Extracting and Using Music Audio Information

Dan Ellis
Laboratory for Recognition and Organization of Speech and Audio
Dept. Electrical Engineering, Columbia University, NY USA

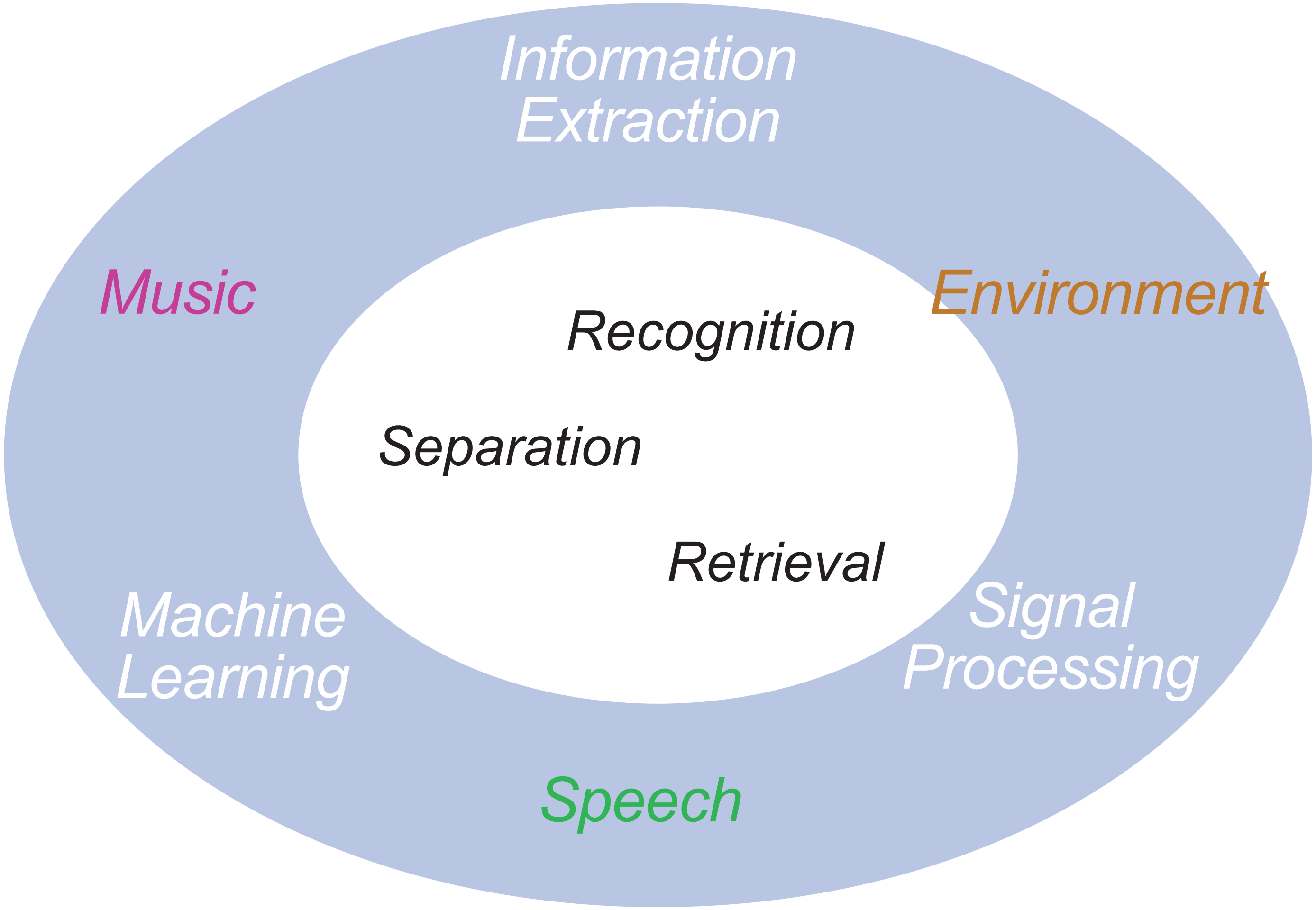
http://labrosa.ee.columbia.edu/

1. Motivation: Music Collections
2. Music Information
3. Music Similarity
4. Music Structure Discovery

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

Columbia University in the City of New York

LabROSA Overview

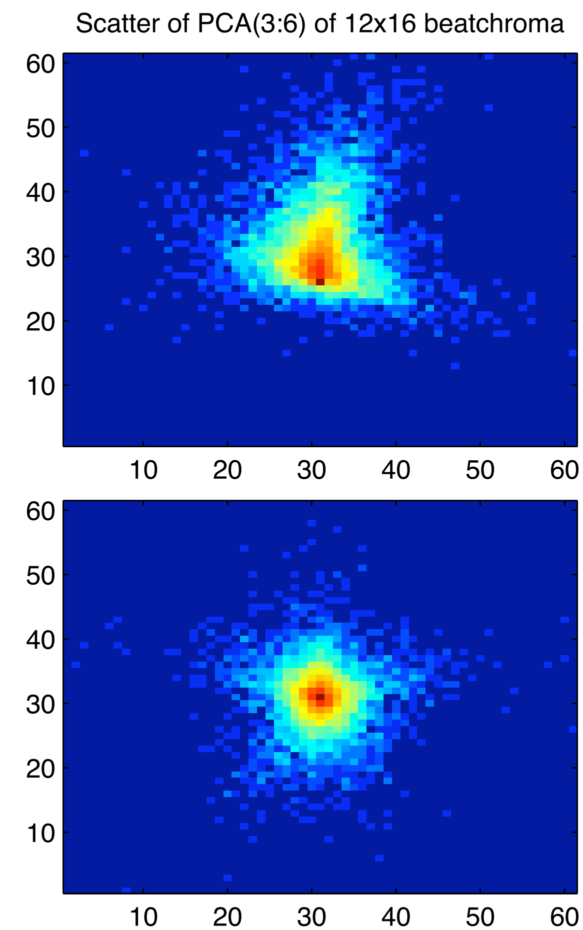


1. Managing Music Collections

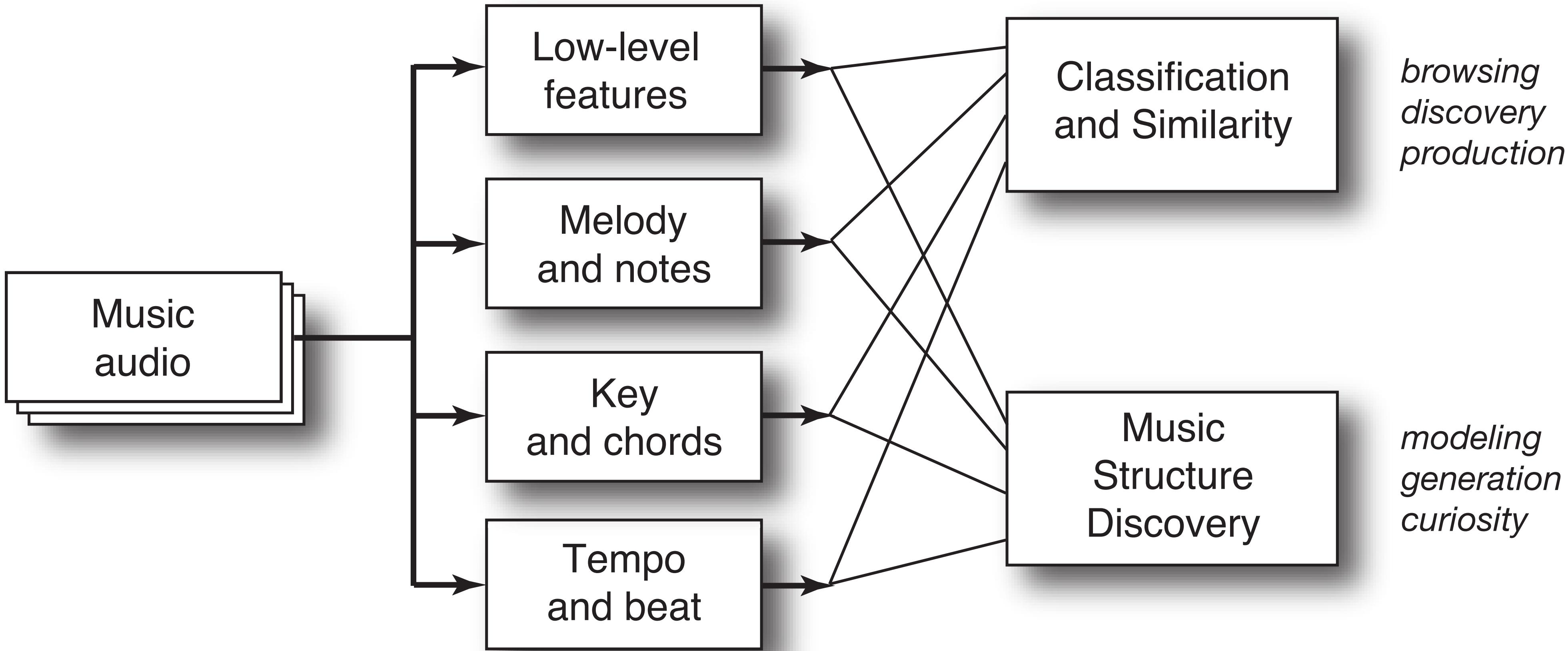
- A lot of music data available
  - e.g. 60G of MP3 ≈ 1000 hr of audio, 15k tracks
- Management challenge
  - how can computers help?
- Application scenarios
  - personal music collection
  - discovering new music
