What can we Learn from Large Music Databases?

Dan Ellis
Laboratory for Recognition and Organization of Speech and Audio
Dept. Electrical Engineering, Columbia University, NY USA
dpwe@ee.columbia.edu

1. Learning Music
2. Music Similarity
3. Melody, Drums, Event extraction
4. Conclusions

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

Columbia University in the City of New York

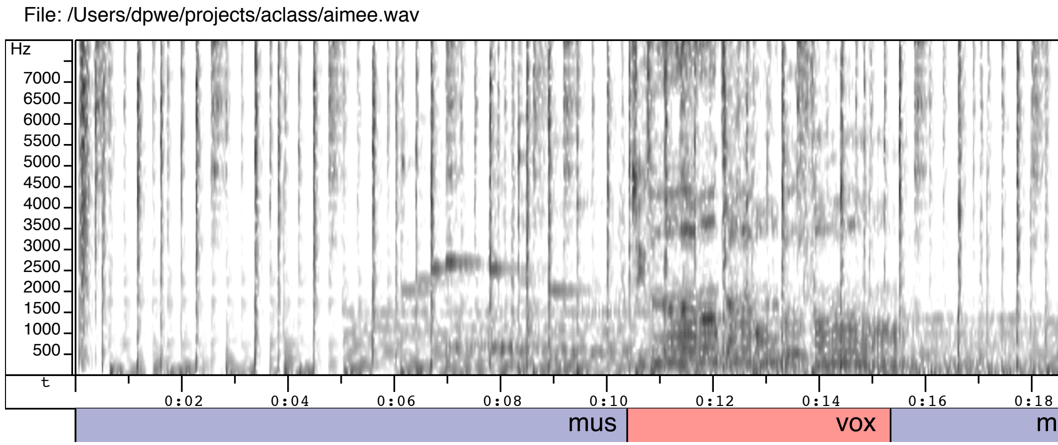
Learning from Music

- 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 / classification
  - computer (assisted) music generation
  - insight into music

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

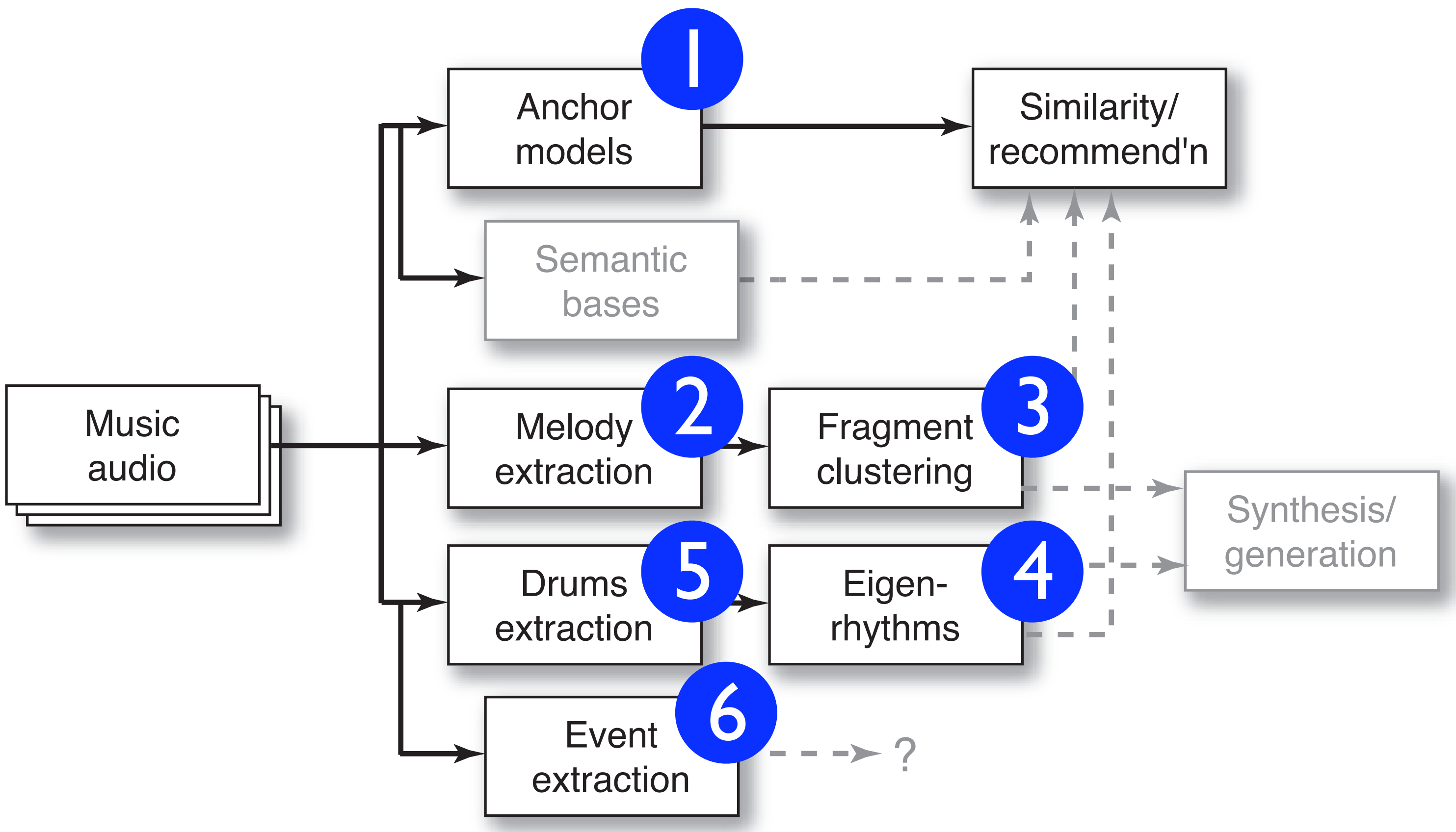
Columbia University in the City of New York

Ground Truth Data



- A lot of **unlabeled** music data available
  - manual annotation is much rarer
- **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?

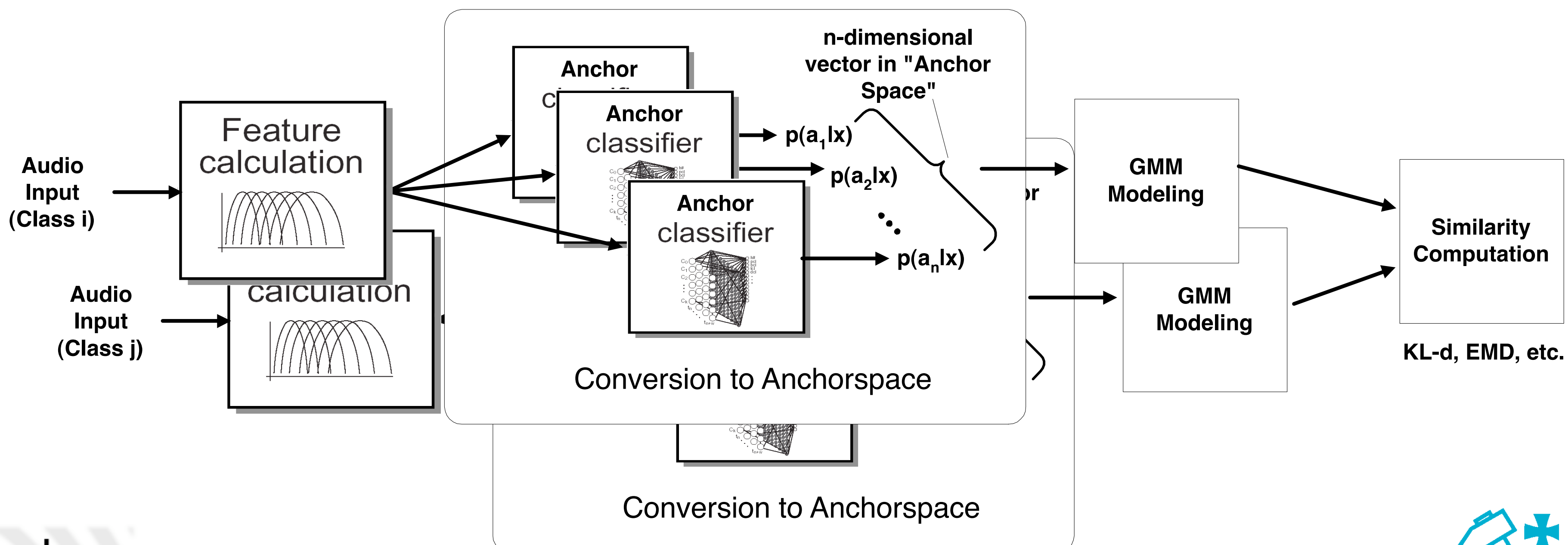
Talk Roadmap



1. Music Similarity Browsing

with Adam Berenzweig

- **Musical information overload**
  - record companies filter/categorize music
  - an automatic system would be less odious
- **Connecting audio and preference**
  - map to a 'semantic space'?

