

MIREX 2008 Audio Classification Tasks

Kris West (kris@onellama.com), J. Stephen Downie (jdownie@uiuc.edu), Michael Mandel (mim@ee.columbia.edu)

Abstract

The Music Information Retrieval Evaluation eXchange (MIREX), run by the International Music Information Retrieval System Evaluation Laboratory (IMIRSEL), has hosted audio classification tasks since its inception in 2005. This year the classification tasks include:

Audio Artist Identification

http://www.music-ir.org/mirex/2008/index.php/Audio_Artist_Identification

- Western pop music collection (MIREX 2007)
- Classical composer collection (MIREX 2007)
- Data collected by IMIRSEL and LabROSA

Audio Genre Classification

http://www.music-ir.org/mirex/2008/index.php/Audio_Genre_Classification

- Western pop music collection (MIREX 2007)
- Data collected by IMIRSEL and LabROSA
- Latin music Collection (*new* collection, MIREX 2008)
- Data collected by Pontifical Catholic University of Paraná and Federal University of Technology of Paraná (cns2@kent.ac.uk) (*The Latin Music Database*, Silla, Koerich, Kaestner, *ISMIR 2008*)

Audio Mood Collection

http://www.music-ir.org/mirex/2008/index.php/Audio_Music_Mood_Classification

- Associated Production Music (APM) (MIREX 2007)
- Ground-truth data validated by IMIRSEL and MTG, UPF

Audio Tag Classification (*new* task for MIREX 2008)

http://www.music-ir.org/mirex/2008/index.php/Audio_Tag_Classification

- MajorMiner game collection (Mandel and Ellis, *ISMIR 2007*)

MIREX classification tasks are organised and discussed on the MRX_COM00 mailing list, sign up at: <https://mail.lis.uiuc.edu/mailman/listinfo/mrx-com00>

Important Considerations

- Splits of collections for any music classification task should be **artist or album filtered** to avoid inflated performance estimates, caused by the matching of an artist's or album's characteristics rather than the intended class' characteristics.
- Valid statistical significance tests should be chosen to compare the performance of algorithms as any apparent differences may be due to characteristics of the collection or test rather than just the algorithms compared.

Evaluation Software

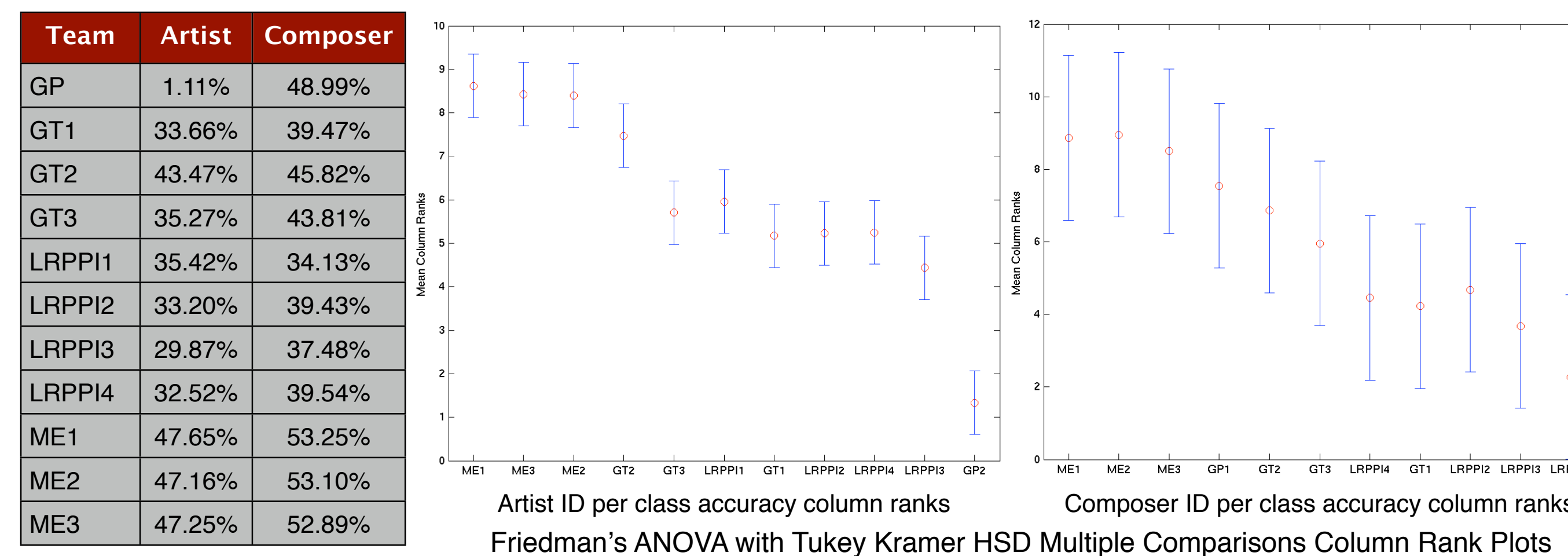
The evaluation software developed to assess and compare performances of submissions to the MIREX 2008 classification tasks will be made available to the MIR community (in a stand alone form) shortly after ISMIR, allowing researchers to duplicate our evaluation procedure and significance tests. Additionally, this software provides facilities to perform artist filtered splits of audio and metadata collections, which non-trivial for tag-based collections and is a step often missed by MIR researchers new to the field. The evaluation software will also be contributed to M2K.

