The pitfalls of competitions

Lasse Becker-Czarnetzki

Heidelberg University
How to lie with statistics
Winter 2019/2020

Dezember 12, 2019


TOP 10 HIGH JUMPS


JAMAICA
TDK
BOLT

OpenML

MATLAB

UCI

Machine Learning Repository

1. Competitions overview
   - What's it about
   - Important techniques
2. What's going wrong?
   - No real standards...
   - Ranking robustness (Everybody wins)
3. What to do?
4. Conclusion

Why do we have competitions

- What did people do before?
  - Use of own datasets
  - No fair/easy comparison
  - Strongly biased results?

- Public data (quality checked)
- Fair comparisons (same conditions)
- Efficient research exchange $\rightarrow$ Progress
- Establishment of good methods, State-of-the-art
- Getting seen, published

Benchmarking principles/Best practices

- Validity
  - Standardized procedure. (Training, test split)
  - Statistical sound procedure.
- Reproduceability
  - Experiments description (e.g Hyperarameters)
  - Data description (e.g Preprocessing)
  - Hardware and software enviroment
- Comparability
  - Identical experimental setup:
    - Benchmark problems
    - Datasets
    - statistcal analysis
  - Use sound metrics to capture relevant difference in performance
  - Avoid bias in data partitions

Best practices

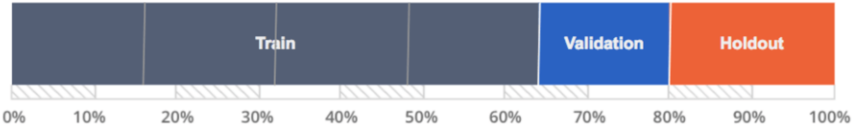
- Get robust ranking
  - Metric based aggregation
  - Sound statistical significance
  - Avoid correlated metrics
- Combat lack of representation
  - Evaluate on broad spectrum of datasets
  - Evaluate on datasets with different statistical properties
    - Number of features
    - Number of classes
    - Noise

Main important techniques

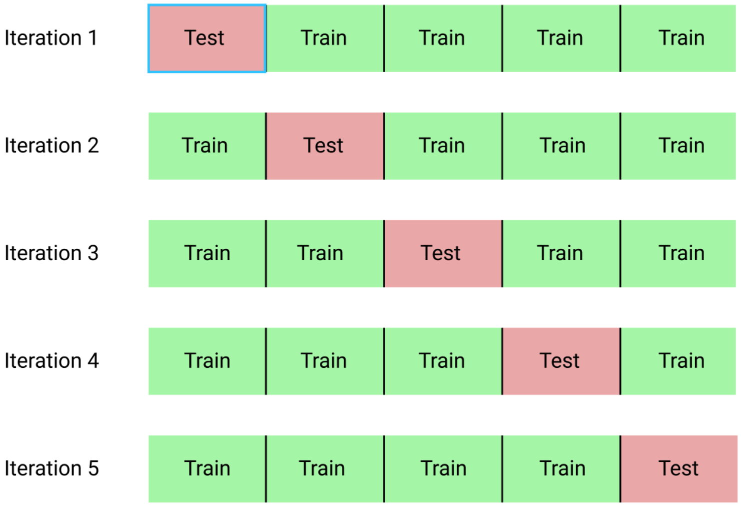
- Hold out method
- Cross-validation
  - K-folds cross validation
  - Leave one out
  - Stratified cross validation
- Bootstrapping

