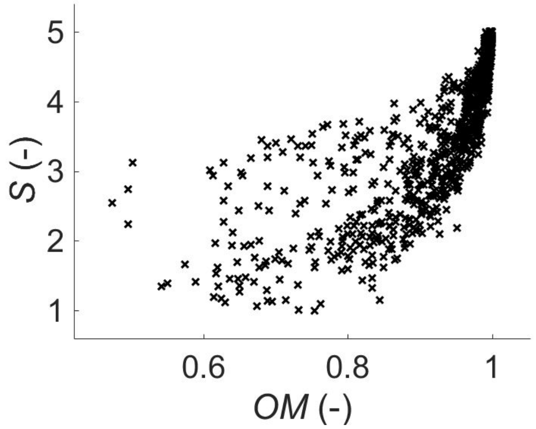
Methodology for Objective Metrics Performance Evaluation...

… and its use for large scale training

Lukáš Krasula
lukas.krasula@univ-nantes.fr

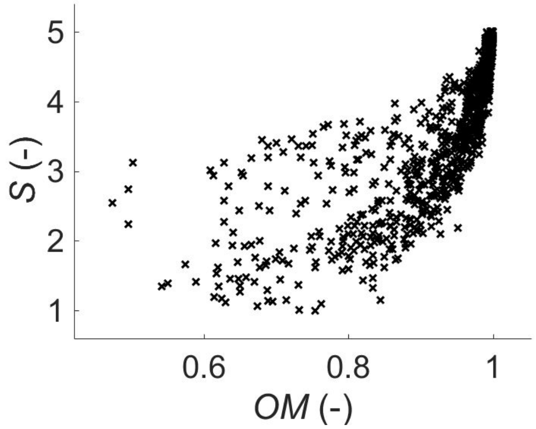
Objective Metrics Performance Evaluation

- Comparing subjective vs. automatically predicted scores (*S* vs. *OM*)



Objective Metrics Performance Evaluation

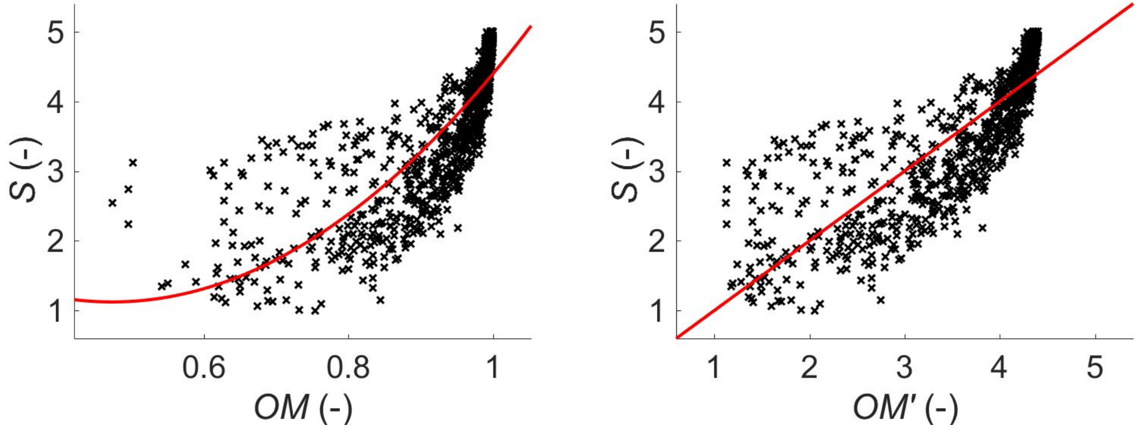
- Comparing subjective vs. automatically predicted scores (*S* vs. *OM*)
- Typical measures [ITU-T Rec. P.1401]
  - Pearson Correlation Coefficient
  - Root Mean Squared Error
  - Outlier Ratio



Objective Metrics Performance Evaluation

- Comparing subjective vs. automatically predicted scores (*S* vs. *OM*)
- Typical measures [ITU-T Rec. P.1401]
  - Pearson Correlation Coefficient
  - Root Mean Squared Error
  - Outlier Ratio

Necessity of mapping to the common scale



Danger of Mapping

- Mapping is not standardized (only required to be monotonic)
- Problems:

Danger of Mapping

- Mapping is not standardized (only required to be monotonic)
- Problems:
  - Different papers provide **different results** obtained **for the same datasets**
    - Reproducibility is questionable

Danger of Mapping

- Mapping is not standardized (only required to be monotonic)
- Problems:
  - Different papers provide **different results** obtained **for the same datasets**
    - Reproducibility is questionable
  - Mapping can bias the results

Correlation for CSIQ database after 3rd order polynomial mapping	SSIM	MS-SSIM
Fitting function coefficients optimized with PLCC (VQEG)	**0.8575**	0.8562
Fitting function coefficients optimized with RMSE (ITU-T Rec. J.149)	0.8581	**0.8859**

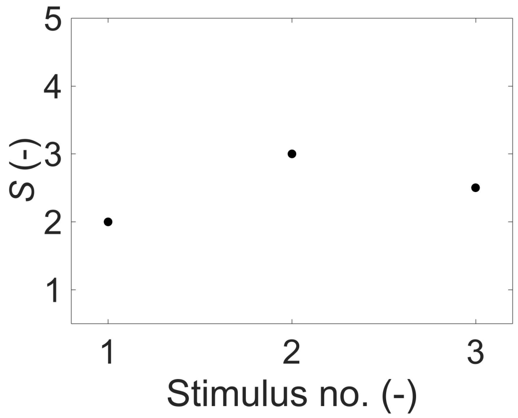
Rank Order Correlation

- Using Rank Order Correlation Coefficients (Spearman’s and/or Kendall’s)
    - Typical solution to the mapping problem - independency towards the monotonic mapping

Rank Order Correlation

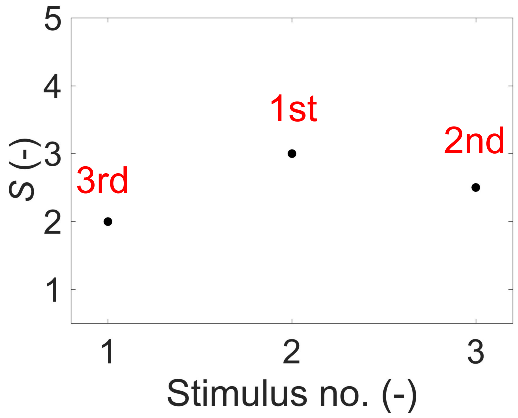
- Using Rank Order Correlation Coefficients (Spearman’s and/or Kendall’s)
  - Typical solution to the mapping problem - independency towards the monotonic mapping

- However...
  - Considering subjective data to be **deterministic**



Rank Order Correlation

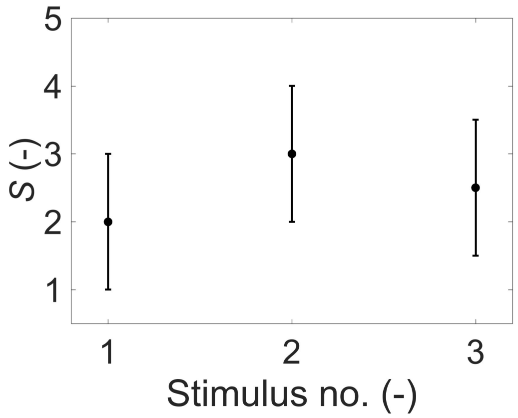
- Using Rank Order Correlation Coefficients (Spearman's and/or Kendall's)
  - Typical solution to the mapping problem - independency towards the monotonic mapping
- However...
  - Considering subjective data to be **deterministic**



