Unraveling uncertainty in benchmarking: Methods and open-source toolkit for analyzing and visualizing challenge results

Manuel Wiesenfarth (German Cancer Research Center)

Joint work with
Annette Kopp-Schneider, Lena Maier-Hein & Annika Reinke

Data science seminar, Heidelberg, 11/06/2019

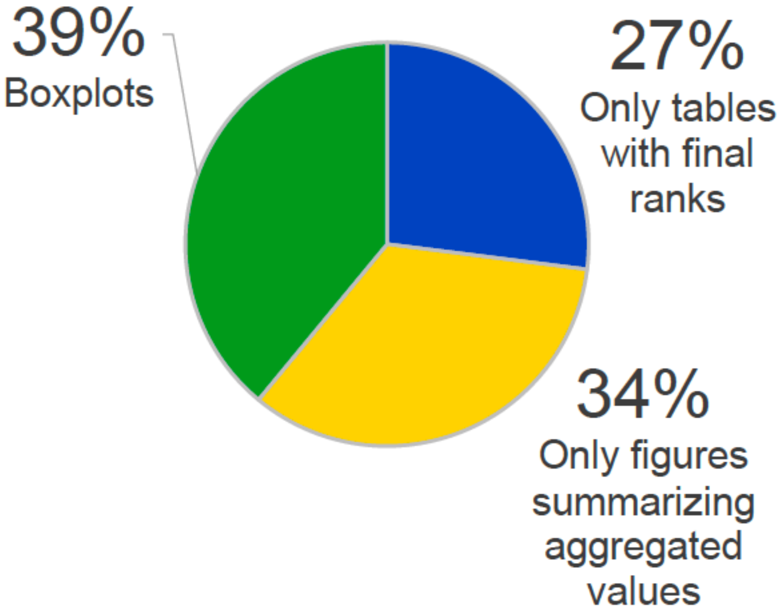
Background

- *Grand challenges* as standard for validation of biomedical image analysis methods in a comparative manner
- Challenges compared to clinical trials
- Lack of common standards in design, analysis and reporting (Maier-Hein et al. *Nature Commun.* 2018)

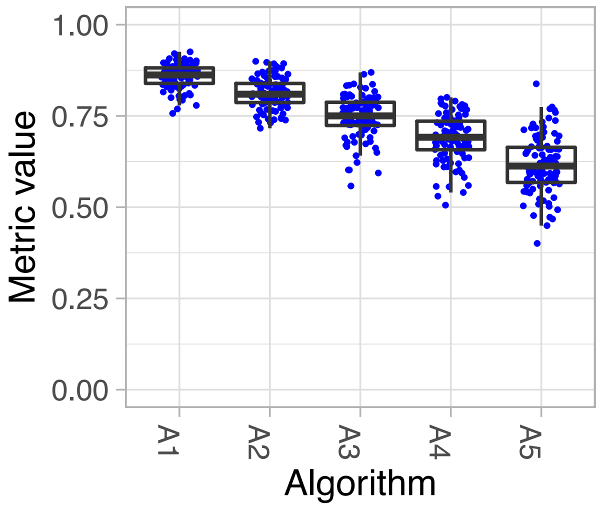
m.wiesenfarth@dkfz.de

Common presentation of results

In 83 challenges analyzed in Maier-Hein et al. Nature Commun. 2018:

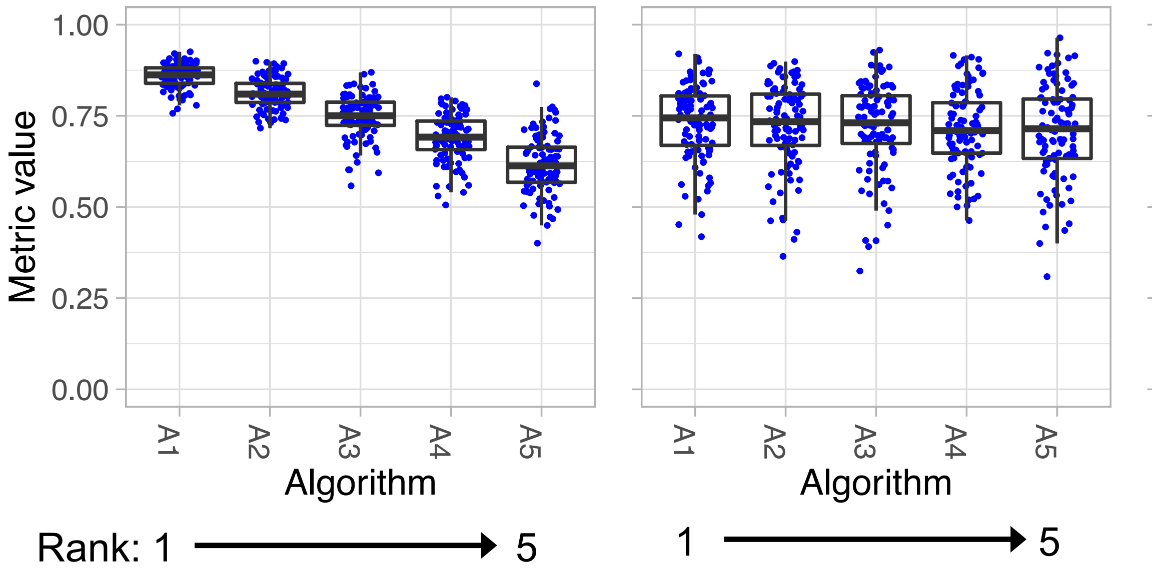


Motivation: Why ranking lists are not enough

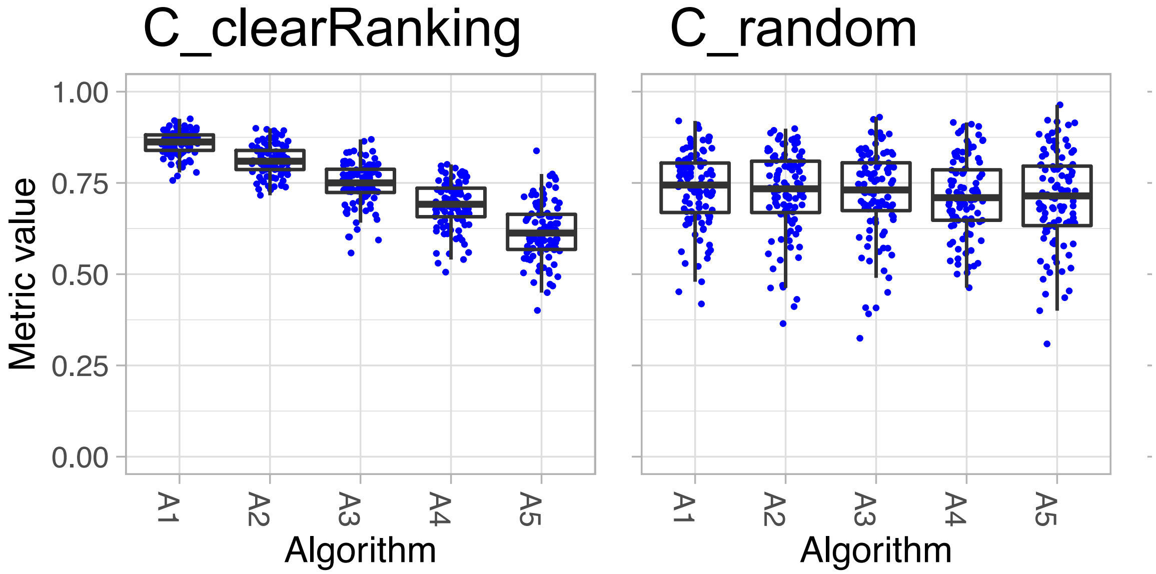


Rank: 1 → 5

Motivation: Why ranking lists are not enough



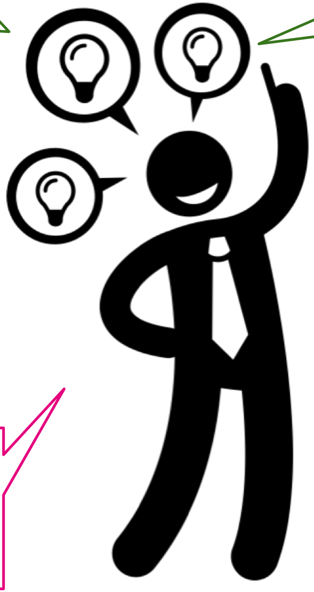
Motivation: Why ranking lists are not enough



The ranking in a challenge may be affected by …

How many test cases are available?

What metric(s) is/are used for performance assessment?



How heterogenous are metric values?

What ranking method is used?

…

Icon created by the Noun Project

m.wiesenfarth@dkfz.de

Contribution

To help challenge organizers and participants gain further insights into algorithms‘ perfomances, we ...

- propose methodology for visualizing results of challenges
- provide an open-source analysis and visualization toolkit

m.wiesenfarth@dkfz.de

Assessment data for a challenge analysis

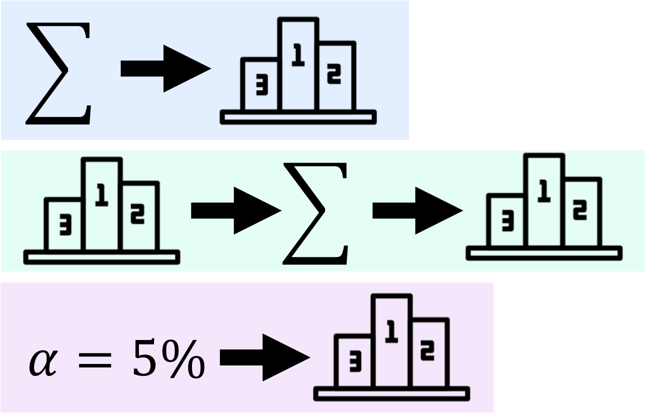
Testcase_ID	Algorithm_name	Metric_value	Task_name
85	A1	0.7952	C_random
15	A4	0.6877	C_clearRanking
81	A3	0.7754	C_random
8	A5	0.6948	C_random
82	A2	0.8576	C_clearRanking
19	A2	0.5556	C_random
84	A1	0.5215	C_random
⋮	⋮	⋮	⋮

m.wiesenfarth@dkfz.de

Ranking methods

Common methods:

- Aggregate-then-rank
- Rank-then-aggregate
- Test based procedures



Different strategies to obtain ranking may lead to different winners

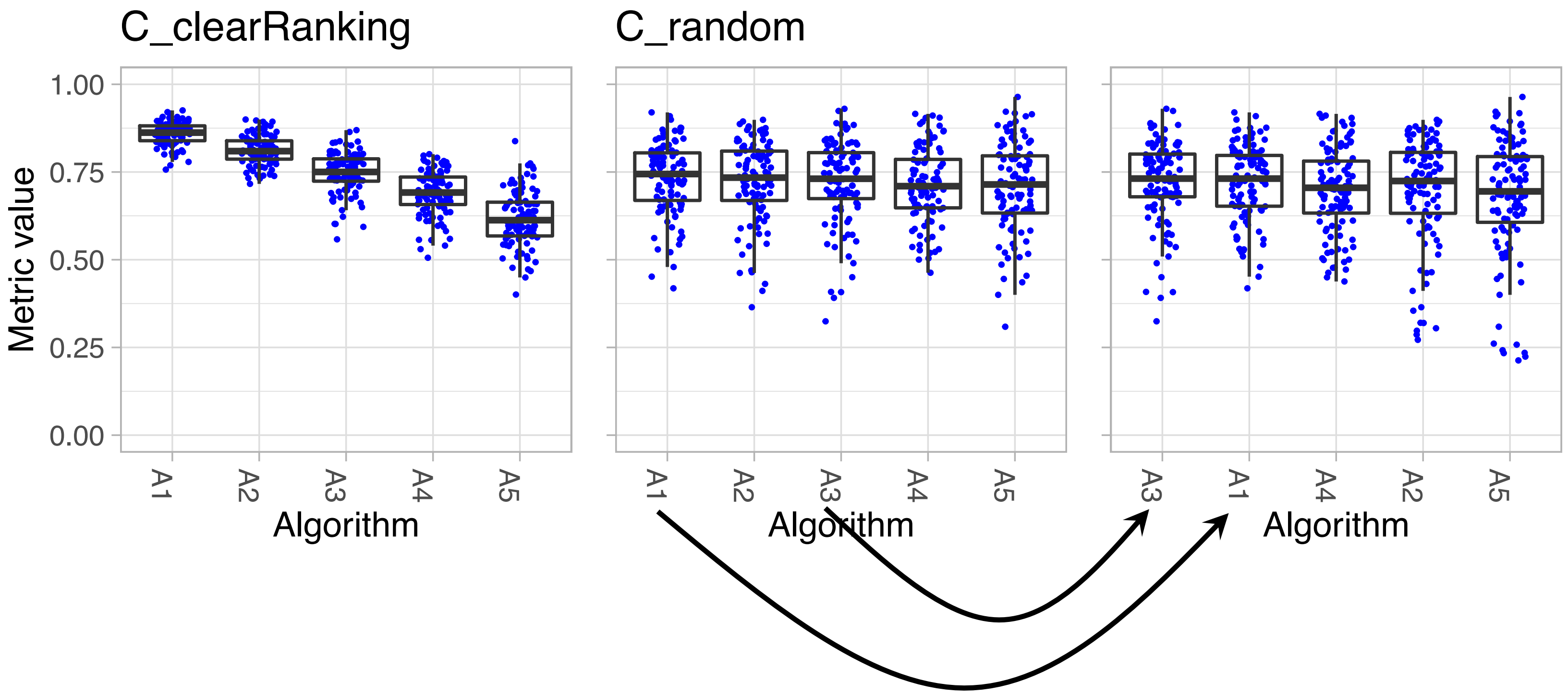
Maier-Hein et al. **Nature Commun. 2018**; Reinke et al. **MICCAI 2018; Wiesenfarth et al., ArXiv 2019**

m.wiesenfarth@dkfz.de

Visualization of raw assessment data: Understand distribution of metric values

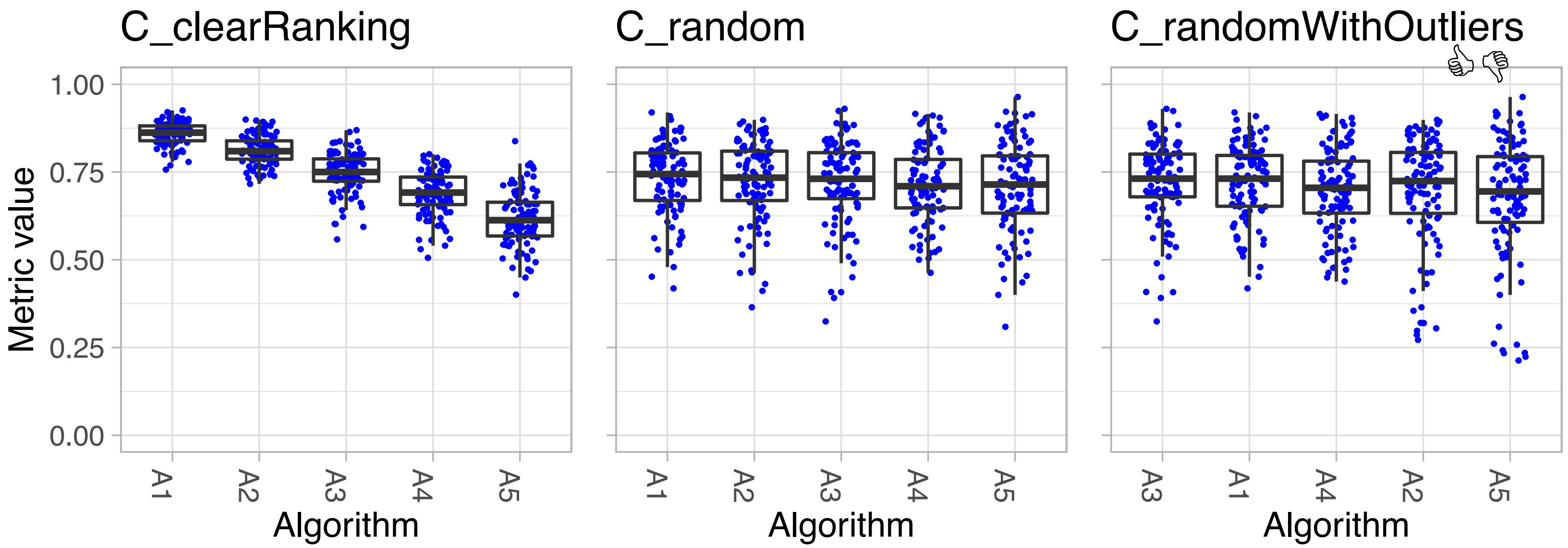
m.wiesenfarth@dkfz.de

Dot- and boxplot: Distribution of metric values



m.wiesenfarth@dkfz.de

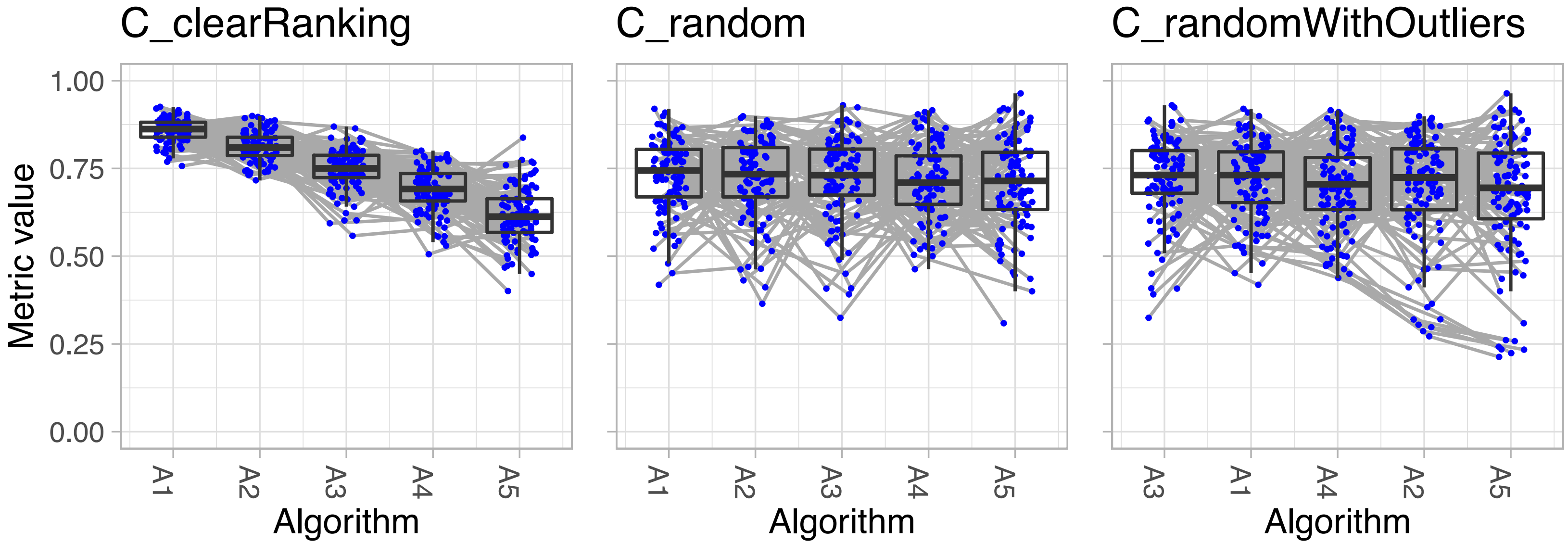
Dot- and boxplot: Distribution of metric values



+ Helps identify implausible values / inconsistencies in the data

m.wiesenfarth@dkfz.de

Dot- and boxplot: Distribution of metric values

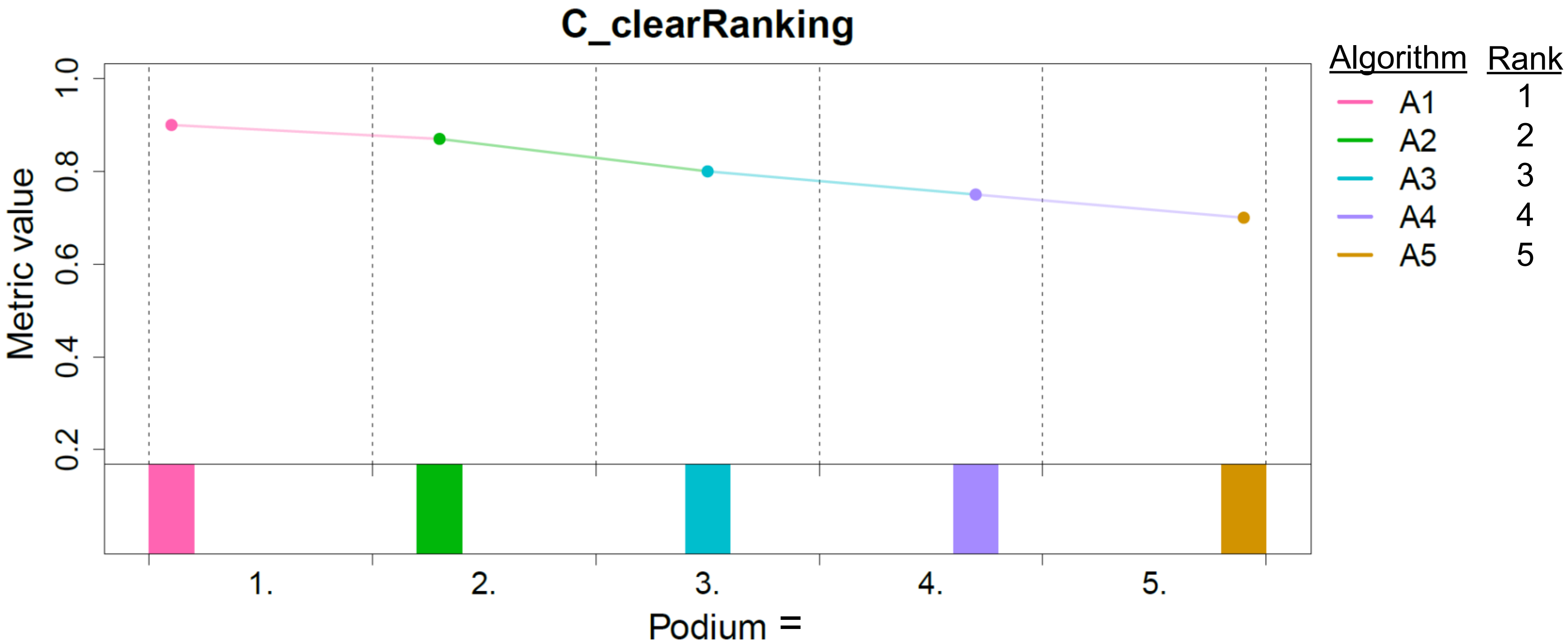


+ Helps identify implausible values / inconsistencies in the data

-- Connection of metric values from the same test case?