  - “music placement”

Learning from Music

- What can we infer from 1000 h of music?
  - common patterns
    sounds, melodies, chords, form
  - what is and what isn't music
- Data driven musicology?
- Applications
  - modeling/description/coding
  - computer generated music
  - curiosity...



The Big Picture

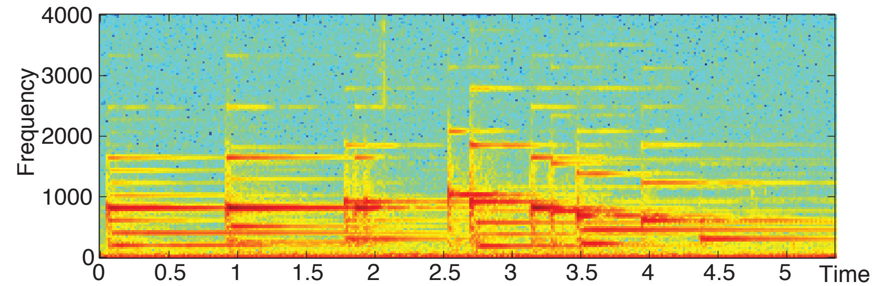


.. so far

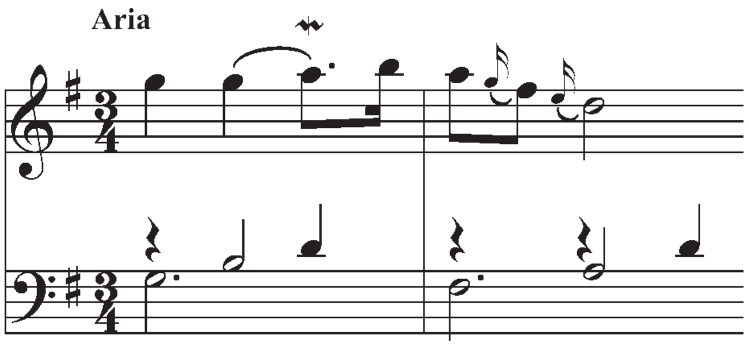
2. Music Information

- How to represent music audio?

- Audio features
  - spectrogram, MFCCs, bases



- Musical elements
  - notes, beats, chords, phrases
  - requires transcription



- Or something inbetween?
  - optimized for a certain task?

Transcription as Classification

Poliner & Ellis '05,'06,'07

- Exchange signal models for data
  - transcription as pure classification problem:

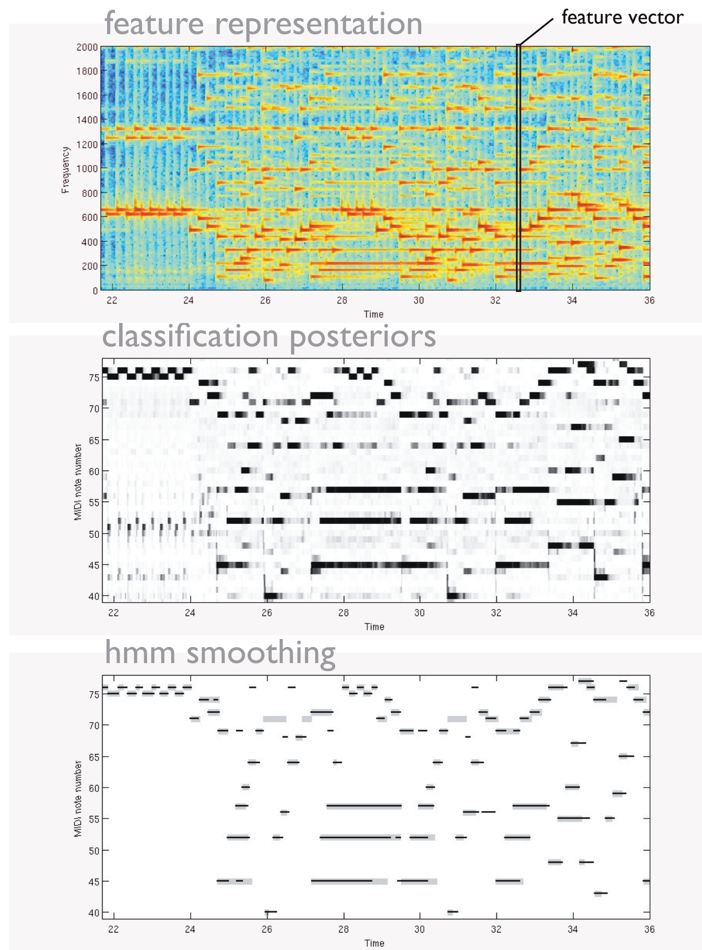
Training data and features:
- MIDI, multi-track recordings, playback piano, & resampled audio (less than 28 mins of train audio).
- Normalized magnitude STFT.

↓

Classification:
- N-binary SVMs (one for ea. note).
- Independent frame-level classification on 10 ms grid.
- Dist. to class bndy as posterior.

↓

Temporal Smoothing:
- Two state (on/off) independent HMM for ea. note. Parameters learned from training data.
- Find Viterbi sequence for ea. note.



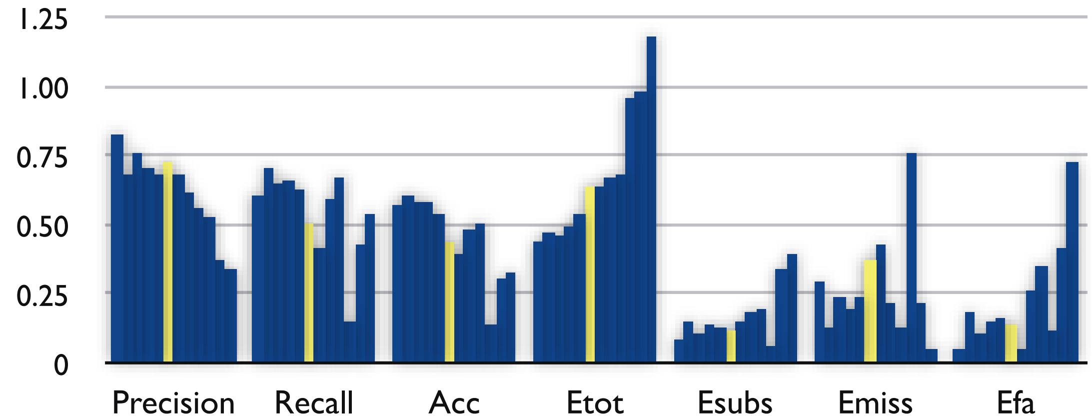
Polyphonic Transcription

MIREX 2007

- Real music excerpts + ground truth

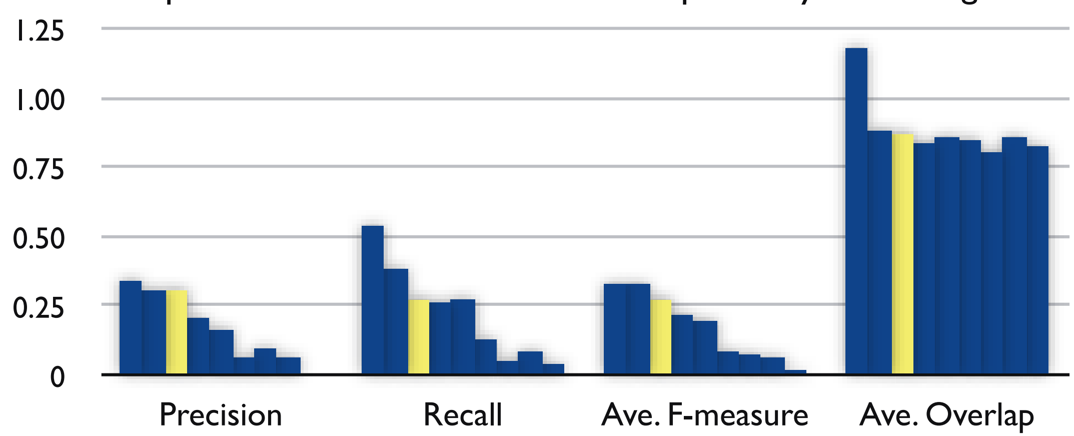
Frame-level transcription

Estimate the fundamental frequency of all notes present on a 10 ms grid



Note-level transcription

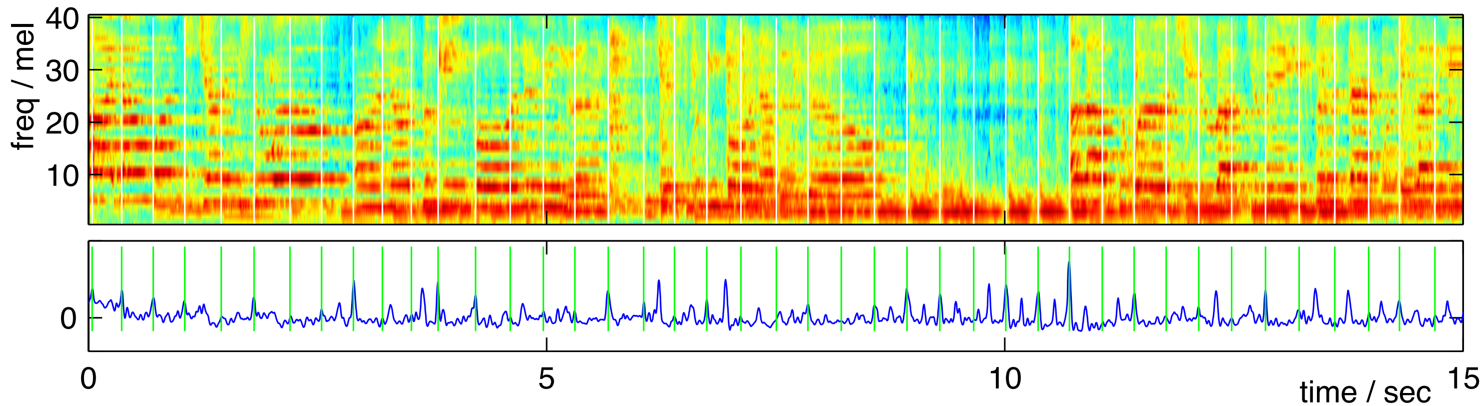
Group frame-level predictions into note-level transcriptions by estimating onset/offset



