

Deep Dive #3: Evaluations

Custom evaluation methods and
leaderboard set-up

Chris van Run

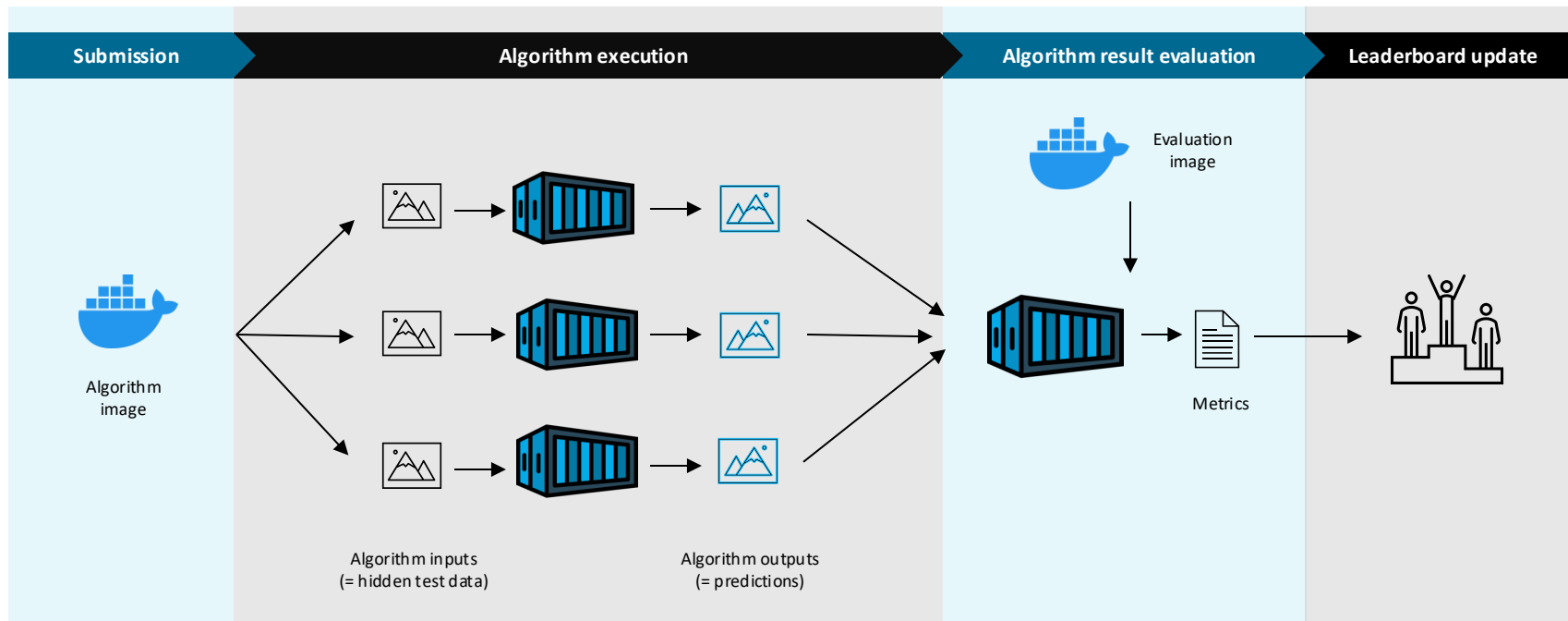
Research Software Engineer @grand-challenge.org

Radboudumc

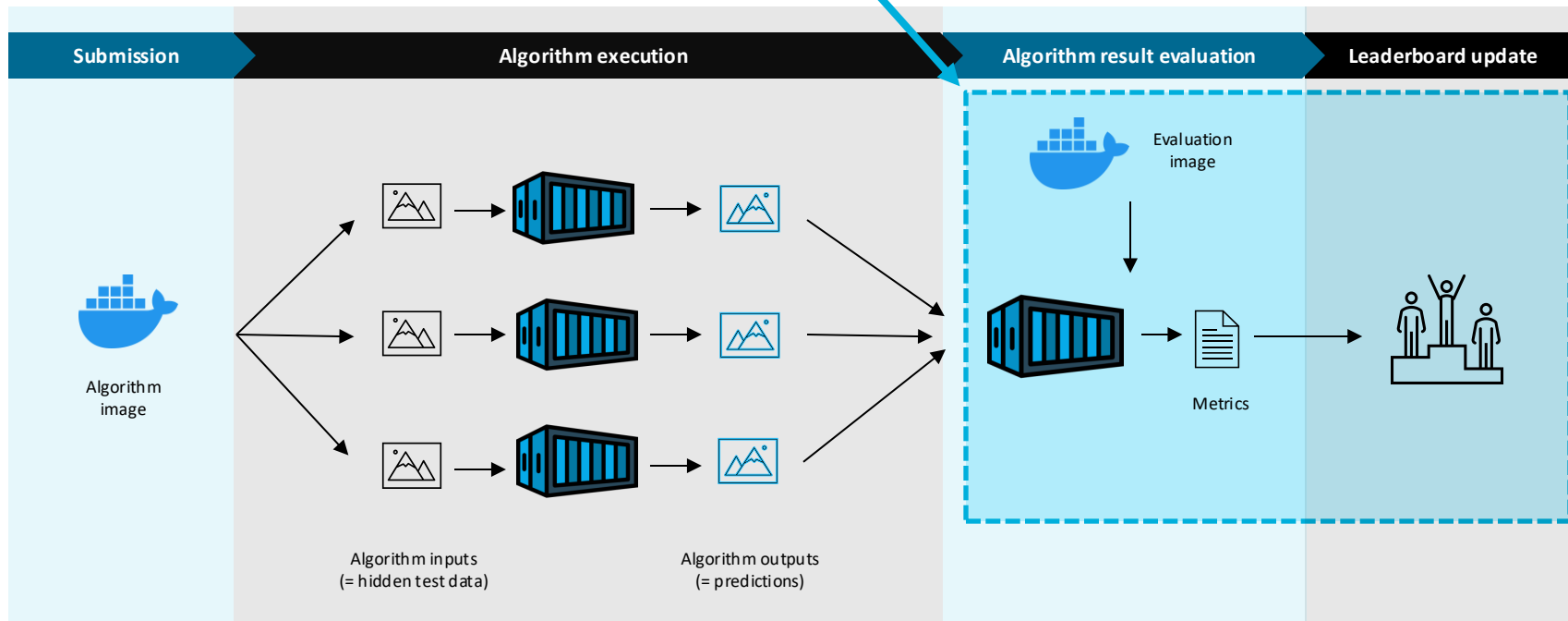
Agenda

Time	Topic
10:00 - 10:30	Welcome and introduction to challenges
10:30 - 11:00	Overview of the GC challenge feature
11:00 - 11:05	Short Break
11:05 - 11:45	Deep dive #1: uploading and managing hidden test data
11:45 - 13:00	Lunch Break
13:00 - 14:15	Deep dive #2: algorithm containers
14:15 - 14:30	Short Break
14:30 - 16:00	Deep dive #3: custom evaluation methods & leaderboard set-up
16:00 - 17:00	Wrap-up , Q&A

Submission workflow



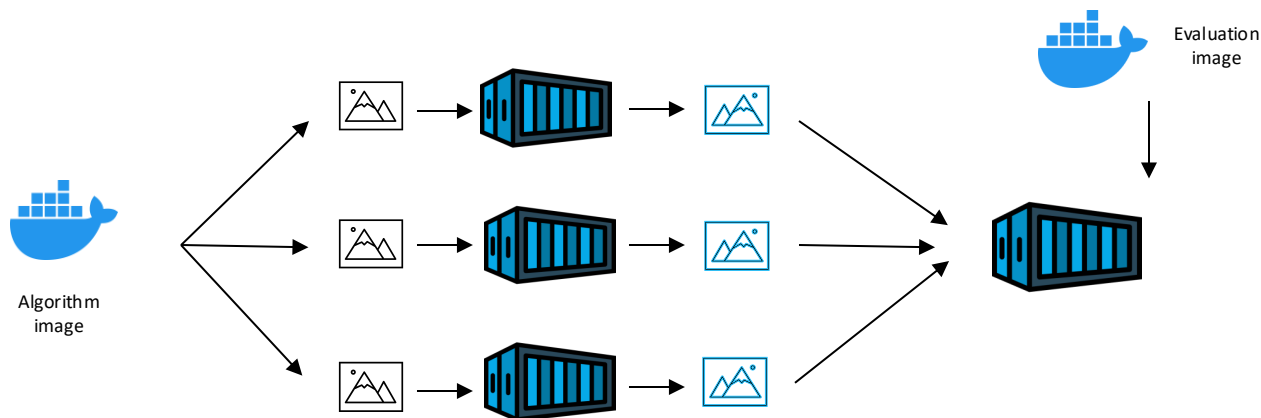
Submission workflow



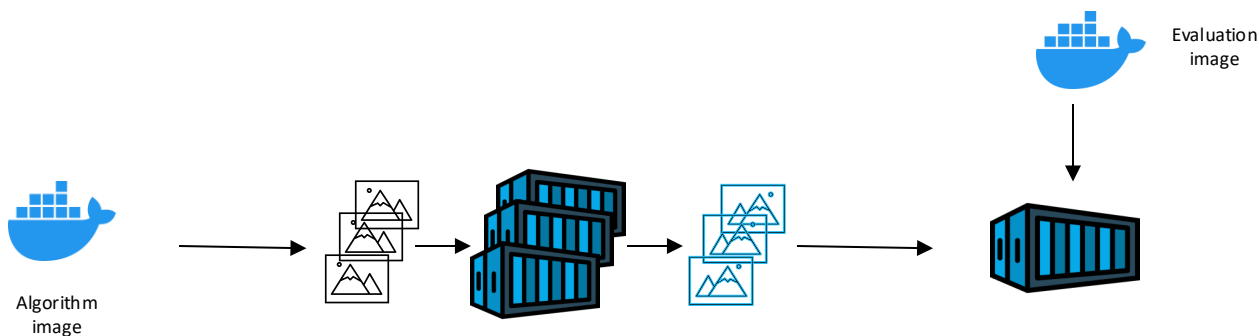
Evaluation

- Every challenge has a unique way of objectively evaluating incoming submissions.
- Scripts require a set of dependencies and computational environments.
- Evaluation Methods are hence container images.

