CROSSMODAL AESTHETICS FROM A FEATURE EXTRACTION PERSPECTIVE: A PILOT STUDY

Alison Mattek
Dartmouth College
Alison.M.Mattek@Dartmouth.edu

Michael Casey
Dartmouth College
mcasey@Dartmouth.edu

ABSTRACT

This paper investigates perceptual relationships between art in the auditory and visual domains. First, we conducted a behavioral experiment asking subjects to assess similarity between 10 musical recordings and 10 works of abstract art. We found a significant degree of agreement across subjects as to which images correspond to which audio, even though neither the audio nor the images possessed semantic content. Secondly, we sought to find the relationship between audio and images within a defined feature space that correlated with the subjective similarity judgments. We trained two regression models using leave- one-subject-out and leave-one-audio-out cross-validation respectively, and exhaustively evaluated each model's ability to predict features of subject-ranked similar images using only a given audio clip's features. A retrieval task used the predicted image features to retrieve likely related images from the data set. The task was evaluated using the ground truth of subjects' actual similarity judgments. Our results show a mean cross-validated prediction accuracy of 0.61 with $p<0.0001$ for the first model, and a mean prediction accuracy of 0.51 with $p<0.03$ for the second model.

1. INTRODUCTION

Art, in any of its modes, affects us. Whether an acrylic or symphonic masterpiece, art has the tendency to attract our attention and stir our sentiments, sometimes in ways that are quite similar across modalities. An attempt to define what a work of art is or to identifying exactly why art affects us the way it does are both ambitious and elusive questions in the field of aesthetics. Yet, these seem to be some of the more progressive objectives of music information retrieval. Once we have diluted a sensuous experience such as listening to a symphony into a concrete string of numbers, the source of our pleasure becomes slightly more objective (though our experience of it may remain quite ineffable). This objectivity has allowed us to examine correlations between sets of songs based on musical features. Perhaps, then, feature extraction could also enlighten us to correlations across domains of art. For example, what features contribute to the phenomenon of a particular painting evoke the same feeling as a particular work of music?

This study attempts to bridge artistic domains from the perspective of feature extraction. If works of art that are emotionally ambiguous and culturally unrelated could still be considered similar, it is very possible that there is objectivity in the similarity that lies at the feature level. This opens up an entirely new question in terms of cross- modal analysis: which auditory features and which visual features are important when considering crossmodal similarities? To simplify the plethora of possibilities, the study focuses on a few standard low-level features: course constant-Q spectrograms of the audio and eight band HSV histograms of the images.

2. RELATED WORK

Congruency across sensory modalities is a subject matter that has been discussed in the field of psychology since the seventies [1]. Cross-modal congruencies have been empirically shown to exist across the auditory and visual domains. This is not to be confused with cross-modal confusion, which is what occurs in individuals suffering from synesthesia. Typical audio-visual cross-modal congruency examples are sounds high in frequency being associated with objects high in space and objects small in volume, or vice versa: sounds low in frequency are associated with objects low in space or objects large in volume. Studies in cross-modal congruencies support the hypothesis that art across different domains may affect us in similar ways.

Translations between visual and auditory art have been attempted in both directions. These attempts are known as music visualization when translating from auditory to visual, and image sonification when translating from visual to auditory. Traditional music player software generally come suited with some means of visualizing the music. Researchers have also devised creative means of attempting the audio to visual translation, including the use of affective photos [2] and self-similarity [3]. Mardirossian and Chew also presented a way to visualize music in two dimensions based on the tonal progressions [4]. The translation in the opposite direction, from images to music, has been investigated using the

geometric characteristics of images to create a time-based sequence that could be translated by musical instruments [5].

Although there has never been an explicit attempt to classify images with audio data (as in the current study), one recent study was able to classify music genre by analyzing the promotional images of the artist [6]. This study used image histograms across three color spaces: RGB, HSV, and LAB to cluster image data into classes of musical genre. All of the above mentioned related worksuggests that there are some consistent perceptual relationships between the auditory and visual domains.

3. BEHAVIORAL STUDY

3.1 Data Collection

The first step in finding similarities across modalities was to find pairs of images and audio that were thought to be similar by a group of subjects. This was done via the behavioral experiment described in this section.

