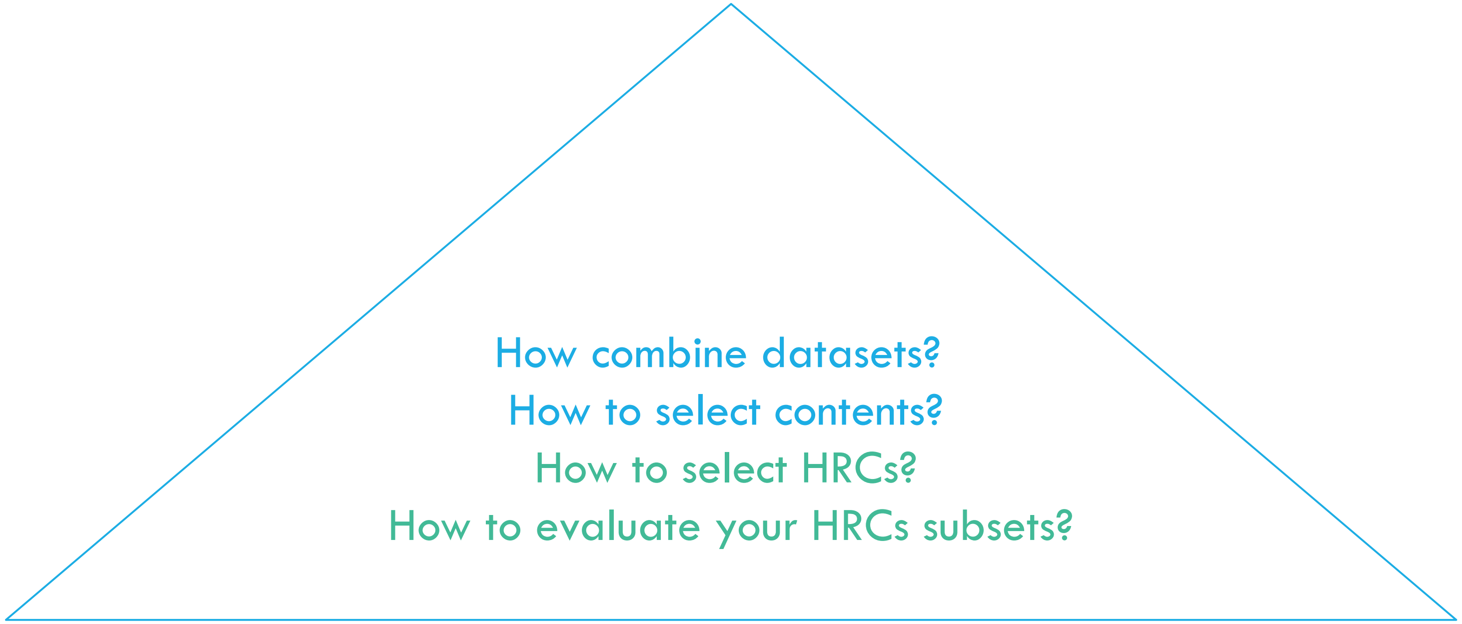
FROM LARGE-SCALE TO SMALL-SCALE DATABASE

Ahmed Aldahdooh, Enrico Masala, Glenn Van Wallendael, Marcus Barkowsky, and Patrick Le Callet

VQEG meeting, May 2017

OBJECTIVE



Identify significant HRCs for:

- Subjective experiments
- Machine-learning-based VQA

Introduction to “Improved Performance Measures for Learning-based Video Quality Assessment Algorithms”

LARGE-SCALE DATABASE

Different correlation scores may be obtained when testing an objective video quality (VQ) measurement using two different databases (and cannot really be averaged)

- Lack of content variety in the databases.
- Lack of different HRCs in the experiments.

Go for Large-scale?

- To evaluate objective measurements that is difficult to achieve in subjective assessment due to Limited HRCs.
  - Agreement of objective measures.
  - Not convenient for frame-based analysis? Consistency within a video sequence.
    - The impact of source contents and the encoder parameters are studied.

SMALL-SCALE DATABASE

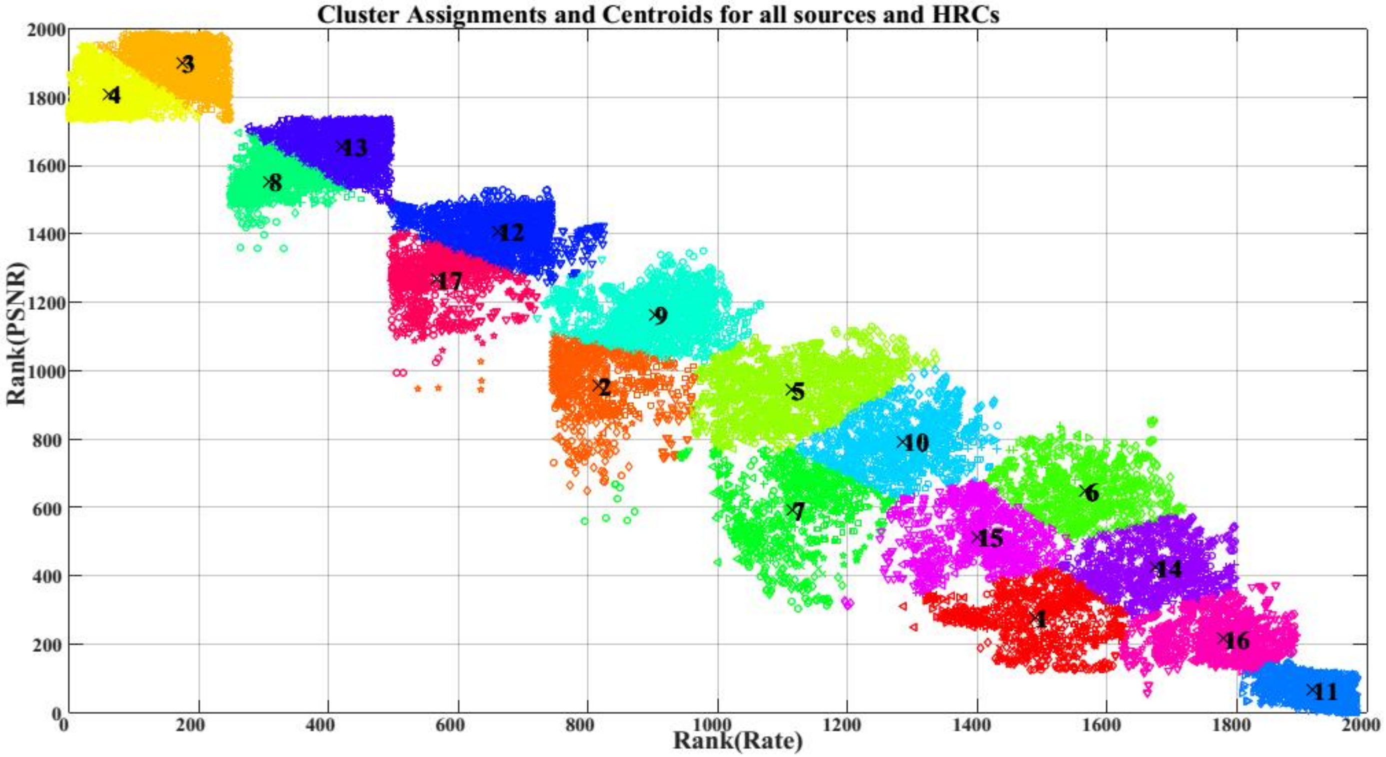
After Analysis: go for small-scale?

