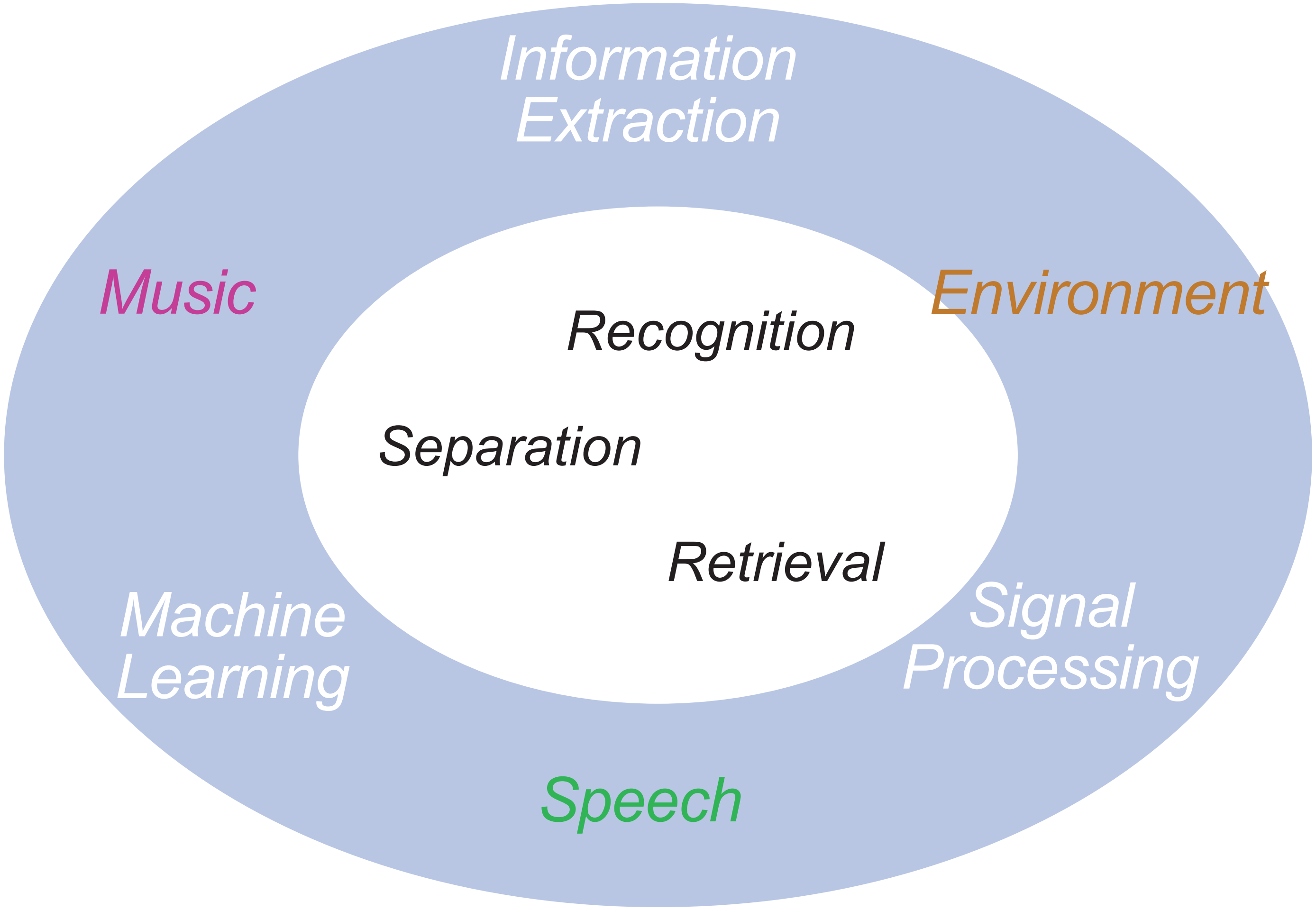
Extracting Information from Music Audio

Dan Ellis
Laboratory for Recognition and Organization of Speech and Audio
Dept. Electrical Engineering, Columbia University, NY USA
http://labrosa.ee.columbia.edu/

1. Motivation: Learning Music
2. Notes Extraction
3. Drum Pattern Modeling
4. Music Similarity

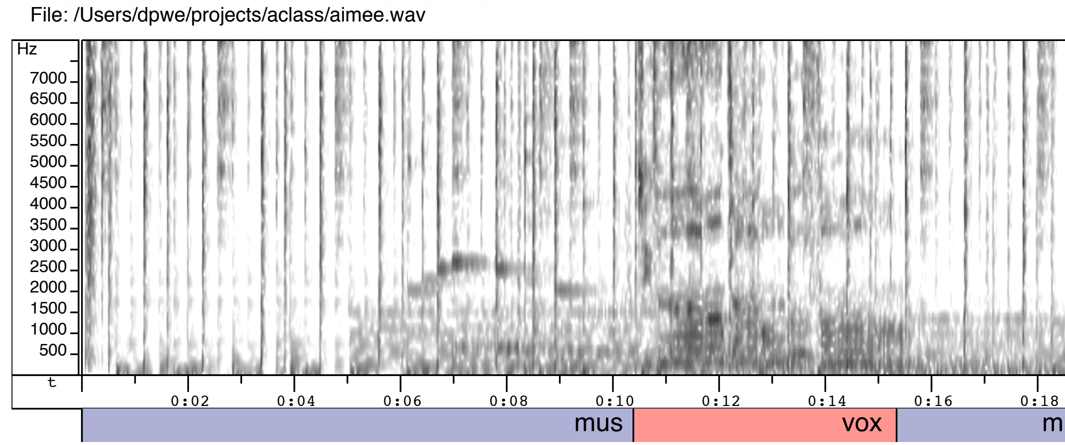
LabROSA Overview



1. Learning from Music

- A lot of music data available
  - e.g. 60G of MP3
    ≈ 1000 hr of audio, 15k tracks
- What can we do with it?
  - implicit definition of 'music'
- Quality vs. quantity
  - Speech recognition lesson:
    10x data, 1/10th annotation, twice as useful
- Motivating Applications
  - music similarity (recommendation, playlists)
  - computer (assisted) music generation
  - insight into music

Ground Truth Data

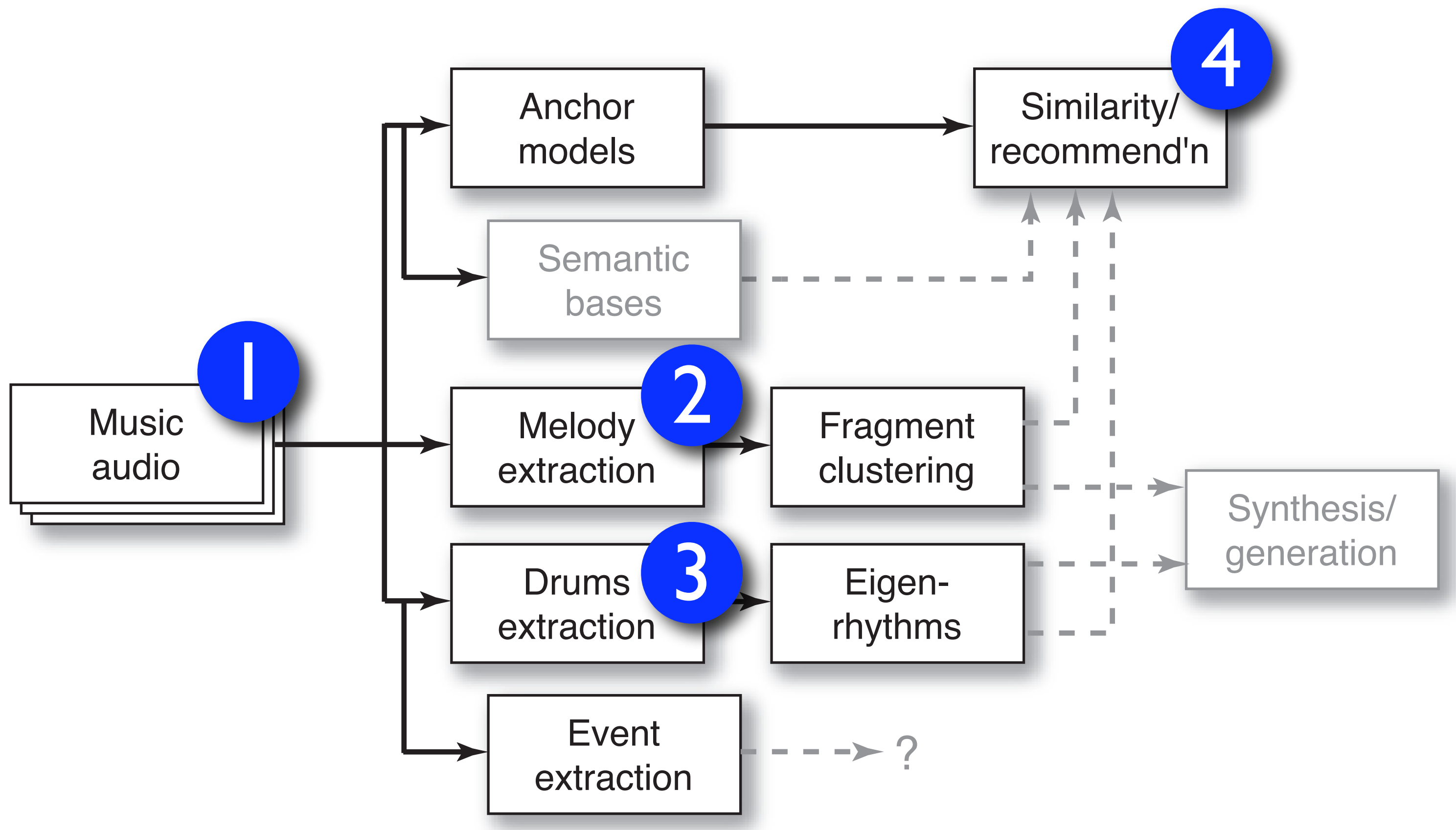


- A lot of **unlabeled** music data available
  - manual annotation is expensive and rare
- **Unsupervised structure discovery possible**
  - .. but labels help to indicate what you want
- **Weak annotation sources**
  - artist-level descriptions
  - symbol sequences without timing (MIDI)
  - errorful transcripts
- **Evaluation requires ground truth**
  - limiting factor in Music IR evaluations?

