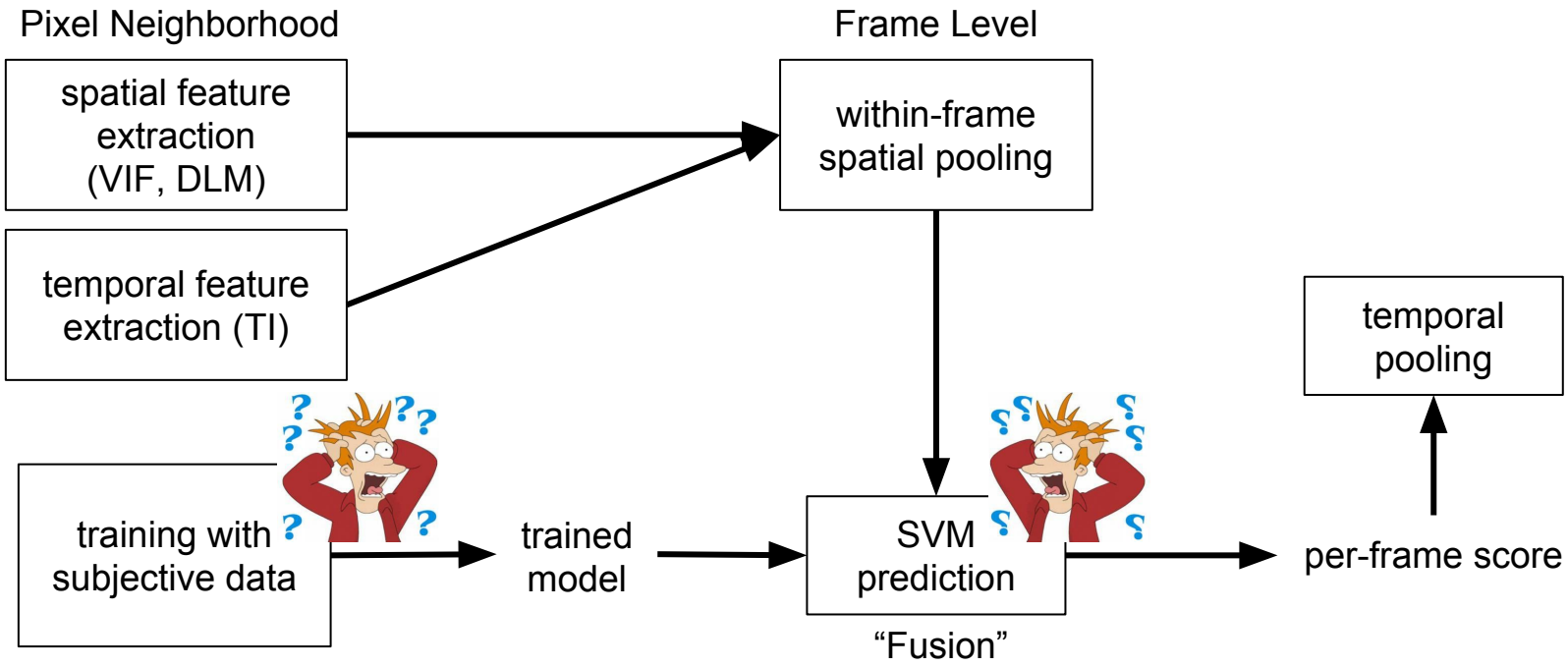
Quantify VMAF Model ~~Uncertainty~~ Variability Using Bootstrapping

Zhi Li, Ioannis Katsavounidis
Netflix

VQEG Madrid 2018

VMAF framework



Bootstrapping - “Resampling with Replacement”

```python
import numpy as np

pop_size = 100000
sample_size = 1000
trials = 100

pop_mean = 5
pop_std = 11
population = np.random.randn(pop_size) * pop_std + pop_mean
sample = population[:sample_size]

means_pop = [np.mean(np.random.choice(population, size=sample_size, replace=True)) for _ in range(trials)]
means_bootstrap = [np.mean(np.random.choice(sample, size=sample_size, replace=True)) for _ in range(trials)]

stds_pop = [np.std(np.random.choice(population, size=sample_size, replace=True)) for _ in range(trials)]
stds_bootstrap = [np.std(np.random.choice(sample, size=sample_size, replace=True)) for _ in range(trials)]

print('std of sample mean: {} (ground truth)'.format(np.std(means_pop)))
print('std of sample mean: {} (bootstrapped)\n'.format(np.std(means_bootstrap)))

print('std of sample std: {} (ground truth)'.format(np.std(stds_pop)))
print('std of sample std: {} (bootstrapped)\n'.format(np.std(stds_bootstrap)))

print('Done.')
```