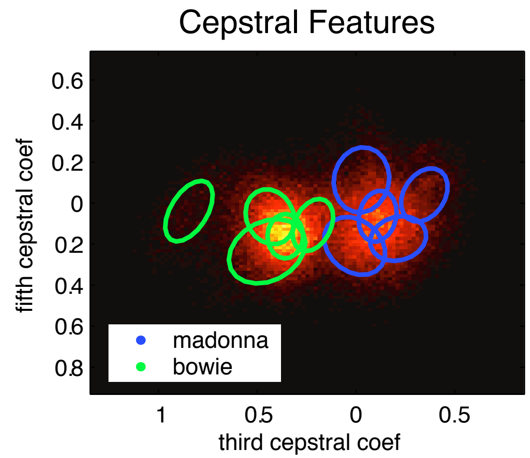


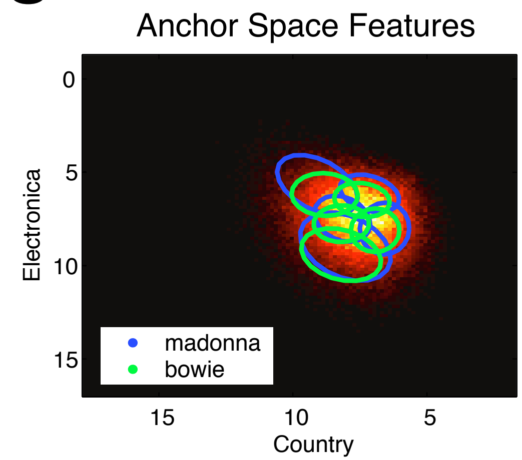
Lab ROSA
Laboratory for the Recognition and Organization of Speech and Audio

Columbia University in the City of New York

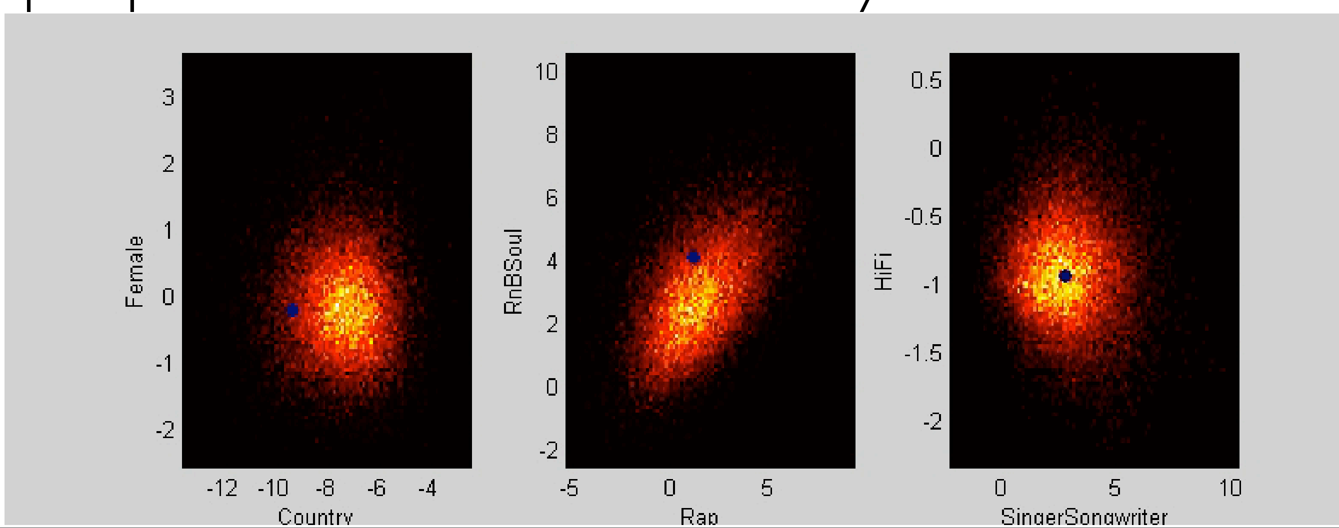
Anchor Space

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

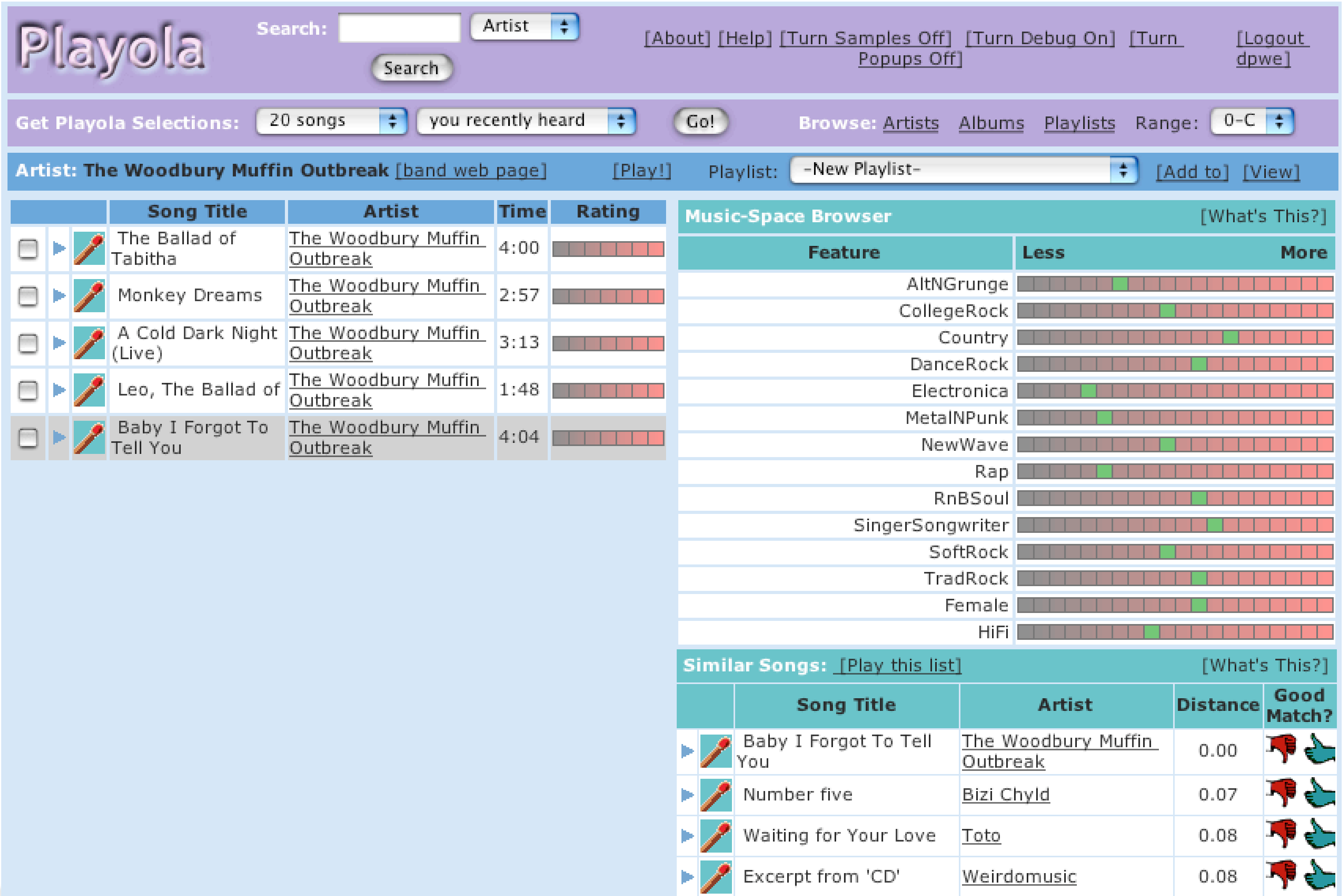




  - properties in distributions? dynamics?



