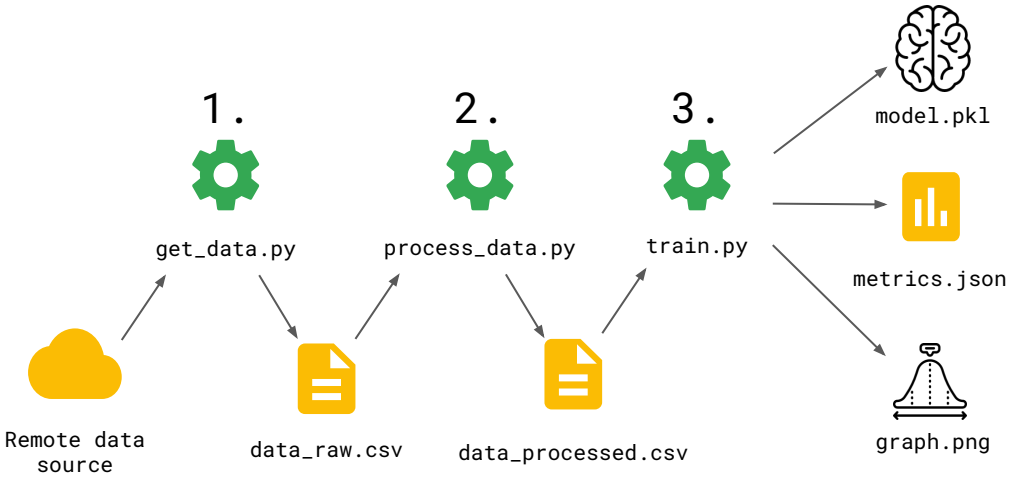


dvc.yml

1.

```
stages:
 get_data:
   cmd: python get_data.py
   deps:
   - get_data.py
   outs:
   - data_raw.csv
 process:
   cmd: python process_data.py
...
```

DVC **101**-style

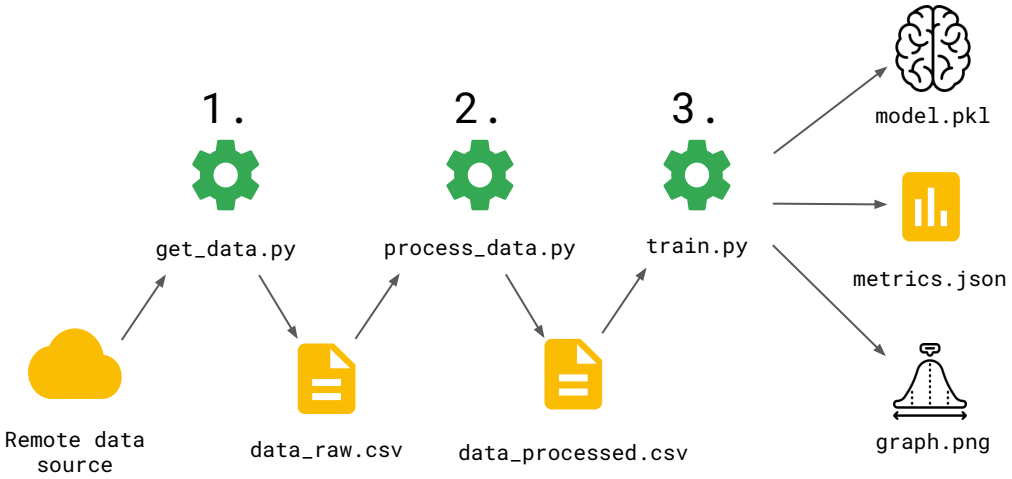


1.

dvc.yml

```
stages:
 get_data:
   cmd: python get_data.py
   deps:
   - get_data.py
   outs:
   - data_raw.csv
 process:
   cmd: python process_data.py
...
```