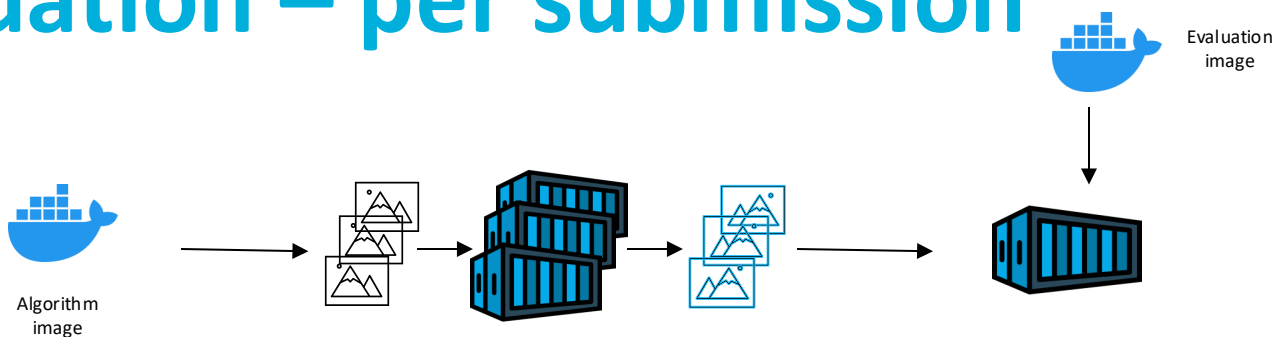
Evaluation – per submission



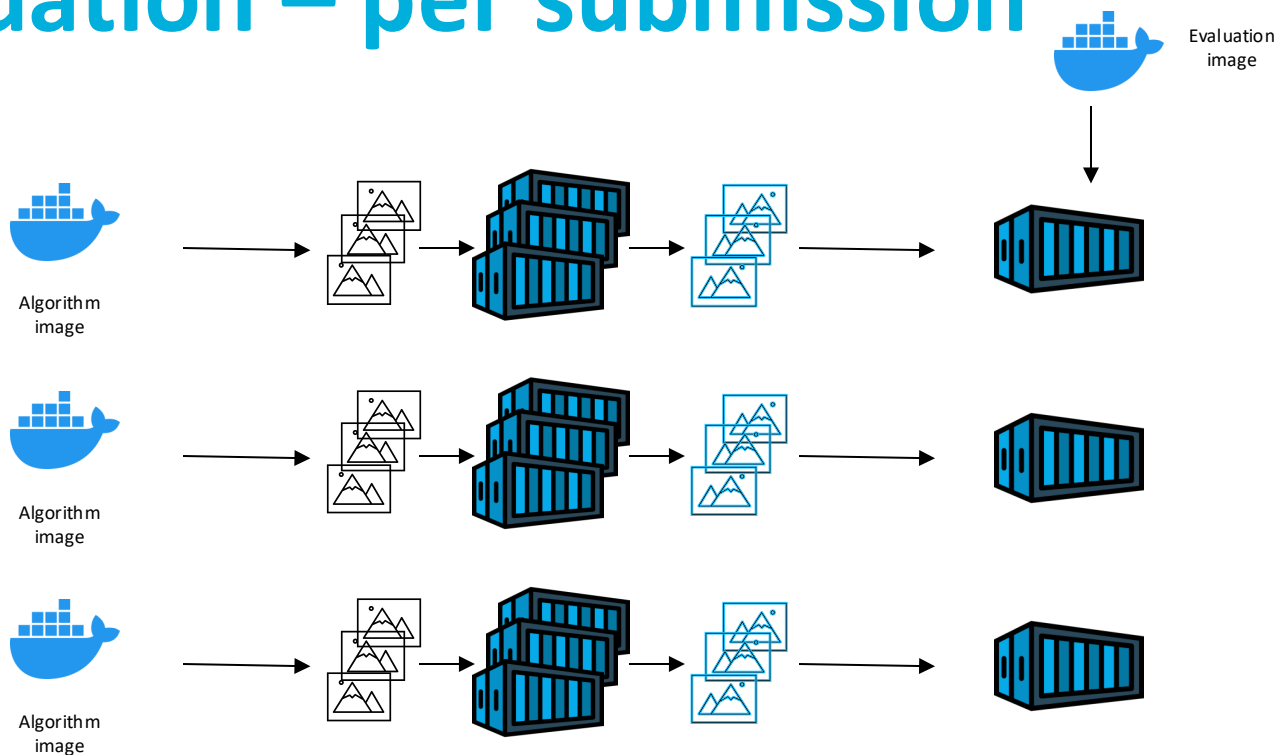
Evaluation – per submission



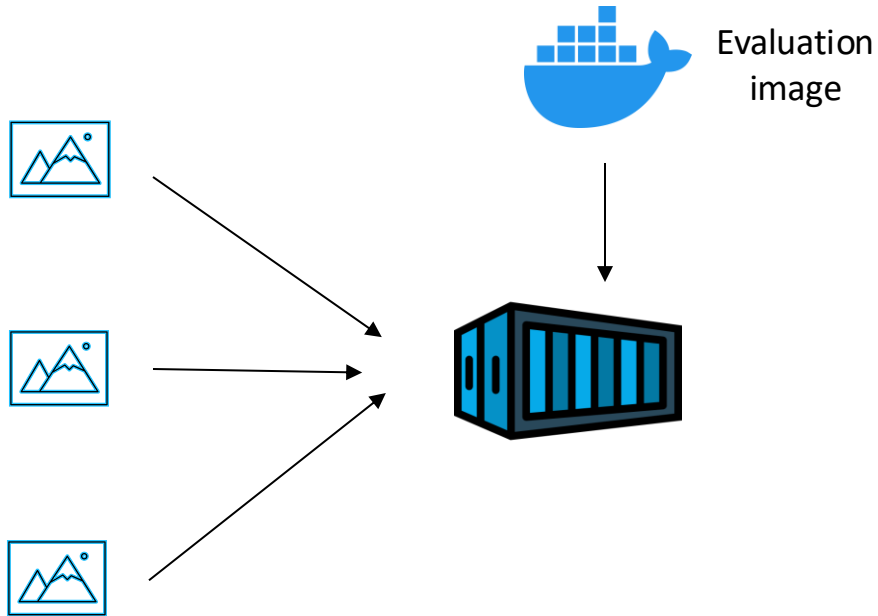
Evaluation – per submission



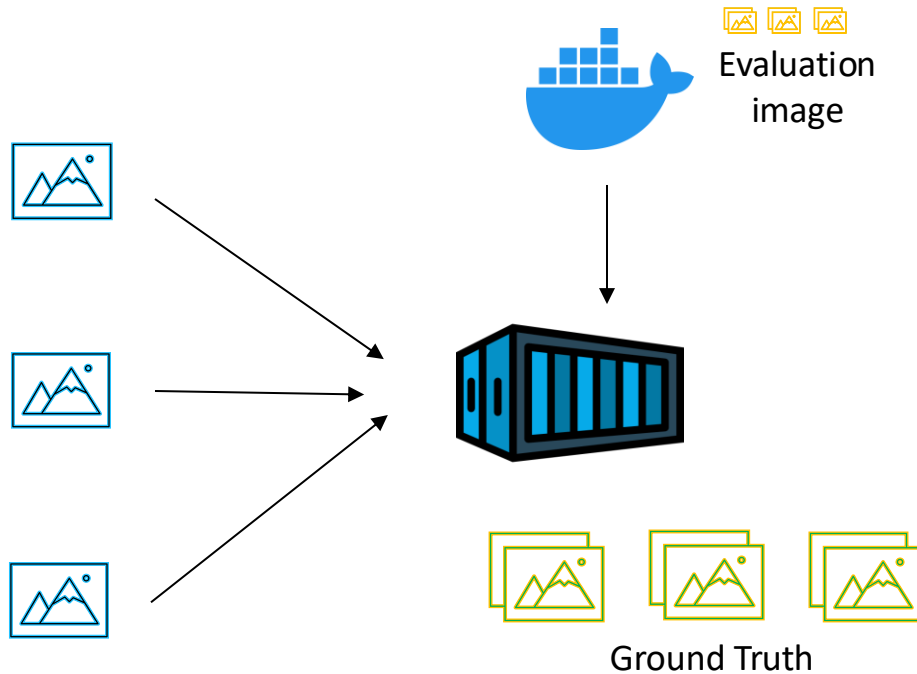