The release will be announced on the music-ir and evalfest lists. Contact kris@onellama.com if you wish to receive a notification or a pre-release copy of the software.

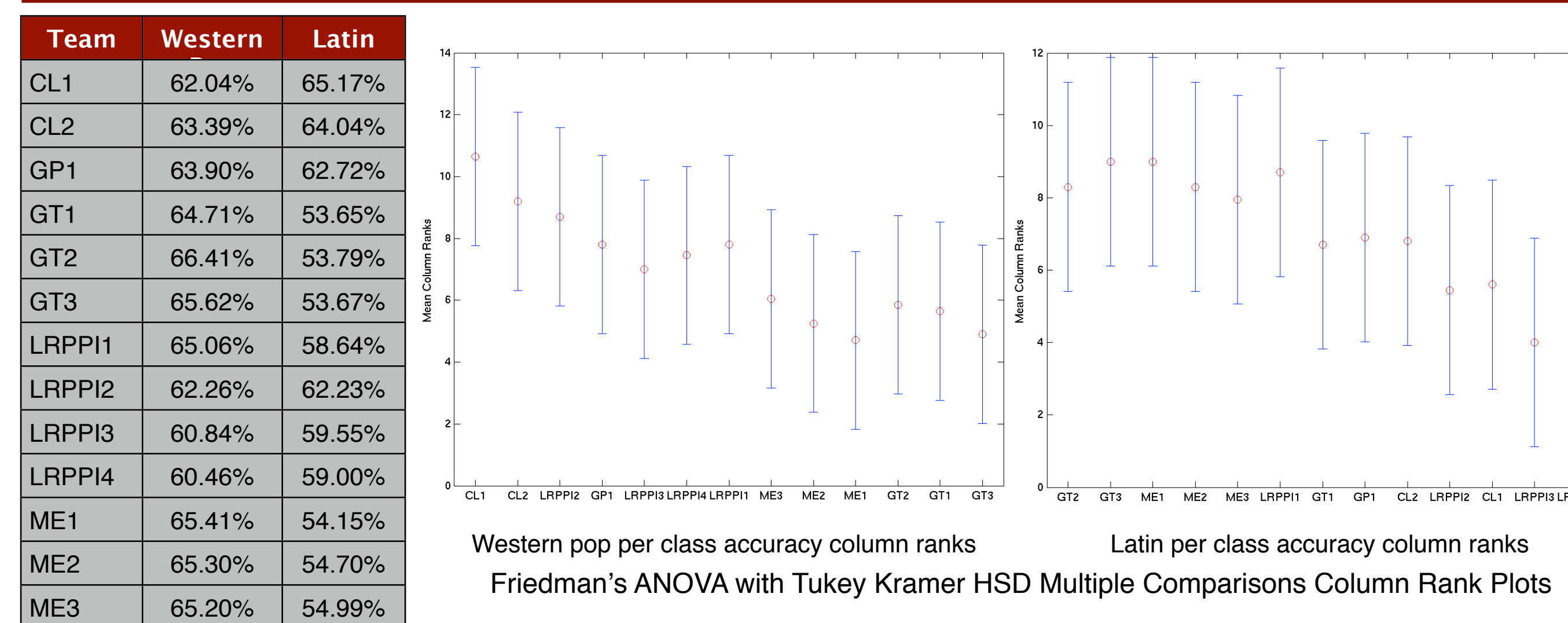
Classification Task Results and Significance Tests

