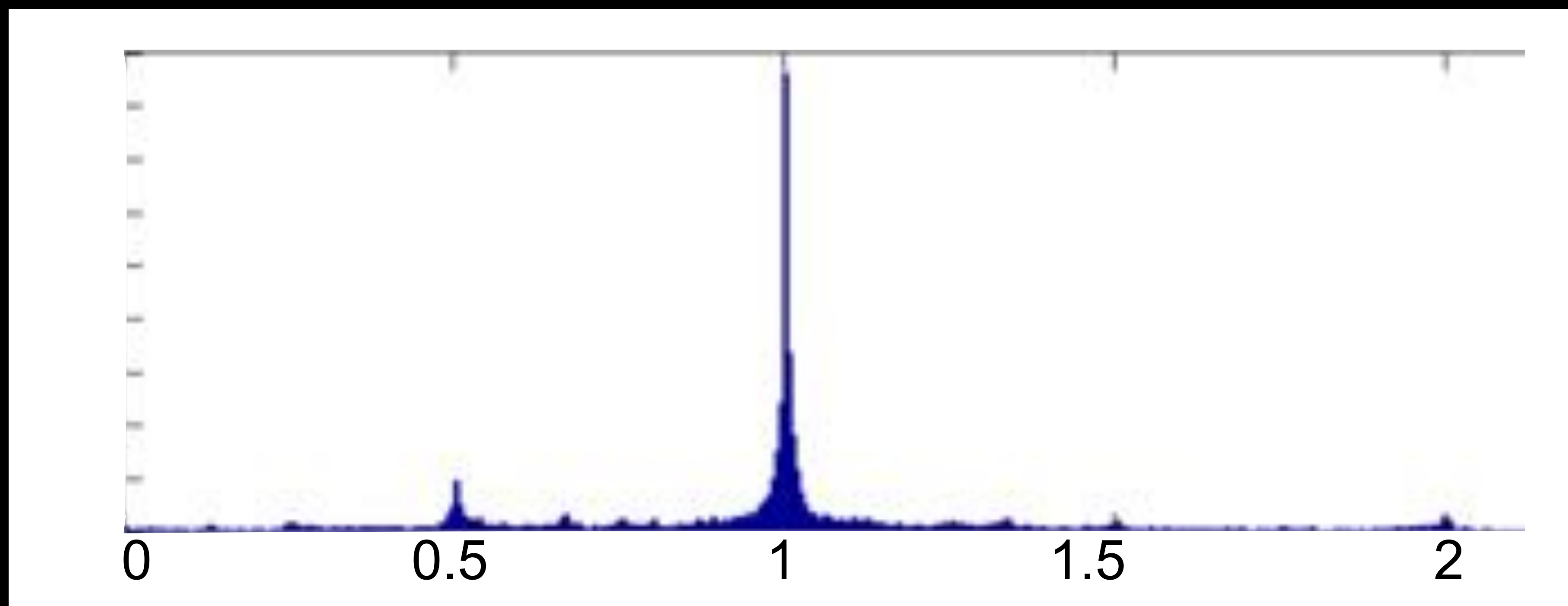
Using prior expectations to improve structural analysis: A cautionary tale

Jordan B. L. Smith
Masataka Goto
National Institute of Advanced Industrial Science and Technology, Japan

Dagstuhl Seminar Stimulus Talk
3 March 2016

ISMIR 2014 tutorial on music structure analysis & evaluation



Histogram of ratio between a song's median segment length and the length of all of its segments

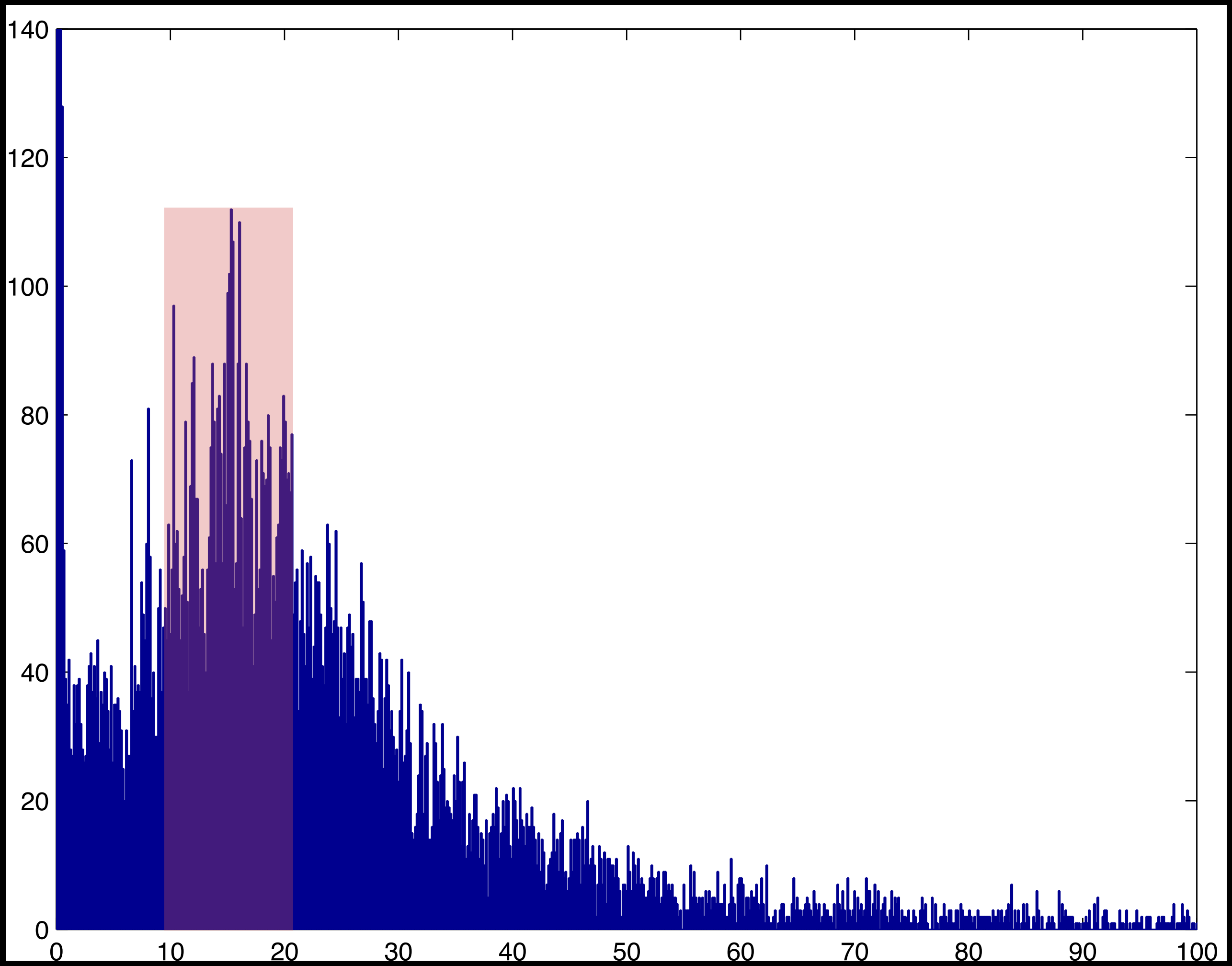
See also Bimbot et al. 2014: "Semiotic Description of Music Structure: an Introduction to the Quaero/Metiss Structural Annotations." AES

Regularities

- In what ways are annotations constrained?
- Can we use these regularities to do improve a structural analysis algorithm?