'Playola' Similarity Browser



Semantic Bases

Brian Whitman

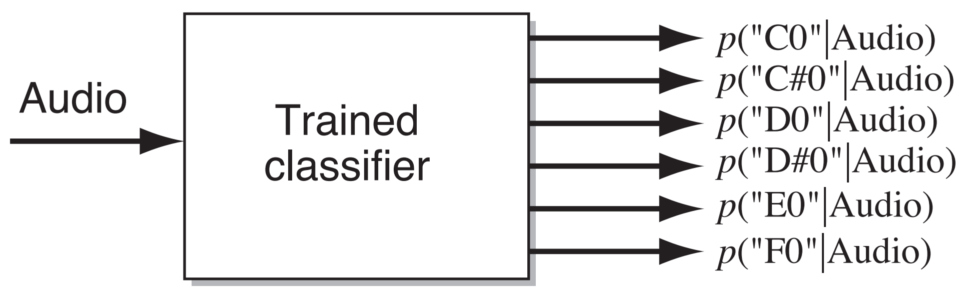
- **What should the 'anchor' dimensions be?**
  - hand-chosen genres? X
  - somehow choose automatically
- **"Community metadata": Use Web to get words/phrases..**
  - .. that are informative about artists
  - .. *and* that can be predicted from audio
- **Refine classifiers to below artist level**
  - e.g. by EM?

adj Term	K-L bits	np Term	K-L bits
aggressive	0.0034	reverb	0.0064
softer	0.0030	the noise	0.0051
synthetic	0.0029	new wave	0.0039
punk	0.0024	elvis costello	0.0036
sleepy	0.0022	the mud	0.0032
funky	0.0020	his guitar	0.0029
noisy	0.0020	guitar bass and drums	0.0027
angular	0.0016	instrumentals	0.0021
acoustic	0.0015	melancholy	0.0020
romantic	0.0014	three chords	0.0019

2. Transcription as Classification

with Graham Poliner

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



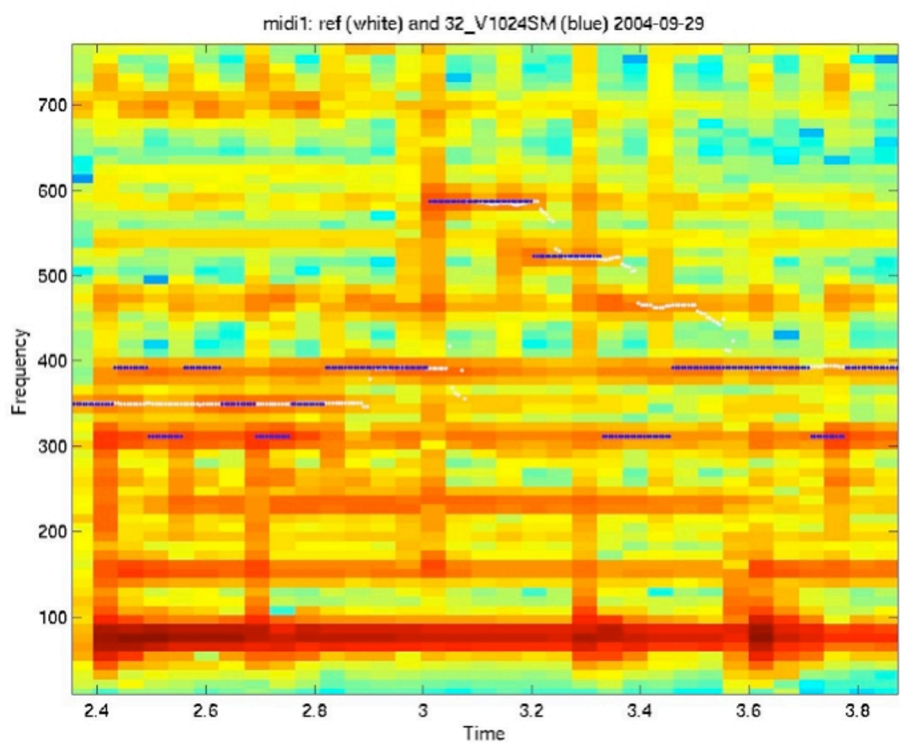
  - single N-way discrimination for "melody"
  - per-note classifiers for polyphonic transcription

Classifier Transcription Results

- Trained on MIDI syntheses (32 songs)
  - SMO SVM (Weka)
- Tested on ISMIR MIREX 2003 set
  - foreground/background separation

Frame-level pitch concordance

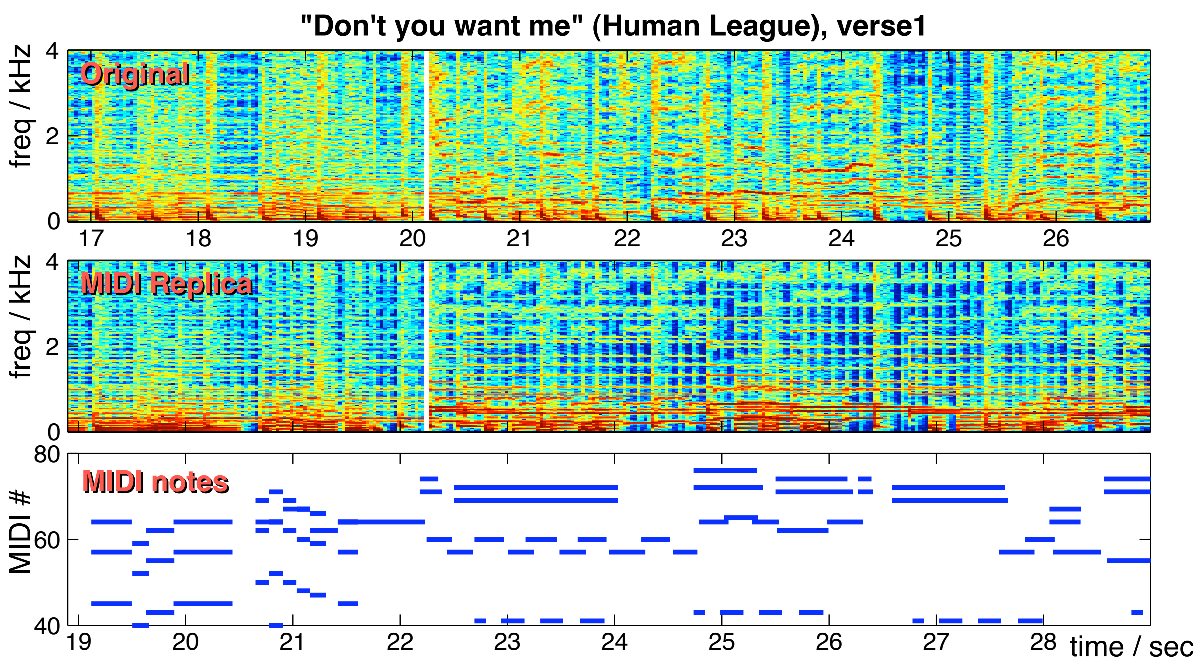
system	*"jazz3"*	*overall*
fg+bg	71.5%	44.3%
just fg	56.1%	45.4%



Forced-Alignment of MIDI

with Rob Turetsky

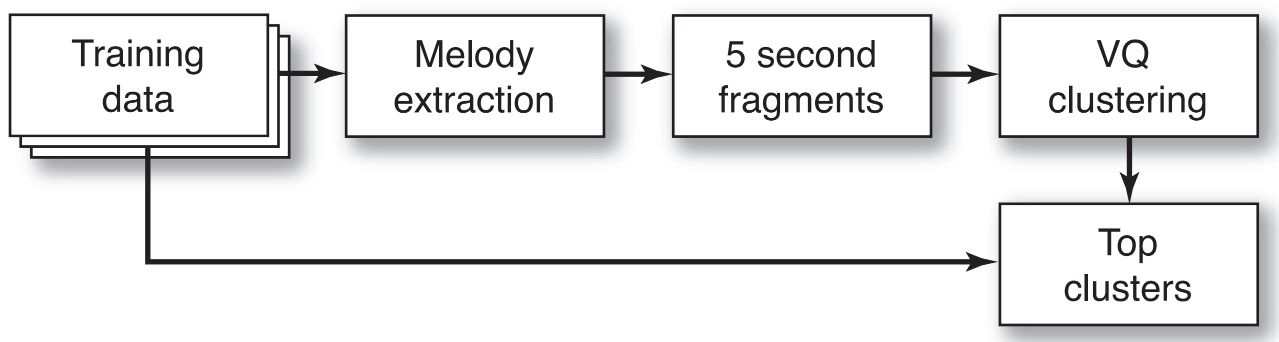
- MIDI is a handy description of music
  - notes, instruments, tracks
  - .. to drive synthesis
- Align MIDI 'replicas' to get GTruth for audio
  - estimate time-warp relation