Lab ROSA
Laboratory for the Recognition and Organization of Speech and Audio

Columbia University in the City of New York

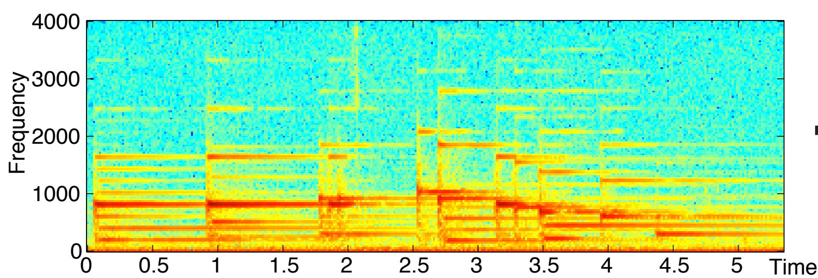
Talk Roadmap

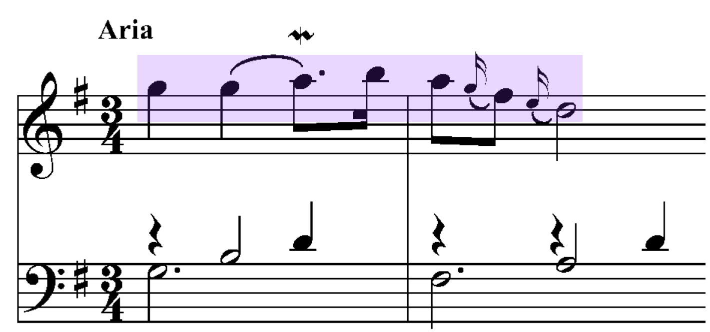


2. Notes Extraction

with Graham Poliner

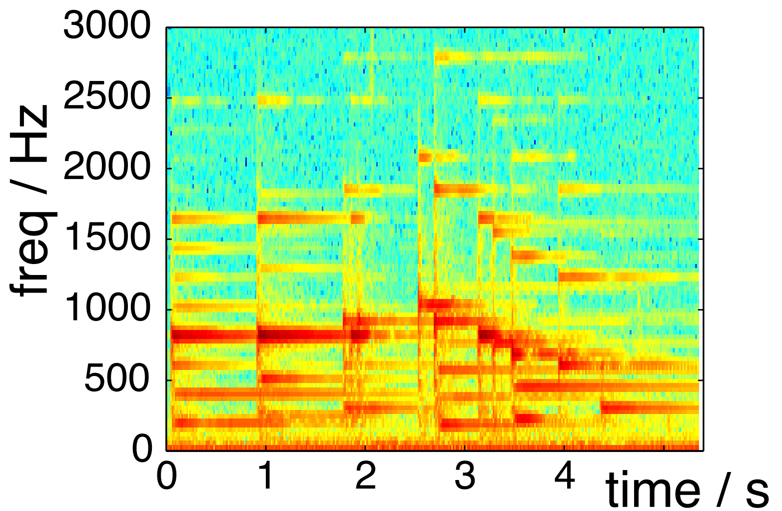
- Audio → Score very desirable
  - for data compression, searching, learning
- Full solution is elusive
  - signal separation of overlapping voices
  - music constructed to frustrate!
- Maybe simplify problem:
  "Dominant Melody" at each time frame

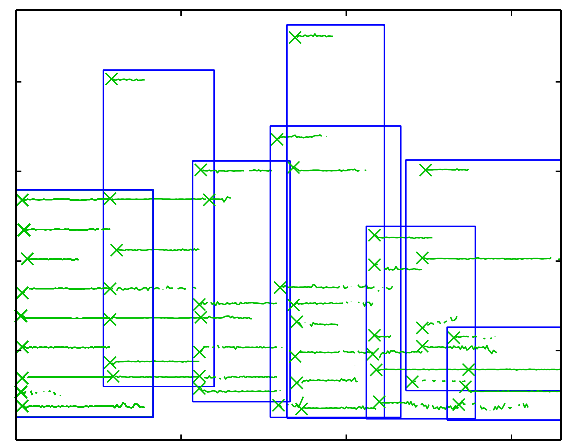




Conventional Transcription

- Pitched notes have harmonic spectra
  → transcribe by searching for harmonics
  - e.g. sinusoid modeling + grouping

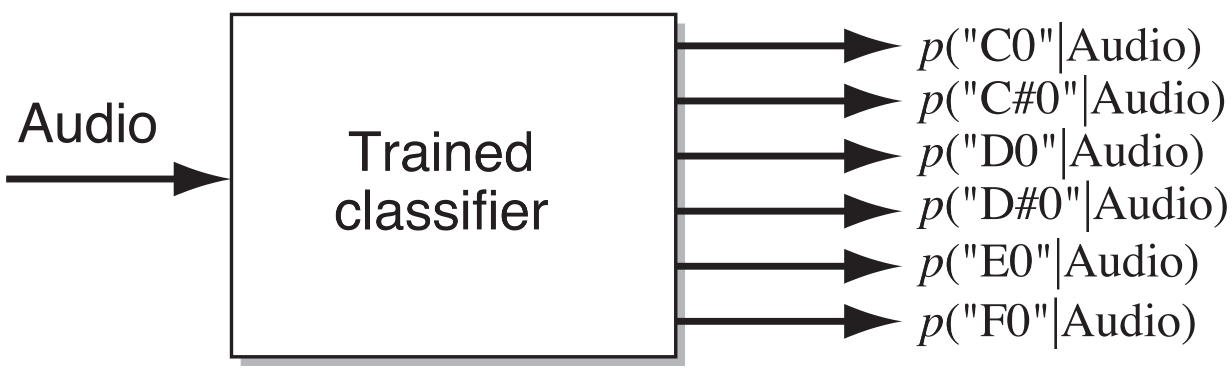




- Explicit expert-derived knowledge

Transcription as Classification

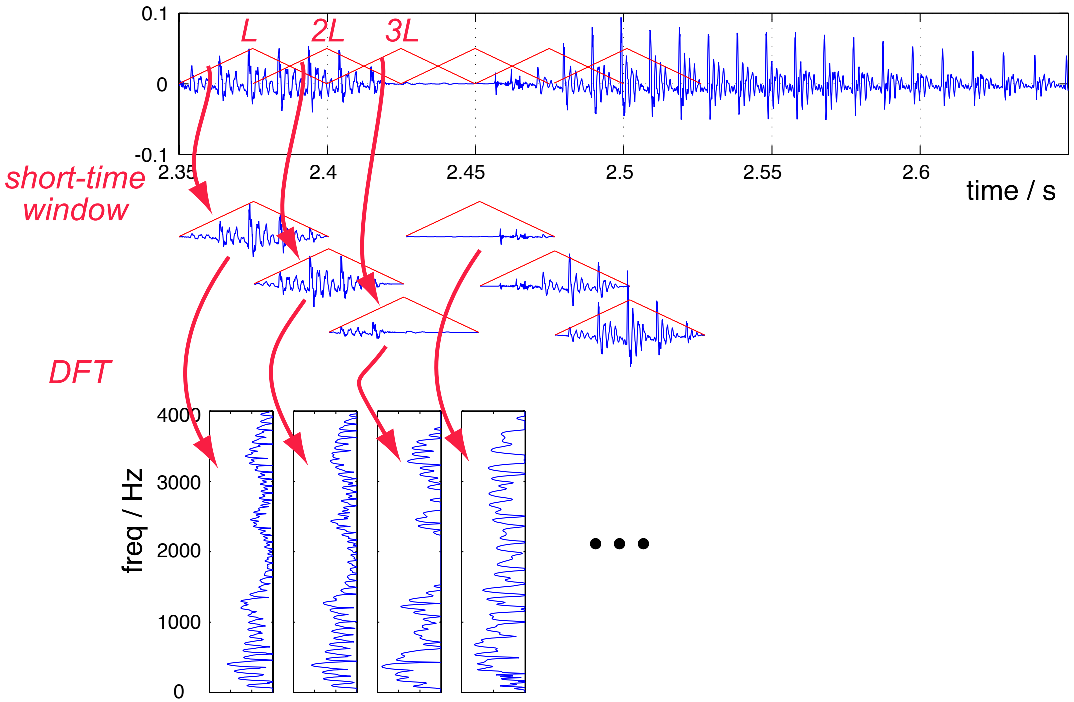
- Signal models typically used for transcription
  - harmonic spectrum, superposition
- But ... trade domain knowledge for data
  - transcription as pure classification problem:

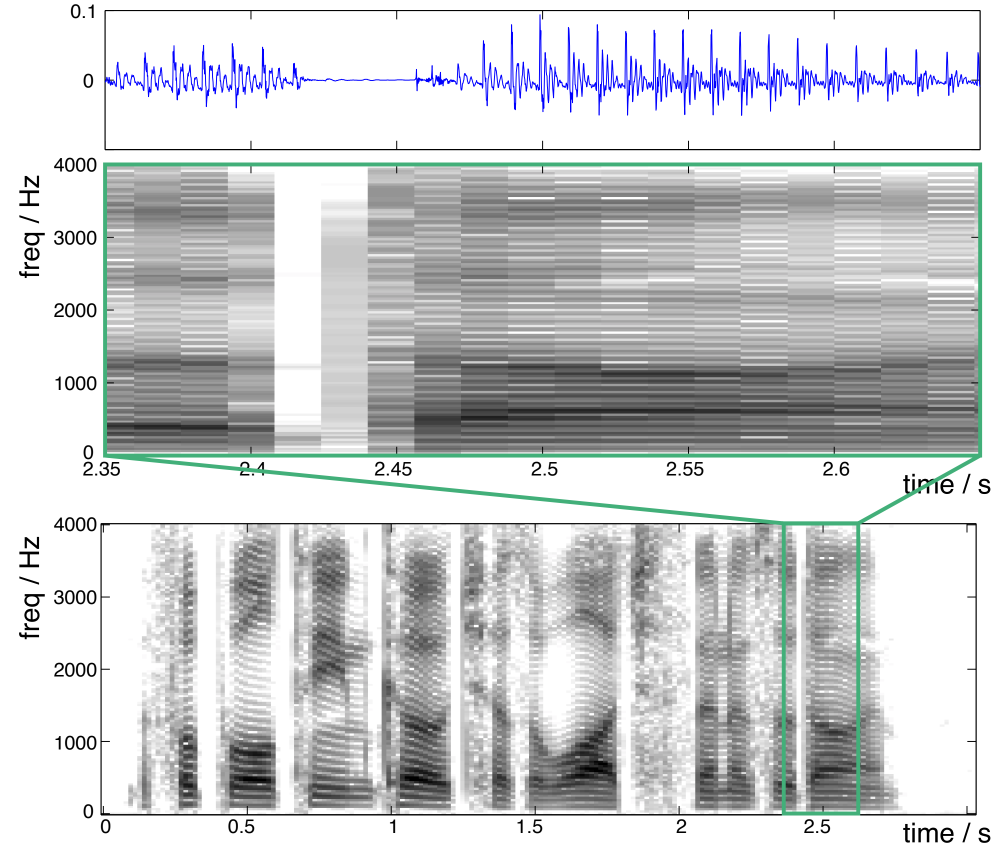


  - single N-way discrimination for "melody"
  - per-note classifiers for polyphonic transcription

Melody Transcription Features

- Short-time Fourier Transform Magnitude (Spectrogram)

