Multiple Annotations and Subjectivity in the Identification of Segment Boundaries in Music

Oriol Nieto
Morwaread M. Farbood

Toronto, ON
October 4th, 2014

MARL
NYU Music and Audio Research Laboratory

CogMIR™

Outline

- Music Segmentation Overview
- Exploring Subjectivity in Segment Boundaries
- Using Multiple References as Ground-Truth
- Conclusions and Discussion

MARL

Outline

- **Music Segmentation Overview**
- Exploring Subjectivity in Segment Boundaries
- Using Multiple References as Ground-Truth
- Conclusions and Discussion

MARL

Music Segmentation Overview

- Goal:
  - Automatically identify the different segments (or sections) of a musical piece.

- Example:

- Motivation:
  - Easier intra-piece navigation in music players.
  - Automatic generation of summaries and/or mash-ups.
  - Large-scale musicological research.

Music Segmentation Evaluation

- Compare estimated boundaries with *ground-truth* boundaries:
  - The *ground-truth* contains a single human annotation per track.
  - A metric (F-measure or Hit Rate) aims at quantifying how successful our algorithms are.

- MIR has been focusing on having large collections of audio data (see Million Song Dataset (Bertin-Mahieux et al. 2011)) and forgotten that we still compare our algorithms with collections that contain a **single** human annotation per track.

- (A SINGLE HUMAN ANNOTATION PER TRACK).

MARL

Overview

- Music Segmentation Overview
- **Exploring Subjectivity in Segment Boundaries**
- Using Multiple References as Ground-Truth
- Conclusions and Discussion

MARL

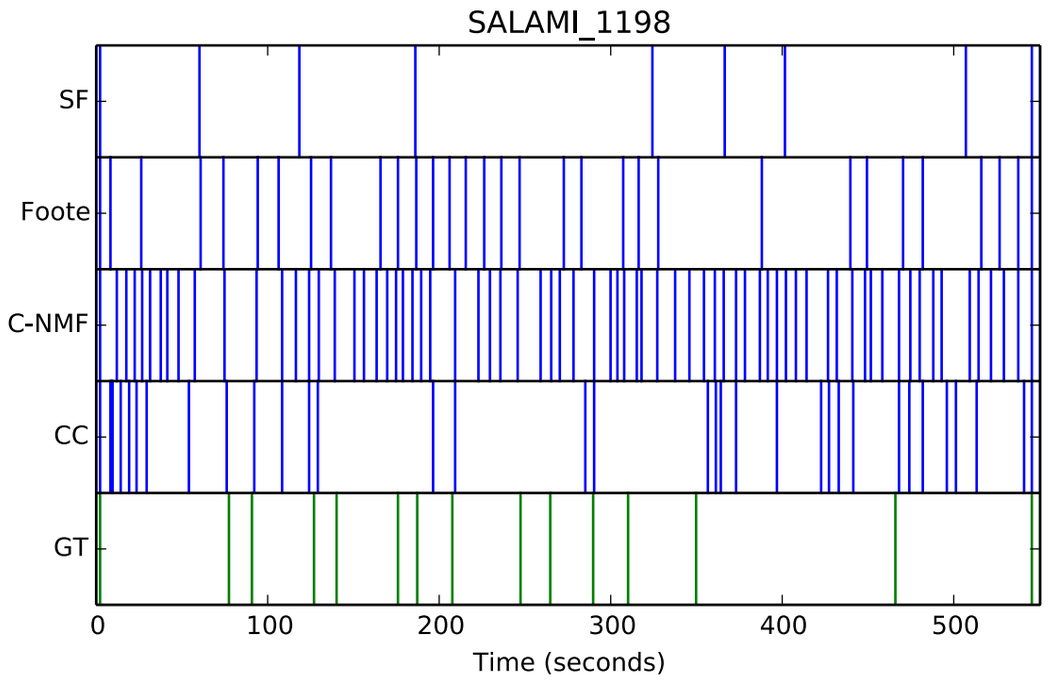
Sujectivity

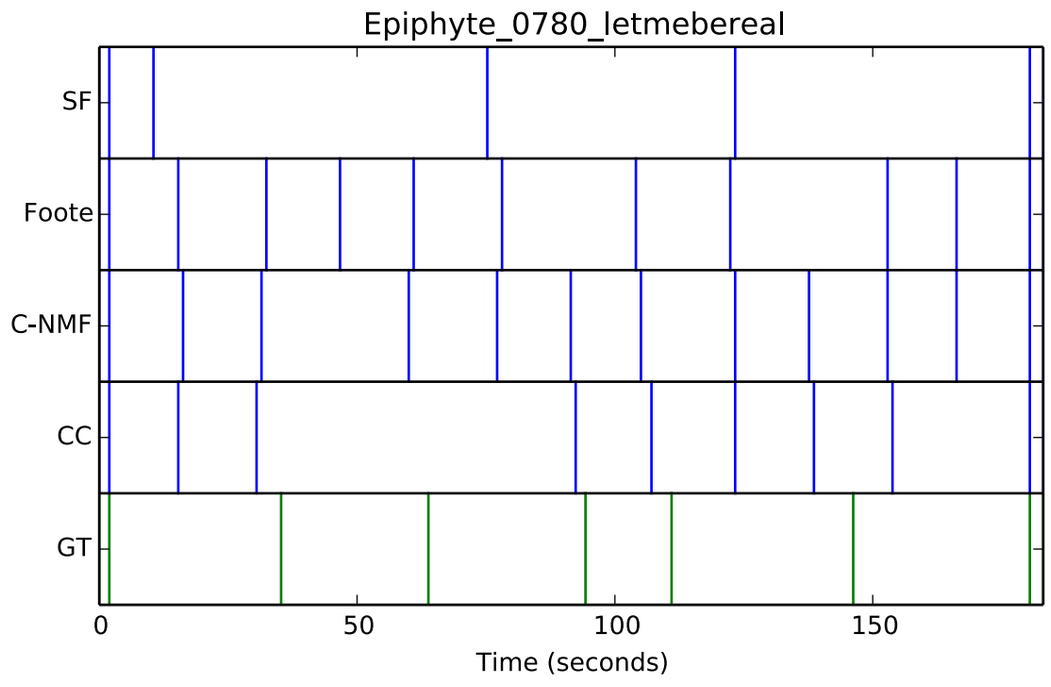
- It has been shown that the perception of segment boundaries in western popular music is highly subjective (Bruderer et al. 2009, Serrà et al. 2014).

