UT Dallas MSP Multimodal Signal Processing Laboratory

# Tradeoff Between Quality And Quantity Of Raters To Characterize Expressive Speech

Alec Burmania, Mohammed Abdelwahab, and Carlos Busso

Multimodal Signal Processing (MSP) lab
The University of Texas at Dallas
Erik Jonsson School of Engineering and Computer Science



UTD

# Labels from expressive speech

- Emotional databases rely on labels for classification
  - Usually obtained via perceptual evaluations
- Lab Setting
  - \+ Allows researcher close control over subjects
  - \- Expensive
  - \- Small demographic distribution
  - \- Smaller corpus size
- Crowdsourcing
  - \+ Can solve some of the above issues
  - \+ Widely tested and used in perceptual evaluations
  - \- Raises issues with rater reliability

amazon mechanical turk

CrowdFlower

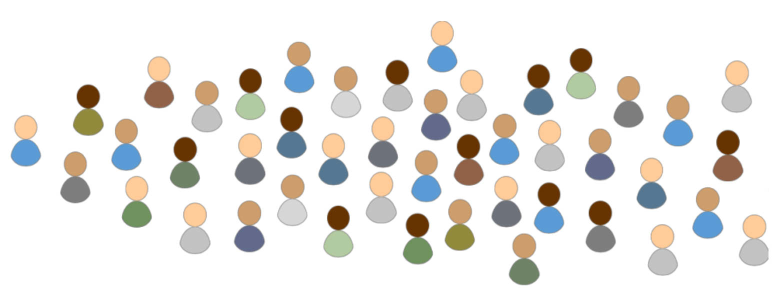
UTD

# Labels from expressive speech

- How do we balance quality and quantity in perceptual evaluations?
  - How many labels is enough?
- Crowdsourcing makes these decisions important

Many Evaluators
&
Low Quality

Few Evaluators
&
High Quality



or

- How does this affect classification?

# Effective Reliability

- Rosenthal et. al[1] proposes Spearman-Brown effective reliability framework for behavioral studies
  - Interprets reliability as a function of quality and quantity
- We use kappa as our metric ($\kappa$) and raters (n)

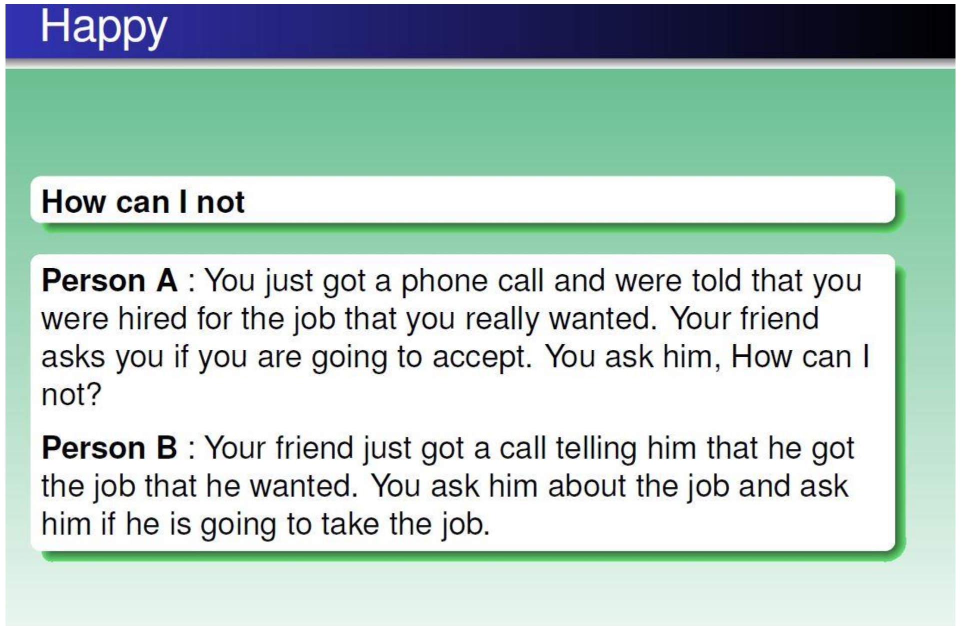
$$\text{Effective Reliability} = \frac{n\kappa}{1+(n-1)\kappa}$$

| | Mean Reliability ($\kappa$) | | | | | | |
|---|---|---|---|---|---|---|---|
| n raters | 0.42 | 0.45 | 0.48 | 0.51 | 0.54 | 0.57 | 0.60 |
| 5 | 78 | 80 | 82 | 84 | 85 | 87 | 88 |
| 10 | 88 | 89 | 90 | 91 | 92 | 93 | **94** |
| 15 | 92 | 92 | 93 | 94 | 95 | 95 | 96 |
| 20 | **94** | 94 | 95 | 95 | 96 | 96 | 97 |

[1] Jinni A Harrigan, Robert Ed Rosenthal, and Klaus R Scherer,The new handbook of methods in nonverbal behavior research.,Oxford University Press, 2005.

# MSP-IMPROV Corpus

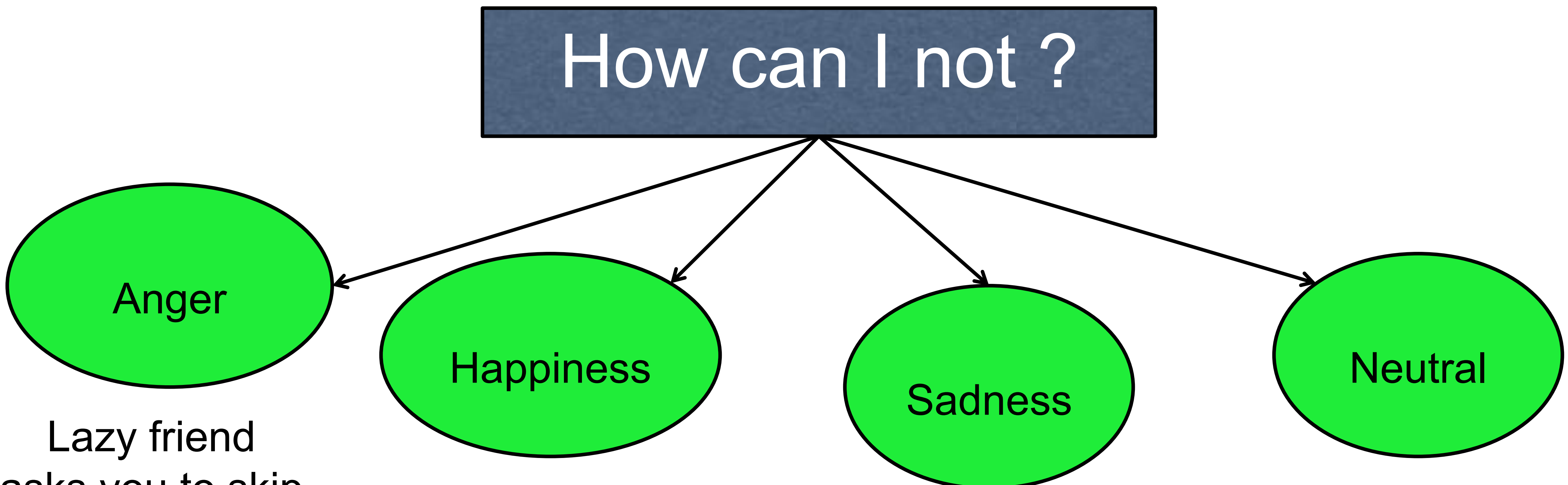
- Recordings of 12 subjects improvising scenes in pairs (>9 hours, 8,438 turns) [2]
- Actors are assigned context for a scene that they are supposed to act out
- Collected for corpus of fixed lexical content but different emotions
- Data Sets
  - Target – Recorded Sentences with fixed lexical content (648)
  - Improvisation – Scene to produce target
  - Interaction – Interactions between scenes




An example scene.