Hold out method

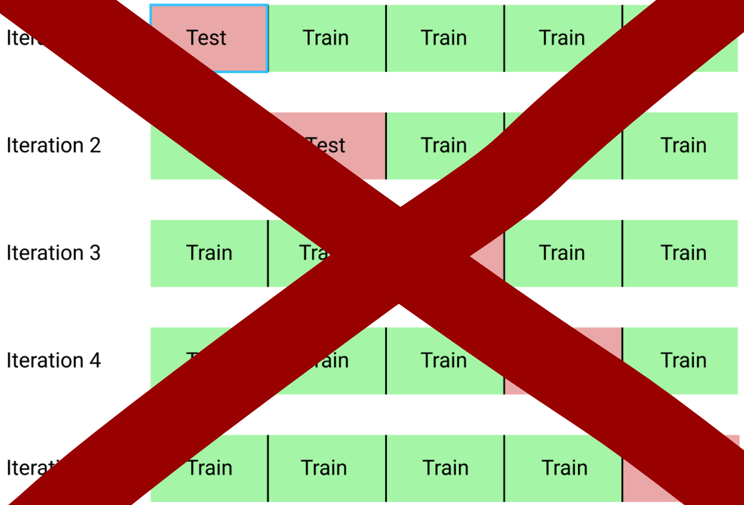
- Split dataset into **train**, **validation**, **test** set
- Don't release test set to prevent data snooping