- We want to:
  - Show that the notion of *ground-truth* annotated by a single human is prone to error.
  - Merge multiple human segment boundary annotations to obtain more robust *ground-truths*.

Selecting Tracks

- From a large collection of >2,000 human annotated tracks:
  - Run multiple boundary retrieval algorithms.
  - Rank them based on a standard evaluation metric (F-measure with a 3 seconds window).
  - Choose the 45 worst performing tracks (i.e. challenging from a machine point of view).
  - Choose the 5 best performing tracks (i.e. trivial from a machine point of view).

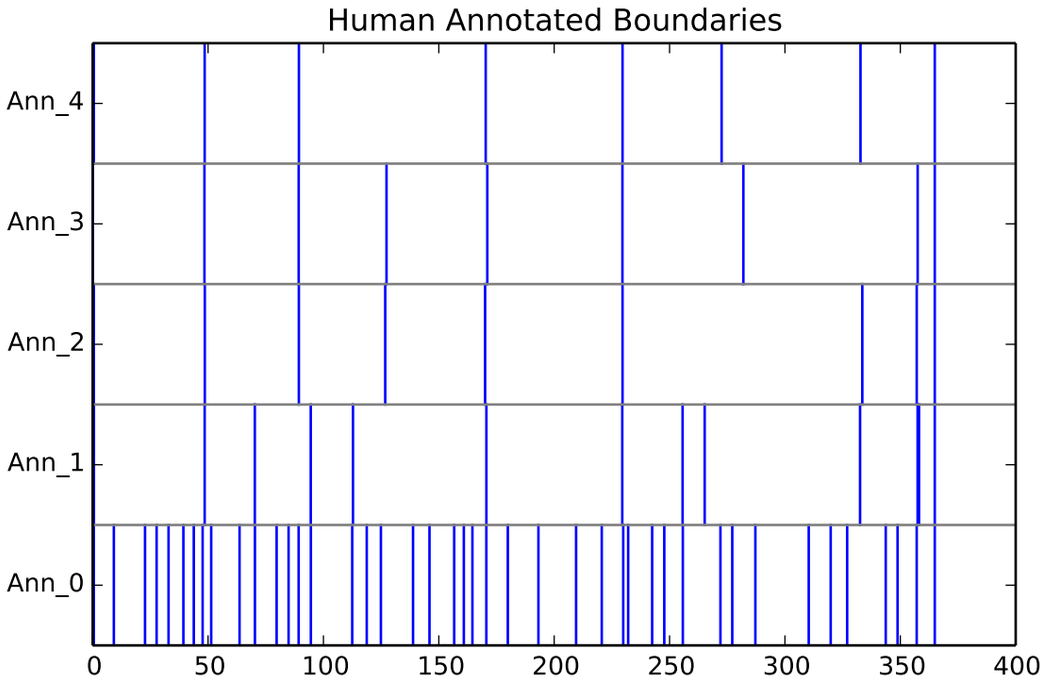




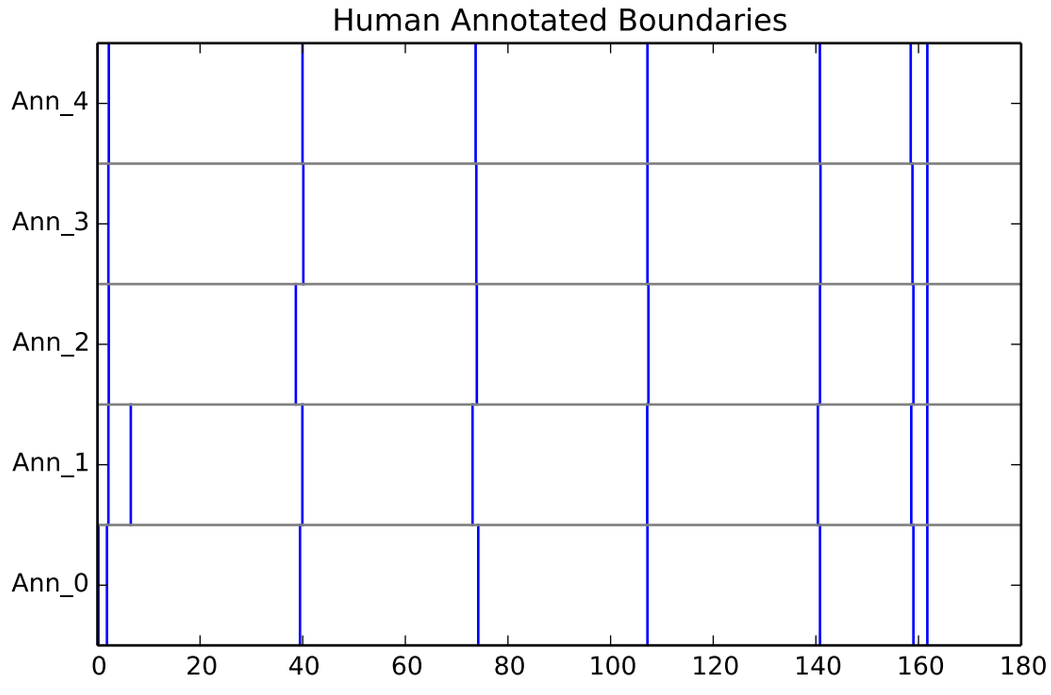
MARL

Collecting Multiple Segment Annotations

- We asked 5 music experts to annotate the 50 selected tracks.
  - Two levels of segmentation: large and small.
- Each track will now contain five additional two-layer segmentation annotations.
- Capture Subjectivity by exploring the variability of the new annotations.