[2]Carlos Busso, Srinivas Parthasarathy, Alec Burmania, Mohammed AbdelWahab, Najmeh Sadoughi, and Emily Mower Provost, "MSP-IMPROV: An acted corpus of dyadic interactions to study emotion perception," IEEE Transactions on Affective Computing, vol. To appear, 2015.

# MSP-IMPROV Corpus




Lazy friend asks you to skip class

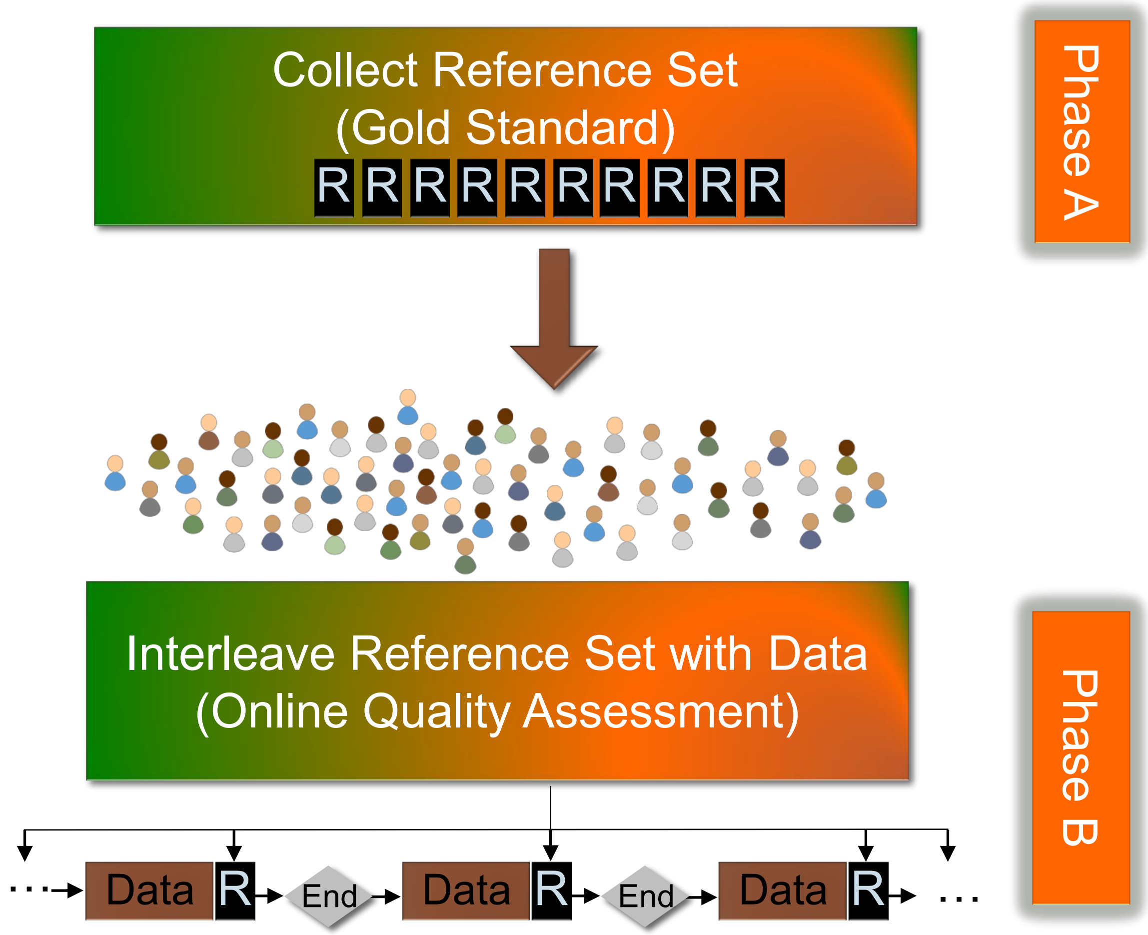
Accepting job offer

Taking extra help when you are failing classes

Using coupon at store

# Perceptual Evaluation

- Idea: Can we verify if a worker is spamming even while lacking ground truth labels for most of the corpus?
- We will focus on  a five class problem (Angry, Sad, Neutral, Happy, Other)




Collect reference set

Trace performance in real time

[3] Alec Burmania, Srinivas Parthasarathy, and Carlos Busso, "Increasing the reliability of crowdsourcing evaluations using online quality assessment," IEEE Transactions on Affective Computing, vol. To appear, 2015.

# Metric: Angular Agreement

- Assign categories (angry, sad, happy neutral, other) as a 5D space (v).
- We calculate the LOWO inter-evaluator agreement

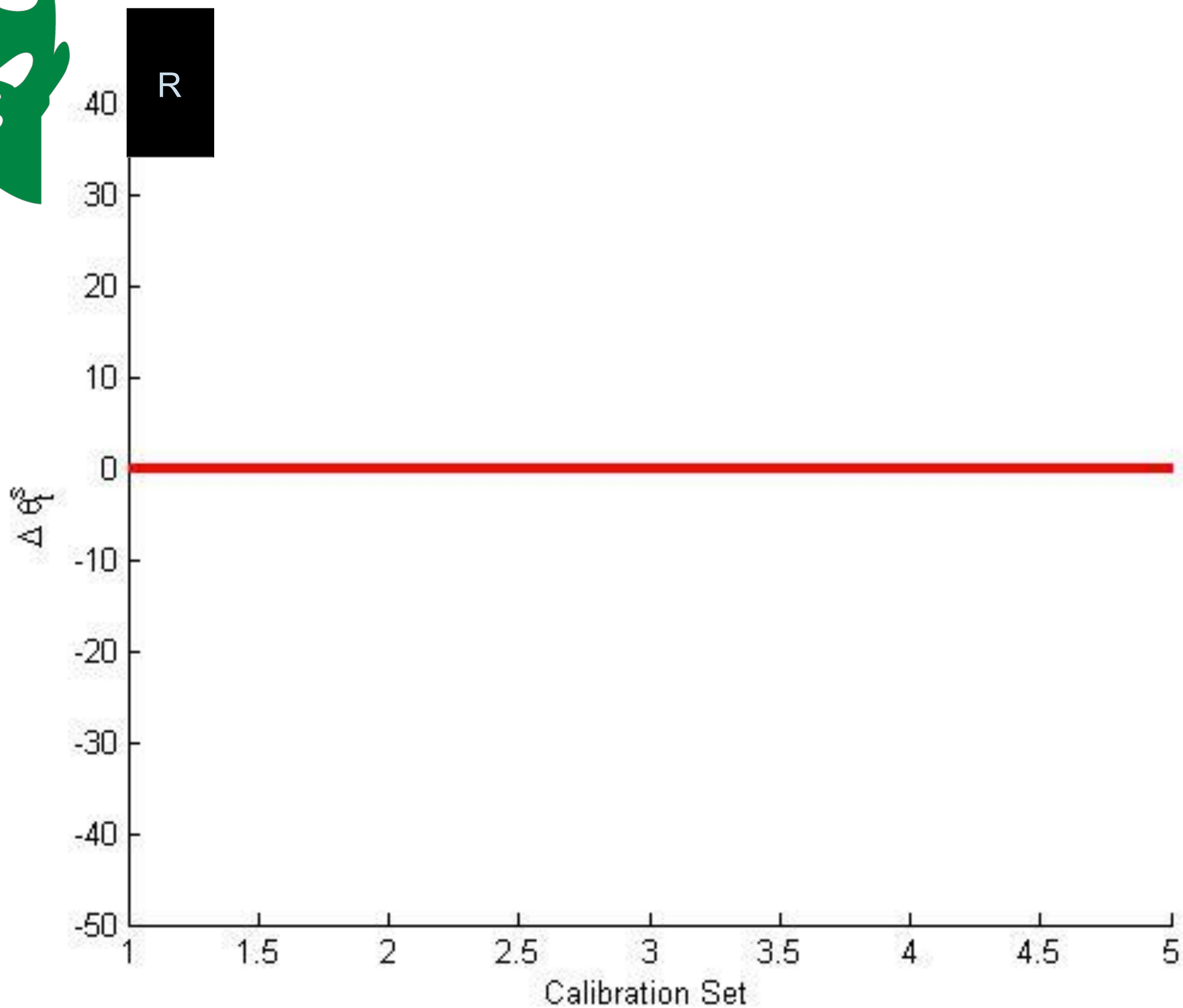
| | |
|---|---|
| 2 | Angry |
| 3 | Sad |
| 0 | Neutral |
| 0 | Happy |
| 0 | Other |

$$Agreement(\theta) = \frac{1}{N}\sum_{i=1}^{N} acos \frac{\overrightarrow{V_{(i)}} \cdot \widehat{V}_i}{\|\overrightarrow{V_{(i)}}\| \|\widehat{V}_i\|}$$