3. Melody Clustering

with Graham Poliner

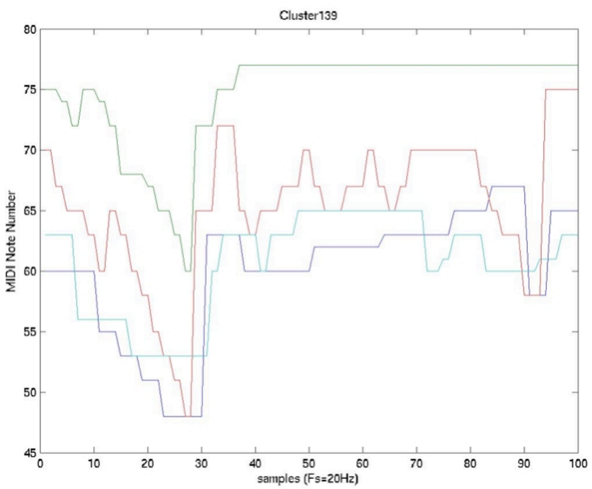
- Goal: Find 'fragments' that recur in melodies
  - .. across large music database
  - .. trade data for model sophistication

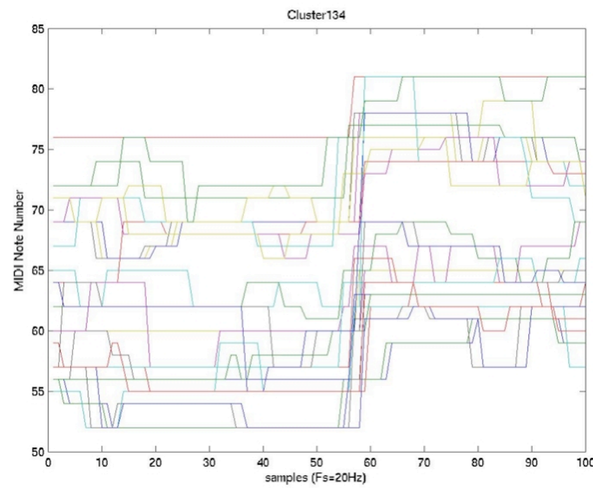


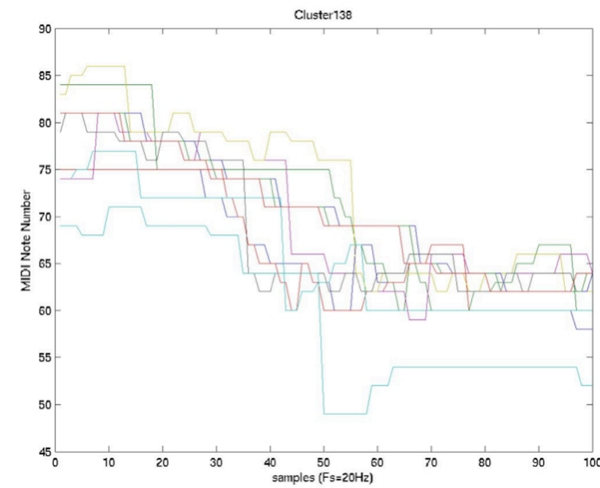
- Data sources
  - pitch tracker, or MIDI training data
- Melody fragment representation
  - DCT(1:20) - removes average, smoothes detail

Melody clustering results

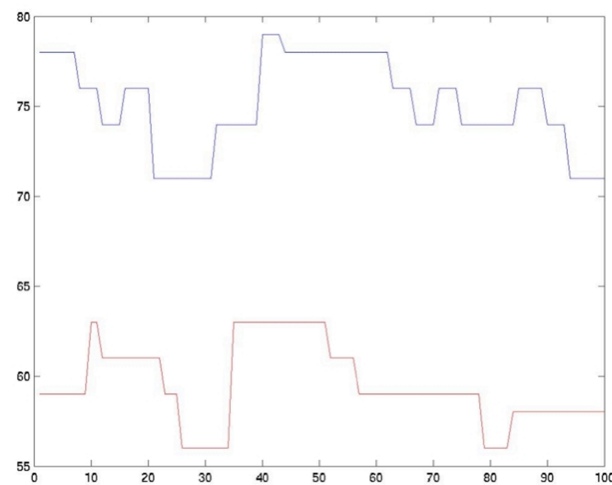
- Clusters match underlying contour:







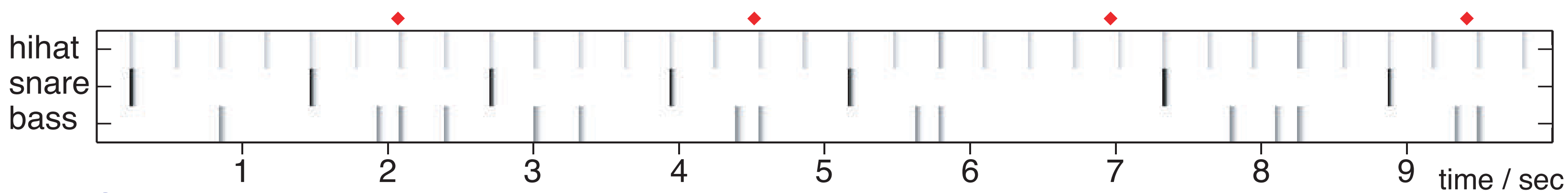
- Finds some similarities:
  - e.g. Pink + Nsync



4. Eigenrhythms: Drum Pattern Space

with John Arroyo

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



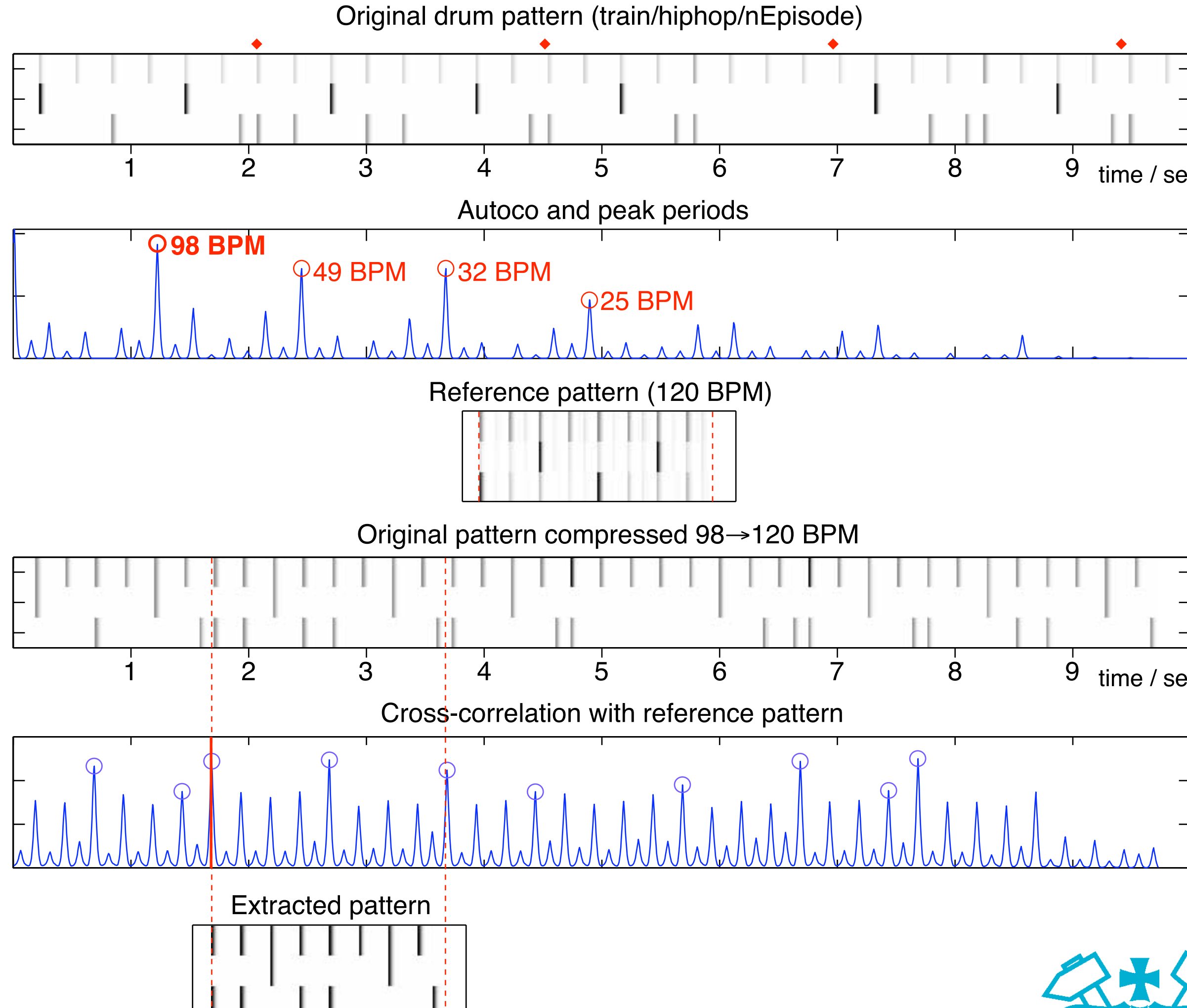
- Eigen-analysis (or ...) to capture variations?
    - by analyzing lots of (MIDI) data, or from audio
- Applications
    - music categorization
    - "beat box" synthesis
    - insight