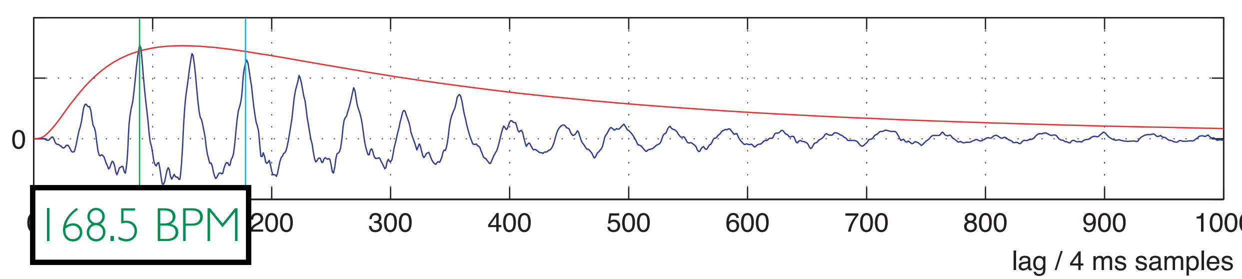
Beat Tracking

Ellis '06,'07

- Goal: One feature vector per 'beat' (tatum)
  - for tempo normalization, efficiency
- "Onset Strength Envelope"
  - $\mathrm{sum}_f(\max(0, \mathrm{diff}_t(\log |X(t, f)|)))$

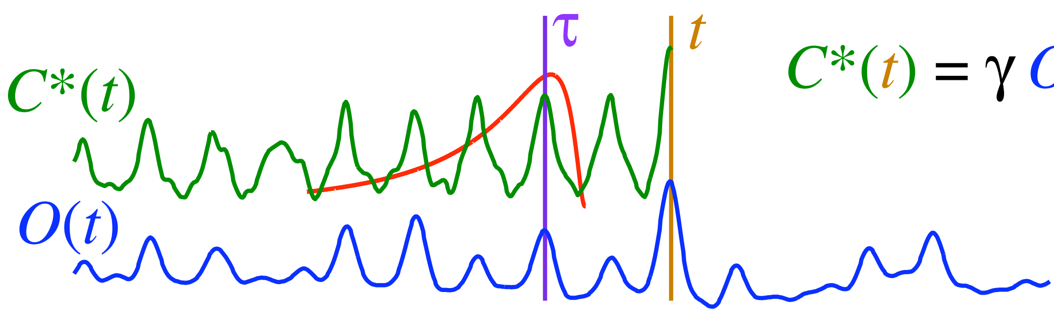


- Autocorr. + window → global tempo estimate



Beat Tracking

- Dynamic Programming finds beat times $\{t_i\}$
  - optimizes $\Sigma_i\, O(t_i) + \alpha\, \Sigma_i\, W\big((t_{i+1} - t_i - \tau_p)/\beta\big)$
  - where $O(t)$ is onset strength envelope (local score)
    $W(t)$ is a log-Gaussian window (transition cost)
    τ_p is the default beat period per measured tempo
  - incrementally find best predecessor at every time
  - backtrace from largest final score to get beats

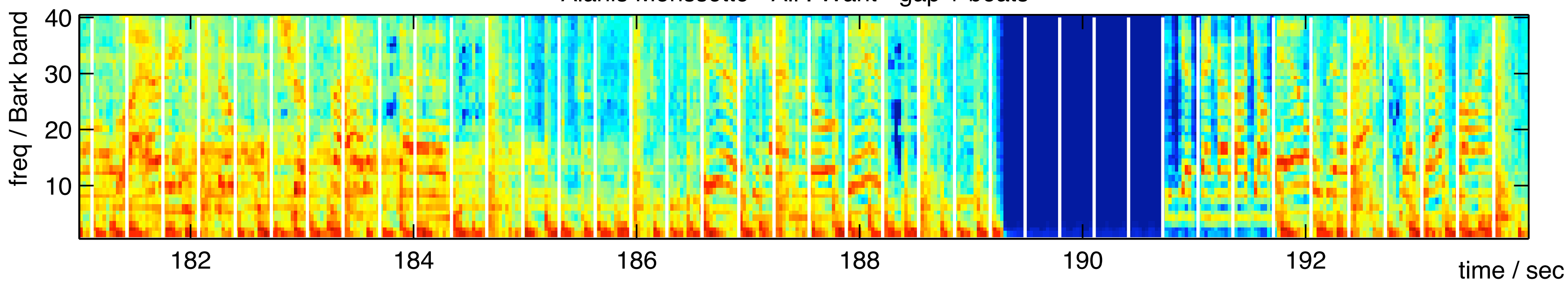


$$C^*(t) = \gamma\, O(t) + (1-\gamma) \max_{\tau}\{W\big((\tau - \tau_p)/\beta\big) C^*(\tau)\}$$

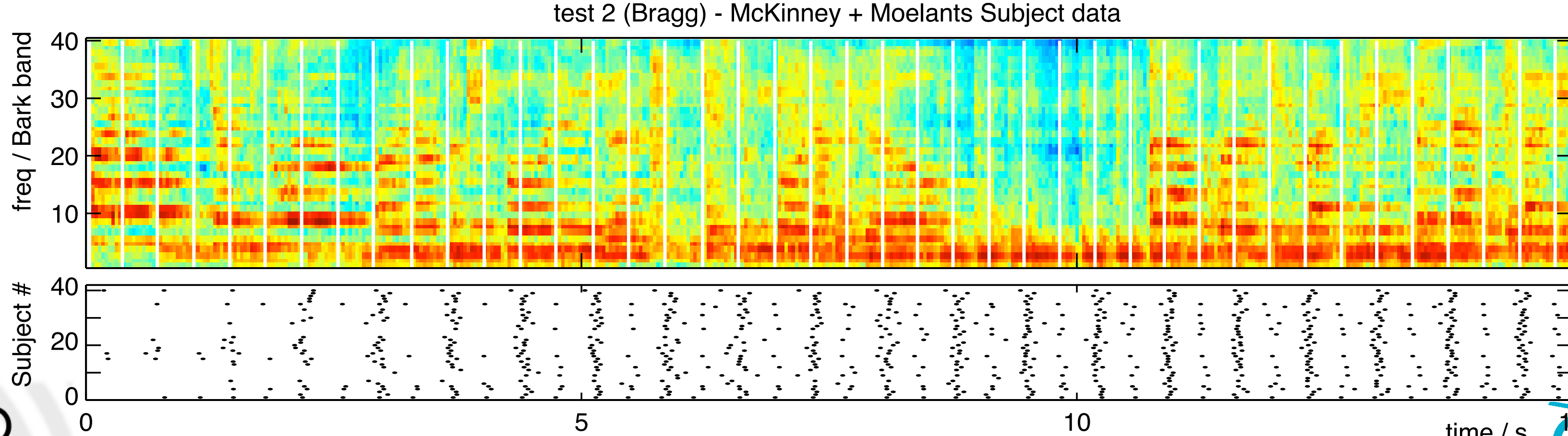
$$P(t) = \operatorname*{argmax}_{\tau}\{W\big((\tau - \tau_p)/\beta\big) C^*(\tau)\}$$