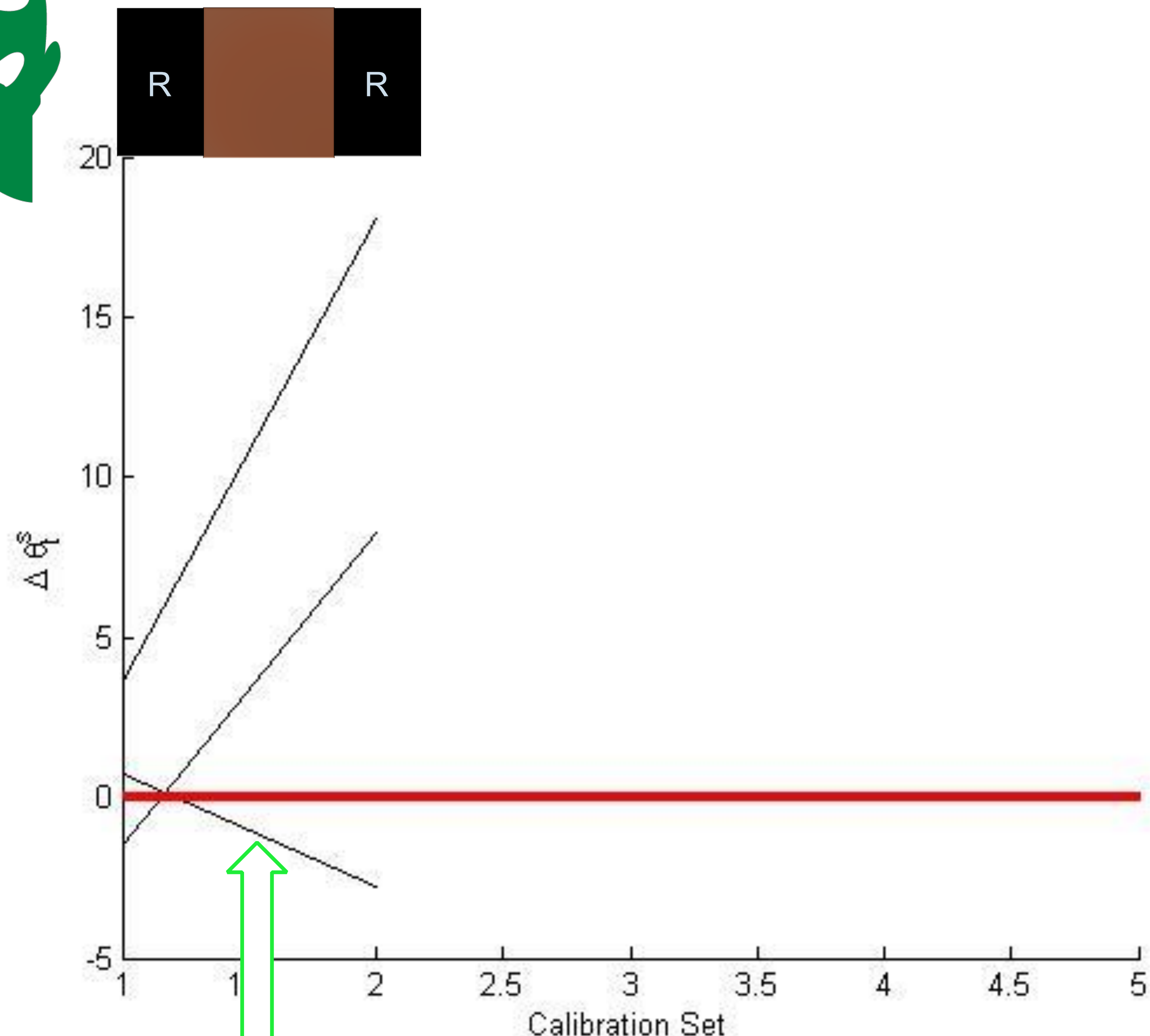
- Assume the rater we are evaluating chooses angry:
- We then recalculate the agreement as above and find the difference:

$$\Delta\theta = \theta_t - \theta_s$$

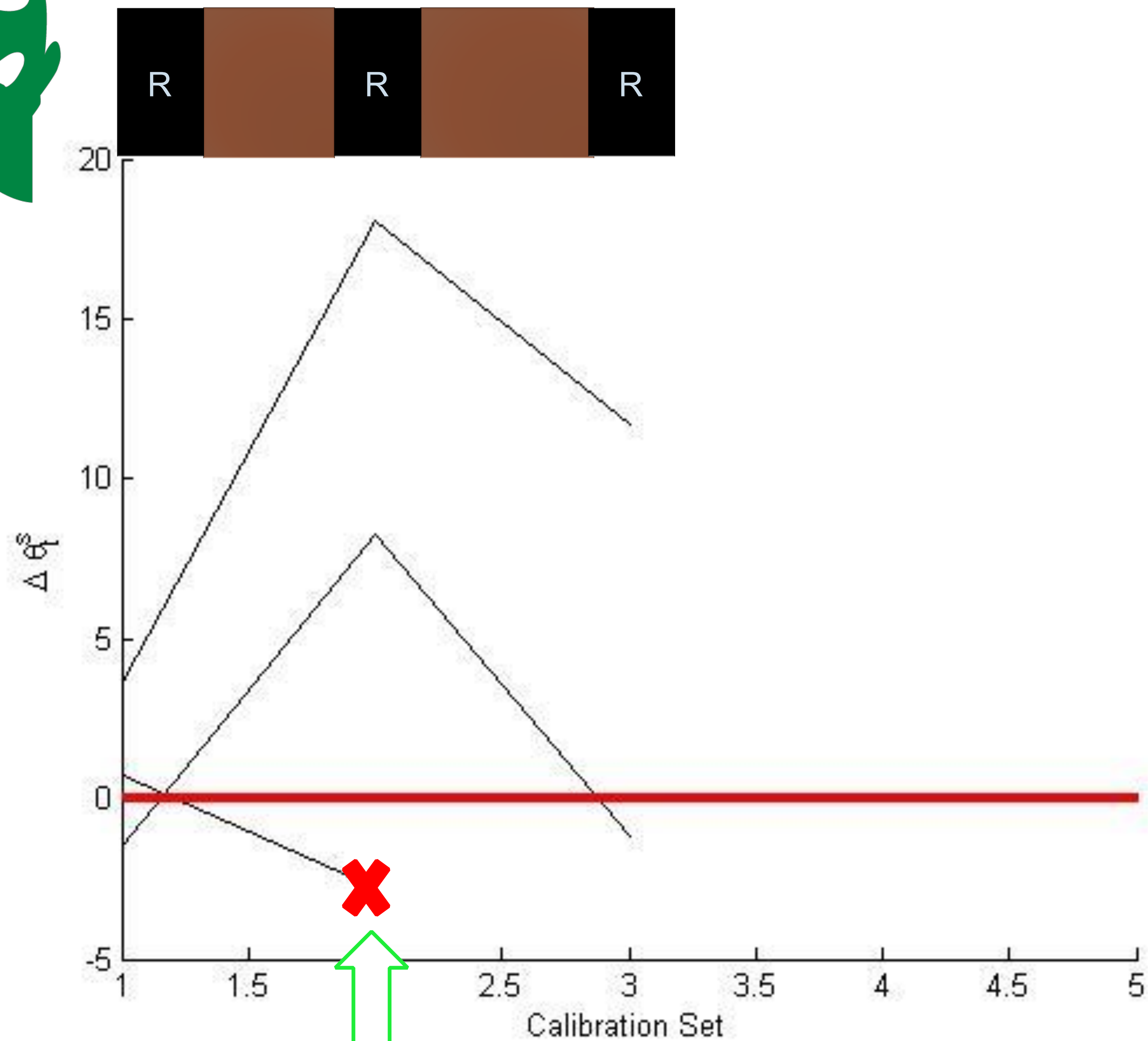
| | |
|---|---|
| 2+1 | Angry |
| 3 | Sad |
| 0 | Neutral |
| 0 | Happy |
| 0 | Other |