Aligning the Data

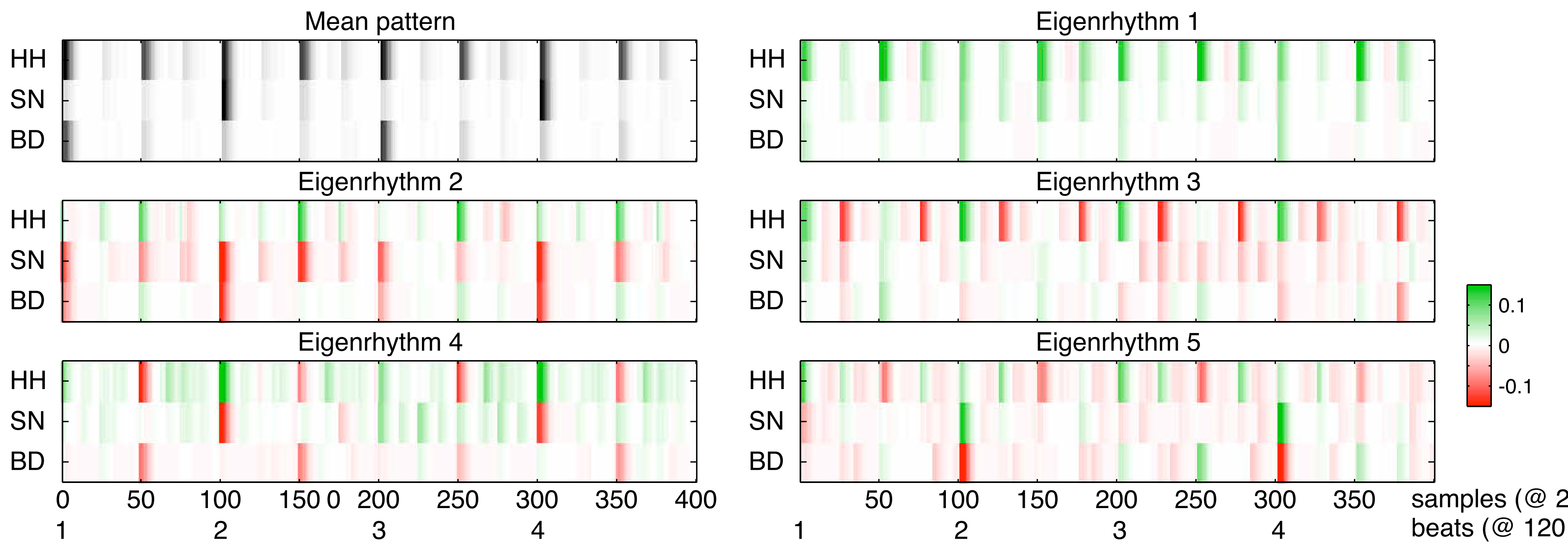
- Need to align patterns prior to modeling...

tempo (stretch):
by inferring BPM &
normalizing

downbeat (shift):
correlate against
'mean' template

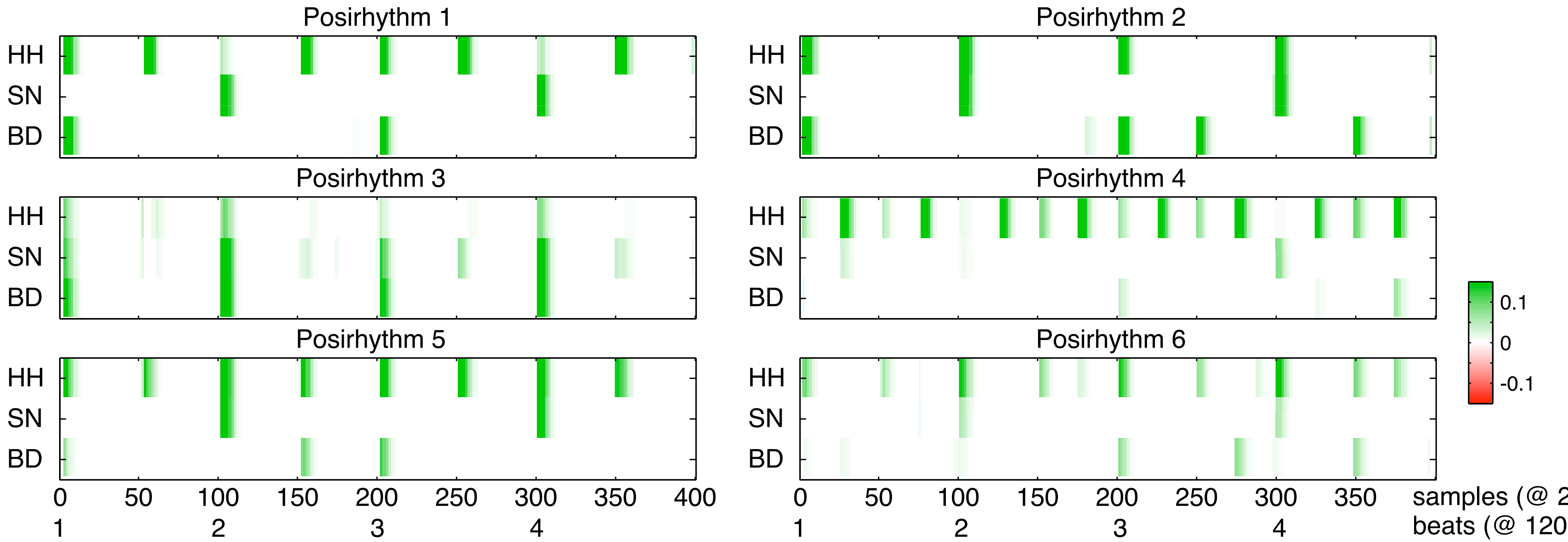


Eigenrhythms (PCA)