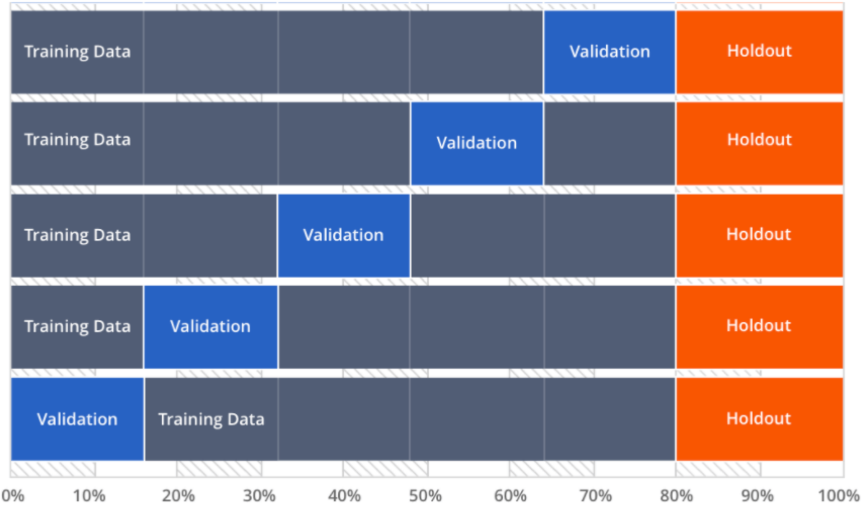
K-Fold Cross Validation



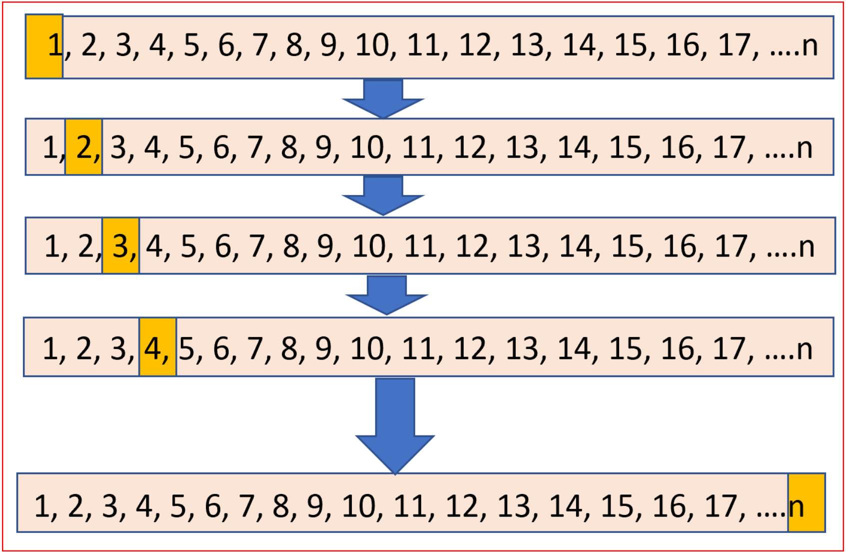
Cross Validation



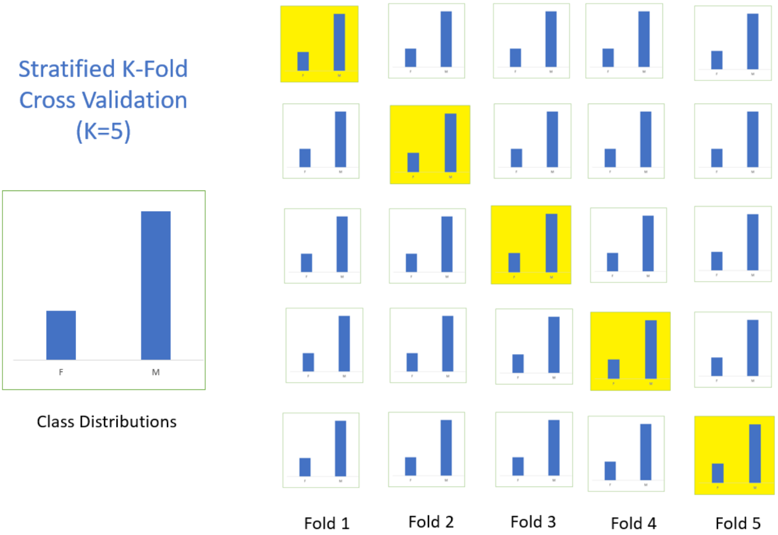
K-Fold Cross Validation



Hold one out


1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, ...n
1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, ...n
1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, ...n
1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, ...n
1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, ...n

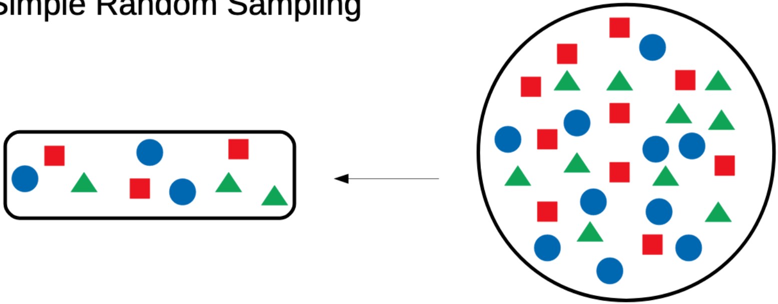
Stratified Cross Validation



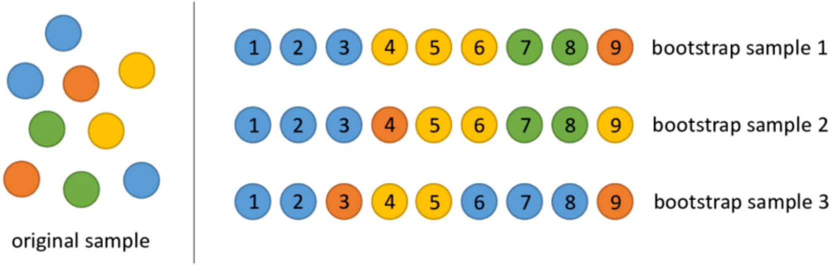
Bootstrapping

Random Sampling

- Simple Random Sampling



Bootstrapping



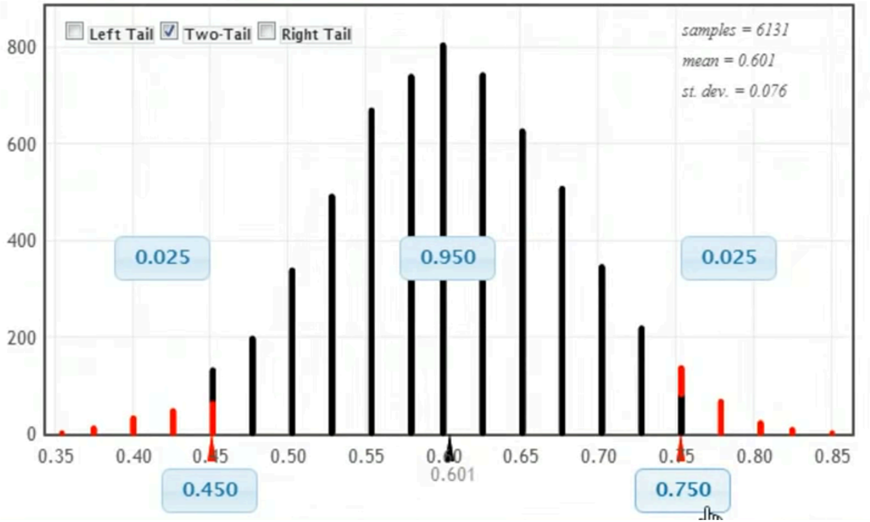
Bootstrapping

- Repeat this process b times (10.000)
- For every resample take some meaningful value (e.g mean)
- Now you can do statistical analysis

