USING PRIORS TO IMPROVE* ESTIMATES OF MUSIC STRUCTURE

Jordan B. L. Smith
Masataka Goto

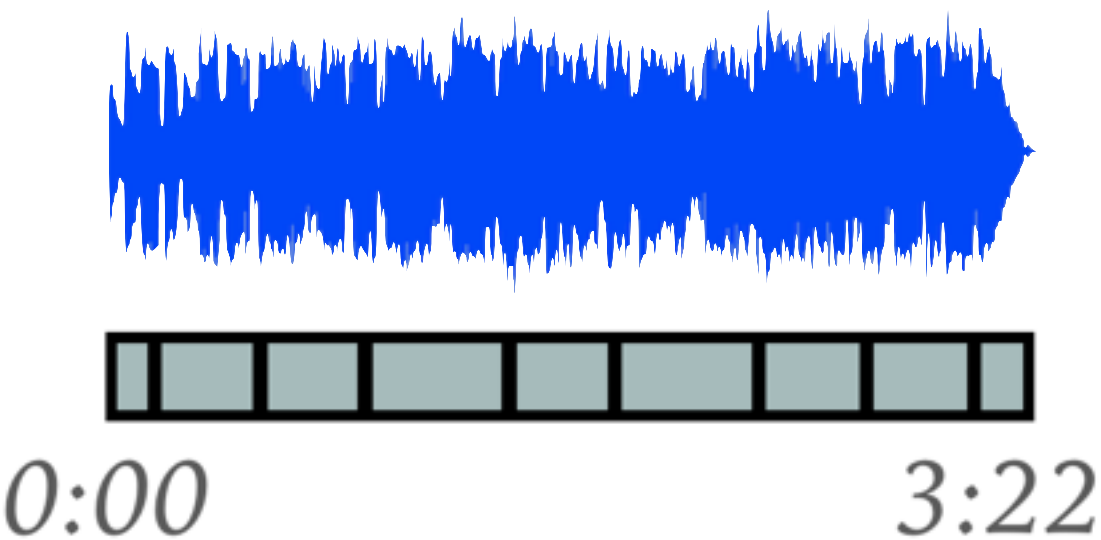
National Institute of Advanced Industrial Science and Technology (AIST), Japan

AIST

Wednesday, August 10th 2016
Oral Session #5: Structure

* or not

WHERE DO BOUNDARIES COME FROM?

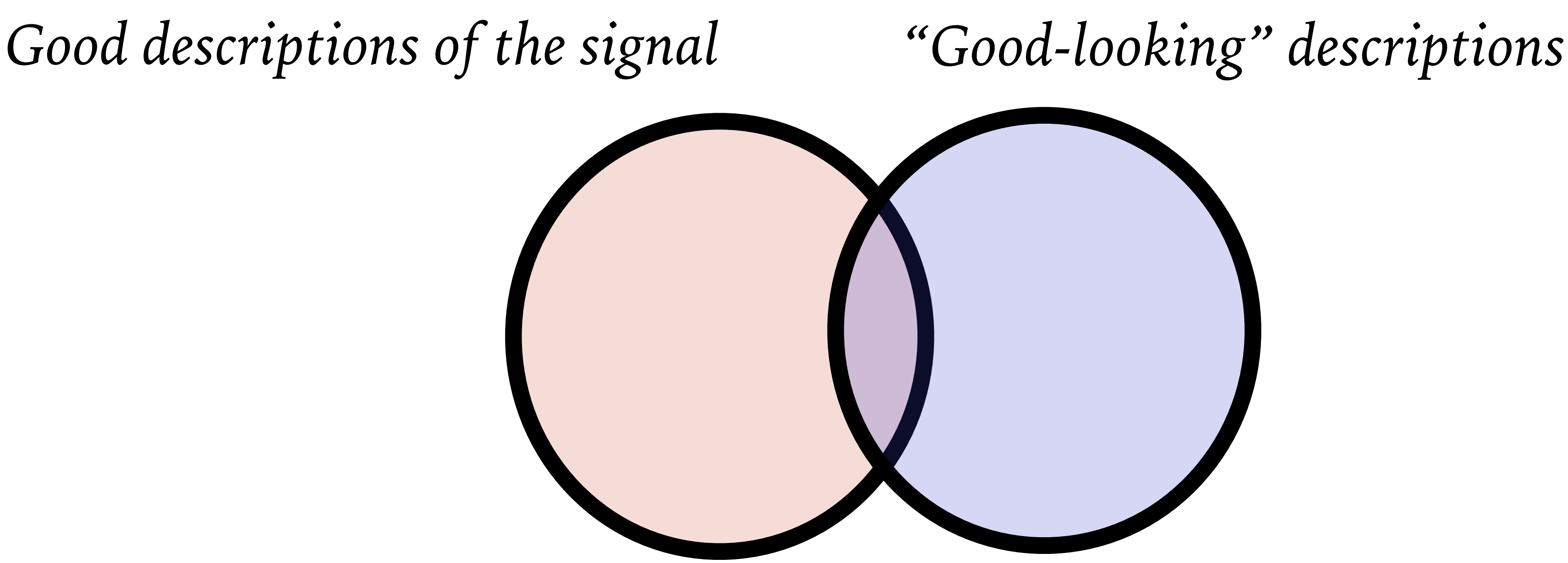


- The music!
  - Sudden changes
  - Repetitions
  - Homogenous stretches

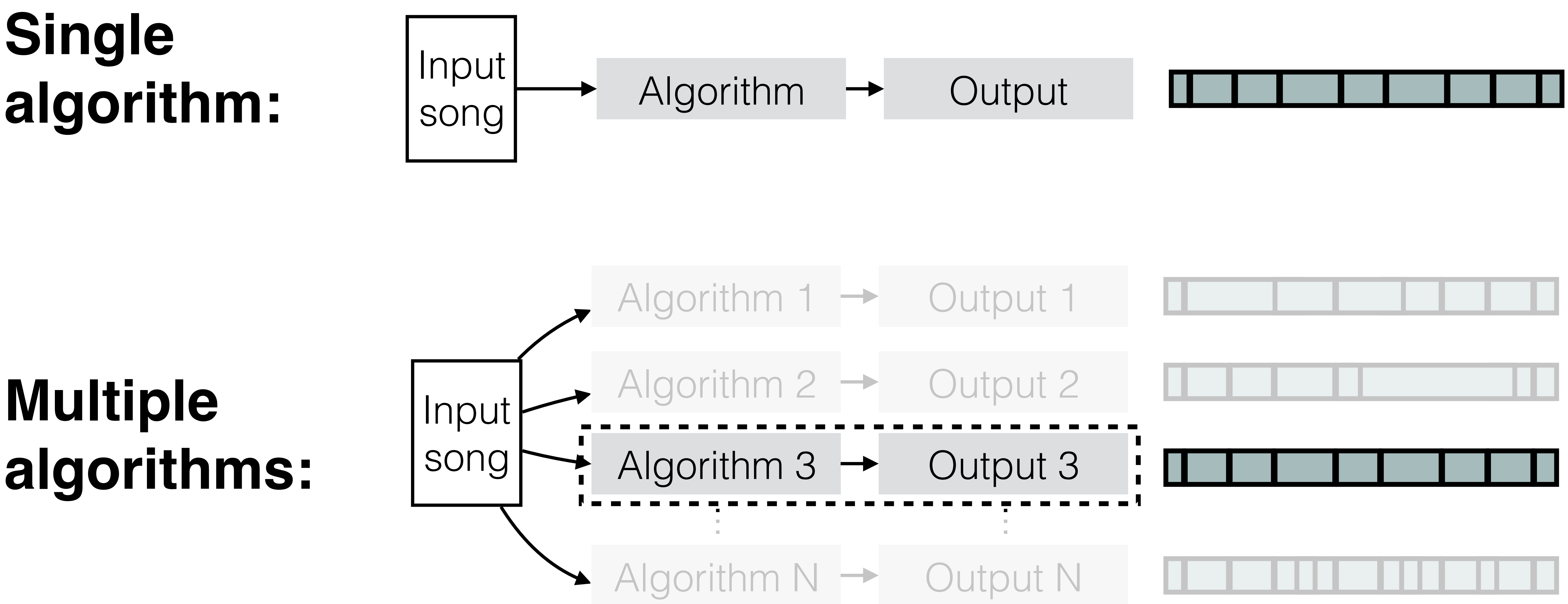
- The listener!
  - Person listens to the above, then decides on best description

MODELING "GOOD-LOOKING" DESCRIPTIONS

- Which is a better description of the piece *L'esempio imperfetta*?