3.1.1 Stimuli

Ten abstract art images by the following artists were chosen for this experiment: Betsy Eby, Gerhard Richter, Giles Hayter, Stephanie Willis, Ian Camleod, Madison Moore, Anne Kavanagh, Ernie Gerzabek, Paul Pulszartti, and Jason Stephen. Figure 1 shows "Blueprint I" by Stephanie Willis. All of the images were constructed either in the late twentieth century or early twenty-first century and all artists are Western, to avoid extreme cultural differences. The images were chosen selectively by the authors to encapsulate a range of colors and symmetries and to avoid any conceptual objects (e.g., figures that resemble a tree or a face). All of the image and audio stimuli used in this experiment can be viewed at: http://alisonmattek.wordpress.com/projects/academic/crossmodal/.

Ten ten-second solo piano clips by the following composers were chosen for this experiment: Handel, Mozart, Liszt, Debussy, Hindemith, Barber, Ligeti, Phillip Glass, Bill Evans, and David Lanz. This list represents Western composers across several centuries. The clips were chosen selectively by the author to encapsulate a range of tempos, pitches, and performers, but the timbre was kept relatively consistent, as all of the clips contained only the piano in the instrumentation.

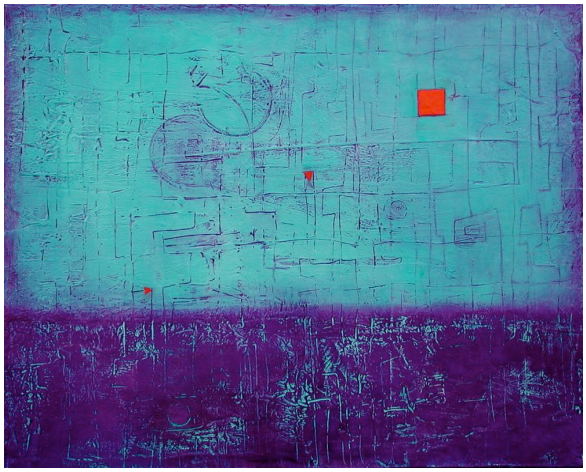

Figure 1. "Blueprint I" by Stephanie Willis

predominantly modern works. In the music selection, had solo piano works been chosen from only the twentieth century as well, there would have been a bias of chromaticism in the harmonic quality of all of the works. In order to achieve more variability in the harmonic structure (that is, to include extremely tonal music), we chose music from previous eras as well. However, the cultural era in which a work was produced is likely a relevant variable, and should be considered in future investigations.

3.1.2 Listening Test

Subjects between the ages of nineteen and thirty years (N = 16, 6 = female, 10 = male) completed a listening test in which they rated the similarities between all pairs of stimuli. Figure 2 shows the graphic user interface for the listening test. Some of the subjects had previous musical training (N = 10, 4 = female, 6 = male). The pairs were presented in a different random order for each subject. The first ten trials of the test were "practice" trials; the subjects were told they could adjust their strategy for choosing a similarity rating during the practice trials. After this, the subjects completed one hundred trials, one for every possible pair of the ten audio clips and ten images. The subjects rated the similarity between each pair on a scale of 1 - 30. 1 - 10 implied "very dissimilar", 11 - 20 implied "average similarity", and 21 - 30 implied "very similar". This 30-point scale was taken from Grey's methodology for multidimensional scaling of musical timbre [7]. The subjects' responses were stored into a ten by ten similarity matrix.

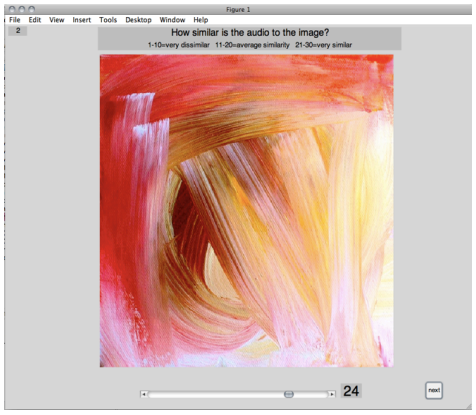

Figure 2. Listening Test GUI

3.2 Data

The results showed correlation across subjects on certain pairs of the audio and images. Figure 3 shows the mean, z-scored similarity matrix across all subjects. High values indicate a pair that was rated as very similar across subjects and low values indicate a pair that was rated as very dissimilar across subjects.