Evaluation – per submission



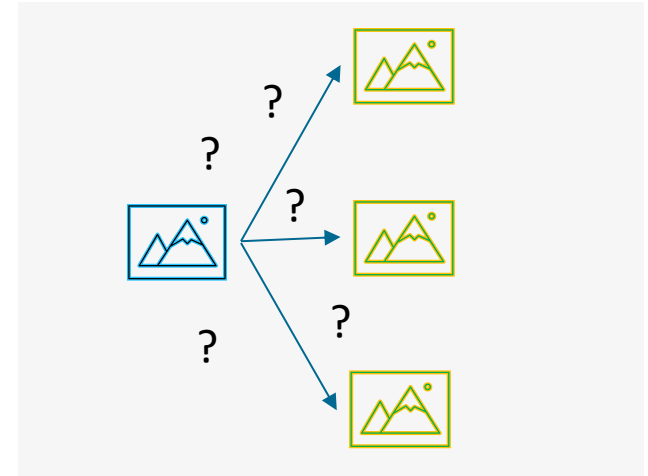
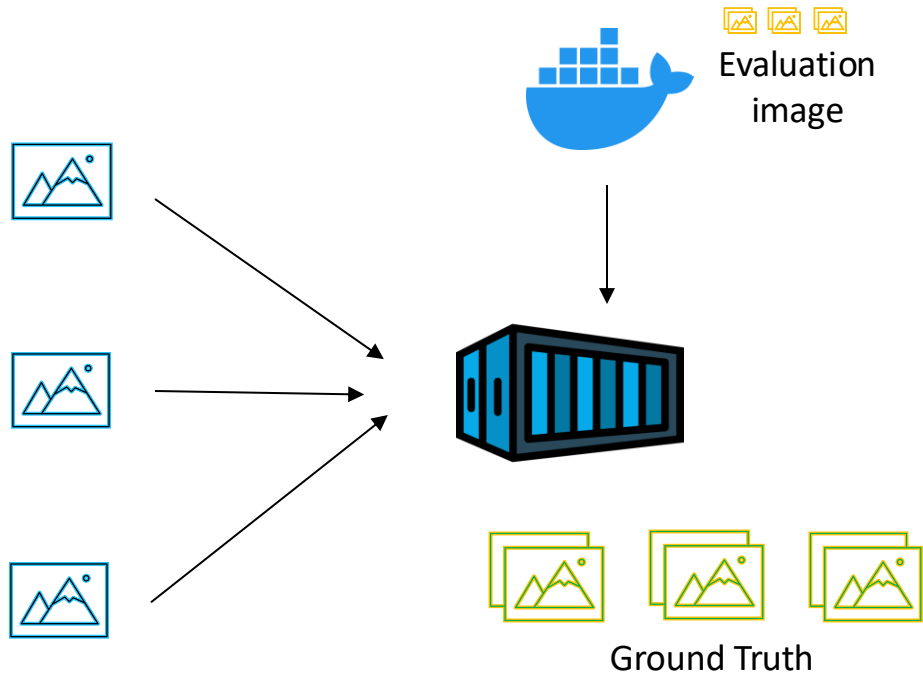
Evaluation – parses outputs as input



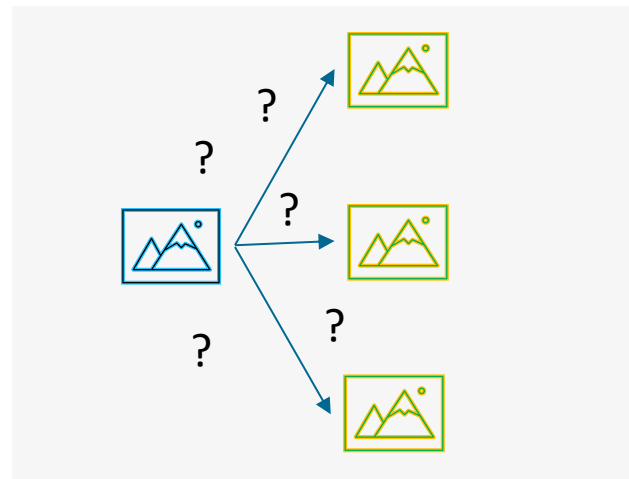
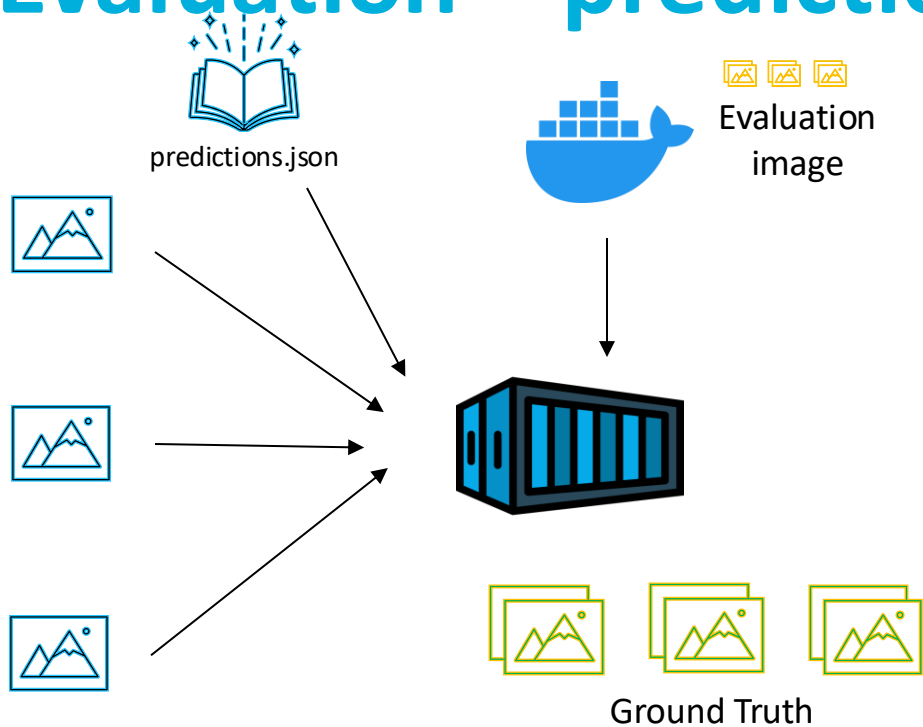
Evaluation – ground truth