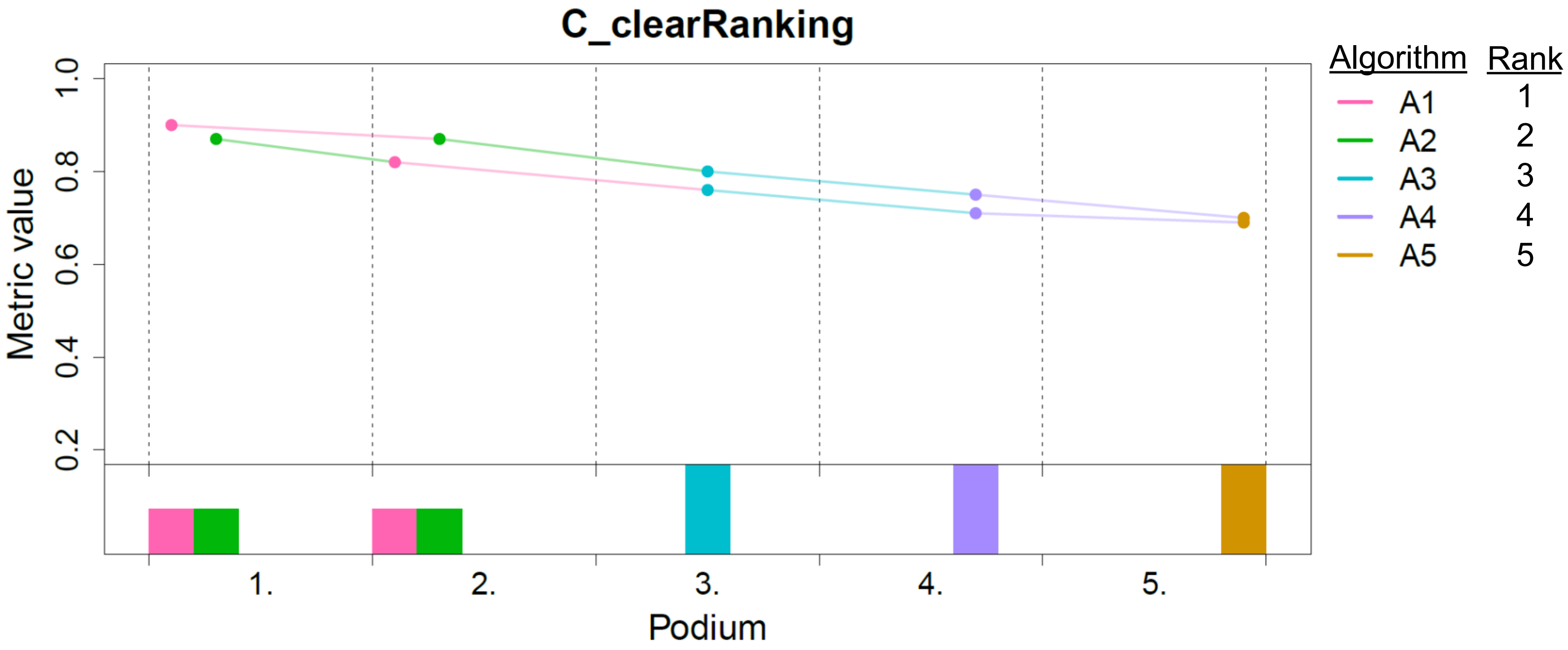
m.wiesenfarth@dkfz.de

Podium plot: Combined view of metric values and ranking



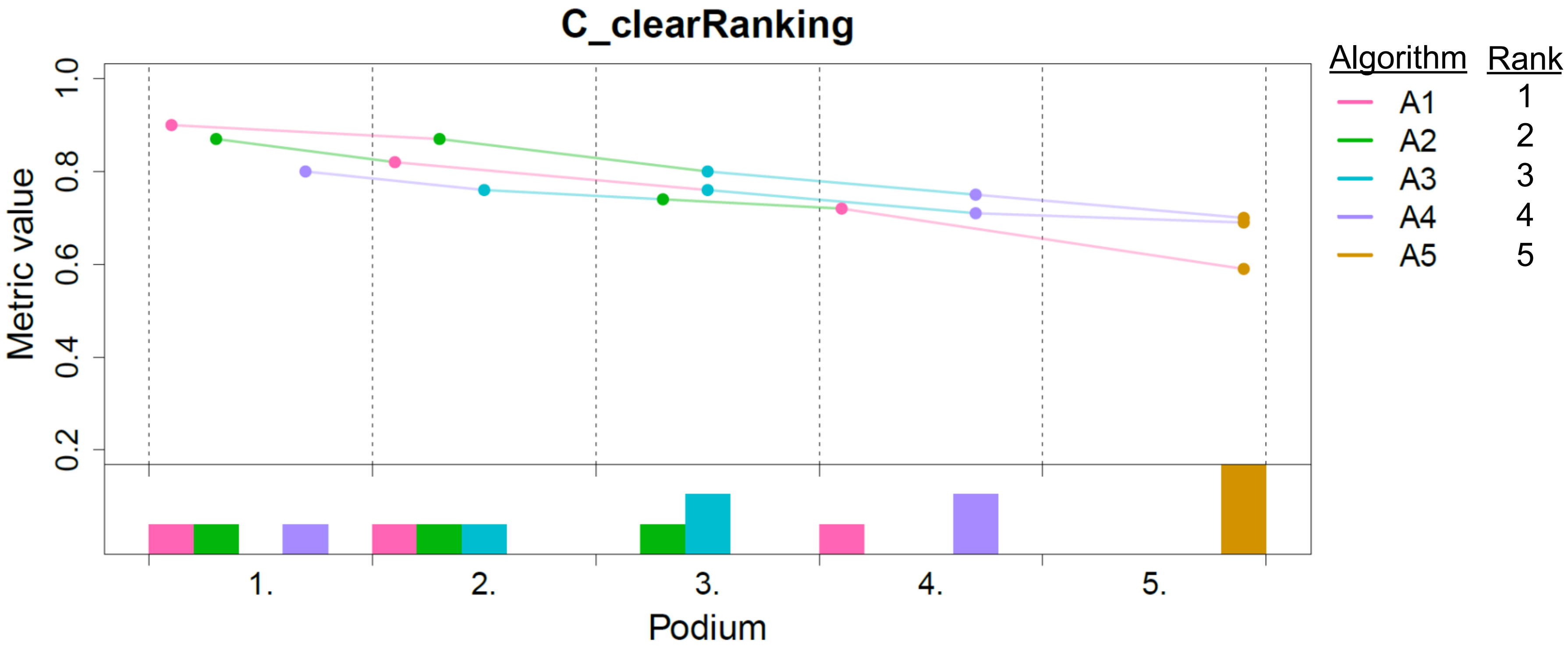
rank for the test case

Podium plot: Combined view of metric values and ranking



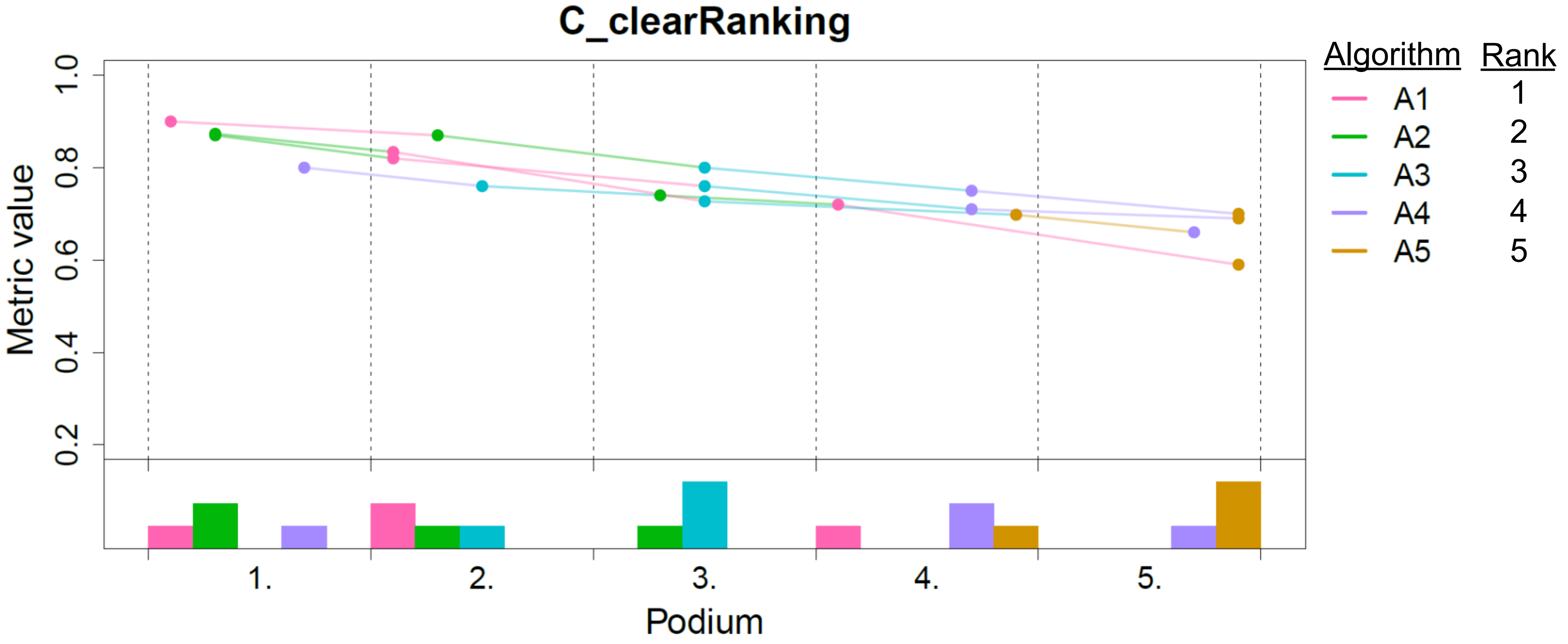
m.wiesenfarth@dkfz.de

Podium plot: Combined view of metric values and ranking



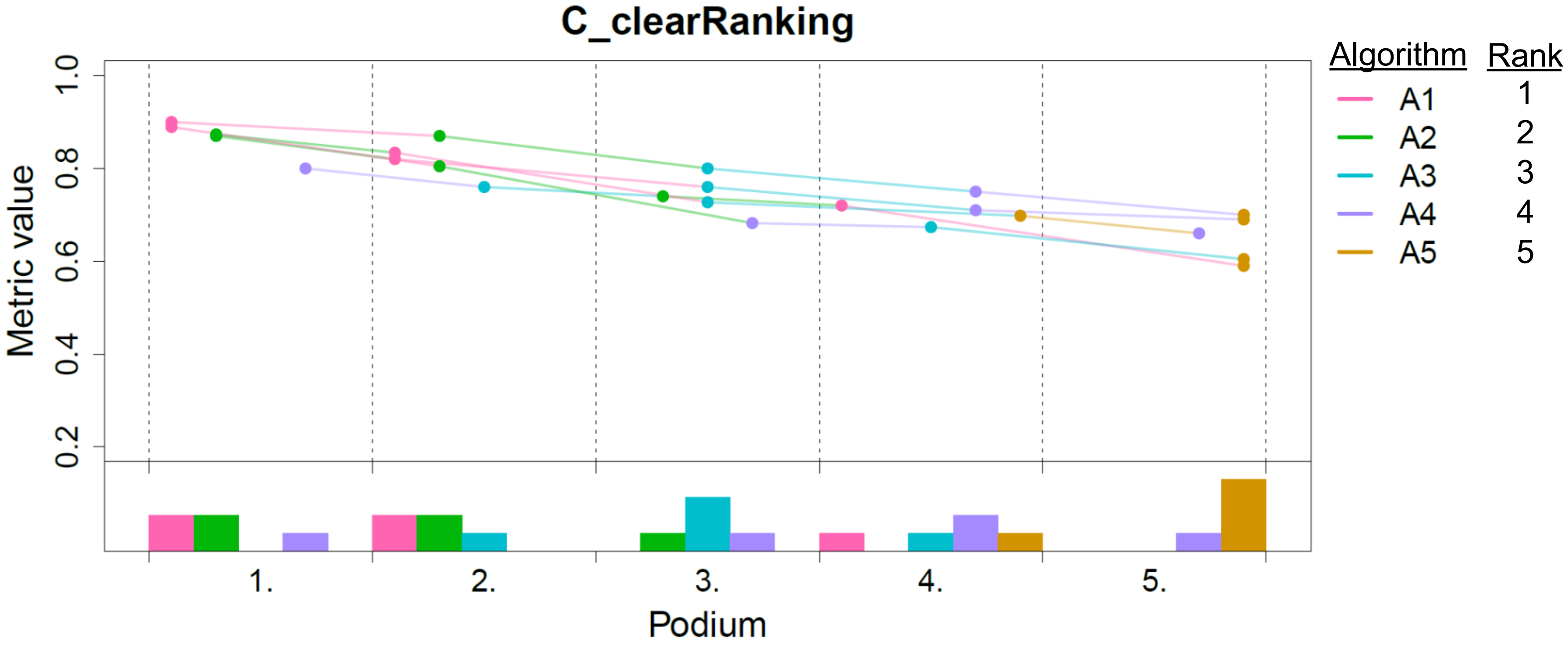
m.wiesenfarth@dkfz.de

Podium plot: Combined view of metric values and ranking

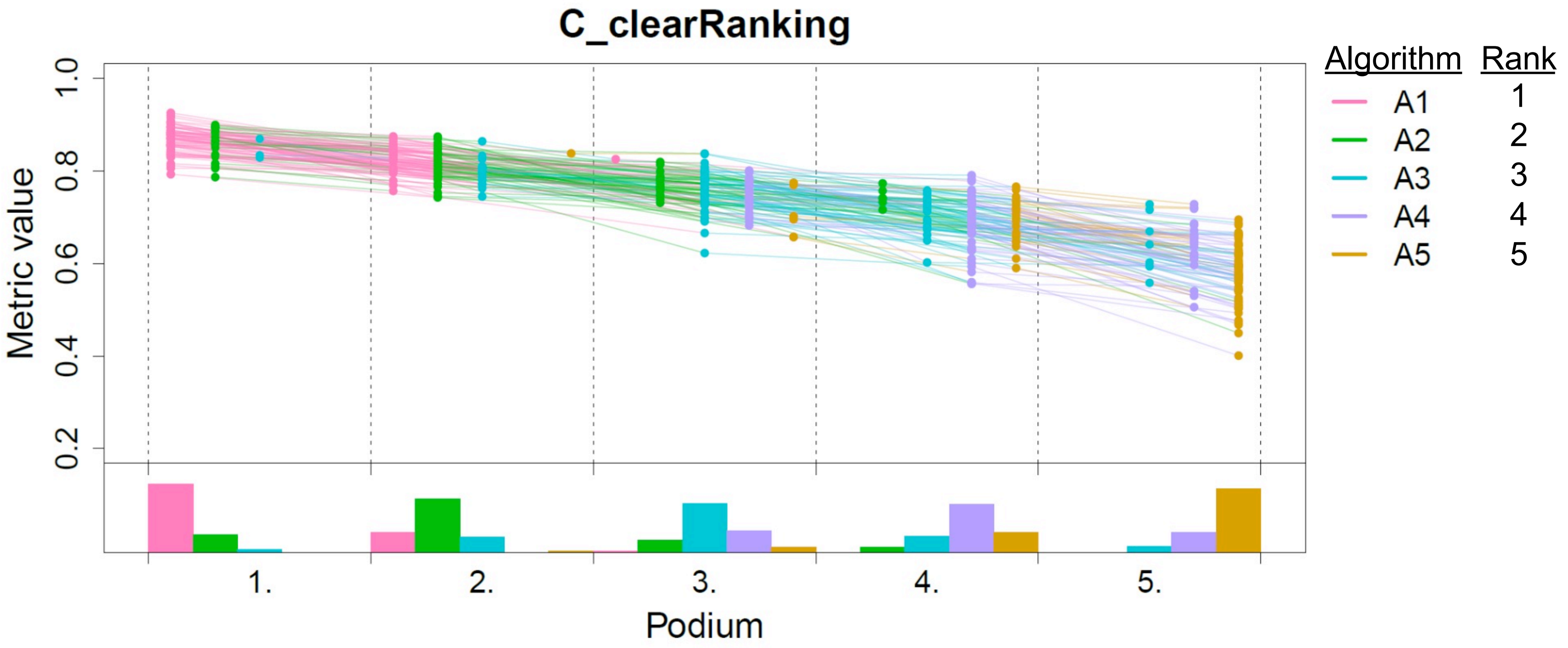