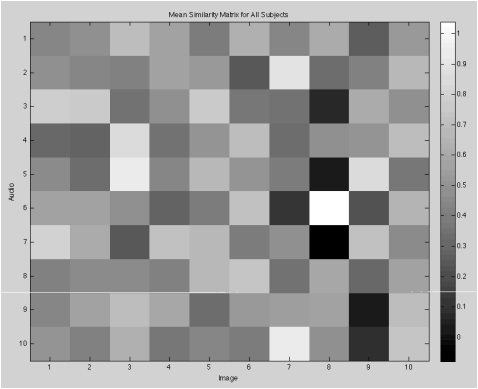

Figure 3. Mean Similarity Matrix for All Subjects

The data was analyzed with plots, covariance matrices, and distance matrices of the z-scored subject responses. Figure 4 shows an analysis of the sixth audio clip, which was an excerpt from Samuel Barber's Excursion No. 1 for solo piano. The plot shows the z- scored subject responses to audio 6 when paired with each of the images, as indicated on the x-axis. What stands out on this plot is that the similarity ratings decrease when audio 6 is compared to image 7, increase when audio 6 is compared to image 8, and decrease again when audio 6 is compared to image 9. In other words, audio 6 was considered to be very similar to image 8, but very dissimilar to image 7 and image 9, with much agreement across subjects.

From this type of analysis on all of the data, the following pairs of images and audio were thought to be similar across subjects: audio 1 and audio 8 were similar to image 6; audio 2 and audio 10 were similar to image 7; audio 3 and audio 7 were similar to image 1, image 4, and image 9; audio 4 and audio 5 were similar to image 3; audio 6 was similar to image 8; and audio 9 was similar to image 10. Images 2 and 5 were not consistently rated as similar to any audio examples. Figure 5 shows image 5, which was not consistently rated as similar or dissimilar to any audio across subjects.

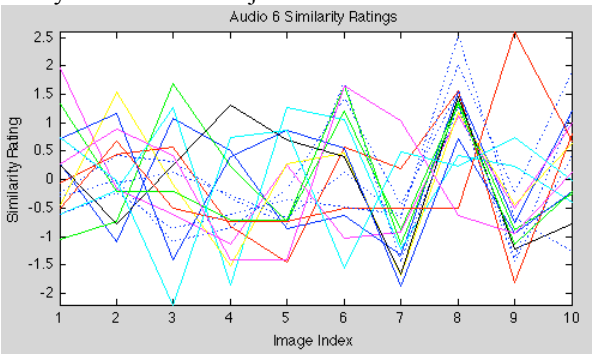

Figure 4. Analysis of Audio 6

Figure 5. "Composition 114-B" by Ian Comleod was not consistently rated as similar to any of the audio examples.

4. IMAGE PREDICTION

Given the subjective cross-modal similarity evaluation, we sought to determine whether there were correspondences in common between the underlying audio and image features spaces. To this end, from the 10 audio clips we extracted average power with a band rate of two constant- Q bands per octave [9]. From the images we extracted eight-band HSV histograms. The HSV representation was chosen over RGB because, like the choice of logarithmic frequency spectrum, the HSV color scale corresponds more closely with human perception than the RGB scale [10]. The HSV values were binned into 3 groups of 8 scalars forming a 24 dimensional vector. The 16 audio bands and 24 image values were independently dimension reduced using a singular value decomposition (SVD) keeping those coefficients corresponding to the first 95% of the total variance in each modality.

4.1 Multivariate Multiple regression

To test the predictability of image features given audio features for an unseen music clip, we performed a retrieval experiment using a cross-validated multivariate multiple regression model [11]. Regression is an optimization method that minimizes the response error for a training set of predictor/response vector pairs (in our case audio features / image features) using a linear model of the form: $\boldsymbol{y}=\boldsymbol{W}^T\boldsymbol{x}+\boldsymbol{b}$, with weight matrix, $\boldsymbol{W}$, predictor variables, $\boldsymbol{x}$, biases $\boldsymbol{b}$, and response variables $\boldsymbol{y}$. Our models consisted of multiple independent variables (audio-feature predictors), and multivariate dependent variables (image-feature responses). Such multivariate multiple regression has previously been applied, in a cross-modal context, to predicting fMRI images corresponding to concrete nouns; where the predictor variables were intermediate vector representations of single words and the response variables were fMRI image voxels [12].

Figure 6 illustrates the method of predicting image features from a regression model trained on audio feature / image feature pairs. Figure 7 shows an example of audio features, a weight matrix, and the predicted response, actual response, and residual images.