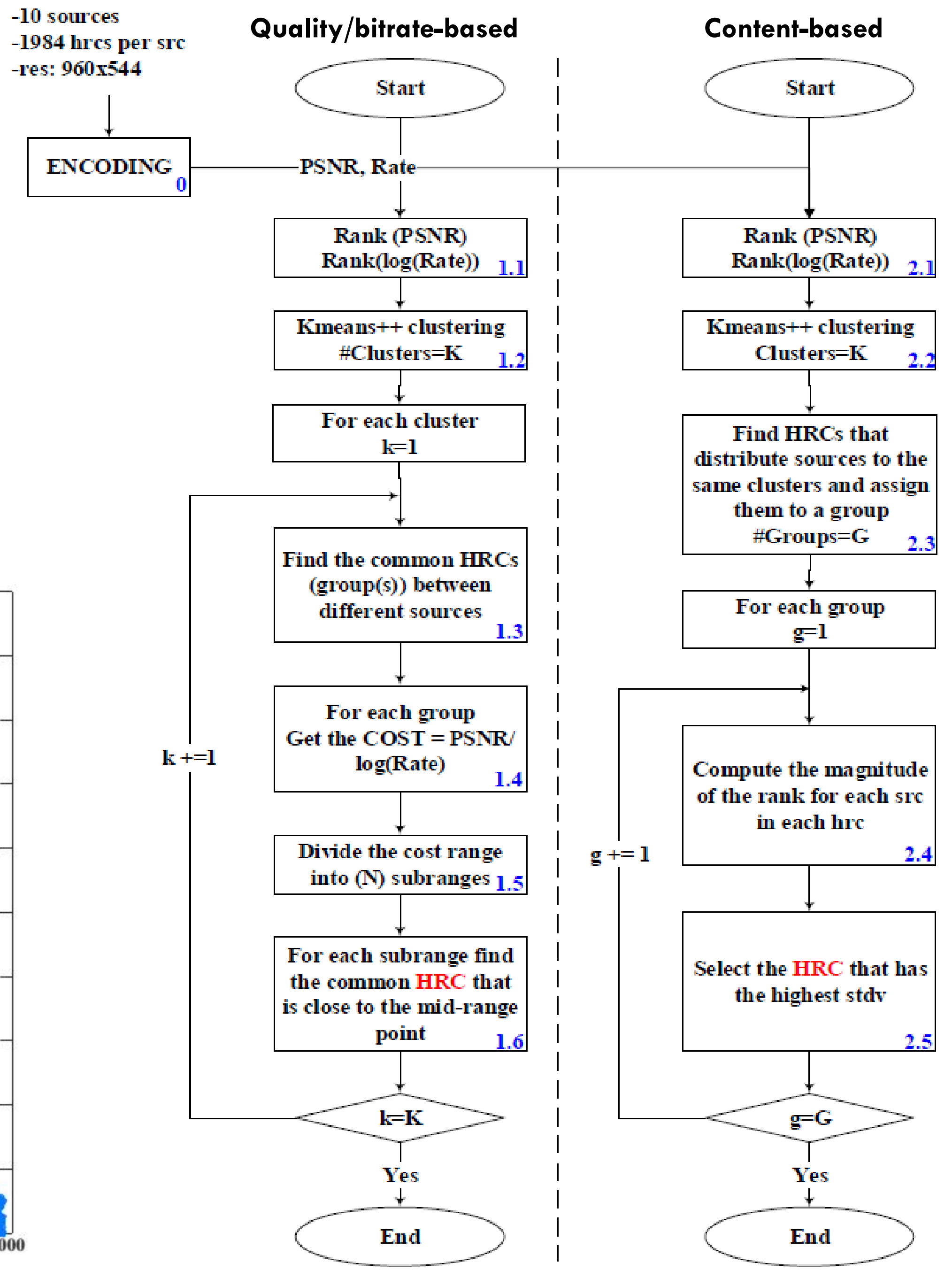
- Identify significant HRCs for:
  - Subjective experiments
  - Machine-learning-based VQA

What we need is to choose HRCs that cover a good variety of targets.