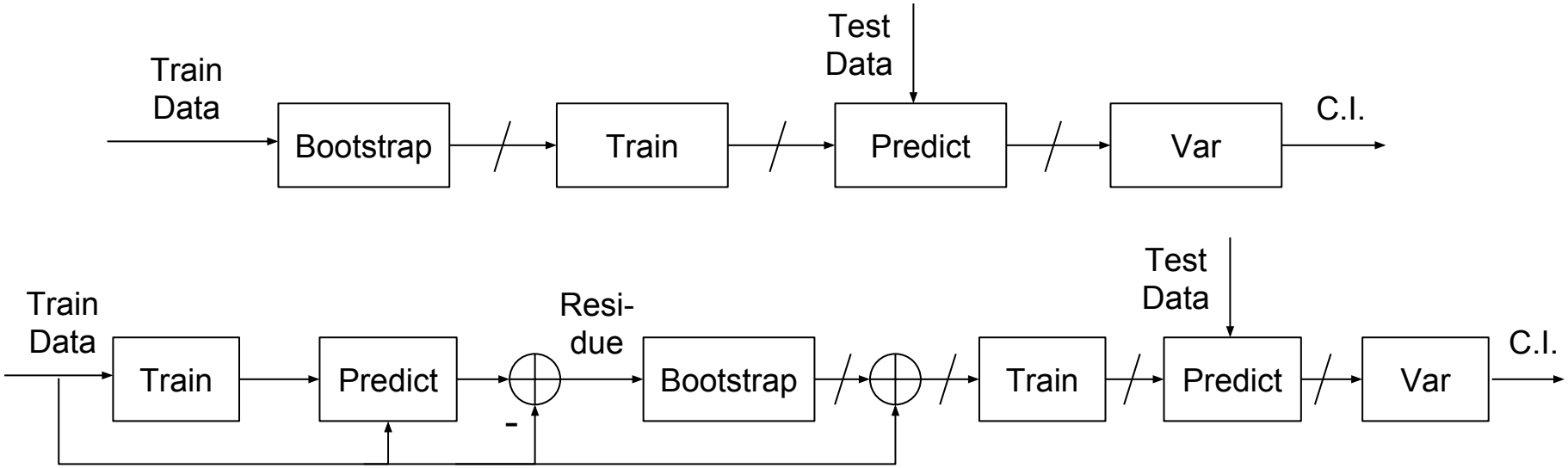
```
std of sample mean: 0.310599353041 (ground truth)
std of sample mean: 0.3649194485 (bootstrapped)

std of sample std: 0.231723205634 (ground truth)
std of sample std: 0.238048033854 (bootstrapped)

Done.
```

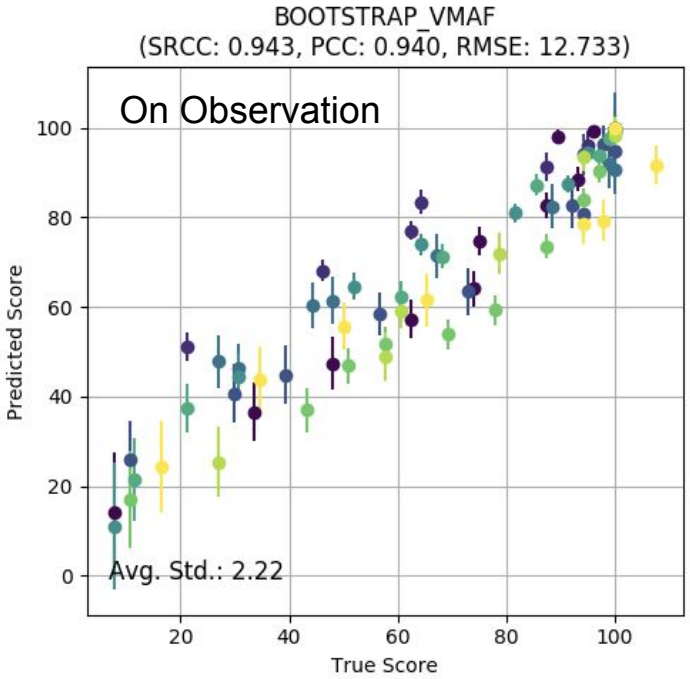
B. Efron, “Bootstrap Methods: Another Look at the Jackknife”, The Annals of Statistics, 1979, Vol. 7, No. 1, 1 - 26