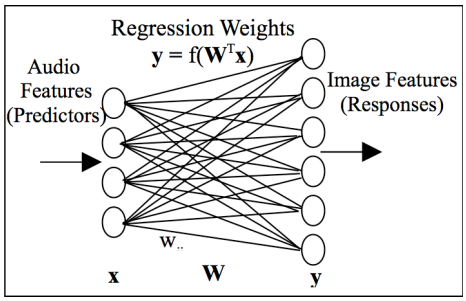

Figure 6. Schematic diagram showing how regression is used to predict response variables from predictor variables. In this paper, the predictor variables are audio features, and the response variables are image features.

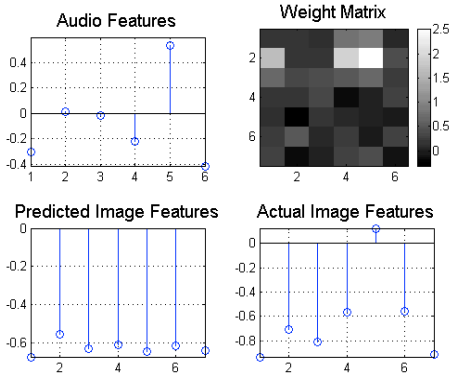

Figure 7. Example of audio features (upper left), trained regression model weights (upper right), predicted image (lower left), and actual image (lower right), for a leave-one-audio-out regression model.

4.2 Training

We trained the regression models using the *mvregress* function from the Statistics Toolbox of the MATLAB numerical scientific package. The training data consisted of the dimension-reduced features of the audio clips as predictor variables and the dimension-reduced image features for each subject's *highest-rated image* (i.e., the most similar image as determined by the similarity judgments) as the response variables. We trained two models: Model 1 was trained using subjects' image response ratings for each audio clip, leaving out one subject's data in each run; Model 2 was trained using all subjects' image response ratings, leaving out one of the audio clips in each run.

4.3 Ground Truth

The data from the behavioral study— i.e. per subject similarity ratings between each audio clip and each image— yielded per subject 10x10 similarity matrices where each row consisted of the image rankings to one audio clip with integer values in the range of 1 to 30. Each subject utilized the scale to a different extent; with some using the full range and others using only part of the available range. To align the different ranges onto a common scale, each row was normalized to the range of 0 to 1. The individual normalized similarity matrices were then averaged yielding a cross-subject mean similarity matrix. From this matrix, a ground truth of relevant images was determined individually for each audio clip, but across subjects, by selecting all images with an average similarity greater than, or equal to, the mean plus one standard deviation of the normalized similarity ratings for that audio clip. This yielded a different number of relevant images for each audio clip ranging from one to three relevant images. These were used as the target images for each audio clip in the retrieval experiments. Note that for Model 1, the ground truth consisted of a mean similarity matrix that excluded the held-out subject; i.e. the test subject's data.

4.4 Prediction

One of the main utilities of regression is that responses can be computed for novel data— such that the response variables interpolate between the training data for previously unseen data. Thus, the trained regression models were used to predict the response variables (image feature vector) for each test feature vector (held-out audio feature vector). A successful interpolation would indicate generality of the model; specifically, the generalization of the subjective cross-modal feature-space mappings, such that the model could be used to predict the human subjective image response to unseen music audio data.

4.5 Evaluation

To evaluate the degree of success of the models' predictions, the set of ground truth images per audio clip was

used in a retrieval task. The two models performed slightly different retrieval tasks: Model 1 left a different subject's predictor/response feature data out per run, for a total of 16 subjects. Here the goal was to was to assess the degree to which an individual subject's responses affect the image prediction result. The model was trained and tested repeatedly, omitting a single subject's data each time, on the set of features corresponding to closest audio-image pairs from the remaining subjects' similarity scores. To test, a response image feature was predicted for each audio feature using the regression weights. The cosine distance was computed between the predicted image feature vector and the set of 10 feature vectors for the 10 images that the test subject ranked in the behavioral experiment. The distances were sorted such that those images whose features were most similar to the predicted features were ranked more highly in the list of retrieved images. Precision and recall values were computed by comparing each ranked image with the relevant image set (ground truth). The recall level was also calculated; i.e. the proportion of ground truth images retrieved for each position in the retrieved image list. The mean precision was calculated by summing over all precision values and dividing by the total number of relevant items across all trials. Additionally, an f-score was computed using the 2P.R/(P+R) statistic and the mean f-score computed in a similar manner as the mean precision. Empirical p-values were computed using the distribution of mean precisions for 10,000 trials of randomly ordered image draws versus image draws ordered by similarity to the regression model's predicted images. The resulting probability is interpreted as the empirical probability that retrieval using randomly permuted image draws performed at least as well as retrieval using regression.