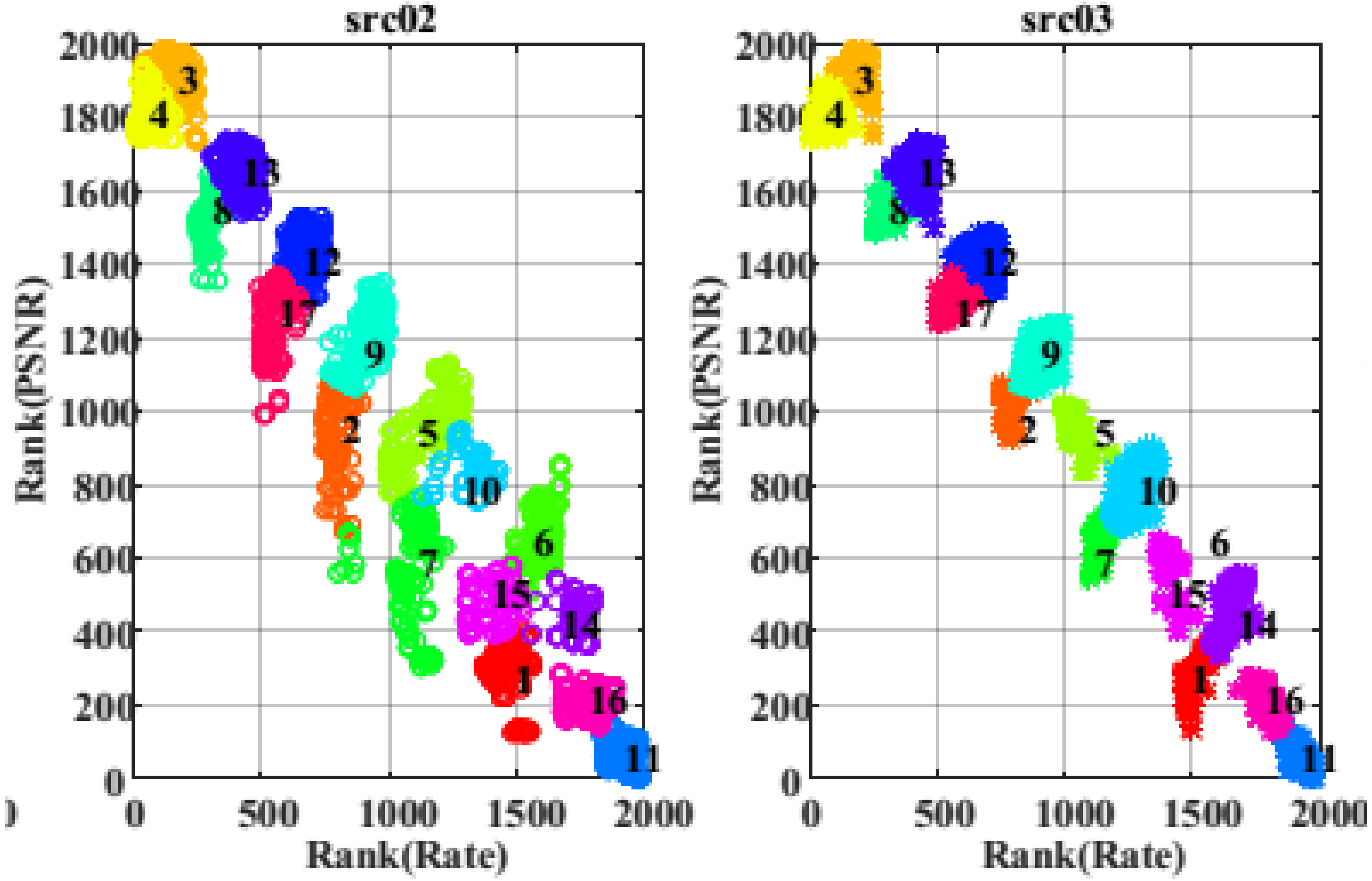
FIRST PART: HRC SELECTION ALGORITHMS

1. QUALITY/BITRATE-DRIVEN HRCS SUBSET
2. CONTENT-DRIVEN HRCS SUBSET

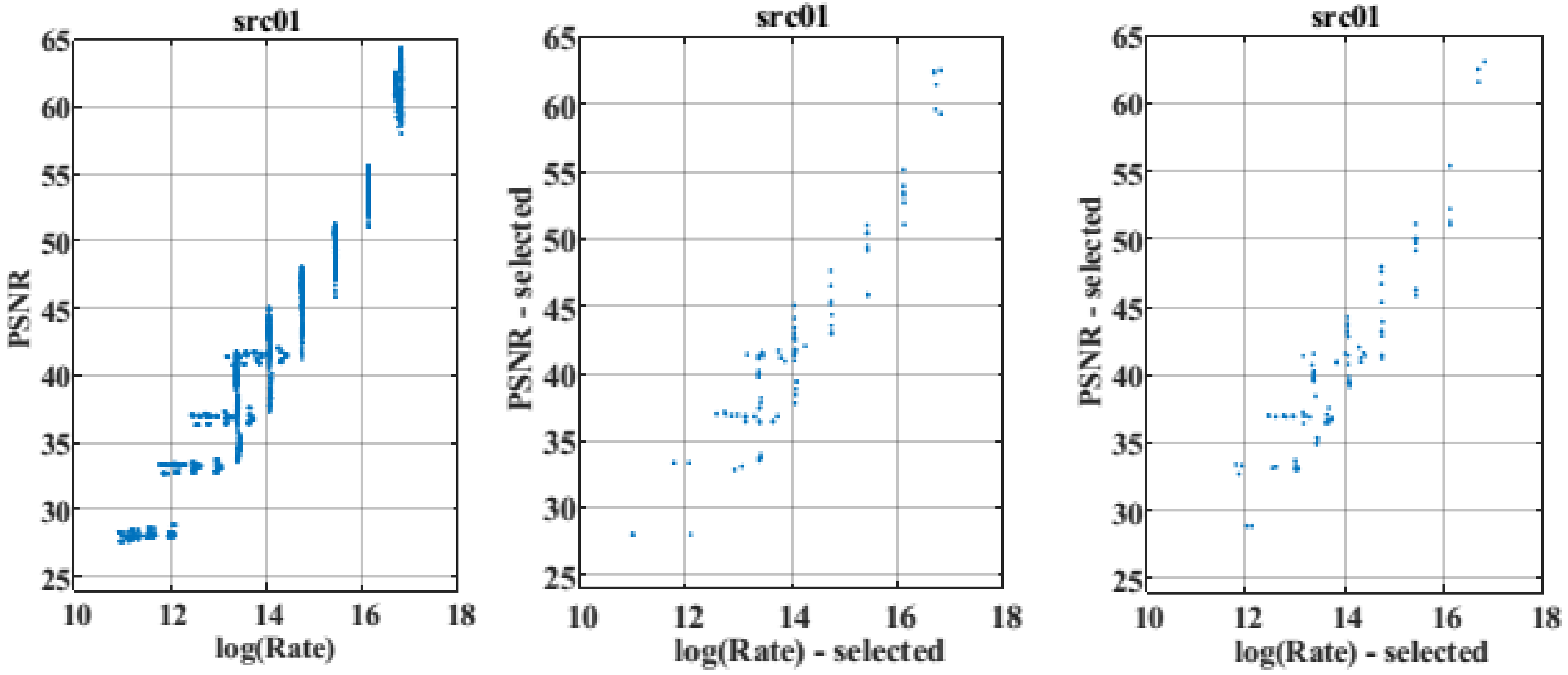




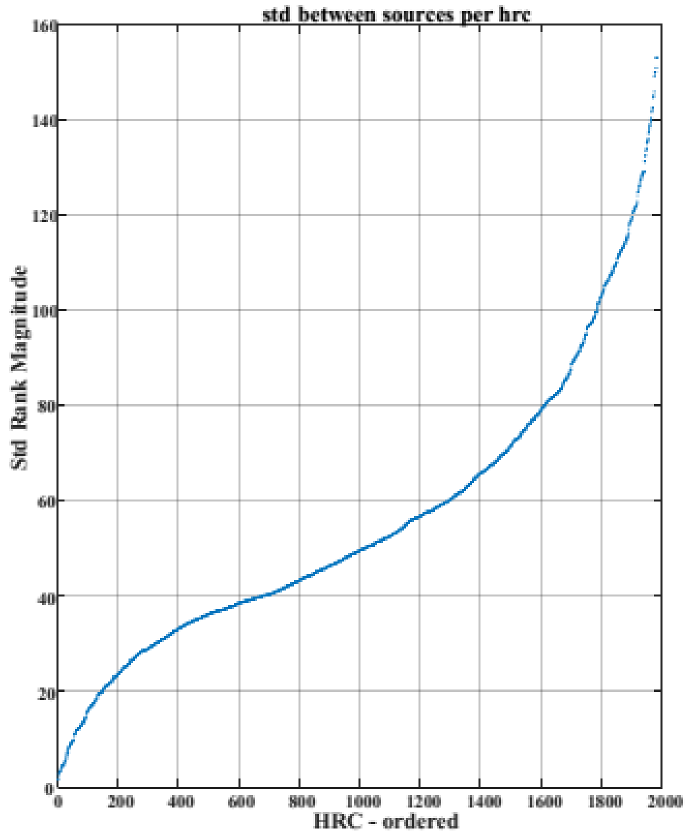
OBSERVATION

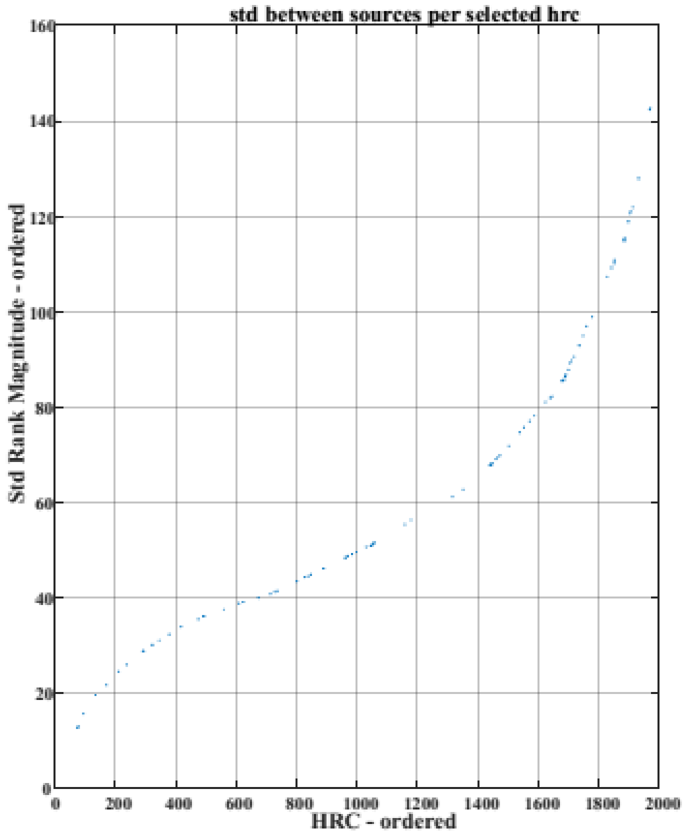


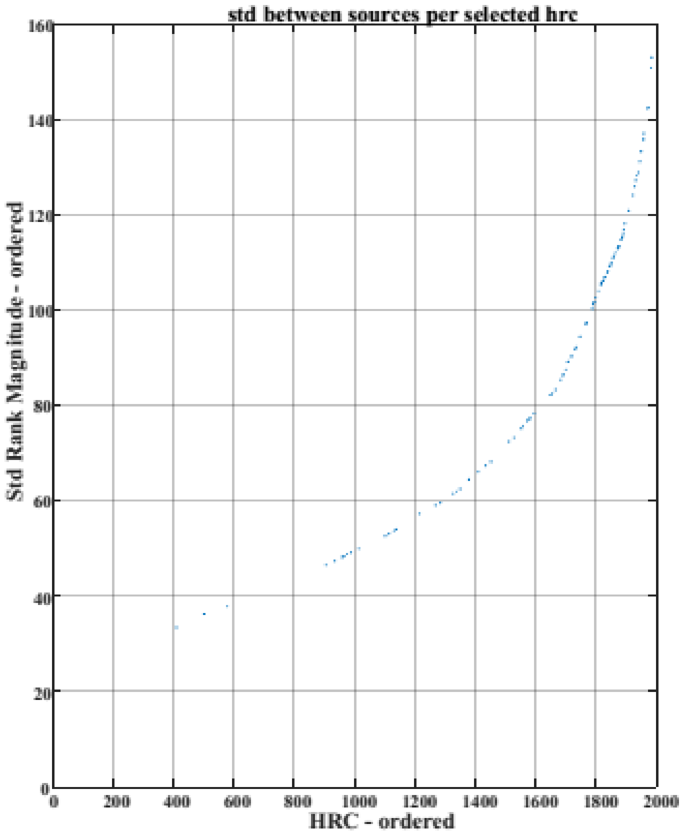