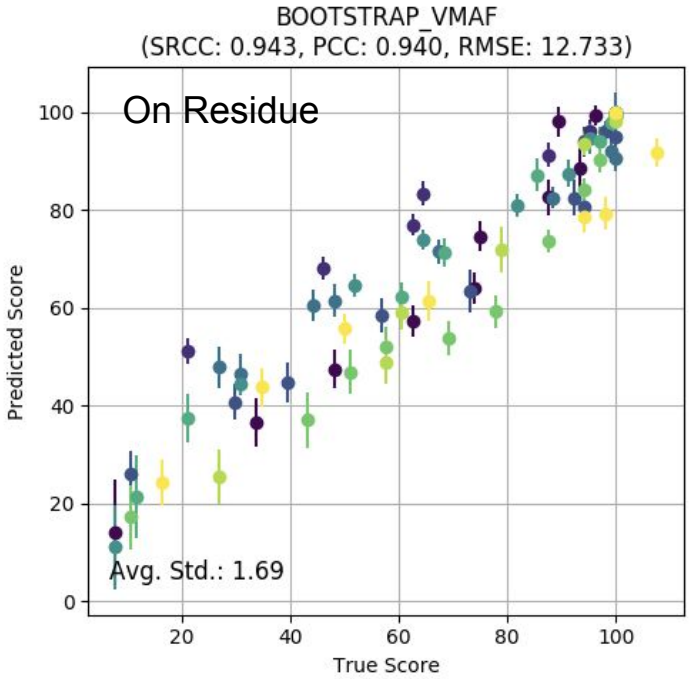
Bootstrapping in Regression Models: Observation (Top) vs. Residue (Bottom)



P. Hall, “On Bootstrap confidence intervals in nonparametric regression”, The Annals of Statistics, 1992, Vol. 20, No. 2, 695 - 711

