Introduction to EEG Decoding for Music Information Retrieval Research

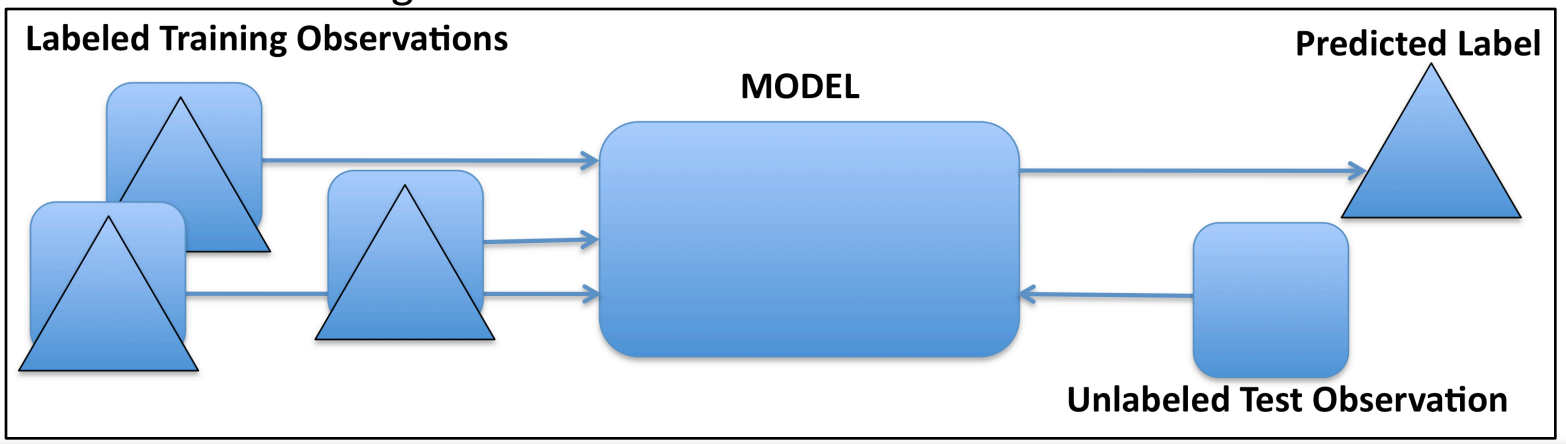
SINGLE-TRIAL CLASSIFICATION

Classification

- **Classification**
  - Fit a statistical model/learning model to a set of labeled observations
  - Use model to predict labels of new (unlabeled) observations
- Classification typically implies **categorical** labels (discrete classes)
- **Regression** problems attempt to assign continuous labels (e.g., numerical scores) to new observations

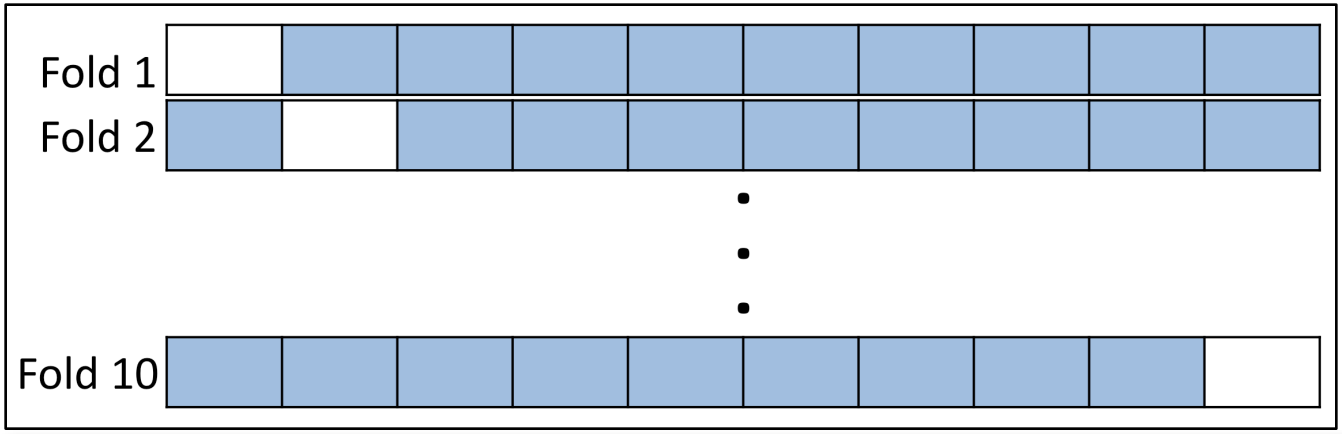
Classification

- **Observation**: One instance of data (e.g., one EEG trial)
- Every observation is composed of a **feature vector** – a set of categorical or numeric descriptors describing the observation. For EEG, this would be the time-sampled voltages, or some transformation of them.
- **Label**: Some specification of the stimulus (stimulus name, stimulus category)
- For the present context, all observations have labels
  - **Training** observations: Observation plus label, used for learning the model
  - **Test** observations: Observation only (label is withheld); label is predicted and later checked against actual label



Cross-validation

- Training and testing on same set of observations may lead to overfitting → partition data into separate training and test sets
- Perform **cross-validation** to get better sense of general performance of classifier.
- **n-fold** cross validation refers to the number of partitions.
  - n=2: Split data in half
  - n=# trials: Test partition contains only one trial. Also known as leave-one-out cross validation (LOOCV)
  - n=10 is fairly standard number of folds (10-fold)



- Every observation is used for testing exactly once

Classifier output

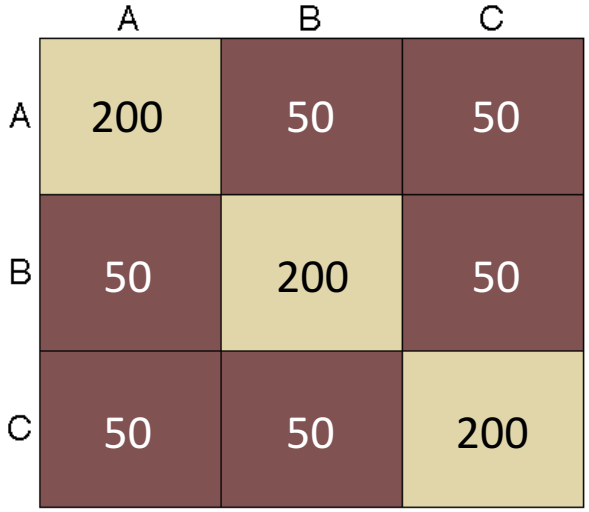
- Every classification task outputs a **confusion matrix** (CM)
  - Rows represent actual labels and columns represent predicted labels*
  - Element in row *i*, column *j* (CM_{ij}) denotes how many observations with actual label *i* were labeled as *j* by the classifier
  - Elements on the diagonal (*i=j*) denote correct classifications
- Row sums express total number of observations actually belonging to a given class