Beat Tracking

- DP will bridge gaps (non-causal)
  - there is always a best path ...



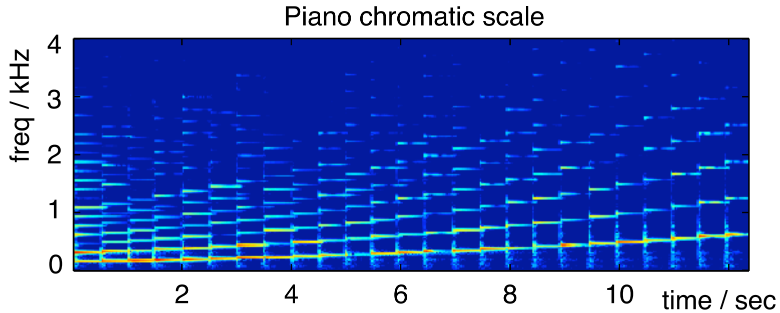
- 2nd place in MIREX 2006 Beat Tracking
  - compared to McKinney & Moelants human data

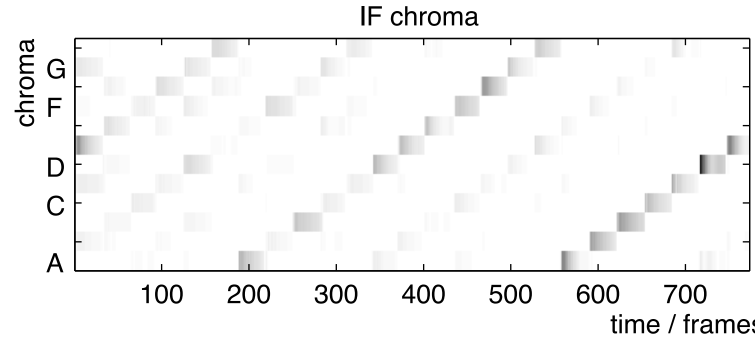


Chroma Features

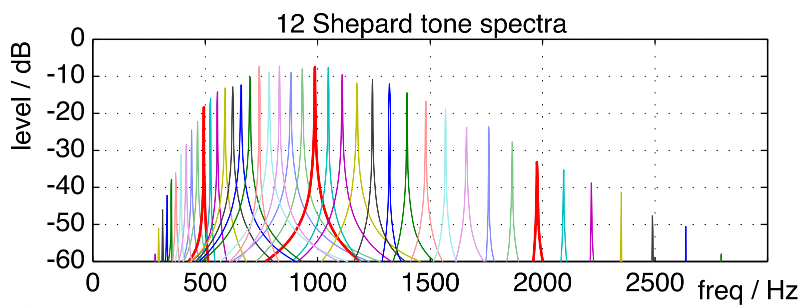
- Chroma features convert spectral energy into musical weights in a canonical octave
  - i.e. 12 semitone bins