m.wiesenfarth@dkfz.de

Podium plot: Combined view of metric values and ranking

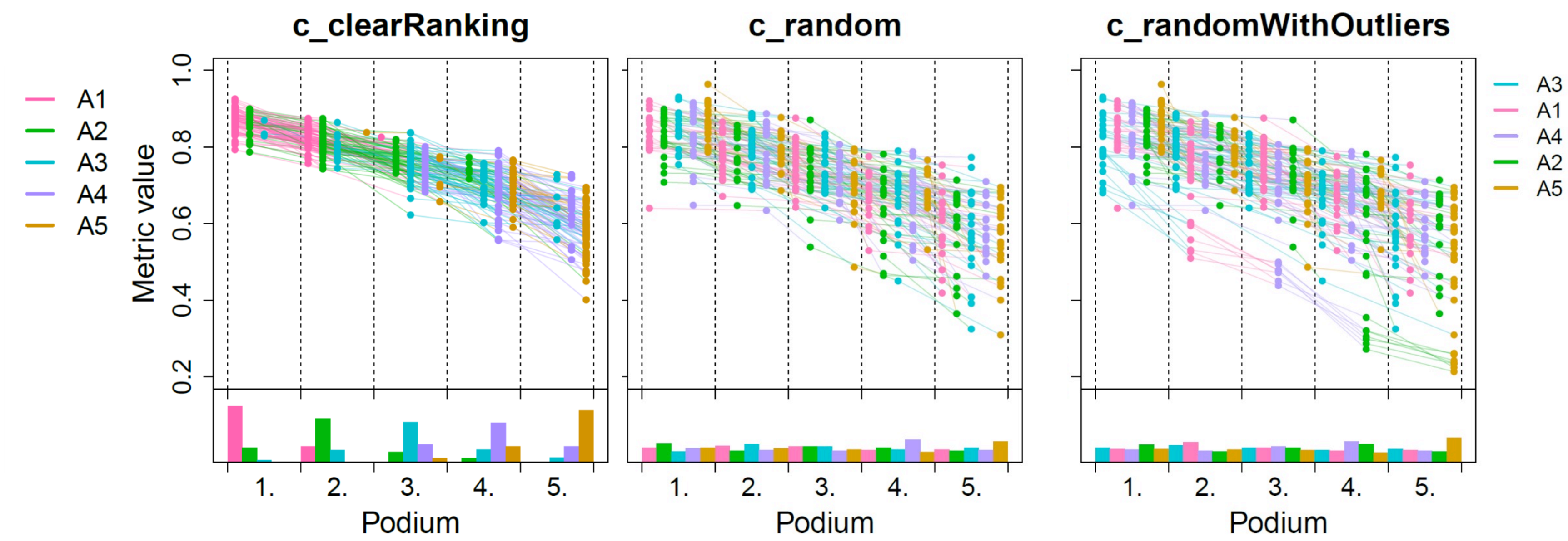


Podium plot: Combined view of metric values and ranking

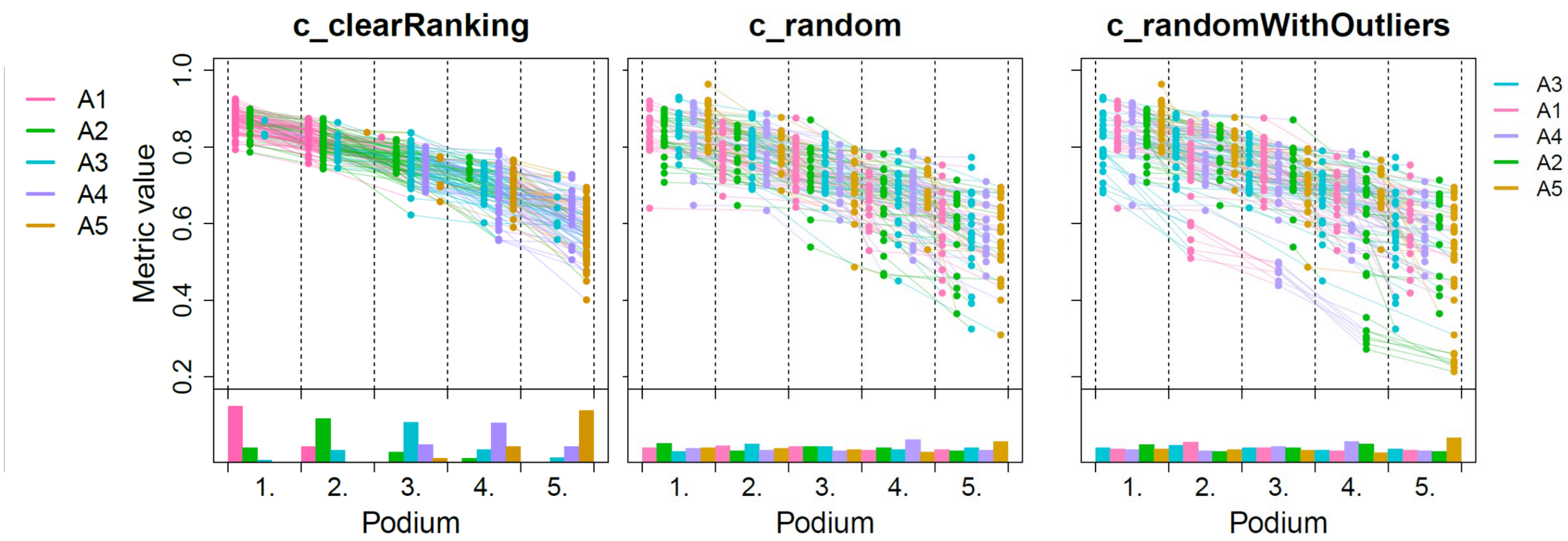


m.wiesenfarth@dkfz.de

Podium plot: Combined view of metric values and ranking



Podium plot: Combined view of metric values and ranking

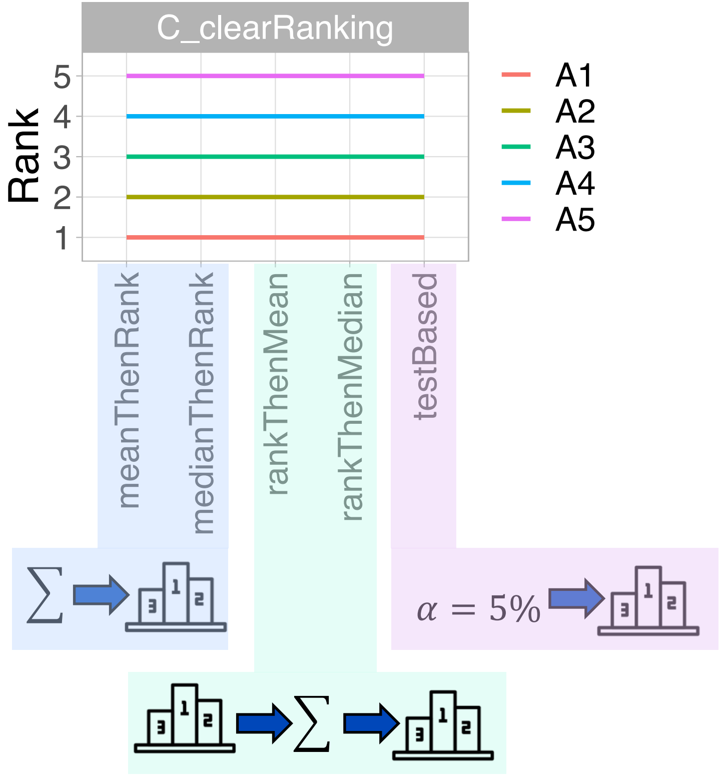


+ Connects metric values from the same test case
+ Allows identification of influential test cases
-- Relatively complex