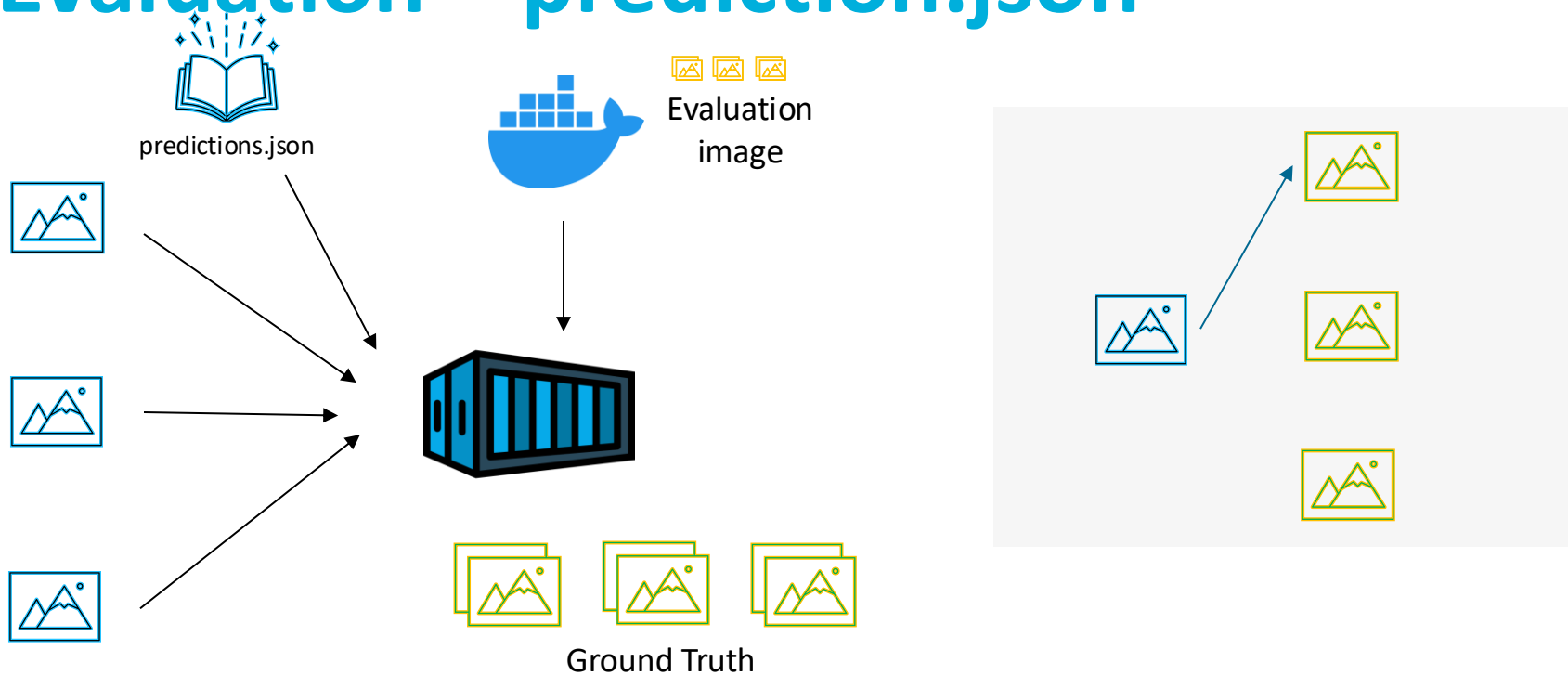
Evaluation – ground truth - matching



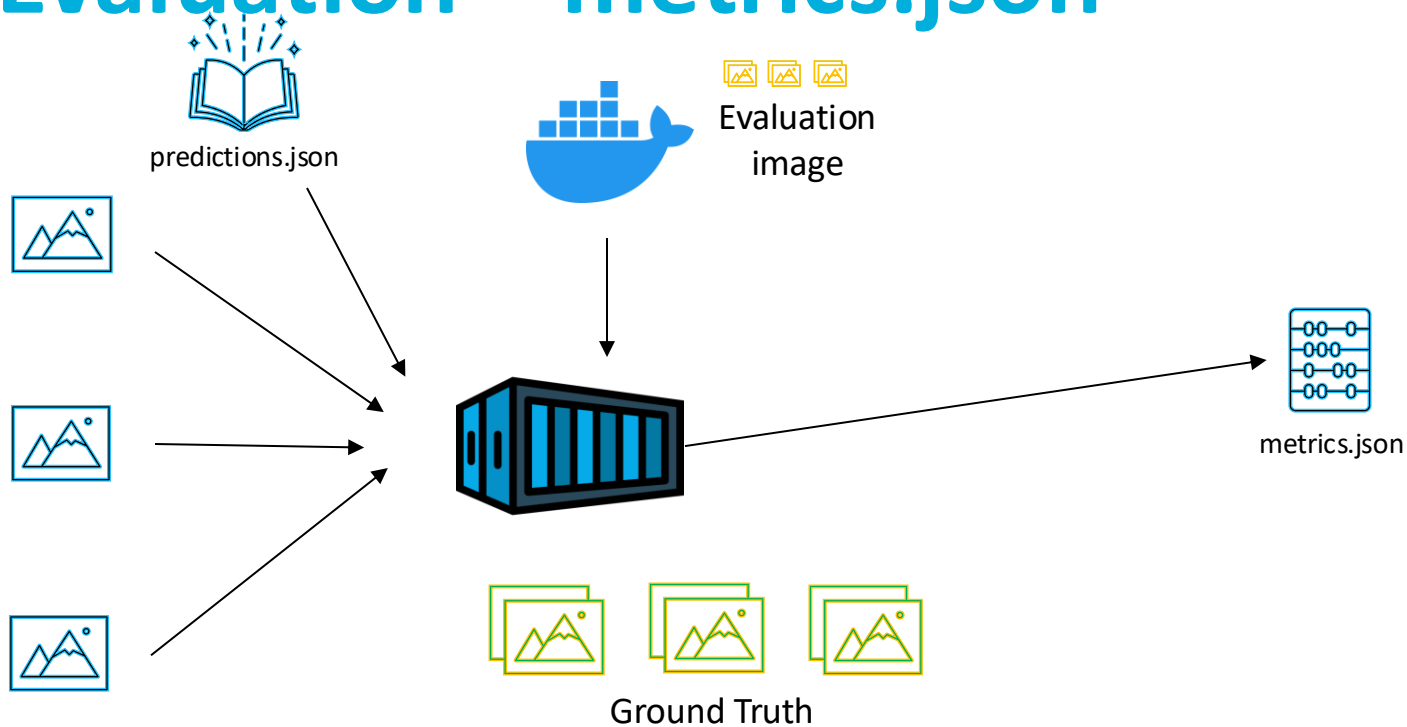
Evaluation – prediction.json



Evaluation – prediction.json



Evaluation – metrics.json



Evaluation – metrics.json

metrics.json

```
{  
  "score": 0.42  
}
```

metrics.json

```
{  
  "results": [  
    {  
      "my_metric": 0  
    },  
    {  
      "my_metric": 1  
    },  
    {  
      "my_metric": 0  
    }  
  ],  
  "aggregates": {  
    "my_metric": 0.33  
  }  
}
```

Evaluation – metrics.json - JSONPath

metrics.json

```
{  
  "score": 0.42  
}
```

JSONPath

Selecting and extracting specific data from JSON objects based on path expressions.

metrics.json

```
{  
  "results": [  
    {  
      "my_metric": 0  
    },  
    {  
      "my_metric": 1  
    },  
    {  
      "my_metric": 0  
    }  
  ],  
  "aggregates": {  
    "my_metric": 0.33  
  }  
}
```

Evaluation – metrics.json - JSONPath

metrics.json

```
{  
  "score": 0.42  
}
```

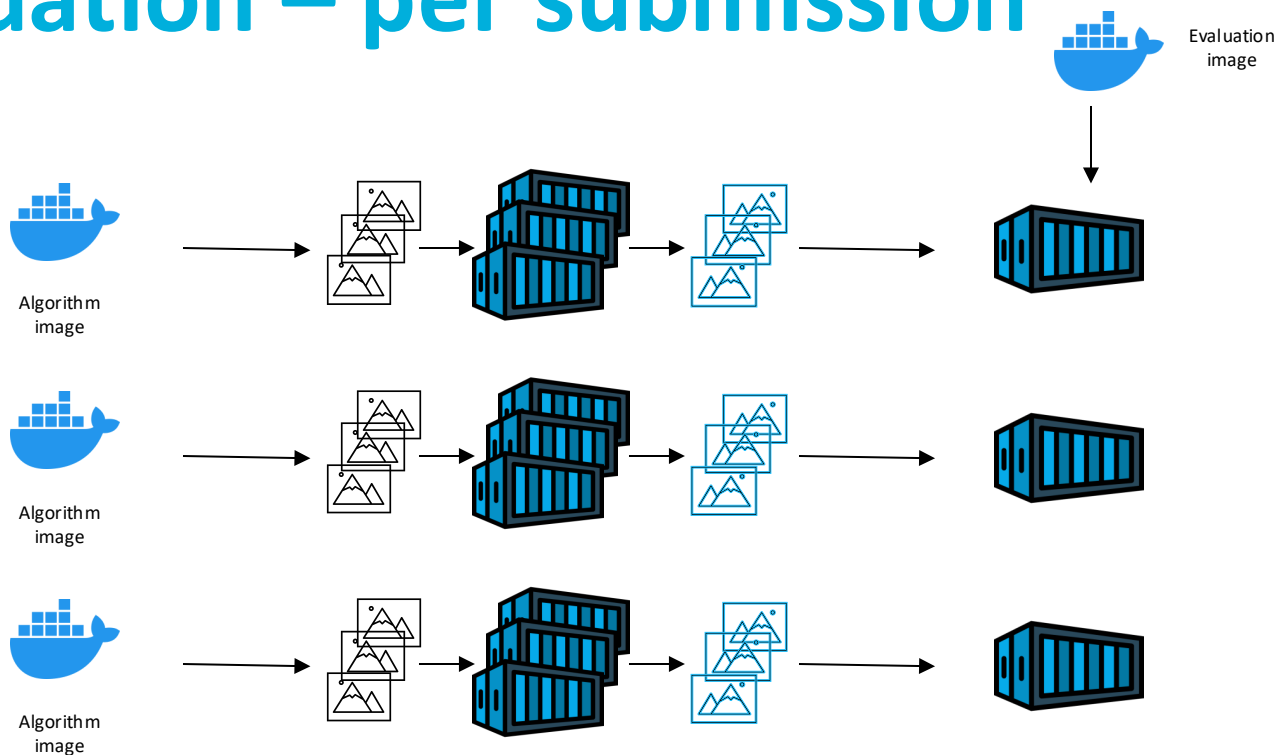
Example 1: `score`

Example 2: `aggregates.my_metric`

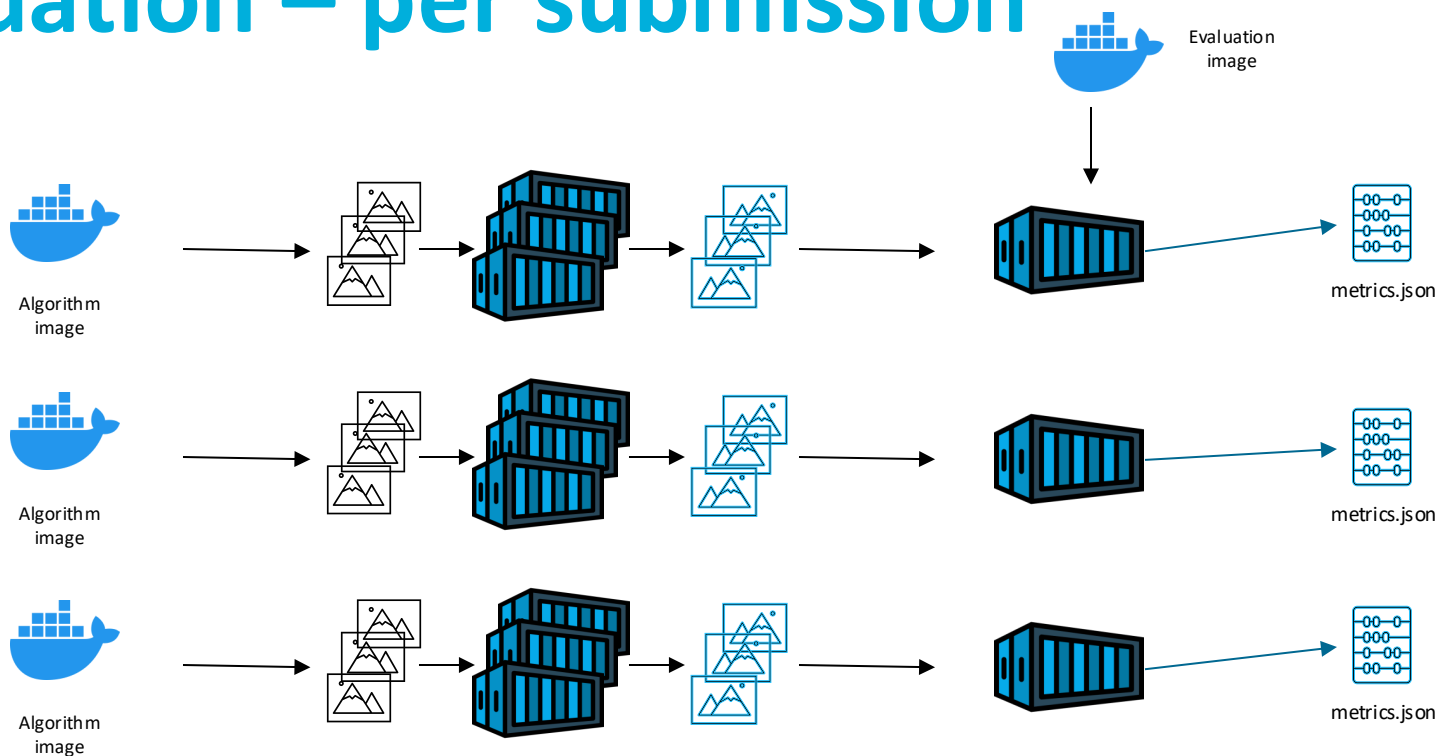
metrics.json

```
{  
  "results": [  
    {  
      "my_metric": 0  
    },  
    {  
      "my_metric": 1  
    },  
    {  
      "my_metric": 0  
    }  
  ],  
  "aggregates": {  
    "my_metric": 0.33  
  }  
}
```