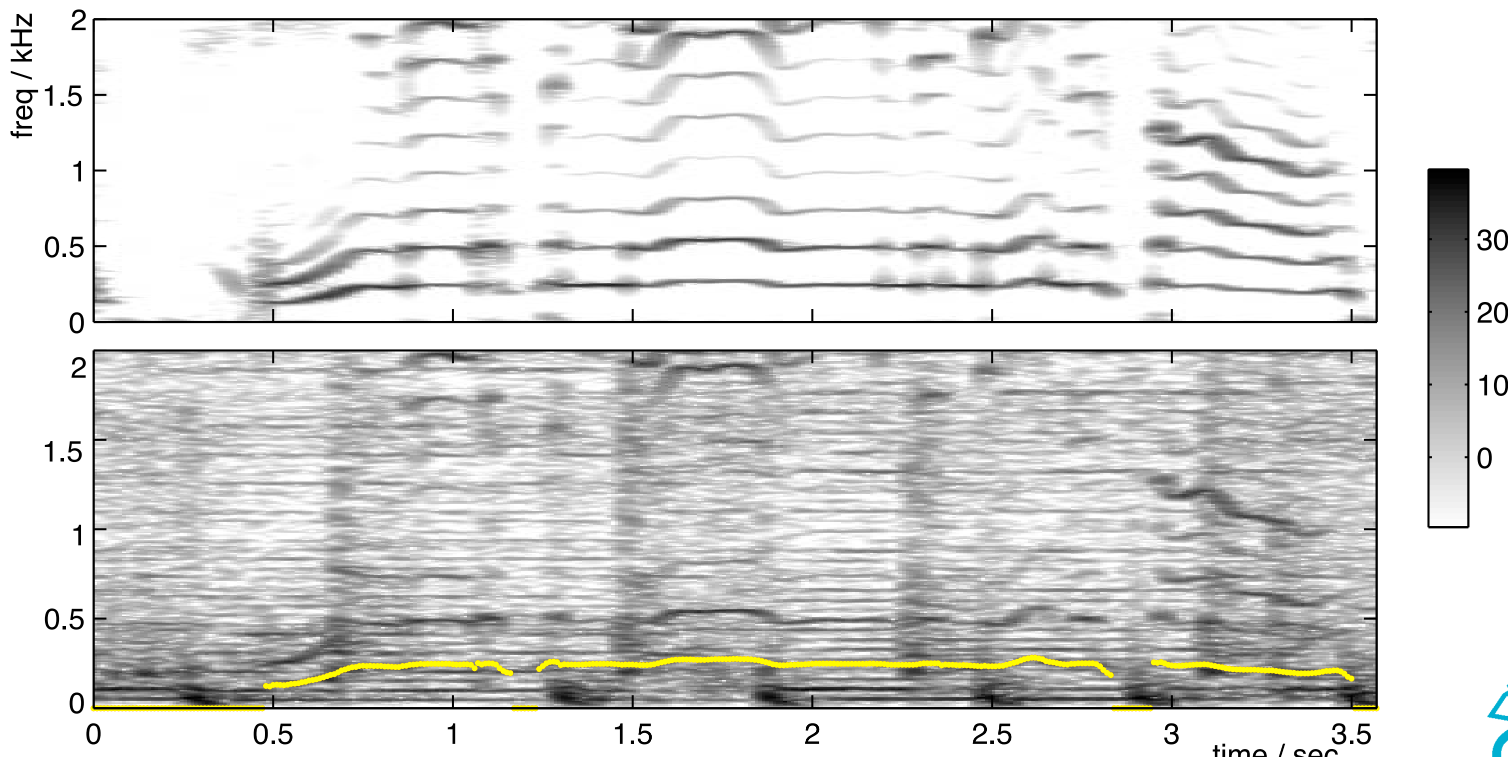




- Standardize over 50 pt frequency window

Training Data

- Need {data, label} pairs for classifier training
- Sources:
  - pre-mixing multitrack recordings + hand-labeling?
  - synthetic music (MIDI) + forced-alignment?



Melody Transcription Results

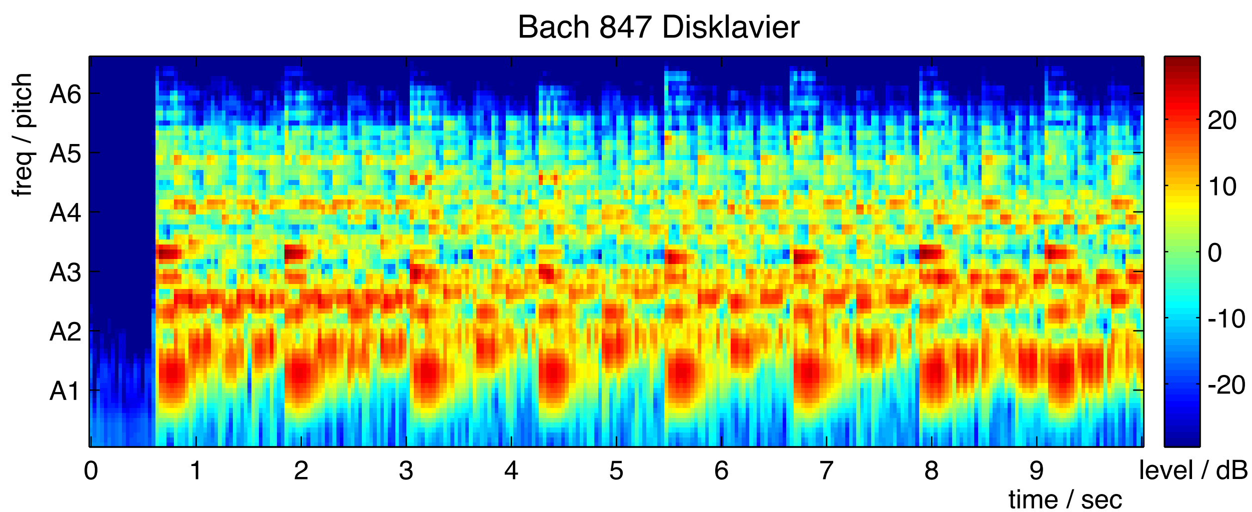
- Trained on 17 examples
  - .. plus transpositions out to +/- 6 semitones
  - All-pairs SVMs (Weka)
- Tested on ISMIR MIREX 2005 set
  - includes foreground/background detection

Rank	Participant	Overall Accuracy	Voicing d'	Raw Pitch	Raw Chroma	Runtime / s
1	Dressler	**71.4%**	**1.85**	68.1%	71.4%	32
2	Ryynänen	64.3%	1.56	**68.6%**	**74.1%**	10970
3	Poliner	61.1%	1.56	67.3%	73.4%	5471
3	Paiva 2	61.1%	1.22	58.5%	62.0%	45618
5	Marolt	59.5%	1.06	60.1%	67.1%	12461
6	Paiva 1	57.8%	0.83	62.7%	66.7%	44312
7	Goto	49.9%*	0.59*	65.8%	71.8%	211
8	Vincent 1	47.9%*	0.23*	59.8%	67.6%	?
9	Vincent 2	46.4%*	0.86*	59.6%	71.1%	251
10	Brossier	3.2%* †	0.14 * †	3.9% †	8.1% †	41

- Example...

Polyphonic Transcription

- Train SVM detectors for every piano note
  - same features & classifier but different labels
  - 88 separate detectors, independent smoothing
- Use MIDI syntheses, player piano recordings



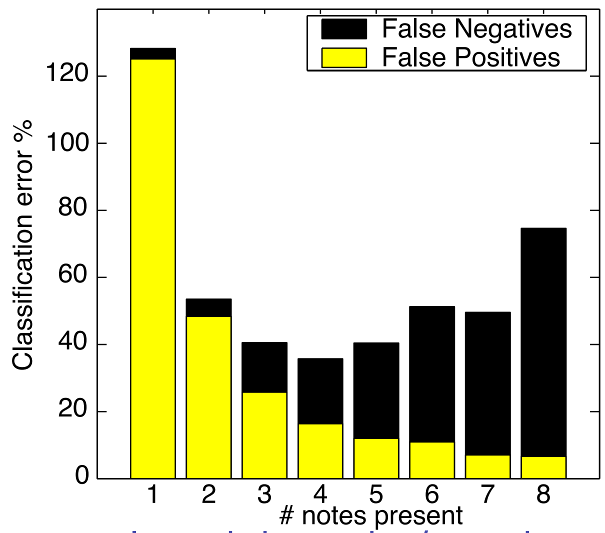
  - about 30 min training data

Piano Transcription Results

- Significant improvement from classifier:
  - frame-level accuracy results:

Algorithm	Errs	False Pos	False Neg	d'
SVM	43.3%	27.9%	15.4%	3.44
Klapuri&Ryynänen	66.6%	28.1%	38.5%	2.71
Marolt	84.6%	36.5%	48.1%	2.35

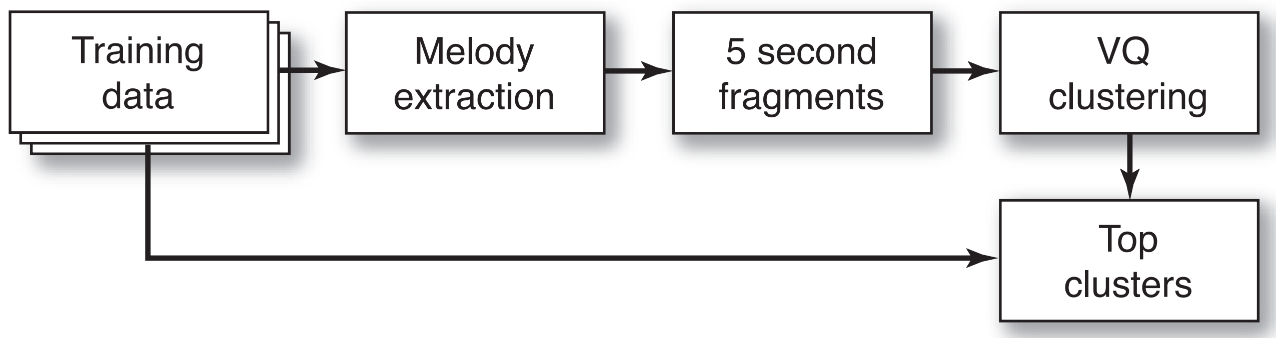
  - Breakdown by frame type:



  - http://labrosa.ee.columbia.edu/projects/melody/

Melody Clustering

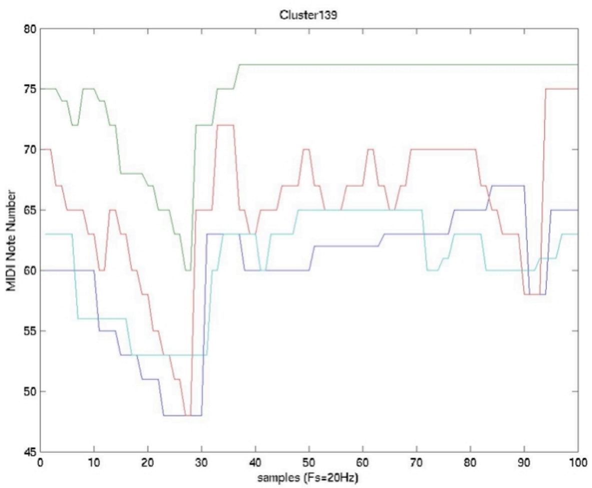
- Goal: Find 'fragments' that recur in melodies
  - .. across large music database
  - .. trade data for model sophistication

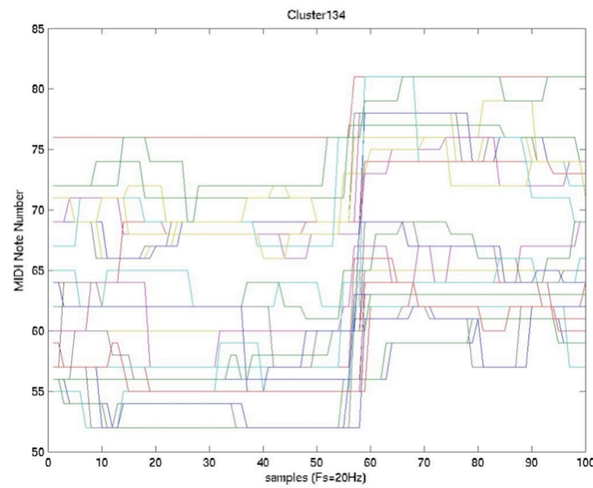


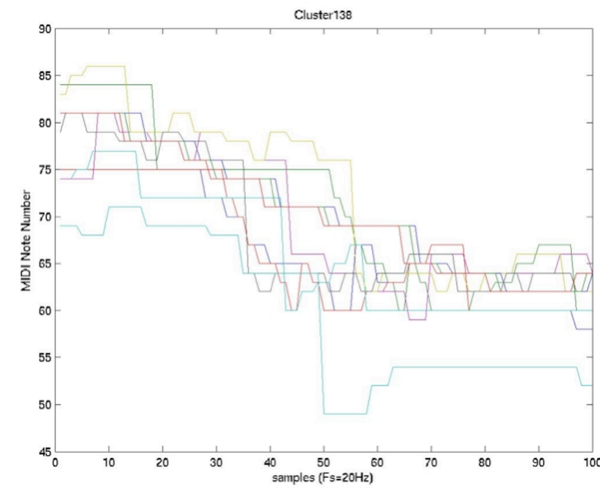
- Data sources
  - pitch tracker, or MIDI training data
- Melody fragment representation
  - DCT(1:20) - removes average, smoothes detail