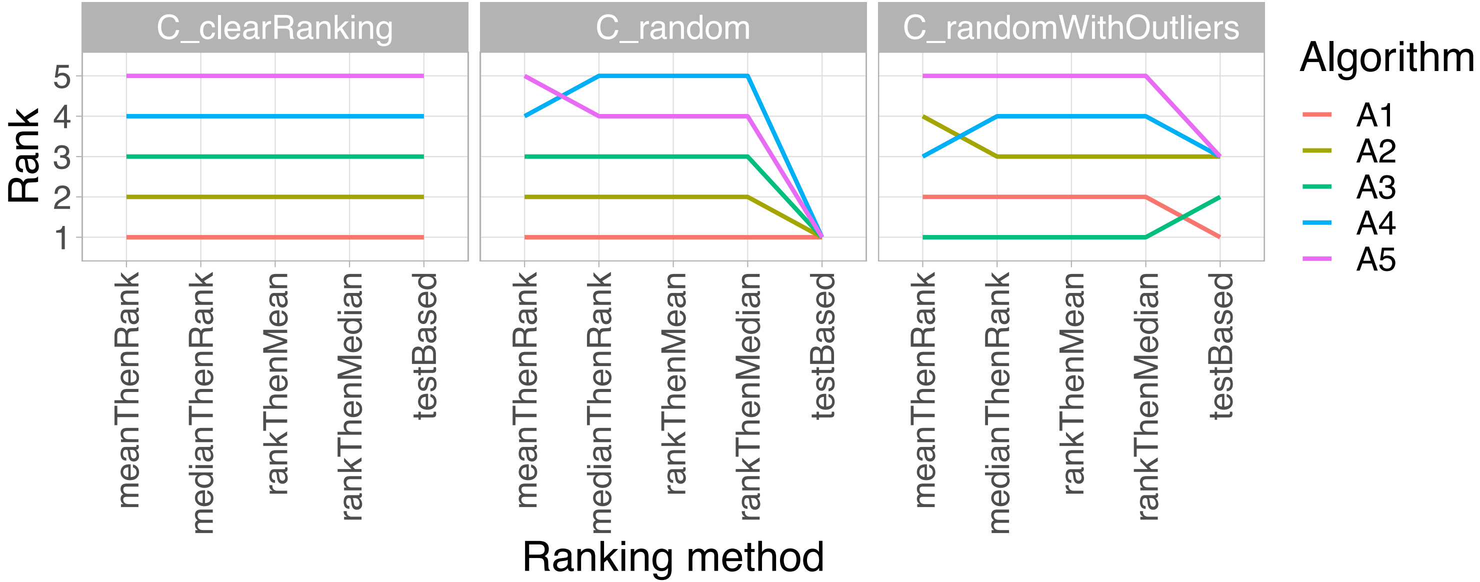
Ranking robustness and uncertainty

m.wiesenfarth@dkfz.de

Line plot: Ranking robustness with regard to ranking method



Line plot: Ranking robustness with regard to ranking method

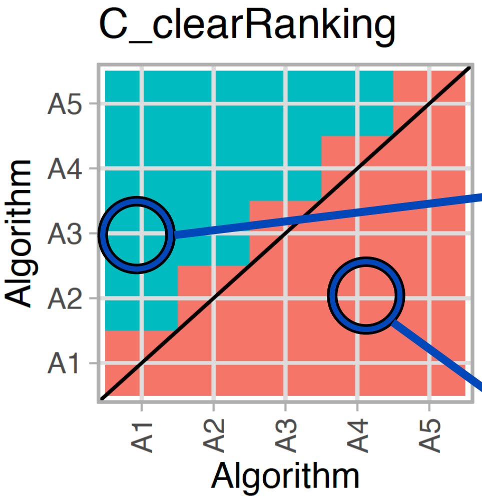


Can be modified to assess robustness with regard to metric, across tasks, …

m.wiesenfarth@dkfz.de

Significance map: Ranking uncertainty using hypothesis tests

Test cases can be considered as a finite sample from a larger population

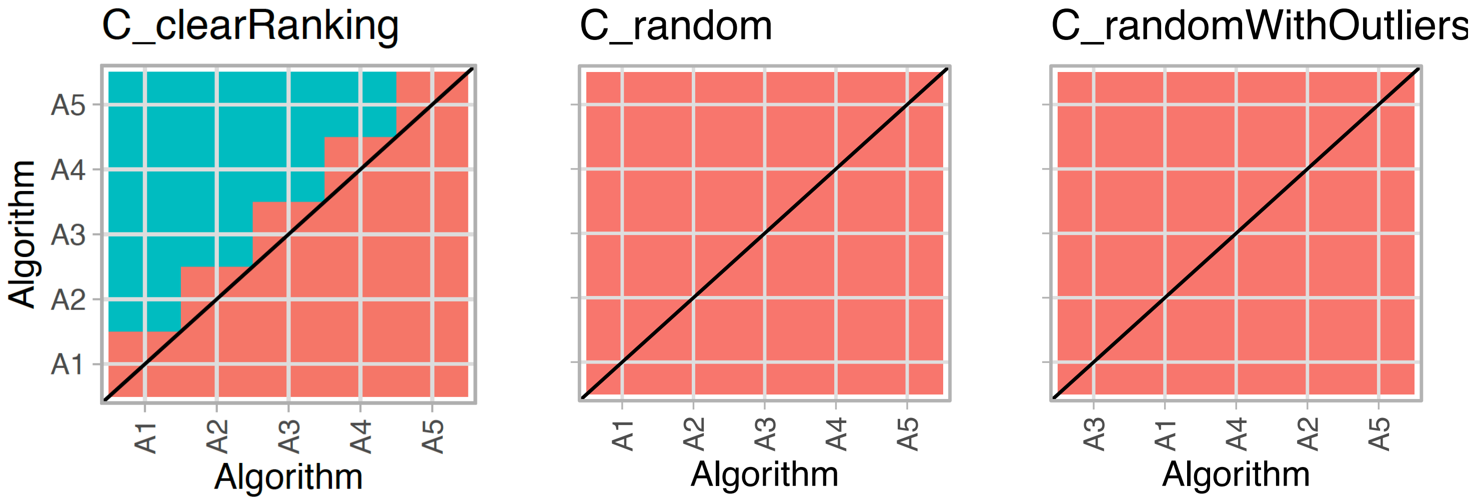


A1 is significantly superior to A3

A4 is not significantly superior to A2

m.wiesenfarth@dkfz.de

Significance map: Ranking uncertainty using hypothesis tests



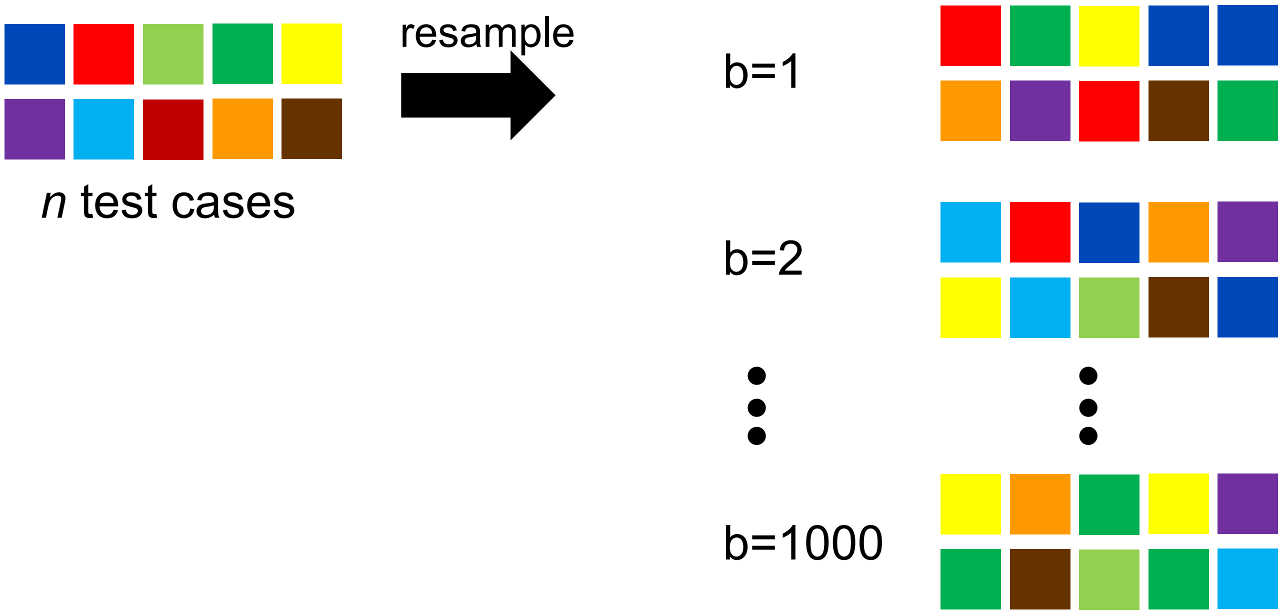
\+ Winner significantly superior to others?

-- Difference between algorithms relevant?