- Friedman's ANOVA is applied to the results in order to equalize the variance inherent in different observations (performance scores over classes or folds).
- Friedman's ANOVA is non-parametric and used in preference to Student's T-test as it does not assume normal distribution of the underlying data.
- Tukey-Kramer HSD multiple comparisons are performed over the Friedman test results to produce a statistically valid pair-wise comparison and to determine if any differences in ranking are significant. Without such a procedure the uncertainty in the pairwise estimates are cumulative and at least one is likely to be wrong.

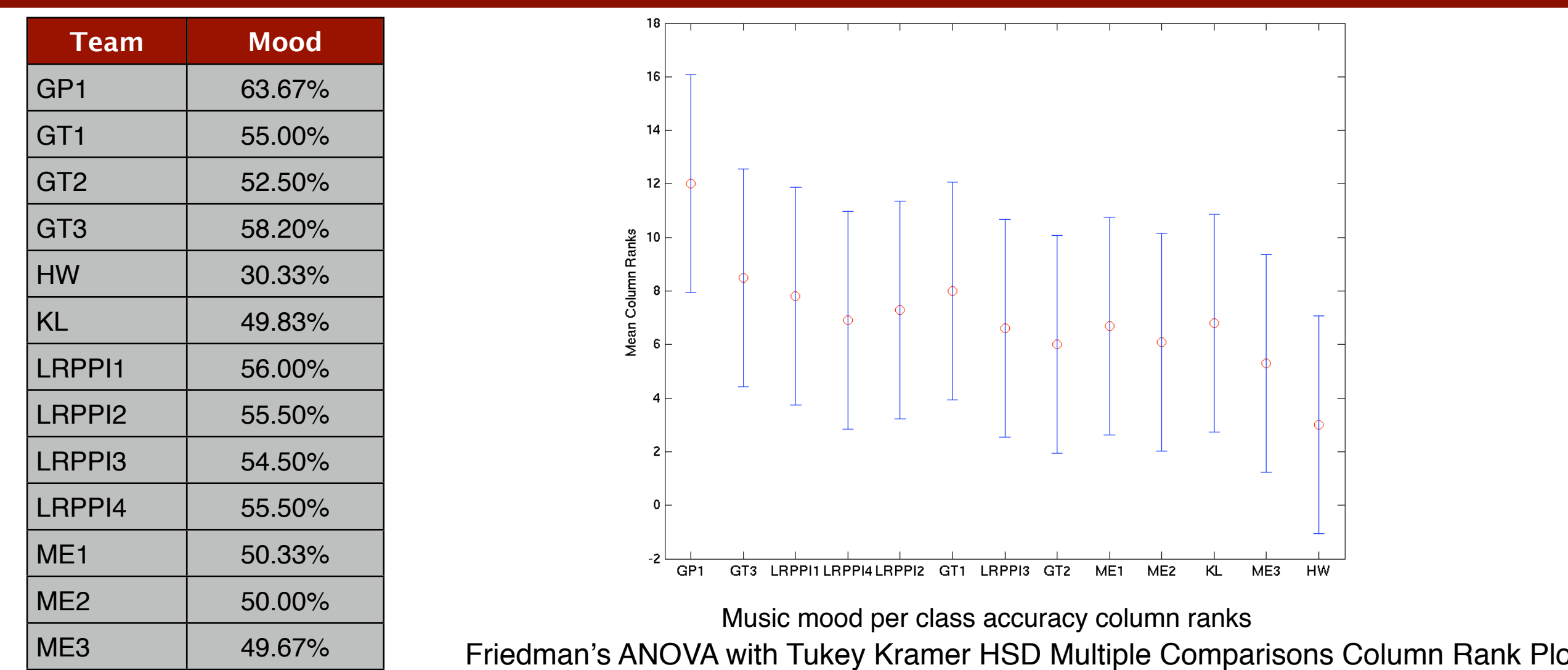
Audio Artist Identification



Audio Genre Classification