RESULTS – 1


src01
PSNR
log(Rate)
src01
PSNR - selected
log(Rate) - selected
src01
PSNR - selected
log(Rate) - selected

RESULTS – 2 - STD. OF THE RANKS MAGNITUDE







SECOND PART: IMPROVED PERFORMANCE MEASURES

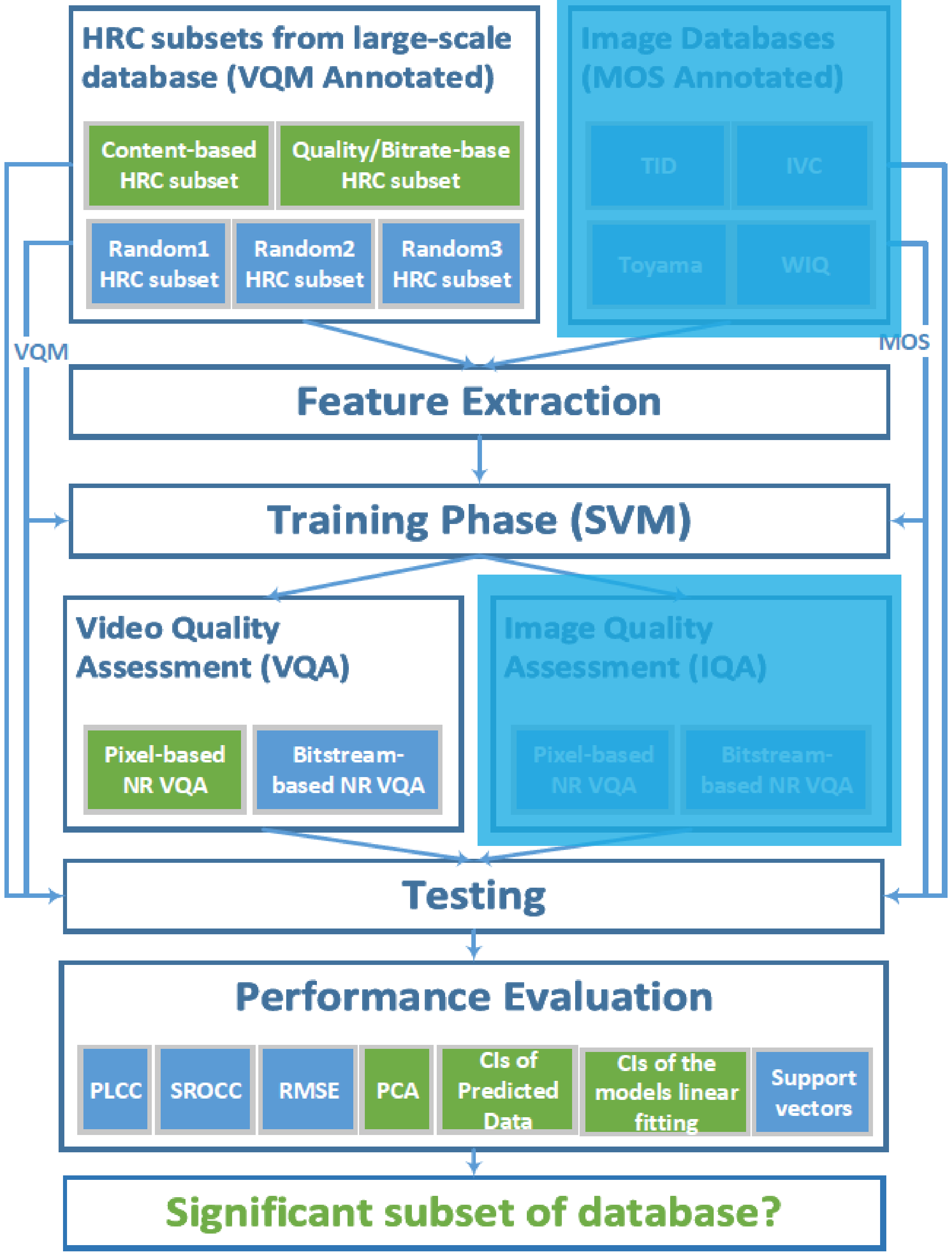
After we have these subset, how do they perform?