DVC **101**-style

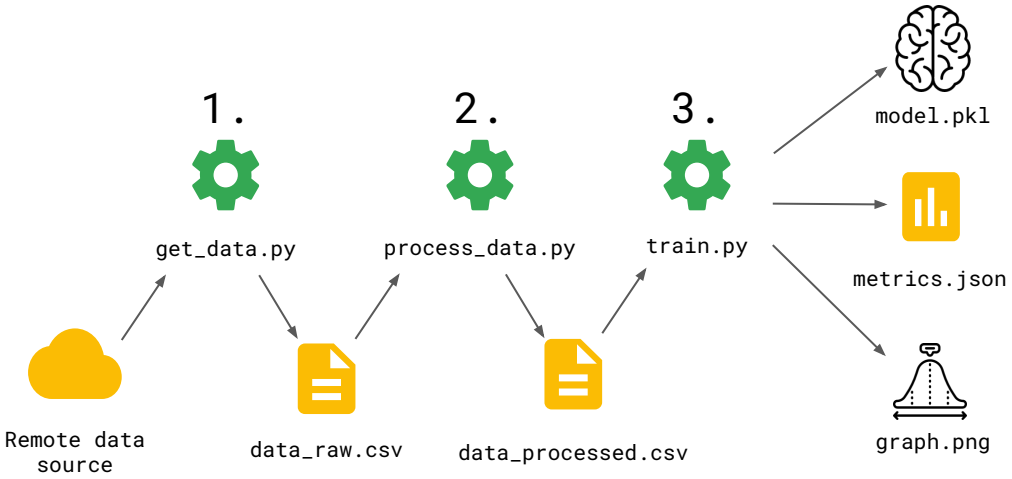


dvc.yml

1.

```
stages:
 get_data:
   cmd: python get_data.py
   deps:
   - get_data.py
   outs:
   - data_raw.csv
 process:
   cmd: python process_data.py
...
```

DVC **101**-style

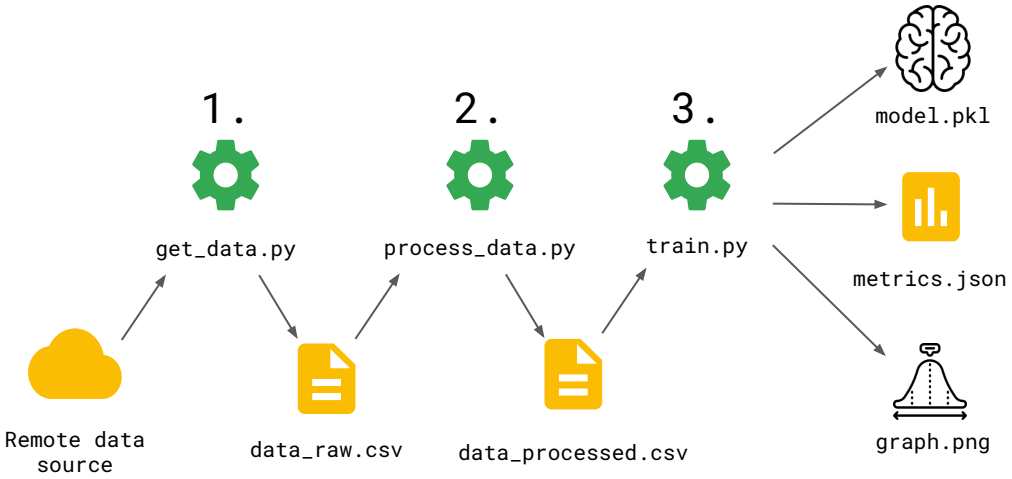


dvc.yml

1.

```
stages:
 get_data:
   cmd: python get_data.py
   deps:
   - get_data.py
   outs:
   - data_raw.csv
 process:
   cmd: python process_data.py
...
```