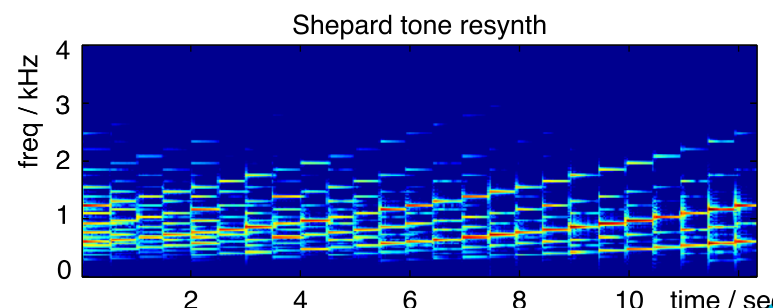
Piano scale





- Can resynthesize as "Shepard Tones"
  - all octaves at once

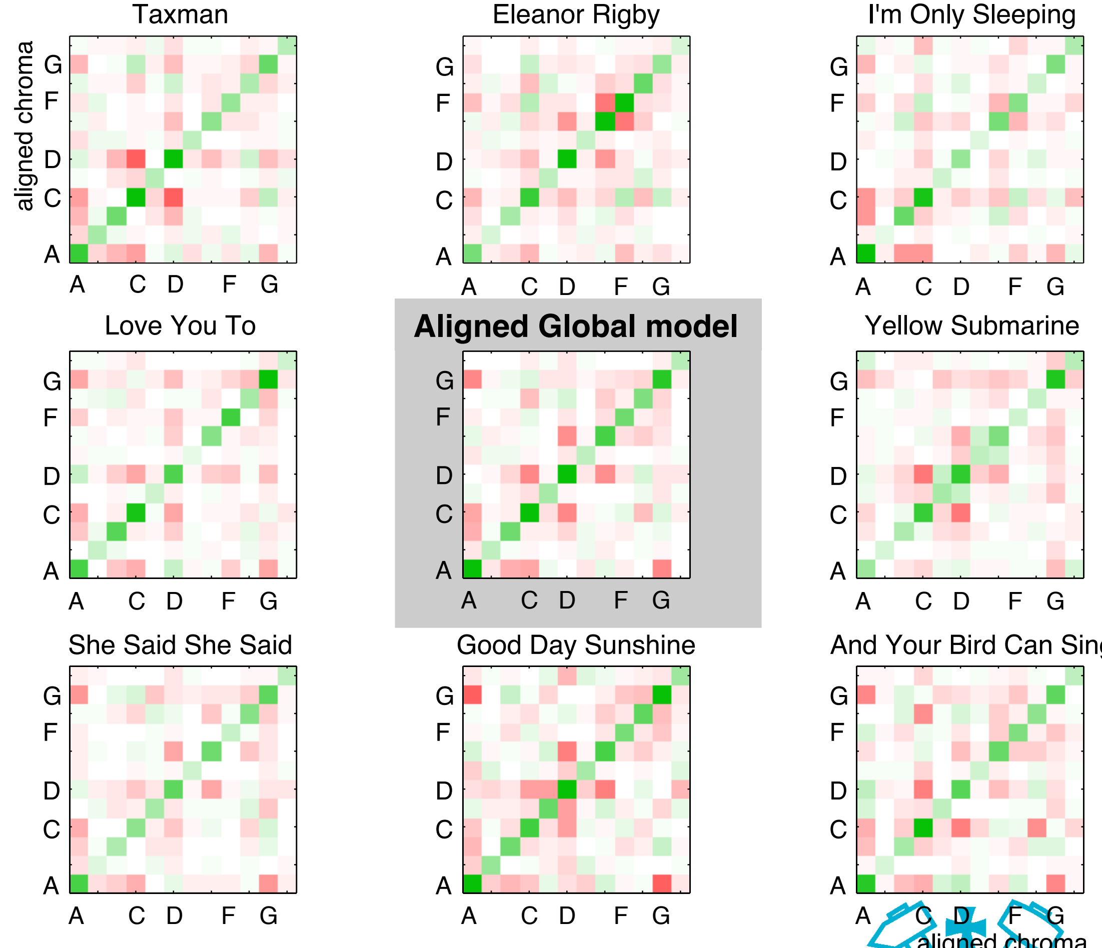




Key Estimation

Ellis ICASSP '07

- Covariance of chroma reflects key
- Normalize by transposing for best fit
  - single Gaussian model of one piece
  - find ML rotation of other pieces
  - model all transposed pieces
  - iterate until convergence

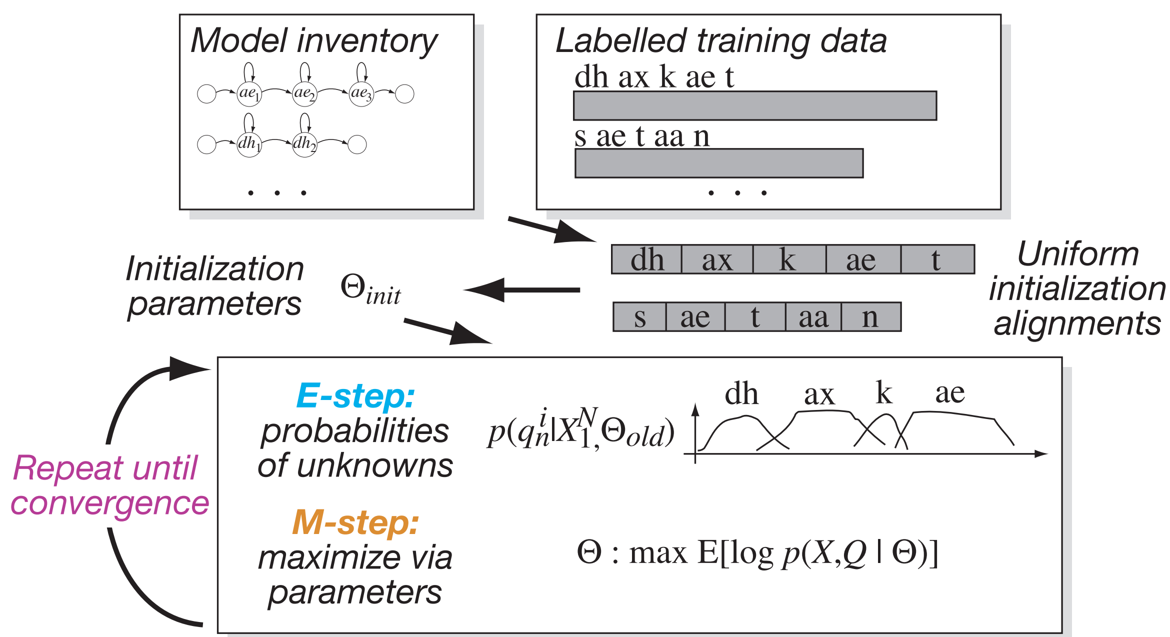


Chord Transcription

Sheh & Ellis '03

- "Real Books" give chord transcriptions
  - but no exact timing
  - .. just like speech transcripts
- Use EM to simultaneously learn and align chord models

```
# 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
```