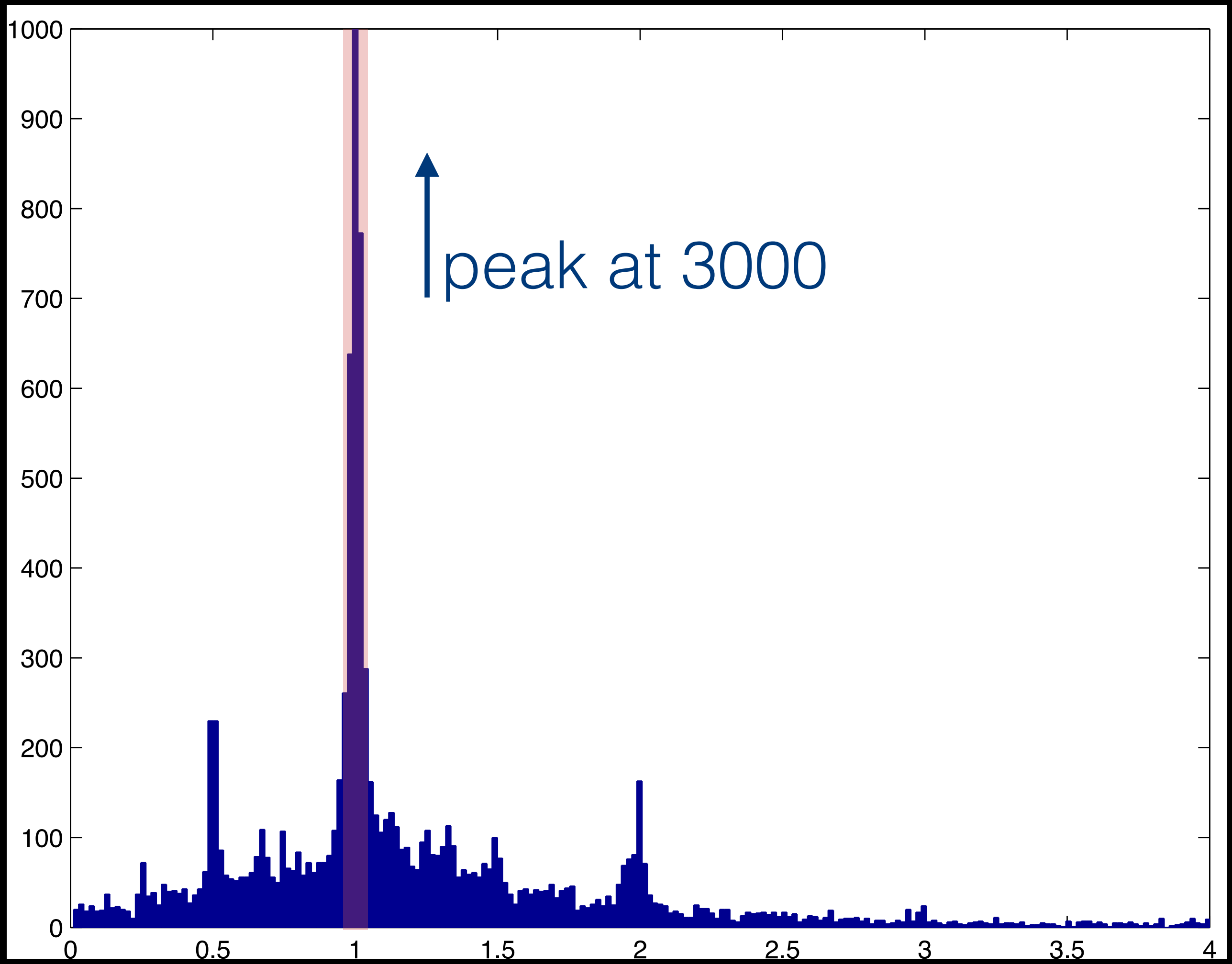
Segment lengths are not independent

Distribution of absolute segment lengths (s)