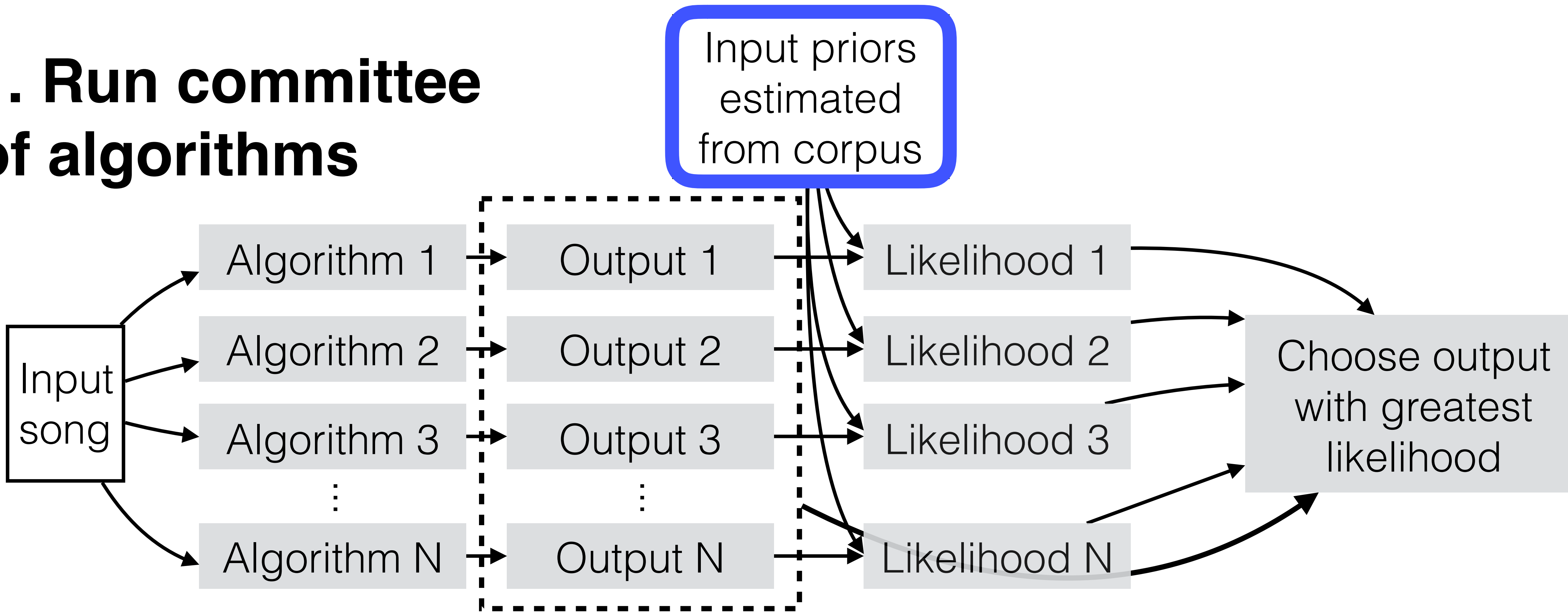
USUAL APPROACH



PROPOSAL

1. Run committee of algorithms

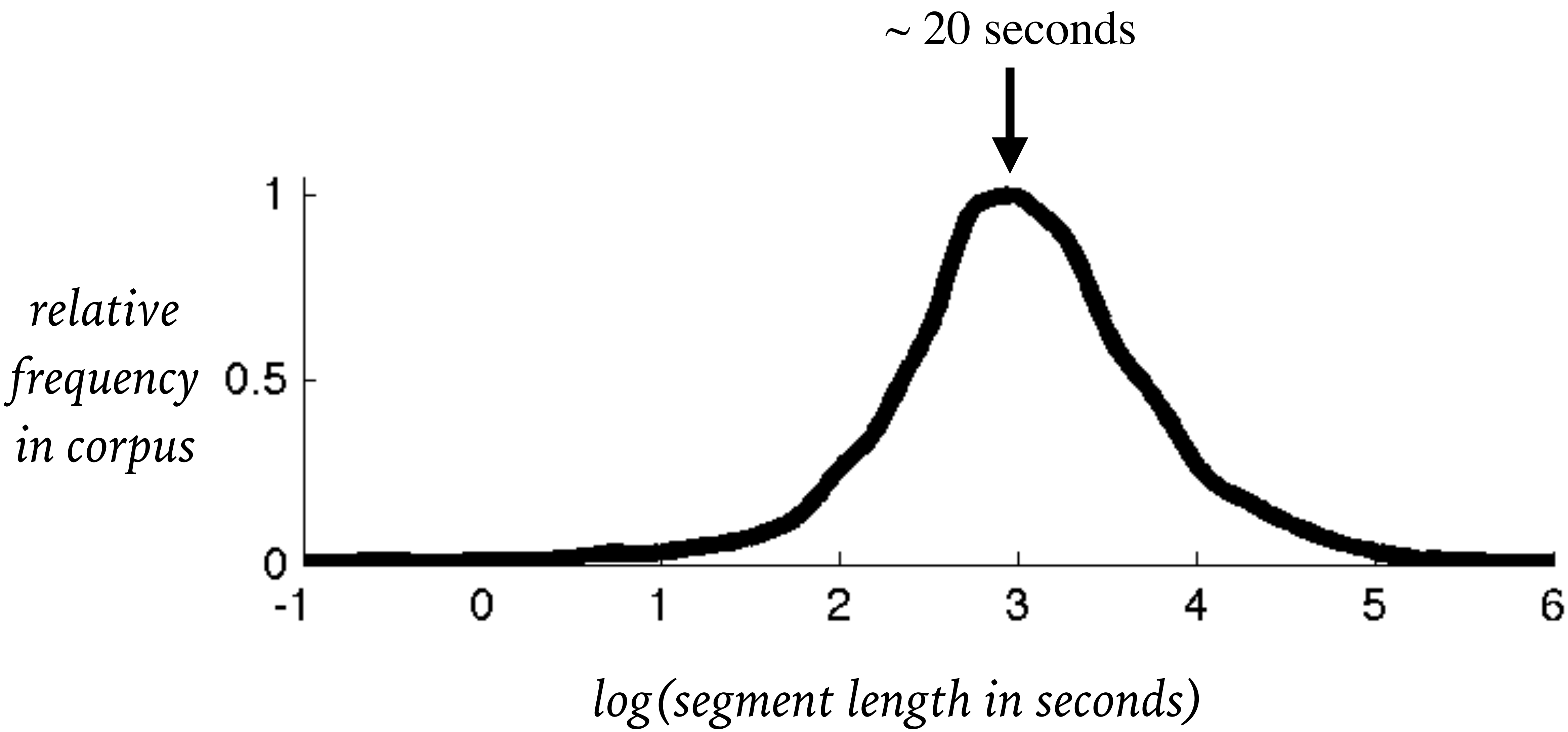
2. Use priors to predict likelihood of outputs