Model 2 was evaluated to test the generality of the model for unseen audio data. For this model, a leave-one-audio-out cross validation paradigm was used. Here, each training iteration omitted the audio / image feature pairs corresponding to one of the audio clips for all subjects. Testing consisted of predicting response image features for each held-out audio feature. As in Model 1, the cosine distance between each predicted image feature vector and the set of ground-truth images for the held-out audio clip yielded a ranked retrieval list of images that was used to calculate precision, recall, f-measure, and p-values, as discussed above.

By leaving one example out for testing, the models used 16-fold and 10-fold cross-validation respectively, a commonly used statistical technique for estimating the generalization power of a given model. Furthermore, 10-fold cross-validation has been shown to be one of the best methods to use for model selection [13].

5. RESULTS

The results of both image prediction experiments are shown in Table 1. We performed a sensitivity analysis by systematically selecting subsets of features from the predictor and response variables used for the regression and retrieval. In Table 1, results are shown both for the full ensemble and the best performing subsets of audio and image features. For the best-performing subset of features, 3 audio dimensions were left out and 2 image dimensions. The p-values for the average precision were $p<0.0001$ for Model 1 and $p<0.03$ for Model 2. Figures 8 and 9 show the precision-recall curves for the two models for $1/10^{th}$ percentile standardized recall levels.

Model	*# trials*	*avg. precision*	*avg. f-score*	*p-value*
1 (full)	**160**	**0.498**	**0.311**	**p<0.0001**
2 (full)	**18**	**0.299**	**0.248**	**0.867**
1 (subset)	**160**	**0.605**	**0.366**	**p<0.0001**
2 (subset)	**18**	**0.511**	**0.321**	**0.028**

Table 1. Cross-validation results for regression model audio-image feature prediction of 16 human subjects' image response data to music stimuli. The subset model used four of seven audio features, and five of seven image features.

Both versions of Model 1 perform significantly better than chance, with the per-subject-validation yielding a significance score of $p<0.0001$ ($p=0$ for 10,000 trials). However, only the feature subset version of Model 2 performed significantly above chance with $p<0.03$. The difference in performance between the two experiments is not wholly surprising. In the first experiment, the predictor/response data for a single subject is left out, but there are still 15 complete sets of audio-image data on which to train the regression model. Figure 8 illustrates the degree to which individual subjects' data influences the overall result. The spread of the mean precision across individual runs is limited. Hence, we conclude that no one subject is contributing significantly more to the result than any other.

Figure 9 illustrates that the spread of results for the held-out audio-image data, across all subjects, varies significantly. This indicates unequal contributions to the model from different audio predictors and their corresponding cross- subject image responses.

6. CONCLUSIONS

The results of this study show that it is possible to predict the relationship between artistic examples from both the audio and visual domains using feature extraction. Our perceptions of art are complex and multidimensional, even within a single domain, so multiple features from

each domain are likely contributing to the similarities perceived across domains. This makes the investigation of cross-modal congruencies within feature spaces particularly challenging.

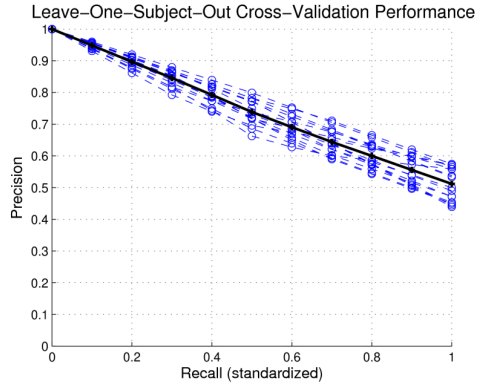

Figure 8. Retrieval performance for Model 1 showing the mean precision of individual runs (dashed lines) and the mean precision taken over all runs (solid line).