Rank Order Correlation

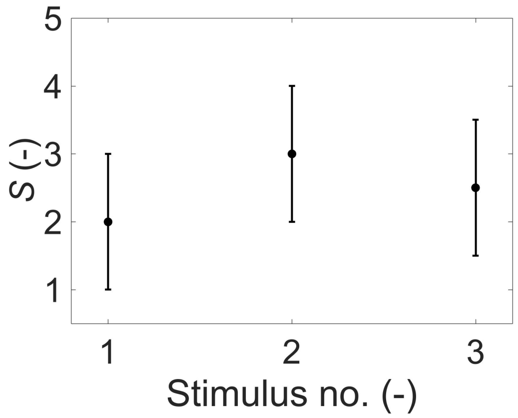
- Using Rank Order Correlation Coefficients (Spearman’s and/or Kendall’s)
  - Typical solution to the mapping problem - independency towards the monotonic mapping

- However...
  - Considering subjective data to be **deterministic**



Rank Order Correlation

- Using Rank Order Correlation Coefficients (Spearman's and/or Kendall's)
  - Typical solution to the mapping problem - independency towards the monotonic mapping
- However...
  - Considering subjective data to be **deterministic**



What is the correct order?

Novel performance evaluation methodology

- Goals:
  - No mapping during the process
  - Considering the uncertainty of the ground truth

Novel performance evaluation methodology

- Goals:
  - No mapping during the process
  - Considering the uncertainty of the ground truth

- Basic premise:
  - Regardless the subjective procedure, we are always able to determine:

Novel performance evaluation methodology

- Goals:
  - No mapping during the process
  - Considering the uncertainty of the ground truth
- Basic premise:
  - Regardless the subjective procedure, we are always able to determine:

(a) Are any two stimuli statistically significantly different in quality?

$$[i,j] \in \mathrm{N} \quad \Leftrightarrow \quad \Pr\{\, S(i) \neq S(j) \,\} < 1-\alpha$$

$$[i,j] \in \mathrm{D} \quad \Leftrightarrow \quad \Pr\{\, S(i) \neq S(j) \,\} \geq 1-\alpha$$

Novel performance evaluation methodology

- Goals:
  - No mapping during the process
  - Considering the uncertainty of the ground truth

- Basic premise:
  - Regardless the subjective procedure, we are always able to determine:

(a) Are any two stimuli statistically significantly different in quality?

$$[i,j] \in N \quad \Leftrightarrow \quad \Pr\{\, S(i) \neq S(j) \,\} < 1-\alpha$$
$$[i,j] \in D \quad \Leftrightarrow \quad \Pr\{\, S(i) \neq S(j) \,\} \geq 1-\alpha$$

(b) If they are, which of them is qualitatively better?

$$[i,j] \in B \quad \Leftrightarrow \quad \Delta S(i,j) = S(i) - S(j) \geq 0, \ \forall \ [i,j] \in D$$
$$[i,j] \in W \quad \Leftrightarrow \quad \Delta S(i,j) = S(i) - S(j) \leq 0, \ \forall \ [i,j] \in D$$

Novel performance evaluation methodology:

Proposed Assumptions

- Reliable metric then

  I. Provides **close** scores for **similar** pairs and **distant** scores for **different**

$$|\Delta OM(i,j)| = |OM(i) - OM(j)| \rightarrow 0, \ \forall \ [i,j] \in N$$

$$|\Delta OM(i,j)| = |OM(i) - OM(j)| \gg 0, \ \forall \ [i,j] \in D$$

Novel performance evaluation methodology:

Proposed Assumptions

- Reliable metric then

I. Provides **close** scores for **similar** pairs and **distant** scores for **different**

$$|\Delta OM(i,j)| = |OM(i) - OM(j)| \rightarrow 0, \ \forall \ [i,j] \in N$$

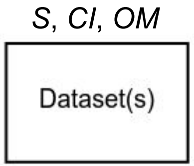
$$|\Delta OM(i,j)| = |OM(i) - OM(j)| \gg 0, \ \forall \ [i,j] \in D$$

II. Provides **higher** score for qualitatively **better** stimulus

$$\mathrm{sign}\{ \Delta OM(i,j) \} = \mathrm{sign}\{ \Delta S(i,j) \}, \ \forall \ [i,j] \in D$$

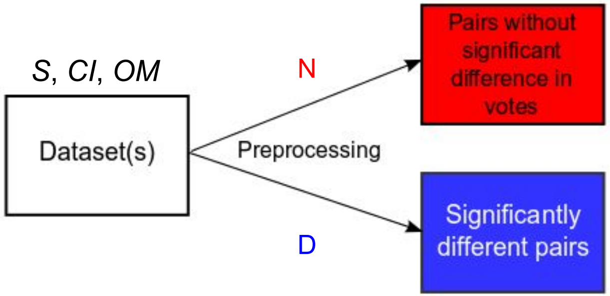
Novel performance evaluation methodology:

Description



Novel performance evaluation methodology:

Description

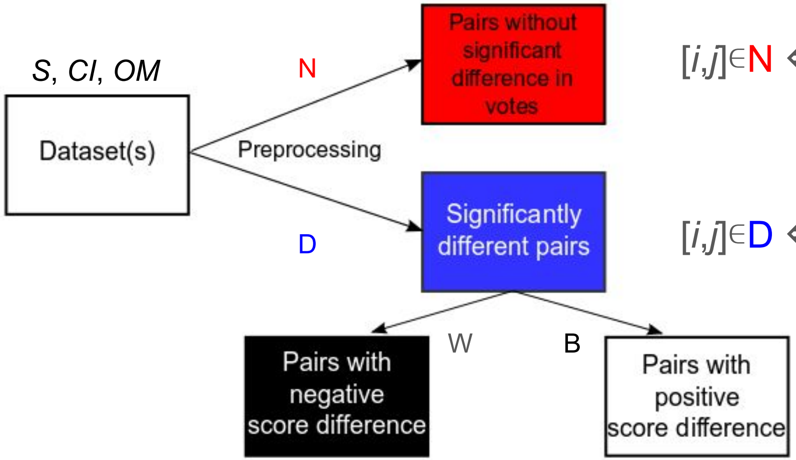


$[i,j] \in \mathrm{N} \Leftrightarrow \Pr\{\, S(i) \neq S(j) \,\} < 1-\alpha$

$[i,j] \in \mathrm{D} \Leftrightarrow \Pr\{\, S(i) \neq S(j) \,\} \geq 1-\alpha$

Novel performance evaluation methodology:

Description



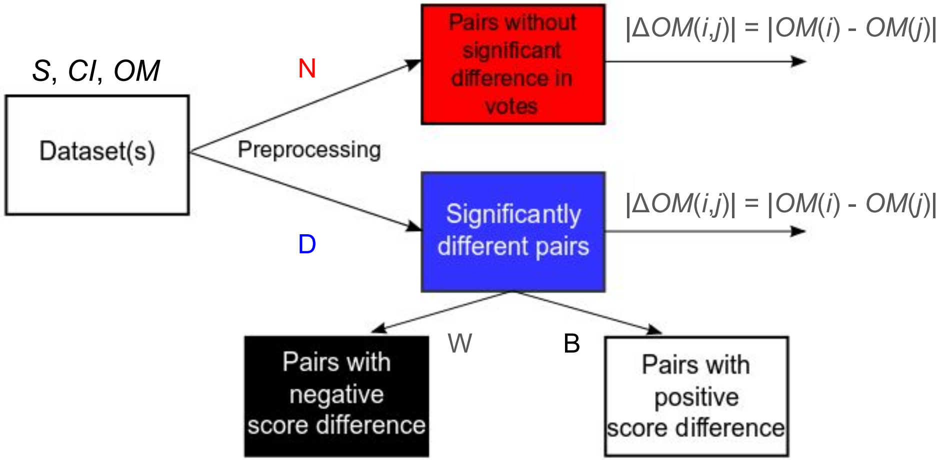
$[i,j] \in N \Leftrightarrow \Pr\{ S(i) \neq S(j) \} < 1-\alpha$