Melody clustering results

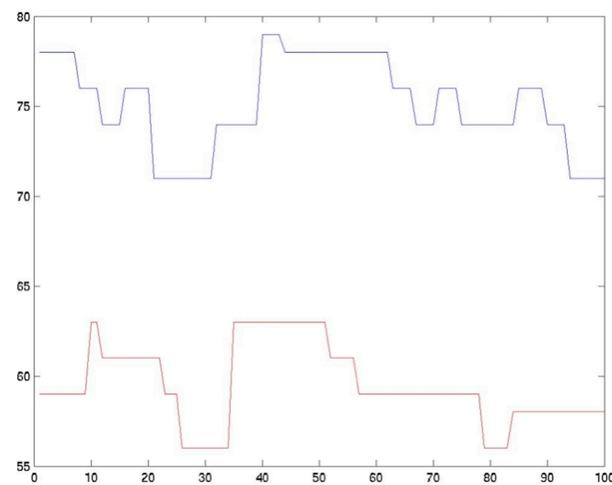
- Clusters match underlying contour:







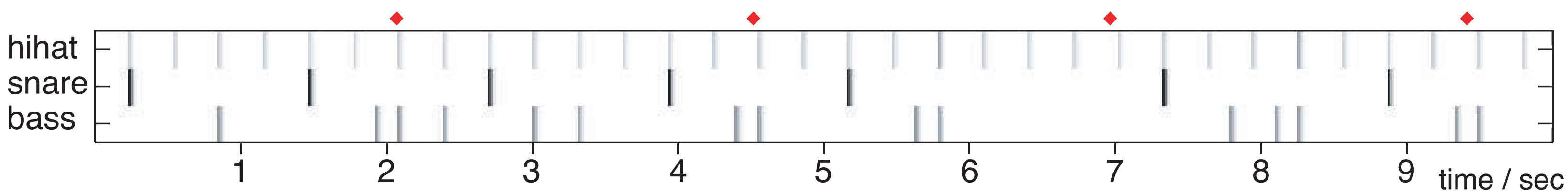
- Some interesting matches:
    - e.g. Pink + Nsync



3. Eigenrhythms: Drum Pattern Space

with John Arroyo

- Pop songs built on repeating "drum loop"
  - variations on a few bass, snare, hi-hat patterns



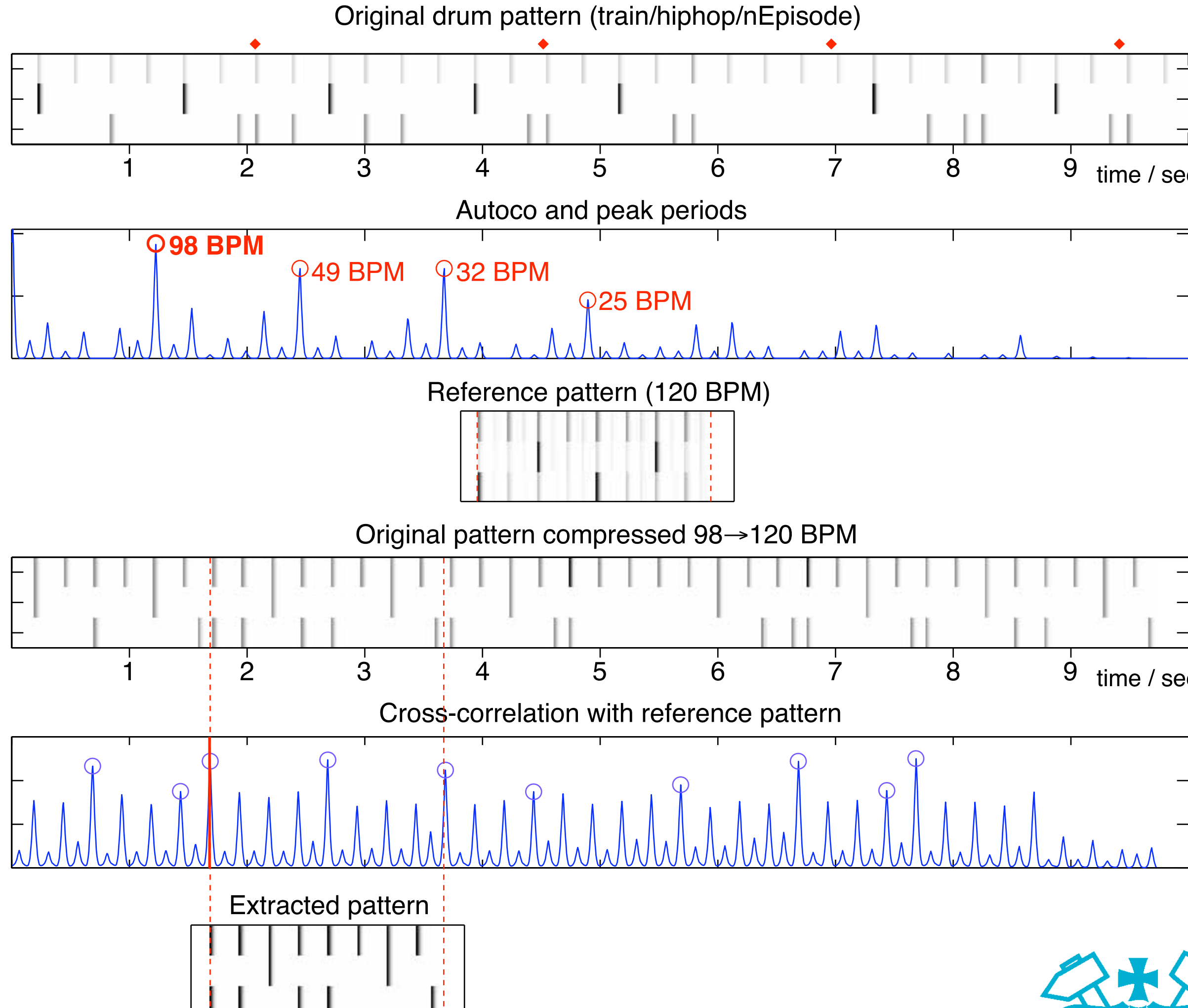
- Eigen-analysis (or ...) to capture variations?
  - by analyzing lots of (MIDI) data, or from audio
- Applications
  - music categorization
  - "beat box" synthesis
  - insight

Aligning the Data

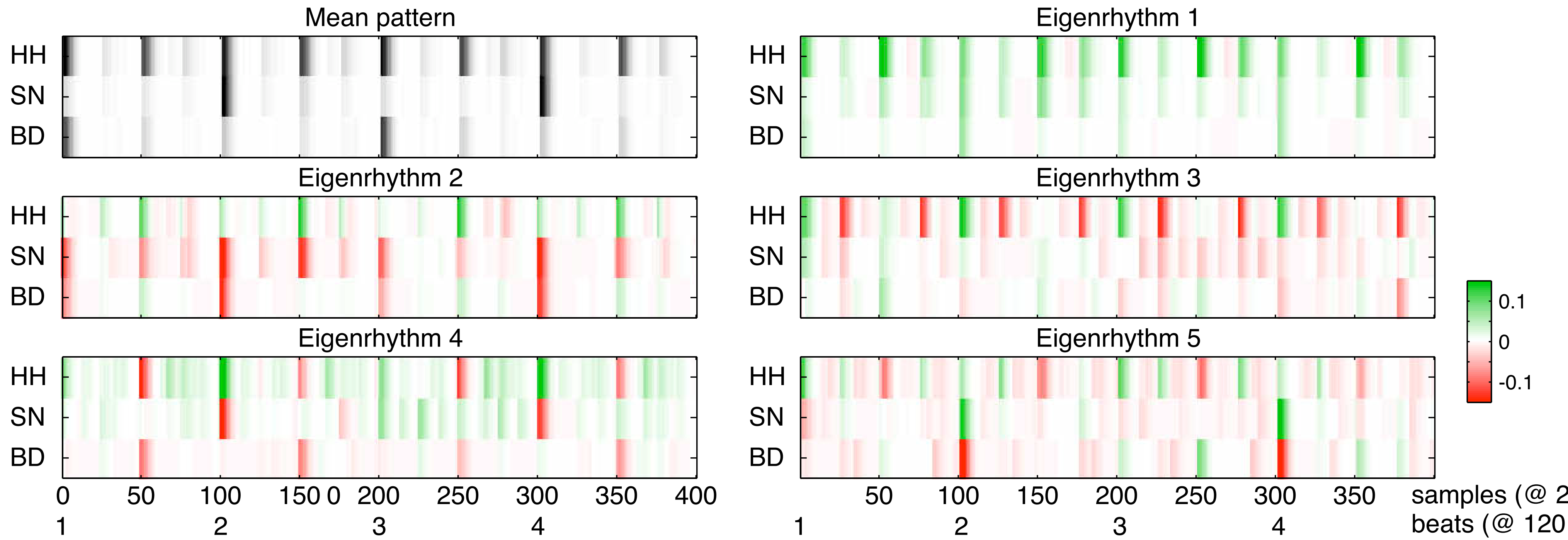
- Need to align patterns prior to modeling...

tempo (stretch):
by inferring BPM &
normalizing

downbeat (shift):
correlate against
'mean' template

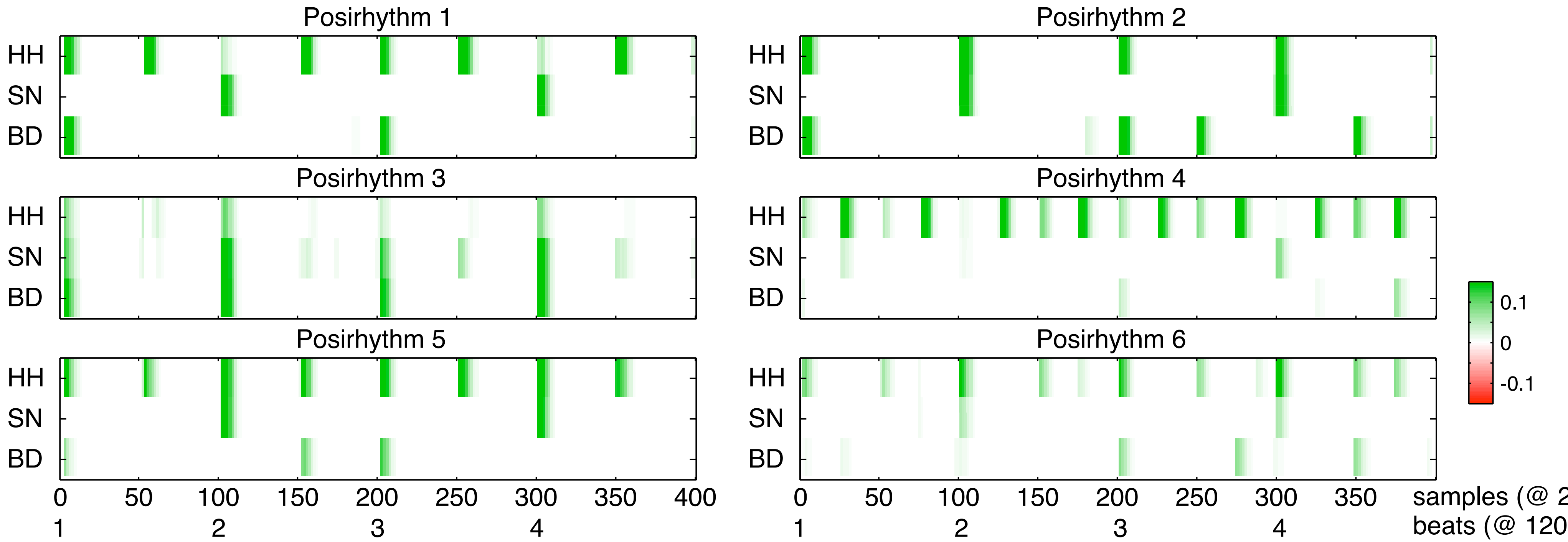


Eigenrhythms (PCA)