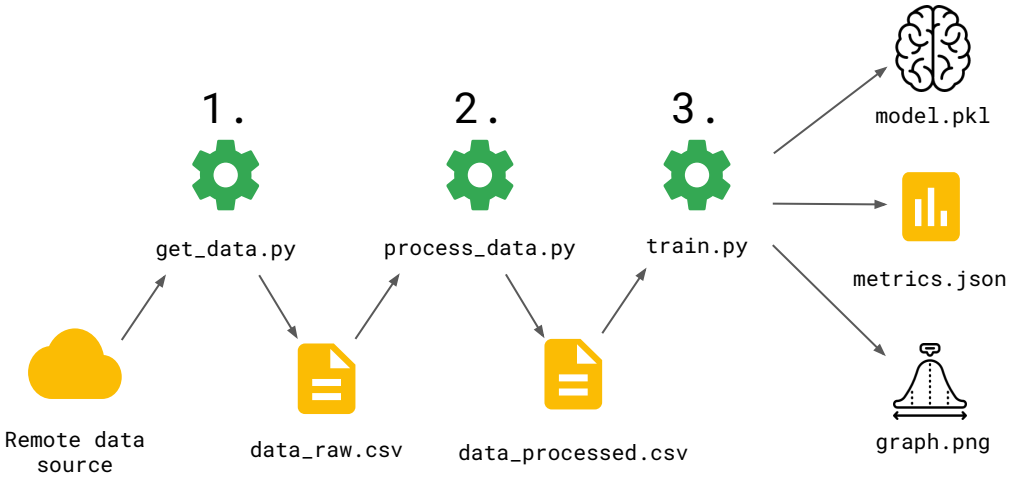
DVC **101**-style



dvc.yml

```
stages:
 get_data:
    cmd: python get_data.py
    deps:
    - get_data.py
    outs:
    - data_raw.csv
 process:
    cmd: python process_data.py
...
```

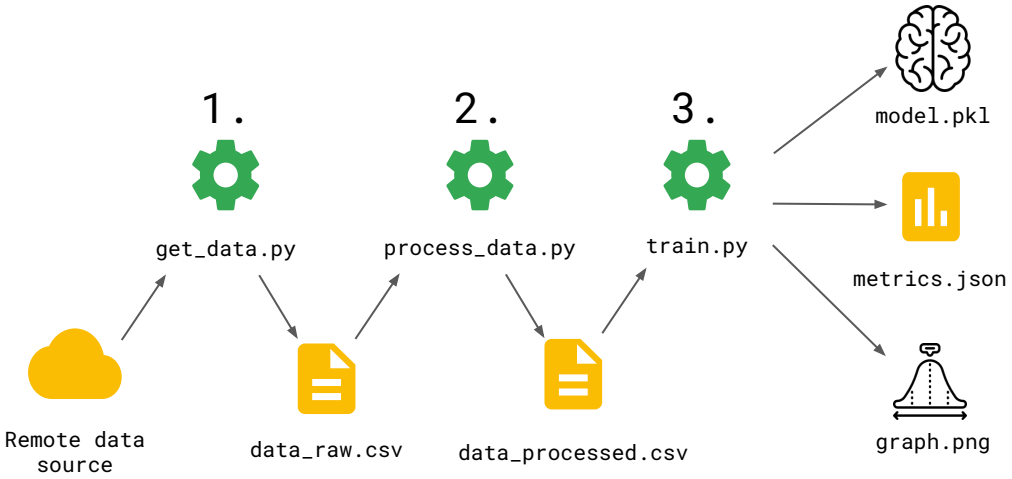
DVC **101**-style



dvc.yml

```
stages:
 get_data:
   cmd: python get_data.py
   deps:
   - get_data.py
   outs:
   - data_raw.csv
 process:
   cmd: python process_data.py
...
```

DVC **101**-style

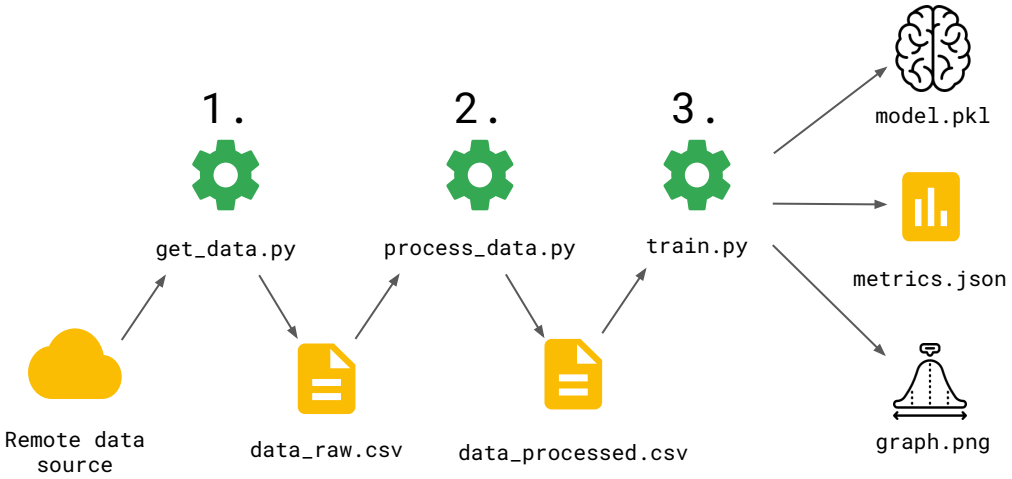


dvc.yml

```
stages:
 get_data:
   cmd: python get_data.py
   deps:
   - get_data.py
   outs:
   - data_raw.csv
 process:
   cmd: python process_data.py
...
```

```
$ dvc metrics diff
```