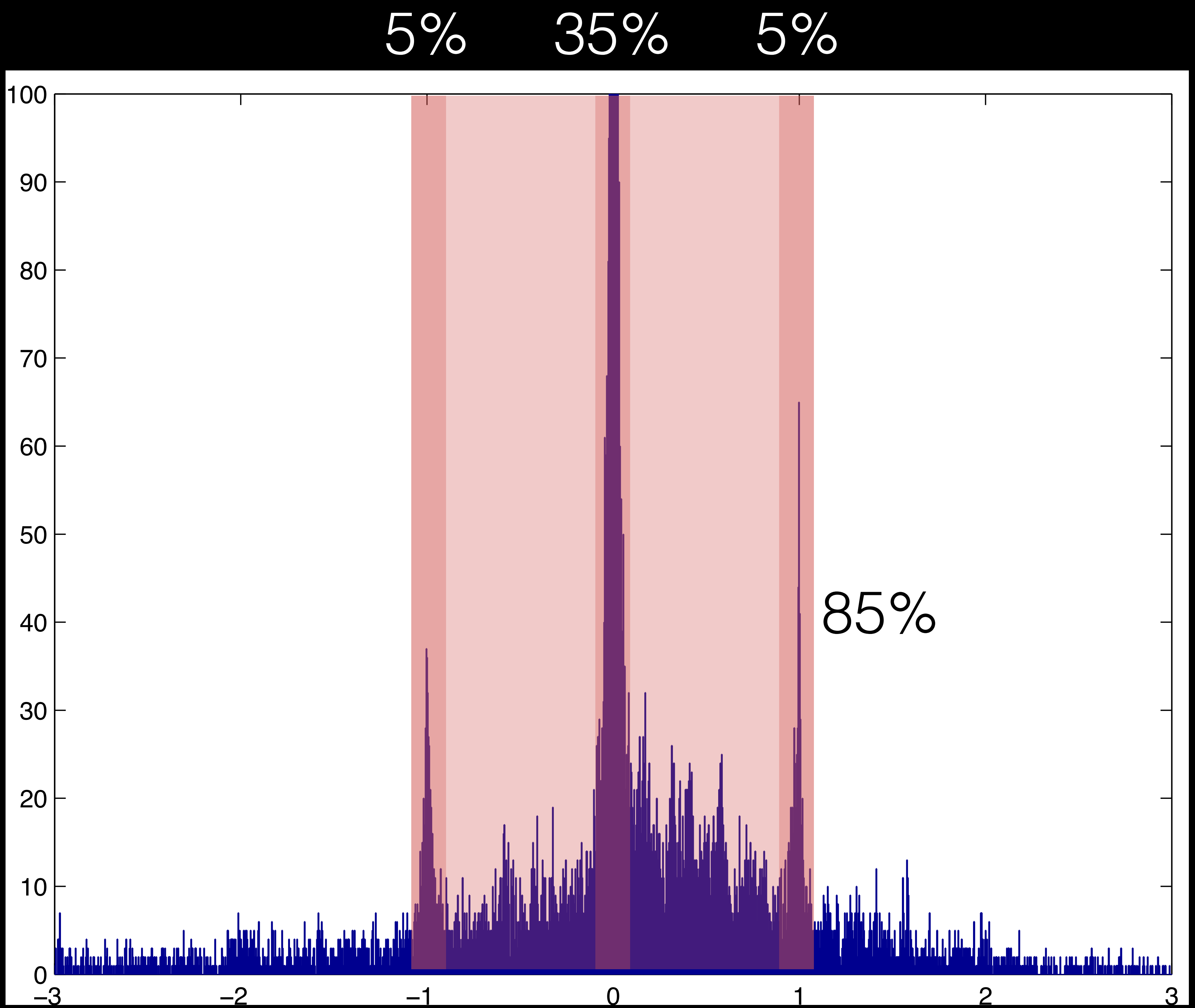
35% of data between 9.5 and 21 seconds

Distribution of median-scaled segment lengths



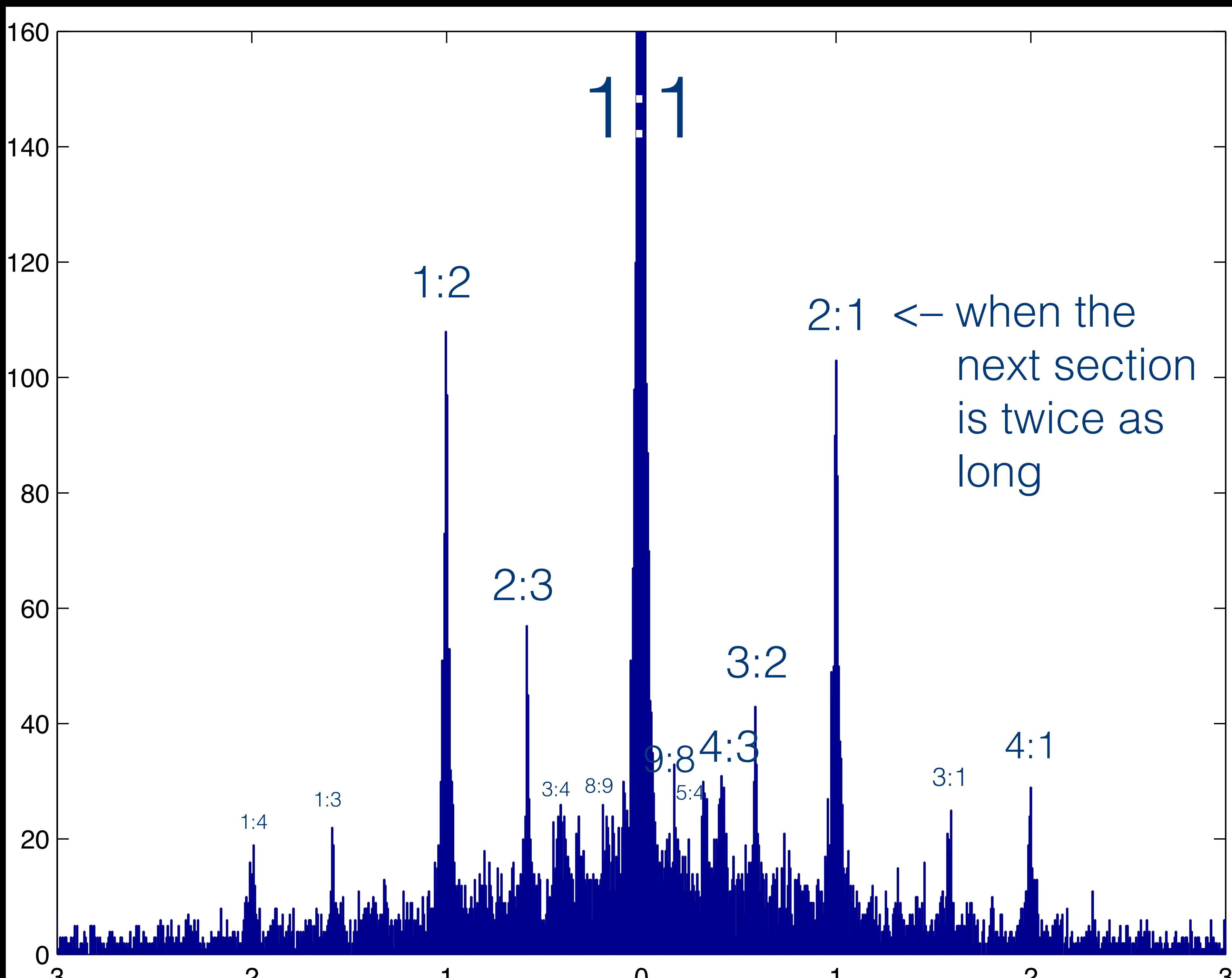
35% of data between 0.96 and 1.04

Distribution of median-scaled segment lengths



Distribution of expected length of next section given previous section

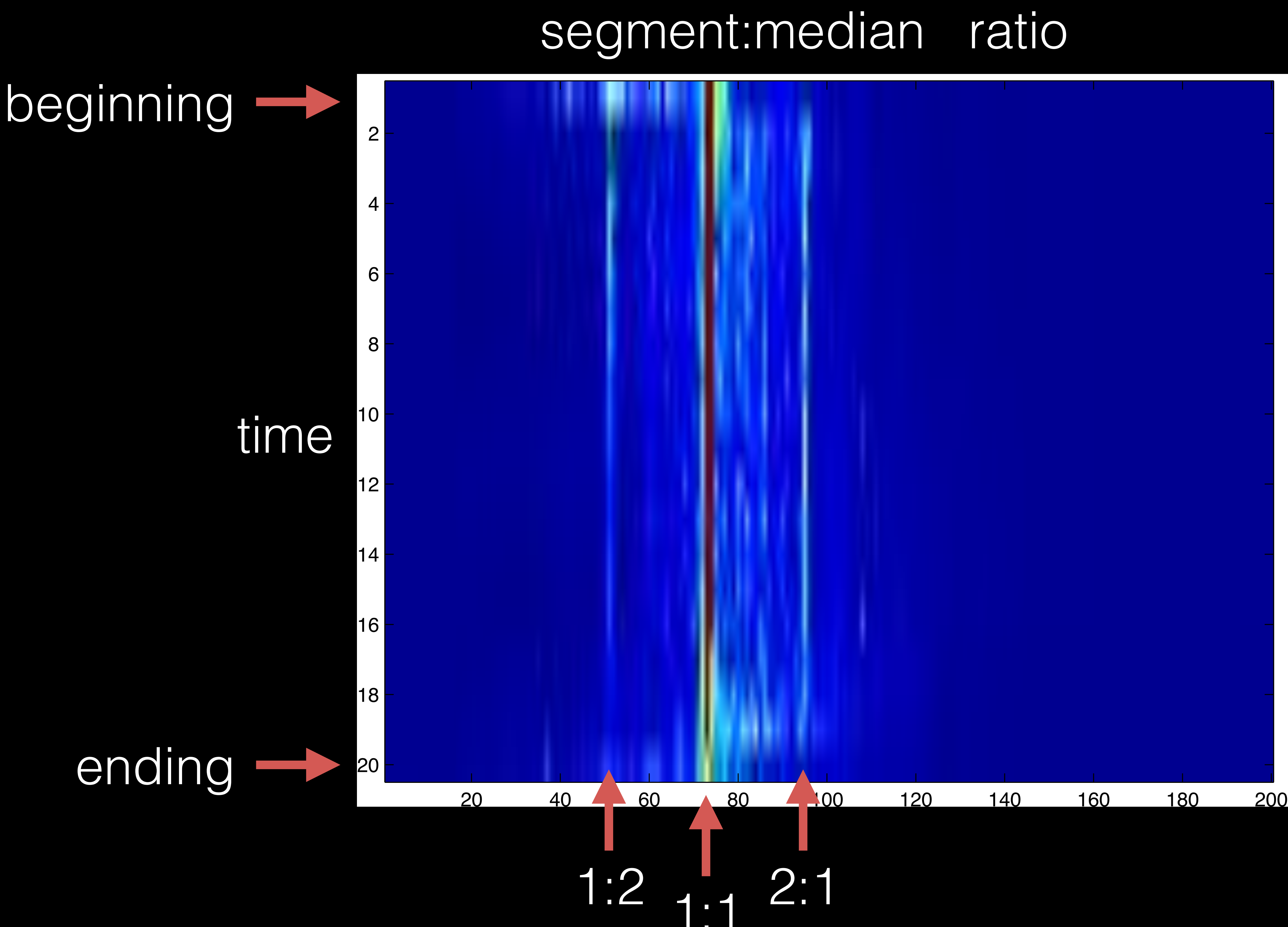
$L_{next}:L_{prev}$



- Conclusion:
  Segment lengths are not independent of each other, neither globally nor sequentially