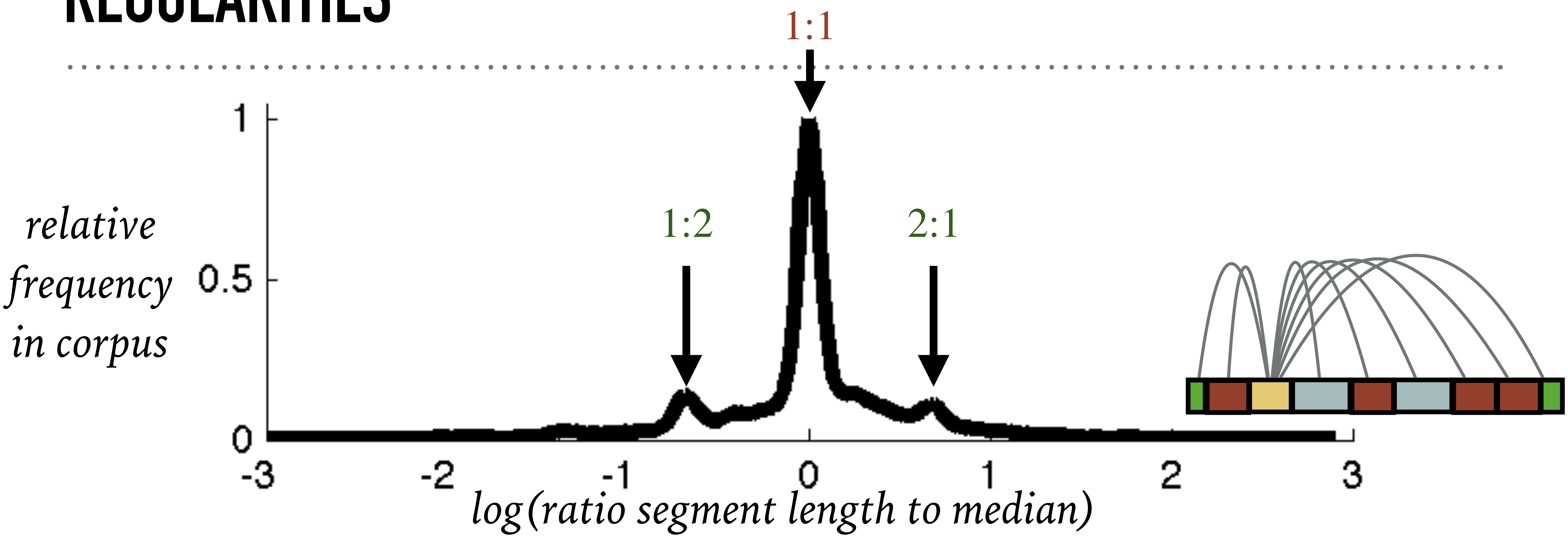
3. Use likelihoods to predict most accurate output

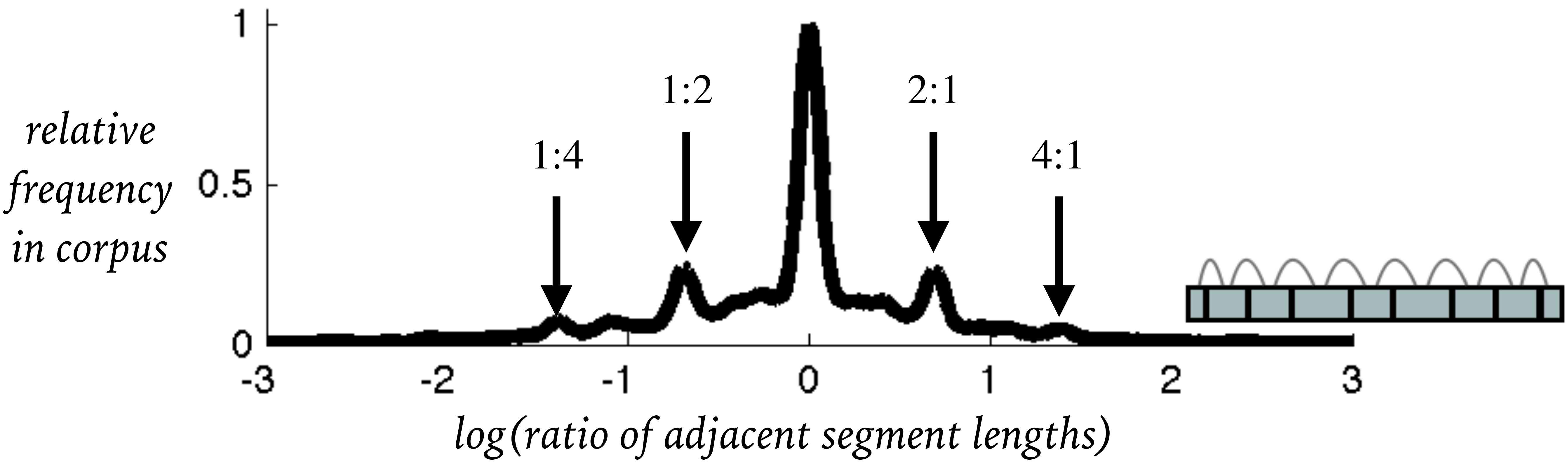
REGULARITIES

- Look at properties of SALAMI annotations



REGULARITIES





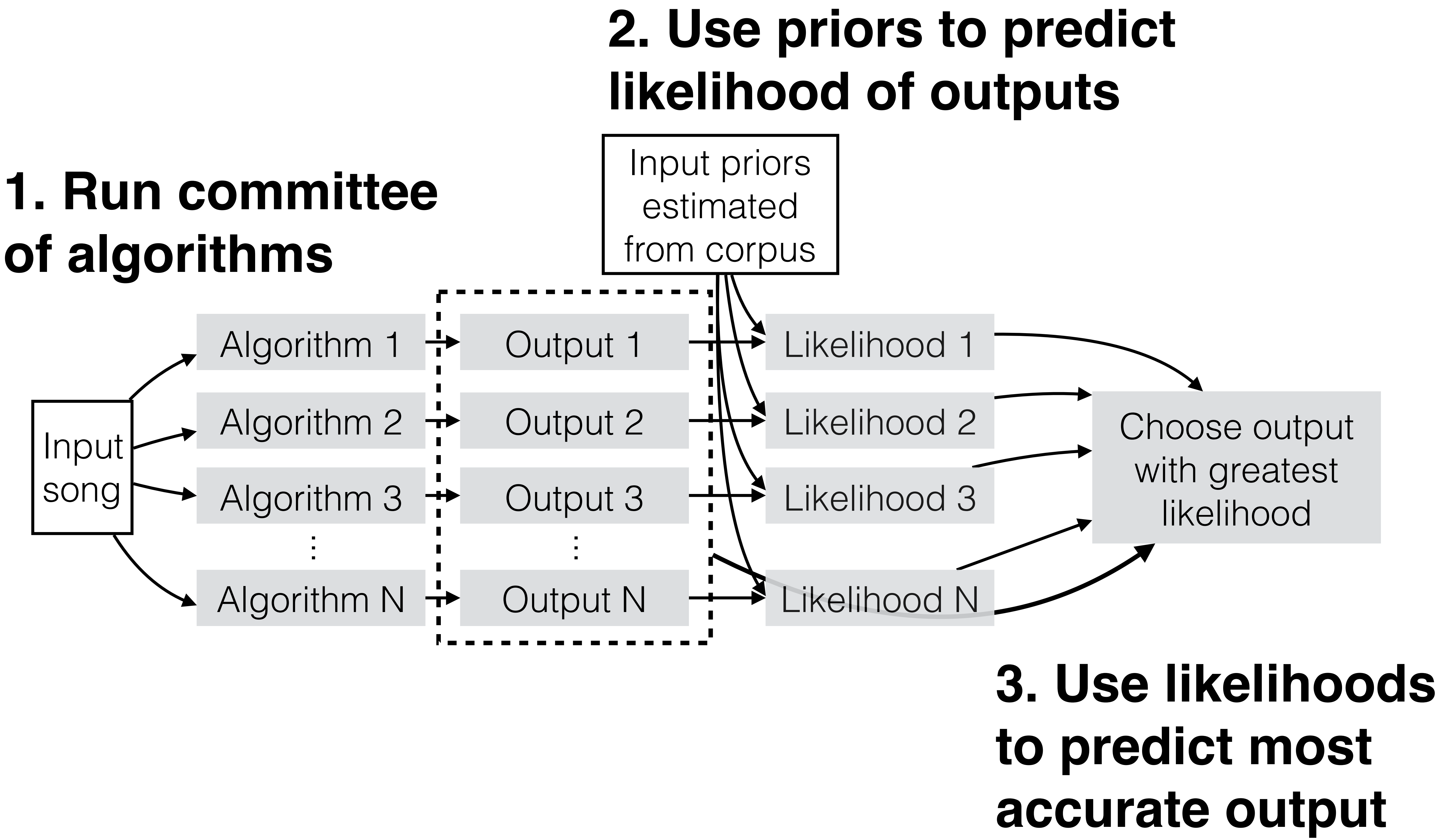
BACKGROUND

- Some strategies to model priors are widespread. E.g.:
    - Force segment length to fall within specific range (say, between 10 and 40 seconds)
    - Encourage segments to be 16, 32, or 64 beats long
- Learning directly from annotated audio is another option:
    - Turnbull et al. (2007) used machine learning to do binary classification of excerpts as boundaries or non-boundaries
    - Ullrich et al. (2014) did the same with neural nets and achieved a huge increase in performance