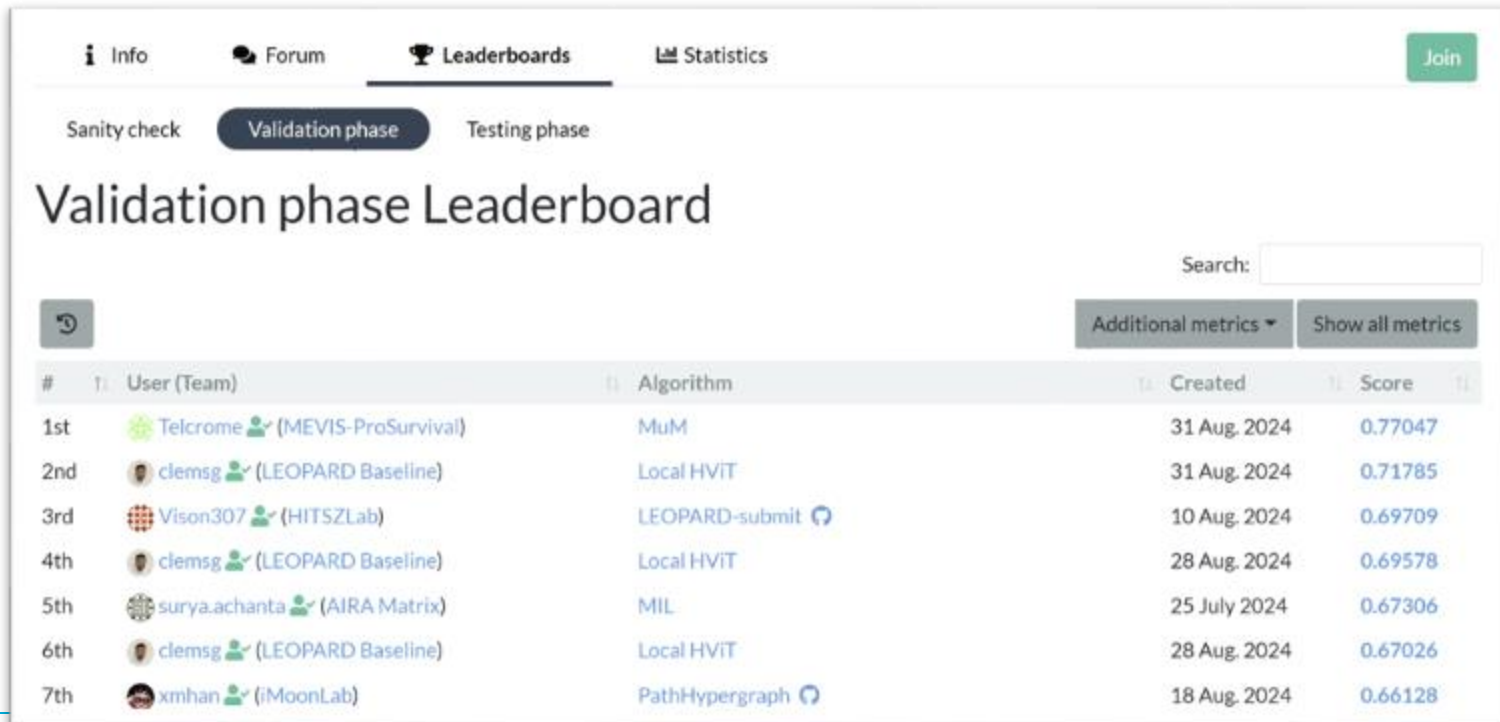
Evaluation – per submission



Evaluation – per submission





Evaluation – leaderboards LEOPARD





Validation phase Leaderboard				
Search: <input type="text"/>				
Additional metrics ▾ Show all metrics				
#	User (Team)	Algorithm	Created	Score
1st	Telcrome (MEVIS-ProSurvival)	MuM	31 Aug. 2024	0.77047
2nd	clemmsg (LEOPARD Baseline)	Local HVIT	31 Aug. 2024	0.71785
3rd	Vison307 (HITSZLab)	LEOPARD-submit	10 Aug. 2024	0.69709
4th	clemmsg (LEOPARD Baseline)	Local HVIT	28 Aug. 2024	0.69578
5th	surya.achanta (AIRA Matrix)	MIL	25 July 2024	0.67306
6th	clemmsg (LEOPARD Baseline)	Local HVIT	28 Aug. 2024	0.67026
7th	xmhan (iMoonLab)	PathHypergraph	18 Aug. 2024	0.66128

Evaluation – leaderboards



#	↑↓ User (Team)	↑↓ Algorithm	↑↓ Created	↑↓ Score	↑↓
1st	 Telcrome  (MEVIS-ProSurvival)	MuM	31 Aug. 2024	0.77047	

Evaluation – leaderboards

#	↑↓ User (Team)	↑↓ Algorithm	↑↓ Created	↑↓ Score
1st	 Telcrome  (MEVIS-ProSurvival)	MuM	31 Aug. 2024	0.77047

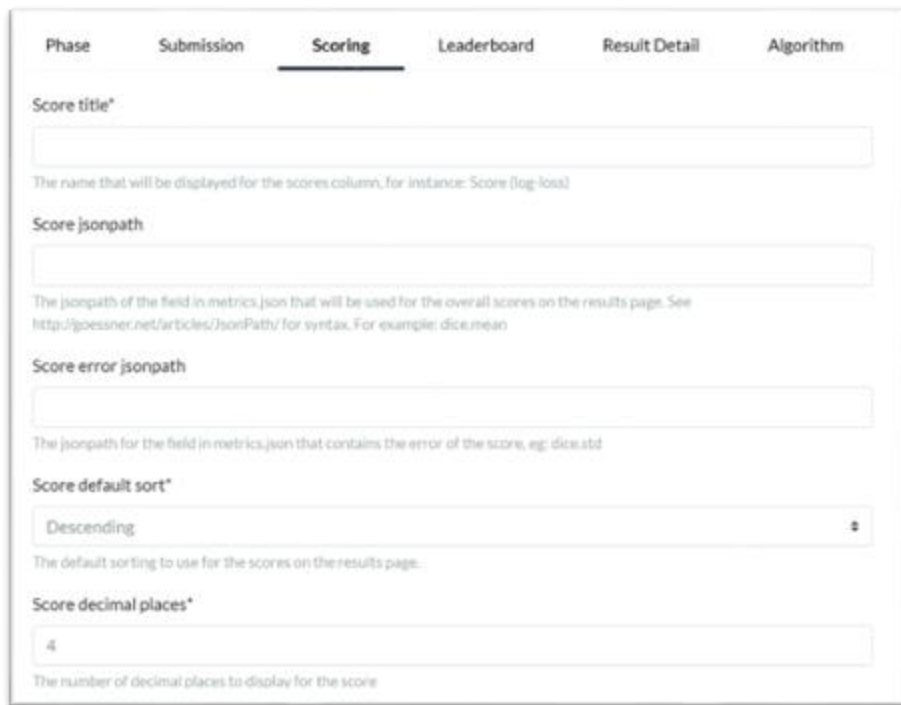
```
{
  "results": [
    {
      "case_id": "case_external_0261.tif",
      "case_id_gt_time": 15.5,
      "case_id_gt_event": 0,
      "case_id_prediction_years_to_recurrence": 9.717811584472656
    },
    {
      "case_id": "case_radboud_0649.tif",
      "case_id_gt_time": 12.22450376,
      "case_id_gt_event": 0,
      "case_id_prediction_years_to_recurrence": 14.979281425476074
    },
    ...
  ],
  "aggregates": {
    "c_index": 0.7704741053903066,
    "c_index_radboud": 0.7783783783783784,
    "c_index_external": 0.7625698324022346
  }
}
```

Evaluation – leaderboards

#	↑↓ User (Team)	↑↓ Algorithm	↑↓ Created	↑↓ Score
1st	 Telcrome  (MEVIS-ProSurvival)	MuM	31 Aug. 2024	0.77047

```
{
  "results": [
    {
      "case_id": "case_external_0261.tif",
      "case_id_gt_time": 15.5,
      "case_id_gt_event": 0,
      "case_id_prediction_years_to_recurrence": 9.717811584472656
    },
    {
      "case_id": "case_radboud_0649.tif",
      "case_id_gt_time": 12.22450376,
      "case_id_gt_event": 0,
      "case_id_prediction_years_to_recurrence": 14.979281425476074
    },
    ...
  ],
  "aggregates": {
    "c_index": 0.7704741053903066,
    "c_index_radboud": 0.7783783783783784,
    "c_index_external": 0.7625698324022346
  }
}
```

Evaluation – leaderboard - settings



The screenshot shows a configuration interface for a leaderboard, specifically the 'Scoring' tab. It contains several input fields and a dropdown menu, each with a label and a descriptive subtitle. The 'Score title' field is empty. The 'Score jsonpath' field is empty, with a subtitle providing a link to a JSONPath tutorial. The 'Score error jsonpath' field is empty. The 'Score default sort' dropdown is set to 'Descending'. The 'Score decimal places' field is set to '4'.