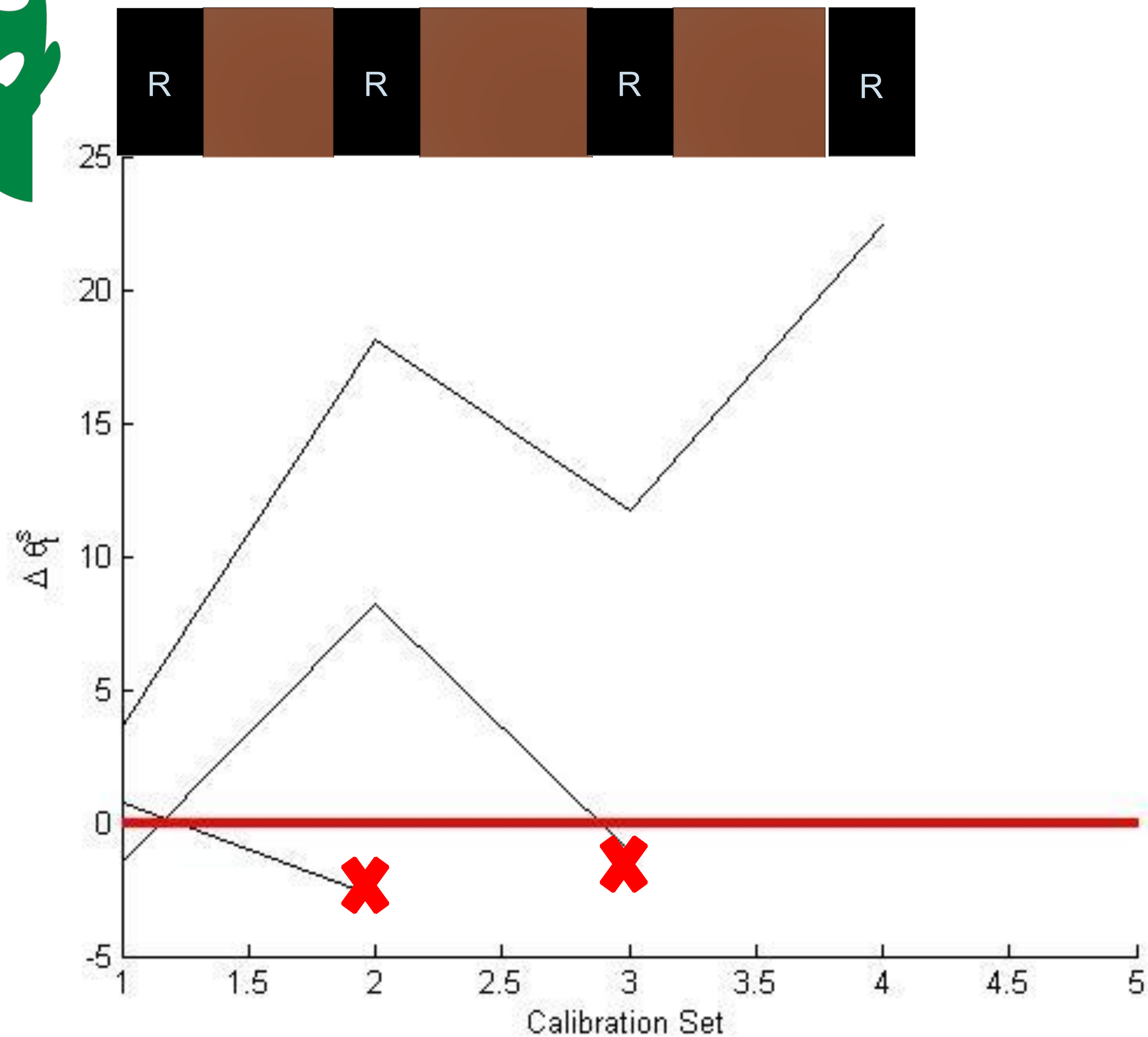

Average Difference
of
Gold Standard

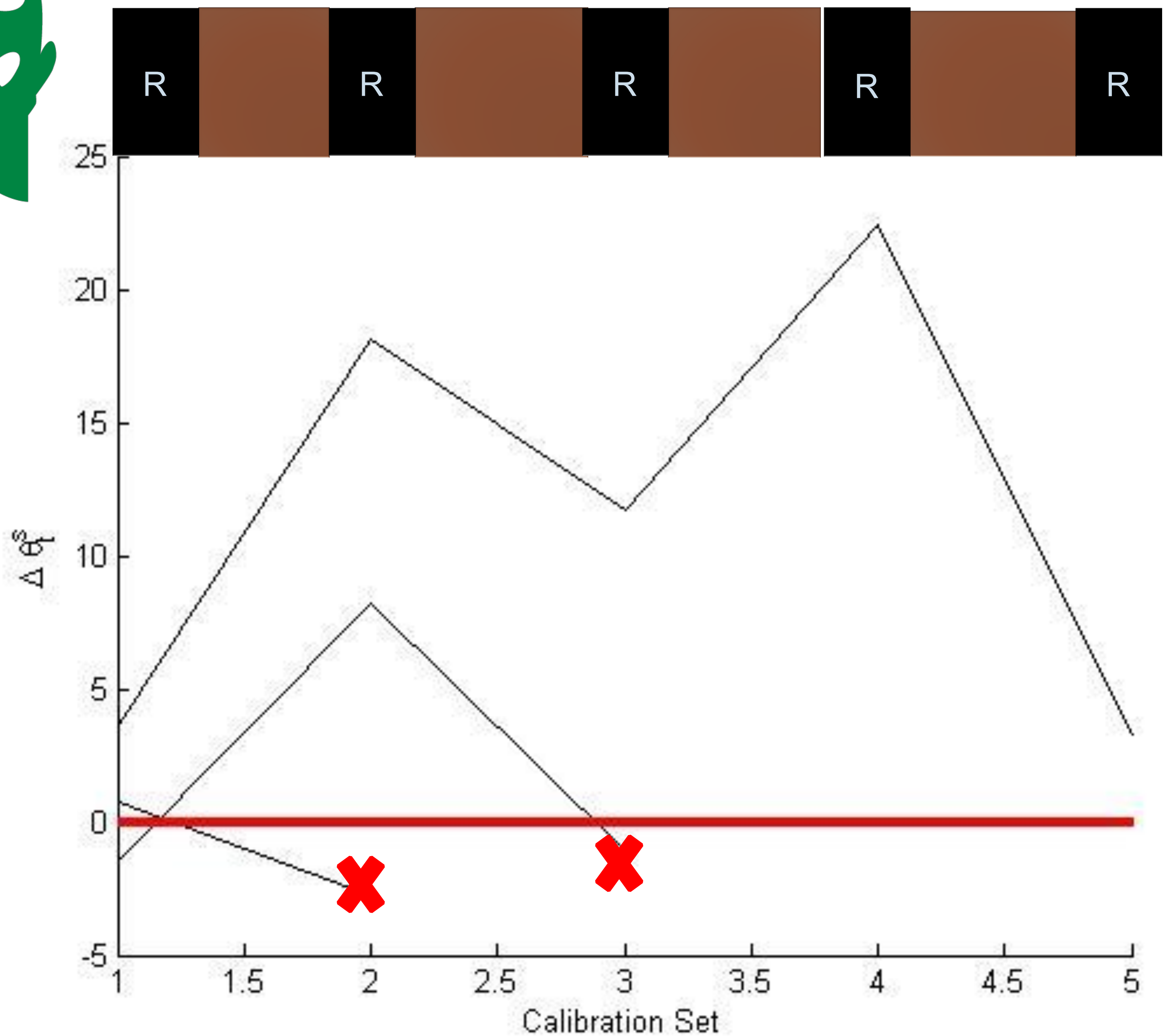


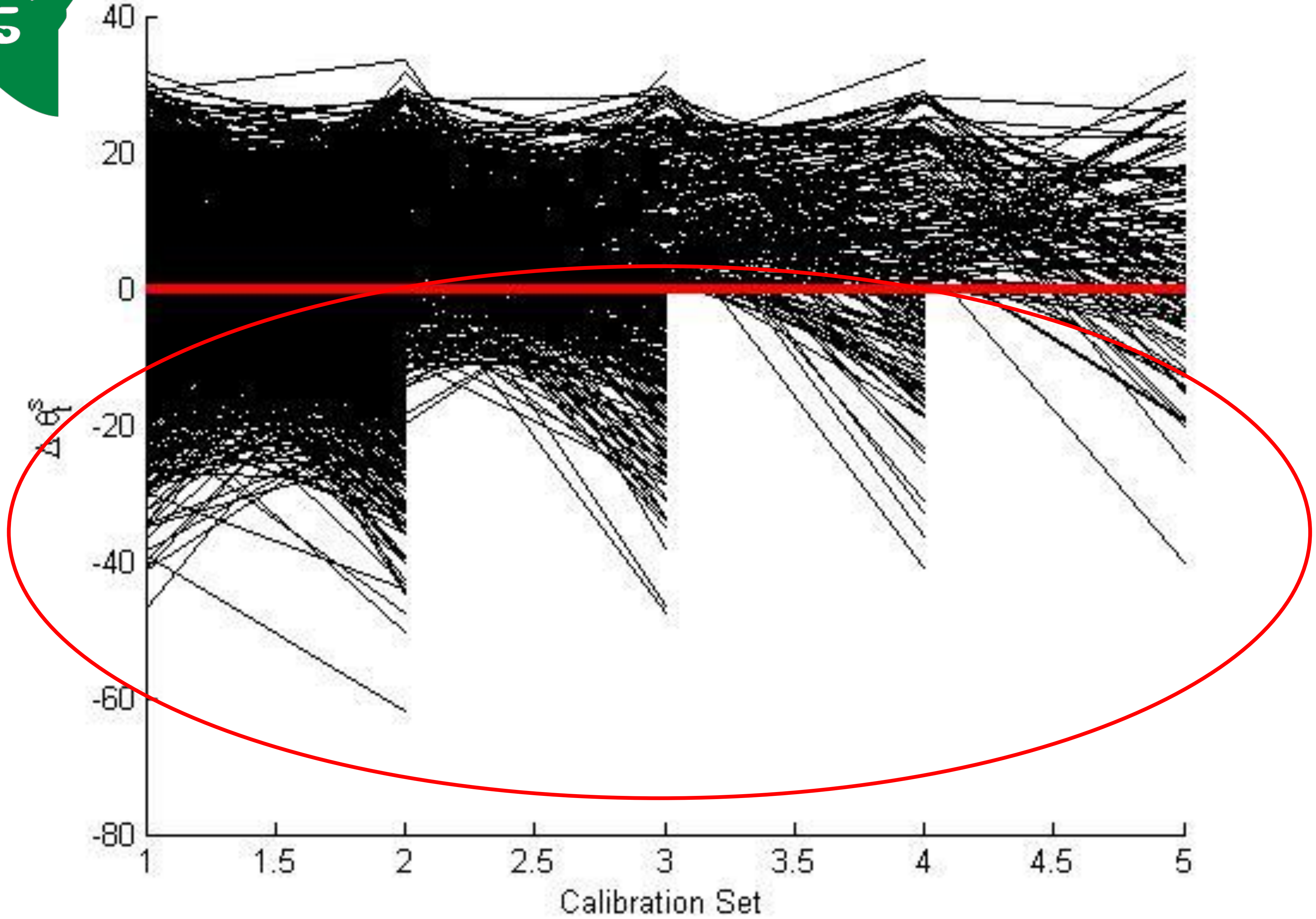

Performance Averaged over first two sets




First Group of Evaluators Removed


R
R
R
R
25
20
15
10
5
0
-5
1
1.5
2
2.5
3
3.5
4
4.5
5