Track from Challenging Group



Track from Control Group

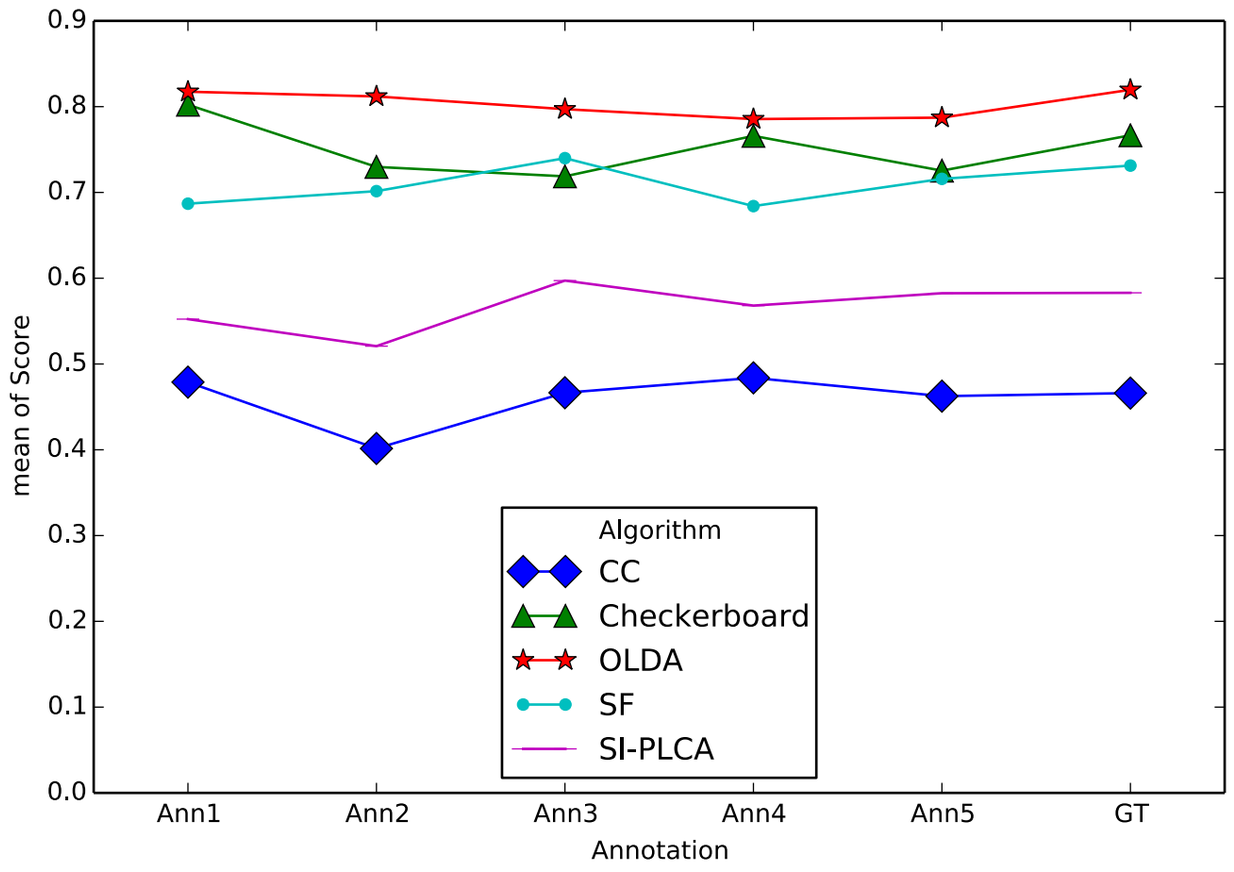
MARL

Analysis of Subjectivity

- We want to analyze the variation of the scores when evaluating the estimated boundaries with the new annotations.

- Use a 2-way ANOVA of average F-measure with algorithms and annotations as factors.

- Start with the control group:



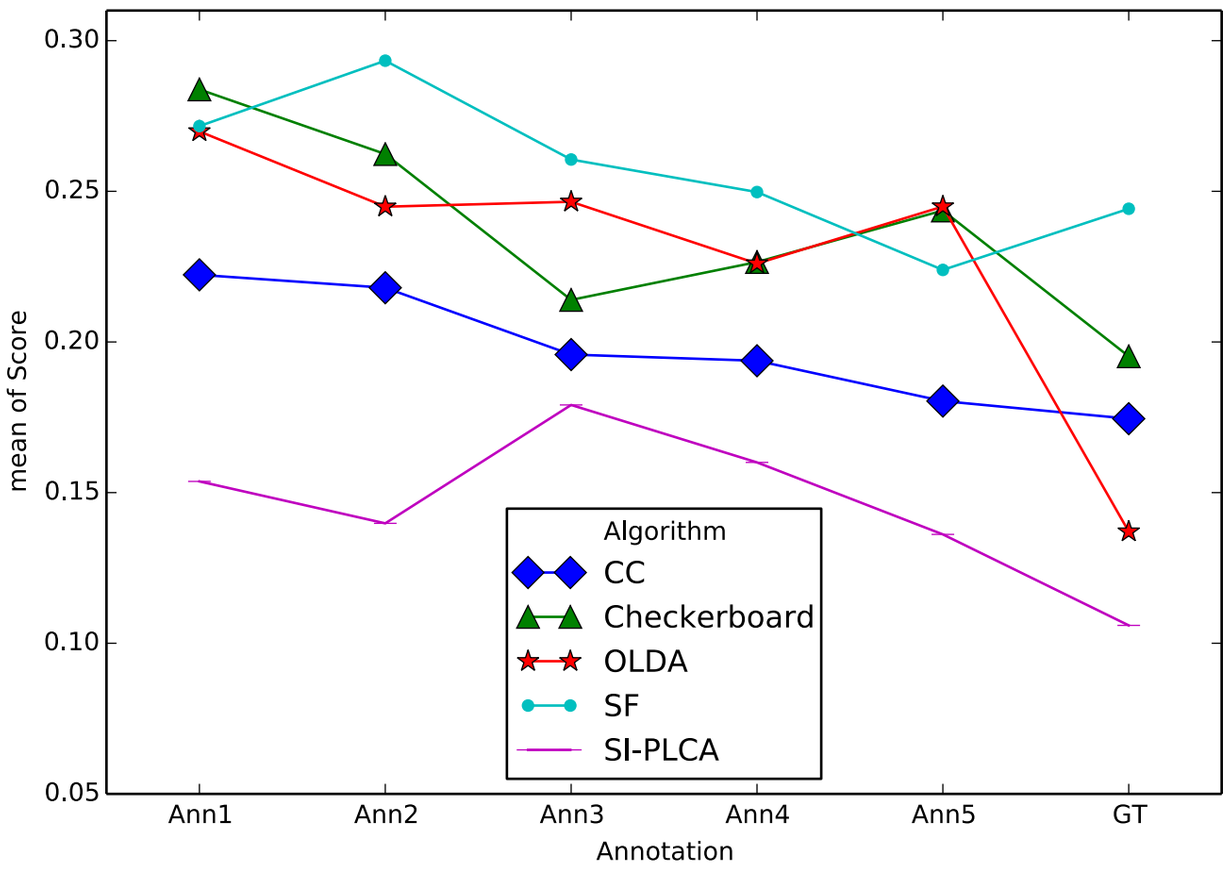
Annotations Effect:
F (5, 120) = .22, p = .95

Interaction:
F(20,120) = .13,p = .99

MARL

Analysis of Subjectivity

- No significant variation for the control group when using different annotations.
- What about the challenging group?