Chord Transcription

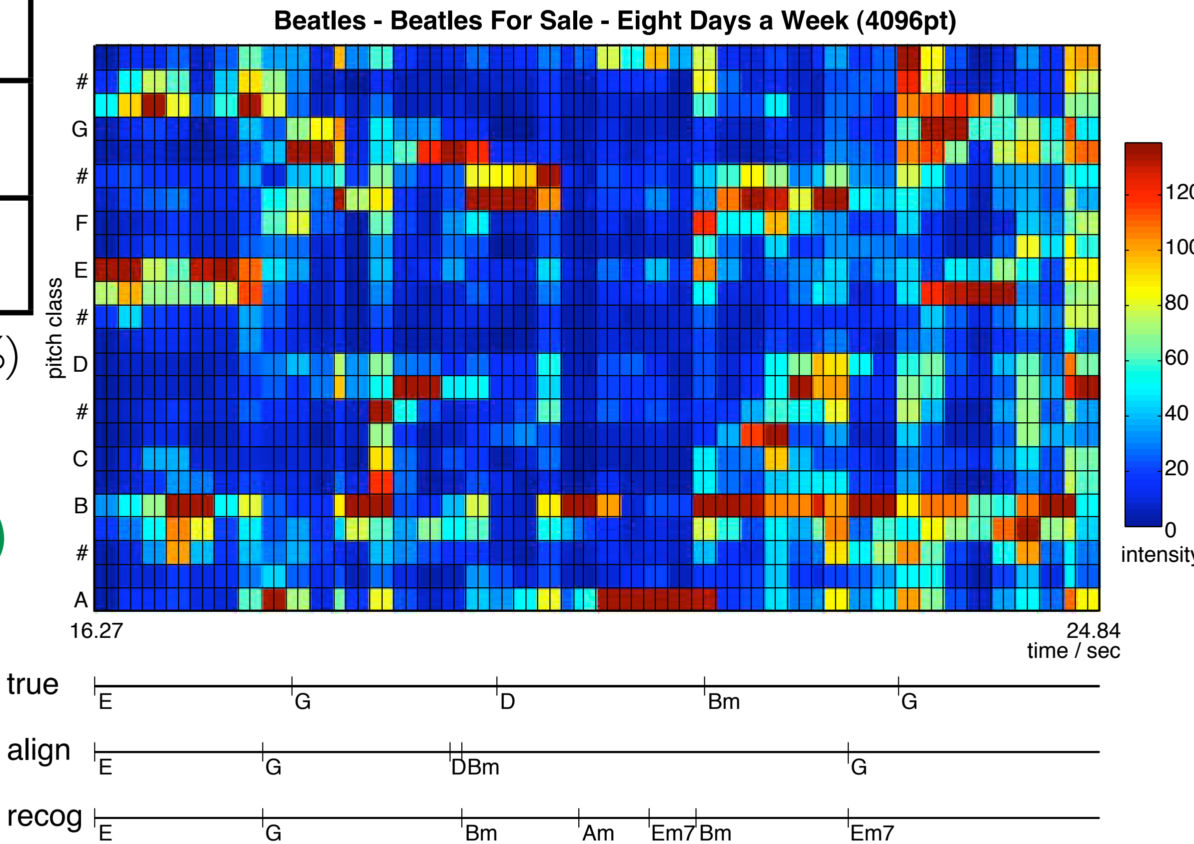
Frame-level Accuracy

Feature	*Recog.*	*Alignment*
MFCC	8.7%	22.0%
PCP_ROT	21.7%	76.0%

(random ~3%)

MFCCs are poor
(can overtrain)

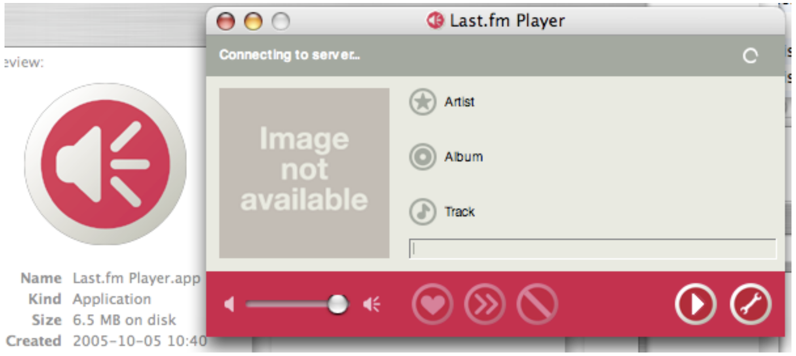
PCPs better
(ROT helps generalization)



- Needed more training data...

3. Music Similarity

- The most central problem...
  - motivates extracting musical information
  - supports real applications (playlists, discovery)
- But do we need content-based similarity?
  - compete with collaborative filtering
  - compete with fingerprinting + metadata

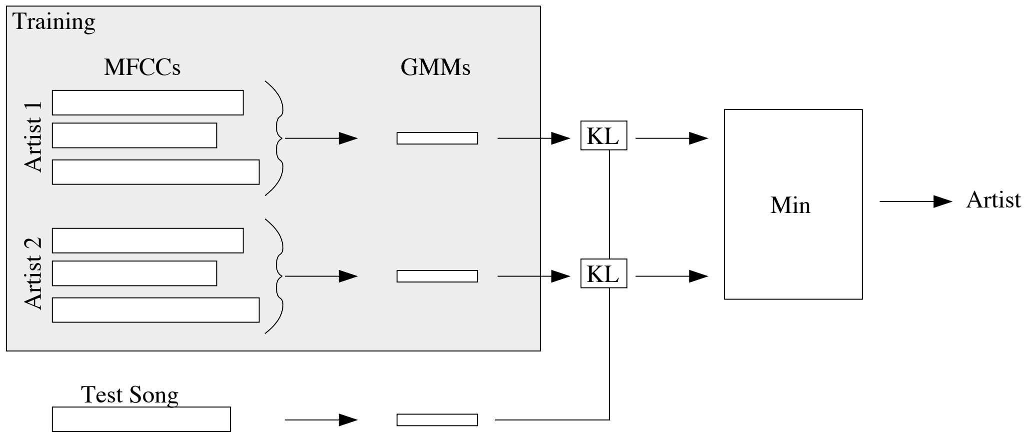


- Maybe ... for the Future of Music
  - connect listeners directly to musicians

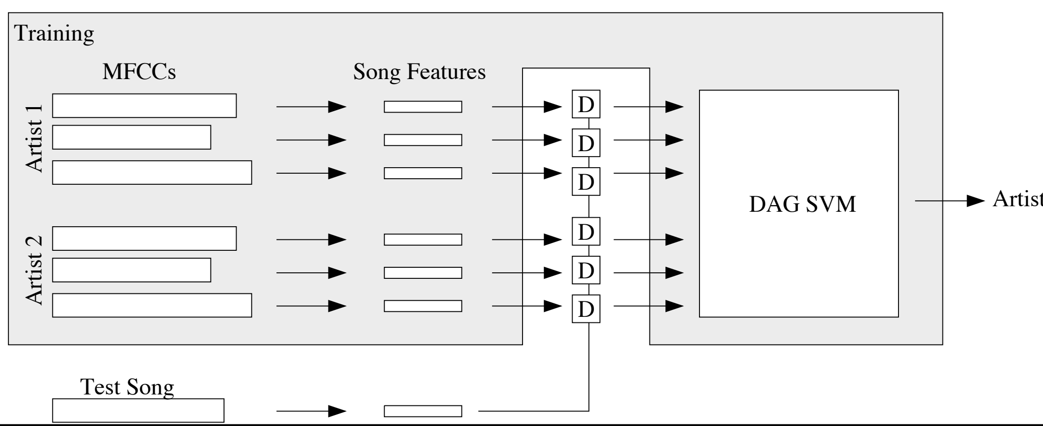
Discriminative Classification

Mandel & Ellis '05

- Classification as a proxy for similarity
- Distribution models...