Goodness?

EXPERIMENTS STEPS

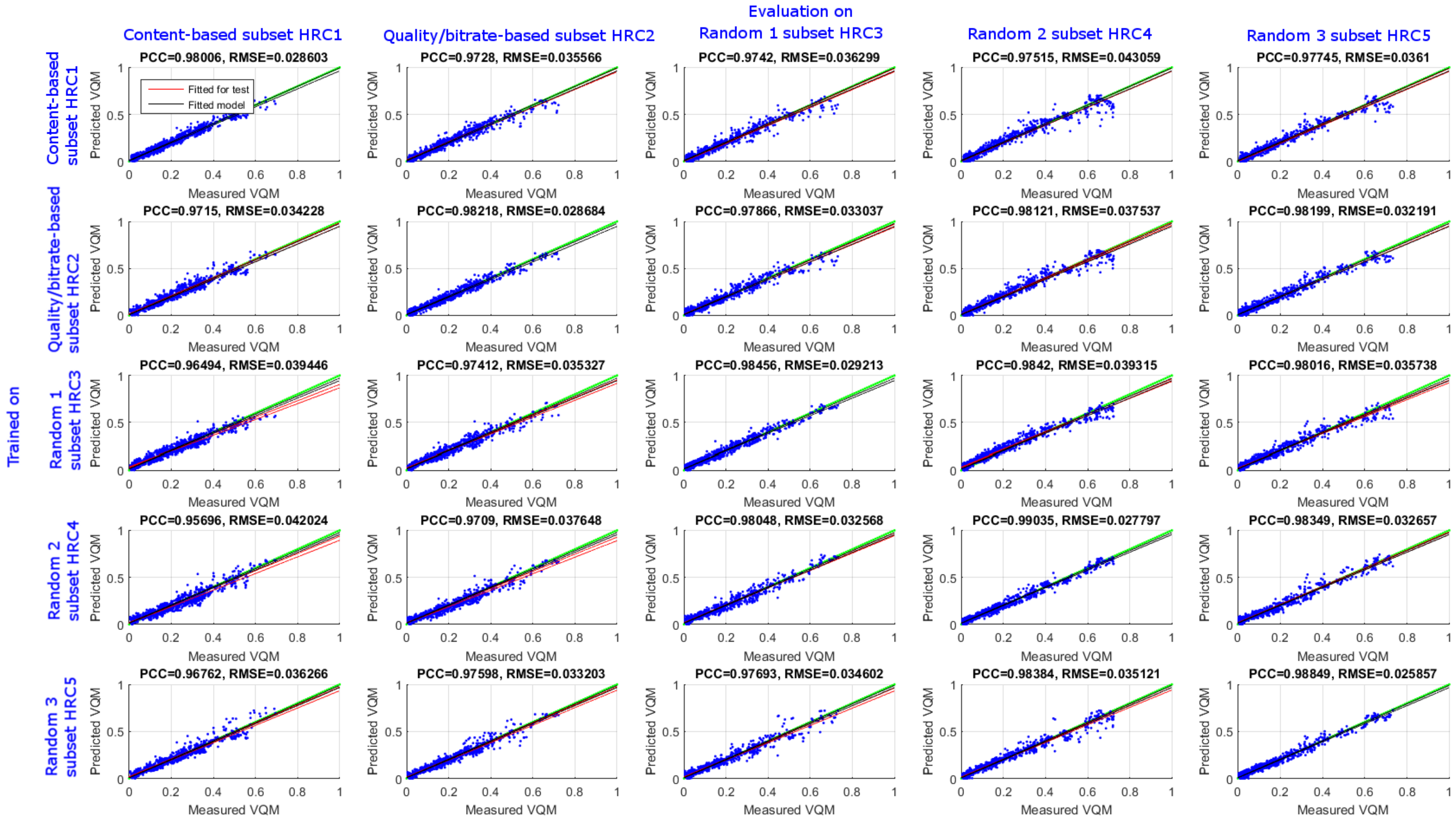
To test the goodness of the elected HRCs subsets.