BACKGROUND

- Other notable examples:
    - Paulus and Klapuri (2009): “Defining a ‘Good’ Structural Description.” Cost function relates to description “quality”.
    - Sargent, Bimbot and Vincent (2011): Estimate median segment length; use to regulate cost function.
    - Rodriguez-Lopez, Volk and Bountoridis (2014): Similar approach, using corpus-estimated priors for melodic segmentation.
    - McFee et al. (2014): Used annotations to optimise their feature representation, then used a standard approach.

PROPOSAL



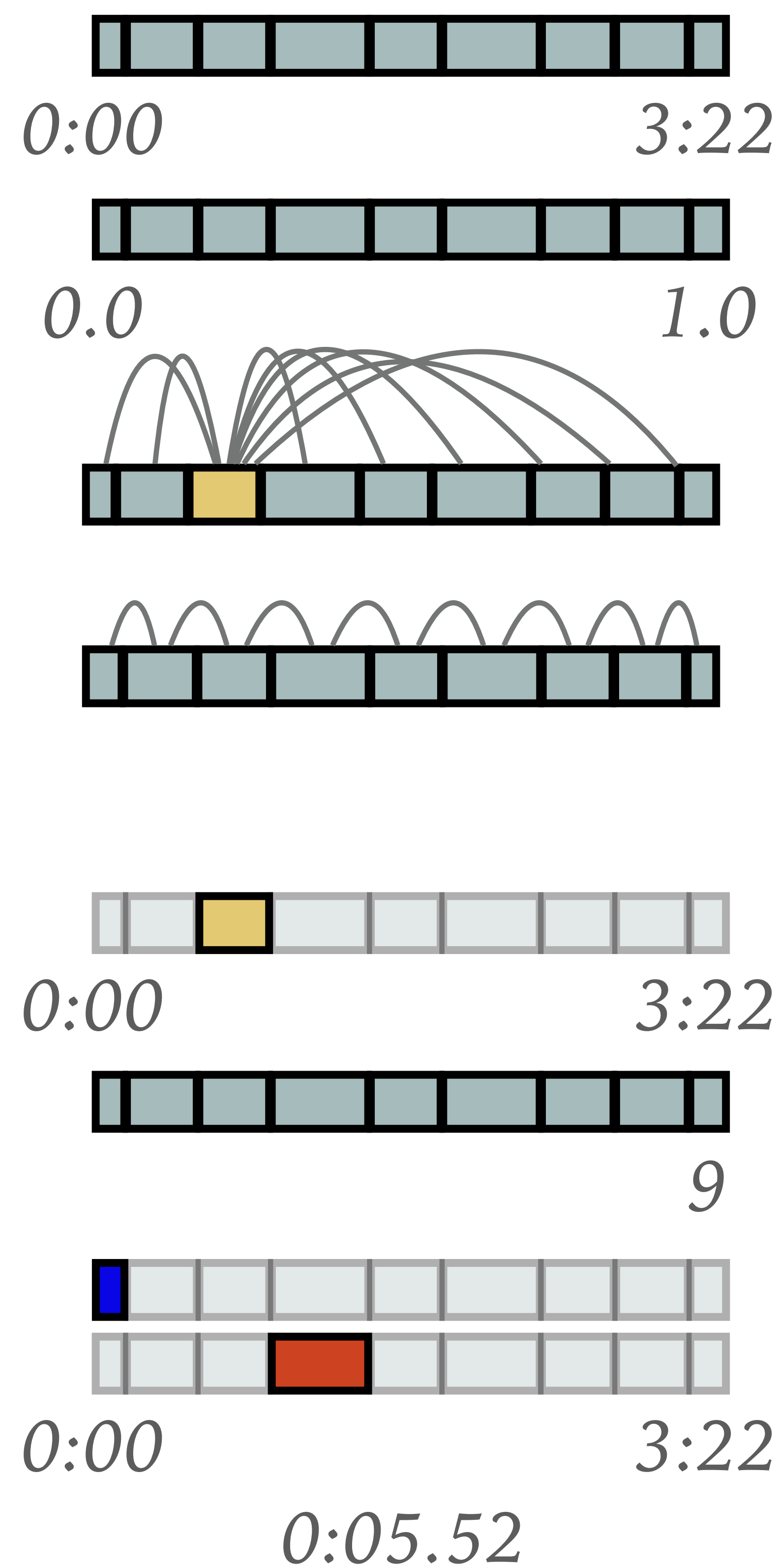
2. Use priors to predict likelihood of outputs
1. Run committee of algorithms
Input priors estimated from corpus
Input song
Algorithm 1
Algorithm 2
Algorithm 3
Algorithm N
Output 1
Output 2
Output 3
Output N
Likelihood 1
Likelihood 2
Likelihood 3
Likelihood N
Choose output with greatest likelihood
3. Use likelihoods to predict most accurate output

1. COMMITTEE OF ALGORITHMS

- Foote (2000) novelty-based segmentation parameters:
    - chroma, MFCC or tempogram features
    - median kernel size
    - checkerboard kernel size
    - novelty function adaptive threshold size
- Serra et al. (2012) structure feature-based segmentation parameters:
    - feature
    - embedded feature dimension size
    - nearest neighbour region
    - adaptive threshold for peak picking
- 40 members altogether
- Used MSAF to run algorithms (Nieto and Bello 2015)

2. SET OF PRIORS

- Per-segment properties:
  - A_1 = Segment length (L_i)
  - A_2 = Fractional segment length (L_i / song length)
  - A_3 = Ratio of L_i to median segment length
  - A_4 = Ratio of adjacent segment lengths (L_i / L_{i+1})
- Per-description properties:
  - A_5 = Median segment length (median of L_i)
  - A_6 = Number of segments
  - A_7 = Minimum segment length
  - A_8 = Maximum segment length
  - A_9 = Standard deviation of segment length