	A	B	C
A	200	50	50
B	50	200	50
C	50	50	200

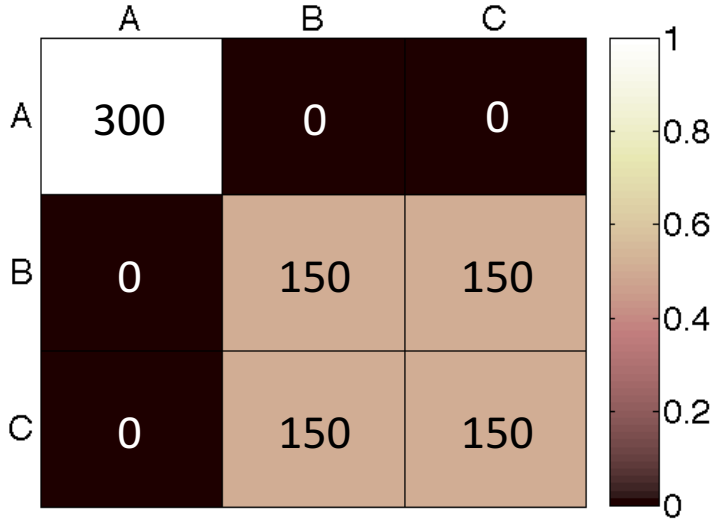
*Transpose is also possible.

Classifier accuracy

- The **accuracy** of the classifier is the percentage of test observations whose labels were predicted correctly.
  - Accuracy = (sum of CM diagonal) / (sum of CM)
- Confusion matrices contain more information than accuracies!



(600) / (600 + 300) = **66.67%**



(600) / (600 + 300) = **66.67%**

Music-based classification studies

- Schaefer et al. (2011): Short, naturalistic music excerpts.
- Vlek et al. (2011): Identify a musical meter imagined over isochronous beats.
- Vlek et al. (2011): A classifier trained on perceived metrical accents can be used to label imagined accents.
- Kaneshiro et al. (2012): Classify tonal function of cadential events in short chord progressions.
- Stober et al. (2014): Classify rhythm families and specific rhythms.
- Treder et al. (2014): Determine which stream in a polyphonic excerpt was attended to using intermittent oddballs.

Practical considerations

- Dealing with inconsistent data
  - Best to have consistent length feature vectors (e.g., same number of time samples, same electrodes involved in classification)
  - May therefore want to use stimuli that are the same length, or epoched to the same length for analysis
- Equal number of trials per category
  - Classification results are easier to interpret when each class contains the same number of observations (trials)
  - If unequal trials are presented, can subset larger classes to balance classes
- Missing values in the data
  - Many classifiers cannot handle missing values (NaNs) in the data
  - Possible solutions: Interpolate missing values; remove the feature (e.g., a bad electrode) from everyone's data. If you have ample trials, it may be possible to omit bad trials from classification.

Practical considerations

- Curse of dimensionality
  - EEG data are high-dimensional!
  - 32 electrodes, 500-ms trials, sampling rate of 100Hz: 32*50= 1,600 features
  - 256 electrodes, 500-ms trials, sampling rate of 1000Hz: 256*500 = 128,000 features
  - Data from adjacent electrodes contain correlated information
  - Data dimensionality may be necessary prior to or during classification
  - Examples: Spatial filtering, orthogonalization of features, feature selection
- Computing statistical significance
  - Present approach: Null binomial distribution (probability of achieving observed number of successes)
  - Can also do permutation test (randomize the labels)

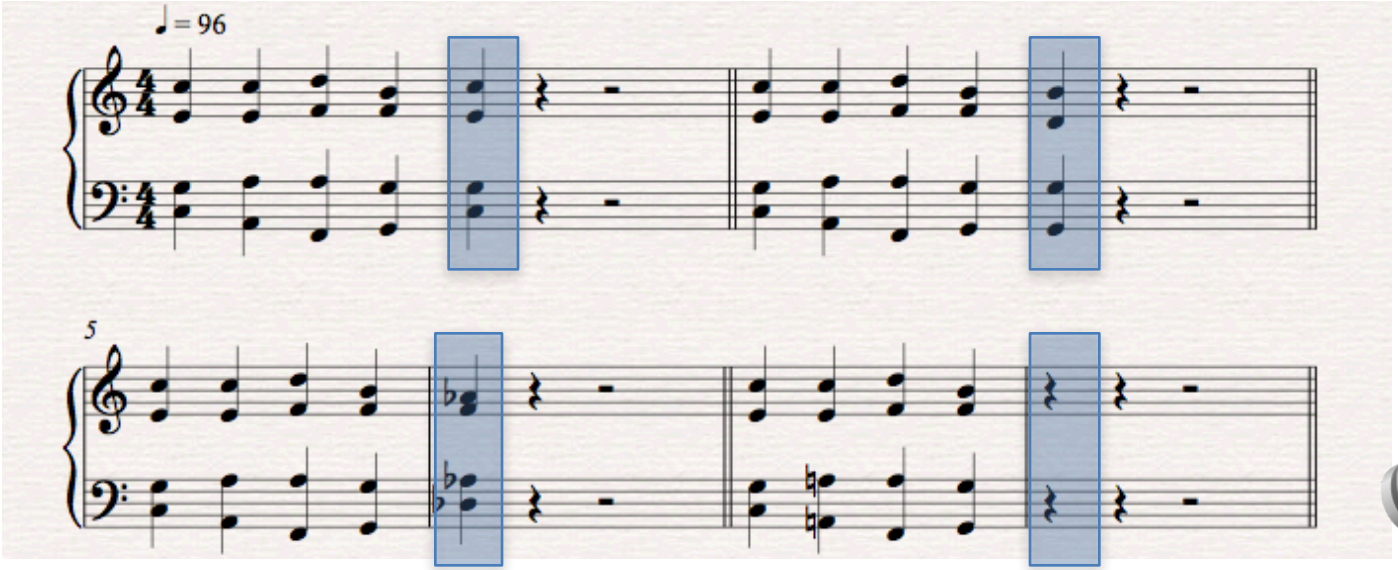
Practical considerations

- Anecdotal observations
  - Want to get at least **~100 trials** per stimulus
  - Classifier accuracy appears to peak around 200-300 trials per stimulus
  - Classification seems to work better when performed **within-participant** (and may be more appropriate for clinical case)
    - Perform separate classifications on each participant's data
    - Interpret mean results across participants
    - However, data can be combined across participants and a single classification performed
  - May want to collect behavioral data separately from EEG
- May be helpful to validate classifier using publicly available datasets from past classification studies

Example: Chord progressions[1]

Stimuli

- Short chord progressions setting up expectation for tonic resolution to cadence
- 4 endings (tonic, dominant, flatted supertonic, silence) x 3 keys (C Major, B Major, F Major) = 12 stimuli
- Stimuli and brain data are publicly available[2]



[1] Kaneshiro et al. (2012). ICMPC. [2] Kaneshiro et al. (2015). Stanford Digital Repository.