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

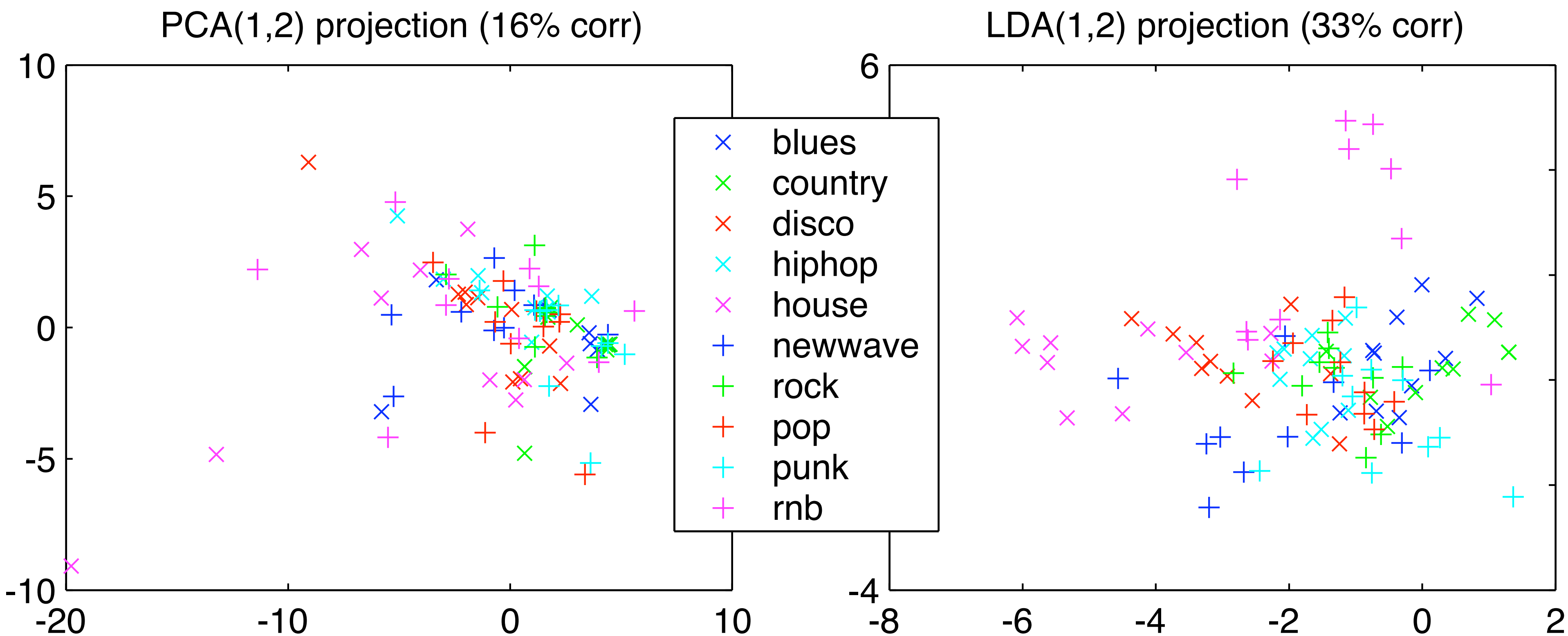
Posirhythms (NMF)



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

Eigenrhythms for Classification

- **Projections** in Eigenspace / LDA space



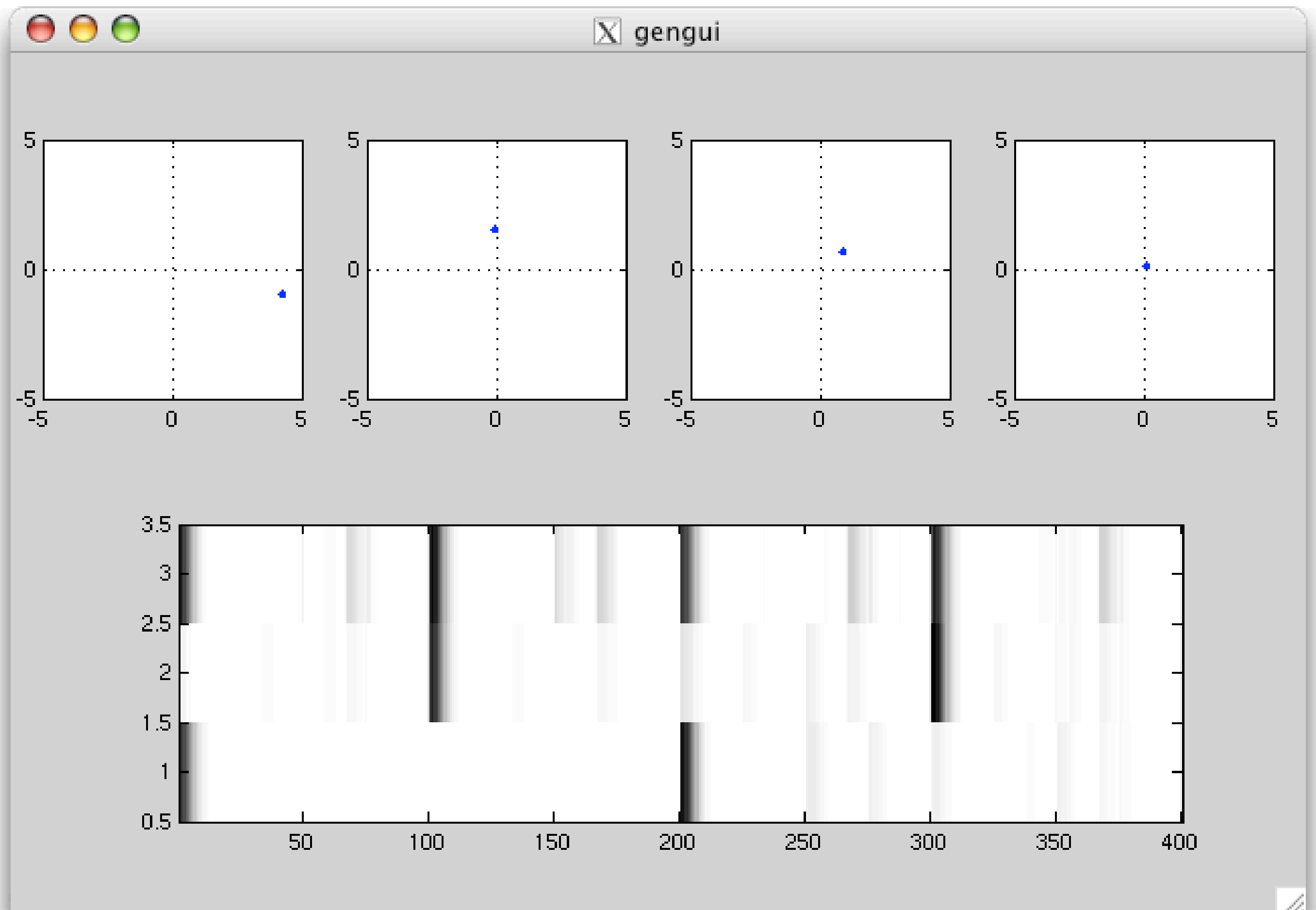
- 10-way **Genre** classification (nearest nbr):
  - PCA3: 20% correct
  - LDA4: 36% correct

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

Columbia University in the City of New York

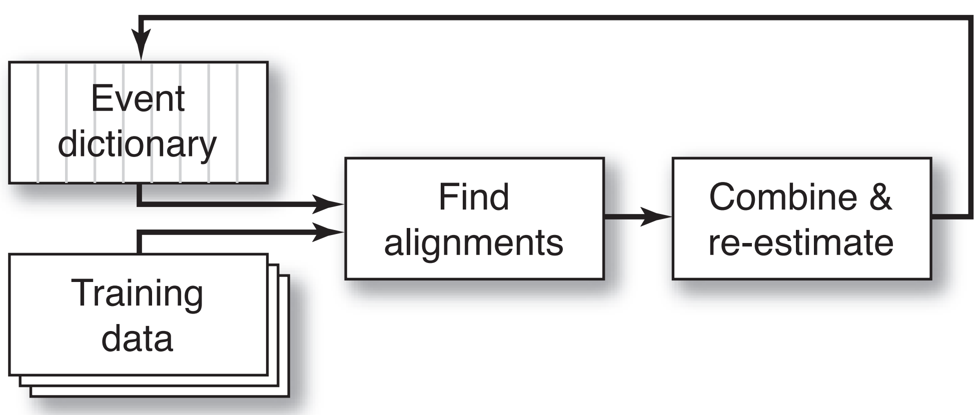
Eigenrhythm BeatBox

- Resynthesize rhythms from eigen-space



5. Event Extraction

- Music often contains many repeated events
  - notes, drum sounds
  - but: usually overlapped...
- Vector Quantization finds common patterns:

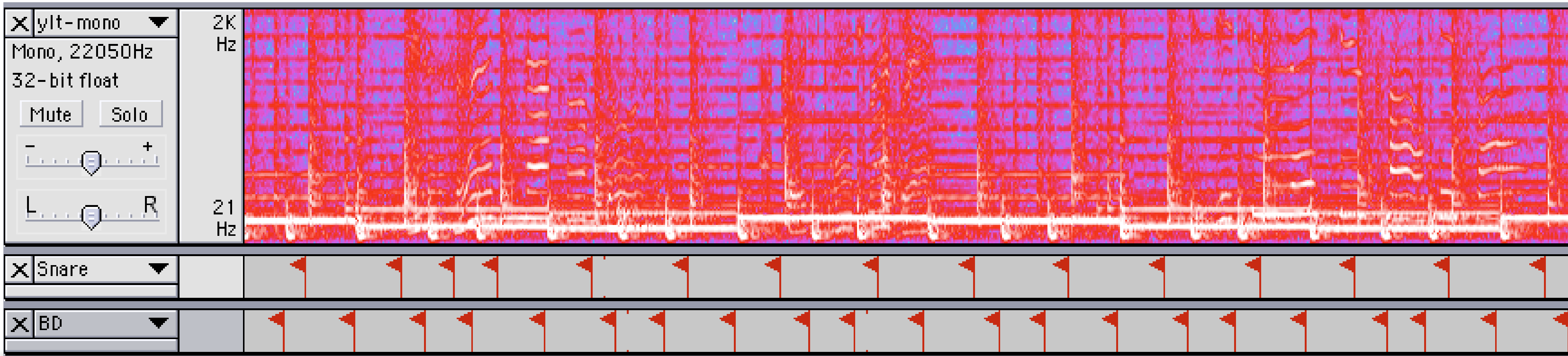


  - representation...
  - aligning/matching...
  - how much coverage required?