Annotations Effect:
$F(5, 1320) = 6.93, p < .01$

Interaction:
$F(20, 1320) = 1.13, p = .3$

MARL

Analysis of Subjectivity

- **Significant variation** when using different annotations for the challenging tracks.
- Therefore:
  - Subjectivity is a relevant problem when evaluating music boundaries.
  - At least on the challenging tracks.
- Can we minimize the subjectivity effect for this task?

Overview

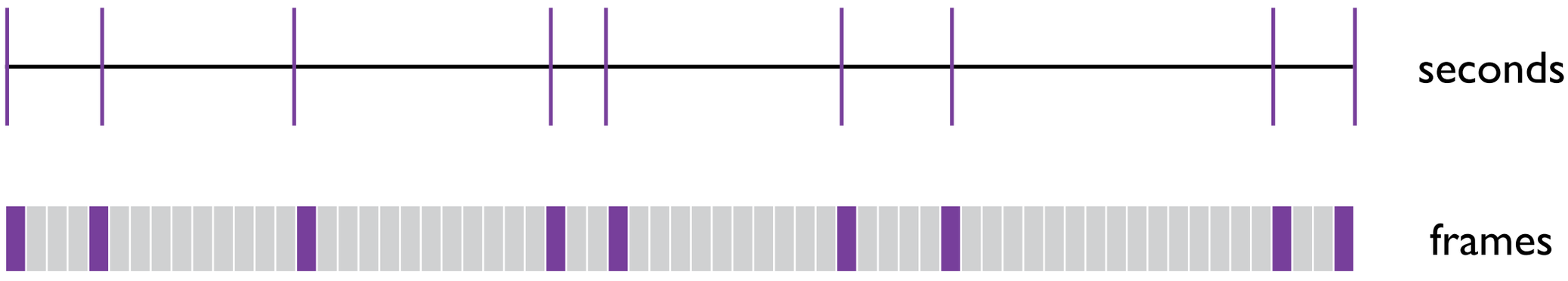
- Music Segmentation Overview
- Exploring Subjectivity in Segment Boundaries
- **Using Multiple References as Ground-Truth**
- Conclusions and Discussion

MARL

Merging Boundaries

- Idea: use all the annotated boundaries as references to overcome differences in perception.
  - Merge them

- How?
  - First, discretize the boundary times.



MARL

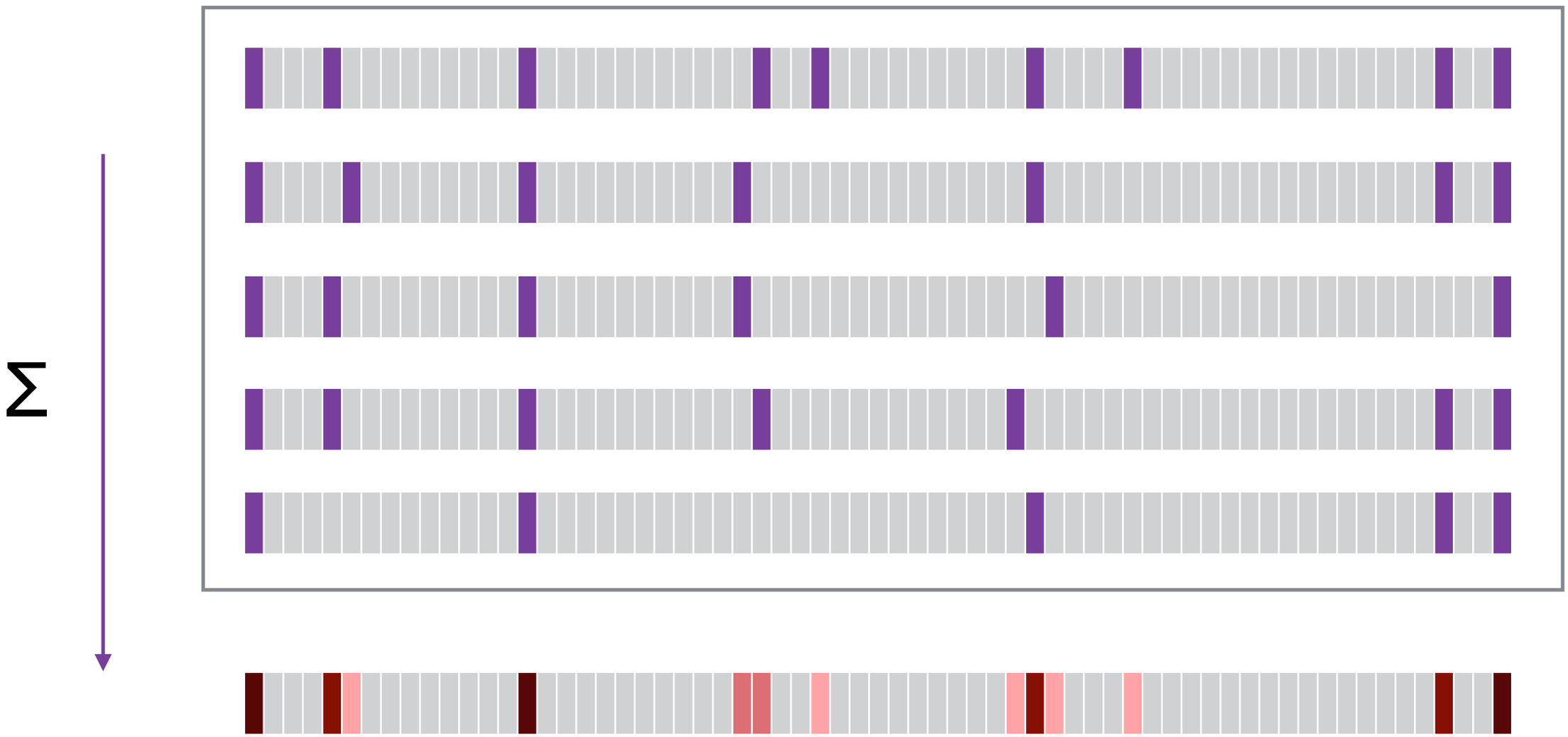
Merging Boundaries

- Various ways of merging the new annotated boundaries:
  - Type I: Flat to Flat
  - Type II: Hierarchical to Flat
  - Type III: Flat to Hierarchical
  - Type IV: Hierarchical to Hierarchical