- Need 20+ Eigenvectors for good coverage of 100 training patterns (1200 dims)
- Eigenrhythms both add and subtract

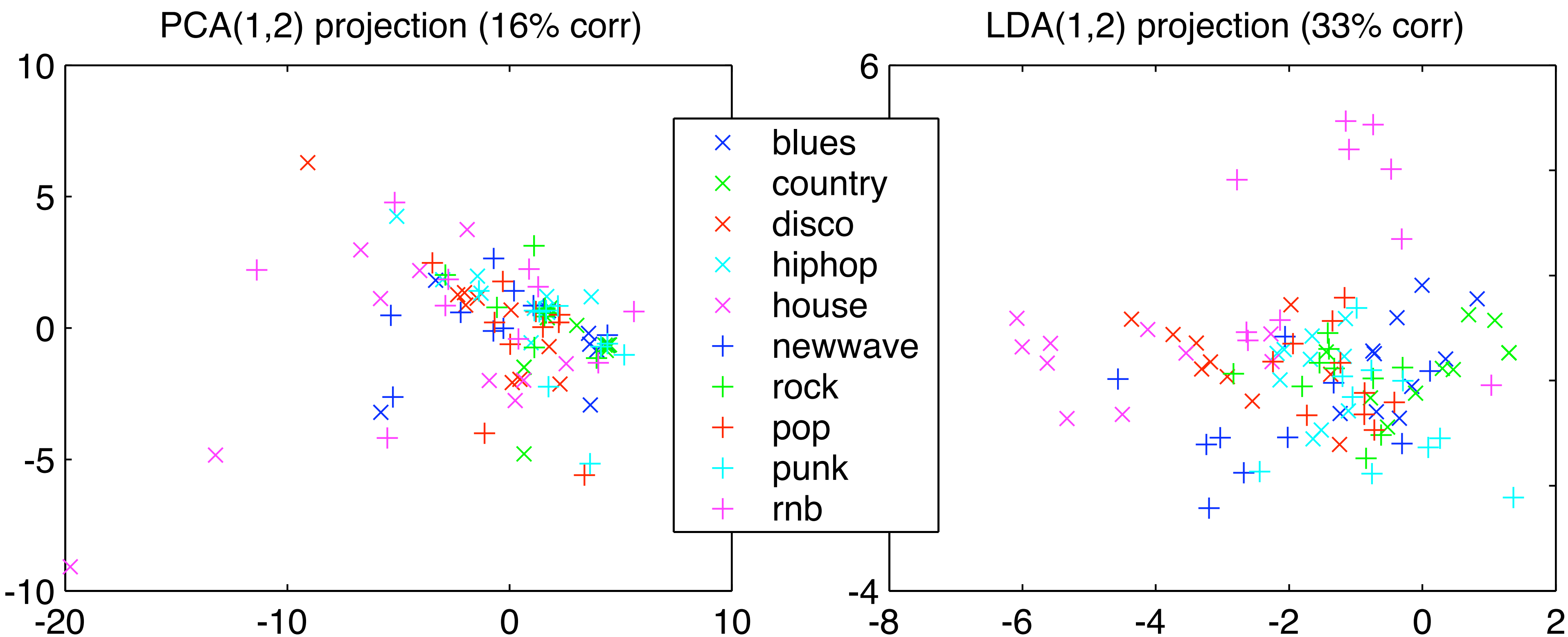
Posirhythms (NMF)



- Nonnegative: only adds beat-weight
- Capturing some structure

Eigenrhythms for Classification

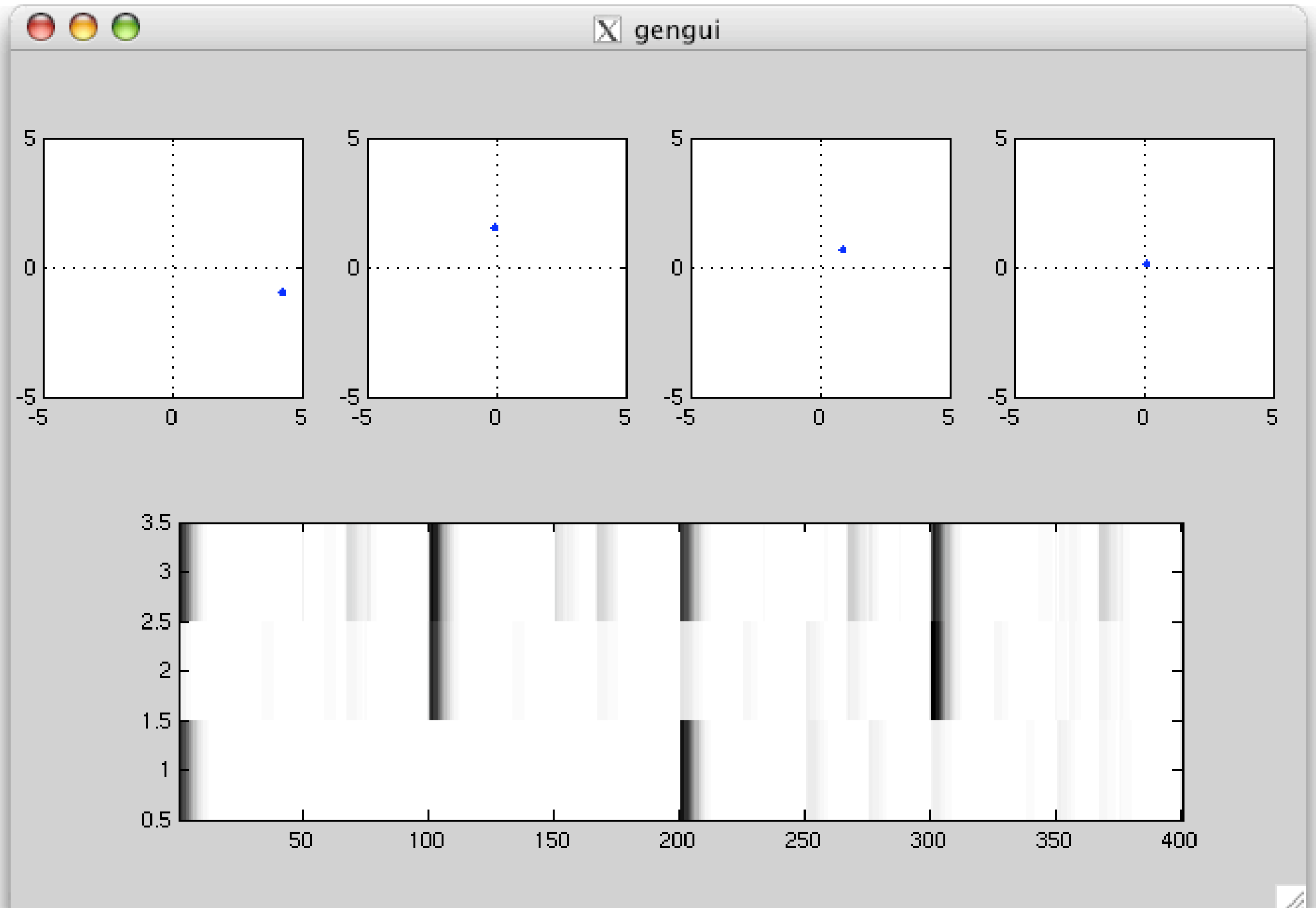
- Projections in Eigenspace / LDA space



- 10-way Genre classification (nearest nbr):
  - PCA3: 20% correct
  - LDA4: 36% correct

Eigenrhythm BeatBox

- Resynthesize rhythms from eigen-space



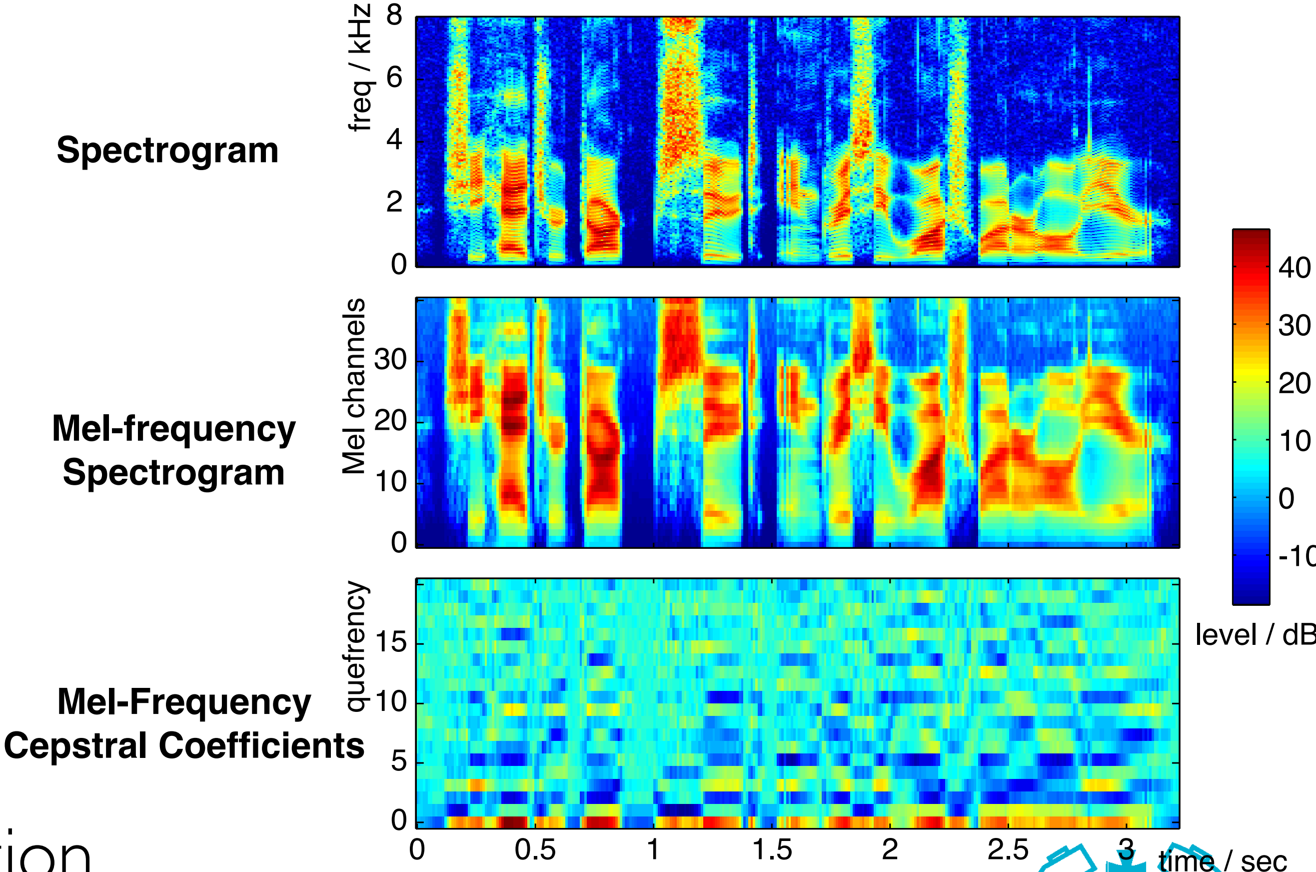
4. Music Similarity

with Mike Mandel and Adam Berenzweig

- Can we predict which songs "sound alike" to a listener?
  - .. based on the audio waveforms?
  - many aspects to subjective similarity
- Applications
  - query-by-example
  - automatic playlist generation
  - discovering new music
- Problems
  - the right representation
  - modeling individual similarity