Calibration Set

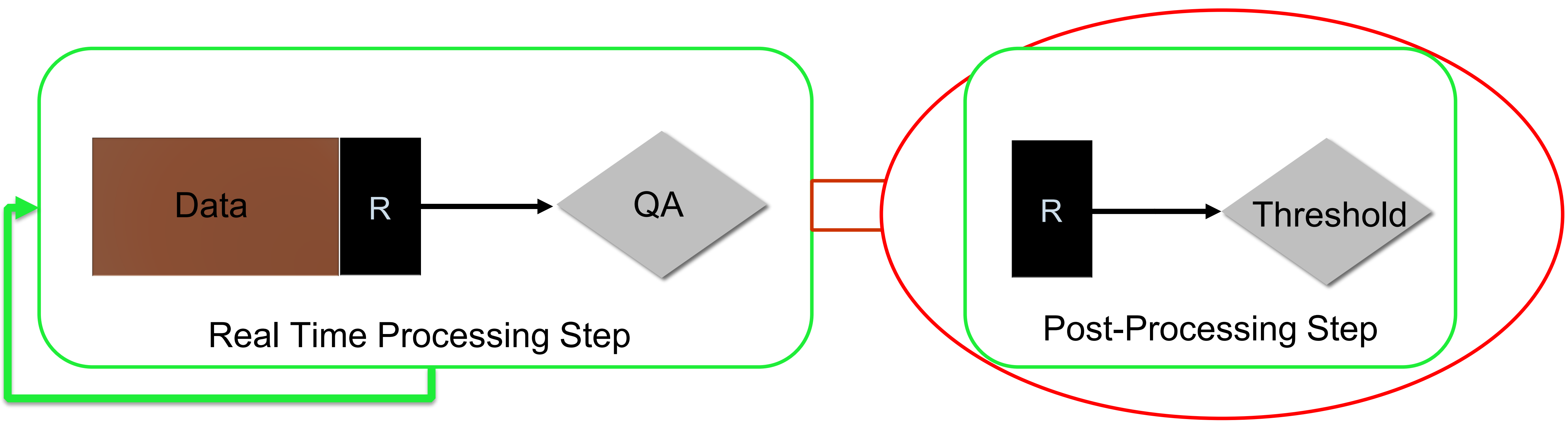

R
R
R
R
R
25
20
15
10
5
0
-5
1
1.5
2
2.5
3
3.5
4
4.5
5
Calibration Set



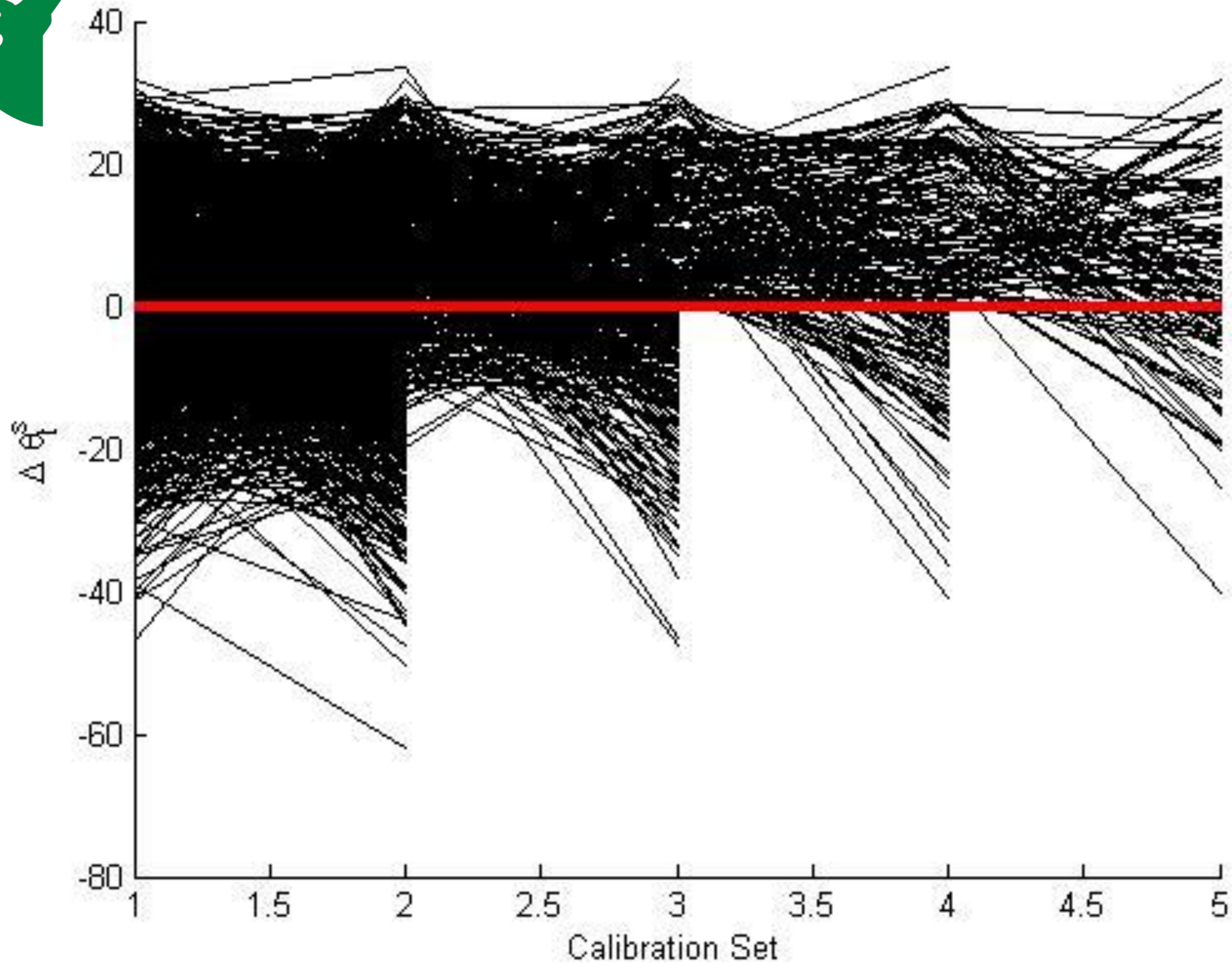

This is still an issue!

# Offline Filtering Process

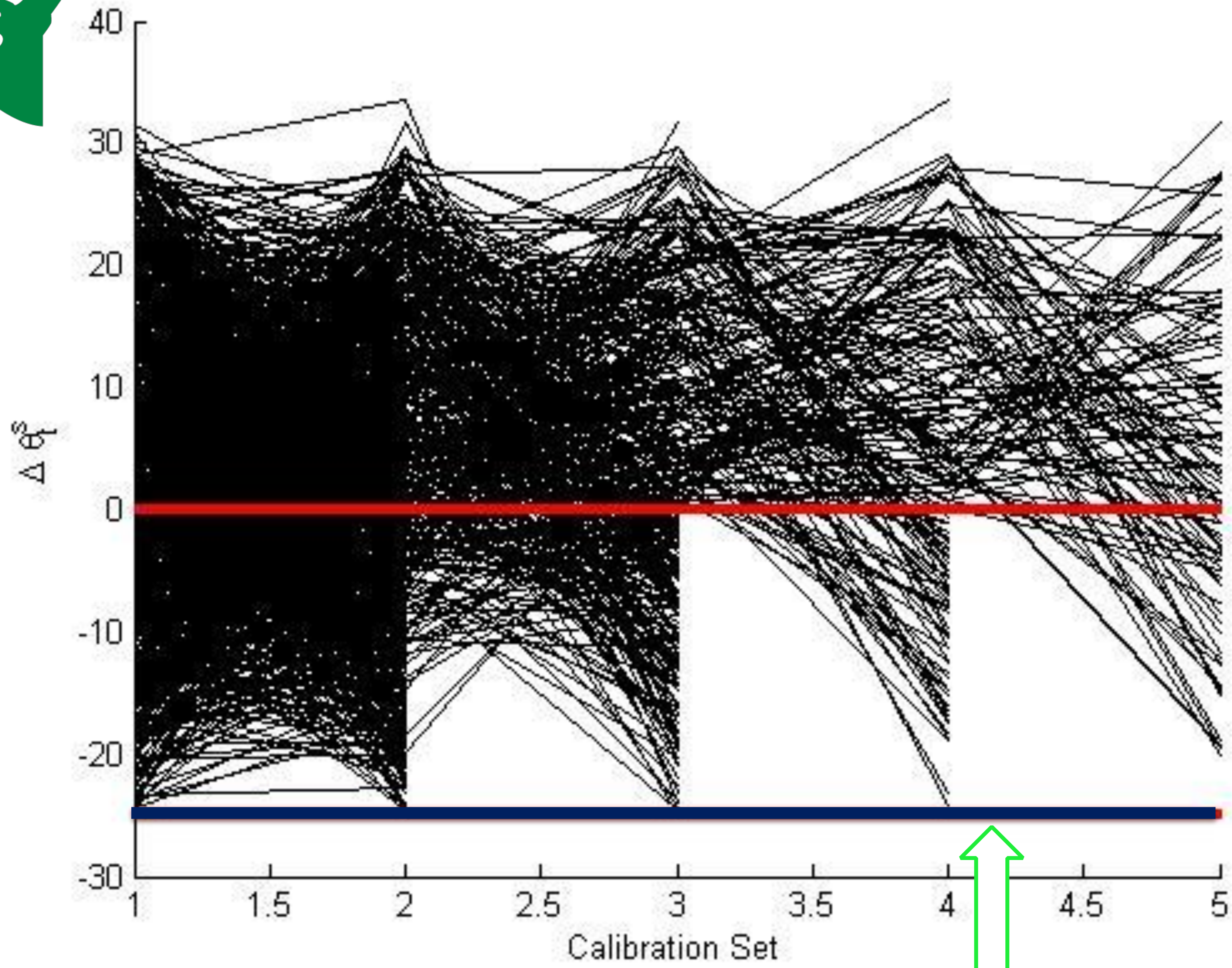
- Because we have the quality at each of the checkpoints, we can filter results that fall below a certain threshold
  - This gives us target sets with an average of number of evaluations >20
- Thus we can filter to have sets with different inter-evaluator agreement
- We choose Angular agreement as our metric (useful for minority emotions)