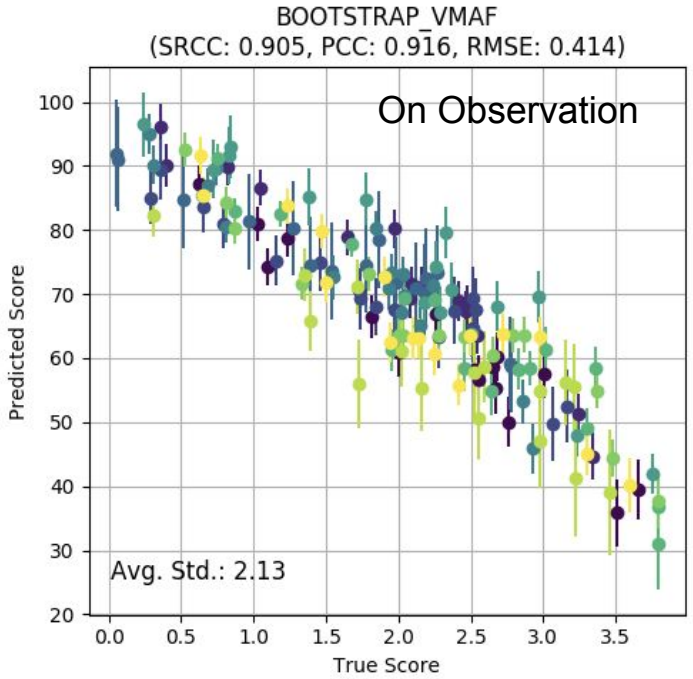
Result: NFLX Public Dataset

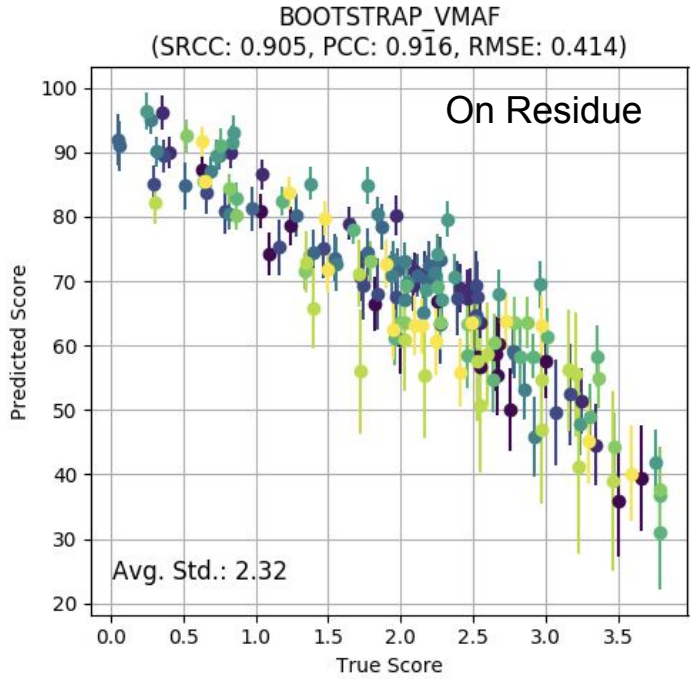




*95% C.I., Bootstrapping based on 20 models

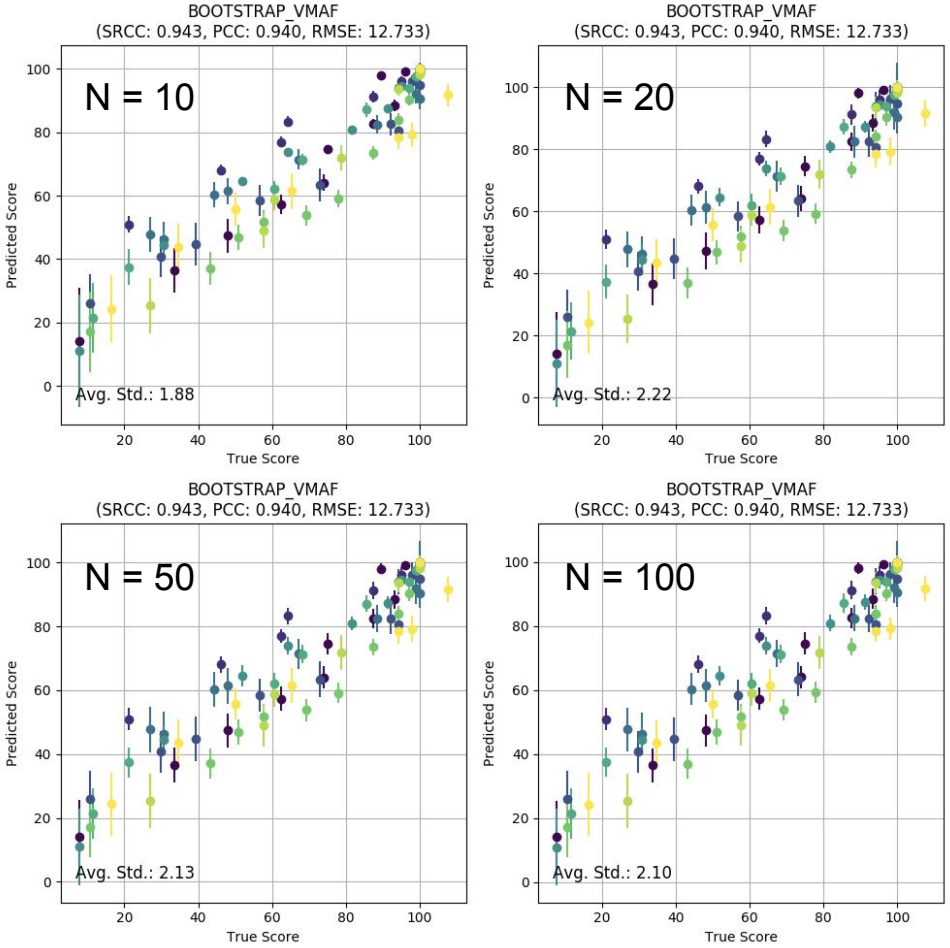
Result: LIVE Mobile Dataset





*95% C.I., Bootstrapping based on 20 models

Result: Impact of Number of Models

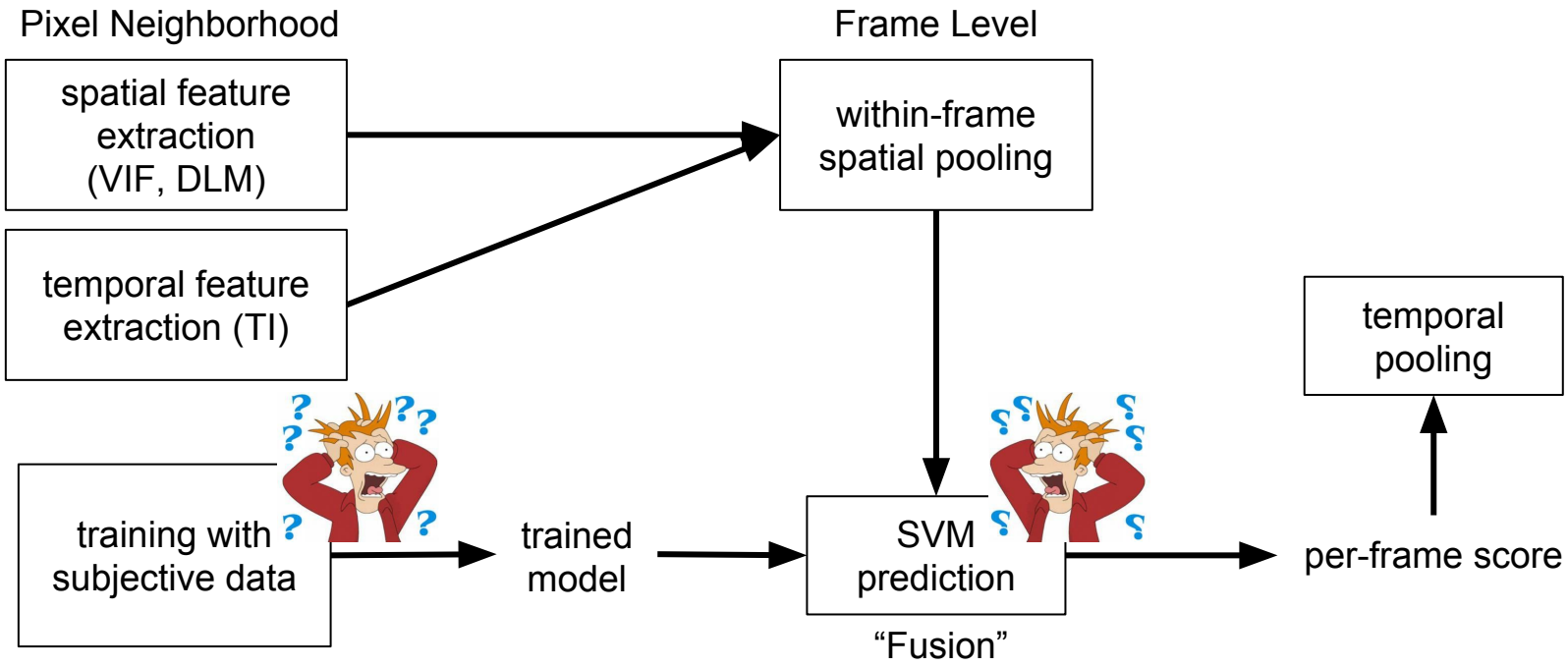