Not to evaluate the prediction models



SHORTCOMING WITH PLCC AND RMSE

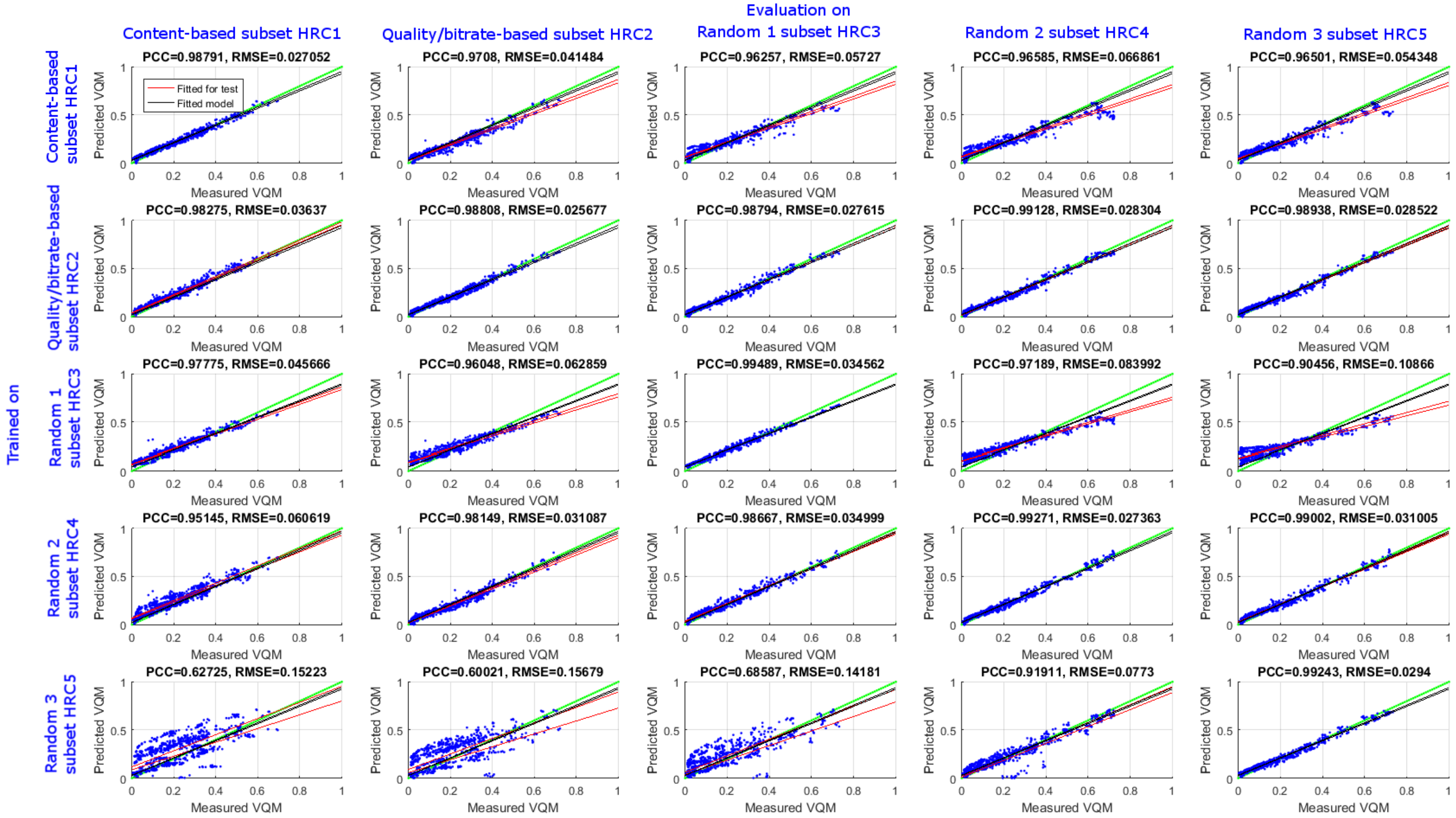
Bitstream-based NR VQA



RMSE & Correlation cannot tell us exactly which HRC set is better

SHORTCOMING WITH PLCC AND RMSE

Pixel-based NR VQA



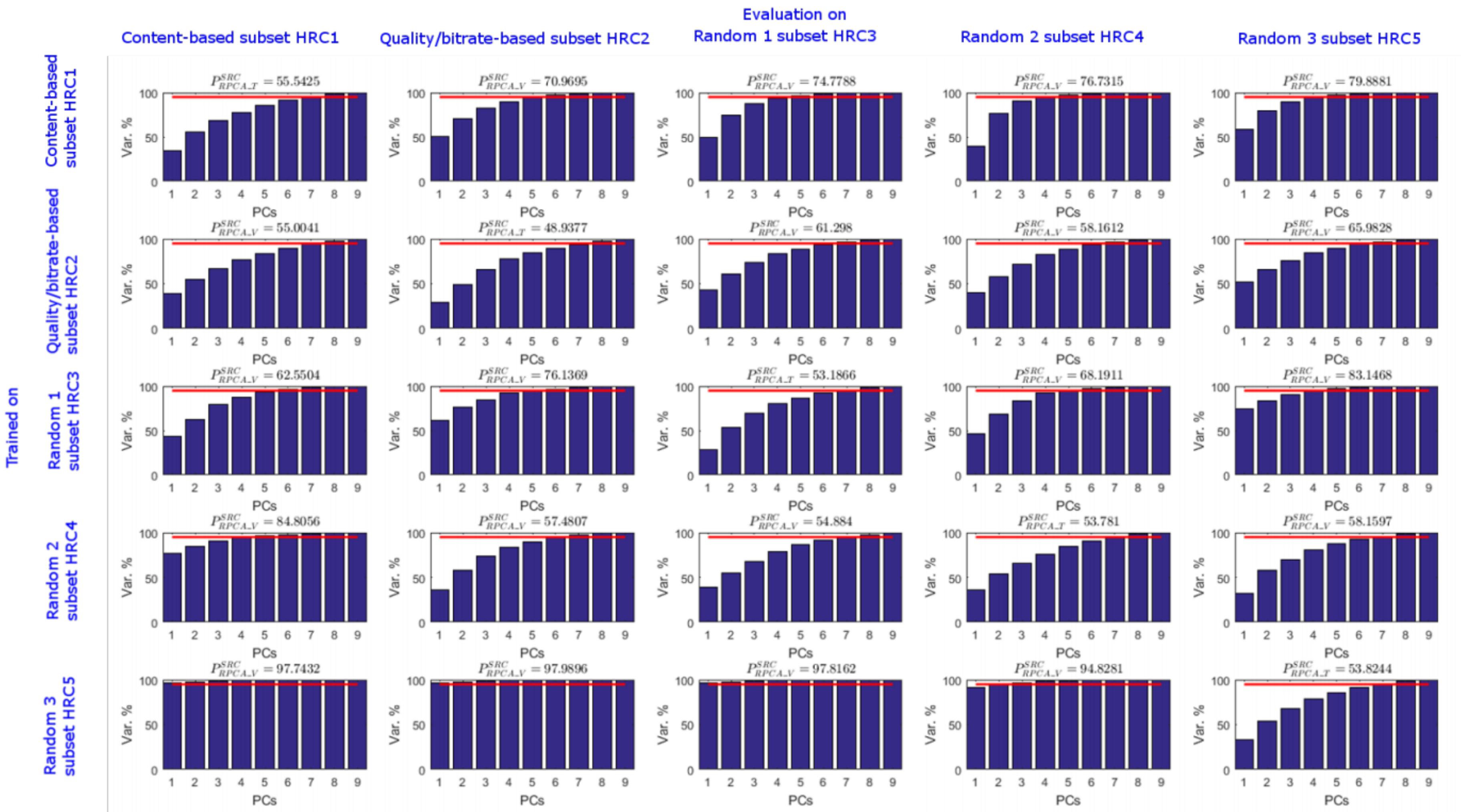
Another Model
RMSE and Correlation can tell us something

(1) RESIDUAL ANALYSIS USING PCA

Pixel-based NR VQA

How residual structured?

Any sign for systematic redundancy in the residual?