MARL

Merging Type I: Flat to Flat

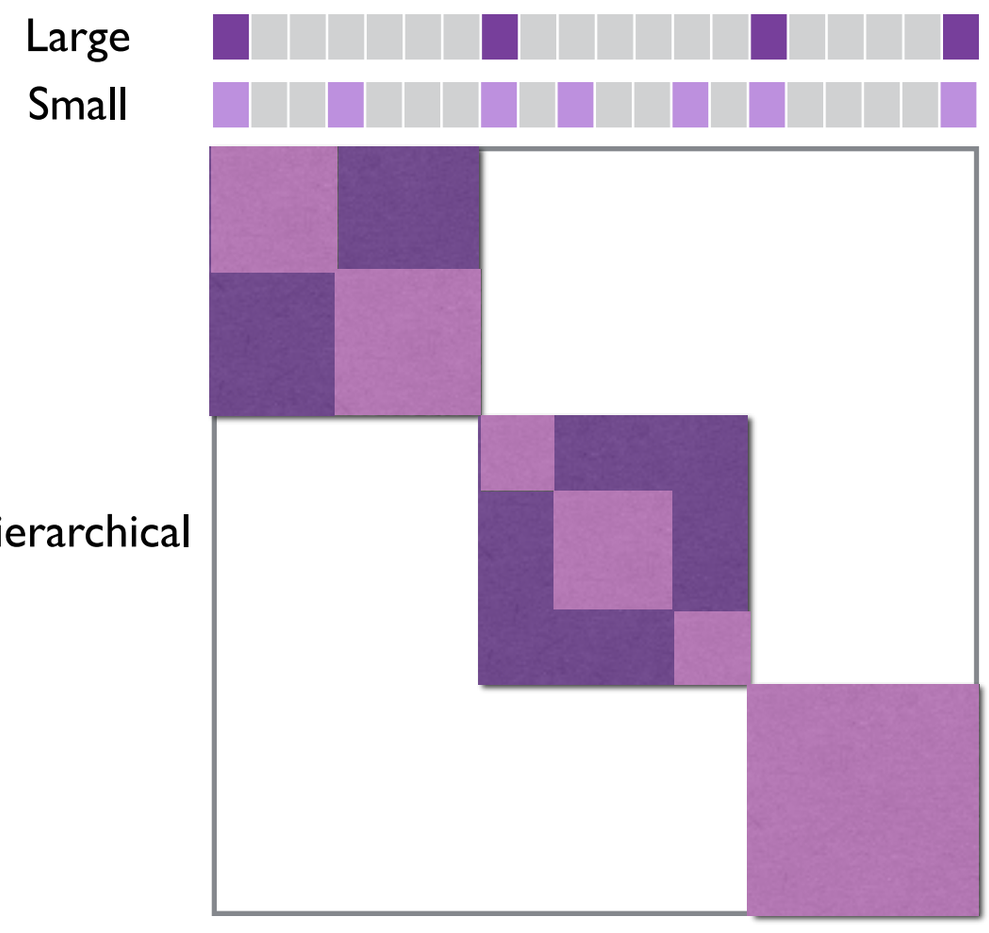
- Take the mean of the discretized boundaries
- Now we have a weight for each boundary (normalized to 1)



MARL

Merging Type II: Hierarchical to Flat

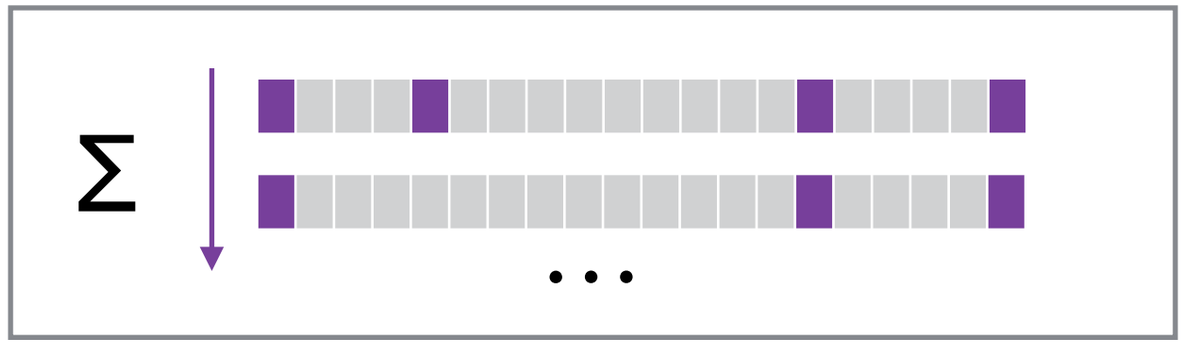
- Treat the large and small scale levels as hierarchical.
  - The large scale boundaries are always a subset of the small scale ones.
- Take the average as in Type I.

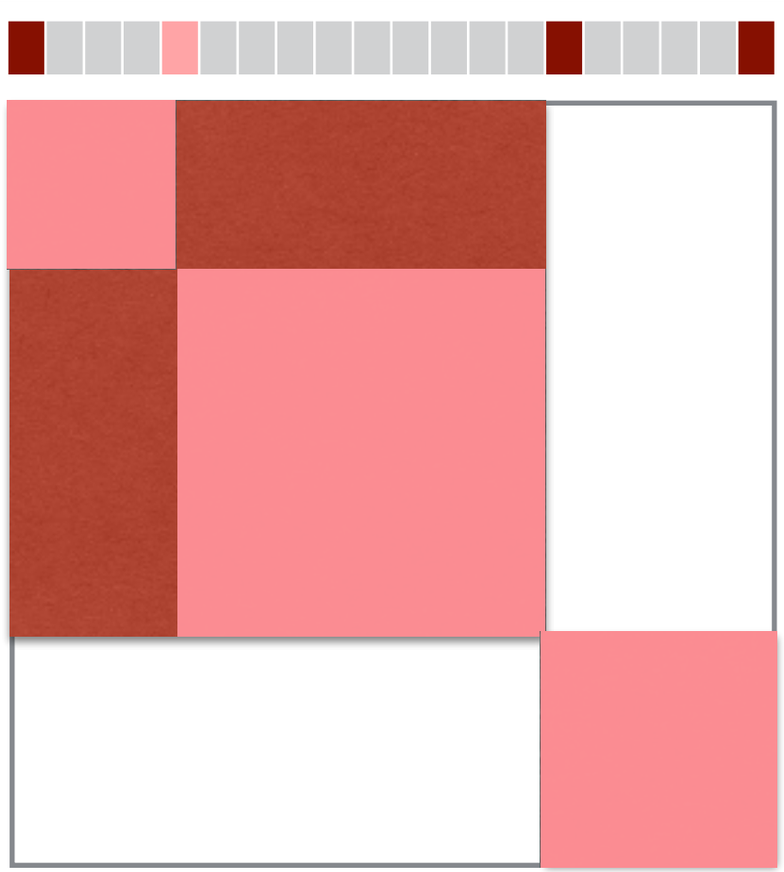


MARL

Merging Type III: Flat to Hierarchical

- Merge down flat boundaries as in Type I (flat to flat).
- Transform the weighted boundaries to a hierarchical annotation.
  - Each unique weight creates a new layer.