$[i,j] \in D \Leftrightarrow \Pr\{ S(i) \neq S(j) \} \geq 1-\alpha$

$[i,j] \in W \Leftrightarrow \Delta S(i,j) = S(i) - S(j) \leq 0$

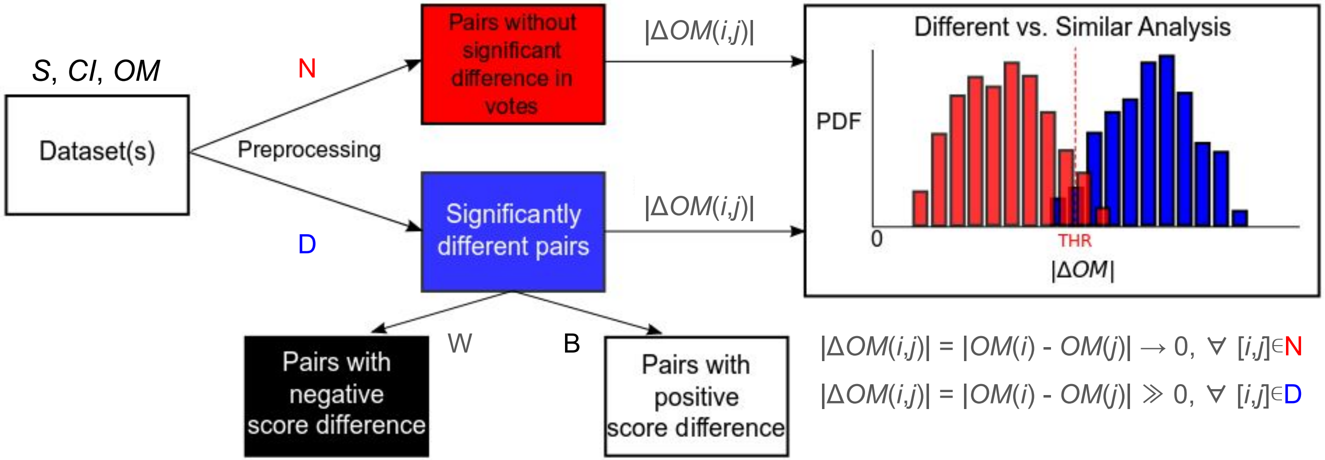
$[i,j] \in B \Leftrightarrow \Delta S(i,j) = S(i) - S(j) \geq 0$

Novel performance evaluation methodology:
Description



Novel performance evaluation methodology:

Description

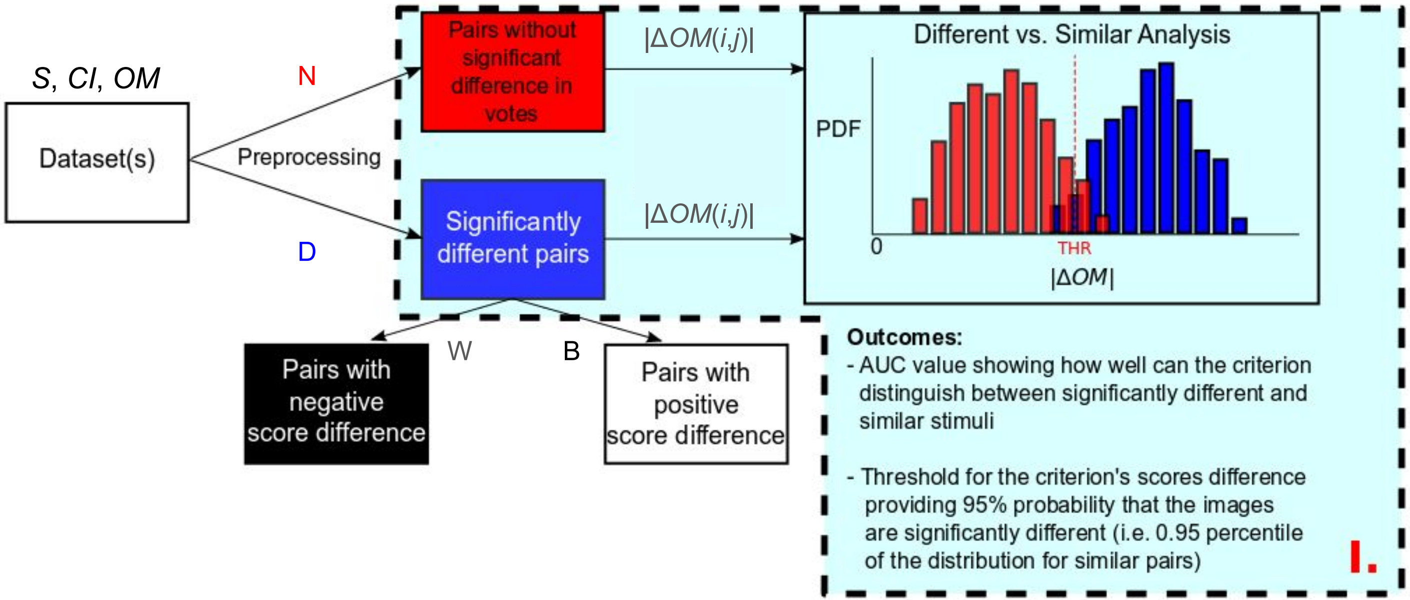


$$|\Delta OM(i,j)| = |OM(i) - OM(j)| \rightarrow 0, \ \forall\ [i,j] \in N$$

$$|\Delta OM(i,j)| = |OM(i) - OM(j)| \gg 0, \ \forall\ [i,j] \in D$$

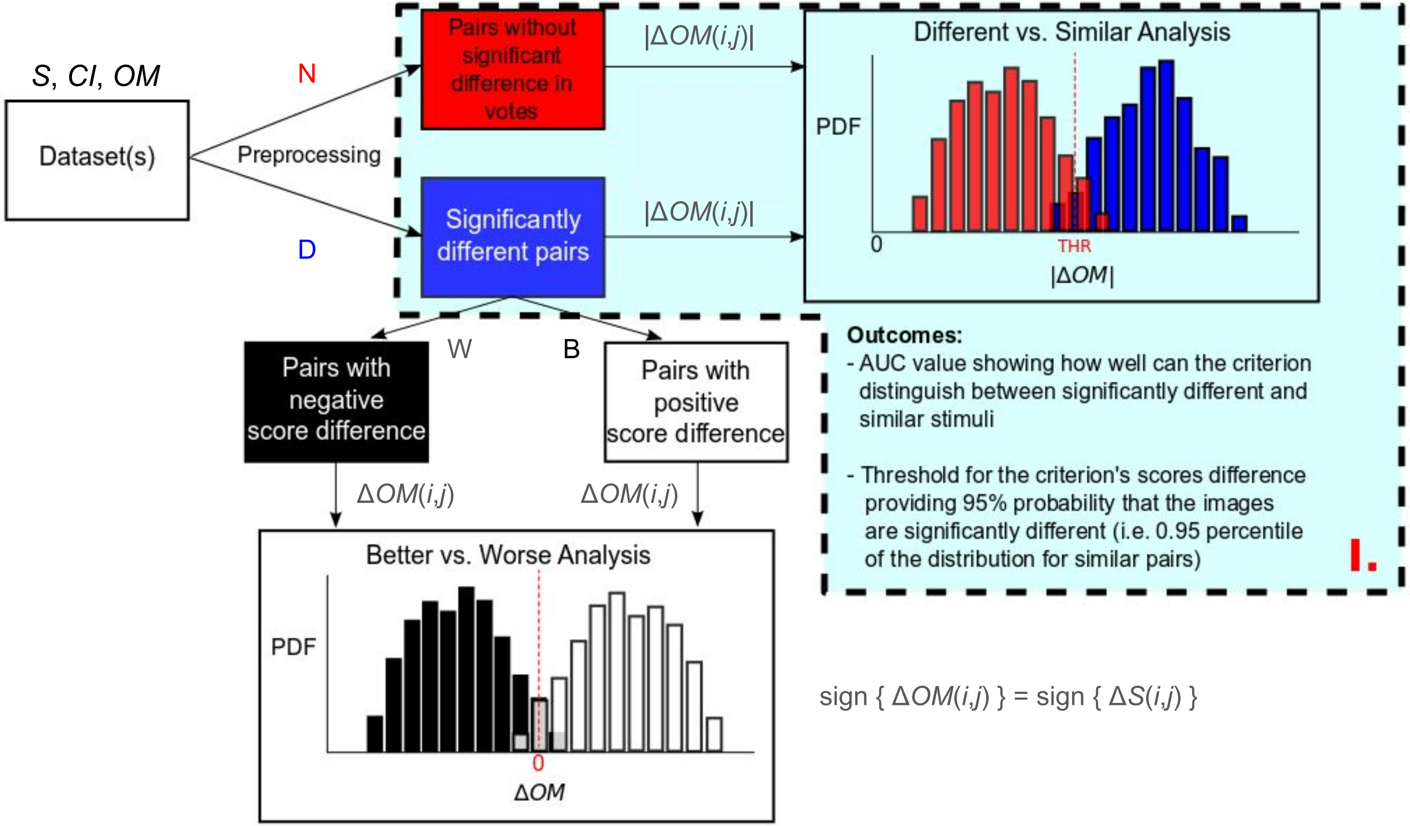
Novel performance evaluation methodology:

Description



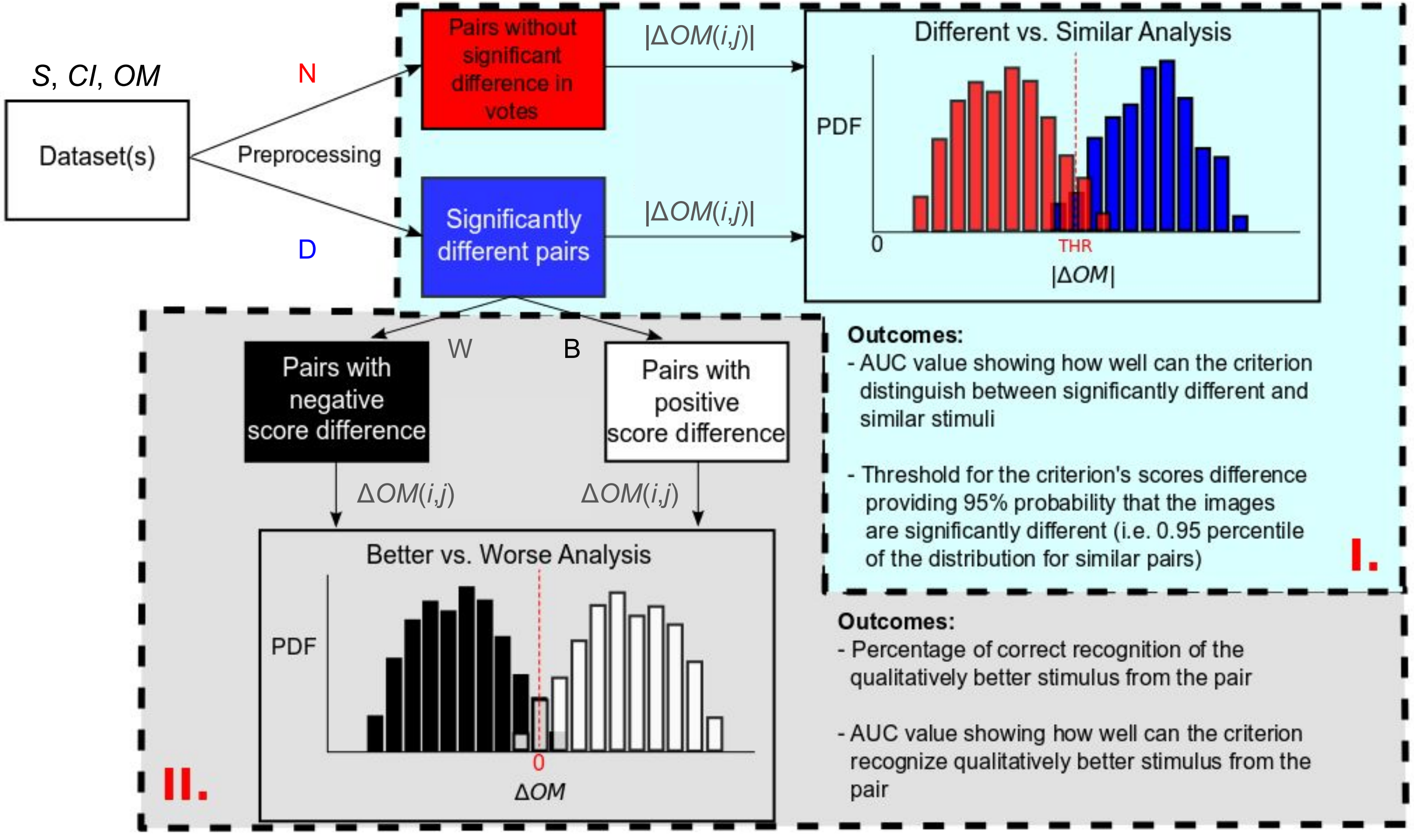
Novel performance evaluation methodology:

Description



Novel performance evaluation methodology:

Description



Novel performance evaluation methodology:

Advantages

- Goals have been fulfilled
  - There is no mapping involved
  - The uncertainty of the subjective scores is considered