Drum Track Extraction

with Ron Weiss, after Yoshii et al. '04

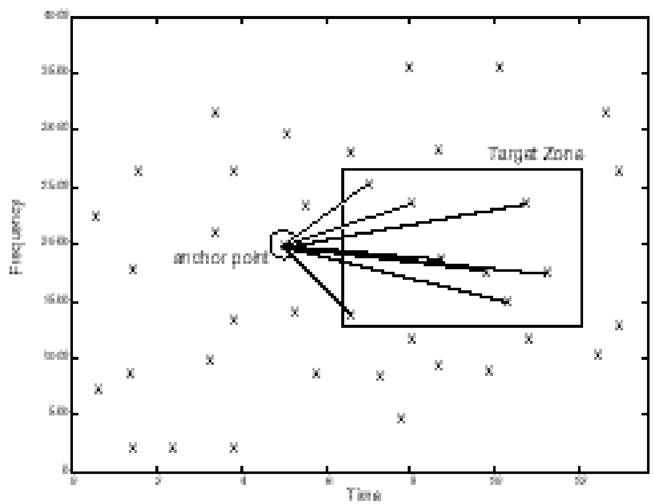
- Initialize dictionary with Bass Drum, Snare
- Match only on a few spectral peaks
  - narrowband energy most likely to avoid overlap
- Median filter to re-estimate template
  - .. after normalizing amplitudes
  - can pick up partials from common notes

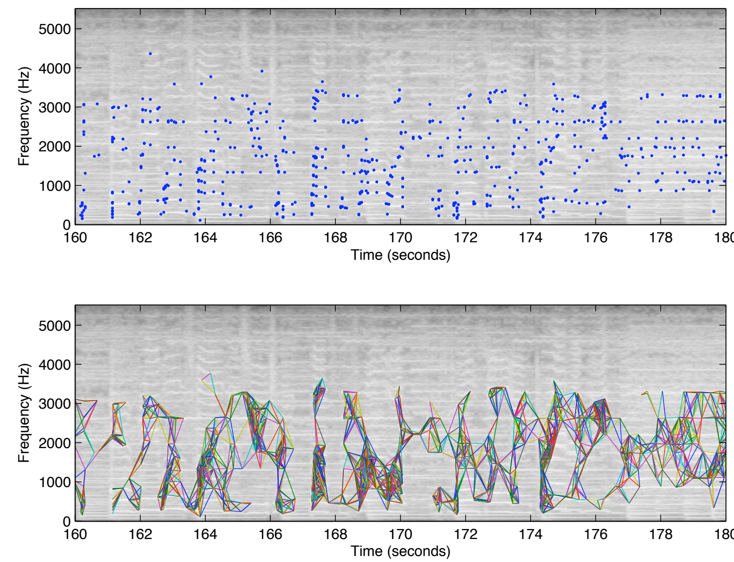


Generalized Event Detection

with Michael Mandel

- Based on 'Shazam' audio fingerprints (Wang'03)