We can control this to produce sets of varying quality

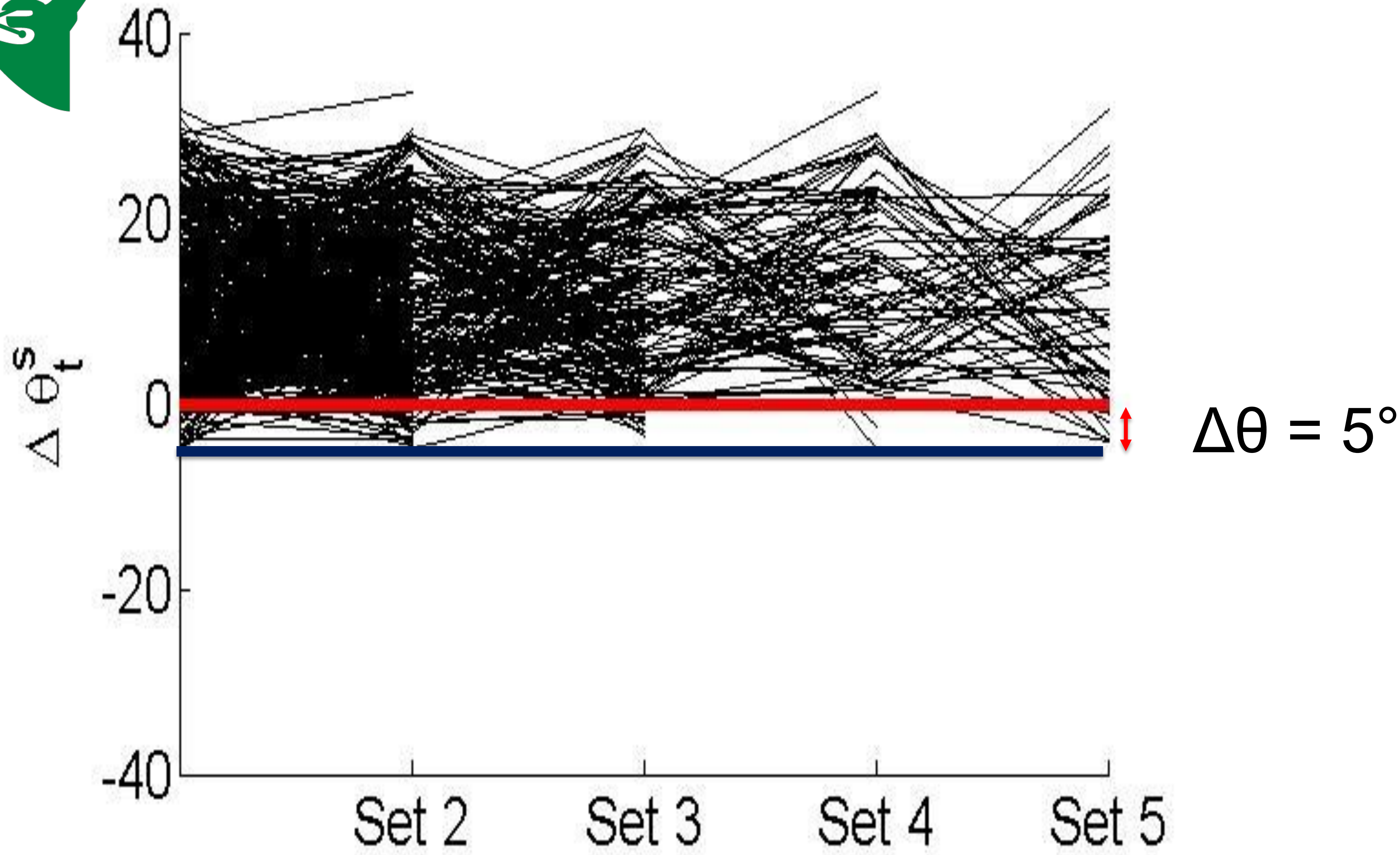


40
20
0
-20
-40
-60
-80
Δ $\theta_i^s$
1
1.5
2
2.5
3
3.5
4
4.5
5
Calibration Set




$\Delta\theta = 25°$

Secondary
Post-processing threshold ($\Delta\theta$)


40
20
0
-20
-40
$\Delta\ \theta_t^s$
Set 2
Set 3
Set 4
Set 5
$\Delta\theta = 5°$

# Rater Quality

Constant sample size

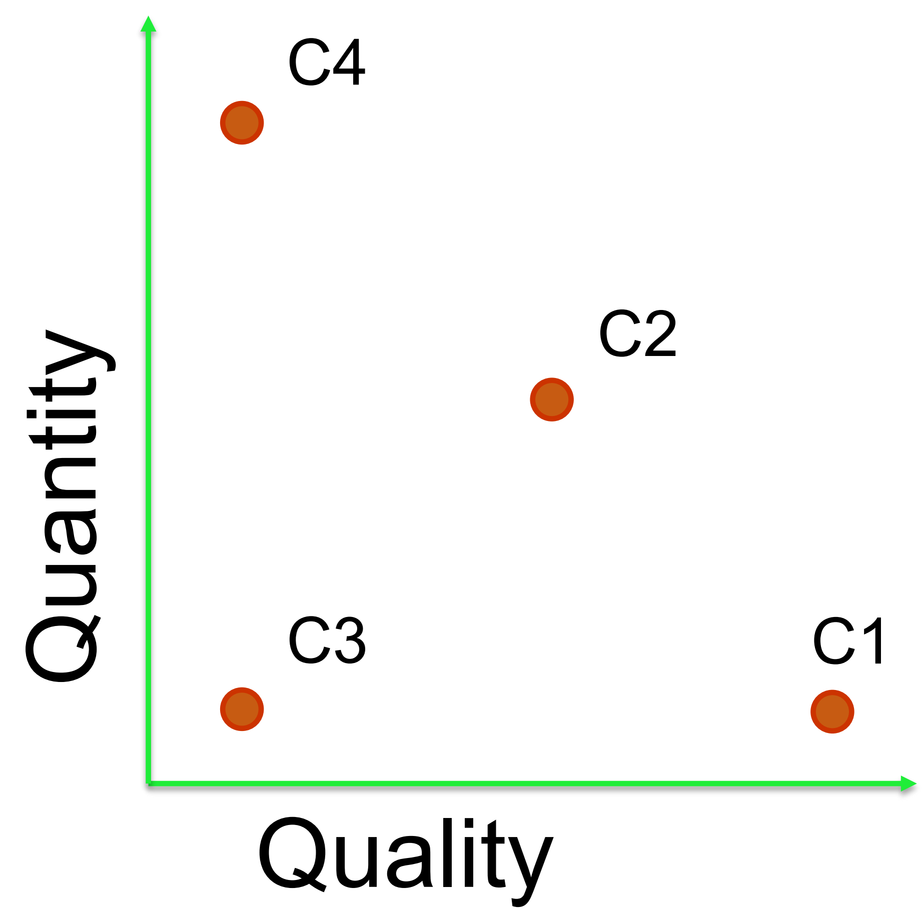
| | 5 Raters | | 10 Raters | | 15 Raters | | 20 Raters | | 25 Raters | |
|---|---|---|---|---|---|---|---|---|---|---|
| Δθ | # sent | κ | # sent | κ | # sent | κ | # sent | κ | # sent | κ |
| 5 | 638 | **0.572** | 525 | 0.558 | 246 | 0.515 | 52 | 0.488 | 0 | - |
| 10 | 643 | 0.532 | 615 | 0.522 | 466 | 0.501 | 207 | 0.459 | 26 | 0.455 |
| 15 | 648 | 0.501 | 643 | 0.495 | 570 | 0.483 | 351 | 0.443 | 112 | 0.402 |
| 20 | 648 | 0.469 | 648 | 0.471 | 619 | 0.463 | 510 | 0.451 | 182 | 0.414 |
| 25 | 648 | 0.452 | 648 | 0.450 | 643 | **0.450** | 561 | 0.440 | 247 | 0.416 |
| 30 | 648 | 0.438 | 648 | 0.433 | 648 | 0.436 | 609 | 0.431 | 298 | 0.410 |
| 35 | 648 | 0.425 | 648 | 0.433 | 648 | 0.426 | 619 | 0.424 | 346 | 0.403 |
| 40 | 648 | 0.420 | 648 | 0.427 | 648 | 0.425 | 629 | 0.423 | 356 | 0.402 |
| 90 | 648 | **0.422** | 648 | 0.419 | 648 | 0.422 | 629 | **0.419** | 381 | 0.409 |