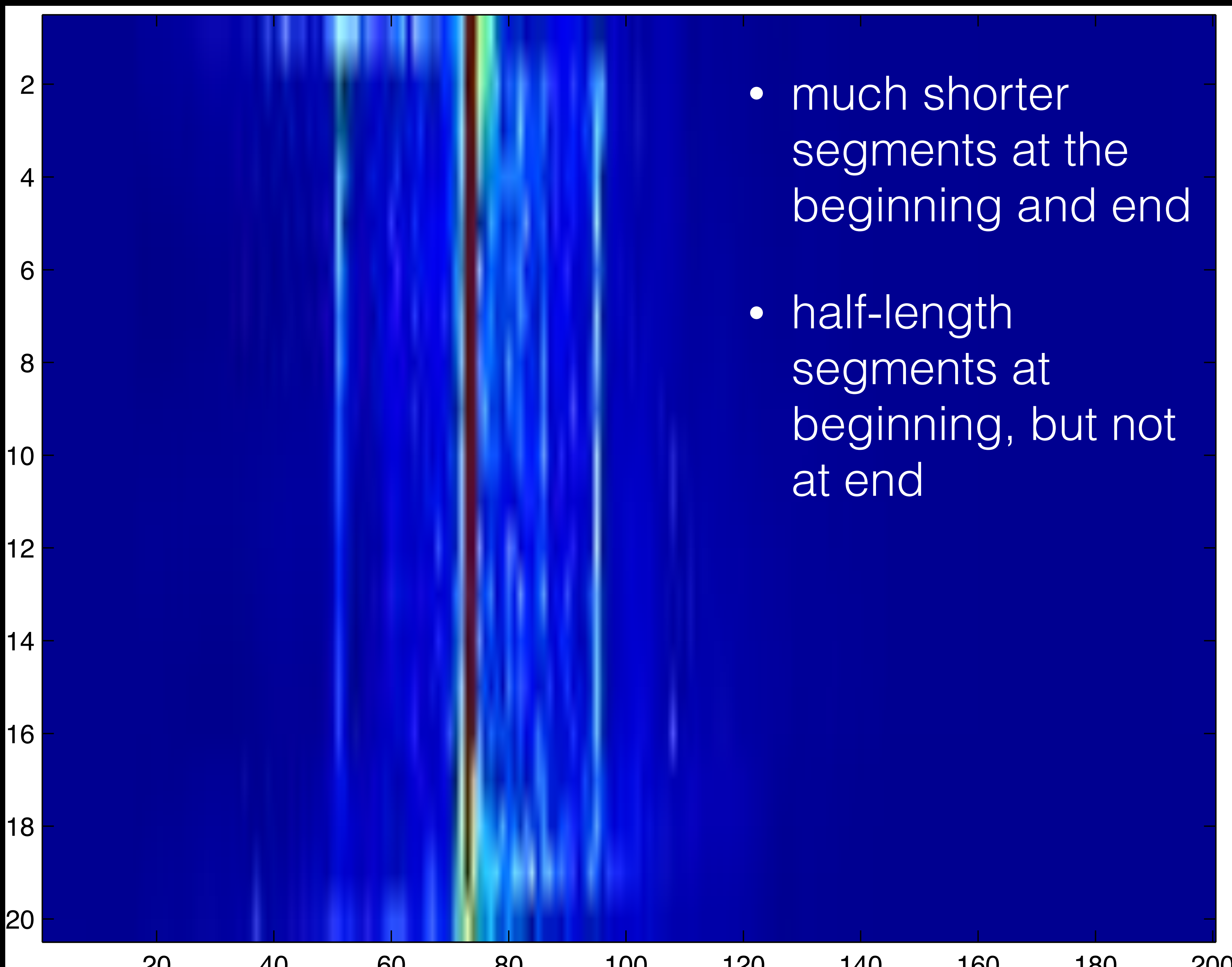
- Next:
  Are they independent of when they occur in a piece of music?

Dependence on location within a piece



Dependence on location within a piece

segment:median ratio



- Conclusion:
  Segment lengths are not independent of each other, neither globally nor sequentially

- Second-order priors may also be relevant (e.g., length of section depending on whether it is the first segment or a segment in the middle)

Prior art

- Turnbull et al. 2007 used a plethora of difference features, and trained a Boosted Decision Stumps learner to classify audio frames as boundaries or not

- Sargent, Bimbot and Vincent 2011 use a Viterbi algorithm which includes "segment length" as a criterion; optimal length could have been set by corpus, but was set by hand. However, Rodriguez-Lopez, Volk and Bountoridis 2014 expand on the algorithm and include many corpus-set priors.

- McFee et al. 2014 used annotations to optimise their feature representation (they devise a transformation that minimises the variance of the feature within segments and maximises it between), then used a standard approach

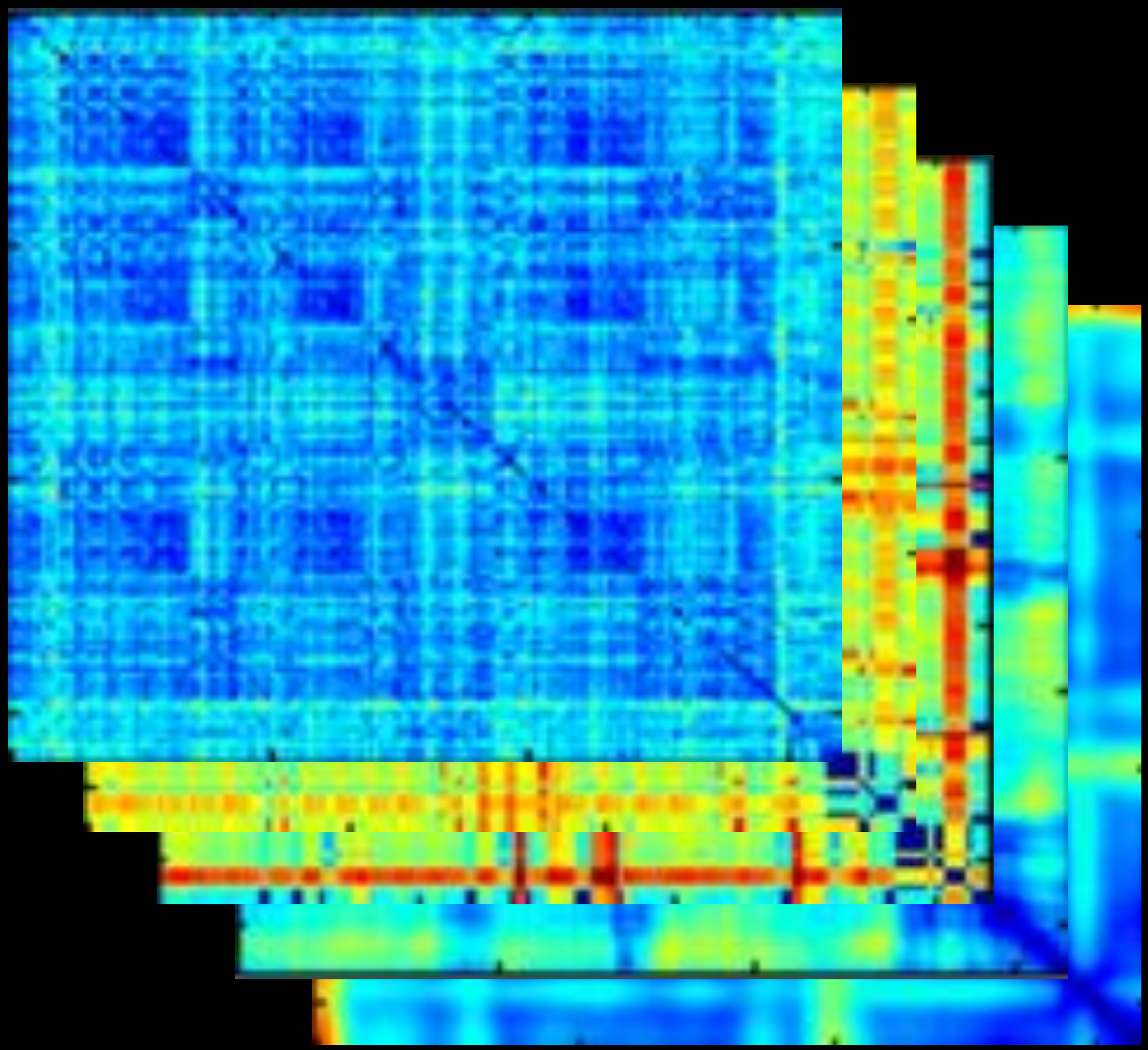
- Ullrich et al. 2014 used convolutional neural nets to do roughly the same thing as Turnbull et al., and using no prior musicological insight, they are now the best-performing segmentation algorithm at MIREX

Application

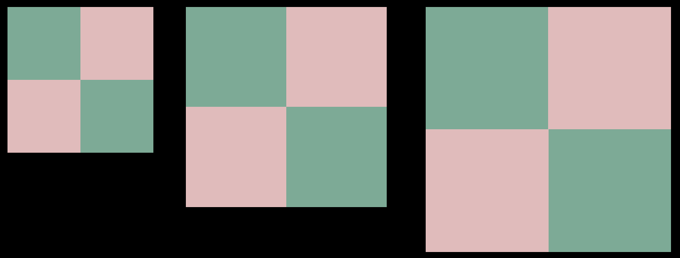
- How can we use this?
    - Use Foote’s novelty approach, multiple times
    - Use fitness to prior distributions to choose which solution to trust

How can we use this?