DVC **101**-style

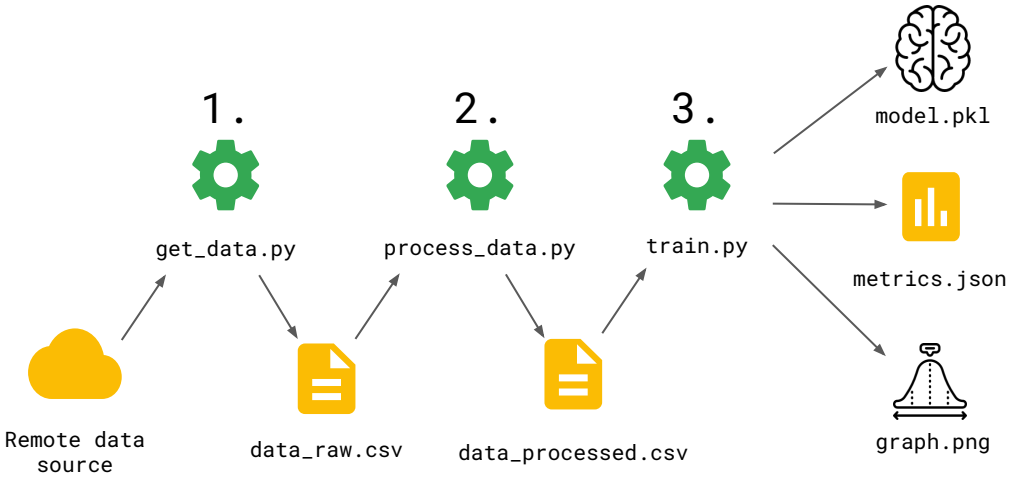


dvc.yml

```
stages:
 get_data:
   cmd: python get_data.py
   deps:
   - get_data.py
   outs:
   - data_raw.csv
 process:
   cmd: python process_data.py
...
```

```
$ dvc metrics diff
$ dvc plots
```

DVC **101**-style



dvc.yml

```
stages:
 get_data:
   cmd: python get_data.py
   deps:
   - get_data.py
   outs:
   - data_raw.csv
 process:
   cmd: python process_data.py
...
```

```
$ dvc metrics diff
$ dvc plots
$ dvc repro
```

CML **101**-style

> .github/workflows/cml.yml

```
...
run: |
  …
  cml publish plot.png --md >> report.md
  cml-send-comment report.md
  …
  cml-runner --cloud gcp <...>
  …
...
```

CML **101**-style

> .github/workflows/cml.yml

```
...
run: |
  …
  cml publish plot.png --md >> report.md
  cml-send-comment report.md
  …
  cml-runner --cloud gcp <...>
  …
...
```

CML **101**-style

> .github/workflows/cml.yml

```
...
run: |
  …
  cml publish plot.png --md >> report.md
  cml-send-comment report.md
  …
  cml-runner --cloud gcp <...>
  …
...
```

CML **101**-style

> .github/workflows/cml.yml

```
...
run: |
  …
  cml publish plot.png --md >> report.md
  cml-send-comment report.md
  …
  cml-runner --cloud gcp <...>
  …
...
```