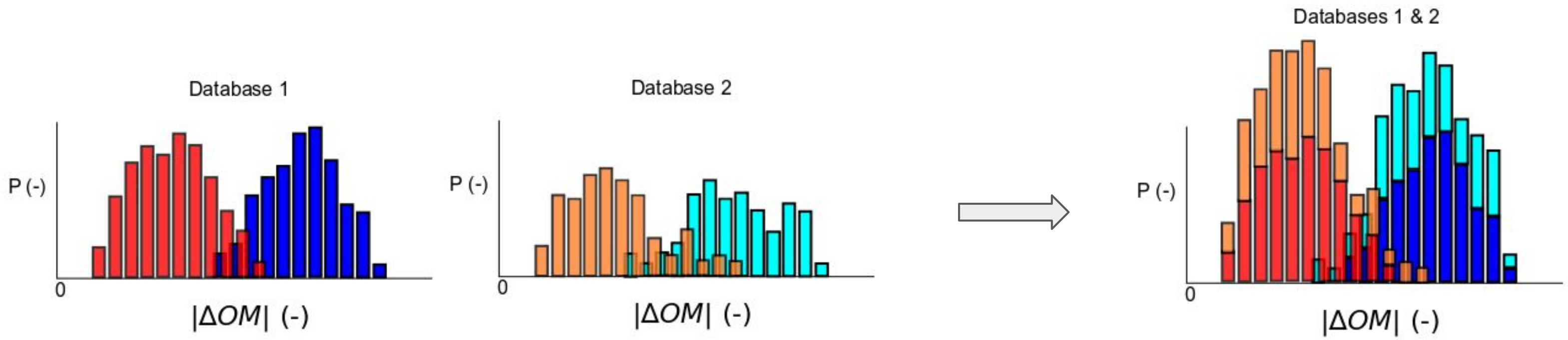
Novel performance evaluation methodology:

Advantages

- Goals have been fulfilled
  - There is no mapping involved
  - The uncertainty of the subjective scores is considered
- Moreover…
  - Universality towards the subjective procedure, scale, and format of the ground-truth data
  - Allows for simple numerical comparisons and testing of statistical significance
  - High statistical power (due to the pair-wise approach)
  - Enables simple and meaningful combination of the data coming from multiple datasets

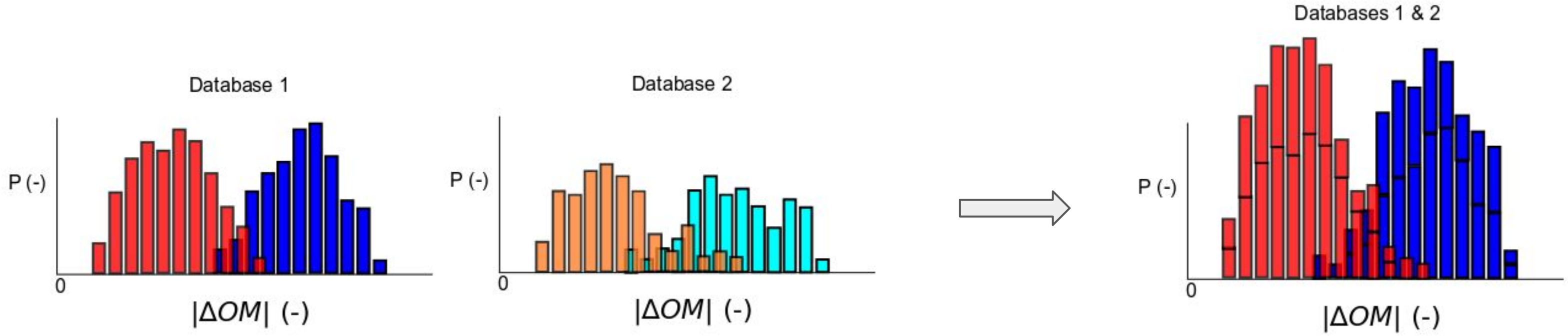
Novel performance evaluation methodology:
Advantages

- Goals have been fulfilled
  - There is no mapping involved
  - The uncertainty of the subjective scores is considered
- Moreover…
  - Universality towards the subjective procedure, scale, and format of the ground-truth data
  - Allows for simple numerical comparisons and testing of statistical significance
  - High statistical power (due to the pair-wise approach)
  - **Enables simple and meaningful combination of the data coming from multiple datasets**



Novel performance evaluation methodology:
Advantages

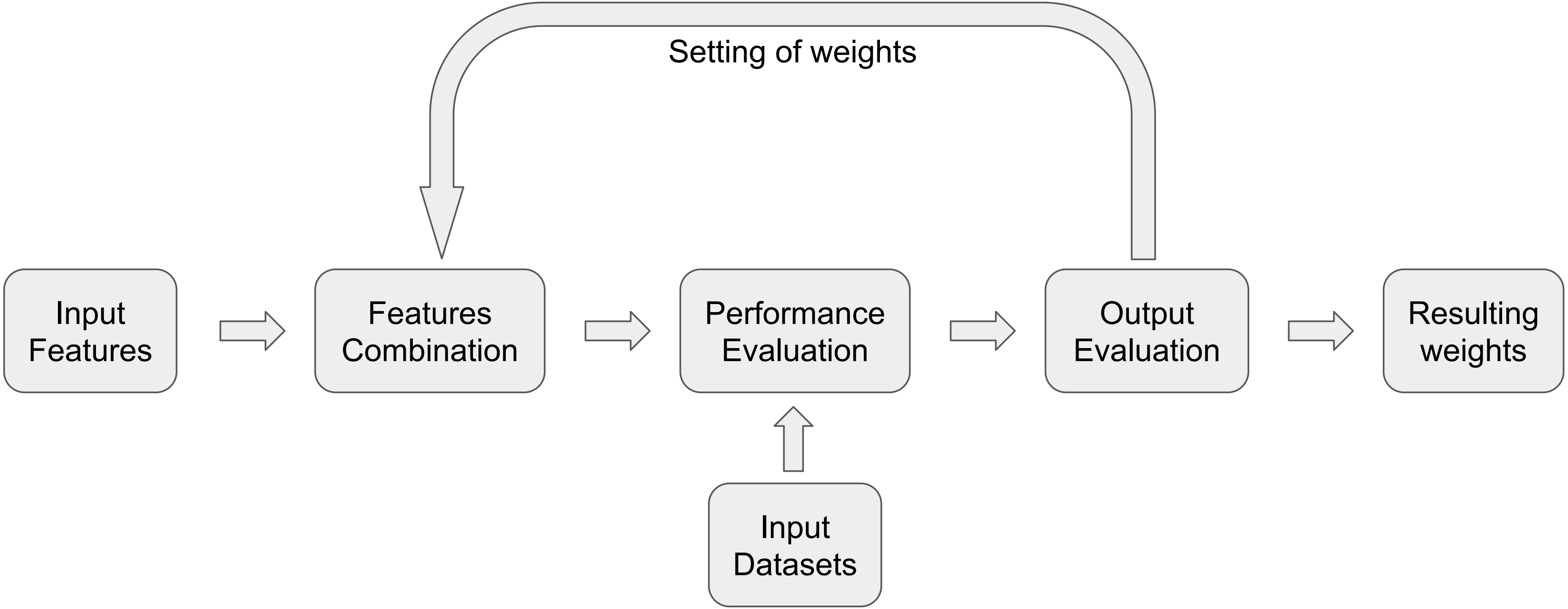
- Goals have been fulfilled
  - There is no mapping involved
  - The uncertainty of the subjective scores is considered
- Moreover…
  - Universality towards the subjective procedure, scale, and format of the ground-truth data
  - Allows for simple numerical comparisons and testing of statistical significance
  - High statistical power (due to the pair-wise approach)
  - **Enables simple and meaningful combination of the data coming from multiple datasets**