Bootstrapping

- Repeat this process b times (10.000)
- For every resample take some meaningful value (e.g mean)
- Now you can do statistical analysis

Bootstrapping



Quality measures

- **Metric Performance** (e.g. Accuracy)
- Model complexity
- Computational complexity
- Scalability
- Sample complexity
- Interpretability

Structure

1 Competitions overview
- What's it about
- Important techniques

2 What's going wrong?
- No real standards...
- Ranking robustness (Everybody wins)

3 What to do?

4 Conclusion

Study on biomedical image analysis competitions

- Study by [Maier-Hein et al. 2018]
- 150 competitions, 549 tasks over 12 years
- Statistical analysis (What are the numbers?)
- Critical analysis
  - Are the challenges sound in procedure
  - **What are they main problems?**
  - **What best practices can combat these?**

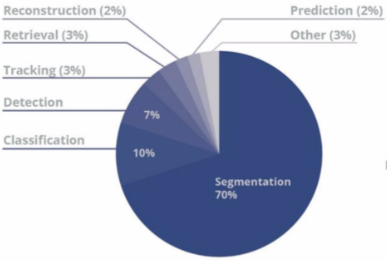


Figure 1: Biomedical image analysis tasks [Maier-Hein et al. 2018]

Why was this necessary

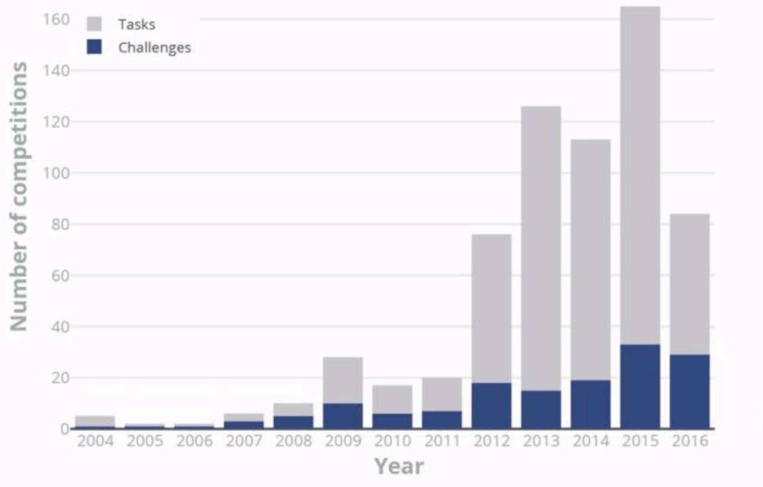


Figure 2: Overview of biomedical image analysis challenges
Maier-Hein et al. 2018

Relevant information not reported

- Authors created list of 53 parameters, that a challenge should report
- 43% of parameters were not reported for 50% of all tasks
- Examples of not reported parameters:
  - 08%: Rank aggregation method
  - 85%: If provided training data was supplemented with other data
  - 66%: Description of gold standard annotations
  - 45%: Annotation aggregation method (Multiple annotators)
  - 19%: Annotator expertise level

Variability in challenge design

- 97 different metrics were used (half of them only on single task)
- 77% No justification for metric use
- 57% Use of single metric to determine winner
- 10 different methods for determining final rank

What can easily change the ranking

- Minor changes in metrics
- Different aggregation methods
- Different annotators
- Removing one test case
- Lack of missing data handling