(2) CONFIDENCE INTERVALS OF PREDICTED DATA

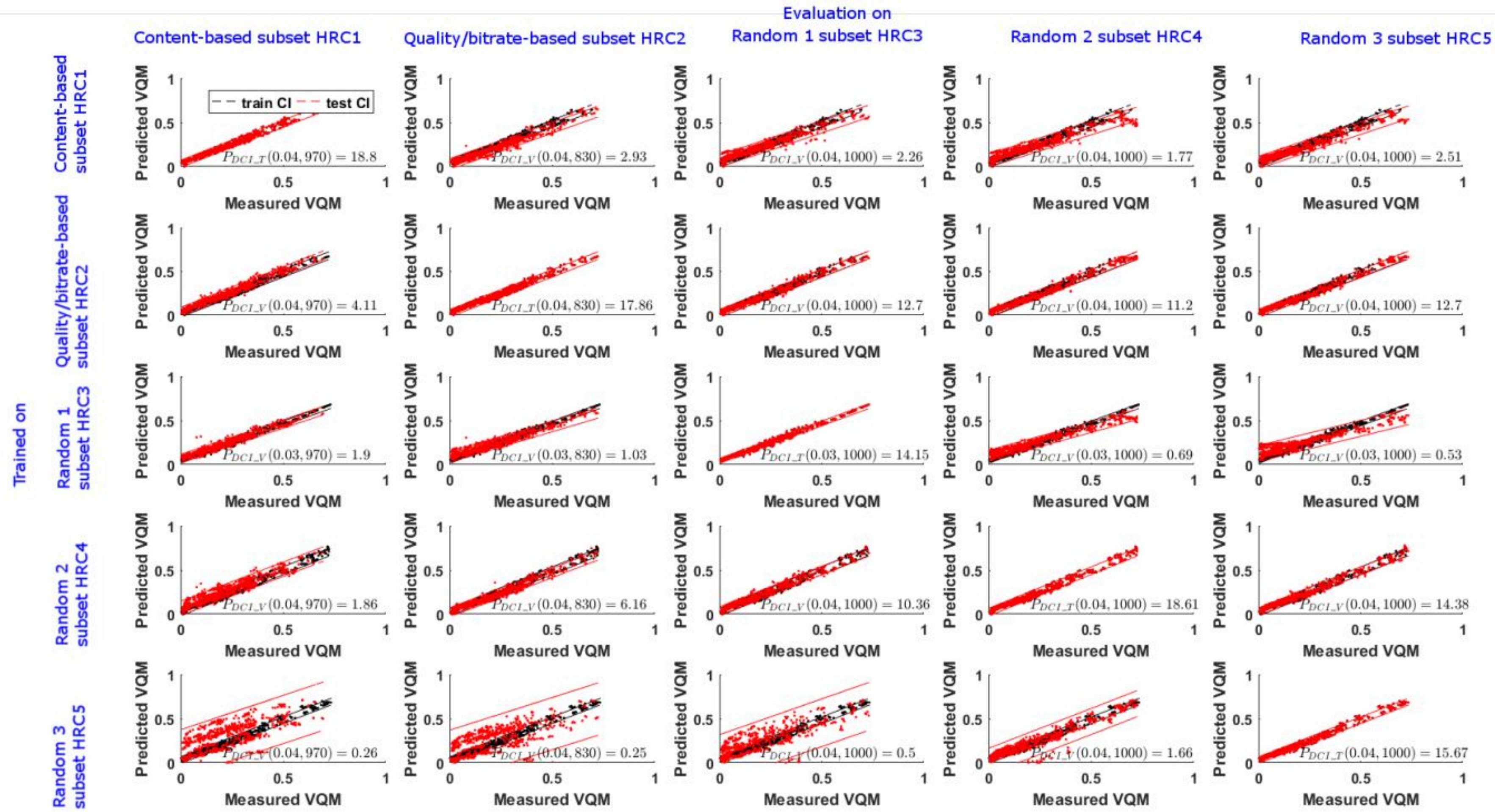
Pixel-based NR VQA

How much of the predicted data lies within CI of the trained model?

<u>**Remember:**</u>

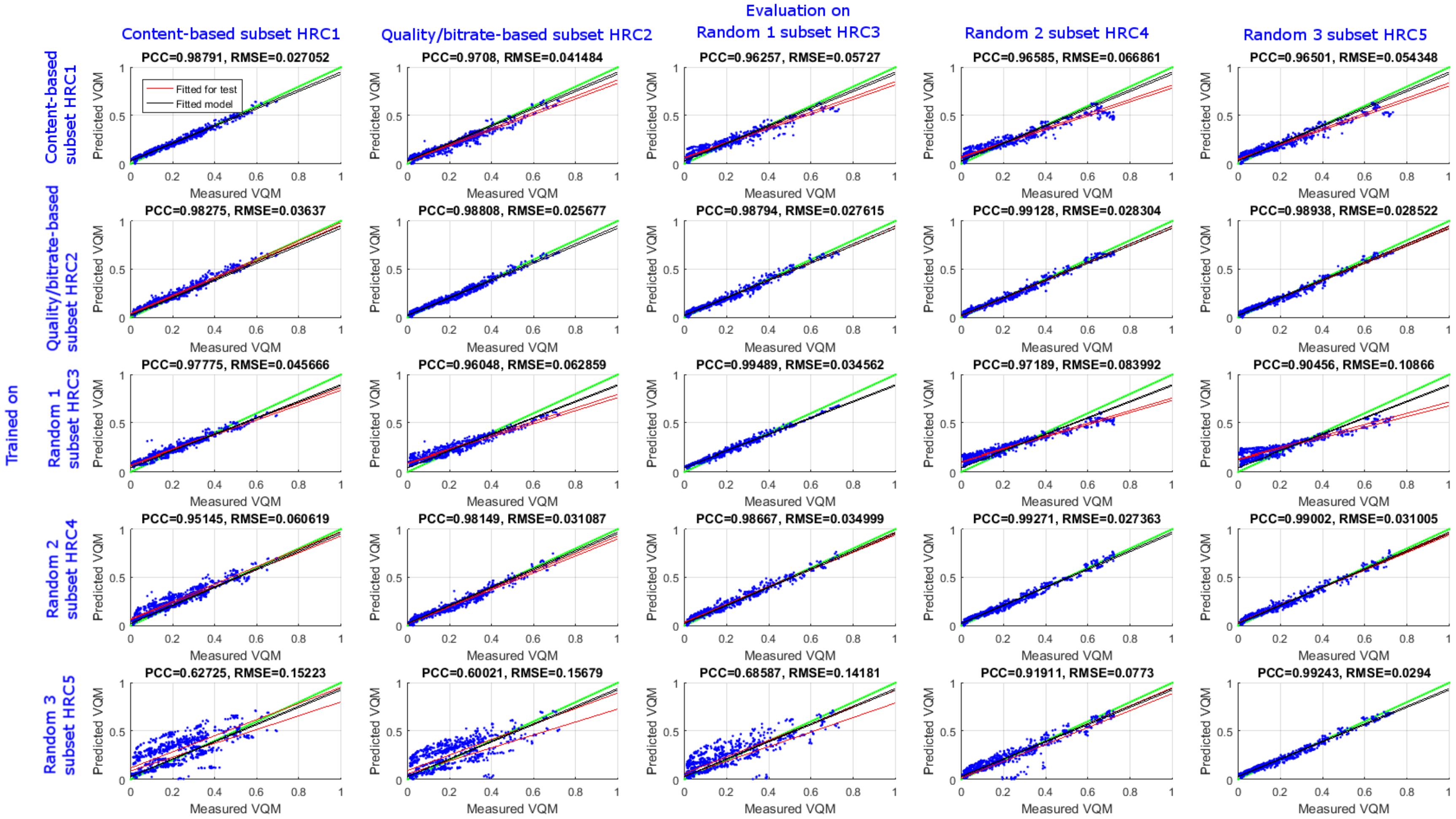
Black lines: CI boundaries of the predicted data of the trained model.

Red Lines: CI boundaries of the predicted data of the validation data



(3) CONFIDENCE INTERVALS OF TRAINED MODELS

Pixel-based NR VQA



Is the model stable when validation data is used?

Remember:

Black lines: CI boundaries of the model coefficient of when training data is used.

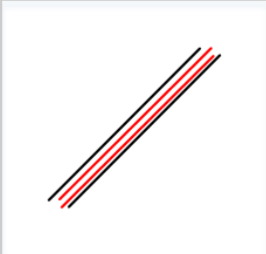
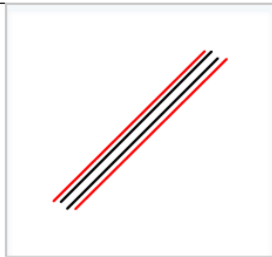
Red Lines: CI boundaries of the model coefficient when validation data is used.

(4) INTERSECTION ANALYSIS

$$G = \frac{i}{\max(b,r)^2}$$

The interaction between black lines and red lines?

The higher the overlap the better.