9 different priors

many log-likelihood values

40 committee members

−5.71	−5.85	−5.48	−8.75	−5.05	−6.63	−1.82	−6.27	−7.48
−5.69	−5.71	−5.76	−8.75	−4.93	−6.63	−1.82	−6.27	−7.42
−4.97	−5.09	−5.65	−7.13	−3.92	−5.34	−1.82	−4.85	−5.22
−4.72	−4.97	−5.06	−6.71	−3.68	−4.98	−1.82	−3.99	−4.17
−5.71	−5.85	−5.48	−8.75	−5.05	−6.63	−1.82	−6.27	−7.48
−5.69	−5.71	−5.76	−8.75	−4.93	−6.63	−1.82	−6.27	−7.42
−4.97	−5.09	−5.65	−7.13	−3.92	−5.34	−1.82	−4.85	−5.22
−4.72	−4.97	−5.06	−6.71	−3.68	−4.98	−1.82	−3.99	−4.17
−4.19	−4.51	−4.08	−5.47	−3.69	−4.55	−1.82	−3.76	−3.63
−4.19	−4.50	−4.07	−5.27	−3.69	−4.55	−1.82	−3.76	−3.63
−4.33	−4.76	−4.10	−5.88	−3.72	−4.72	−1.82	−3.66	−3.58
−4.33	−4.75	−3.99	−5.89	−3.76	−4.72	−1.82	−3.66	−3.60
−4.19	−4.51	−4.08	−5.47	−3.69	−4.55	−1.82	−3.76	−3.63
−4.19	−4.50	−4.07	−5.27	−3.69	−4.55	−1.82	−3.76	−3.63
−4.33	−4.76	−4.10	−5.88	−3.72	−4.72	−1.82	−3.66	−3.58
−4.33	−4.75	−3.99	−5.89	−3.76	−4.72	−1.82	−3.66	−3.60
−5.61	−6.37	−6.04	−8.75	−3.91	−6.63	−1.82	−5.67	−6.60
−6.27	−6.10	−6.32	−10.28	−5.27	−6.73	−1.82	−6.40	−8.73
−4.38	−4.71	−4.27	−5.81	−3.66	−4.72	−1.82	−3.95	−3.81
−4.58	−4.98	−4.57	−6.09	−3.69	−4.98	−1.82	−3.99	−4.12
−5.61	−6.37	−6.04	−8.75	−3.91	−6.63	−1.82	−5.67	−6.60
−6.27	−6.10	−6.32	−10.28	−5.27	−6.73	−1.82	−6.40	−8.73
−4.38	−4.71	−4.27	−5.81	−3.66	−4.72	−1.82	−3.95	−3.81
−4.58	−4.98	−4.57	−6.09	−3.69	−4.98	−1.82	−3.99	−4.12
−4.20	−4.52	−4.22	−5.68	−3.64	−4.55	−1.82	−3.76	−3.63
−4.21	−4.51	−4.21	−5.68	−3.64	−4.55	−1.82	−3.71	−3.63
−4.33	−4.72	−4.15	−5.87	−3.72	−4.72	−1.82	−3.71	−3.60
−4.34	−4.71	−4.22	−6.10	−3.69	−4.72	−1.82	−3.74	−3.63
−4.20	−4.52	−4.22	−5.68	−3.64	−4.55	−1.82	−3.76	−3.63
−4.21	−4.51	−4.21	−5.68	−3.64	−4.55	−1.82	−3.71	−3.63
−4.33	−4.72	−4.15	−5.87	−3.72	−4.72	−1.82	−3.71	−3.60
−4.34	−4.71	−4.22	−6.10	−3.69	−4.72	−1.82	−3.74	−3.63
−6.27	−6.10	−6.32	−10.28	−5.27	−6.73	−1.82	−6.40	−8.73
−6.27	−6.10	−6.32	−10.28	−5.27	−6.73	−1.82	−6.40	−8.73
−6.27	−6.10	−6.32	−10.28	−5.27	−6.73	−1.82	−6.40	−8.73
−6.27	−6.10	−6.32	−10.28	−5.27	−6.73	−1.82	−6.40	−8.73
−6.27	−6.10	−6.32	−10.28	−5.27	−6.73	−1.82	−6.40	−8.73
−6.27	−6.10	−6.32	−10.28	−5.27	−6.73	−1.82	−6.40	−8.73
−6.27	−6.10	−6.32	−10.28	−5.27	−6.73	−1.82	−6.40	−8.73
−6.27	−6.10	−6.32	−10.28	−5.27	−6.73	−1.82	−6.40	−8.73

How to choose
an output
based on the priors?

3. USING PRIORS TO PREDICT BEST ANSWER

- Grab bag of techniques:
  - Maximize an individual prior (A_1 through A_9)
  - Maximize combination of priors:
    - sum of the prior likelihoods
    - minimum of A_1 through A_9
    - use a linear model to predict *f*-measure based on all likelihoods
    - use a higher-order linear model (interactions / quadratic models)

PROPOSAL

40 members

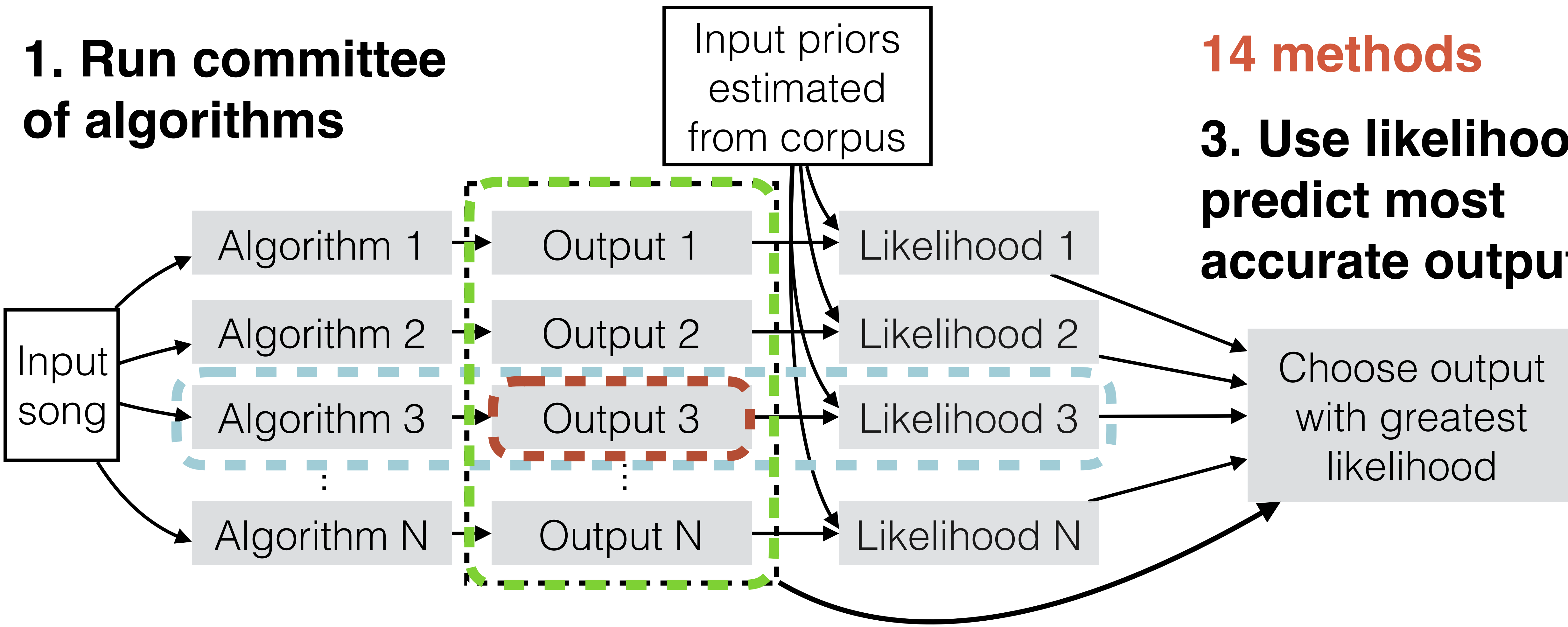
1. Run committee of algorithms

9 likelihoods

2. Use priors to predict likelihood

14 methods

3. Use likelihoods to predict most accurate output



Compare:

Average committee quality (min)

Parameter optimization (baseline)

Magically pick the best one (theoretical max)

RESULTS: FOOTE AND SERRA COMMITTEE ON PUBLIC SALAMI

System	f-measure (+/-3 seconds)	f-measure (+/- 0.5 seconds)	
A_1	0.4230	0.1051	*Individual priors*
A_2	0.4156	0.0958	
A_3	0.4176	0.1140	
A_4	0.4194	0.1072	
A_5	0.3597	0.0863	
A_6	0.3781	0.0991	
A_7	0.0603	0.0124	
A_8	0.3907	0.0961	
A_9	0.3956	0.0950	
$\sum A_i$	0.4260	0.1093	*Multiple priors*
$\min A_i$	0.4206	0.1046	
Linear model	0.4399	0.0845	*Linear models*
Interactions	0.4451	0.0688	
Quadratic	**0.4494**	0.0739	
Committee mean	0.2826	0.0691	
Baseline	0.4439	**0.1151**	
Theoretical max	0.6015	0.2572	

- A_1 - Segment length
- A_2 - Fractional segment length
- A_3 - Ratio to median segment length
- A_4 - Ratio of adjacent segment lengths
- A_5 - Median segment length
- A_6 - Number of segments
- A_7 - Minimum segment length
- A_8 - Maximum segment length
- A_9 - Standard deviation of segment length

EXPERIMENT #2: MIREX COMMITTEE

- Could a more diverse committee of state-of-the-art algorithms do better?
- Run the same experiment with new committee:
    - Set of 23 MIREX participants, 2012–2014.