-- Little power to detect differences when # test cases small

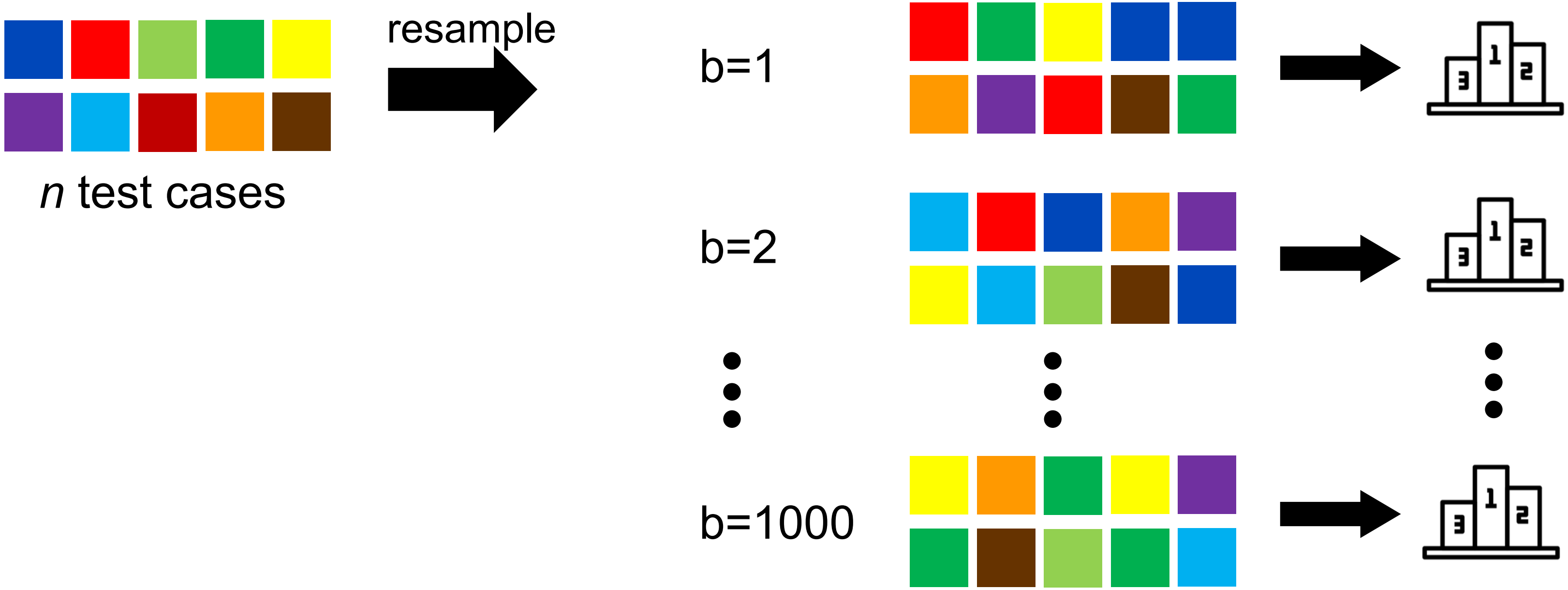
Ranking uncertainty using bootstrapping

1. Use available data set to generate 1000 bootstrap data sets



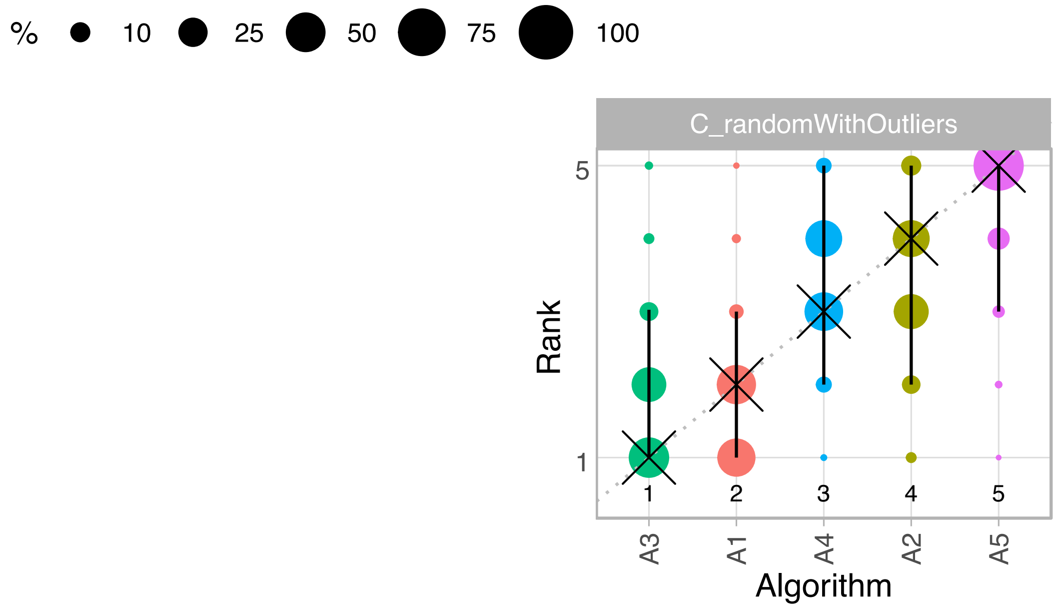
Ranking uncertainty using bootstrapping

1. Use available data set to generate 1000 bootstrap data sets
2. Perform ranking on each bootstrap data set



 m.wiesenfarth@dkfz.de

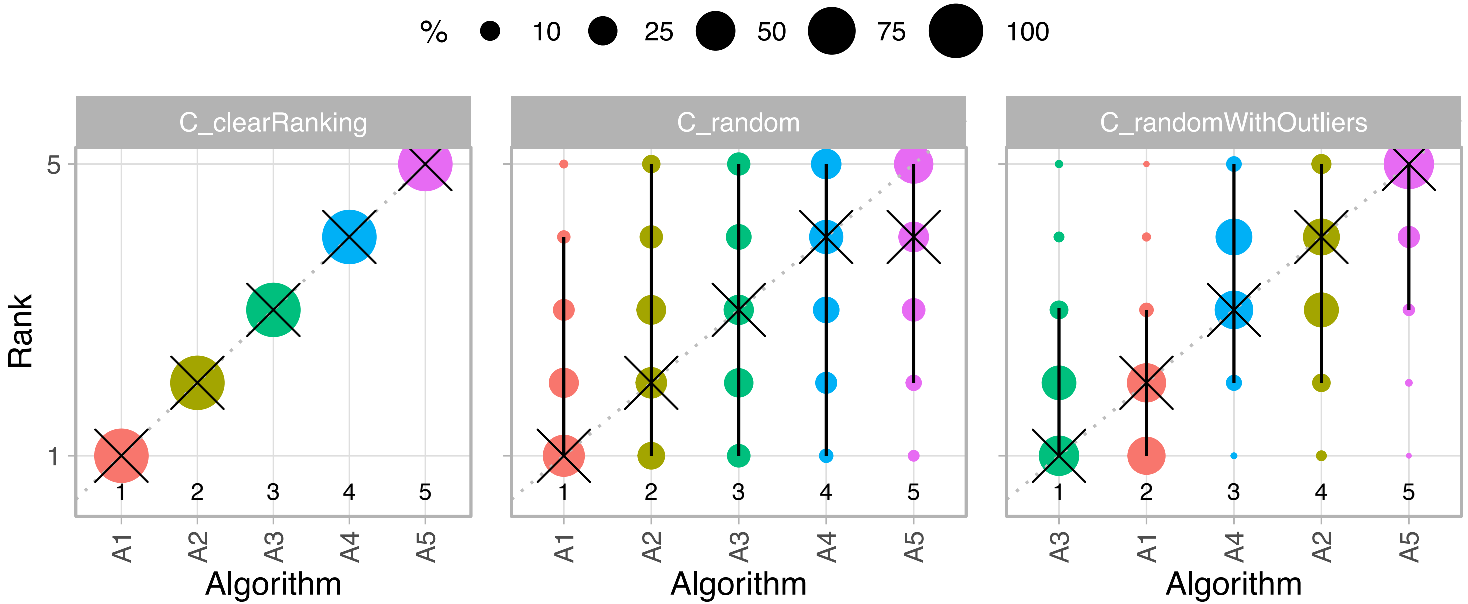
Blob plots: Ranking uncertainty using bootstrapping



+ What range of ranks for each algorithm is supported by the data?

m.wiesenfarth@dkfz.de

Blob plots: Ranking uncertainty using bootstrapping



+ What range of ranks for each algorithm is supported by the data?

-- Not sensible for very few test cases

-- Difference between algorithms relevant?

m.wiesenfarth@dkfz.de

Multi-task challenges

m.wiesenfarth@dkfz.de

Multi-task challenges

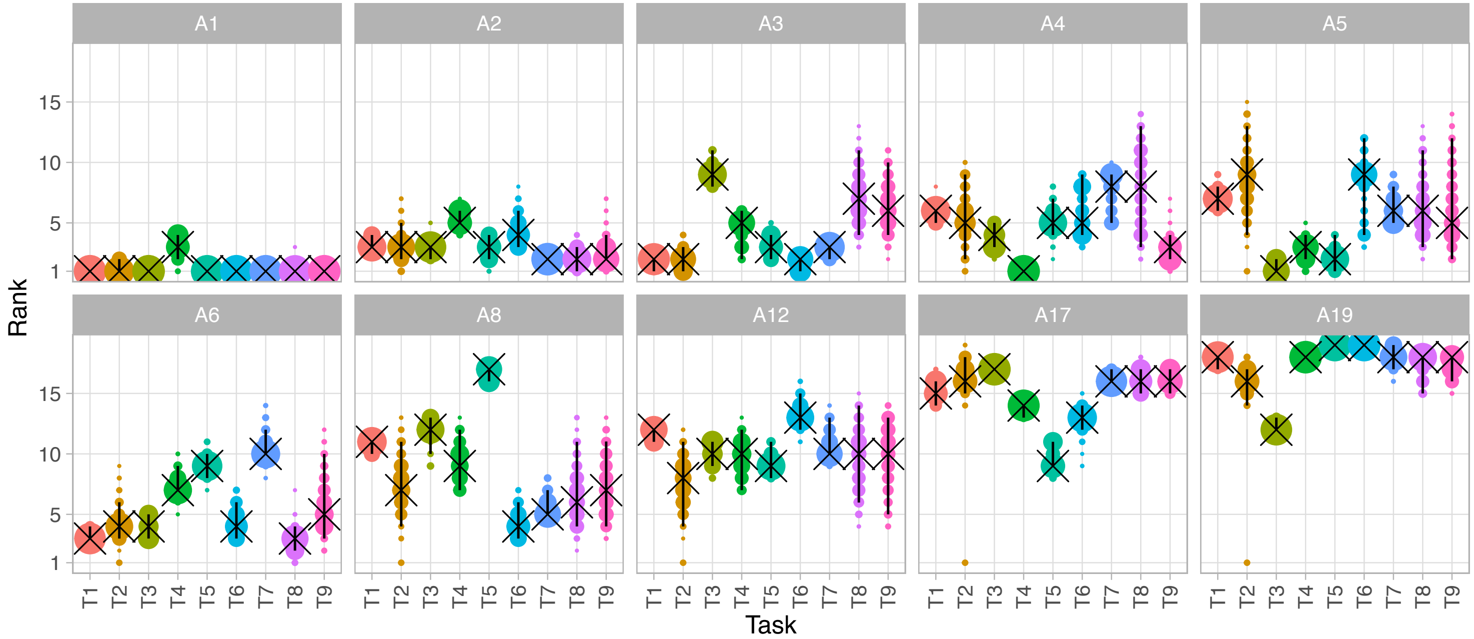
- Allow assessment of …
  - generalizability of algorithms across tasks
  - which task yields clear separation of algorithms
  - similarity of tasks with regard to rankings
  - …

- For illustration:
  - Subset of algorithms and tasks of *Medical Segmentation Decathlon* (2018)
  - Anonymized algorithms A_i, anonymized tasks T1-T9