Kendall's Tau

- Rank correlation coefficient
- Ordinal association between two measured quantities.
- Takes first ranking as starting point
- Looks how often does the second ranking break the first

$$\tau = \frac{S}{\frac{n(n-1)}{2}}$$

$$\tau \in [-1; 1]$$

$$S = concordants - disconcordants$$

	First	Second	Third	Fourth
Result A	1	2	3	4
Result B	4	1	2	3
Pairs	(1,4)	(2,1)	(3,2)	(4,3)

$$S = concordants - disconcordants$$

	First	Second	Third	Fourth
Result A	1	2	3	4
Result B	4	1	2	3
Pairs	(1,4)	(2,1)	(3,2)	(4,3)

Compare

(1,4)	(2,1)	(3,2)	(4,3)
(2,1)	(3,2)	(4,3)	
(3,2)	(4,3)		

concordant	disconcordant
	3
2	
1	

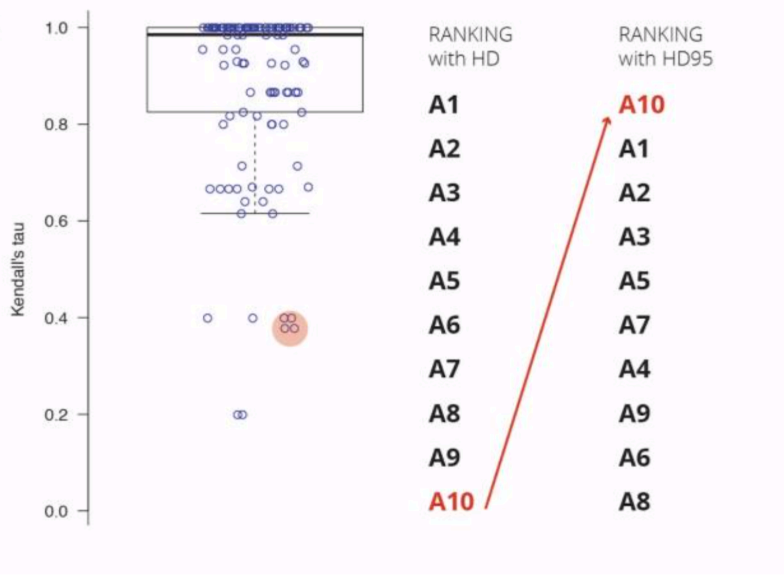


Figure 3: Ranking robustness, metric based [Maier-Hein et al. 2018]

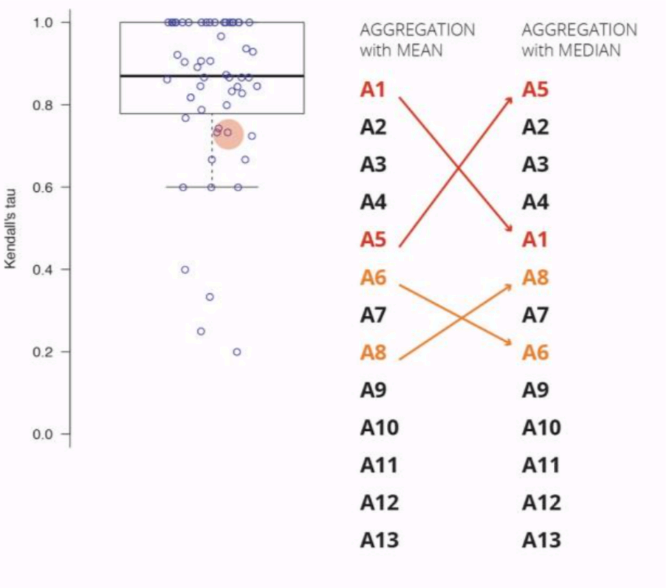


Figure 4: Ranking robustness, mean or median [Maier-Hein et al. 2018]