RESULTS: MIREX COMMITTEE ON MIREX SALAMI

System	f-measure (+/-3 seconds)	f-measure (+/- 0.5 seconds)
A_1	0.6273	0.2733
A_2	0.3487	0.0996
A_3	0.3487	0.0996
A_4	0.3487	0.0996
A_5	0.3916	0.1385
A_6	0.3768	0.1594
A_7	0.3487	0.0996
A_8	0.4662	0.1356
A_9	0.4233	0.1514
$\sum A_i$	0.6273	0.2733
$\min A_i$	0.6273	0.2733
Linear model	0.5591	0.4005
Interactions	0.6273	0.4005
Quadratic	0.6273	0.4005
Committee mean	0.4447	0.1697
Baseline	**0.6273**	**0.4005**
Theoretical max	0.7345	0.5157

Individual priors

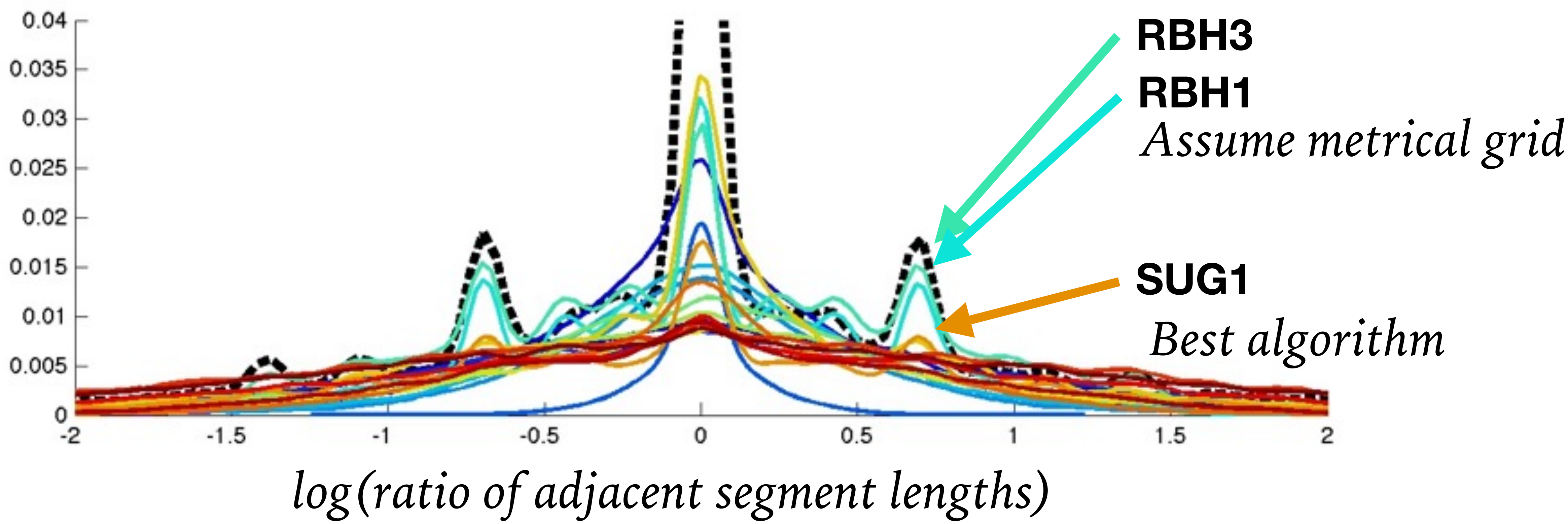
Multiple priors

Linear models

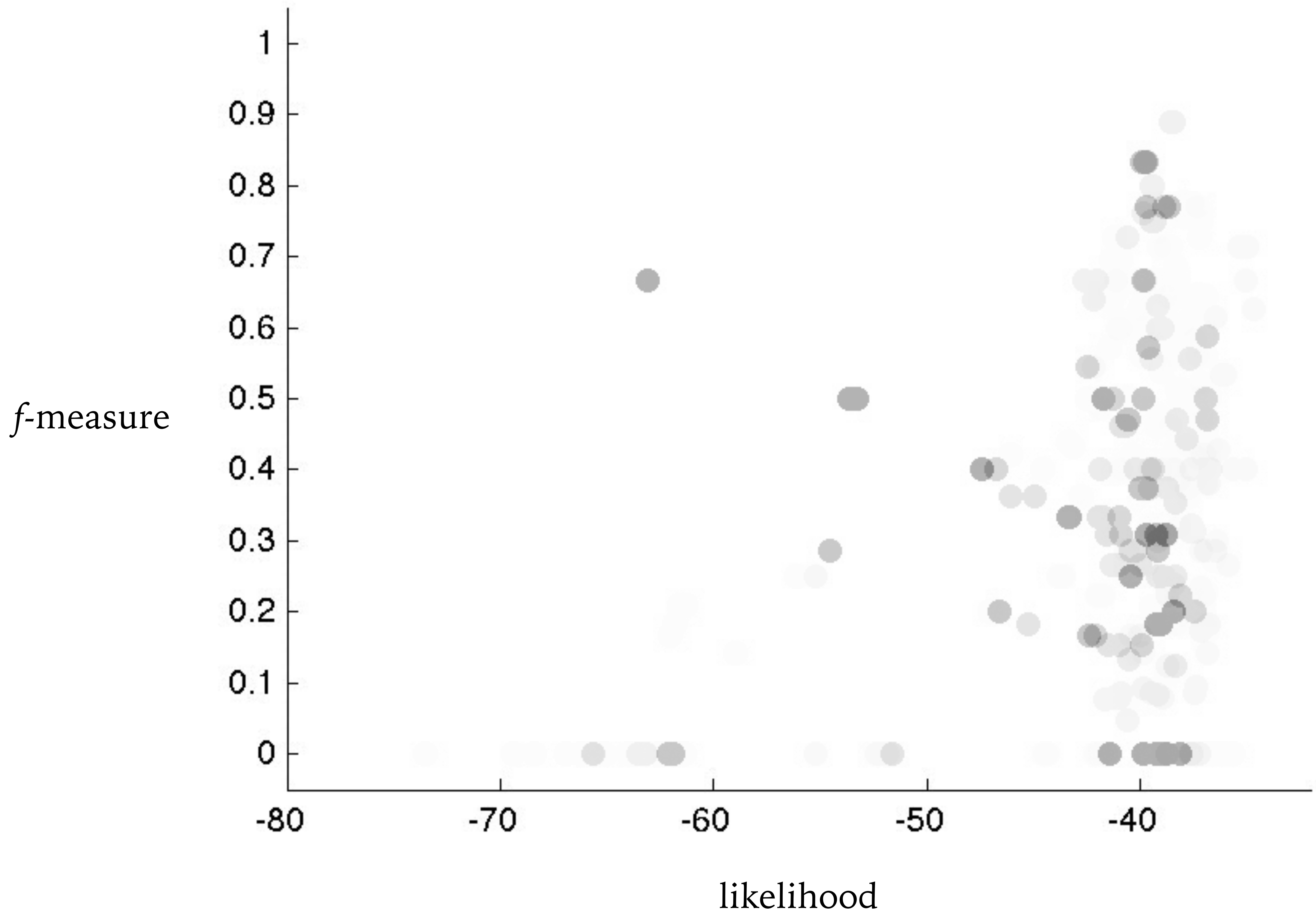
A_1 - Segment length
A_2 - Fractional segment length
A_3 - Ratio to median segment length
A_4 - Ratio of adjacent segment lengths
A_5 - Median segment length
A_6 - Number of segments
A_7 - Minimum segment length
A_8 - Maximum segment length
A_9 - Standard deviation of segment length

FAILURE ANALYSIS: EXISTING FIT TO PRIORS

- The method doesn't work. Why not?
    - Are the algorithms already producing "good-looking" descriptions?