Interlude: Live Demo – Part 2

- basic analysis of the epoched data
  - from responses to short chord progressions https://purl.stanford.edu/js383fs8244
  - using MNE-Python: https://github.com/mne-tools/mne-python

- jupyter notebook available at: https://github.com/sstober/ismir2016eeg-tutorial

Example: Chord progressions

- Procedure
  - Pilot data from 2 participants
  - Each participant heard each stimulus at least 108 times (tonic progressions were presented more)
  - Stimuli presented in single-key blocks
  - Interested in single chord event for rare, deviant stimuli → each participant completed **12 20-minute subsessions across 4 visits to the lab** (not a feasible design!)
- Analysis
  - Data records were filtered and cleaned
  - Classification performed using two approaches
  - Classifications performed within-participant; reported results are the average across participants

Classification tasks

- 4-class: What was the tonal function of the cadential event?
  - Classes: Tonic, repeated dominant, flatted supertonic, silence (grouped across keys)
  - Expected result: Significantly above chance. Most listeners can process tonal function independent of musical key.
- 3-class: In which key was the progression presented?
  - Classes: C Major, F Major, B Major (grouped across tonal functions)
  - Expected result: Around chance level. Most listeners are not perceptually sensitive to global key.
- 12-class: Which cadential event was presented, **and** in which key?
  - Classes: No grouping of stimuli.
  - Expected result: May expect confusions among chord events sharing tonal functions (e.g., responses to C-Major tonic may be confused with responses to B-Major tonic).

Classification approach 1

- 10-fold LDA with PC decomposition
  - Simplified version of classifier used in Kaneshiro et al. (2012)
  - Dimensionality reduction
    - Initial feature vector for each trial is concatenated channels by time, length 124 x 39 = 4,836.
    - PCA performed on the trials x feature matrix (all trials) using SVD. PCs are computed along the feature dimension.
  - 10-fold classification using the first 200 PCs (~20x reduction in data dimensionality) of the data

Classification approach 2

- Convolutional Neural Network (CNN):
  - 1 pre-trained spatial filter
  - linear SVC on top as basic classifier
  - 9-fold nested cross-validation (4-fold tuning)

- Pre-training technique:
  Similarity-Constraint Encoding (SCE)
  - learn signal filters that lead to **distinguishing** (temporal) patterns for the different classes

Classification approach 2

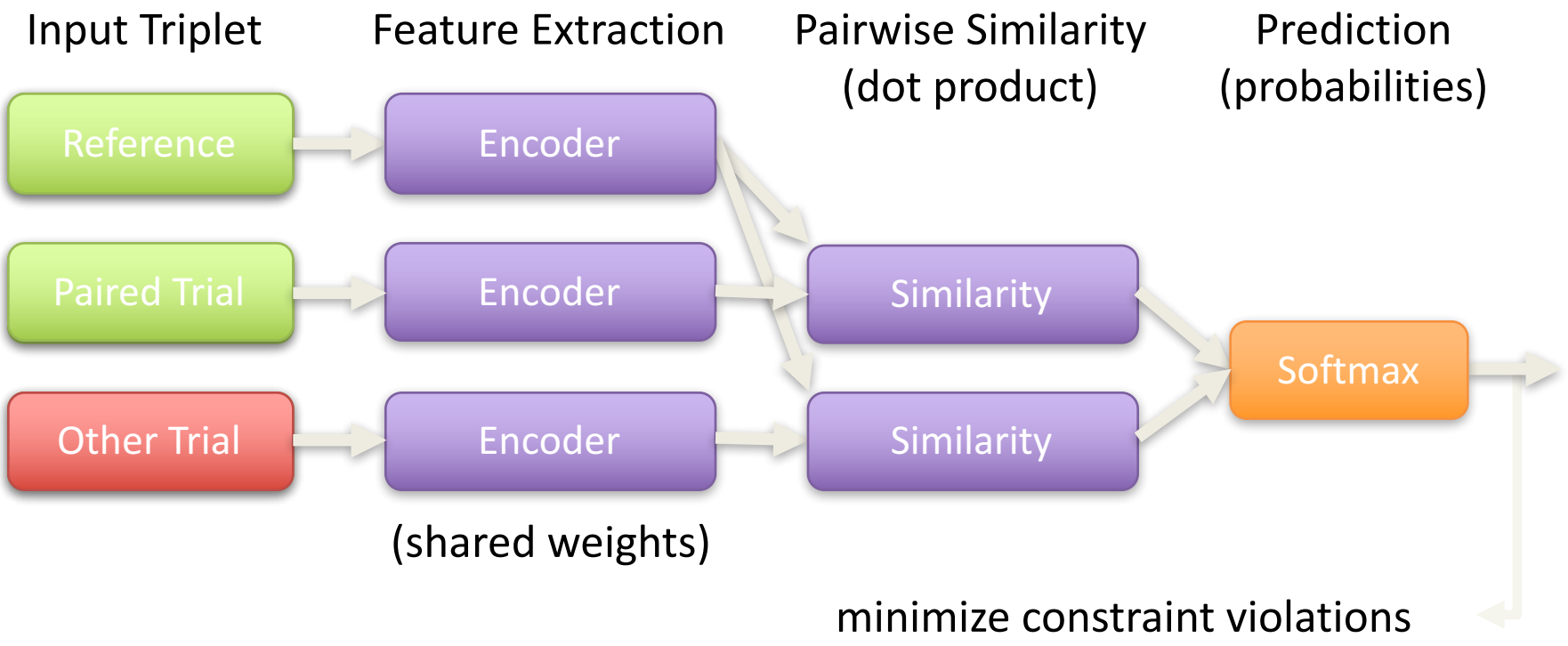
- motivated by relative constraints used for metric learning:
  - for all paired trials (A,B) + trial C from other class:

$$sim(A,B) > sim(A,C)$$

- many combination for (A,B) and C
- favors features that are representative and allow to distinguish classes