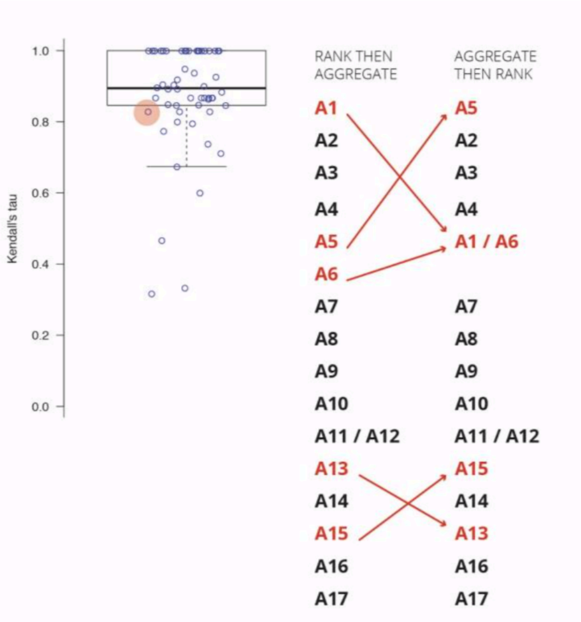


Figure 5: Ranking robustness, aggregation method **[Medic]**

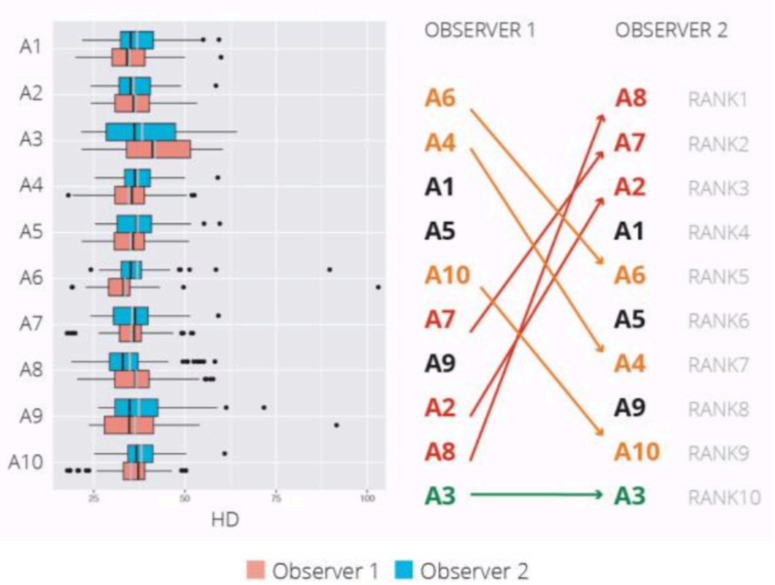


Figure 6: Ranking robustness, annotator (HD) [Maier-Hein et al. 2018]

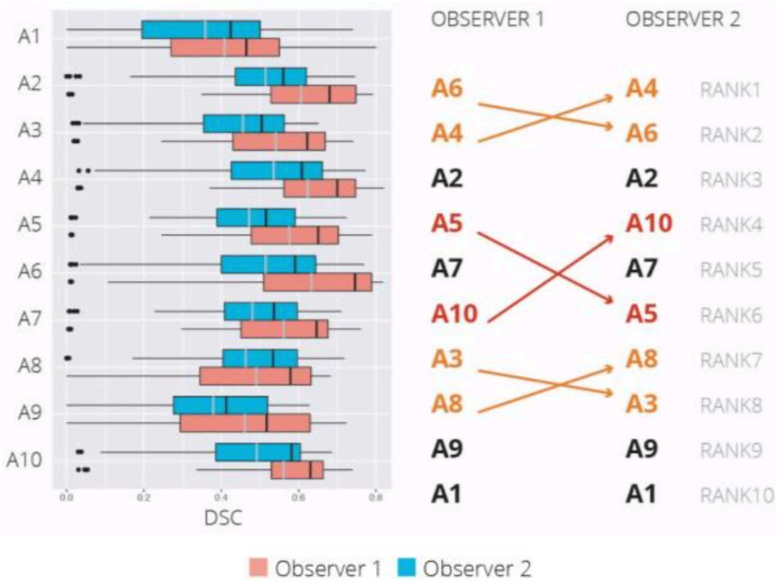


Figure 7: Ranking robustness, annotator (DSC) [Maier-Hein et al. 2018]

Bootstrapping experiments

- Bootstrap experiments on single-metric rankings
- Compare robustness of variables
- Resample (1000 times) check if original winner is still the winner