Phase	Submission	Scoring	Leaderboard	Result Detail	Algorithm
Score title*					
<input type="text"/>					
The name that will be displayed for the scores column, for instance: Score (log-loss)					
Score jsonpath					
<input type="text"/>					
The jsonpath of the field in metrics.json that will be used for the overall scores on the results page. See http://goessner.net/articles/JsonPath/ for syntax. For example: dice.mean					
Score error jsonpath					
<input type="text"/>					
The jsonpath for the field in metrics.json that contains the error of the score, eg: dice.std					
Score default sort*					
<div>Descending</div>					
The default sorting to use for the scores on the results page.					
Score decimal places*					
<div>4</div>					
The number of decimal places to display for the score					

- Score title
- Score JSONPath
- Score error JSONPath
- Score default sort
- Score decimal places

Evaluation – leaderboard - settings

```
{
  "results": [
    {
      "case_id": "case_external_0261.tif",
      "case_id_gt_time": 15.5,
      "case_id_gt_event": 0,
      "case_id_prediction_years_to_recurrence": 9.717811584472656
    },
    {
      "case_id": "case_radboud_0649.tif",
      "case_id_gt_time": 12.22450376,
      "case_id_gt_event": 0,
      "case_id_prediction_years_to_recurrence": 14.979281425476074
    },
    ...
  ],
  "aggregates": {
    "c_index": 0.7704741053903066,
    "c_index_radboud": 0.7783783783783784,
    "c_index_external": 0.7625698324022346
  }
}
```

Score title*

The name that will be displayed for the scores column, for instance: Score (log-loss)

Evaluation – leaderboard - settings

```
{
  "results": [
    {
      "case_id": "case_external_0261.tif",
      "case_id_gt_time": 15.5,
      "case_id_gt_event": 0,
      "case_id_prediction_years_to_recurrence": 9.717811584472656
    },
    {
      "case_id": "case_radboud_0649.tif",
      "case_id_gt_time": 12.22450376,
      "case_id_gt_event": 0,
      "case_id_prediction_years_to_recurrence": 14.979281425476074
    },
    ...
  ],
  "aggregates": {
    "c_index": 0.7704741053903066,
    "c_index_radboud": 0.7783763763763764,
    "c_index_external": 0.7625698324022346
  }
}
```

Score title*

The name that will be displayed for the scores column, for instance: Score (log-loss)

Score jsonpath

The jsonpath of the field in metrics.json that will be used for the overall scores on the results page. See

Evaluation – leaderboard - settings

```
{
  "results": [
    {
      "case_id": "case_external_0261.tif",
      "case_id_gt_time": 15.5,
      "case_id_gt_event": 0,
      "case_id_prediction_years_to_recurrence": 9.717811584472656
    },
    {
      "case_id": "case_radboud_0649.tif",
      "case_id_gt_time": 12.22450376,
      "case_id_gt_event": 0,
      "case_id_prediction_years_to_recurrence": 14.979281425476074
    },
    ...
  ],
  "aggregates": {
    "c_index": 0.7704741053903066,
    "c_index_radboud": 0.7783763763763764,
    "c_index_external": 0.7625698324022346
  }
}
```

Score title*

The name that will be displayed for the scores column, for instance: Score (log-loss)

Score jsonpath

The jsonpath of the field in metrics.json that will be used for the overall scores on the results page. See <http://goessner.net/articles/JsonPath/> for syntax. For example: dice.mean

Score error jsonpath

The jsonpath for the field in metrics.json that contains the error of the score, eg: dice.std

Evaluation – leaderboard - settings

```
{
  "results": [
    {
      "case_id": "case_external_0261.tif",
      "case_id_gt_time": 15.5,
      "case_id_gt_event": 0,
      "case_id_prediction_years_to_recurrence": 9.717811584472656
    },
    {
      "case_id": "case_radboud_0649.tif",
      "case_id_gt_time": 12.22450376,
      "case_id_gt_event": 0,
      "case_id_prediction_years_to_recurrence": 14.979281425476074
    },
    ...
  ],
  "aggregates": {
    "c_index": 0.7704741053903066,
    "c_index_radboud": 0.7783763763763764,
    "c_index_external": 0.7625698324022346
  }
}
```

Score title*

The name that will be displayed for the scores column, for instance: Score (log-loss)

Score jsonpath

The jsonpath of the field in metrics.json that will be used for the overall scores on the results page. See <http://goessner.net/articles/JsonPath/> for syntax. For example: dice.mean

Score error jsonpath

The jsonpath for the field in metrics.json that contains the error of the score, eg: dice.std

Score default sort*

The default sorting to use for the scores on the results page.

Evaluation – leaderboard - settings

```
{
  "results": [
    {
      "case_id": "case_external_0261.tif",
      "case_id_gt_time": 15.5,
      "case_id_gt_event": 0,
      "case_id_prediction_years_to_recurrence": 9.717811584472656
    },
    {
      "case_id": "case_radboud_0649.tif",
      "case_id_gt_time": 12.22450376,
      "case_id_gt_event": 0,
      "case_id_prediction_years_to_recurrence": 14.979281425476074
    },
    ...
  ],
  "aggregates": {
    "c_index": 0.7704741053903066,
    "c_index_radboud": 0.7783763763763764,
    "c_index_external": 0.7625698324022346
  }
}
```

Score title*

The name that will be displayed for the scores column, for instance: Score (log-loss)

Score jsonpath

The jsonpath of the field in metrics.json that will be used for the overall scores on the results page. See <http://goessner.net/articles/JsonPath/> for syntax. For example: dice.mean

Score error jsonpath

The jsonpath for the field in metrics.json that contains the error of the score, eg: dice.std

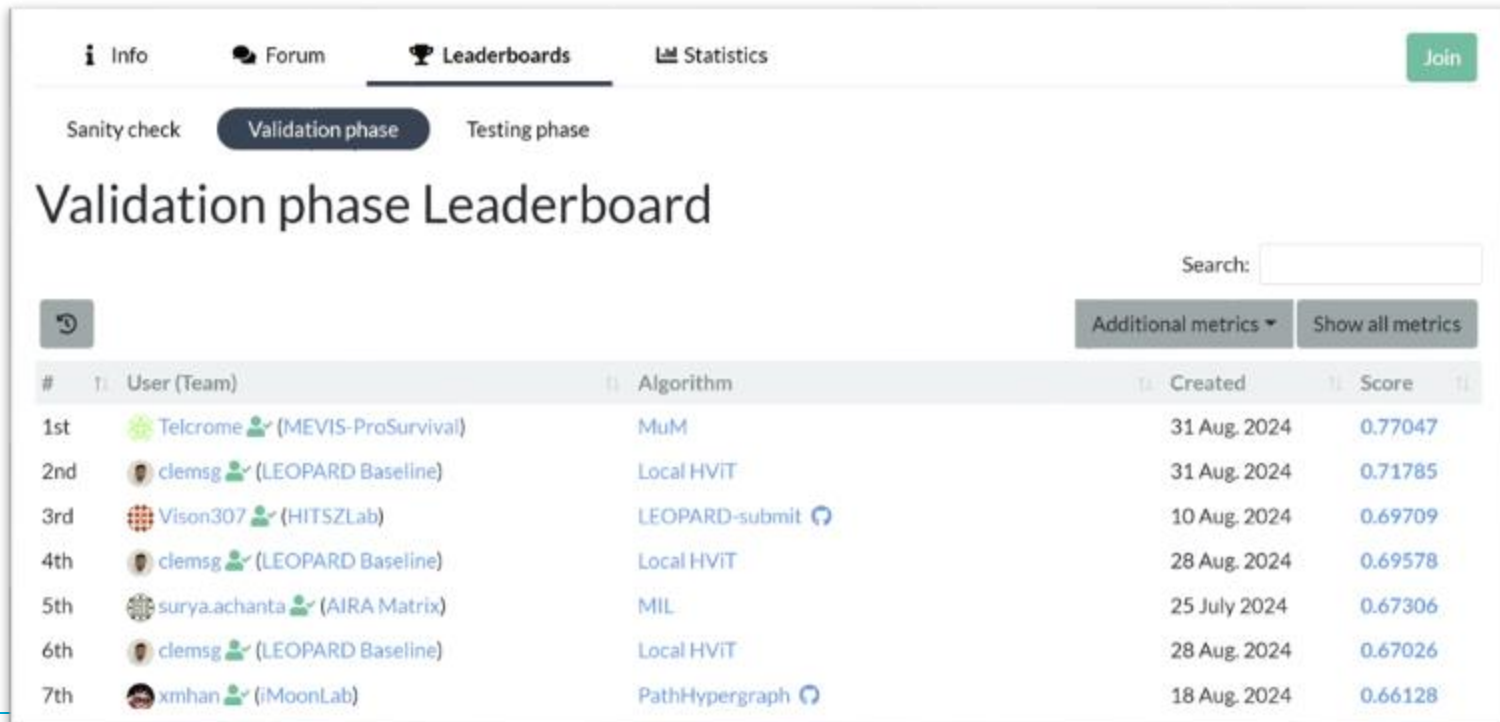
Score default sort*

The default sorting to use for the scores on the results page.