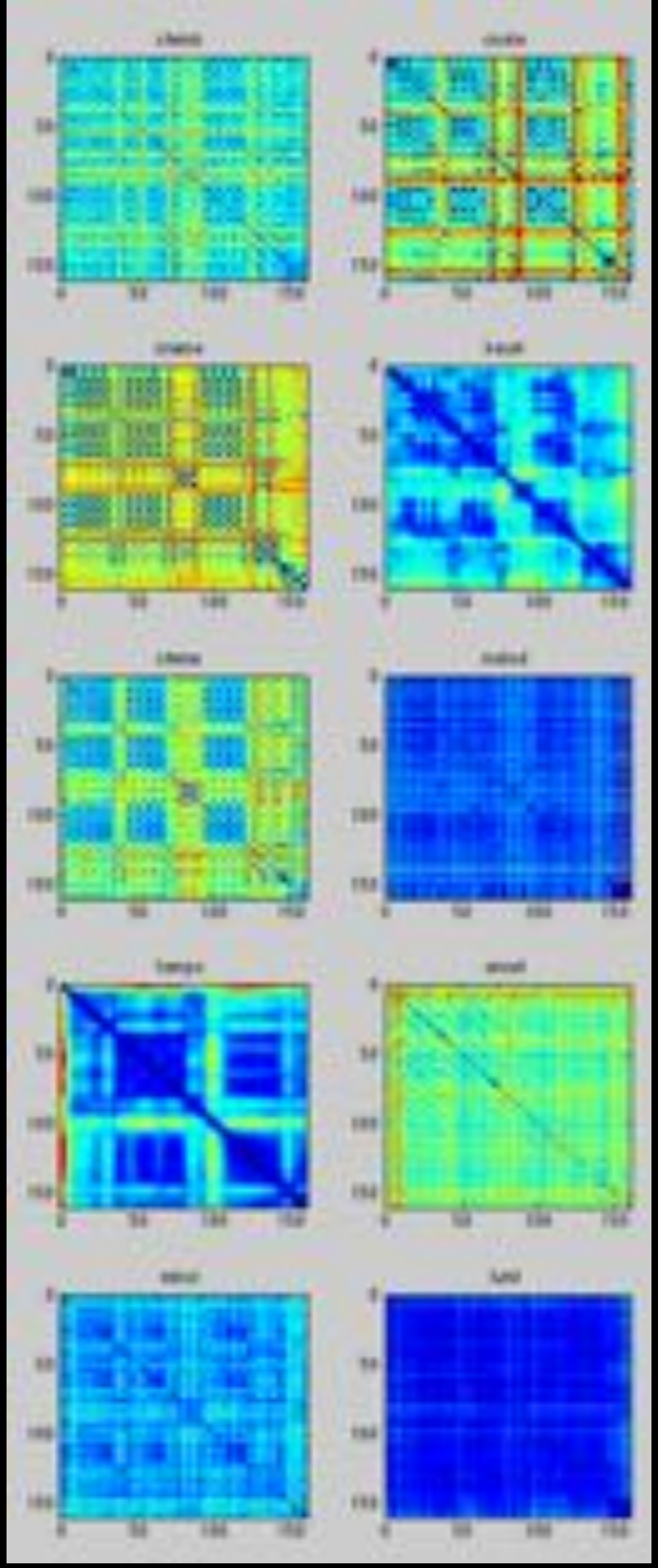
Feature



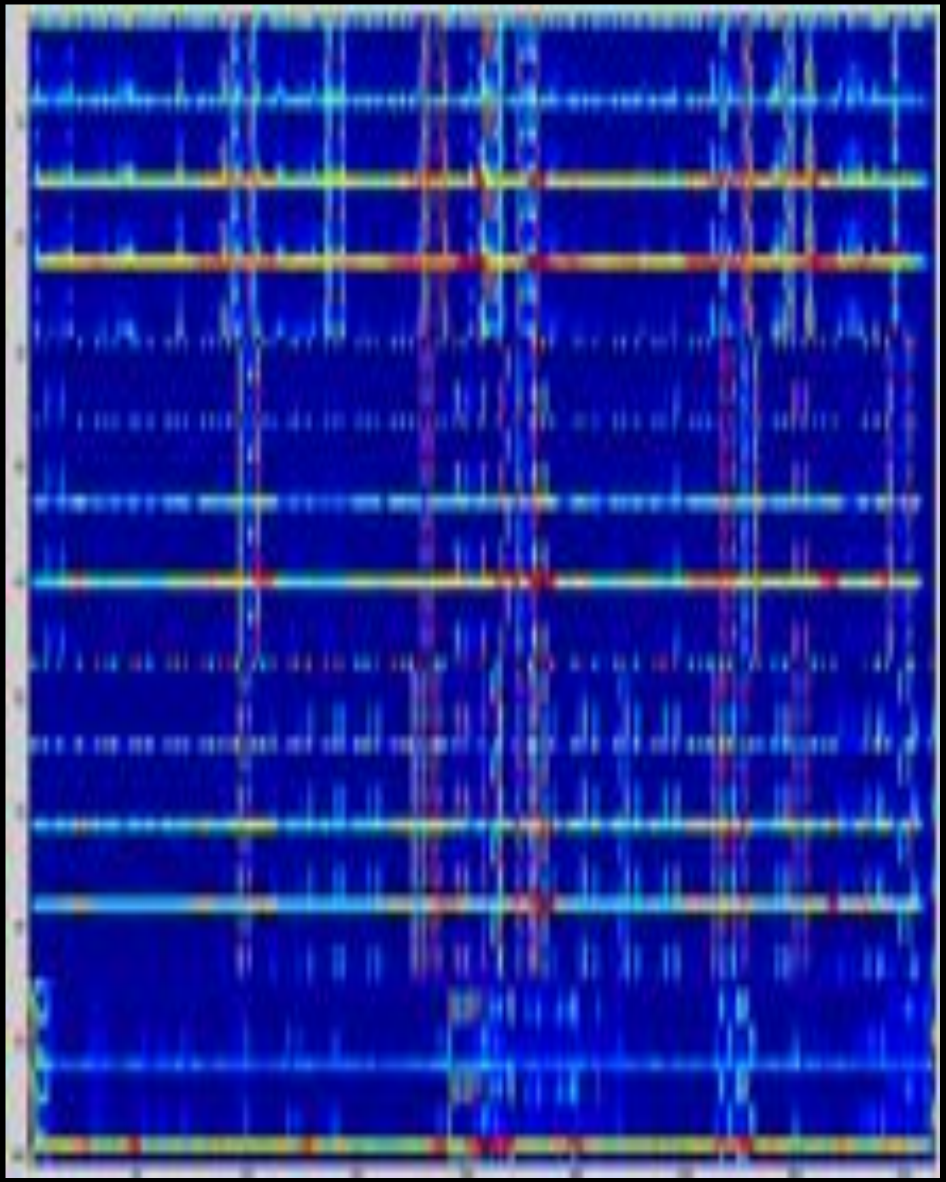
Kernel



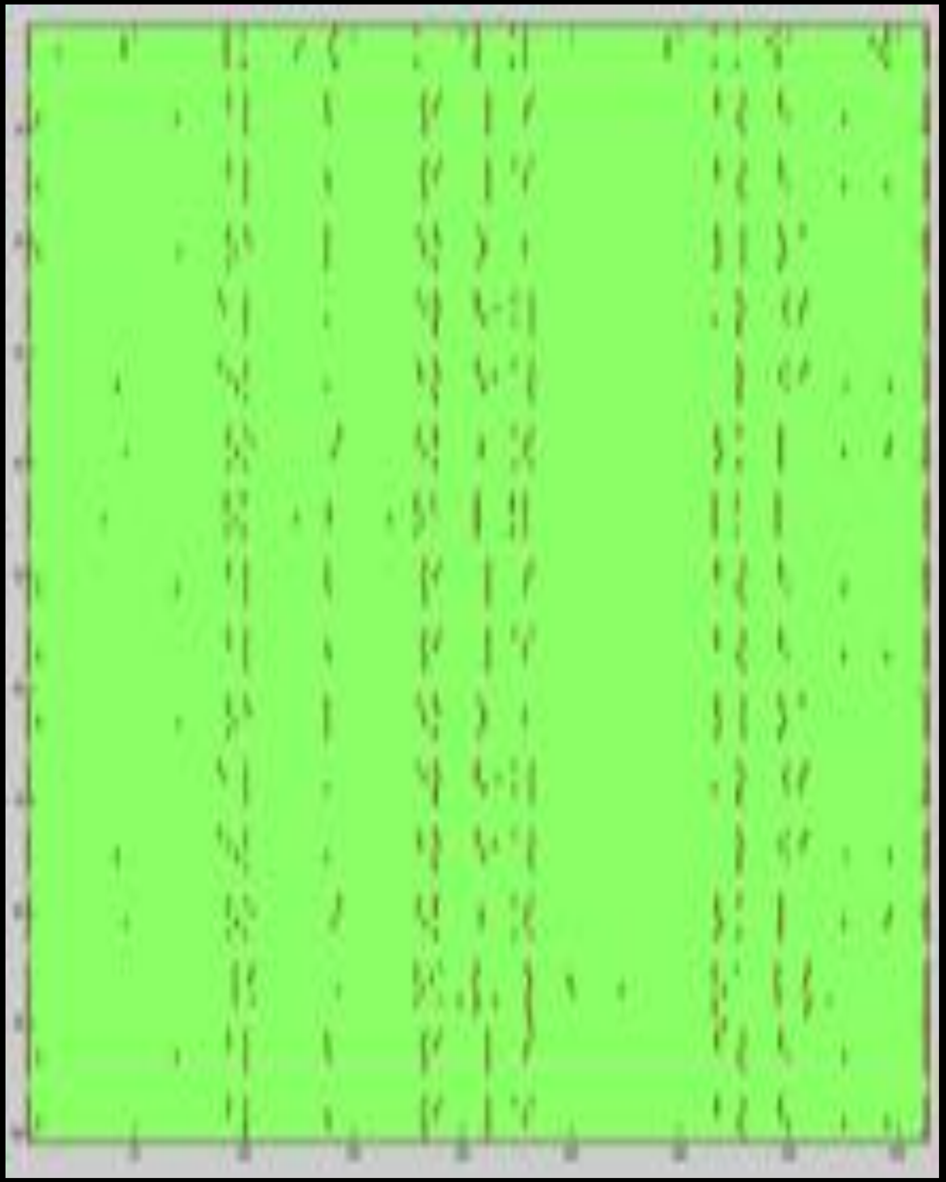
Peak picking approaches



10 features



48 novelty functions



6 peak picking methods

Experiment

- Estimate boundaries using 1000s of parameter settings
  - BASELINE: use train/test split to choose best parameter setting