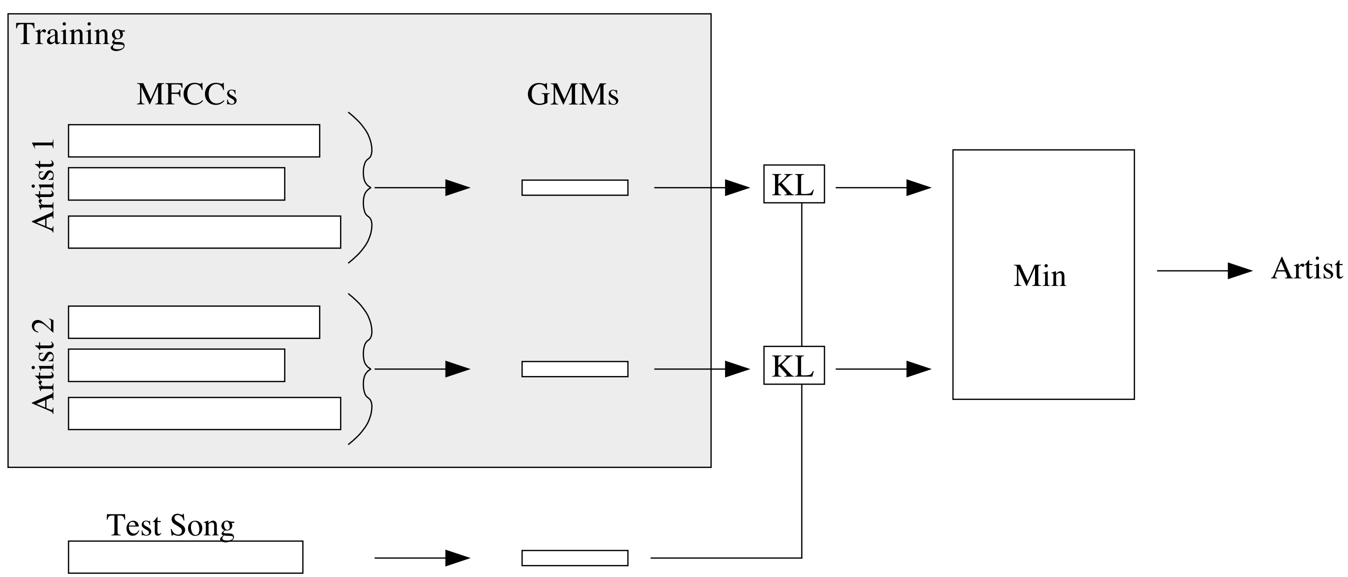
Music Similarity Features

- Need "timbral" features:
  Mel-Frequency Cepstral Coeffs (MFCCs)
  - auditory-like frequency warping
  - log-domain
  - discrete cosine transform orthogonalization



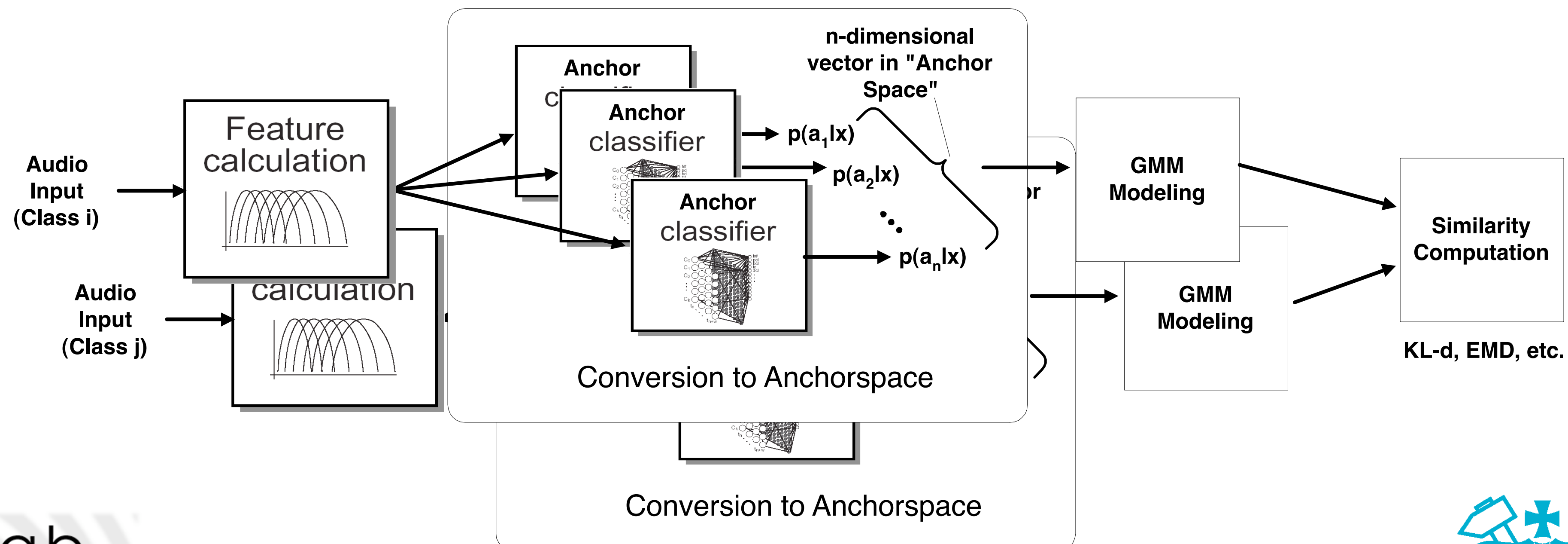
Timbral Music Similarity

- Measure similarity of feature distribution
  - i.e. collapse across time to get density $p(x_i)$
  - compare by e.g. KL divergence
- e.g. Artist Identification
  - learn artist model $p(x_i \mid artist\ X)$ (e.g. as GMM)
  - classify unknown song to closest model



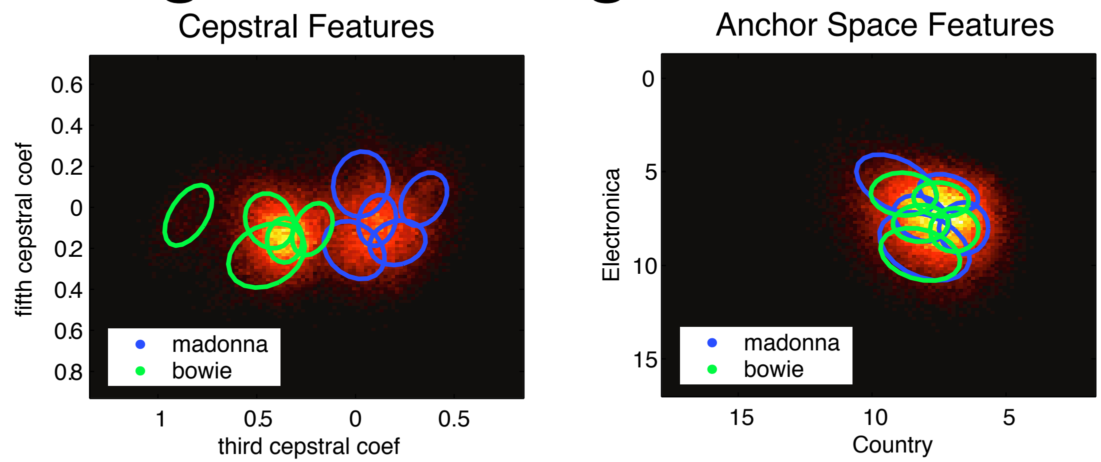
"Anchor Space"

- Acoustic features describe each song
  - .. but from a signal, not a perceptual, perspective
  - .. and not the differences between songs
- Use genre classifiers to define new space
  - prototype genres are "anchors"