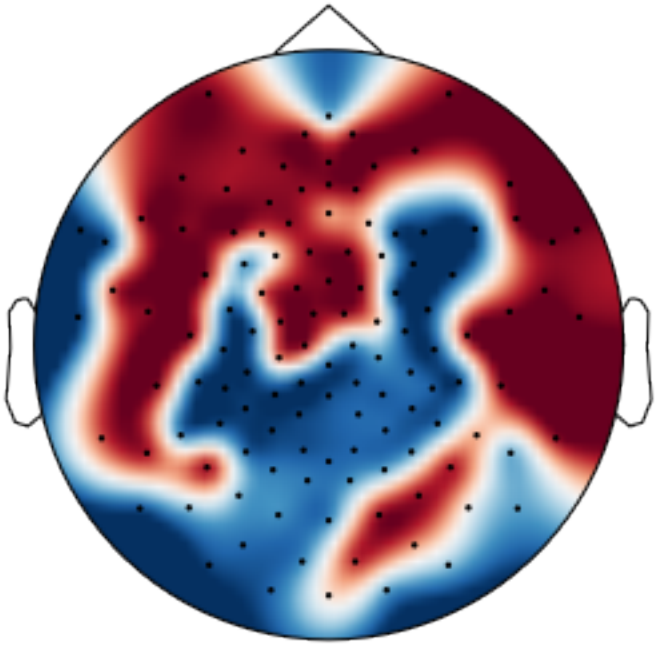
Classification approach 2

Similarity-Constraint Encoder

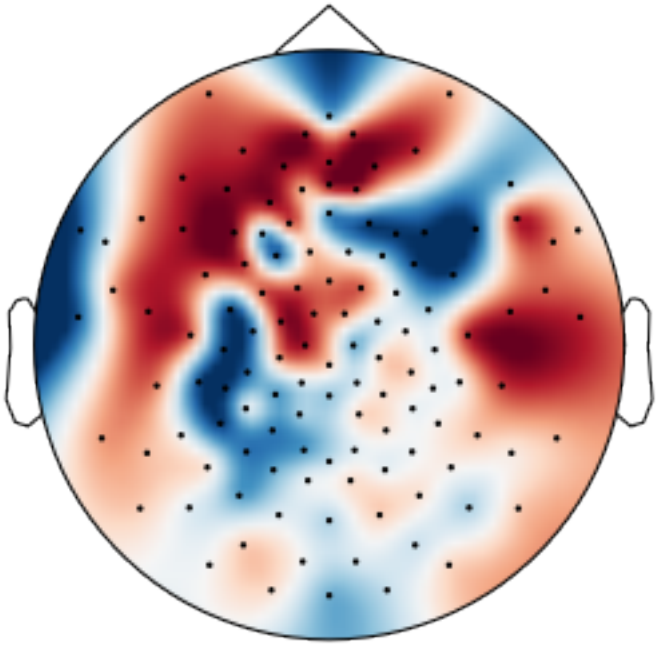


Classification approach 2

- pre-trained filter per subject:



Subject P01



Subject P02

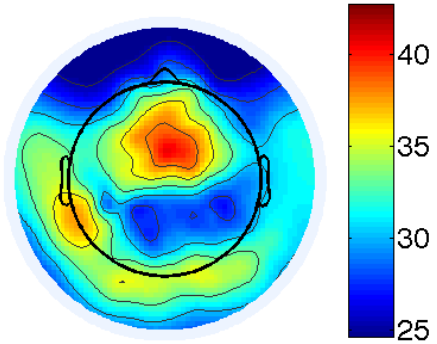
Classification results

- Results are mean accuracy across both participants.
- Statistical significance computed under the null distribution of the binomial distribution
  - Number of successes where chance level is 1/nClasses
  - Sample size: Number of trials in **one test fold** (conservative approach)

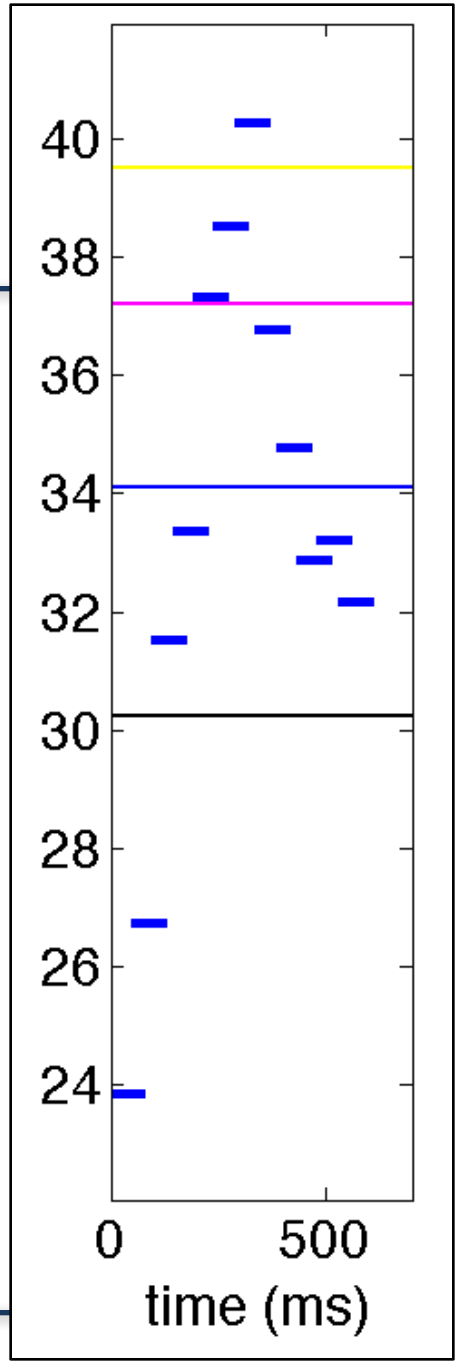
	4-class	3-class	12-class
Chance level	25.00%	33.33%	8.33%
Method 1	**42.57%**	**37.36%**	**14.36%**
	$p < 10^{-5}$	$p = 0.15$	$p = 0.01$
Method 2	**41.74%**	**36.96%**	**14.66%**
	$p < 10^{-4}$	$p = 0.20$	$p = 0.01$

Other analyses[1]

- Break down the problem: Classify spatial and/or temporal subsets of the response to identify where and when classification is successful.



Mean single-electrode rates, 4-class



Mean temporally resolved rates, 4-class

[1] Kaneshiro et al. (2012). ICMPC

Conclusions: Classification

- Tonal processing has been well studied using ERPs but can also be assessed using a classification approach
- Tonal function of cadential events can be labeled from single EEG trials significantly above chance
- Results generalize across musical keys
- Different classification methods achieve similar classifier accuracies
- Useful spatiotemporal features of the brain response can be identified using classifier features, as well as by breaking down the problem

REPRESENTATIONAL SIMILARITY ANALYSIS

Representational Similarity Analysis

Representational Similarity Analysis (RSA)[1,2]

- Use pairwise (dis-)similarities among a set of items to compare their structure across representations
  - Representational Dissimilarity Matrix (RDM) summarizes all pairwise dissimilarities
  - RDM is typically a (*1 – correlation*) matrix
- Compare brain responses, behavioral assessments, computational models, etc. through each modality's RDM.
- Uses visualization methods (MDS, dendrograms) to display the representational space.

[1] Kriegeskorte et al. (2008). Frontiers in Systems Neuroscience. [2] Kriegeskorte et al. (2008). Neuron.