  - PROPOSAL: pick the estimated solution with the greatest prior likelihood

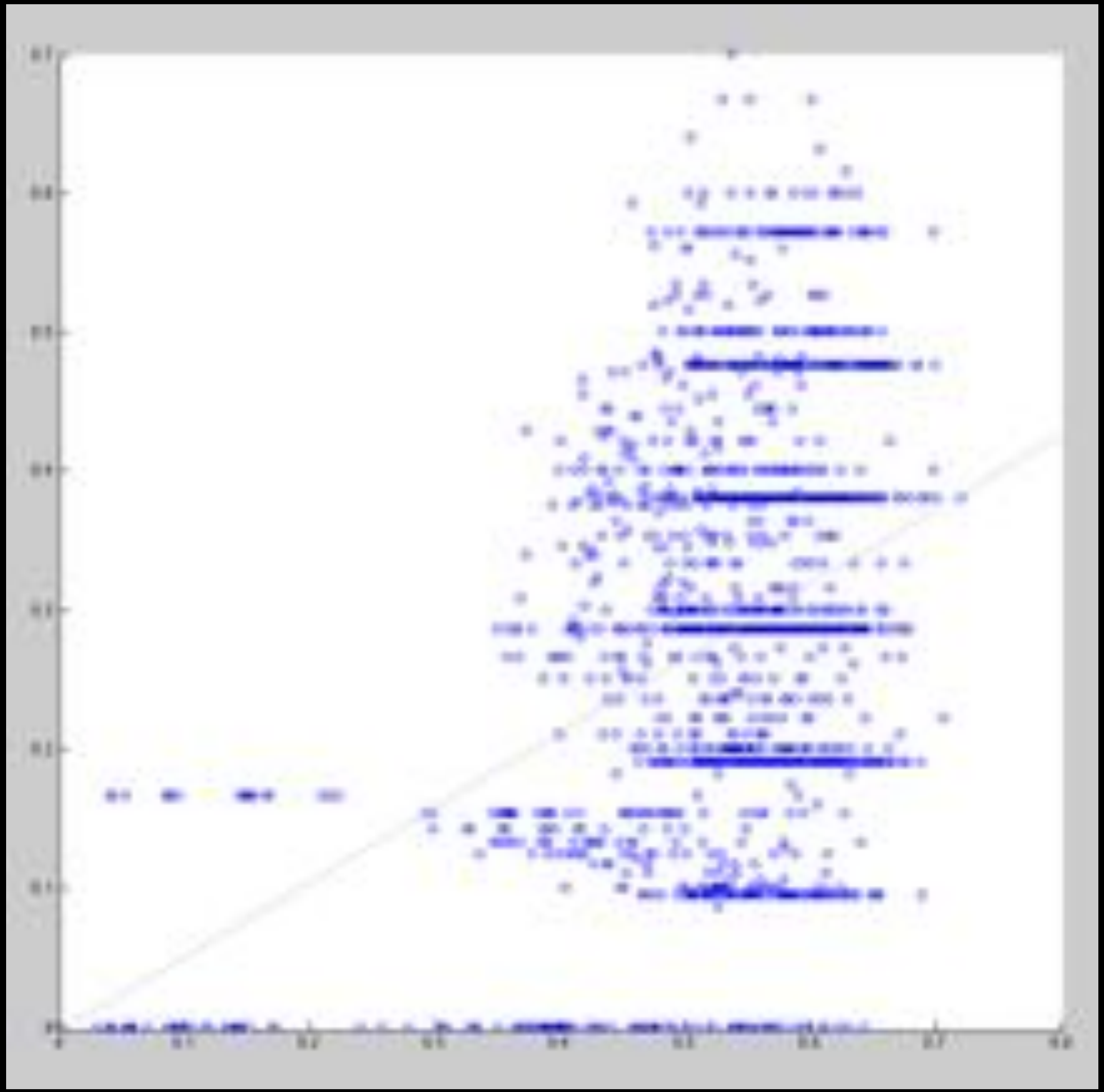
	boundary f-measure @ 3sec.
Baseline approach: pick best parameter set	**0.47**
Range of performance in MIREX:	**~0.45–0.61**
Our approach: pick solution with greatest prior likelihood	**0.37**

Reality check

- Do the fitness values I'm generating correlate with the quality of the analyses?