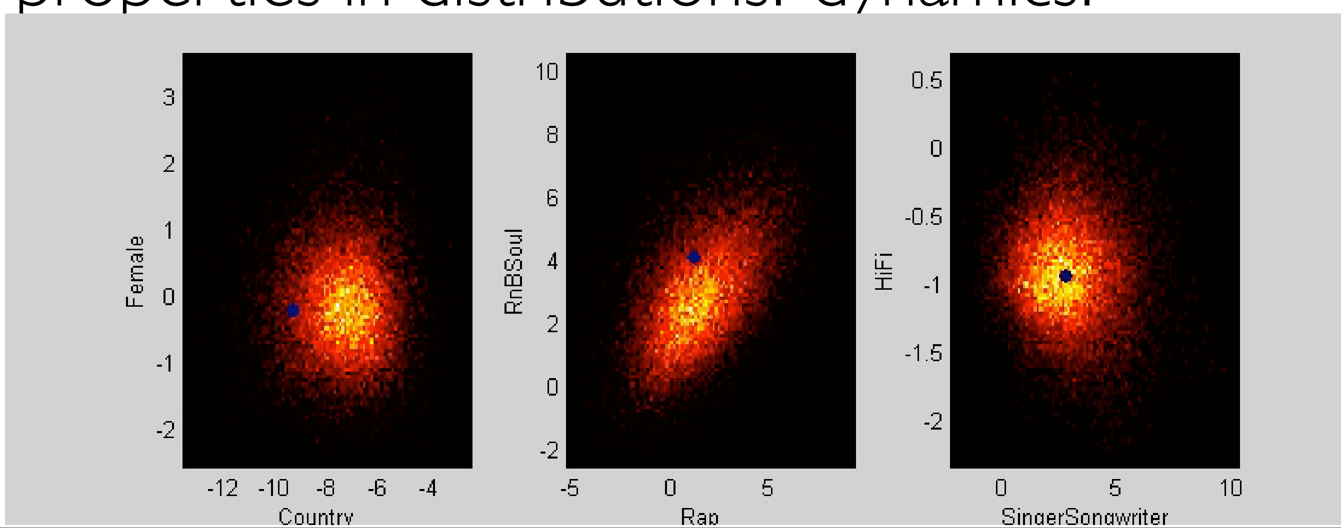


Anchor Space

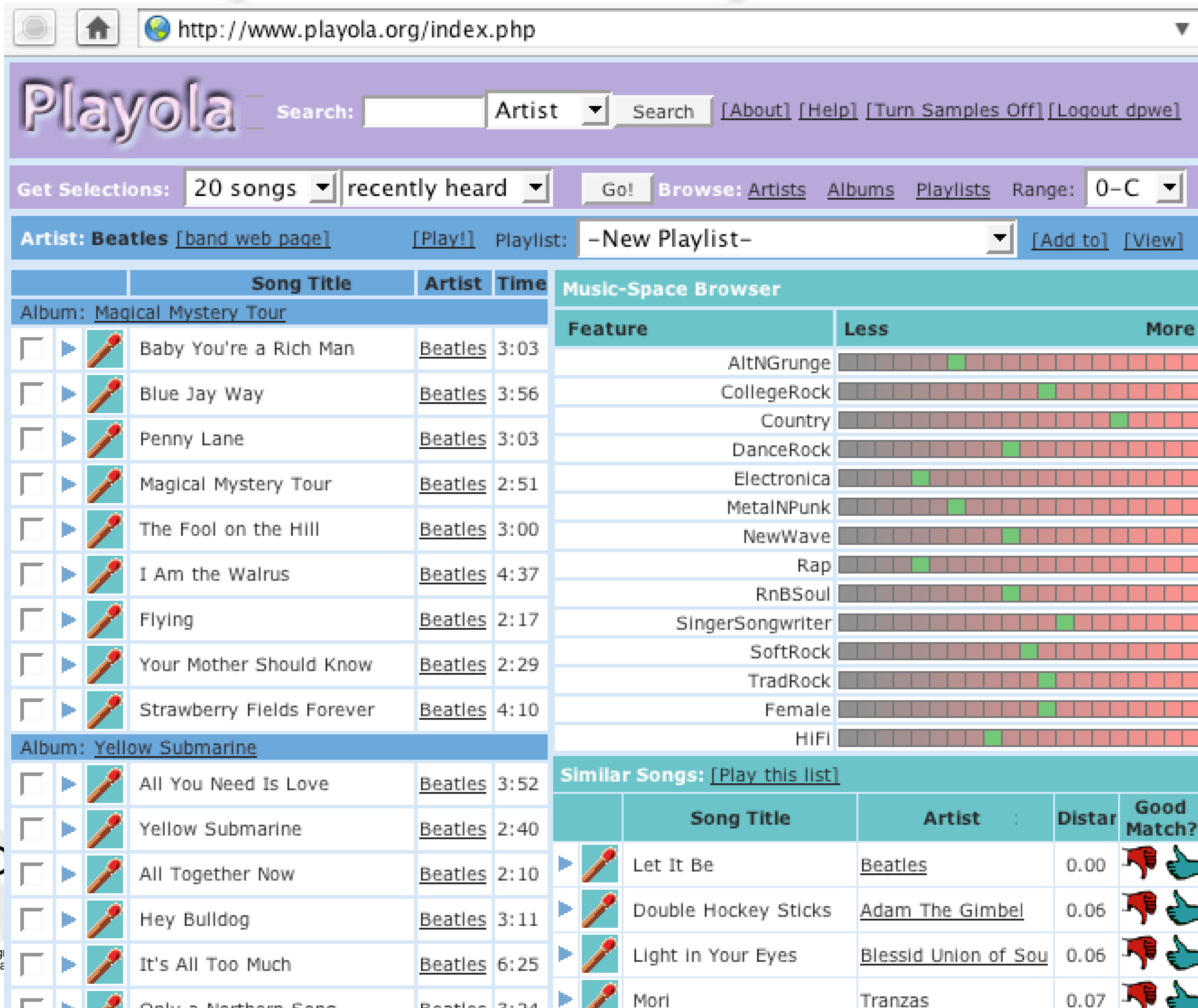
- Frame-by-frame high-level categorizations
  - compare to raw features?



  - properties in distributions? dynamics?



‘Playola’ Similarity Browser



Ground-truth data

- Hard to evaluate Playola's 'accuracy'
  - user tests...
  - ground truth?
- "Musicseer" online survey:
  - ran for 9 months in 2002
  - > 1,000 users, > 20k judgments
  - http://labrosa.ee.columbia.edu/projects/musicsim/

Which artist is most similar to:
Janet Jackson?

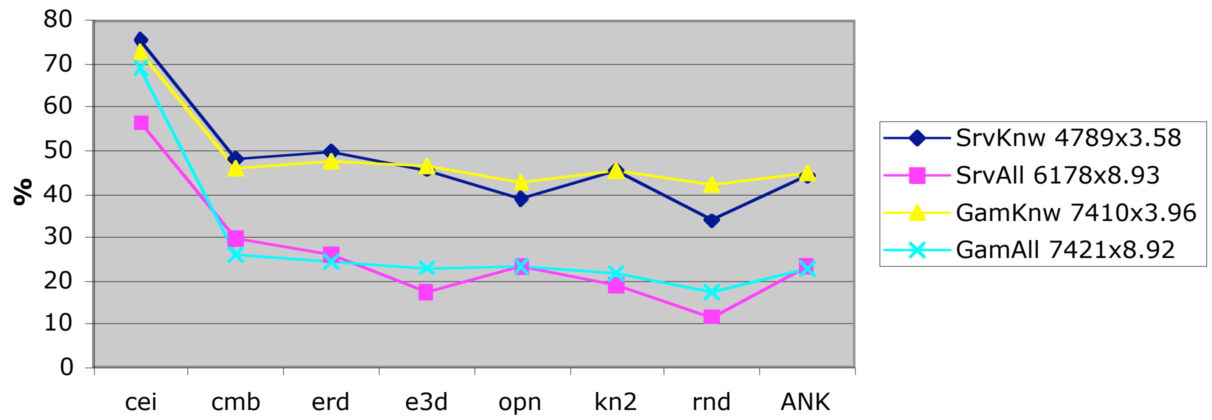
1. R. Kelly
2. Paula Abdul
3. Aaliyah
4. Milli Vanilli
5. En Vogue
6. Kansas
7. Garbage
8. Pink
9. Christina Aguilera

Evaluation

- Compare Classifier measures against Musicseer subjective results
  - "triplet" agreement percentage
  - Top-N ranking agreement score: $s_i = \sum_{r=1}^{N} \alpha_r^r \alpha_c^{k_r}$
  - "Average Dynamic Recall" ?(Typke et al.)
  - First-place agreement percentage
    - simple significance test

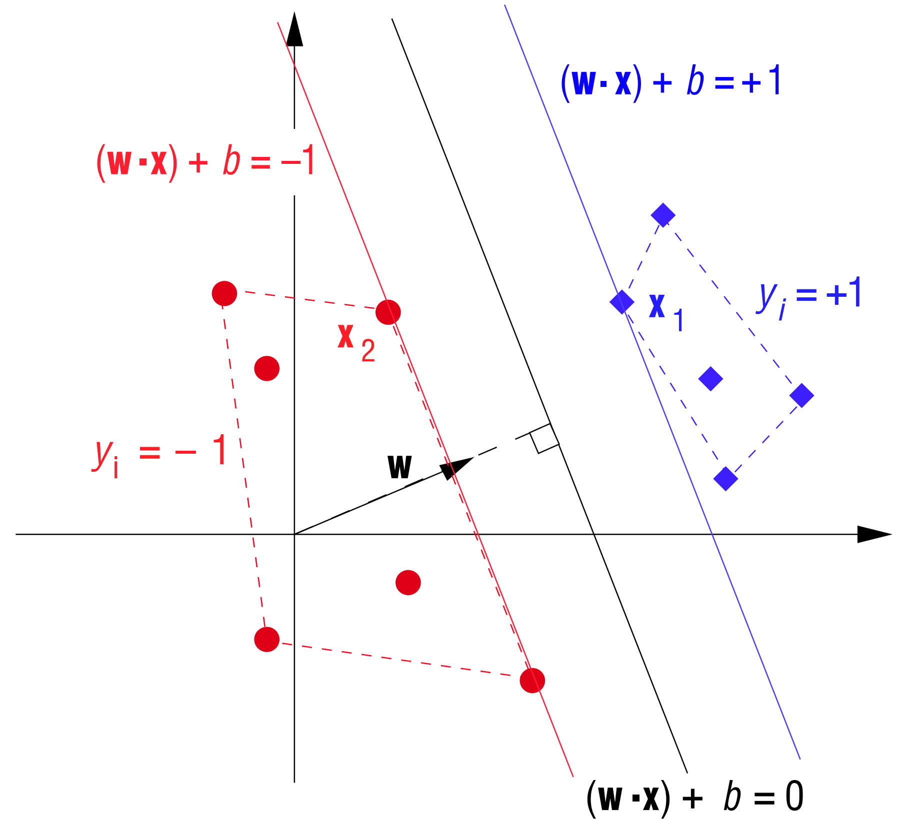
$$\alpha_r = \left(\frac{1}{2}\right)^{\frac{1}{3}}$$

$$\alpha_c = \alpha_r^2$$



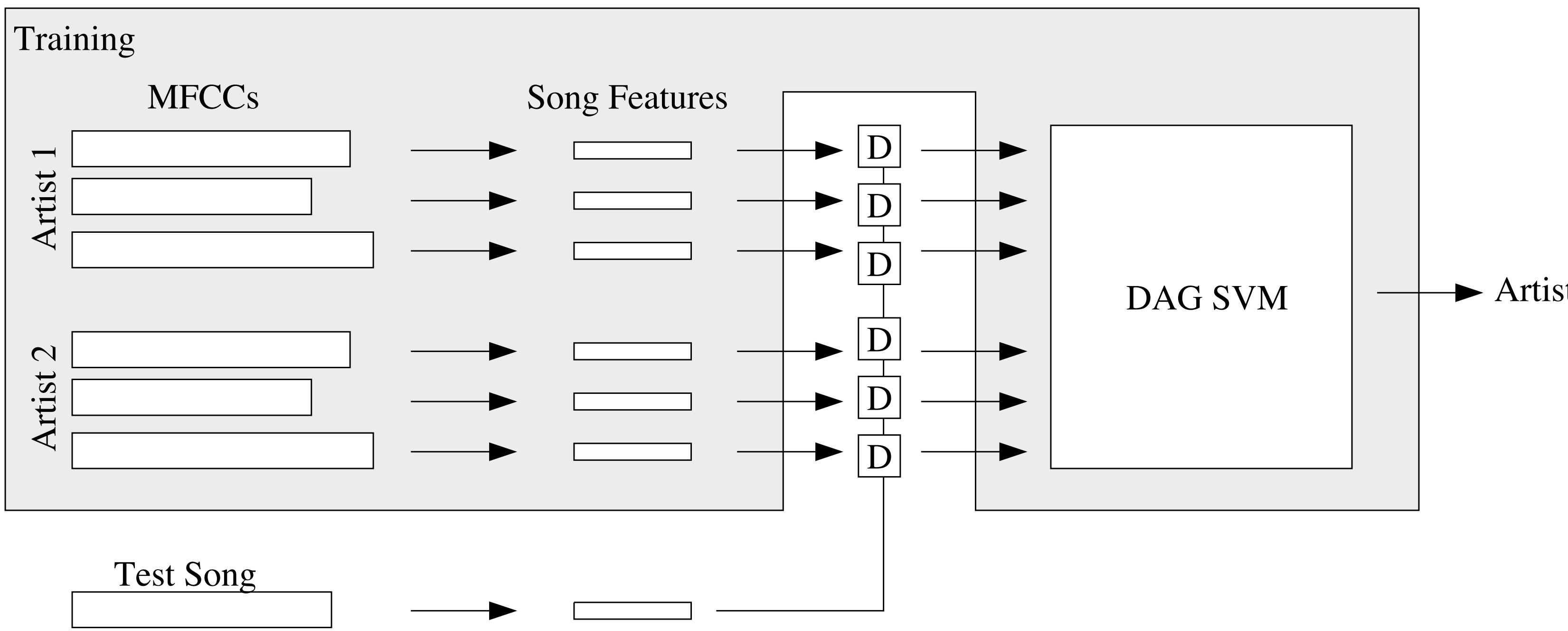
Using SVMs for Artist ID

- Support Vector Machines (SVMs) find hyperplanes in a high-dimensional space
  - relies only on matrix of distances between points
  - much 'smarter' than nearest-neighbor/overlap
  - want diversity of reference vectors...