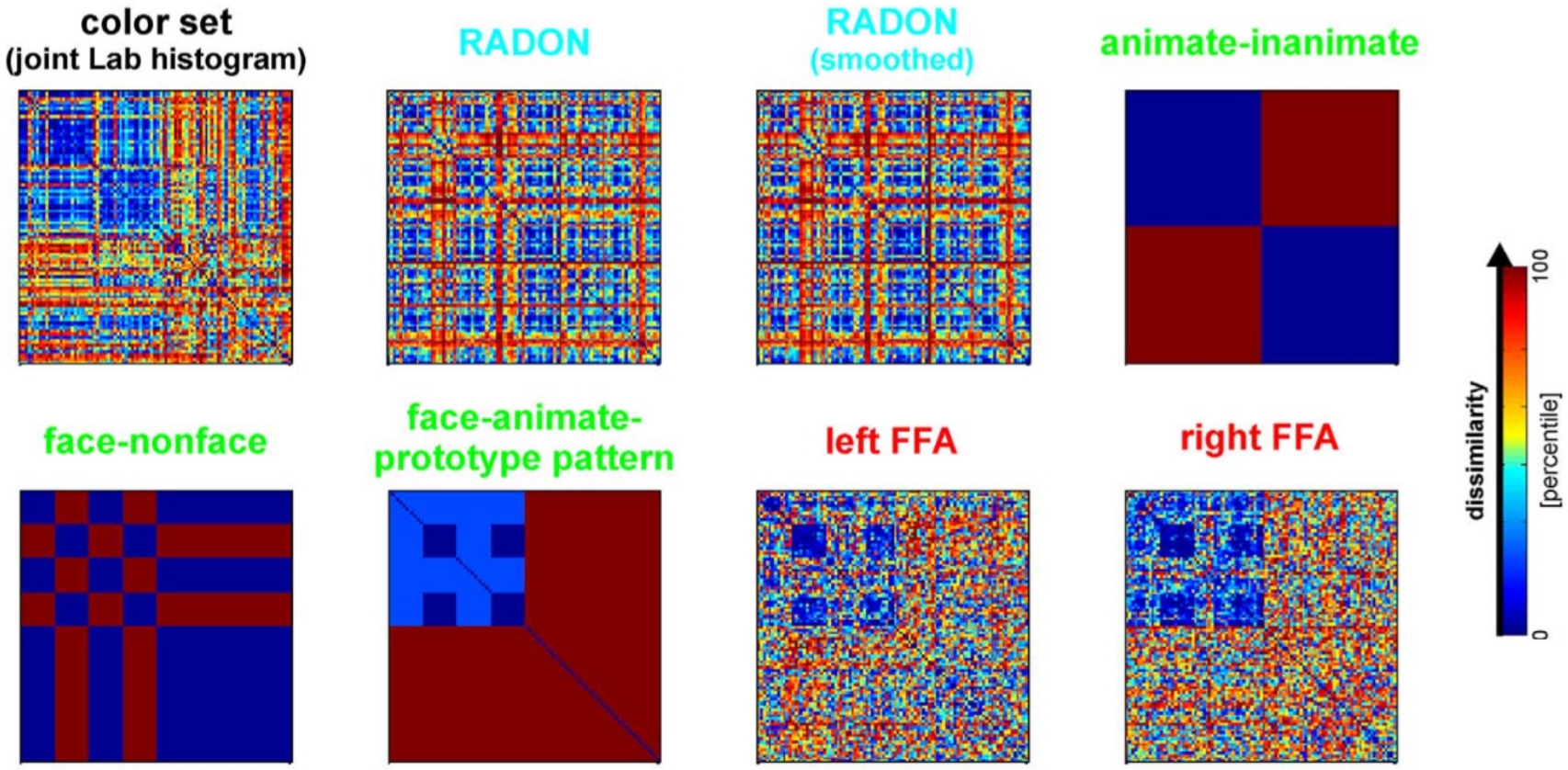
Representational Similarity Analysis

First introduced by Kriegeskorte et al. (2008)[1] with visual responses recorded using fMRI

- Set of 92 images with hierarchical structure
  - Top level: Animate/inanimate
  - Animate splits into human/animal faces/bodies
  - Inanimate splits into natural and man-made
- Derived categorical structure with similarities between cortical (fMRI) responses and computational models; human and monkey[2] cortical responses

[1] Kriegeskorte et al. (2008). Frontiers in Systems Neuroscience. [2] Kriegeskorte et al. (2008). Neuron.

RSA example: fMRI correlations (vision) – Kriegeskorte et al., 2008[1]



RDMs of a set of 92 images from simple (black) and complex (blue) computational models, conceptual models (green), and brain responses (red).

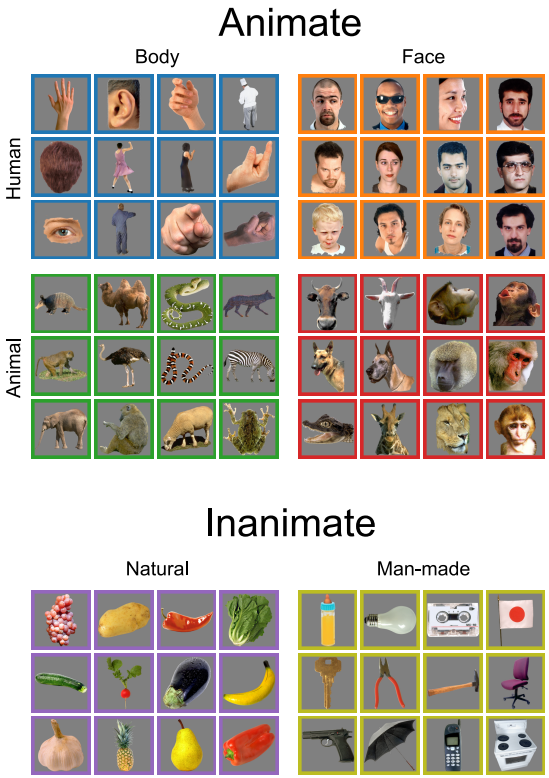
[1] Kriegeskorte et al. (2008). Frontiers in Systems Neuroscience.