*95% C.I., Bootstrapping on observations

Observations

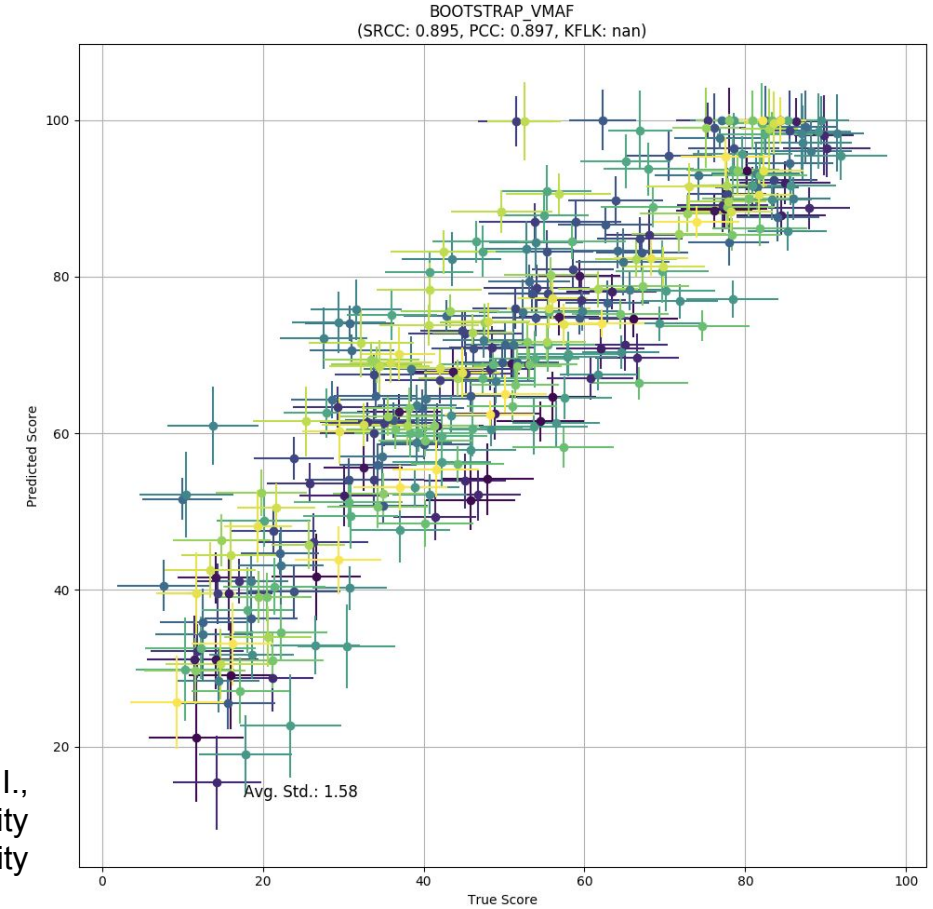
- VMAF v0.6.1 has tighter confidence interval on the high-score range compared to the low-score range
- Bootstrapping on the residues can yield tighter confidence interval than bootstrapping on the observations themselves most of the time (occasionally it can be the other way around)
- The estimated confidence interval is quite robust against the number of models used in bootstrapping

VMAF framework



Open questions

- How to combine the two variabilities?
- How to make the trade-off?



95% C.I.,
Horizontal: Subjective Score Variability
Vertical: Regression Model Variability