Song-Level SVM Artist ID

- Instead of one model per artist/genre, use *every* training song as an 'anchor'
  - then SVM finds best support for each artist



Artist ID Results

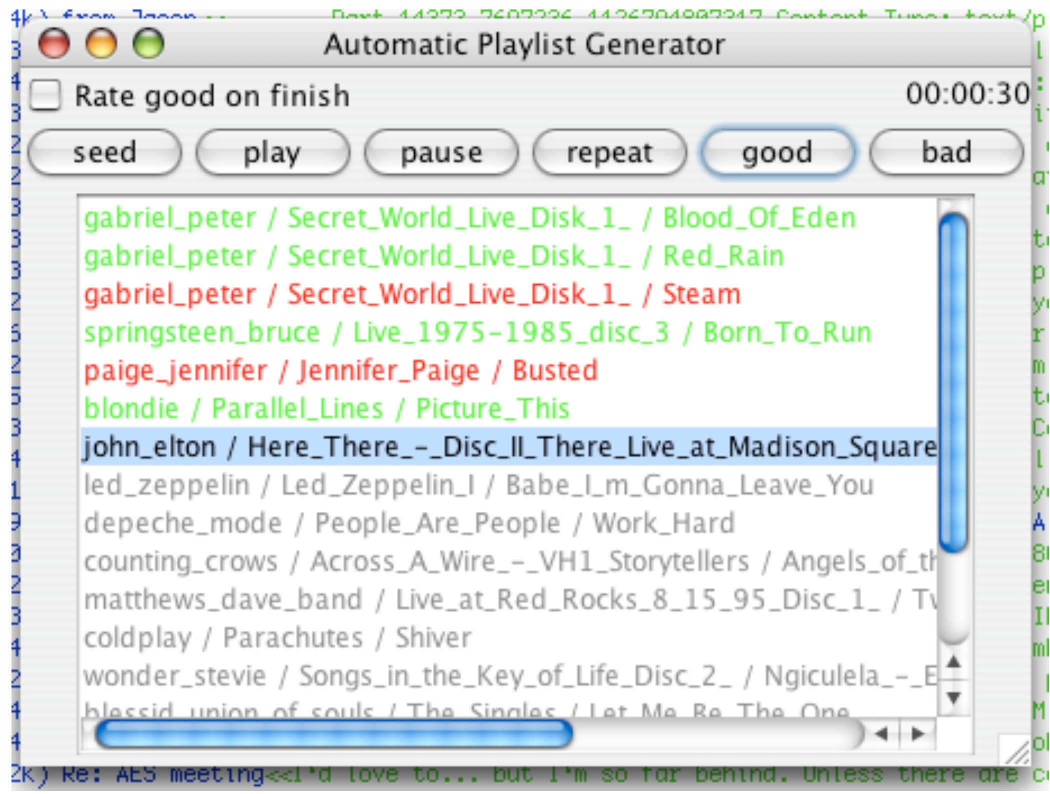
- ISMIR/MIREX 2005 also evaluated Artist ID
- 148 artists, 1800 files (split train/test) from ‘uspop2002’
- Song-level SVM clearly dominates
  - using only MFCCs!

MIREX 05 Audio Artist (USPOP2002)

Rank	Participant	Raw Accuracy	Normalized	Runtime / s
1	Mandel	**68.3%**	**68.0%**	10240
2	Bergstra	59.9%	60.9%	86400
3	Pampalk	56.2%	56.0%	4321
4	West	41.0%	41.0%	26871
5	Tzanetakis	28.6%	28.5%	2443
6	Logan	14.8%	14.8%	?
7	Lidy	Did not complete		

Playlist Generation

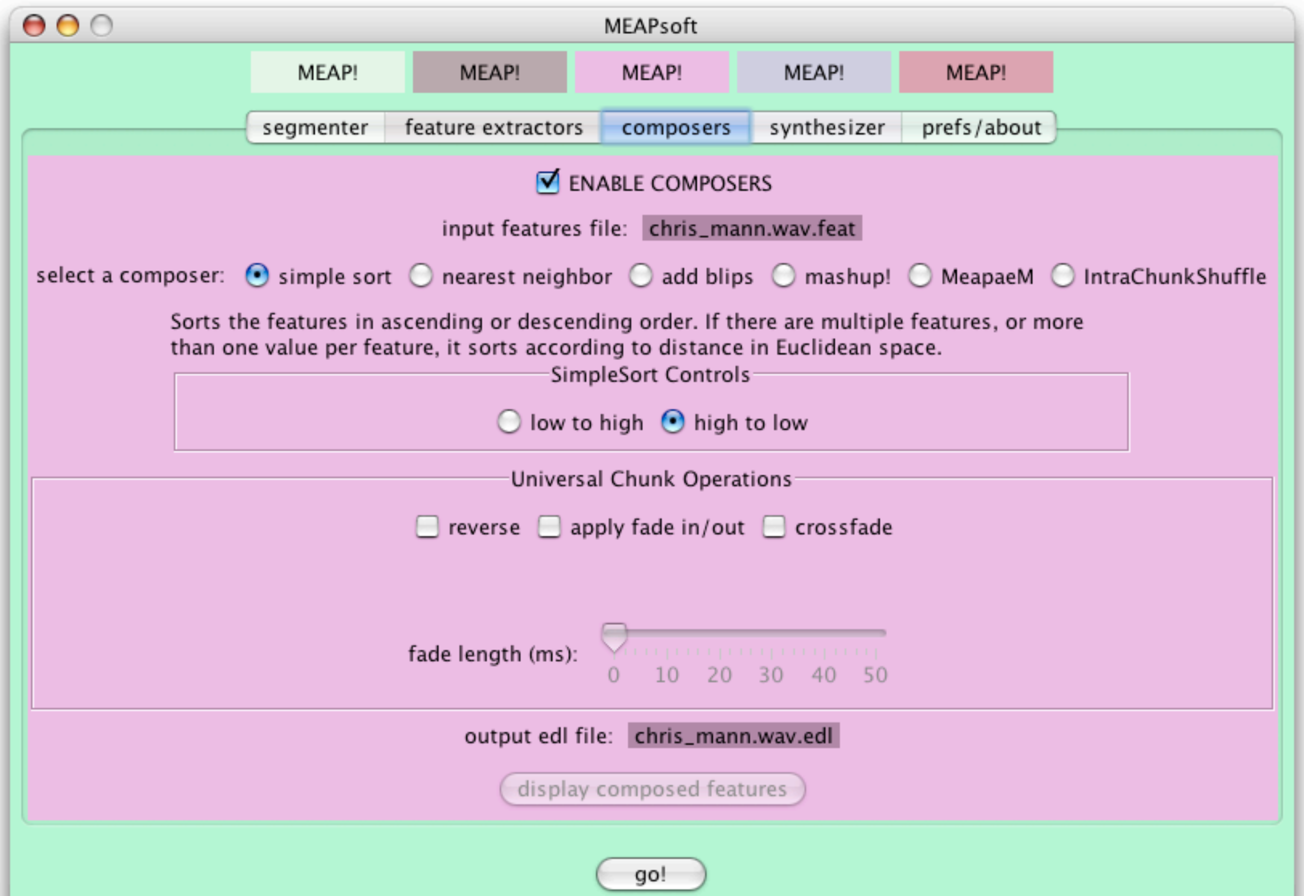
- SVMs are well suited to "active learning"
  - solicit labels on items closest to current boundary
- Automatic player with "skip" = Ground truth data collection
  - active-SVM automatic playlist generation



5. Artistic Application

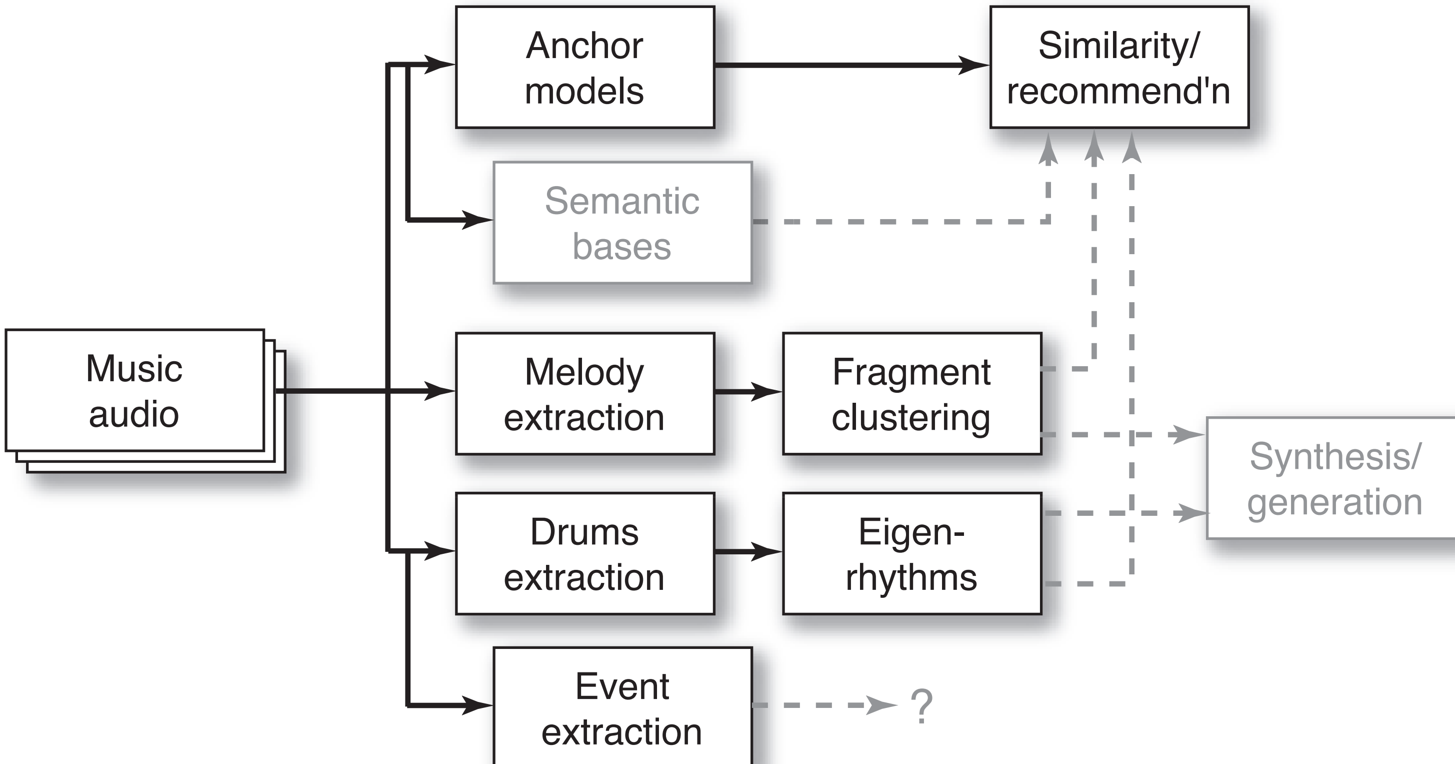
with Douglas Repetto, Ron Weiss, and the rest of the MEAP team

- "Compositional" applications of automatic music analysis



- music reformulation – automatic mashup generator

Conclusions



- Lots of data
  + noisy transcription
  + weak clustering
  ⇒ musical insights?