Increasing agreement due to filter

Decreasing samples meeting size criteria

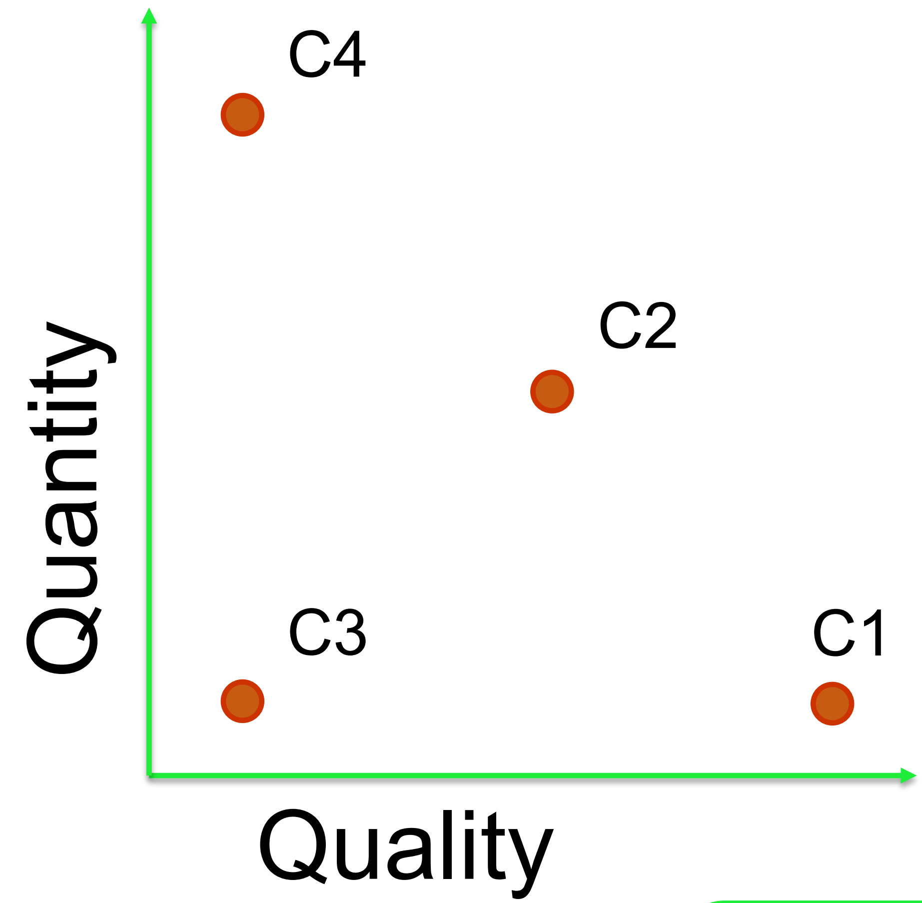
# Experimental Setup

- Let's choose 4 scenarios which tradeoff quality and quantity, asses their effective reliabilities and classification performance
  - **Case 1:** High Quality, Low Quantity
    - 5 degree filter, and 5 Raters (κ = 0.572)
  - **Case 2:** Moderate Quality, Moderate Quantity
    - 25 Degree Filter, 15 raters (κ = 0.450)
  - **Case 3:** Low Quality, Low Quantity
    - No Filter, 5 Raters (κ = 0.422)
  - **Case 4:** Low Quality, High Quantity
    - No Filter, 20 Raters (κ = 0.419)



# Classification

- Five Class Problem (Angry, Sad, Neutral, Happy, Other)
  - Excluded turns w/o majority vote agreement
  - Acoustic Features IS 2013 - OPENSMILE



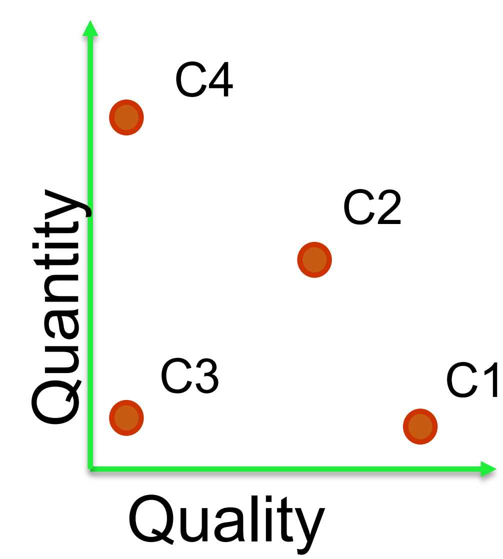

Feature Extraction → CAE Feature Selection → Forward Feature Selection → SVM Classifier → 6F-SI Cross Validation

D = 6373 | D = 1000 | D = 50

# Results

| | Common Turns in all Cases | | | | |
|---|---|---|---|---|---|
| | # Turns | Acc. (%) | Pre. (%) | Rec. (%) | F-score(%) |
| Case 1 | 514 | 47.39 | 46.53 | 47.39 | 46.96 |
| **Case 2** | 514 | **48.23** | **47.42** | **48.23** | **47.82** |
| Case 3 | 514 | 47.07 | 46.62 | 47.07 | 46.84 |
| Case 4 | 514 | 47.88 | 47.17 | 47.88 | 47.52 |

| | **EF Reliability** | **Reliability Rank** | **F-Score Rank** |
|---|---|---|---|
| Case 1 | 87 | 3 | 3 |
| Case 2 | 92 | 2 | **1** |
| Case 3 | 78 | 4 | 4 |
| Case 4 | 94 | 1 | 2 |

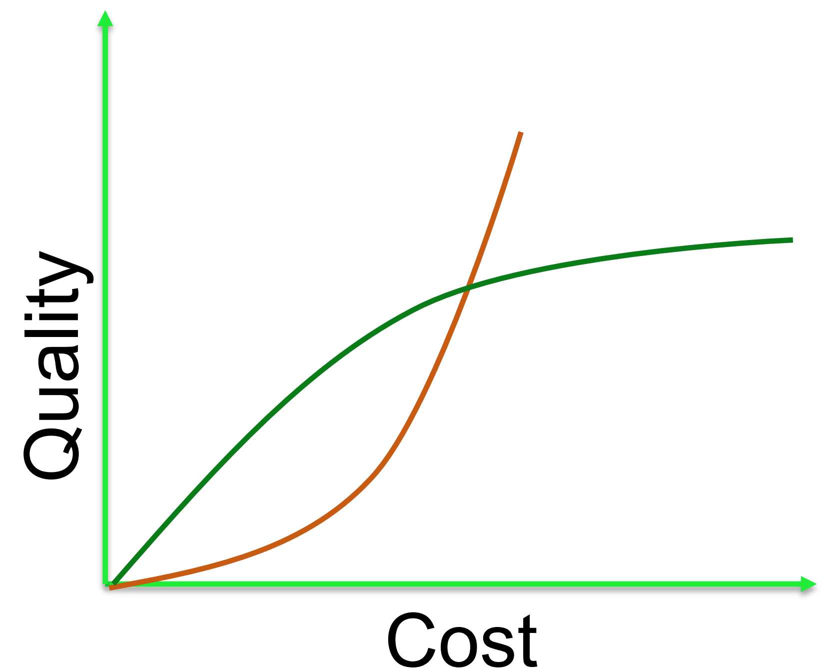


# Discussion

- Relatively small differences appear in labels (<10%)
  - "Wisdom of the crowd" seems to be useful for emotion
- Cost
  - Accuracy desired may be a function of cost
    - Is it worth 4x cost for minor improvement?
    - What is the cost of quality?

**Label Differences**

| | Case 1 | Case 2 | Case 3 | Case 4 |
|---|---|---|---|---|
| Case 1 | - | 26 | **40** | 32 |
| Case 2 | - | - | 32 | 10 |
| Case 3 | - | - | - | 36 |
| Case 4 | - | - | - | - |



# What does this mean?

- We can establish a rough crowdsourcing framework for emotion

# Questions?

Interested in the MSP-IMPROV database?
Come visit us at msp.utdallas.edu and click “Resources”

# References


[1] Jinni A Harrigan, Robert Ed Rosenthal, and Klaus R Scherer,The new handbook of methods in nonverbal behavior research.,Oxford University Press, 2005.

[2]Carlos Busso, Srinivas Parthasarathy, Alec Burmania, Mohammed AbdelWahab, Najmeh Sadoughi, and Emily Mower Provost, "MSP-IMPROV: An acted corpus of dyadic interactions to study emotion perception," IEEE Transactions on Affective Computing, vol. To appear, 2015.

[3] Alec Burmania, Srinivas Parthasarathy, and Carlos Busso, "Increasing the reliability of crowdsourcing evaluations using online quality assessment," IEEE Transactions on Affective C omputing, vol. To appear, 2015.