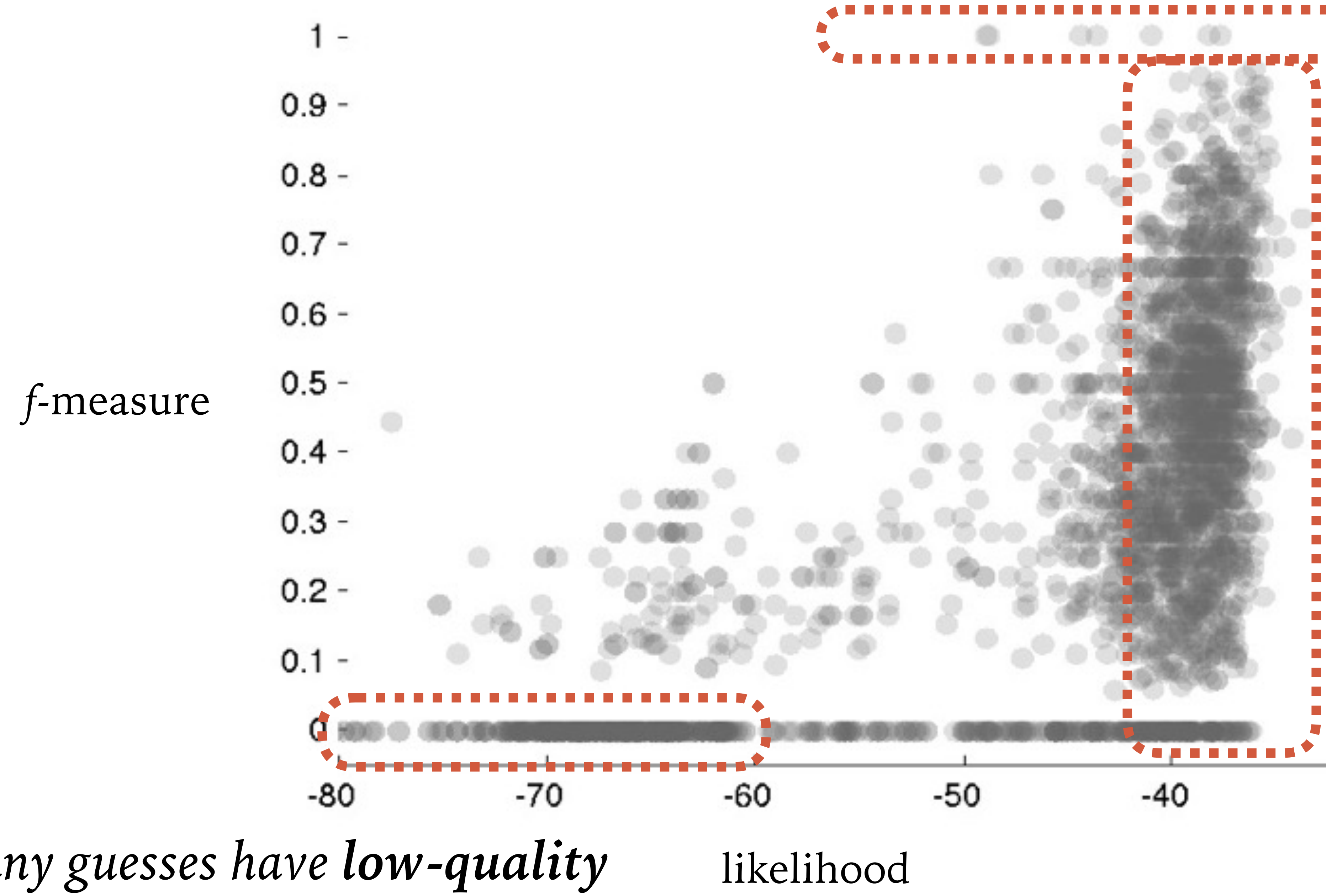


FAILURE ANALYSIS: CORRELATION BETWEEN FITNESS AND ACCURACY



FAILURE ANALYSIS: CORRELATION BETWEEN FITNESS AND ACCURACY

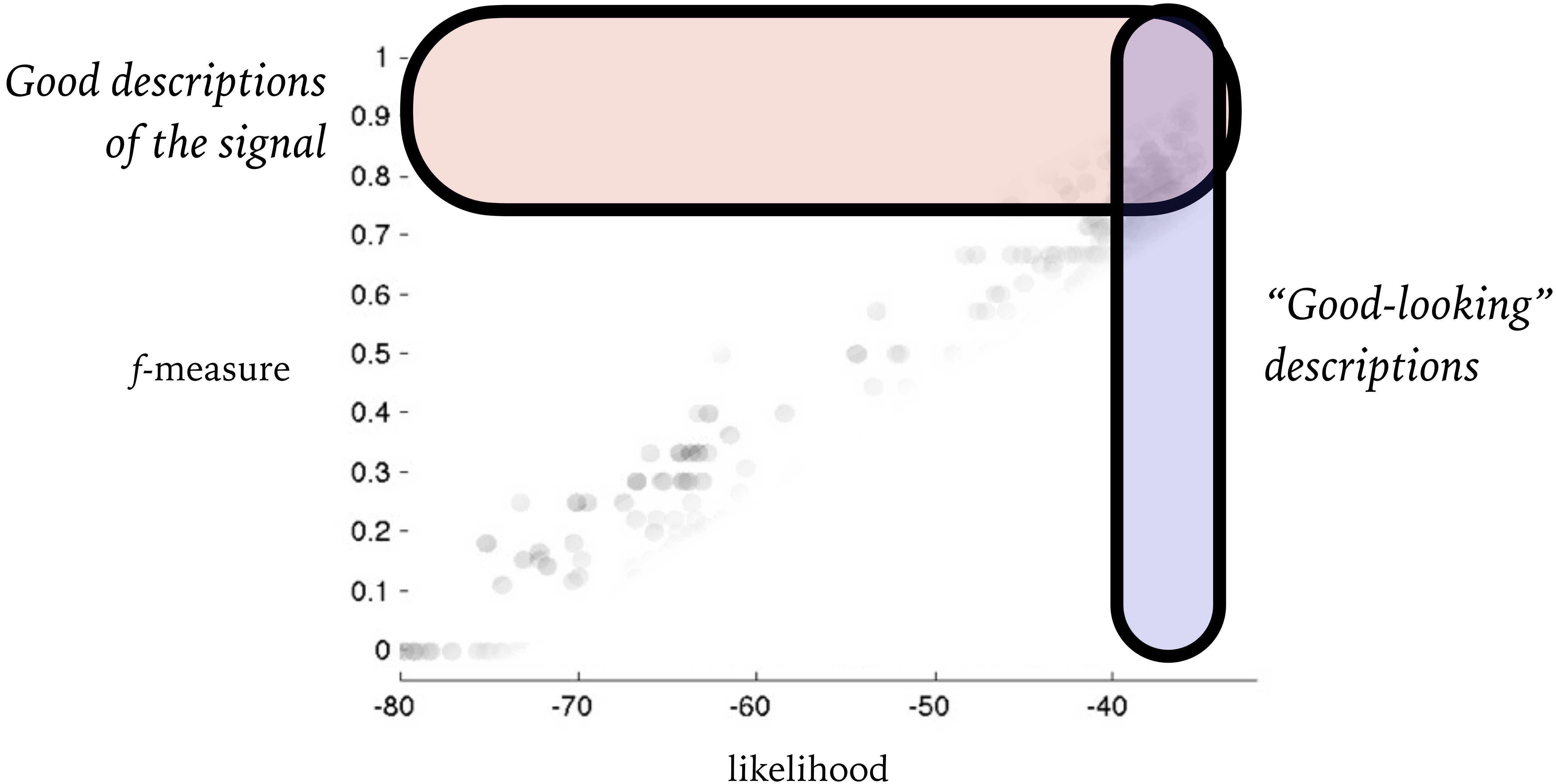
Annotations don't ***all*** *have high fitness*