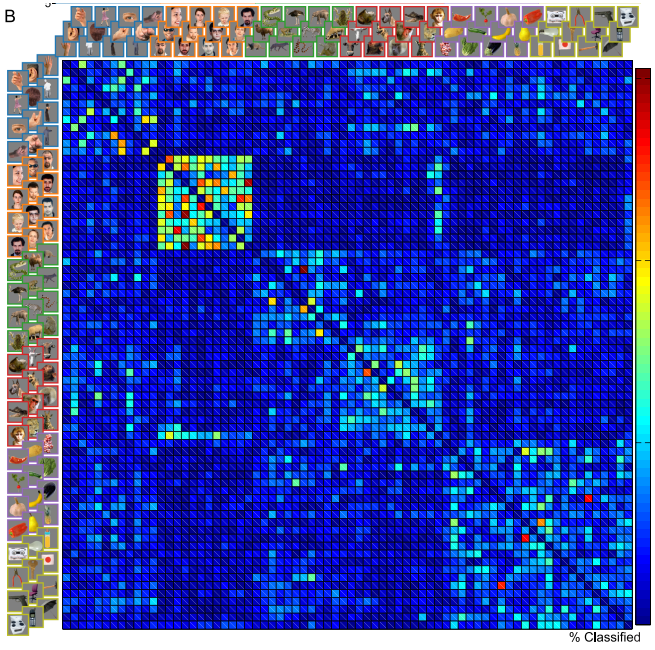
RSA: Classification

- EEG classification setting
  - Cannot simply correlate single EEG trials due to noise and data dimensionality
  - Classification is a useful tool for deriving RDMs from EEG
    - Similar inputs (stimuli) produce similar outputs (brain responses)
    - If EEG responses to stimuli *x* and *y* are often confused by the classifier, the responses must be fairly similar
    - If responses to *x* and *z* are easy to classify, the brain responses must be more distinct
  - Can do all classifications of all stimulus pairs – accuracies reflect distance
  - **Misclassifications** of EEG trials may reflect stimulus similarity – can use the confusion matrix and multi-category classification!

RSA example: EEG classification (vision)[1]

- Multi-category classification on EEG-recorded responses to 72 images from Kriegeskorte (2008) study.
- 10 participants each completed 72 trials of each image.
- Confusion matrix converted to RDM and visualized using MDS and dendrograms.
- Raw and preprocessed brain data are publicly available.[2]



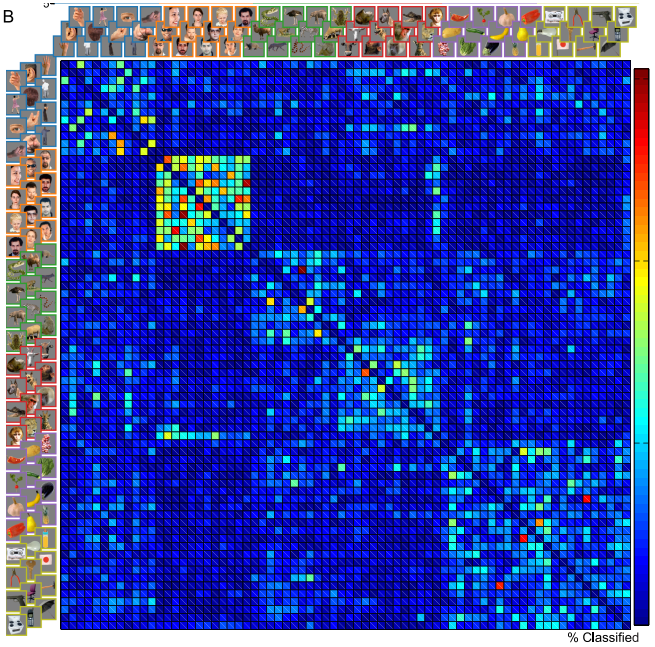


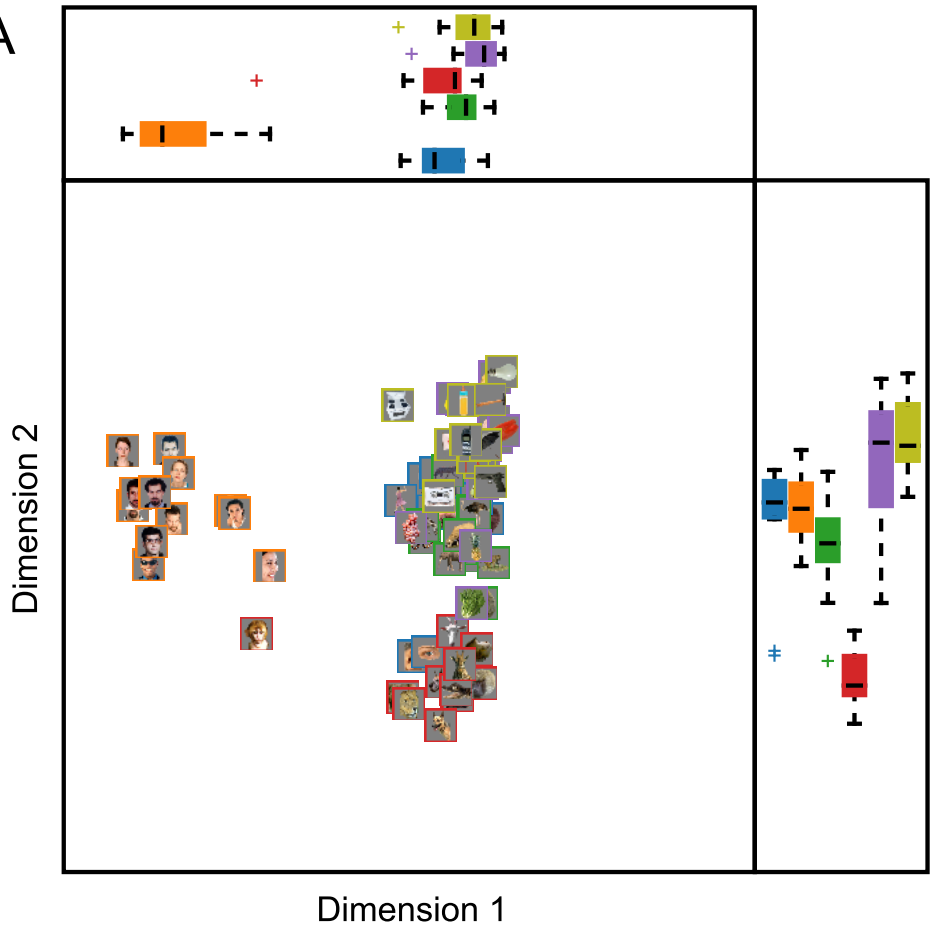
[1] Kaneshiro et al. (2015). PLoS ONE.

[2] Kaneshiro et al. (2015). Stanford Digital Repository.

RSA example: EEG classification (vision)

Can visualize the structure of the confusion matrix using MDS.