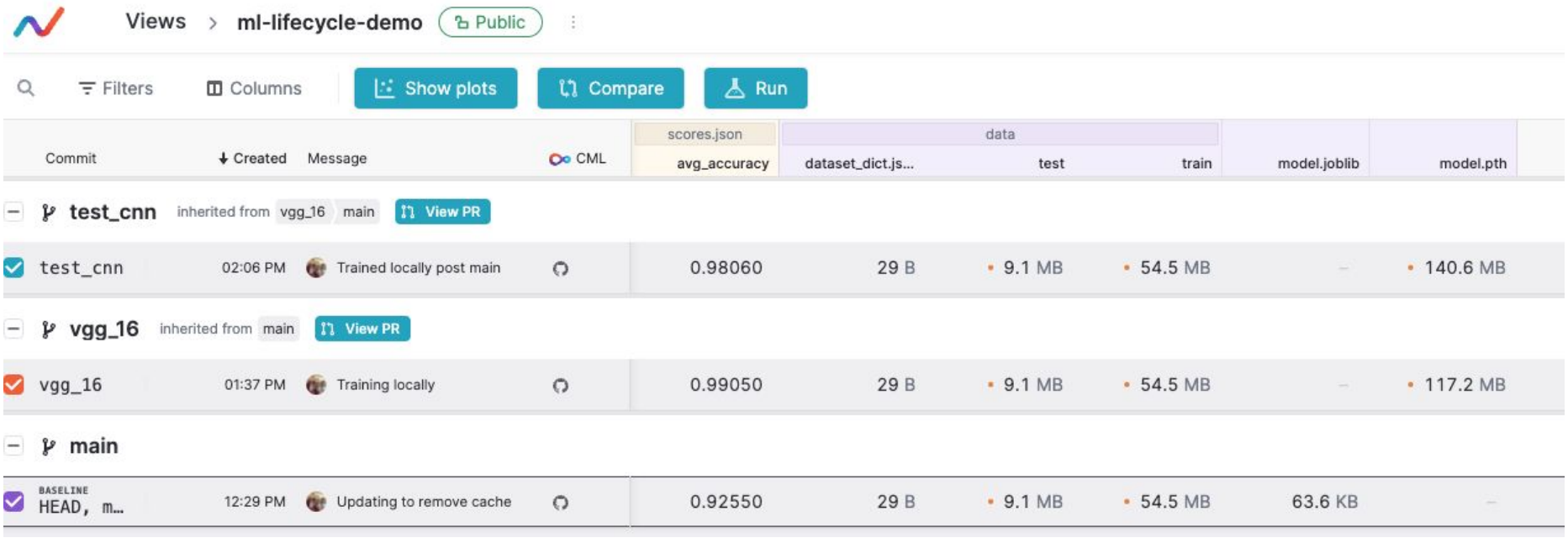
```
Running stage 'train':
> python train.py
100%|                                        | 60000/60000 [00:06<00:00, 9382.31ex/s]
100%|                                        | 10000/10000 [00:01<00:00, 9591.72ex/s]
/Users/londogard/miniforge3/envs/simswap/lib/python3.9/site-packages/sklearn/linear_model/_logistic.p
y:814: ConvergenceWarning: lbfgs failed to converge (status=1):
STOP: TOTAL NO. of ITERATIONS REACHED LIMIT.

Increase the number of iterations (max_iter) or scale the data as shown in:
    https://scikit-learn.org/stable/modules/preprocessing.html
Please also refer to the documentation for alternative solver options:
    https://scikit-learn.org/stable/modules/linear_model.html#logistic-regression
  n_iter_i = _check_optimize_result(
Accuracy: 0.9255

To track the changes with git, run:

        git add data/dataset_dict.json.dvc data/test.dvc data/train.dvc
Use `dvc push` to send your updates to remote storage.
(simswap)  londogard@MBPsontorHampus  main
```

DVC-Studio **101**-style



github.com/Lundez/ml-lifecycle-demo

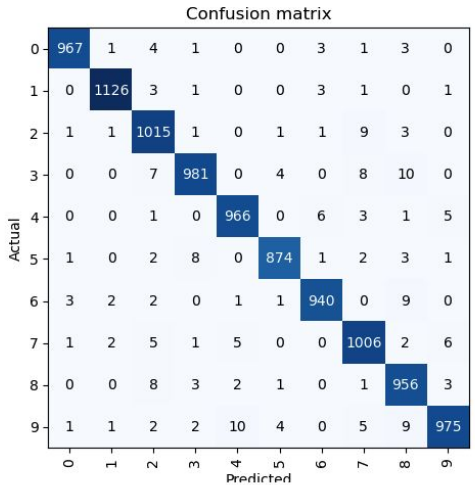
1 comment on commit `539d16e`

github-actions bot commented on 539d16e 2 hours ago

Metrics

Path	avg_accuracy
scores.json	0.9806

Confusion Matrix



Diff w/ main

Path	Metric	main	workspace	Change
scores.json	avg_accuracy	0.9255	0.9806	0.0551

github.com/Lundez/ml-lifecycle-demo

Questions?

github.com/Lundez/ml-lifecycle-demo

studio.iterative.ai/user/Lundez/views/ml-lifecycle-demo-uzx0yqj5io

@hlondogard lundez & londogard hampus.londogard@afry.com

1. **Neural net training is a leaky abstraction**
2. **Neural net training fails silently**

\- Andrej Karpathy
(Head of AI @Tesla)