Audio Mood Classification



Audio Tag Classification

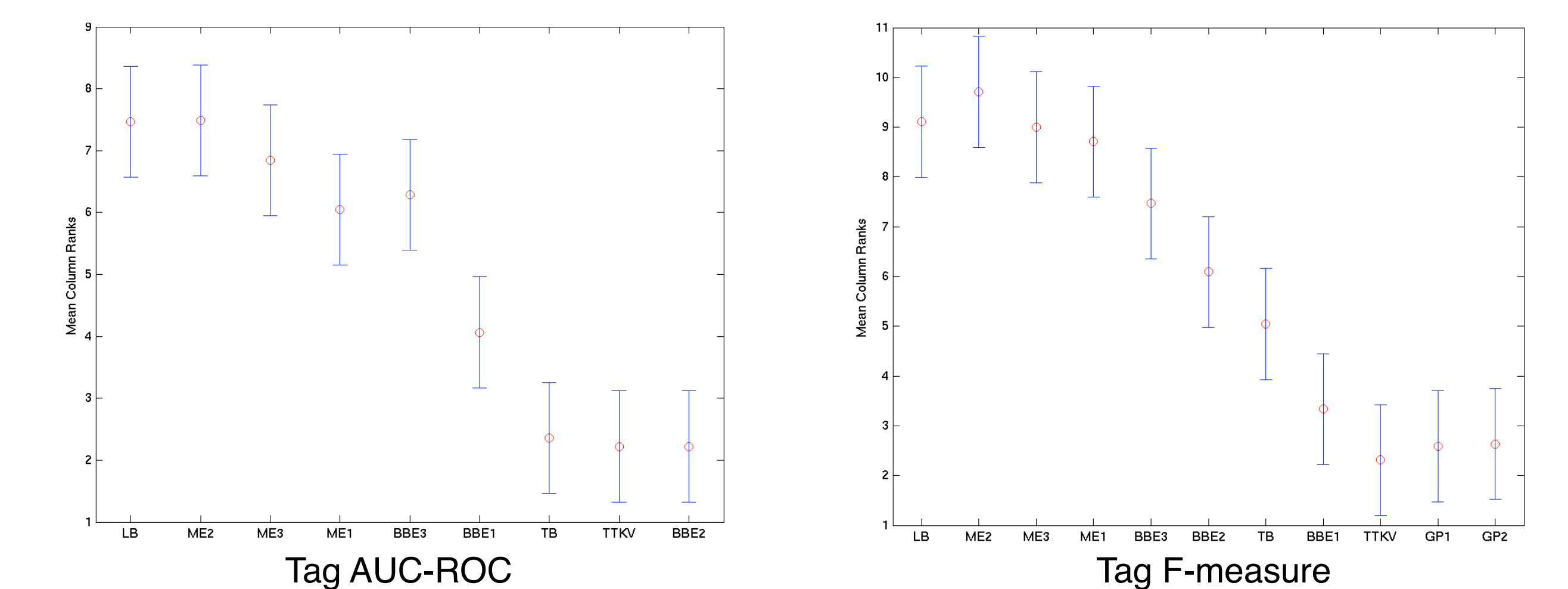
- lected by the MajorMiner game (Mandel and Ellis, 2007)
- players label 10-second clips with arbitrary textual descriptions called tags.
- score points when others describe the same clips with the same tag.
- experiments include tags verified by ≥ 2 players on ≥ 35 clips.
- 45 tags qualify, total of 9000 verifications on 2200 clips.
- MajorMiner data does not include negative labels.
- a negative example of a particular tag is a clip on which another tag has been verified, but the tag in question has not.
- Multiple Evaluation metrics including Accuracy and F-measure (per tag) and Area Under the ROC Curve (AUC-ROC) for both tags and tracks.