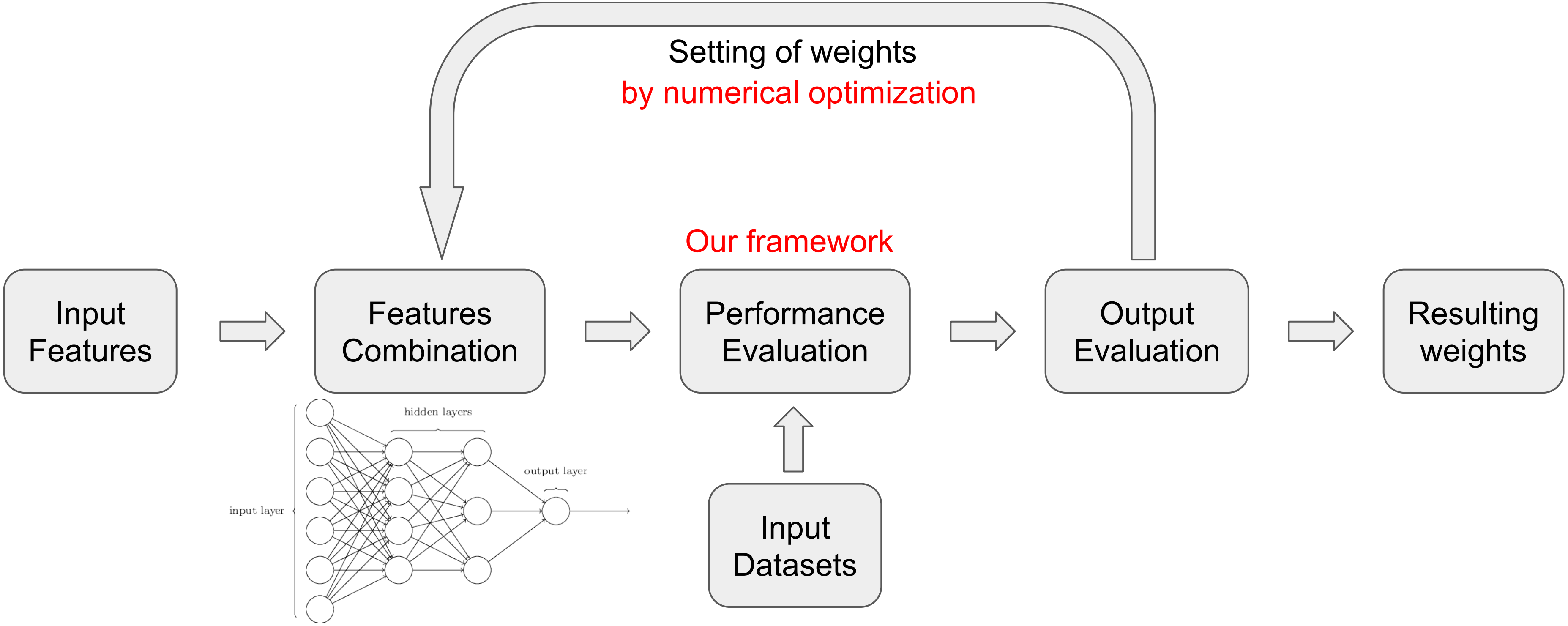
Novel performance evaluation methodology:
Advantages

- Goals have been fulfilled
  - There is no mapping involved
  - The uncertainty of the subjective scores is considered
- Moreover…
  - Universality towards the subjective procedure, scale, and format of the ground-truth data
  - Allows for simple numerical comparisons and testing of statistical significance
  - High statistical power (due to the pair-wise approach)
  - **Enables simple and meaningful combination of the data coming from multiple datasets**
    - No inter-experiment mapping necessary
    - Overall performance can be easily determined
    - Increase of number of training/testing points in orders of magnitude - deep learning etc.

Using the framework for objective metrics training



Using the framework for objective metrics training

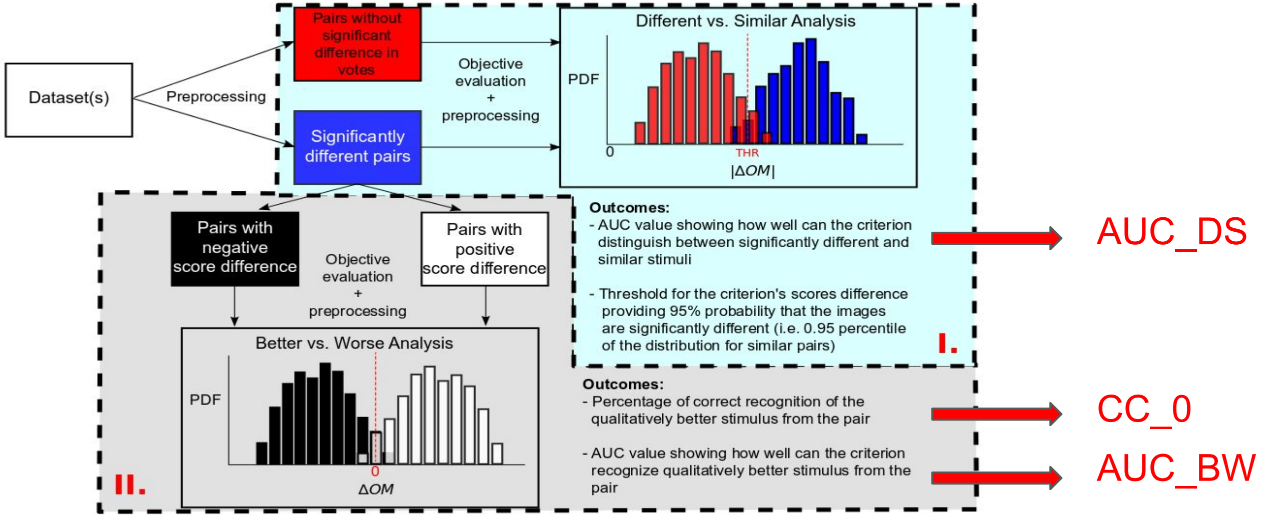


Preliminary results

- Publicly available VMAF (Video Multi-Method Assessment Fusion) package
  - VMAF features (VIF on 4 scales, Detail Loss, Motion)
- 18 datasets (9 used for training, 9 for testing)
- 1 hidden layer, 6 neurons, RELU activation function

Preliminary results

- Publicly available VMAF (Video Multi-Method Assessment Fusion) package
  - VMAF features (VIF on 4 scales, Detail Loss, Motion)
- 18 datasets (9 used for training, 9 for testing)
- 1 hidden layer, 6 neurons, RELU activation function



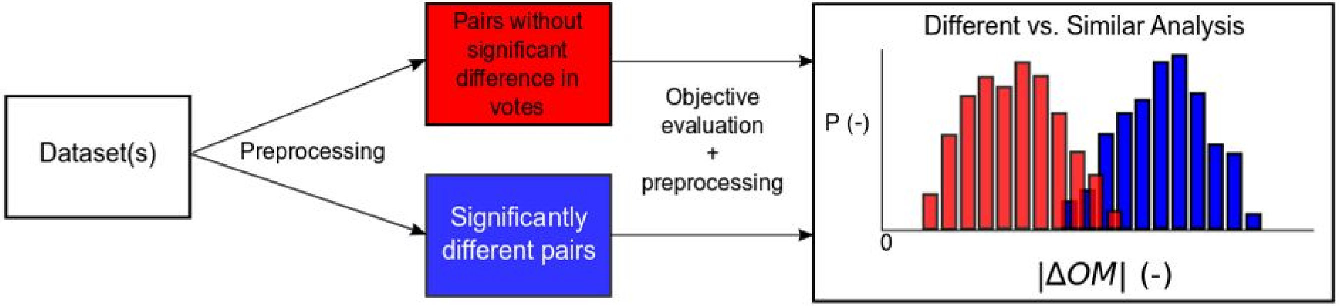
Preliminary results

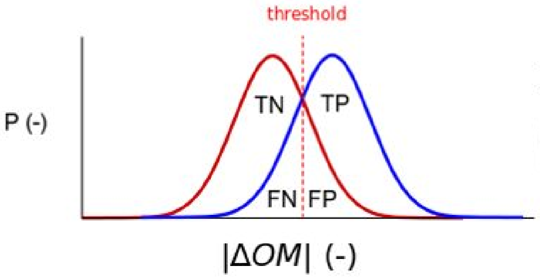
- Publicly available VMAF (Video Multi-Method Assessment Fusion) package
  - VMAF features (VIF on 4 scales, Detail Loss, Motion)
- 18 datasets (9 used for training, 9 for testing)
- 1 hidden layer, 6 neurons, RELU activation function