- This is one song, with one fitness measure, with correlation 0.77.

- The mean across all songs is between 0.5 and 0.65 for all the fitness measures.



How to combine fitness values?

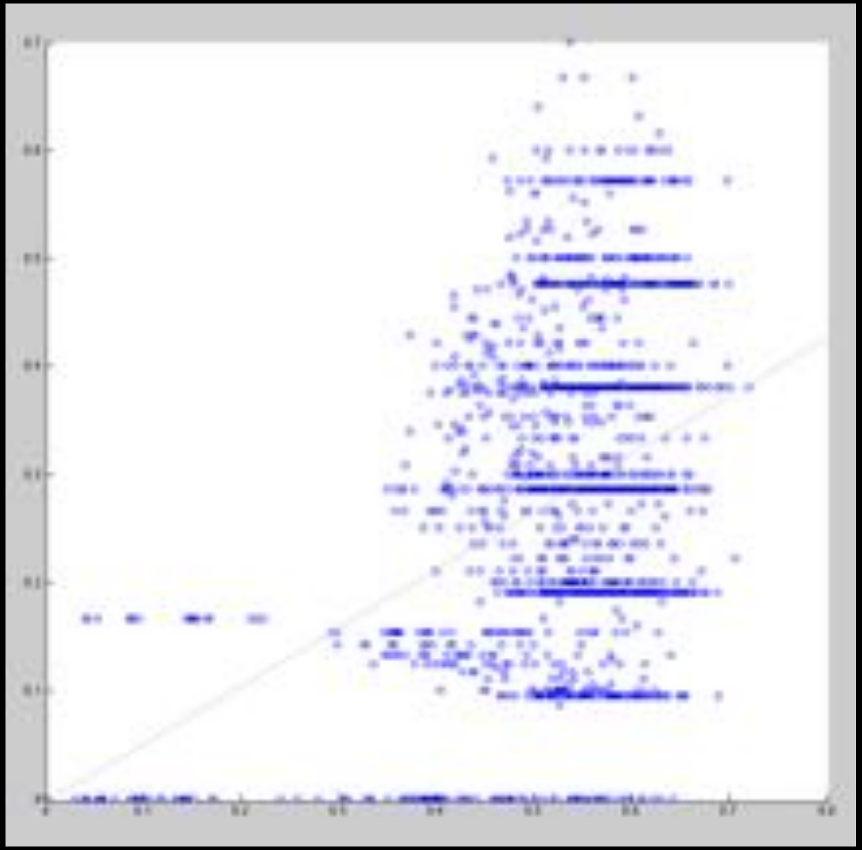
Method	mean *f*-measure
Simple parameter tuning	0.47
Use product of all (7) fitness values	0.37
Sum of all fitness values	0.36
Prior on absolute segment length only	0.35
Prior on number of segments	0.31
Prior on ratio between successive segments	0.21
Prior on ratio to median segment length	0.15

Learn a linear model to predict f-measure from fitness values

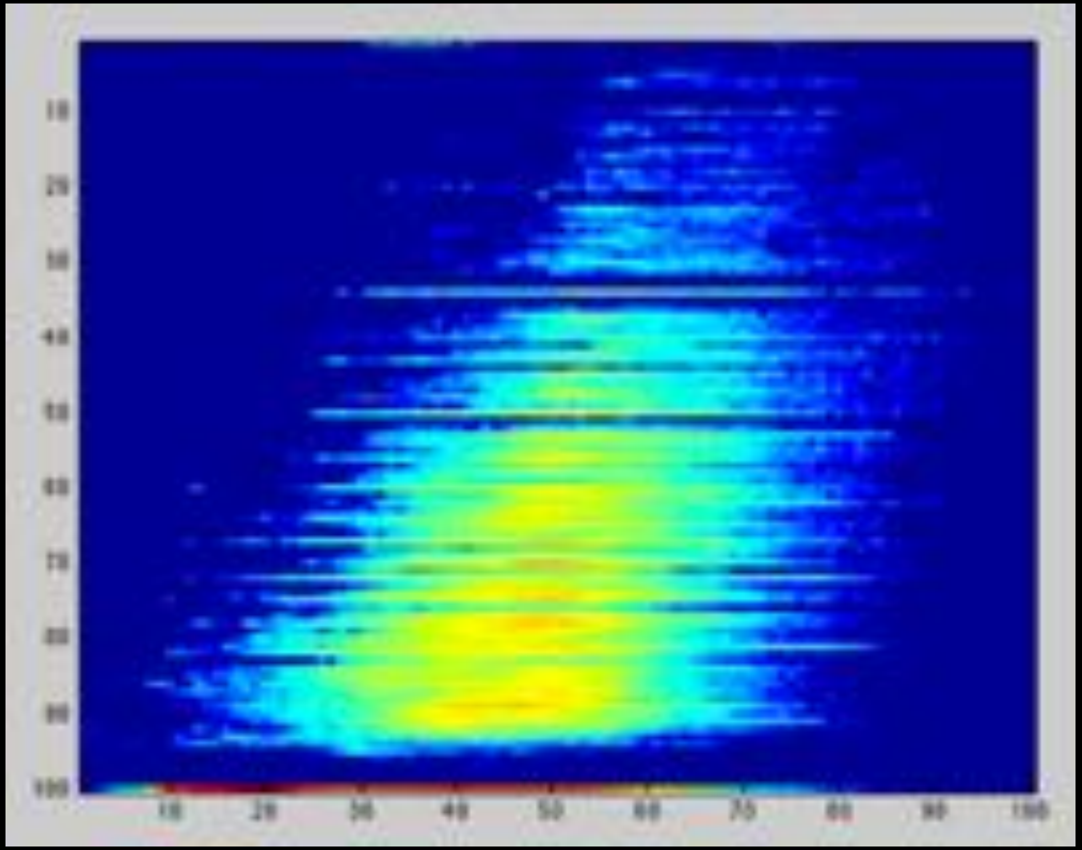
Linear model	~0.37
Interactions model	~0.43
Quadratic model	?

Sample output

Before:



Now:



Other improvements

1. **Estimate more priors**.

    Remove characterization part, and just use actual histogram.
    (I use KDE)

    Look at more properties of the annotations, and at conditional probabilities.