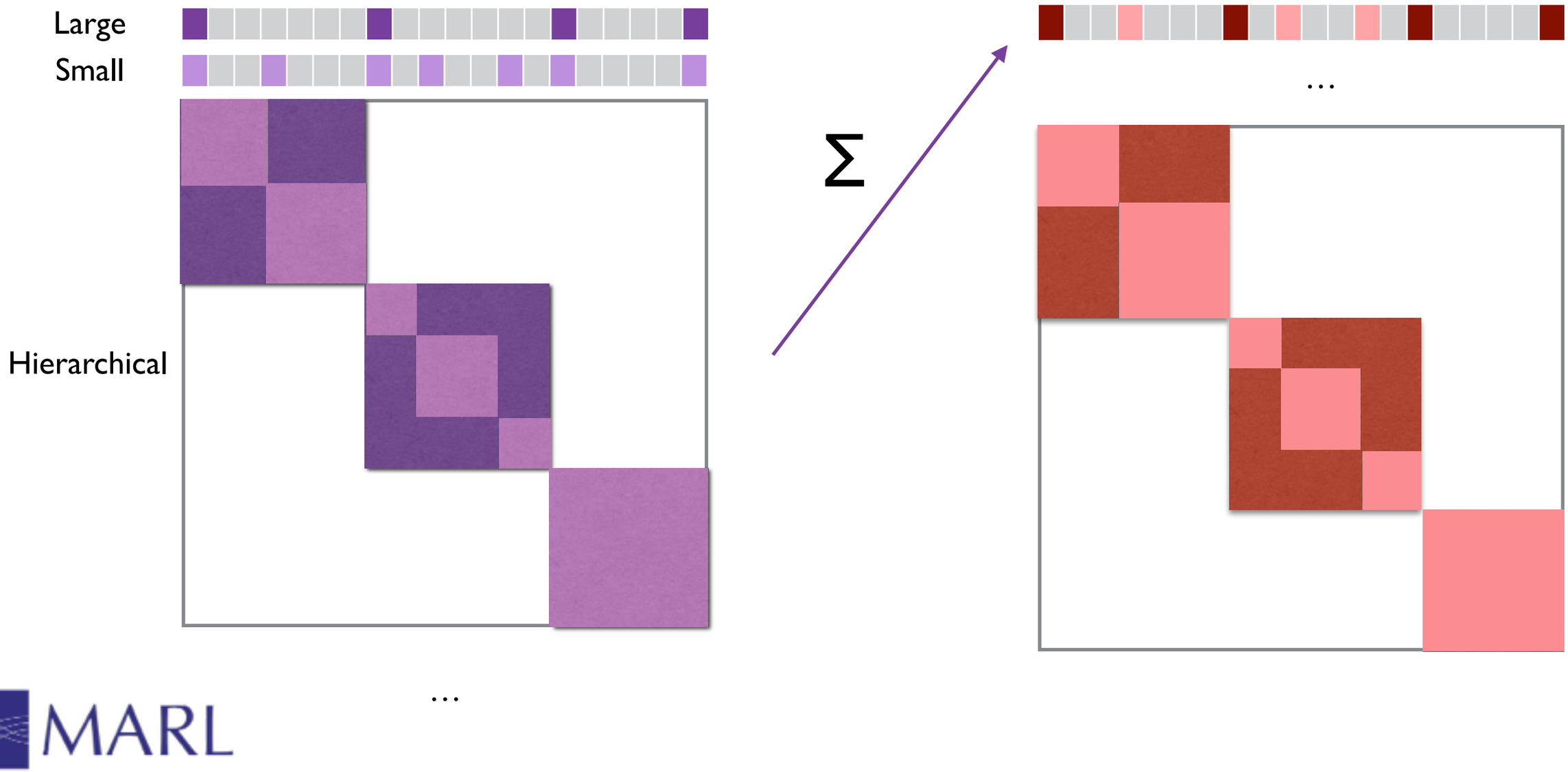




MARL

Merging Type VI: Hierarchical to Hierarchical

- Merge down the hierarchical annotations like in Type II (hierarchical to flat).
- Build hierarchy based on the new weighted annotations like in Type III.



MARL

Robustness of Merged Boundaries

- In order to test the robustness of this aggregation, we divide annotations into sets of 3:
  - 5 annotators, dividing them into sets of 3. ($\binom{5}{3} = 10$)
  - Similar to cross-validation