Score decimal places*

The number of decimal places to display for the score

Evaluation – leaderboards

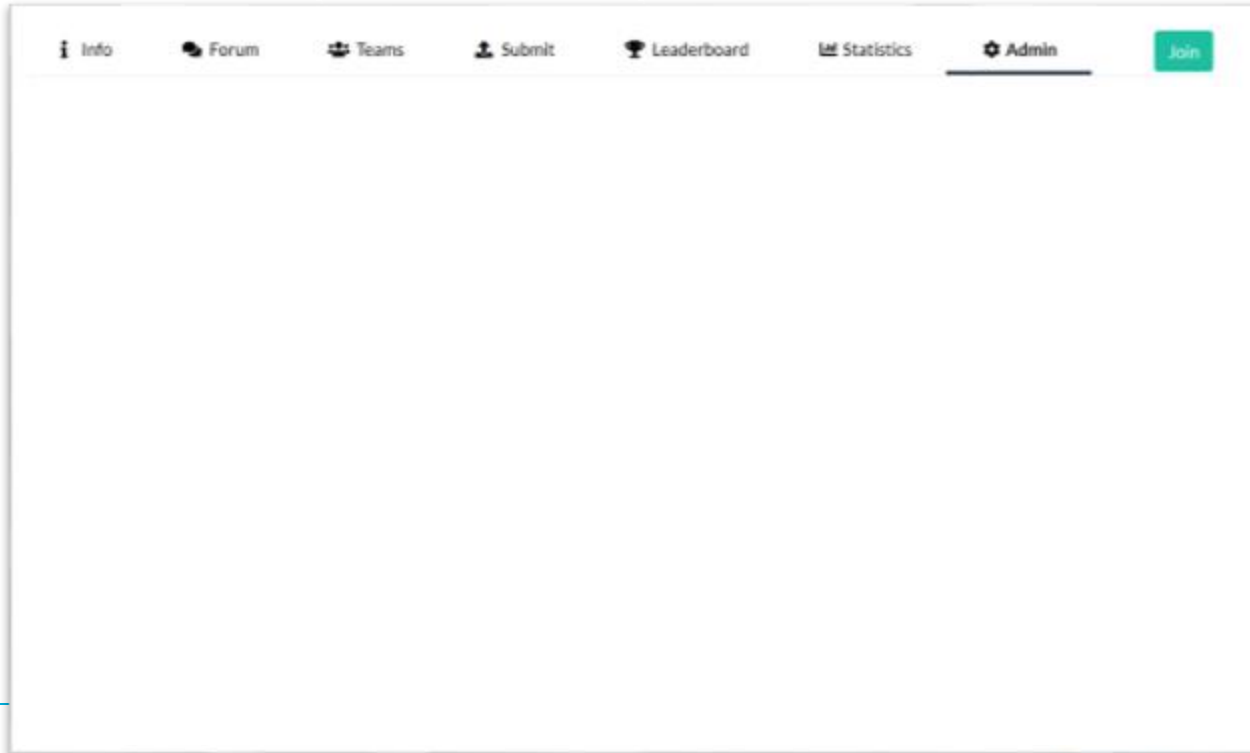


Validation phase Leaderboard				
Search: <input type="text"/>				
Additional metrics ▾ Show all metrics				
#	User (Team)	Algorithm	Created	Score
1st	Telcrome (MEVIS-ProSurvival)	MuM	31 Aug. 2024	0.77047
2nd	clemmsg (LEOPARD Baseline)	Local HVIT	31 Aug. 2024	0.71785
3rd	Vison307 (HITSZLab)	LEOPARD-submit	10 Aug. 2024	0.69709
4th	clemmsg (LEOPARD Baseline)	Local HVIT	28 Aug. 2024	0.69578
5th	surya.achanta (AIRA Matrix)	MIL	25 July 2024	0.67306
6th	clemmsg (LEOPARD Baseline)	Local HVIT	28 Aug. 2024	0.67026
7th	xmhan (iMoonLab)	PathHypergraph	18 Aug. 2024	0.66128

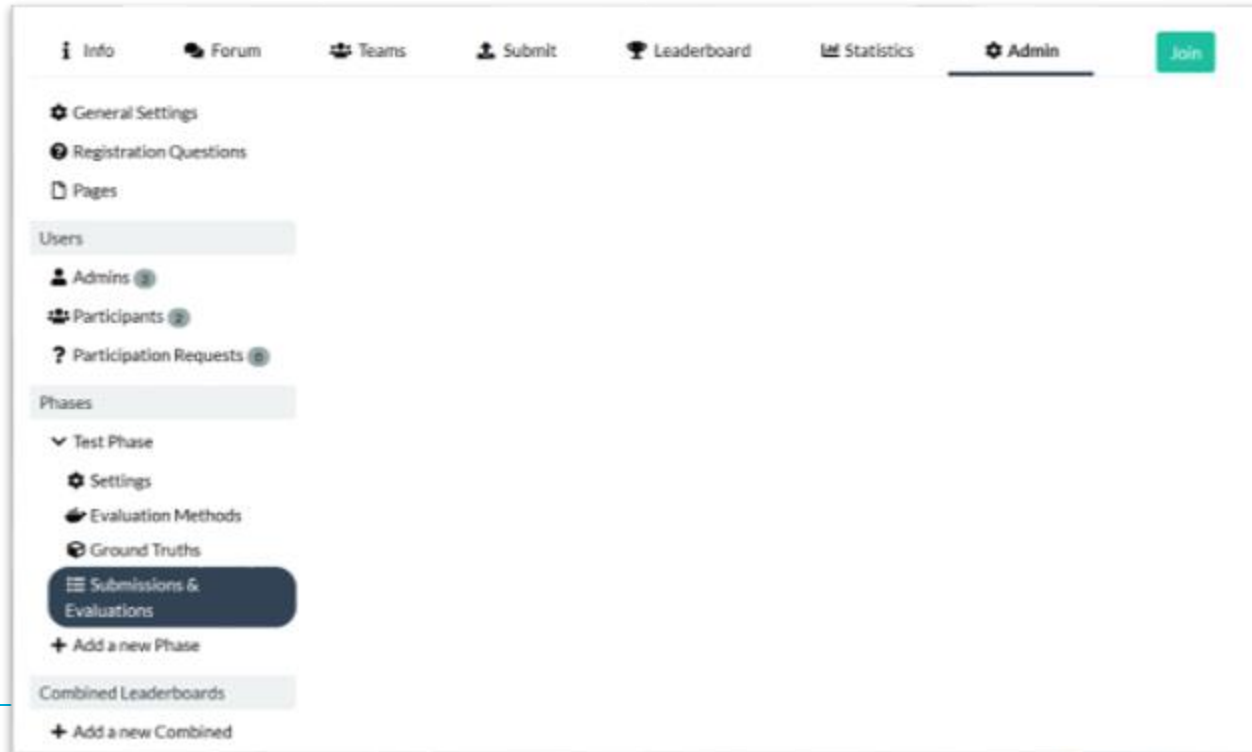
Challenge Admin – keeping track

- Keeping track of submissions and their evaluations
- Challenge Admin: **Submissions & Evaluations**

Challenge Admin - Submissions



Challenge Admin - Submissions



Evaluations - details

Submissions and Evaluations for Test Phase

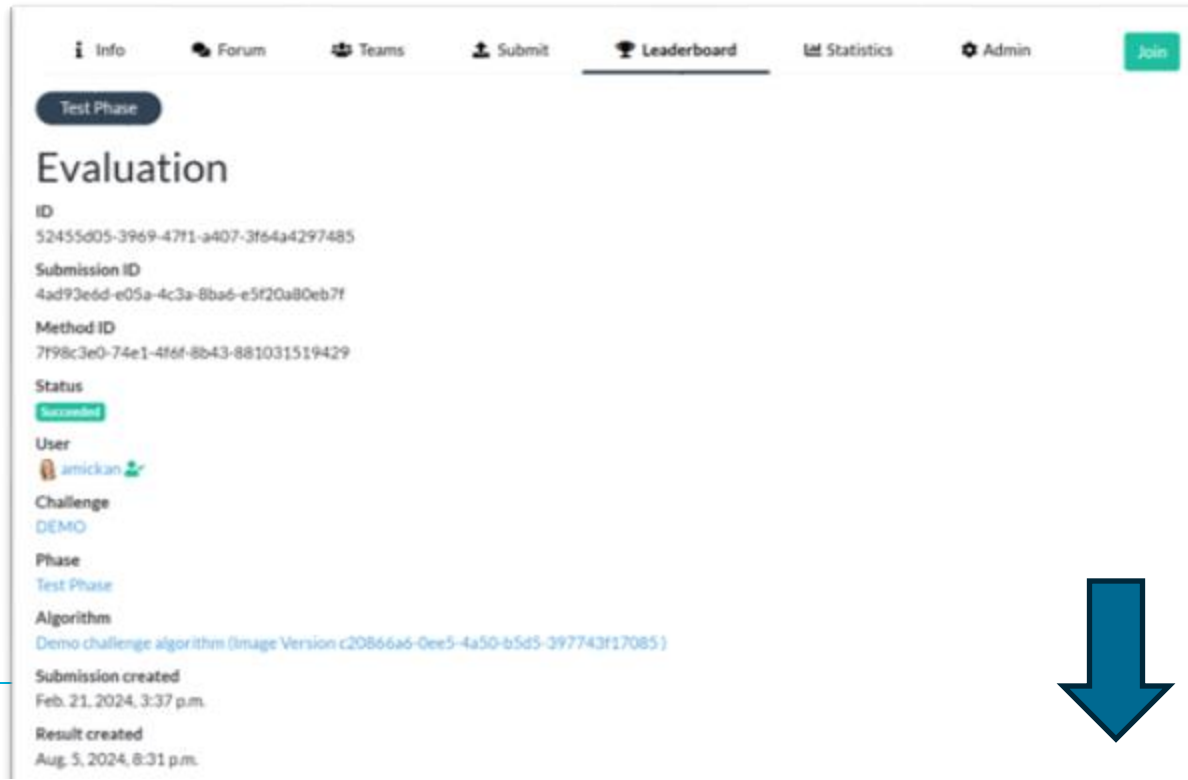
Show 30 entries Search:

Submission ID	Evaluation ID	Evaluation Created	User	Status	Hide/Publish	Algorithm Results
4ad93e6d	52455d05	Aug. 5, 2024, 8:31 p.m.	amickan	Successful	Hide Result	Demo challenge algorithm
4ad93e6d	f92b92bf	July 30, 2024, 10:38 a.m.	amickan	Failed		Demo challenge algorithm
4ad93e6d	3a90e2f9	July 30, 2024, 10:11 a.m.	amickan	Successful	Hide Result	Demo challenge algorithm
4ad93e6d	5a3a3e26	Feb. 21, 2024, 3:37 p.m.	amickan	Failed		Demo challenge algorithm
e82c08f7	1cd6515b	Feb. 21, 2024, 2:39 p.m.	amickan	Failed		Demo challenge algorithm
5950edfe	6bf66783	Feb. 21, 2024, 12:09 p.m.	amickan	Failed		Demo challenge algorithm
50a3e29b	8b2479ff	Feb. 21, 2024, 10:42 a.m.	amickan	Failed		Demo Vessel Segmentation

Evaluations - details

Submission ID	Evaluation ID	Evaluation Created	User	Status	Hide/Publish	Algorithm Results
4ad93e6d	52455d05	Aug. 5, 2024, 8:31 p.m.	amickan	Successful	Hide Result	Demo challenge algorithm
4ad93e6d	f92b92bf	July 30, 2024, 10:38 a.m.	amickan	Failed		Demo challenge algorithm
4ad93e6d	3a90e2f9	July 30, 2024, 10:11 a.m.	amickan	Successful	Hide Result	Demo challenge algorithm
4ad93e6d	5a3a3e26	Feb. 21, 2024, 3:37 p.m.	amickan	Failed		Demo challenge algorithm
e82c08f7	1cd6515b	Feb. 21, 2024, 2:39 p.m.	amickan	Failed		Demo challenge algorithm
5950edfe	6bf66783	Feb. 21, 2024, 12:09 p.m.	amickan	Failed		Demo challenge algorithm
50a3e29b	8b2479ff	Feb. 21, 2024, 10:42 a.m.	amickan	Failed		Demo Vessel Segmentation

Evaluations - details



Info Forum Teams Submit **Leaderboard** Statistics Admin [Join](#)

Test Phase


Evaluation

ID
52455d05-3969-47f1-a407-3f64a4297485

Submission ID
4ad93e6d-e05a-4c3a-8ba6-e5f20a80eb7f

Method ID
7f98c3e0-74e1-4f6f-8b43-881031519429

Status
Successful

User
 amickan

Challenge
DEMO

Phase
Test Phase

Algorithm
Demo challenge algorithm (Image Version c20866a6-0ee5-4a50-b5d5-397743f17085)

Submission created
Feb. 21, 2024, 3:37 p.m.

Result created
Aug. 5, 2024, 8:31 p.m.

dboudumc

Metrics

```
{
  "results": [
    {
      "my_metric": 1
    }
  ],
  "aggregates": {
    "my_metric": 1
  }
}
```

Evaluation Admin

Contact User

[Message User](#)

Predictions

[Download the predictions.json file for this evaluation](#)

[Go to the results of Demo challenge algorithm](#)

Visibility

[This result is published on the leaderboard\(s\)](#)

[Exclude this result from the leaderboard\(s\)](#)

Logs

Runtime Metrics



Metrics

```
{
  "results": [
    {
      "my_metric": 1
    }
  ],
  "aggregates": {
    "my_metric": 1
  }
}
```

Evaluation Admin

Contact User

[Message User](#)

Predictions

[Download the predictions.json file for this evaluation](#)

[Go to the results of Demo challenge algorithm](#)

Visibility

[This result is published on the leaderboard\(s\)](#)

[Exclude this result from the leaderboard\(s\)](#)

Logs

Runtime Metrics



Metrics

```
{
  "results": [
    {
      "my_metric": 1
    }
  ],
  "aggregates": {
    "my_metric": 1
  }
}
```

Evaluation Admin

Contact User

Message User

Predictions

Download the predictions.json file for this evaluation

Go to the results of Demo challenge algorithm

Visibility

This result is published on the leaderboard(s)

Exclude this result from the leaderboard(s)

Logs

Runtime Metrics



Metrics

```
{
  "results": [
    {
      "my_metric": 1
    }
  ],
  "aggregates": {
    "my_metric": 1
  }
}
```

Evaluation Admin

Contact User

[Message User](#)

Predictions

[Download the predictions.json file for this evaluation](#)

[Go to the results of Demo challenge algorithm](#)

Visibility

[This result is published on the leaderboard\(s\)](#)

[Exclude this result from the leaderboard\(s\)](#)

Logs

Runtime Metrics



Metrics

```
{
  "results": [
    {
      "my_metric": 1
    }
  ],
  "aggregates": {
    "my_metric": 1
  }
}
```

Evaluation Admin

Contact User

[Message User](#)

Predictions

[Download the predictions.json file for this evaluation](#)

[Go to the results of Demo challenge algorithm](#)

Visibility

[This result is published on the leaderboard\(s\)](#)

[Exclude this result from the leaderboard\(s\)](#)

Logs

Runtime Metrics



Metrics

```
{
  "results": [
    {
      "my_metric": 1
    }
  ],
  "aggregates": {
    "my_metric": 1
  }
}
```

Evaluation Admin

Contact User

[Message User](#)

Predictions

[Download the predictions.json file for this evaluation](#)

[Go to the results of Demo challenge algorithm](#)

Visibility

[This result is published on the leaderboard\(s\)](#)

[Exclude this result from the leaderboard\(s\)](#)

Logs

Runtime Metrics



Evaluations - details

Submission ID	Evaluation ID	Evaluation Created	User	Status	Hide/Publish	Algorithm Results
4ad93e6d	52455d05	Aug. 5, 2024, 8:31 p.m.	amickan	Successful	Hide Result	Demo challenge algorithm
4ad93e6d	f92b92bf	July 30, 2024, 10:38 a.m.	amickan	Failed		Demo challenge algorithm
4ad93e6d	3a90e2f9	July 30, 2024, 10:11 a.m.	amickan	Successful	Hide Result	Demo challenge algorithm
4ad93e6d	5a3a3e26	Feb. 21, 2024, 3:37 p.m.	amickan	Failed		Demo challenge algorithm
e82c08f7	1cd6515b	Feb. 21, 2024, 2:39 p.m.	amickan	Failed		Demo challenge algorithm
5950edfe	6bf66783	Feb. 21, 2024, 12:09 p.m.	amickan	Failed		Demo challenge algorithm
50a3e29b	8b2479ff	Feb. 21, 2024, 10:42 a.m.	amickan	Failed		Demo Vessel Segmentation

Evaluations – Algorithm Image Errors

Evaluation Admin

Contact User

[Message User](#)

Prerequisite Jobs

ID	Created	Status
6818386a-9da0-4fc6-80a8-8528fcbdf102	Feb. 21, 2024, 2:39 p.m.	Failed

Predictions

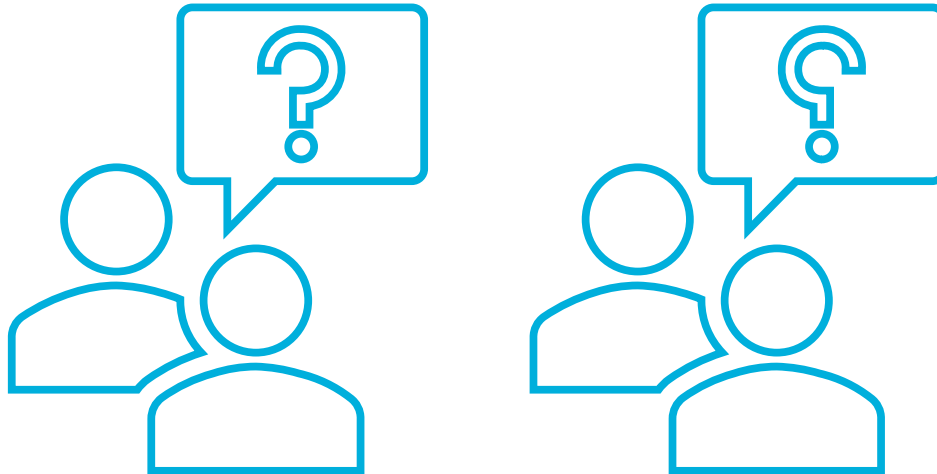
[» Go to the results of Demo challenge algorithm](#)

Logs

Stdout

No logs found on stdout

Questions & Answers



Hands-on #3: Evaluations and Leaderboards

1. Uploading of an evaluation method
 2. Submitting an algorithm
 3. Debug submissions / evaluations
 4. Configure leaderboard
- Instruction page on workshop2024.grand-challenge.org

Agenda

Time	Topic
10:00 - 10:30	Welcome and introduction to challenges
10:30 - 11:00	Overview over the GC challenge feature
11:00 - 11:05	Short Break
11:05 - 11:45	Deep dive #1: uploading and managing hidden test data
11:45 - 13:00	Lunch Break
13:00 - 14:15	Deep dive #2: algorithm containers
14:15 - 14:30	Short Break
14:30 - 16:00	Deep dive #3: custom evaluation methods & leaderboard set-up
16:00 - 17:00	Wrap-up , Q&A