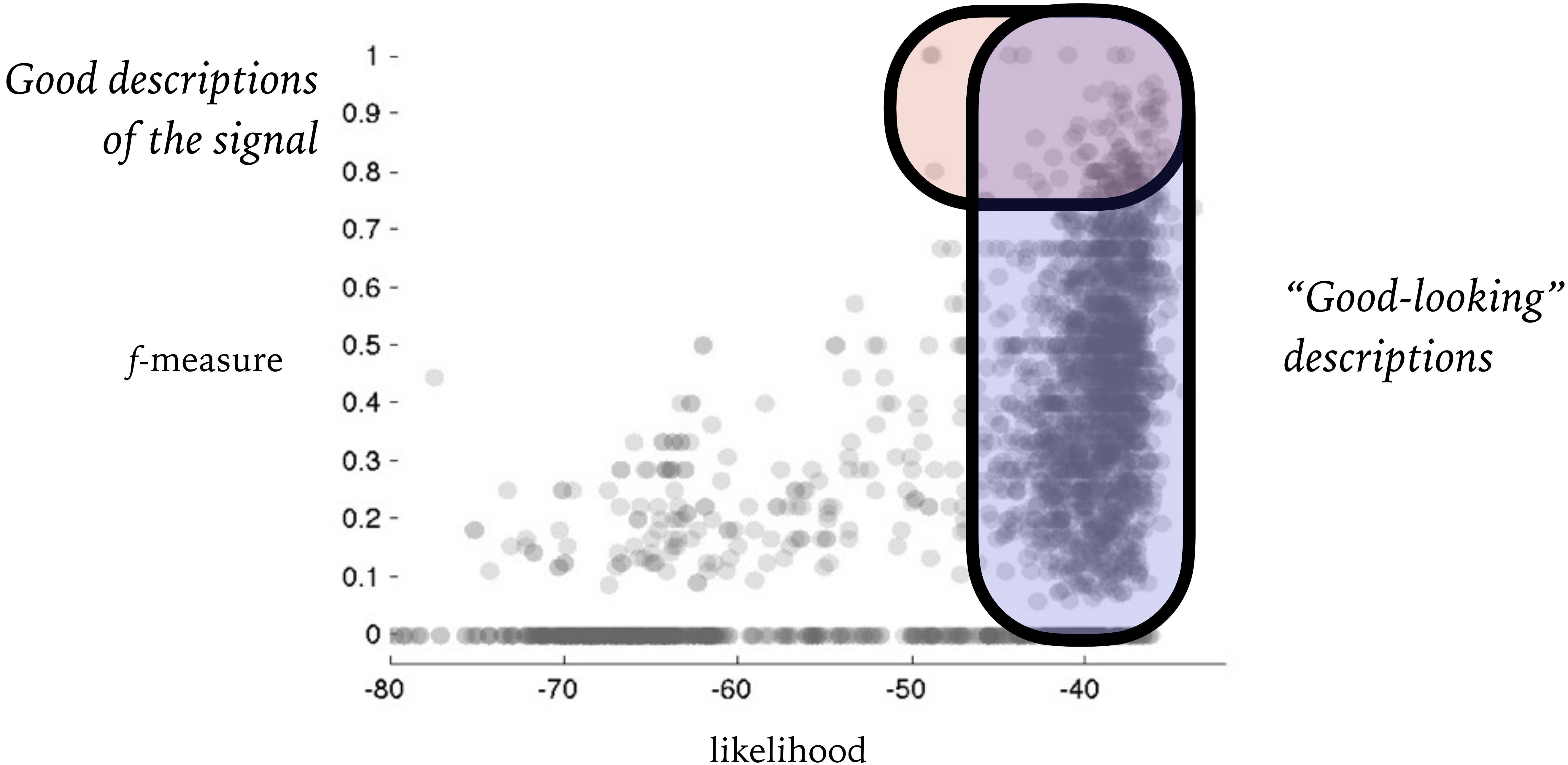
Most *output is already in the same high-likelihood region*

Many guesses have ***low-quality and low fitness****, boosting correlation unhelpfully*

FANTASY

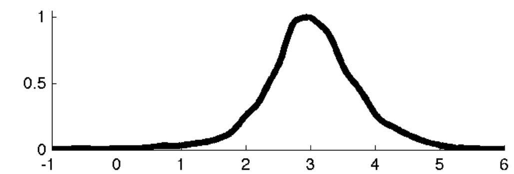

Good descriptions of the signal
“Good-looking” descriptions
f-measure
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
-80
-70
-60
-50
-40
likelihood

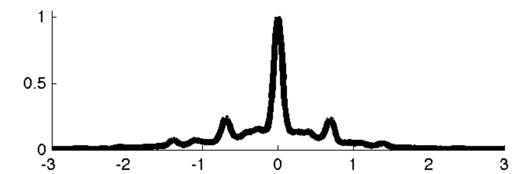
REALITY


Good descriptions of the signal
f-measure
"Good-looking" descriptions
likelihood
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
-80
-70
-60
-50
-40

CONCLUSION

- Annotations have strong regularities:
  - Restricted segment scale
  - Regular segment proportions





- These seem to be **not useful** for post-hoc algorithm improvement…
  - …but they may still be useful if modeled at earlier stages in an algorithm
- Cause of failure: algorithm output already **very good looking**!
  - Good signal-derived descriptions already fall into space of plausible descriptions

REFERENCES

- Jonathan Foote. Automatic audio segmentation using a measure of audio novelty. In *Proceedings of the IEEE International Conference on Multimedia & Expo*, 452–455, 2000.
- Brian McFee, Oriol Nieto, and Juan Pablo Bello. Hierarchical evaluation of segment boundary detection. In *Proceedings of ISMIR*, Málaga, Spain, 2015.
- Oriol Nieto and Juan Pablo Bello. Systematic exploration of computational music structure research. In *Proceedings of ISMIR*, New York, NY, USA, 2016.
- Jouni Paulus and Anssi Klapuri. Music structure analysis using a probabilistic fitness measure and a greedy search algorithm. *IEEE Transactions on Audio, Speech & Language Processing*, 17(6):1159–1170, 2009.
- Marcelo Rodríguez-López, Anja Volk, and Dimitrios Bountouridis. Multi-strategy segmentation of melodies. In *Proceedings of ISMIR*, 207–212, Taipei, Taiwan, November 2014.
- Gabriel Sargent, Frédéric Bimbot, and Emmanuel Vincent. A regularity-constrained Viterbi algorithm and its application to the structural segmentation of songs. In *Proceedings of ISMIR*, 483–488, Miami, FL, USA, 2011.
- Joan Serrà, Meinard Müller, Peter Grosche, and Josep Ll. Arcos. Unsupervised detection of music boundaries by time series structure features. In *Proceedings of the AAAI International Conference on Artificial Intelligence*, 1613–1619, Toronto, Canada, 2012.
- Jordan B. L. Smith, J. Ashley Burgoyne, Ichiro Fujinaga, David De Roure, and J. Stephen Downie. Design and creation of a large-scale database of structural annotations. In *Proceedings of ISMIR*, 555–560, Miami, FL, USA, 2011.
- Douglas Turnbull, Gert Lanckriet, Elias Pampalk, and Masataka Goto. A supervised approach for detecting boundaries in music using difference features and boosting. In *Proceedings of ISMIR*, 51–54, Vienna, Austria, 2007.
- Karen Ullrich, Jan Schlüter, and Thomas Grill. Boundary detection in music structure analysis using convolutional neural networks. In *Proceedings of ISMIR*, 417–422, Taipei, Taiwan, November 2014.
- MIREX. http://www.music-ir.org/mirex/wiki/MIREX_HOME

THANKS!

Special thanks to Juan Pablo Bello, Elaine Chew, Meinard Müller and Schloss Dagstuhl

AIST

SCHLOSS DAGSTUHL
Leibniz-Zentrum für Informatik