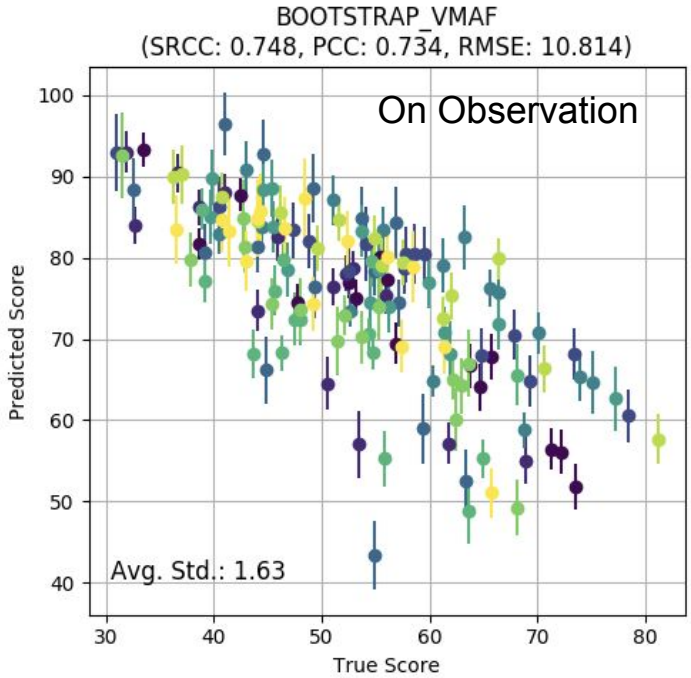
Coming up next to VMAF open-source repo

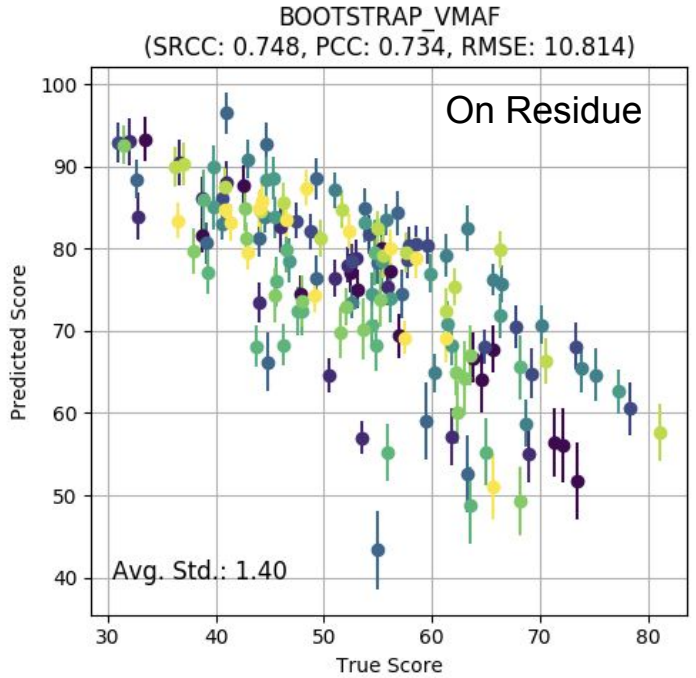
- Release of new 4K model (Monday's talk)
- Confidence interval of VMAF model (this talk)
- Enhanced temporal features
- New HDR model

Backup Slides

NETFLIX

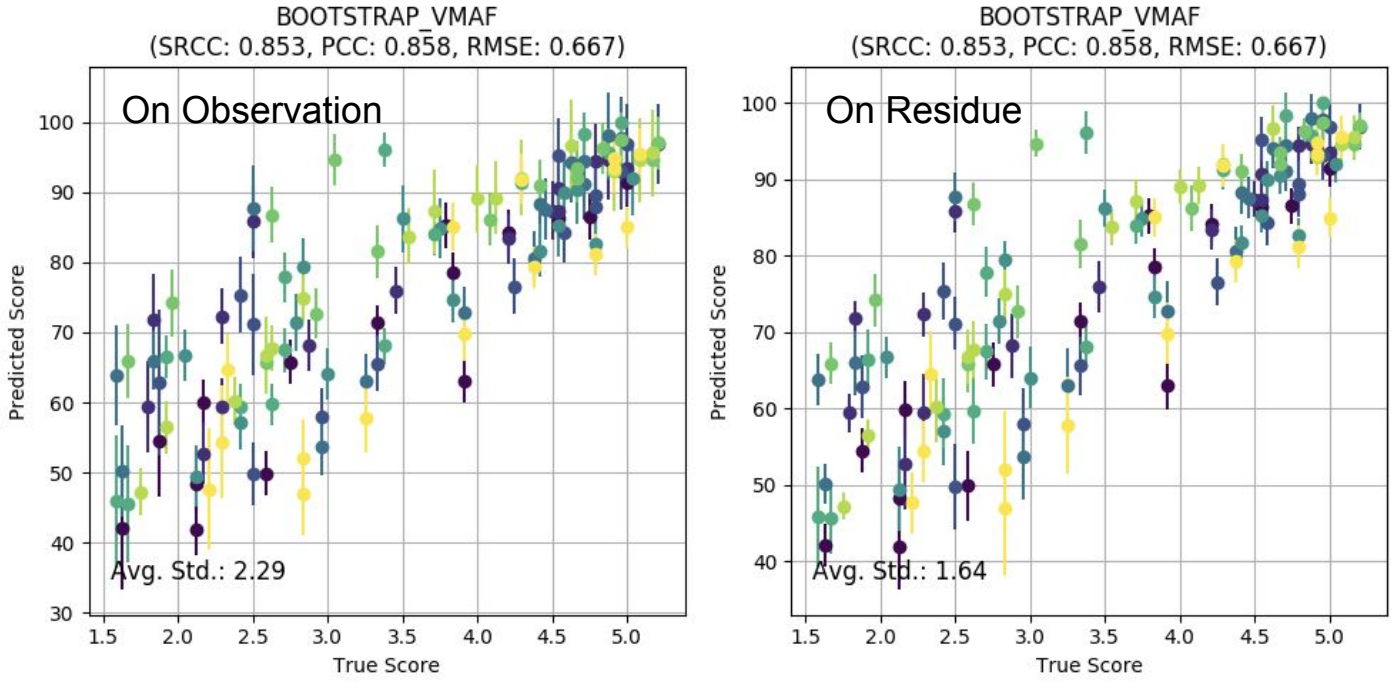
Result: LIVE Video Dataset





*95% C.I., Bootstrapping based on 20 models

Result: VQEGHD3 Dataset



*95% C.I., Bootstrapping based on 20 models