Significance Tests

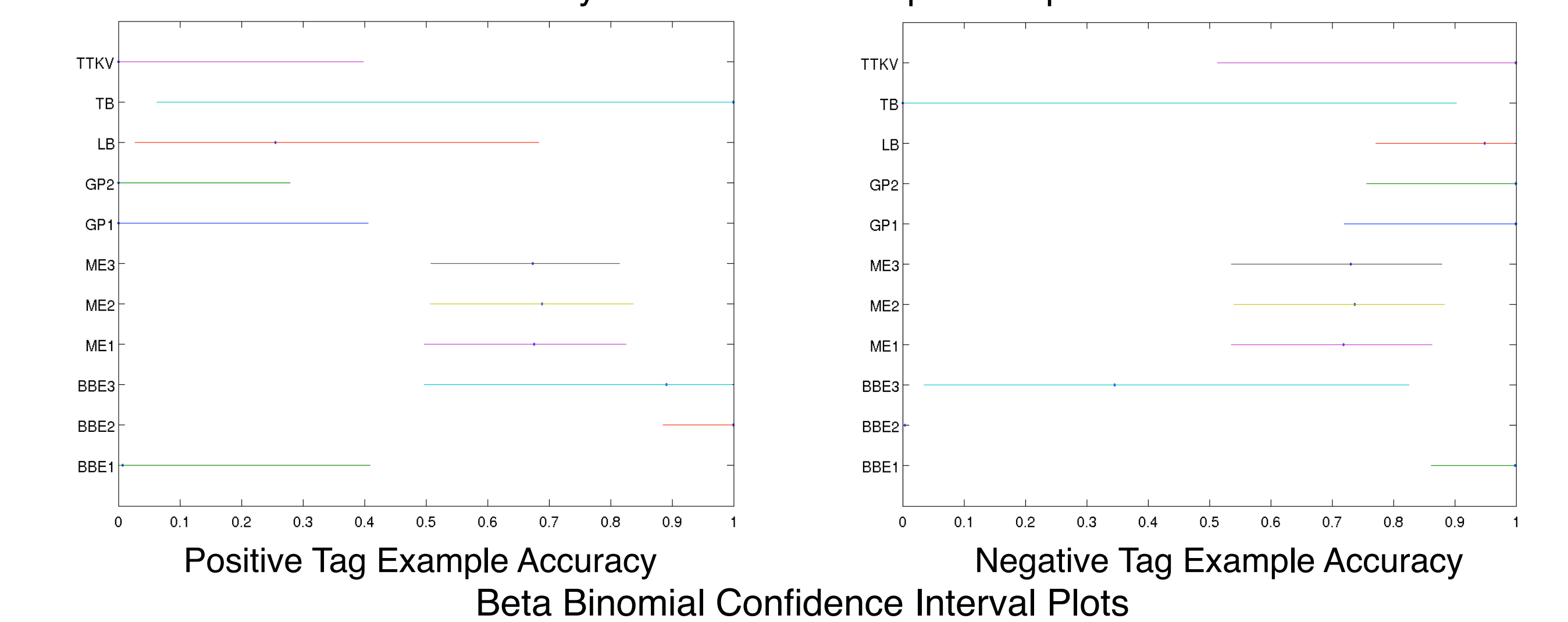
- Friedman's testing with Tukey-Kramer HSD
- Beta-Binomial testing
- Empirical Bayes method for estimating the probability of a set of exchangeable binomial random variables θ_i . (Gelman et al., 2003)
- Hierarchical nature allows the θ_i s to share information, so that if one tag doesn't have many observations, it shrinks its estimate towards the mean of the prior distribution.

Results

Measure	BBE1	BBE2	BBE3	ME1	ME2	ME3	GP1	GP2	LB	TB	TTKV
Average Tag Positive Example Accuracy	0.05	1.00	0.85	0.67	0.68	0.66	0.04	0.03	0.28	0.91	0.03
Average Tag Negative Example Accuracy	0.99	0.00	0.37	0.71	0.73	0.73	0.98	0.98	0.94	0.09	0.97
Average Tag F-Measure	0.06	0.15	0.19	0.24	0.26	0.26	0.03	0.02	0.28	0.15	0.04
Average Tag Accuracy	0.91	0.09	0.43	0.71	0.73	0.72	0.90	0.89	0.90	0.17	0.90
Average AUC-ROC Clip	0.82	0.49	0.81	0.77	0.79	0.78	n/a	n/a	0.84	0.69	0.78
Average AUC-ROC Tag	0.66	0.50	0.74	0.75	0.77	0.76	n/a	n/a	0.77	0.50	0.50
Overall Beta-Binomial Positive Example Accuracy	0.01	1.00	0.89	0.68	0.69	0.67	0.00	0.00	0.26	1.00	0.00
Overall Beta-Binomial Negative Example Accuracy	1.00	0.00	0.35	0.72	0.74	0.73	1.00	1.00	0.95	0.00	1.00



Friedman's ANOVA with Tukey Kramer HSD Multiple Comparisons Column Rank Plots



Positive Tag Example Accuracy Beta Binomial Confidence Interval Plots Negative Tag Example Accuracy