	Custom Neural Network:	VMAF (trained on one of the datasets):
Test set	AUC_DS = **0.7869**	AUC_DS = 0.7586
	AUC_BW = **0.9550**	AUC_BW = 0.9490
	CC_0 = **0.8963**	CC_0 = 0.8951
Test + Train sets	AUC_DS = **0.7646**	AUC_DS = 0.7230
	AUC_BW = **0.9551**	AUC_BW = 0.9469
	CC_0 = **0.8957**	CC_0 = 0.8954

Thank you for your attention!

ROC Analysis



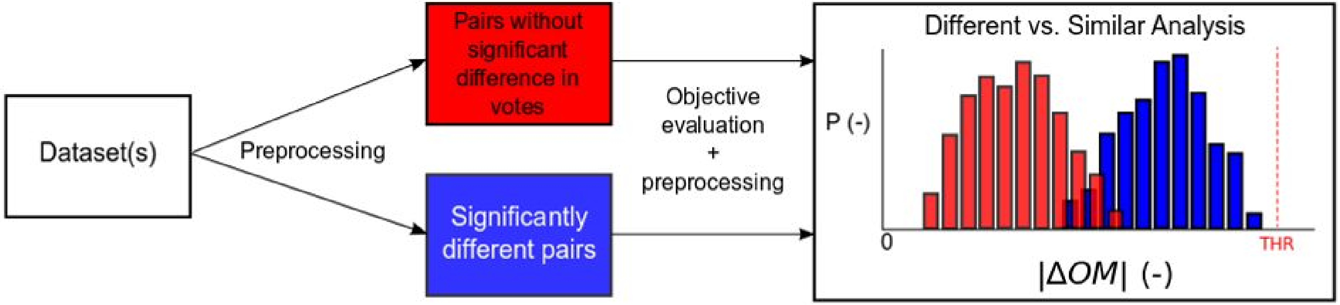


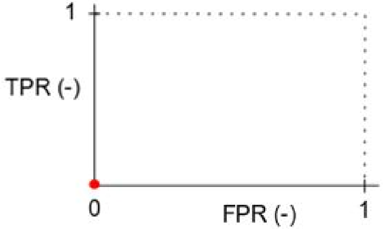
TN - true negative
TP - true positive
FN - false negative
FP - false positive

True positive rate
TPR = TP / (TP+FN)

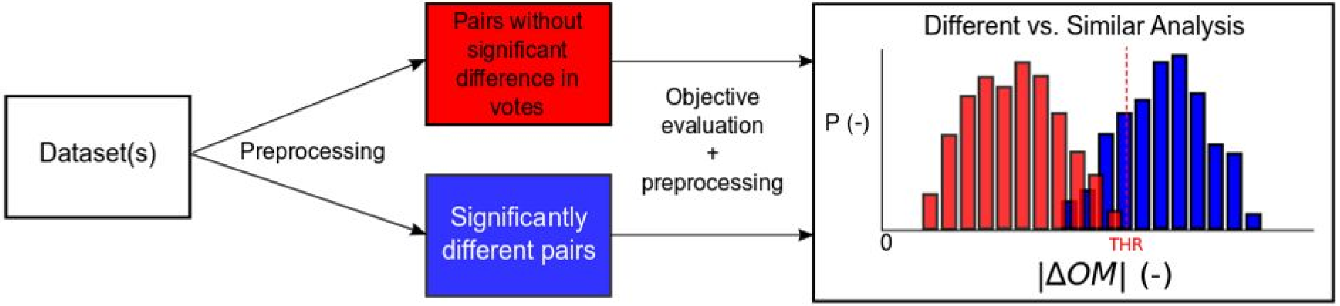
False positive rate
FPR = FP / (FP+TN)