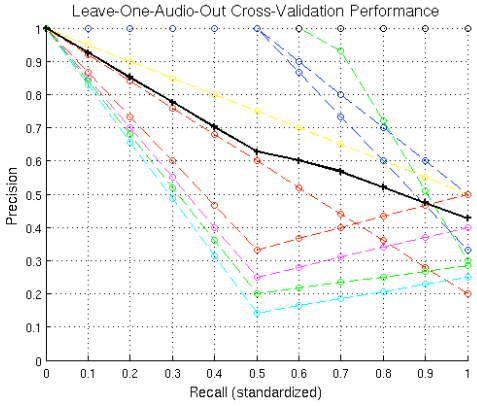

Figure 9. Retrieval performance for Model 2 showing the mean precision of individual runs (dashed lines) and the mean precision taken over all runs (solid line). Here, the model is trained on all subjects' most similar audio-image feature pairs for left-in audio.

Further research can investigate the correlations between multiple features of audio and images. The choice of features in this study was somewhat arbitrary, but seemed like an intuitive place to start. The techniques used here demonstrated the use of low-level features. However, the complexity of the problem suggests that many more features are contributing to the relationship between domains.

A primary limitation of the results of this study is a possible lack of generalizability due to the small size of the data set. The data set was kept small out of consideration for the behavioral experiment design. The subjects had to give similarity ratings for all possible combinations of visual and auditory art, which amounted to 100 trials total. With this amount of stimuli, the behavioral test took 30-40 minutes. Adding more stimuli would cause the behavioral test to increase in length exponentially. Considering the attention span of subjects is important in this regard, because an experiment that was much longer could have compromised the integrity of the responses.

Research in the area of cross-modal congruencies provides a step towards understanding the perceptual processes related to cross-modal binding. Our minds our constantly receiving input streams from various senses and must use them to create the continuous and whole experience of consciousness. Identifying how modality- specific features relate and integrate across domains is a fundamental part of the discovery of our constant reality, *e pluribus unum*.

7. REFERENCES

[1] Marks, Lawrence E. "On cross-modal similarity: Audio-visual interactions in speeded discrimination." *Journal of Experimental Psychology: Human Perception and Performance*. Vol. 13, No. 3, pp. 384-394, 1987.

[2] Chen, Chin-Han, Ming-Fang Weng, Shyh-Kang Jang, and Yung-Yu Chuang. "Emotion Based Music Visualization Using Photos." *Advances in Multimedia Modeling*. Springer, Berlin: 2008.

[3] Cooper, M. and J. Foote. "Visualizing Music and Rhythm via Self-Similarity." *Proceedings ICMC*, 2002.

[4] Mardiossian, Arpi and Elaine Chew. "Visualizing Music: Tonal Progressions and Distributions." *Proceedings ISMIR*, 2007.

[5] Yeo, Woon Seung and Jonathon Berger. "Application of Image Sonification to Music."*Proceedings of ISMIR,* 2005.

[6] Libeks, Janis and Douglas Turnbull. "Exploring 'Artist Image' Using Content-Based Analysis of Promotional Photos." *Proceedings of the International Computer Music Conference*, 2010.

[7] Grey, John M. "Multidimensional perceptual scaling of musical timbres." *Journal of the Acoustical Society of America*. Vol. 61, Issue 5, pp. 1270 – 1277, 1979.

[8] Hoffman, Thomas, Jan Puzicha, and Michael I. Jordan. "Learning from dyadic data." *Proceedings of the 1998 conference on Advances in Neural Information Processing Systems II*. MIT Press, Cambridge, MA: 1999.

[9] Tzanetakis, George and P. Cook. "Musical Genre Classification of Audio Signals." *IEEE Transactions on Speech and Audio Processing*, Vol. 10, No. 5, July 2002.

[10] Deselaers, T., D. Keysers, and H. Ney. "Features of image retrieval: An experimental comparison." *Information Retrieval*, 2008.

[11] McCullagh, P., and J.A. Nelder, *Generalized Linear Models,* 2nd edition, Chapman&Hall/CRC Press, 1990.

[12] Mitchell, Tom M., S. V. Shinkareva, A. Carlson, K. Chang, V. Malave, R. Mason, and M. A. Just. "Predicting Human Brain Activity Associated with the Meaning of Nouns." *Science*, Vol. 320, No. 5880, 2008.

[13] Kohavi, Ron. "A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection." *International Joint Conference on Artificial Intelligence* (IJCAI), 1995.