MSD 2018, https://medicaldecathlon.com

m.wiesenfarth@dkfz.de

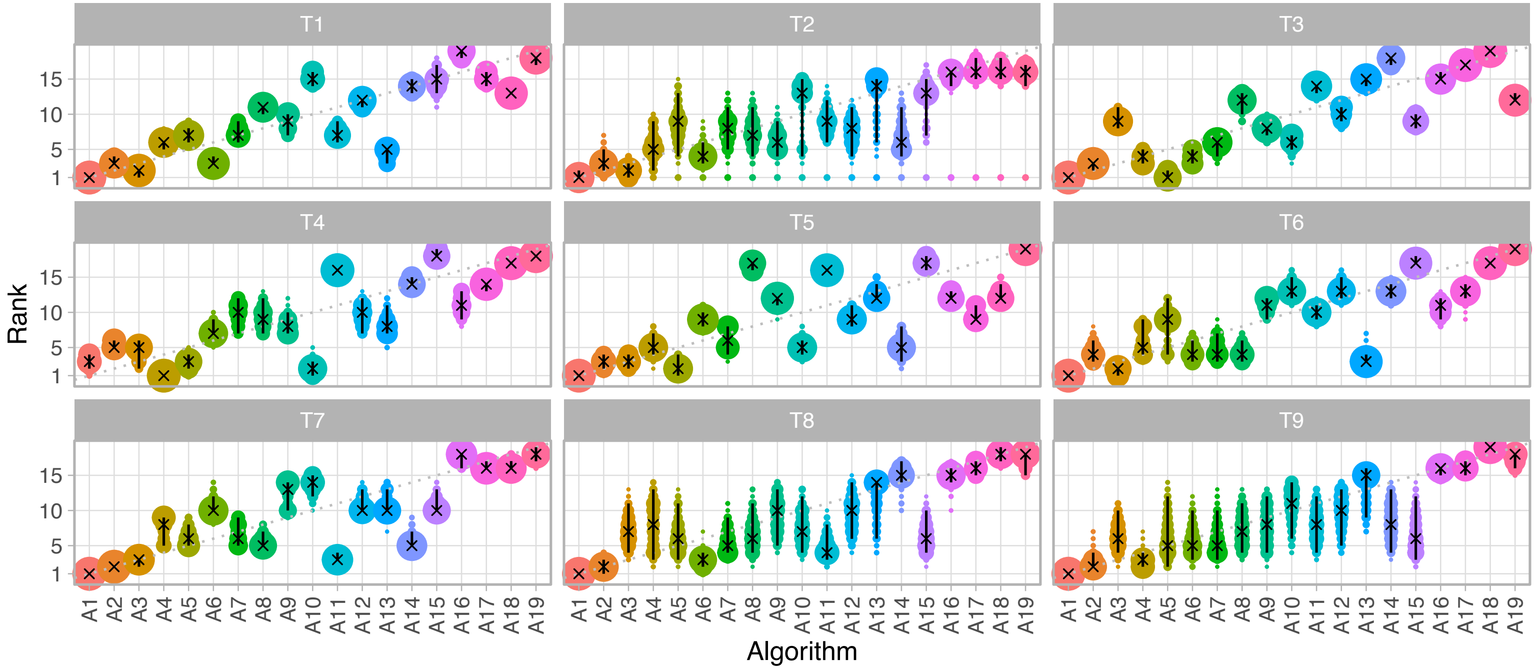
Blob plot stratified by algorithm



+ What range of ranks is supported by the data within and across tasks?
+ In which task does algorithm rank favorably or unfavorably?

m.wiesenfarth@dkfz.de

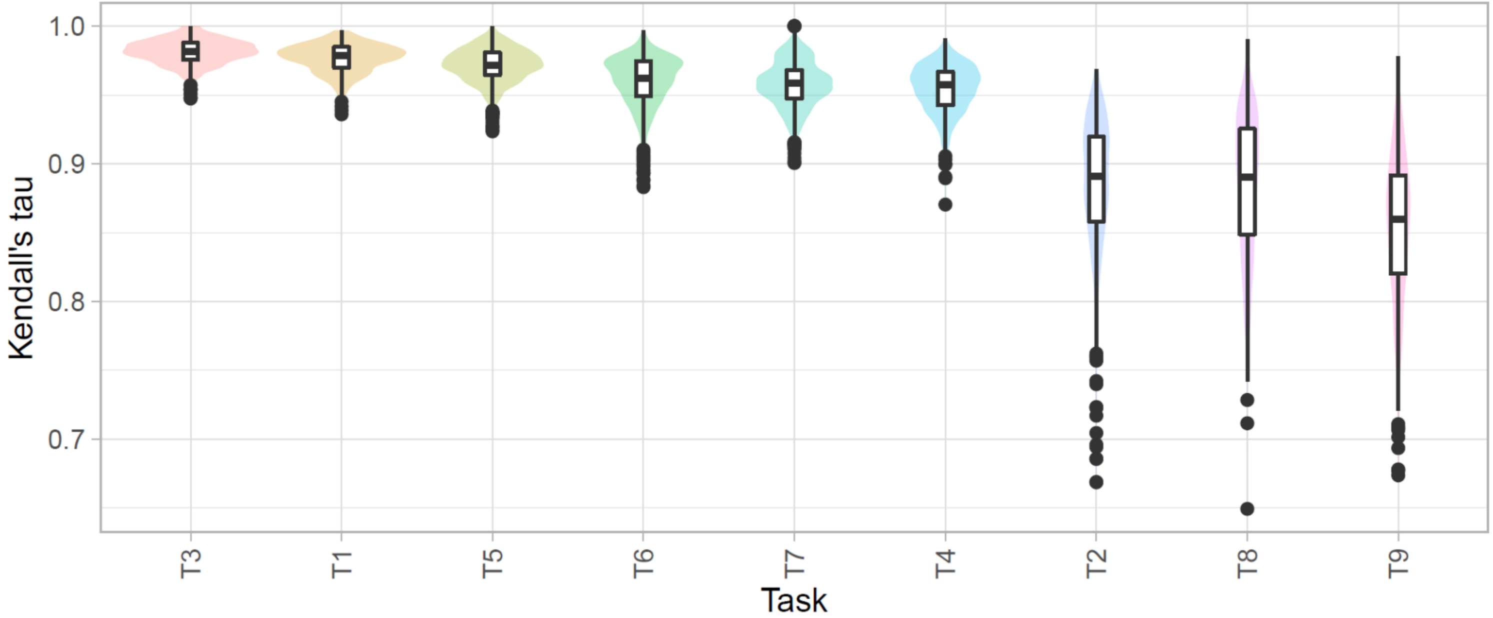
Blob plot stratified by task



+ Which task yields clear separation of algorithms?

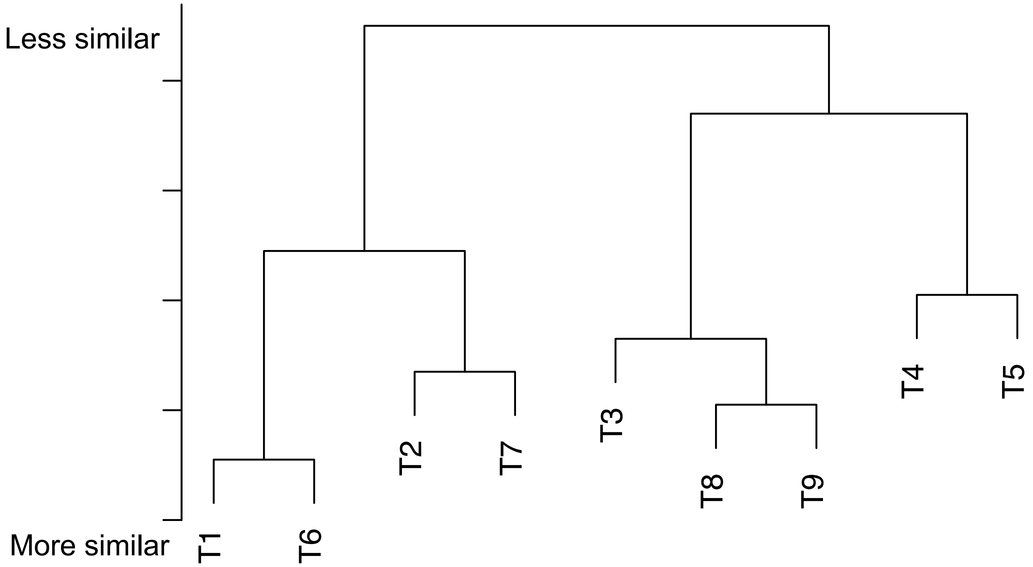
m.wiesenfarth@dkfz.de

Violin plot: Overall ranking stability of a task



+ Compares ranking stability of entire ranking lists across tasks

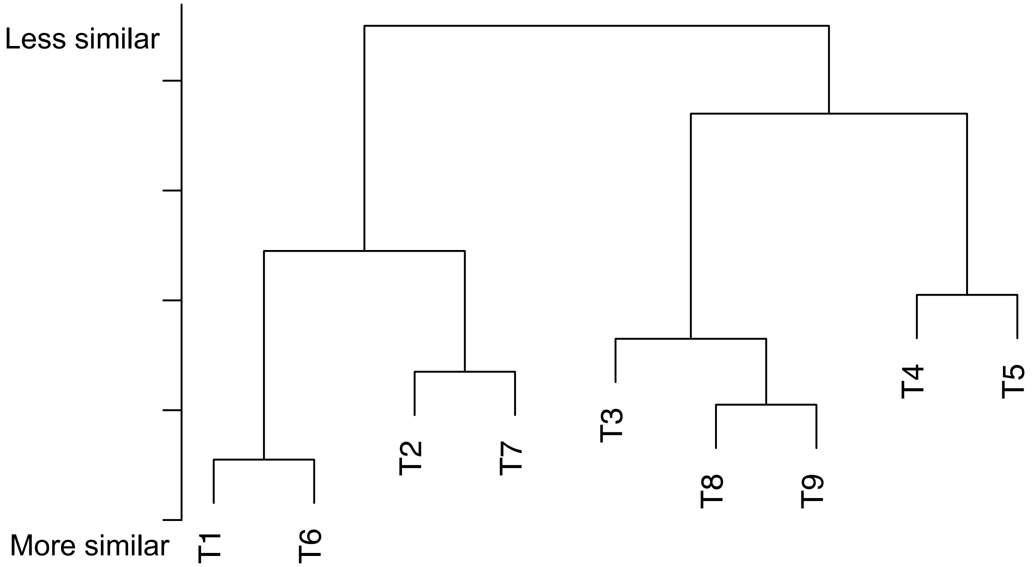
Dendrogram: Similarity of tasks with regard to ranking