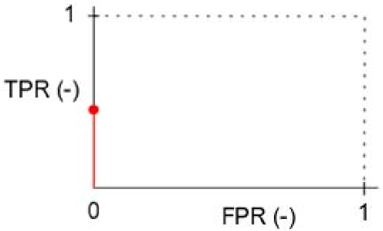
ROC Analysis



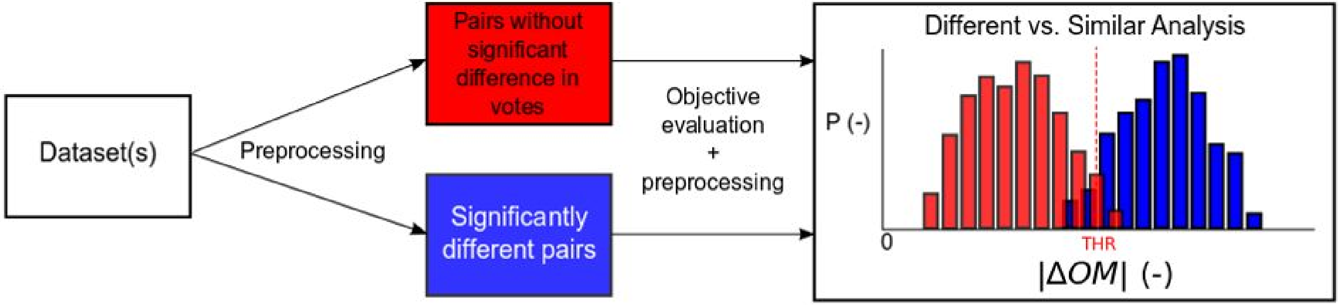


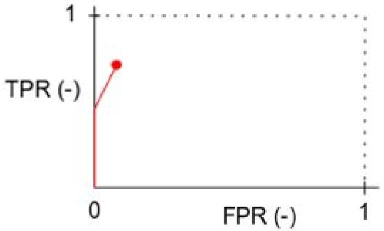
ROC Analysis



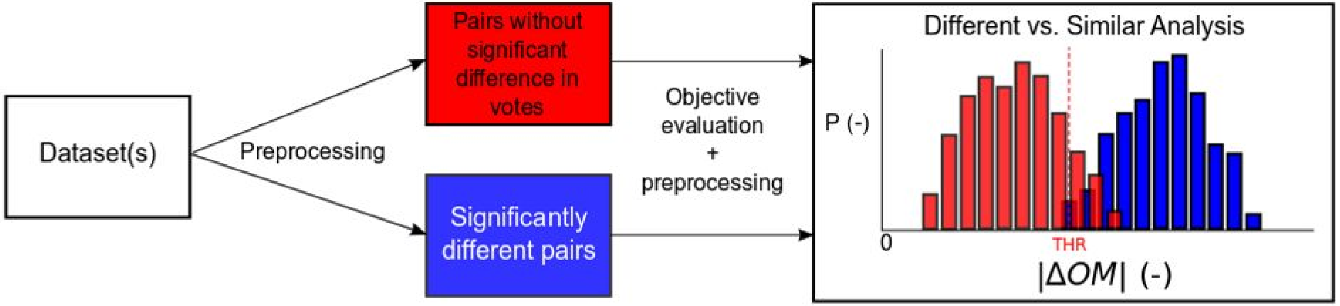


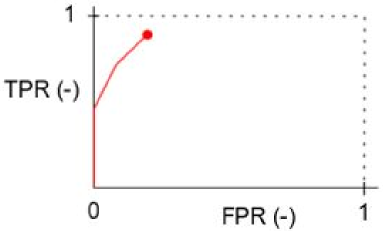
ROC Analysis



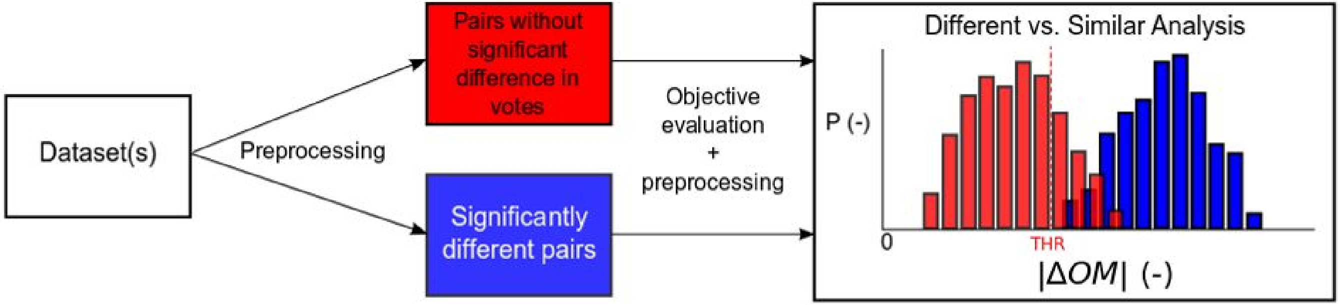


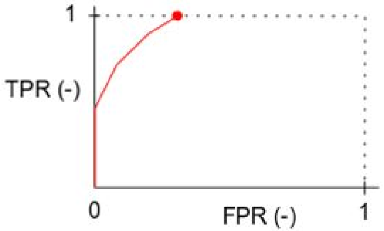
ROC Analysis



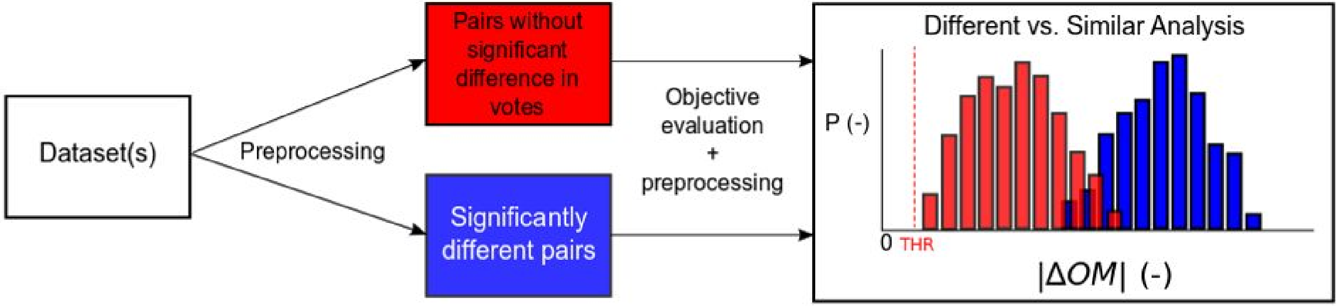


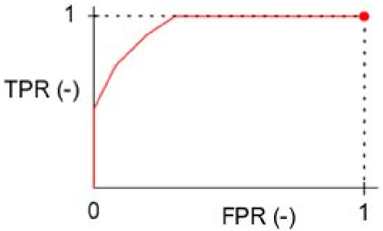
ROC Analysis



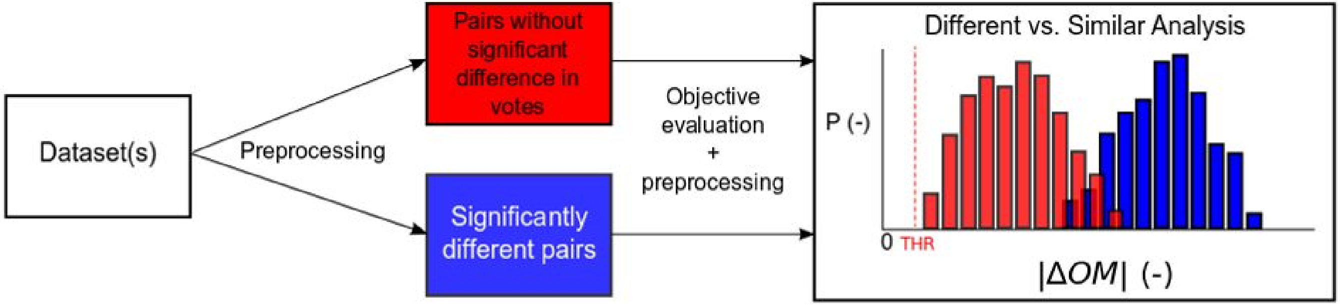


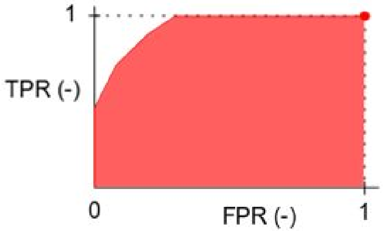
ROC Analysis



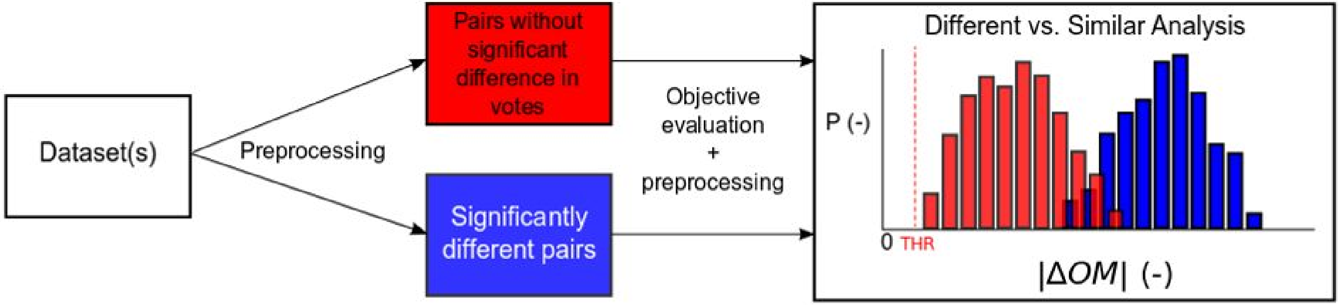


ROC Analysis

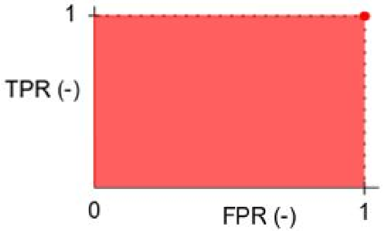




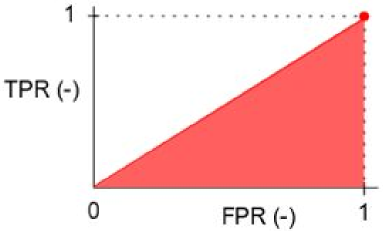
ROC Analysis



AUC = 1



AUC = 0.5



AUC = 0.85