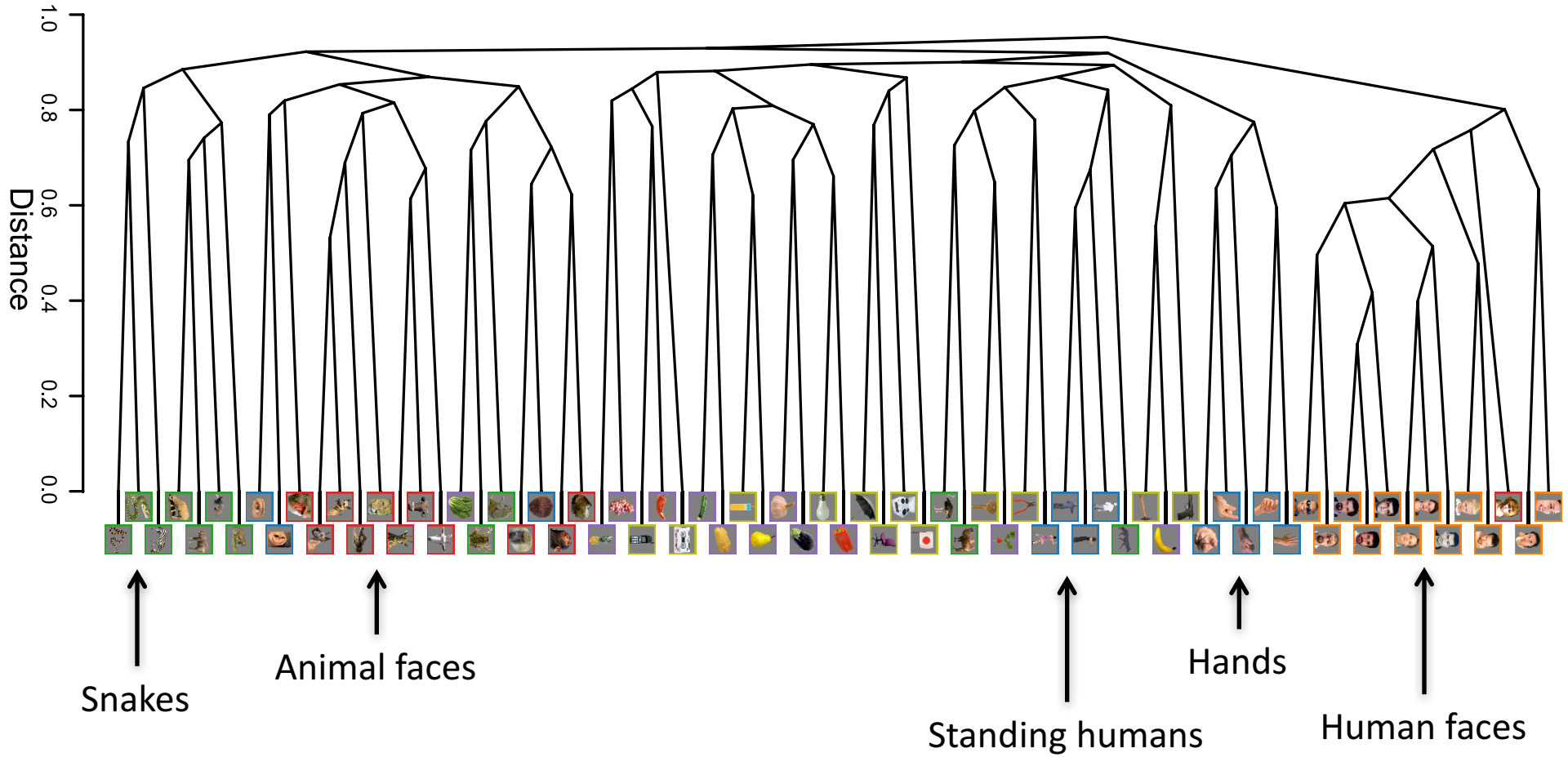




Kaneshiro et al. (2015). PLoS ONE.

RSA example: EEG classification (vision)

Can visualize hierarchical structure of confusion matrix using dendrograms.



Kaneshiro et al. (2015). PLoS ONE.

RSA example: Chord progressions

- Most RSA studies to date have involved vision
- Can we use RSA to gain insights into the structure of musical stimuli?
- Returning to chord progressions
  - 4 cadential events (tonic, dominant, flat II, silence)
  - 3 musical keys (C Major, B Major, F Major)
  - 2 classification methods (PCA/LDA, CNN/SCE)
  - Classification by chord function was more successful than classification by key

RSA example: Chord progressions

Let's visualize results from the chord progression classifications. Recall that classification by chord function was more successful than classification by key.

Procedure[1]

1. Convert confusion matrices to estimated conditional probability matrices by scaling each row to sum to 1
2. Scale the matrix by self-similarity of each class by dividing each row by the value along the diagonal.
3. Symmetrize the matrix – here we take the geometric mean of the matrix and its transpose.
4. Now we have a set of similarities S for all pairs of classes. Let's convert to distance by computing D = 1-S
5. Use publicly available R implementation[2] to visualize conditional probability matrices.

[1] Kaneshiro et al. (2015). PLoS One. [2] Kim (2015). GitHub.

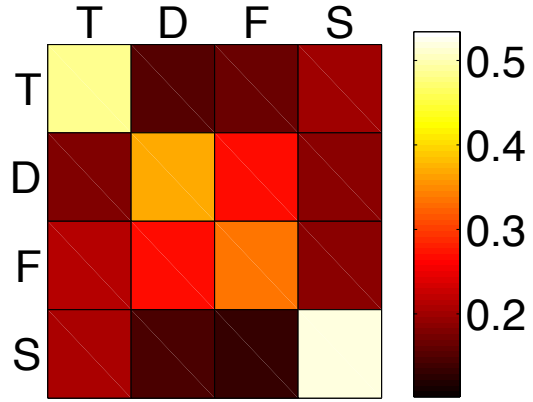
Results: 4-class classification

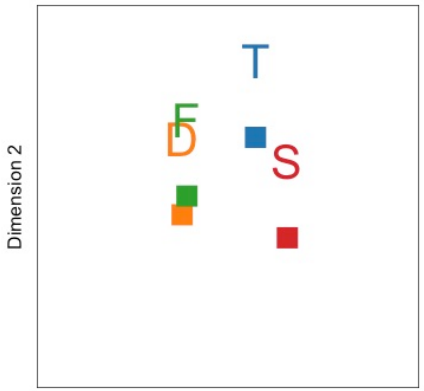
Confusion matrix (normalized)