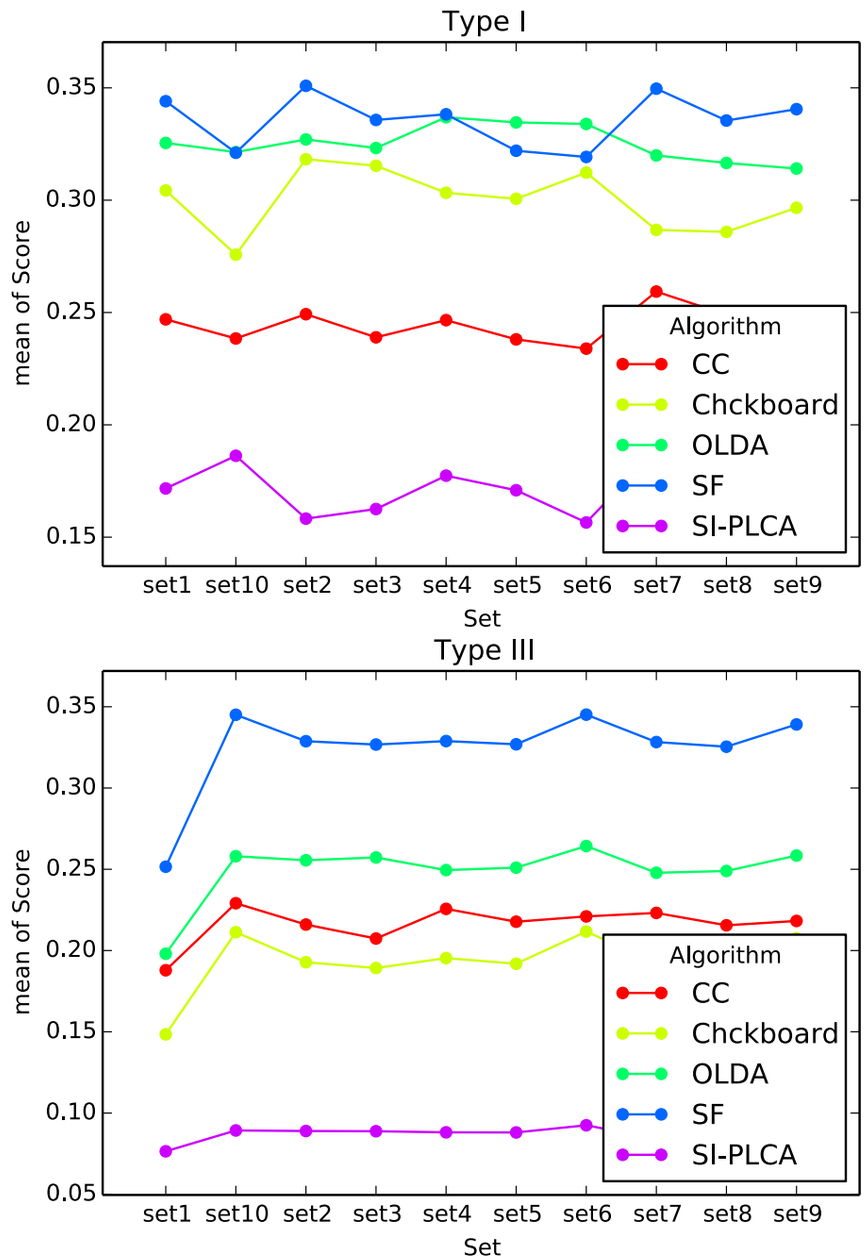
- For each of 10 sets, we merge their annotations using the four different types (types I, II, III and IV).
- For each type, compute two-way ANOVA with algorithm and sets as factors.
  - Aim to obtain similar results to the ones of the control group.

MARL

Robustness of Merged Boundaries

Sets of 3

Merge Type	$F(9, 2200)$	p-value
I	.23	.99
II	.42	.92
III	3.35	$< .01$
IV	1.56	.12



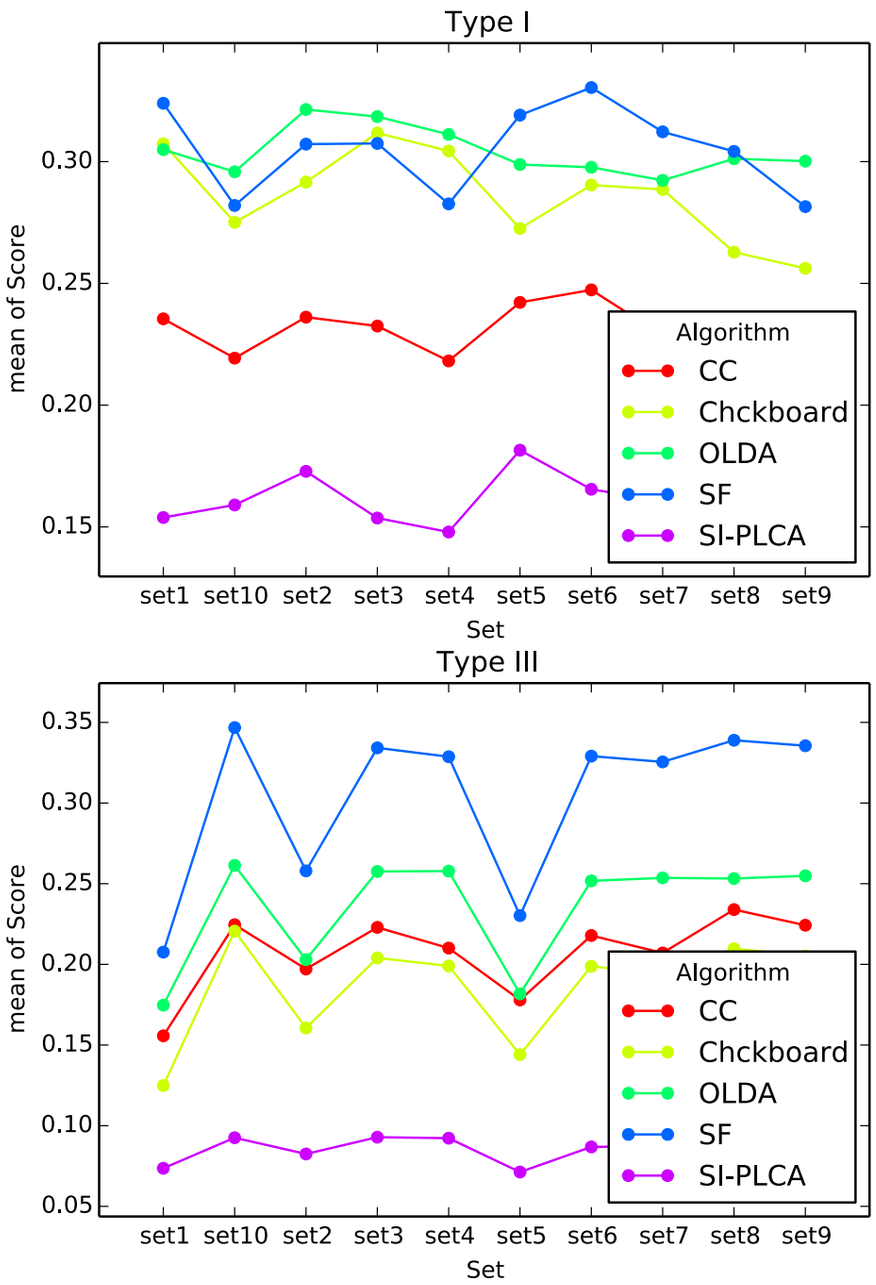
- Except type III none of the scores significantly vary depending on the set chosen.
- No conflicts in marginal means in types III and IV.