Bootstrapping experiments

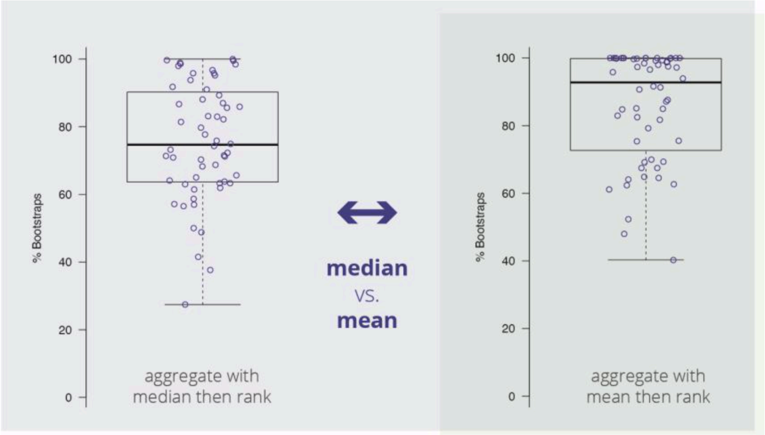


Figure 8: Robustness comparison median vs mean [**Medic**]

Bootstrapping experiments

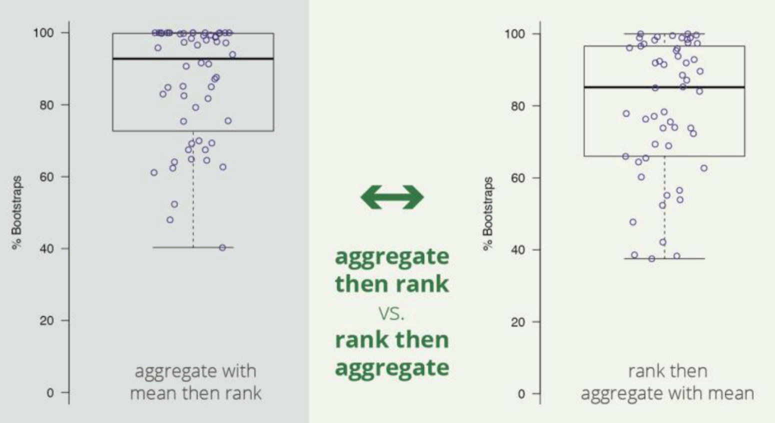


Figure 9: Robustness comparison aggregation methods **[Medic]**

Structure

Recommended best practices

- Incomplete reporting
  - $\rightarrow$ Instantiate full parameter list
- Low annotation quality
  - $\rightarrow$ Use multiple annotators
  - $\rightarrow$ Provide clear guidelines for annotations
- Suboptimal metric(s)
  - $\rightarrow$ Sound metric for task/challenge goal
  - $\rightarrow$ Be aware of biases
  - $\rightarrow$ Maybe check for ranking robustness
- Ranking and uncertainty
  - $\rightarrow$ Metric-based aggregation $>$ case based aggregation
  - $\rightarrow$ Mean $>$ Median
  - $\rightarrow$ Quantify the uncertainties, (annotations, rankings)
  - $\rightarrow$ Report inter-observer variability
  - $\rightarrow$ Perform bootstrapping to quantify ranking stability

Structure

Conclusion

- Competitions are very popular for benchmarking
- Winner might not be the best
- Consider other factors than challenge ranking
- Transparency is key
- Research for good standard practices needed
- Incentives to use these practices needed.

Thank You for Listening

Any Questions?

Bibliography

Hoffmann, Frank, Torsten Bertram, Ralf Mikut, Markus Reischl, and Oliver Nelles (2019). "Benchmarking in classification and regression". In: *WIREs Data Mining and Knowledge Discovery* 9.5, e1318. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/widm.1318. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/widm.1318.

Maier-Hein, Lena et al. (2018). "Is the winner really the best? A critical analysis of common research practice in biomedical image analysis competitions". In: *CoRR* abs/1806.02051. arXiv: 1806.02051. URL: http://arxiv.org/abs/1806.02051.