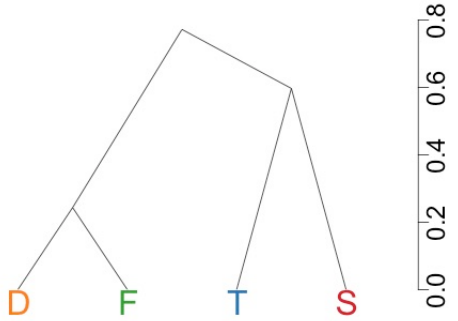
MDS

Dendrogram

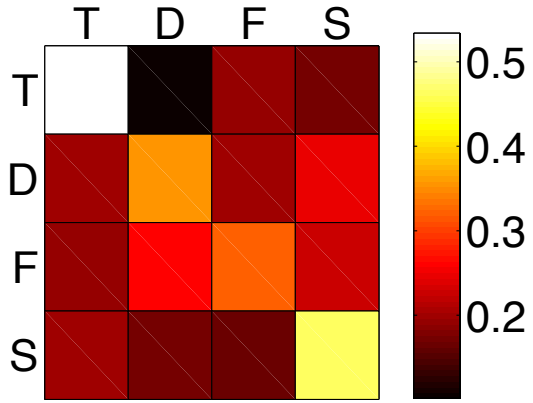
Method 1

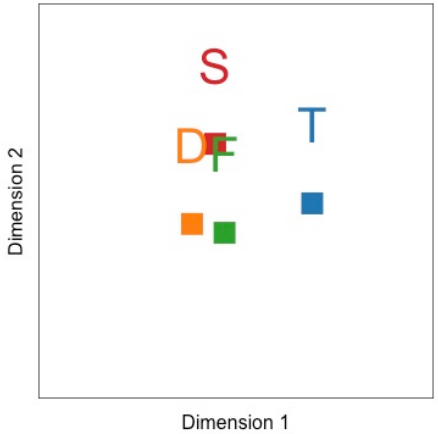


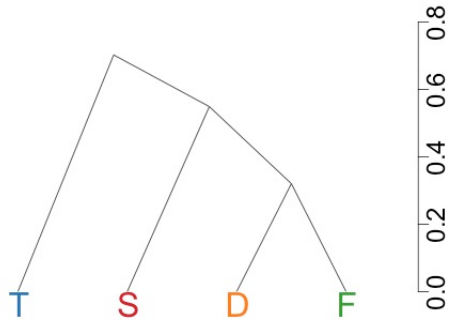




Method 2







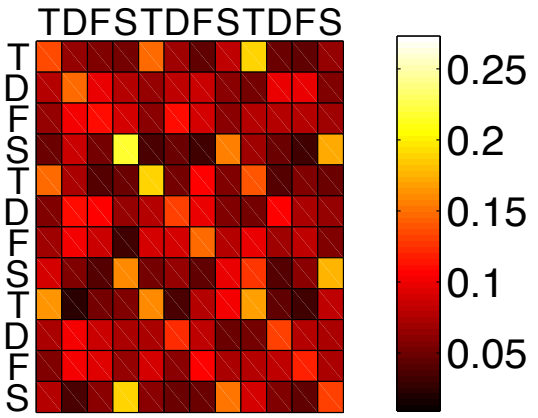
Results: 12-class classification

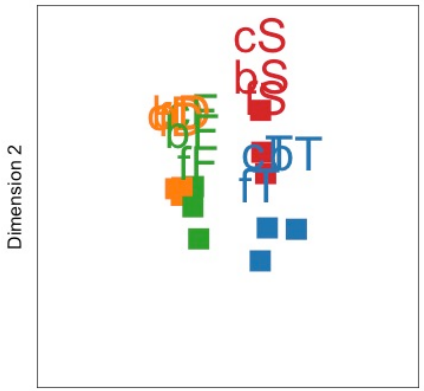
Confusion matrix (normalized)

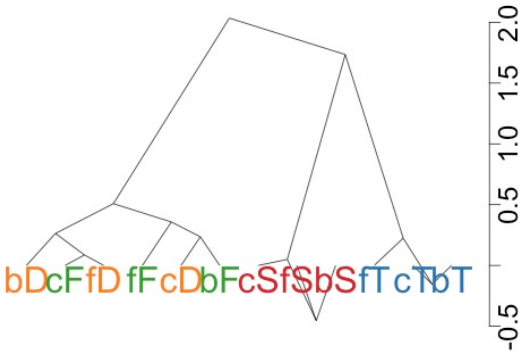
MDS

Dendrogram

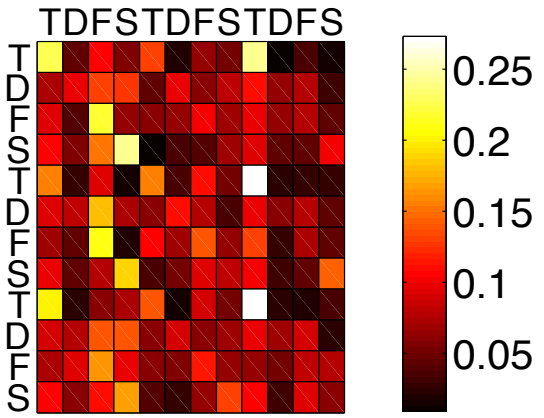
Method 1

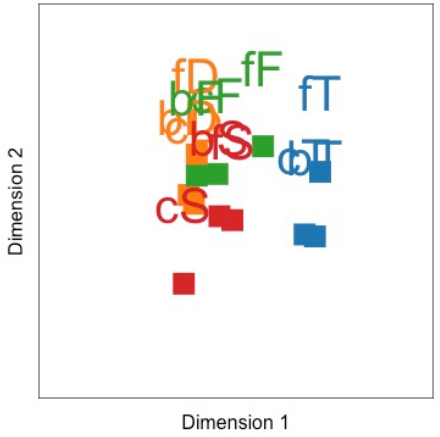


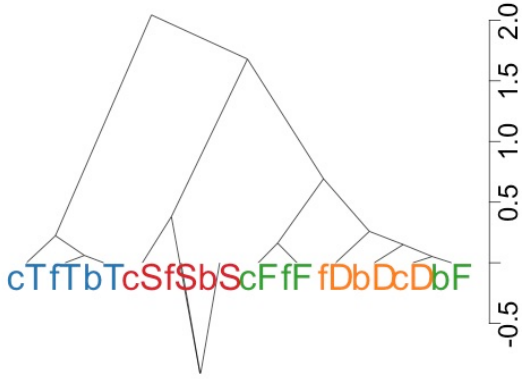




Method 2







Conclusions: RSA

- Classification confusion matrices can be used to assess relationships between the classes, as represented in the brain response
- With chord progressions, highly expected and highly unexpected stimuli produced the most distinct (easy to classify) brain responses
- Mildly unexpected stimuli produced brain responses that were more difficult for the classifier to differentiate
- More in-depth RSA results from this data to be presented at CogMIR on Friday!

Practical considerations: RSA

- RSA will be more meaningful if stimulus set has some structure
  - Vision experiment: 6 image categories with hierarchical category structure
  - Chord progressions: 4 musical functions x 3 musical keys
- There are several ways to compute RDMs from brain data
  - Pairwise correlations (fMRI)
  - Pairwise classification rates
  - Multi-class confusion matrix (provides self-similarity measure)
  - Sample-wise voltage differences in averaged ERPs[1]

[1] Groen et al. (2012). Frontiers in Computational Neuroscience.

Conclusions: Classification and RSA

- Classification is a **multivariate** approach to data analysis – can make use of high-dimensional response data (e.g., dense-array EEG)
- RSA approach can handle large stimulus sets
- RSA approach allows us to look beyond classifier accuracy to assess the **structure** of the stimuli
- RSA allows brain responses to be directly compared with other response modalities (e.g., behavioral ratings, behavioral discrimination, computational models of stimuli)