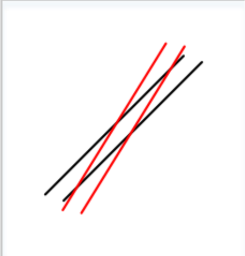
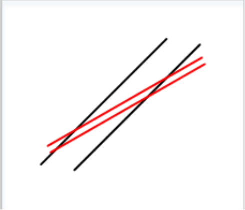
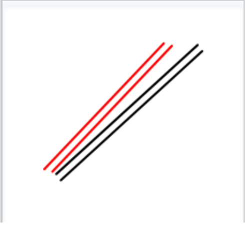
Case	Icon	Condition	Note
1		$b = r = i$	Typical case for validating on the training data, this is considered the perfect fitting, i.e. all three areas are identical. Refer for example to the main diagonal $X(n,n)$ in Fig. 6. In this case, $G = \frac{1}{\max(b,r)}$. To compare between different models or data, the lower the $\max(b,r)$, i.e. the smaller the larger CI, the better.
2		$r = i$	The validation data is better predicted than the training data and the CI lie completely within the boundaries of the trained model. This is likely to be a default of the validation data and thus reduces the goodness as compared to Case 1. In this case, $G = \frac{r}{b^2}$.
3		$b = i$	The validation data is less well predicted than the training data but the validation CI covers completely the training CI. This is considered a case of overfitting of the model and should thus be penalized compared to case 1. In this case, $G = \frac{b}{r^2}$.

(4) INTERSECTION ANALYSIS (CONT.)

$$G = \frac{i}{\max(b, r)^2}$$

The interaction between black lines and red lines?

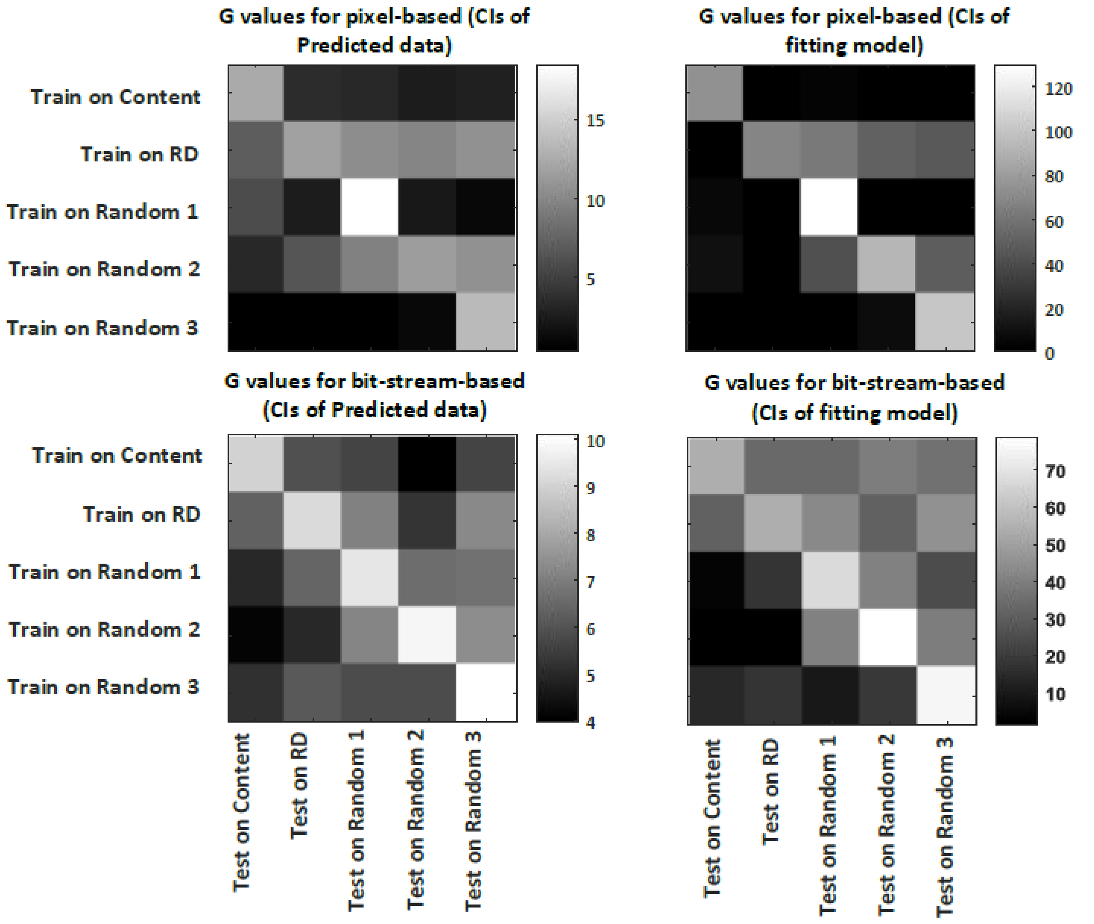
The higher the overlap the better.

4		$b \approx r$	This is the typical case of slight deviation between training and validation. The goodness depends mainly on the intersection area. In this case, $G = \frac{i}{\max(r,b)^2}$.
5		$b >> r$	These cases indicate a larger misalignment either of . the training CI or the validation CI with respect to the model fit and thus are a combination of case 4 with the cases 2 and 3 respectively. In these cases, the smaller intersection penalizes the goodness compared to the case 4 as the value of i is smaller in $G = \frac{i}{\max(r,b)^2}$
6		$b << r$	
7		$i = 0$	This is the worst case, the validation data does succeed in being predicted by the model, thus $G = 0$. Please note that this may also be an indication of a missing alignment between the training and validation data. An additional alignment step may be required in particular for models that were trained on different conditions (e.g. different video encoder).

(4) INTERSECTION ANALYSIS (CONT.)

Pixel-based NR VQA

Bitstream-based NR VQA



PERFORMANCE MEASURE COMPARISON

Performance measure	Pixel-based NR VQA (Proposed)					Bit-stream-based NR VQA				
	Content	**RD**	**Rand 1**	**Rand 2**	**Rand 3**	**Content**	**RD**	**Rand 1**	**Rand 2**	**Rand 3**
PLCC Cross-dataset	3	1	4	2	5	4	1	3	5	2
PLCC Leave-one-out	2	1	5	3	4	2	4	3	5	1
PLCC Challenging HRCs	2	1	3	4	5	1	2	3	5	4
RMSE Cross-dataset	3	1	4	2	5	5	1	4	3	2
RMSE Leave-one-out	2	1	4	3	5	3	5	4	1	2
RMSE Challenging HRCs	1	2	3	4	5	1	3	5	2	4
$P^{\text{SRC}}_{\text{RPCA\_T}}(\frac{n}{m}, m)$, $P^{\text{SRC}}_{\text{RPCA\_V}}(\frac{n}{m}, m)$	3	1	4	2	5	1	2	1	1	3
$P_{\text{DCI\_V}}(\delta, n)=\frac{i}{o}$	3	1	4	2	5	2	1	3	5	4
$P^{(\text{b,r,i})}_{\text{GModel}}$	3	1	4	2	5	2	1	3	4	5
$P^{(\text{b,r,i})}_{\text{GData}}$	3	1	4	2	5	5	1	2	3	4
Average	**2.5**	**1.1**	3.9	2.6	4.9	**2.6**	**2.1**	3.10	3.4	3.10

THANKS!

Questions