2. Experiment with **boundary-merging methods.**

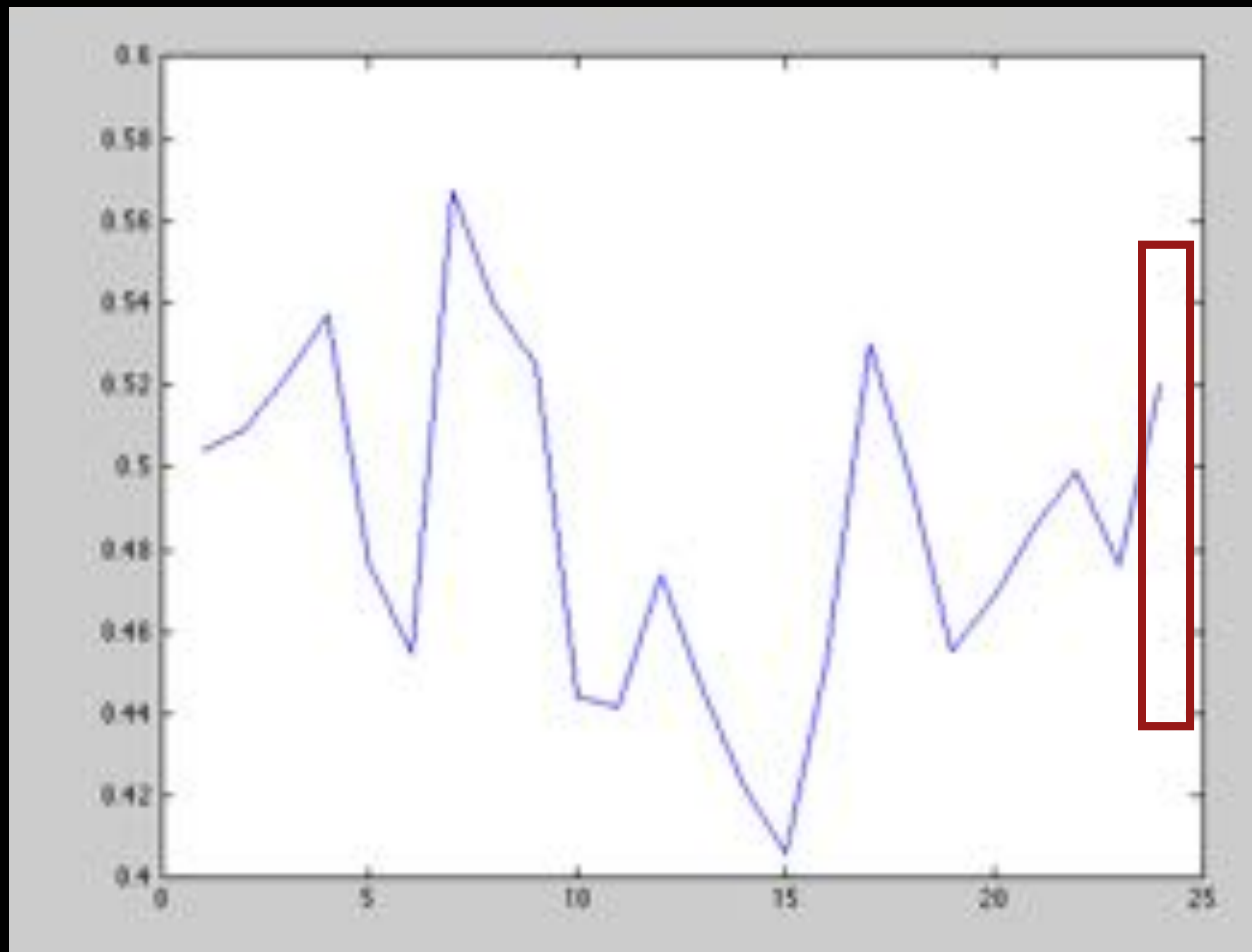
    After creating merged descriptions, compute the fitness of these and continue as usual.

3. Experiment with focusing power on top percentile of fitness values.
(Goal is not actually to predict f-measure, but select best one.)

Nothing worked!
What if the problem is
garbage-in, garbage-out?

Reality check: try to predict winning MIREX entry from fitness values

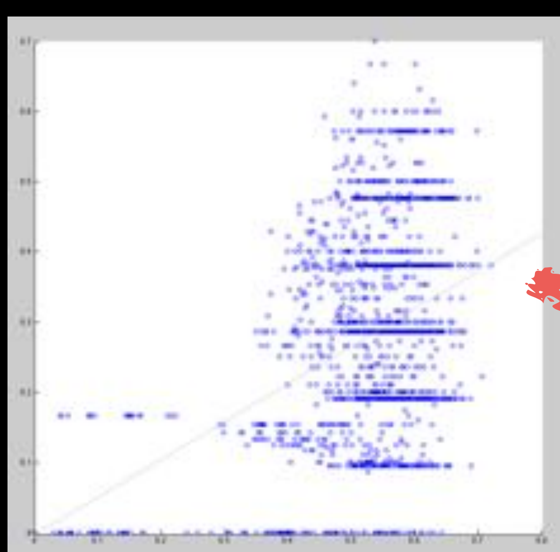
Performance of all MIREX submissions, 2012–2014

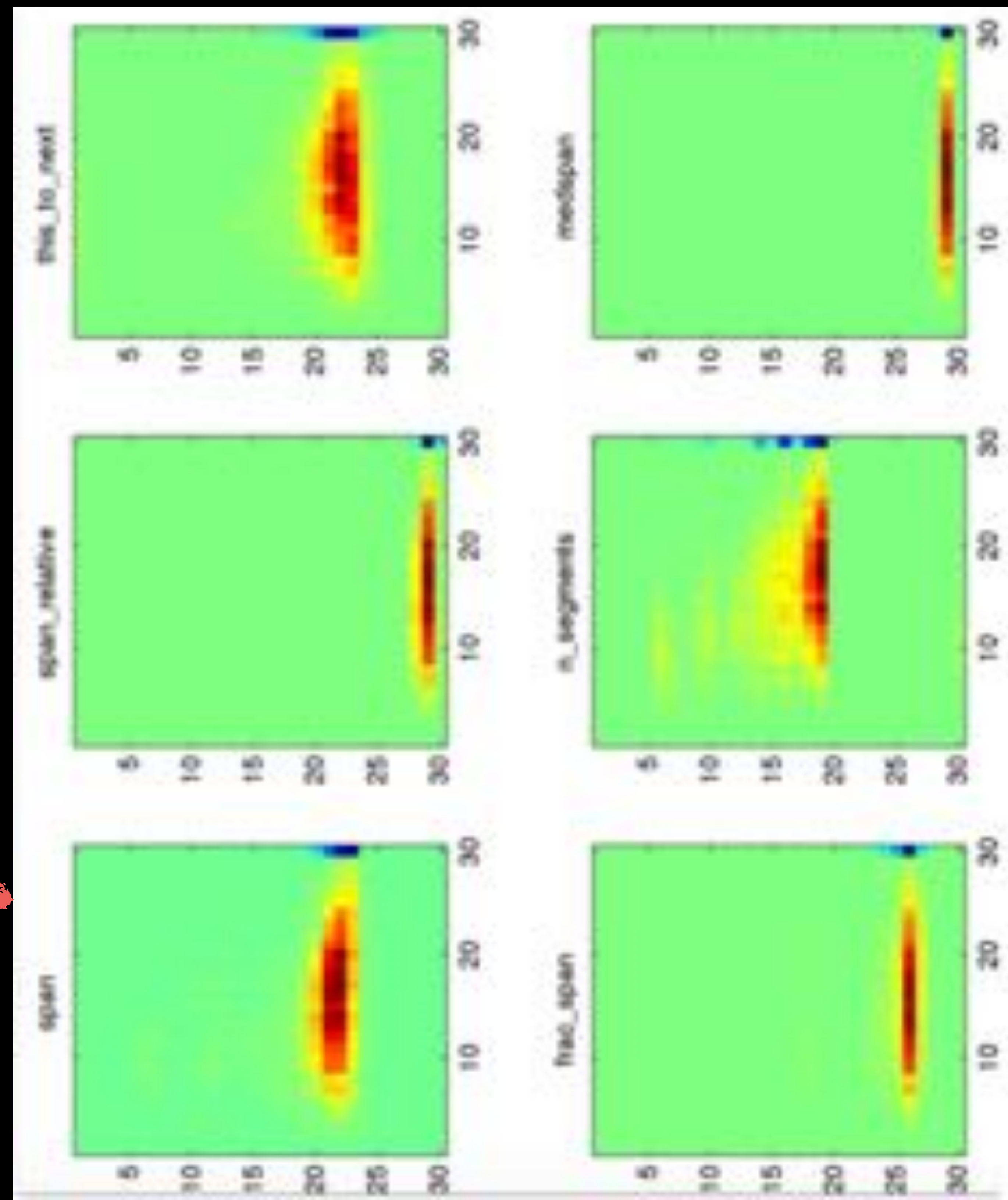


Performance of method that learns from fitnesses

Meanwhile, an SVM learned to always pick Ullrich et al. 2014

Reality check: how 'fit' are the annotations themselves?





Conclusion

- Method simply does not work!
  - Does not find the gem in a grab-bag of approaches
  - Does not find the gem in a committee of state-of-the-art approaches
- Output of state-of-the-art algorithms are already as ‘fit’ as annotations, without explicit training

Lesson:

- Check reality sooner!

Questions

- Do the experiment convince you that the approach cannot work?
- Do you think the approach will be useful for you?
- What is the underlying reason for the method not working?