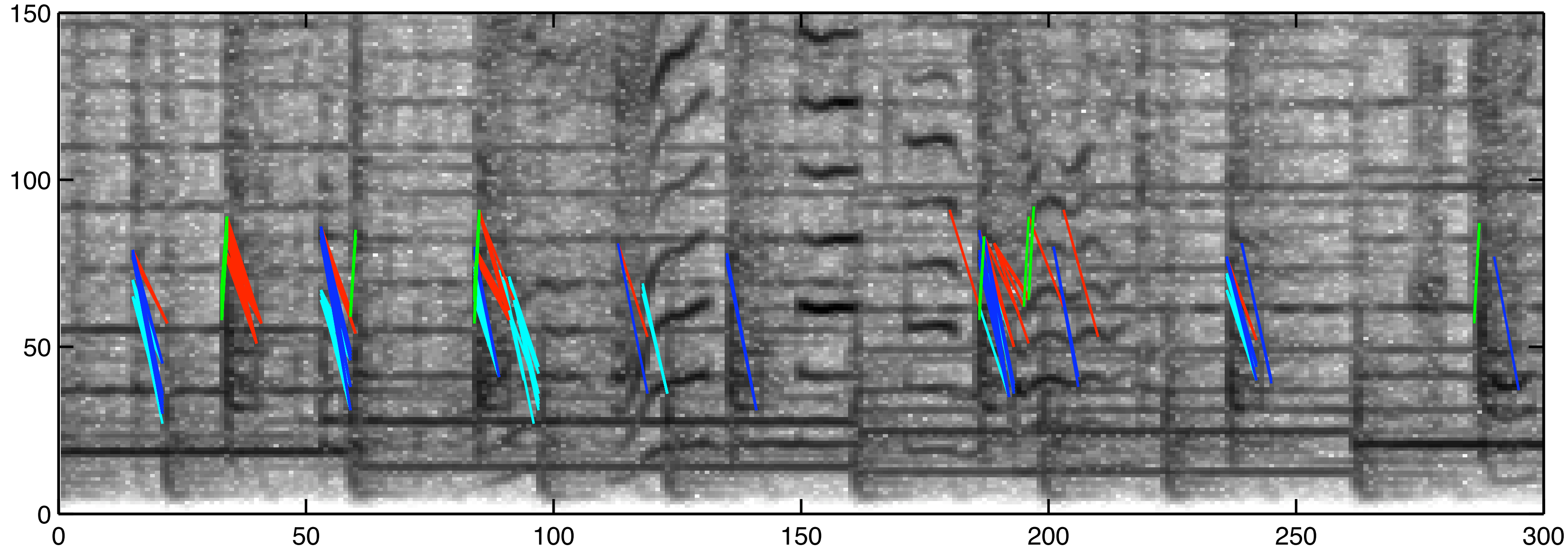




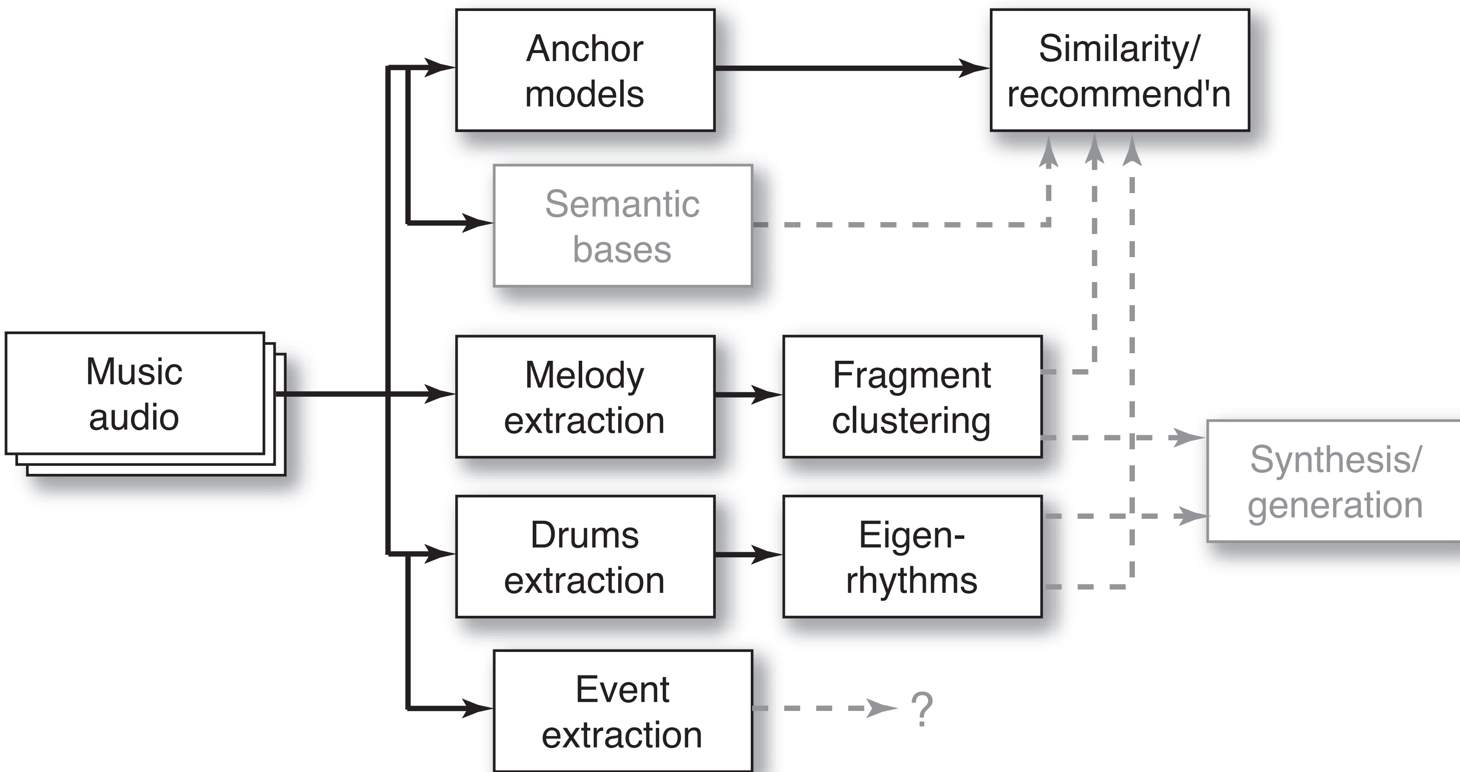
  - relative timing of F_1-F_2-ΔT triples discriminates pieces
  - narrowband features to avoid collision (again)
- **Fingerprint events, not recordings: choose top triples, look for repeats**
  - rank reduction of triples x time matrix

Event detection results

- **Procedure**
  - find hash triples
  - cluster them
  - patterns in hash co-occurrence = events?



Conclusions



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

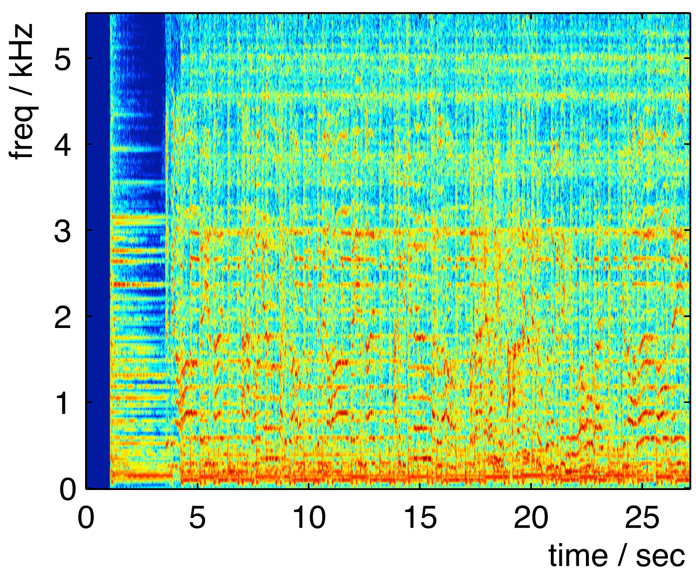
Approaches to Chord Transcription

with Alex Sheh

- **Note transcription, then note→chord rules**
  - like labeling chords in MIDI transcripts
- **Spectrum→chord rules**
  - i.e. find harmonic peaks, use knowledge of likely notes in each chord
- **Trained classifier**
  - don't use any "expert knowledge"
  - instead, learn patterns from labeled examples
- **Train ASR HMMs with chords ≈ words**

Chord Sequence Data Sources

- **All we need are the chord sequences for our training examples**
  - Hal Leonard "Paperback Song Series"
    - manually retyped for 20 songs: "Beatles for Sale", "Help", "Hard Day's Night"




```
# The Beatles - A Hard Day's Night
#
G Cadd9 G F6 G Cadd9 G F6 G C D G C9 G
G Cadd9 G F6 G Cadd9 G F6 G C D G C9 G
Bm Em Bm G Em C D G Cadd9 G F6 G Cadd9 G
 F6 G C D G C9 G D
G C7 G F6 G C7 G F6 G C D G C9 G Bm Em Bm
 G Em C D
G Cadd9 G F6 G Cadd9 G F6 G C D G C9 G
C9 G Cadd9 Fadd9
```

    - hand-align chords for 2 test examples

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

Columbia University in the City of New York

Chord Results

- Recognition weak, but forced-alignment OK

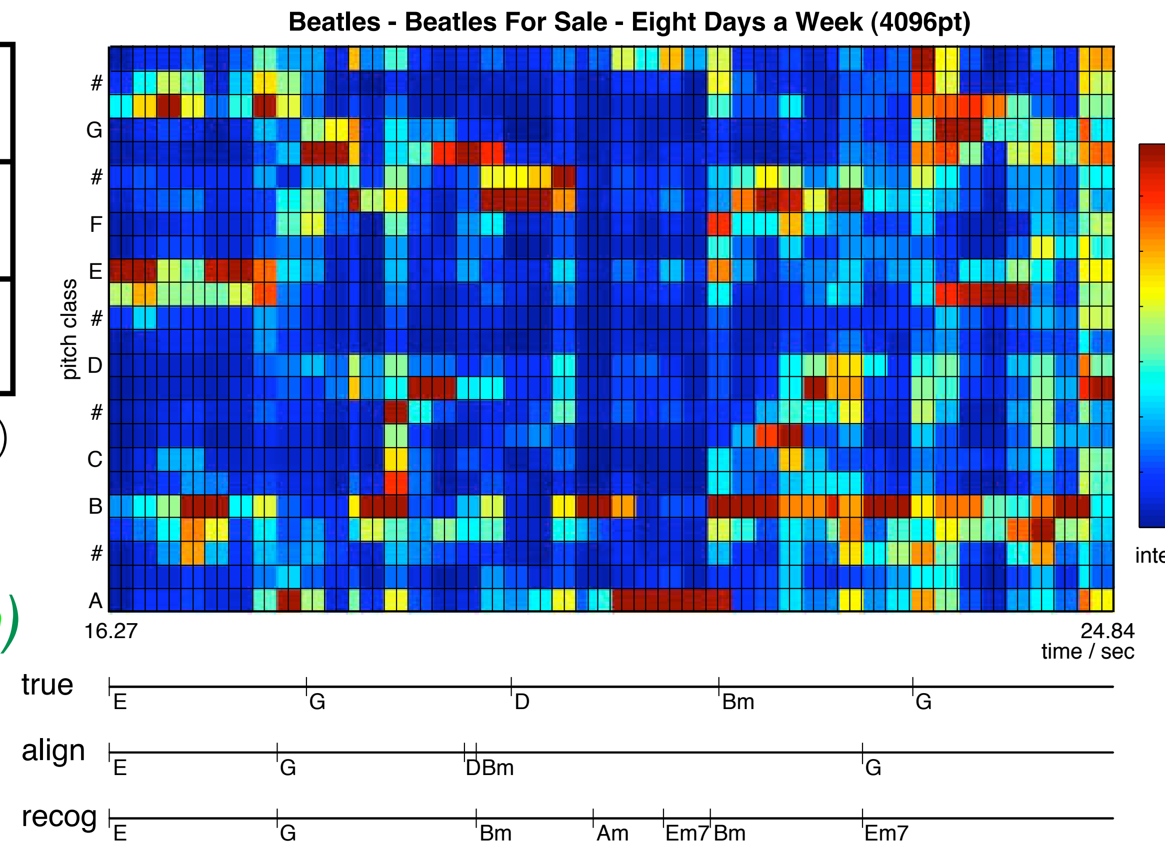
Frame-level Accuracy

Feature	*Recognition*	*Alignment*
MFCC	8.7%	22.0%
PCP_ROT	21.7%	76.0%

(random ~3%)

MFCCs are poor (can overtrain)

PCPs better (ROT helps generalization)



What did the models learn?

- Chord model centers (means) indicate chord 'templates':