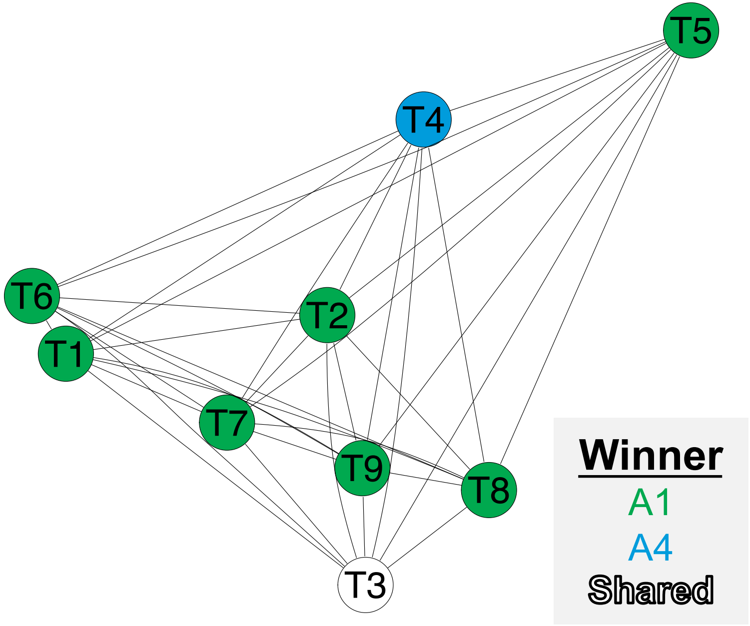


+ Group tasks according to similarity of their ranking lists

m.wiesenfarth@dkfz.de

Dendrogram/Network: Similarity of tasks with regard to ranking

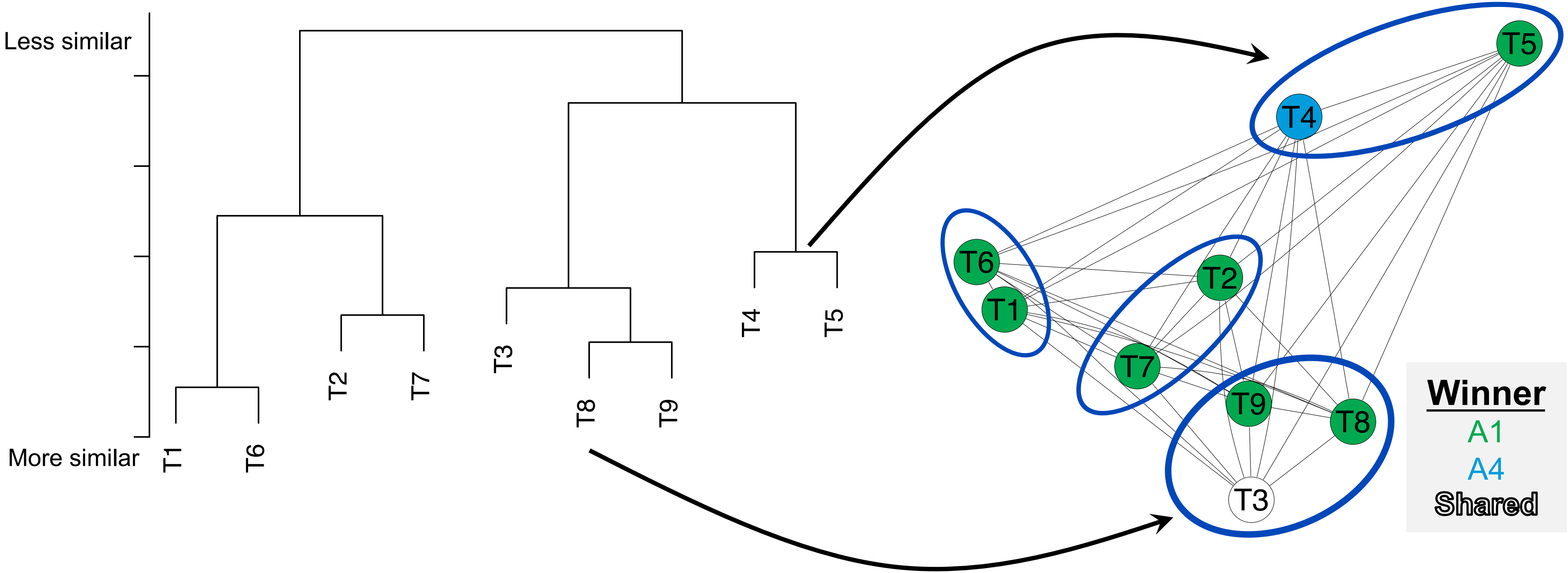




+ Group tasks according to similarity of their ranking lists

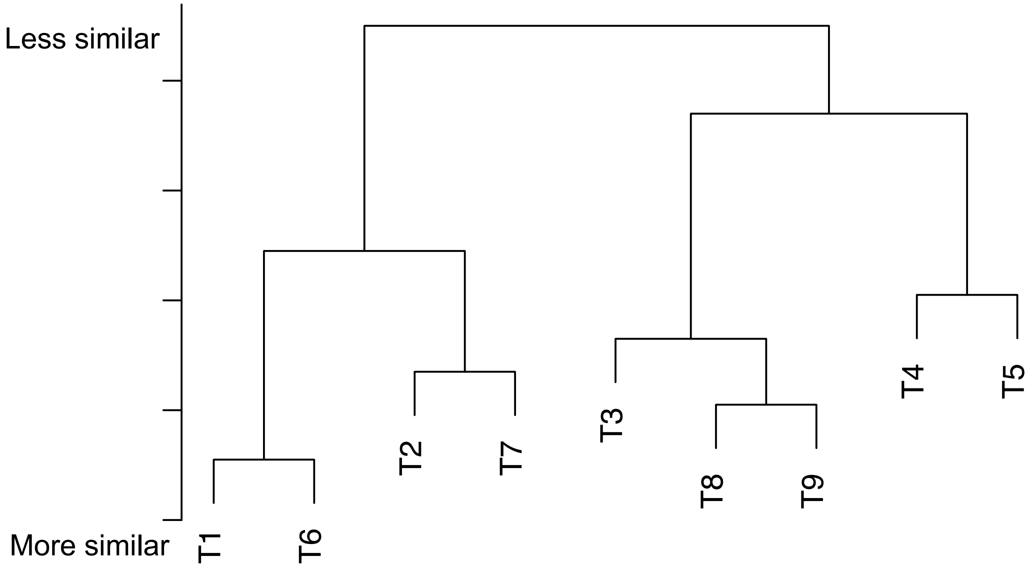
m.wiesenfarth@dkfz.de

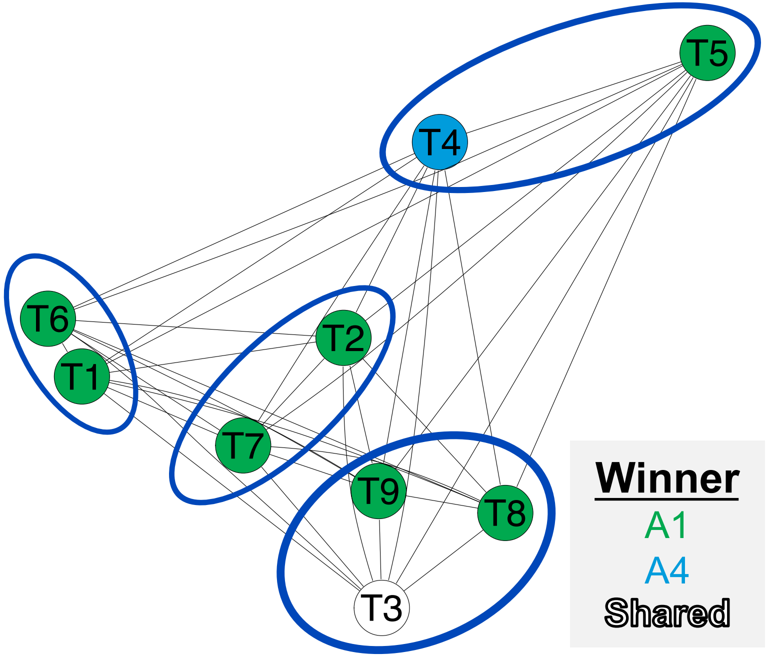
Dendrogram/Network: Similarity of tasks with regard to ranking



+ Group tasks according to similarity of their ranking lists

Dendrogram/Network: Similarity of tasks with regard to ranking





+ Group tasks according to similarity of their ranking lists
-- Depend on chosen distance measure and agglomeration method/layout engine

m.wiesenfarth@dkfz.de

Open-source framework: *challengeR*

m.wiesenfarth@dkfz.de

Open-source toolkit *challengeR*

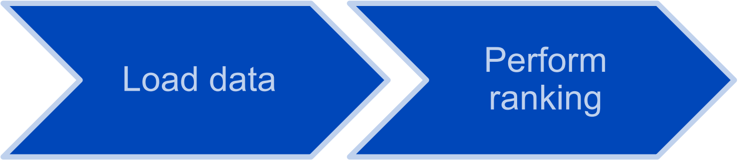
Load data

Testcase_ID	Algorithm_name	Metric_value	Task_name
85	A1	0.7952	c_random
15	A4	0.6877	c_clearRanking
81	A3	0.7754	c_random
8	A5	0.6948	c_random
82	A2	0.8576	c_clearRanking
19	A2	0.5556	c_random
84	A1	0.5215	c_random

```
data_matrix = read.csv(file.choose())
challenge = data_matrix %>% as.challenge(
                         case="Testcase_ID", algorithm="Algorithm_name",
                         value="Metric_value",
                         by="Task_name", # (only for multi-task)
                         smallBetter = FALSE)
```


m.wiesenfarth@dkfz.de

Open-source toolkit *challengeR*




```
ranking = challenge %>% testThenRank(...)
ranking = challenge %>% aggregateThenRank(...)
ranking = challenge %>% rankThenAggregate(...)

# Consensus ranking for multiple tasks:
  meanRanks = ranking %>% consensus(method = "euclidean")
```

Open-source toolkit *challengeR*

Load data → Perform ranking → Perform bootstrapping

```
ranking_bootstrapped = ranking %>% bootstrap(...)
```


m.wiesenfarth@dkfz.de

Open-source toolkit *challengeR*

Load data → Perform ranking → Perform bootstrapping → Generate report

```
report(ranking_bootstrapped,
       consensus=meanRanks,
       format = "PDF",
       ...)
```


m.wiesenfarth@dkfz.de

Open-source toolkit *challengeR*

Alternatively all in a single call:

```
data_matrix %>% as.challenge(…) %>%
       test(…) %>% rank(…) %>% bootstrap(…) %>% report(…)
```


m.wiesenfarth@dkfz.de

Open-source toolkit *challengeR*

Load data → Perform ranking → Perform bootstrapping → Generate report

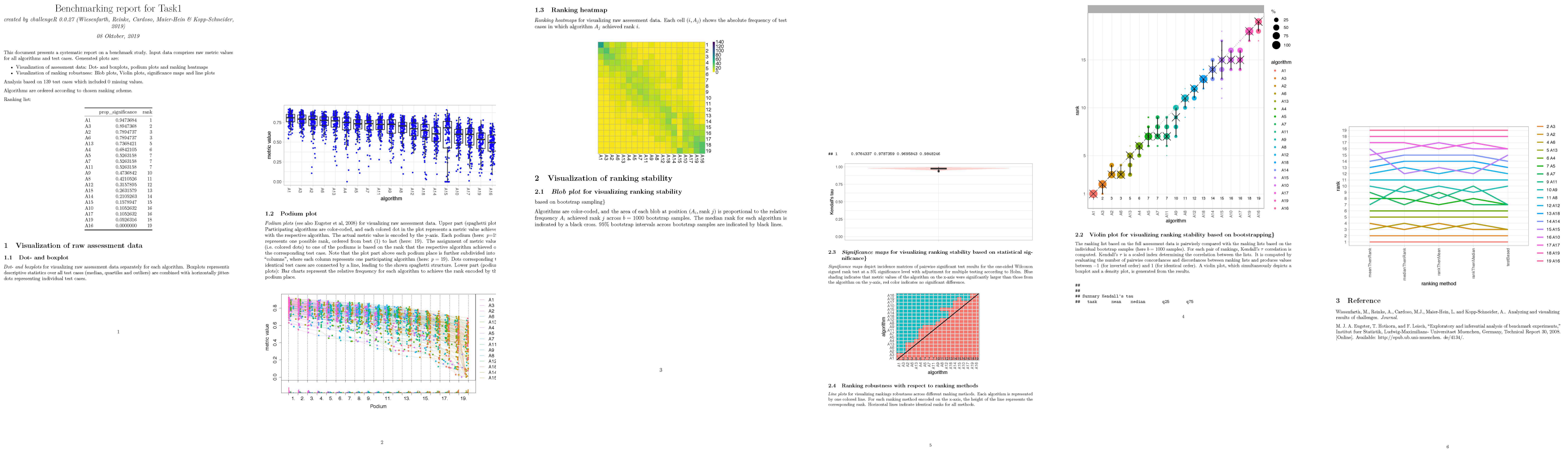
Alternatively all in a single call:

```
data_matrix %>% as.challenge(…) %>%
       aggregate(…) %>% rank(…) %>% bootstrap(…) %>% report(…)
```


m.wiesenfarth@dkfz.de

Open-source toolkit *challengeR*

Load data → Perform ranking → Perform bootstrapping → Generate report



m.wiesenfarth@dkfz.de

Open-source toolkit *challengeR:* Single plots

`ranking %>%`
- `boxplot(...)`
- `podium(...)`
- `rankingHeatmap(...)`
- `significanceMap(...)`
- …

`ranking_bootstrapped %>%`
- `violin(...)`
- …

m.wiesenfarth@dkfz.de

Summary

- Uncertainty and robustness of rankings needs to be assessed

- Open-source toolkit (R package)
  - Available on `https://github.com/wiesenfa/challengeR`
  - Computes rankings for diverse ranking methods
  - Generates an automatic report

- Wiesenfarth, Reinke, Landmann, Cardoso, Maier-Hein & Kopp-Schneider (2019). Methods and open-source toolkit for analyzing and visualizing challenge results. *arXiv preprint arXiv:1910.05121*

m.wiesenfarth@dkfz.de

Summary

- Anyone already tried to use it?

- Feedback / problems / feature requests?

- Contributions welcome!

m.wiesenfarth@dkfz.de