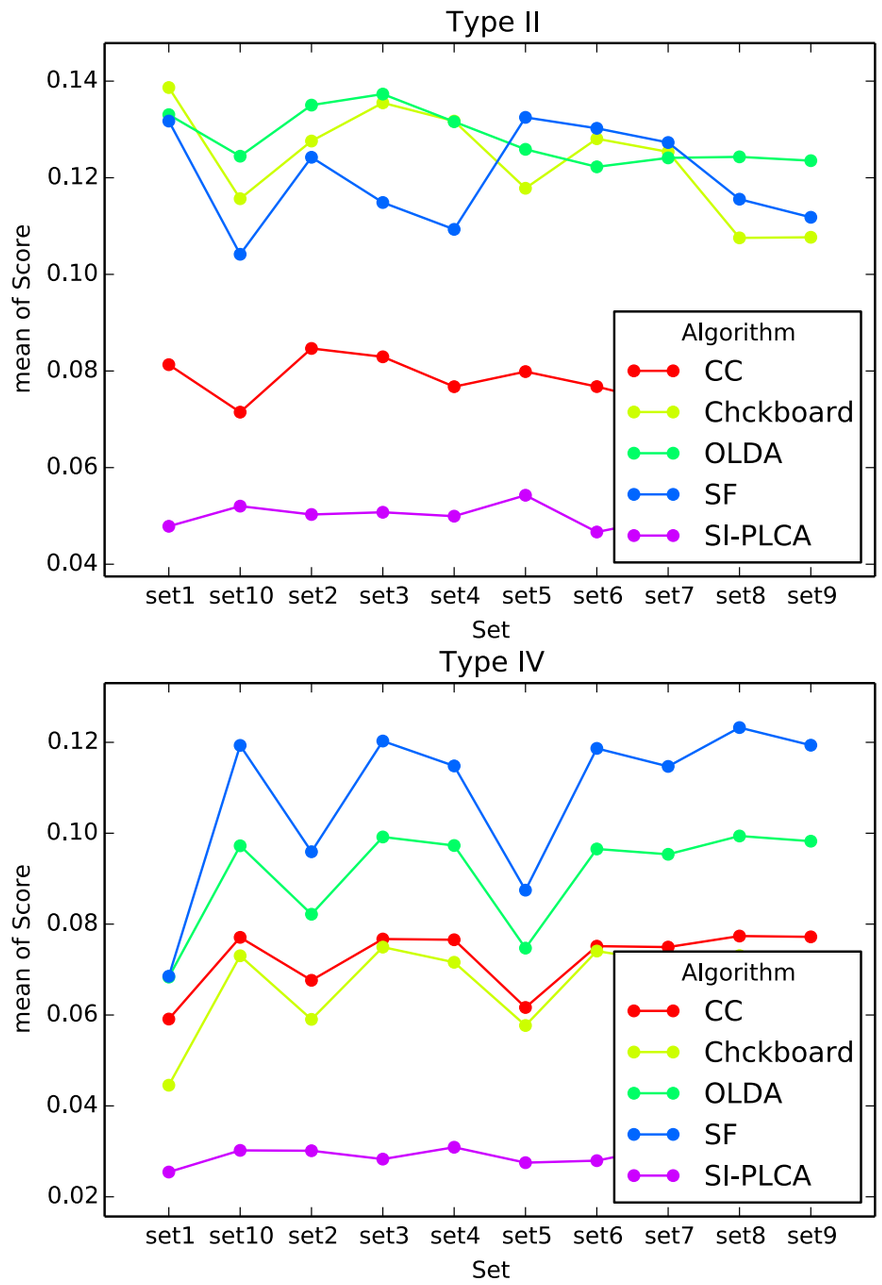
MARL

Robustness of Merged Boundaries

Sets of 2

Merge Type	$F(9, 2200)$	p-value
I	.68	.71
II	.97	.46
III	12.71	$< .01$
IV	7.35	$< .01$





- Types I and II do not significantly vary depending on the set chosen.
- No conflicts in marginal means in types III and IV.

MARL

Overview

- Music Segmentation Overview
- Exploring Subjectivity in Segment Boundaries
- Using Multiple References as Ground-Truth
- **Conclusions and Discussion**

MARL

Conclusions

- *Ground-truth* with a single human annotation per track is prone to error.
- Subjectivity is a significant problem when evaluating music boundaries of challenging tracks.
- Merging annotations can significantly alleviate the subjectivity problem:
  - 4 types of merging.
  - Types I and II do not statistically vary (like in the control group).
  - Types III and IV seem to vary consistently (i.e. they may be reliable as well).
  - Ground-truth of two different boundary annotations already more robust (but three seem more robust than two).
- Subjectivity might not be a problem on the simpler tracks.

Open Source

- Replicate these results!
- All source code in:
  - https://github.com/urinieto/msaf/

References

- Bruderer, M. J., Mckinney, M. F., & Kohlrausch, A. (2009). The Perception of Structural Boundaries in Melody Lines of Western Popular Music. Musicæ Scientiæ, 13(2), 273–313.
- Levy, M., & Sandler, M. (2008). Structural segmentation of musical audio by constrained clustering. IEEE Transactions on Audio, Speech, and Language Processing, 16(2), 318–326. doi:10.1109/TASL.2007.910781
- Ni, Y., McVicar, M., Santos-Rodriguez, R., & De Bie, T. (2013). Understanding Effects of Subjectivity in Measuring Chord Estimation Accuracy. IEEE Transactions on Audio, Speech, and Language Processing, 21(12), 2607–2615. doi:10.1109/TASL.2013.2280218
- Nieto, O., & Jehan, T. (2013). Convex Non-Negative Matrix Factorization For Automatic Music Structure Identification. In Proc. of the 38th IEEE International Conference on Acoustics Speech and Signal Processing (pp. 236–240). Vancouver, Canada.
- Serrà, J., Müller, M., Grosche, P., & Arcos, J. L. (2014). Unsupervised Music Structure Annotation by Time Series Structure Features and Segment Similarity. IEEE Transactions on Multimedia, Special Issue on Music Data Mining, 16(5), 1229 – 1240. doi:10.1109/TMM.2014.2310701
- Smith, J. B., Burgoyne, J. A., Fujinaga, I., De Roure, D., & Downie, J. S. (2011). Design and Creation of a Large-Scale Database of Structural Annotations. In Proc. of the 12th International Society of Music Information Retrieval (pp. 555–560). Miami, FL, USA.
- Weiss, R., & Bello, J. P. (2011). Unsupervised Discovery of Temporal Structure in Music. IEEE Journal of Selected Topics in Signal Processing, 5(6), 1240–1251.

Questions?

- Subjectivity is a significant problem in music segmentation.
- Merging annotations alleviates the subjectivity problem.
- Source code: https://github.com/urinieto/msaf/

MARL