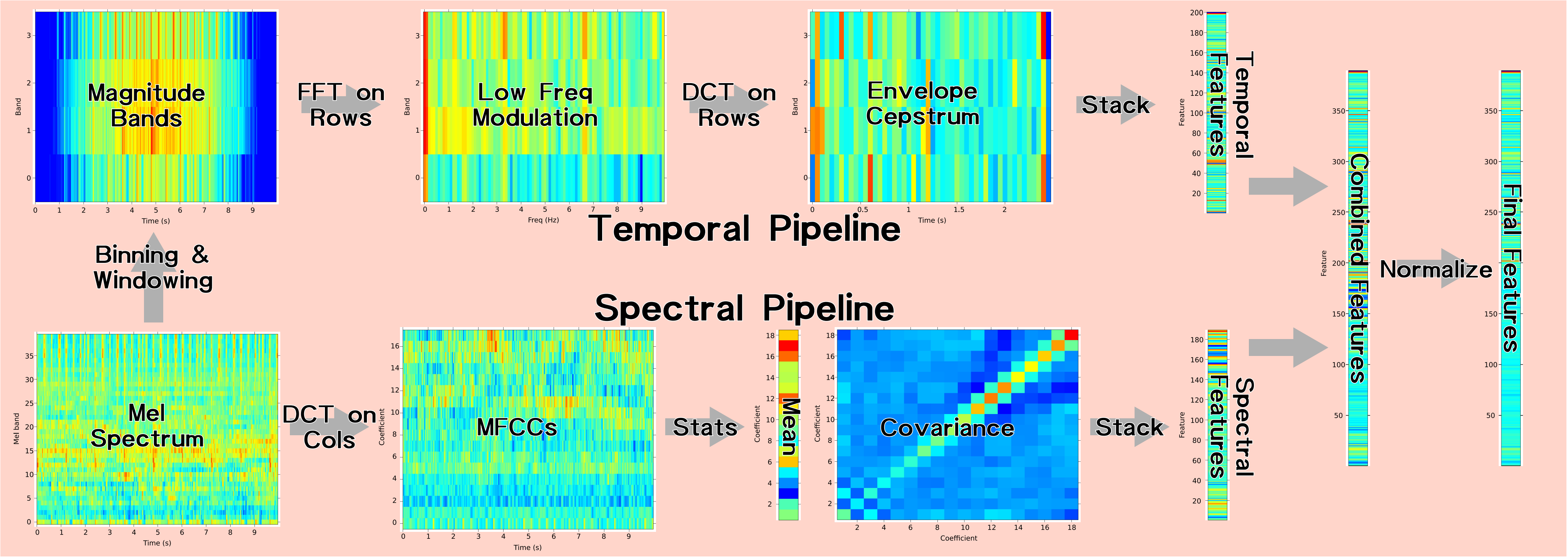
- vs. SVM



Segment-Level Features

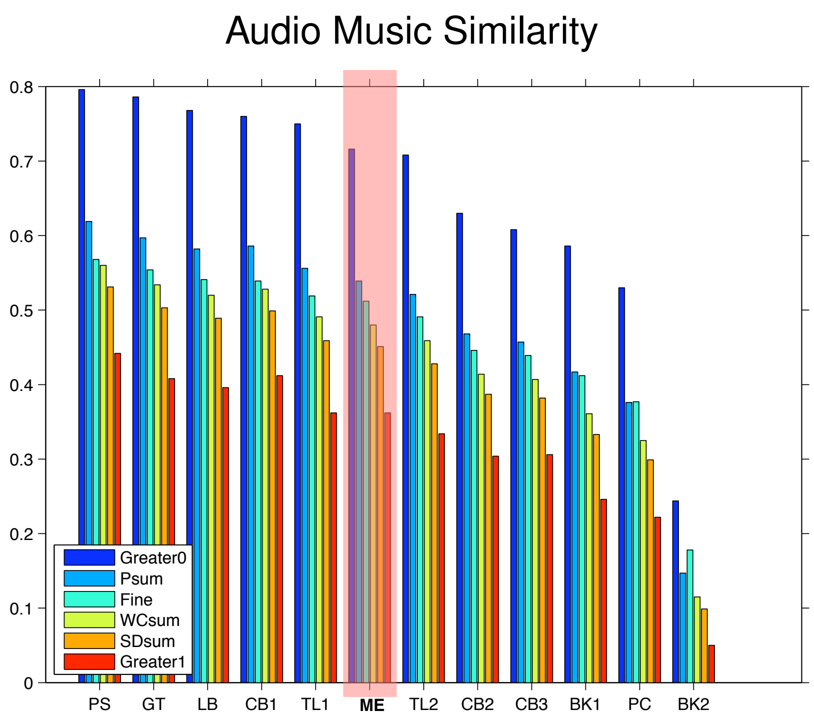
Mandel & Ellis '07

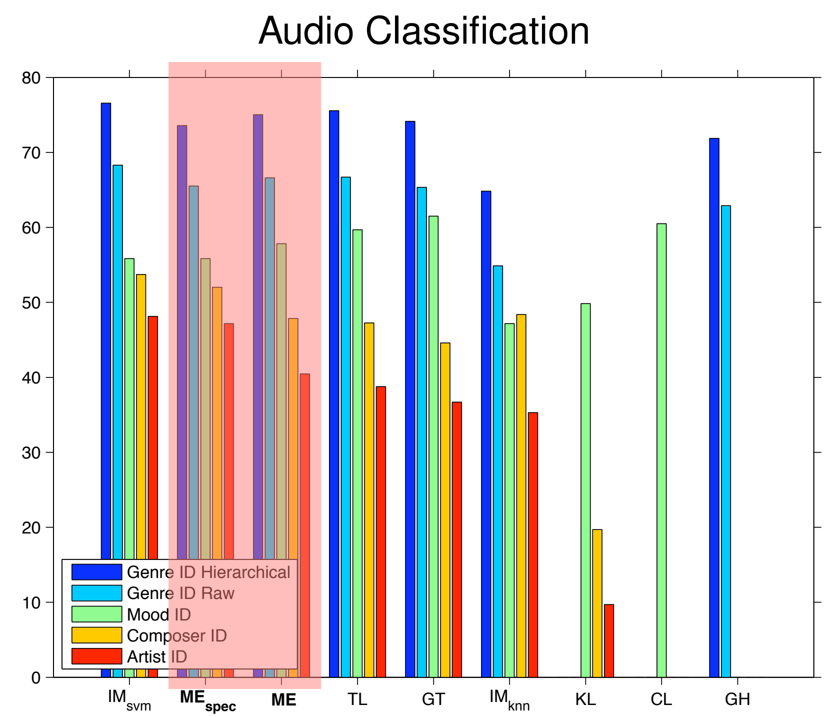
- Statistics of spectra and envelope define a point in feature space
  - for SVM classification, or Euclidean similarity...



MIREX’07 Results

- One system for similarity and classification



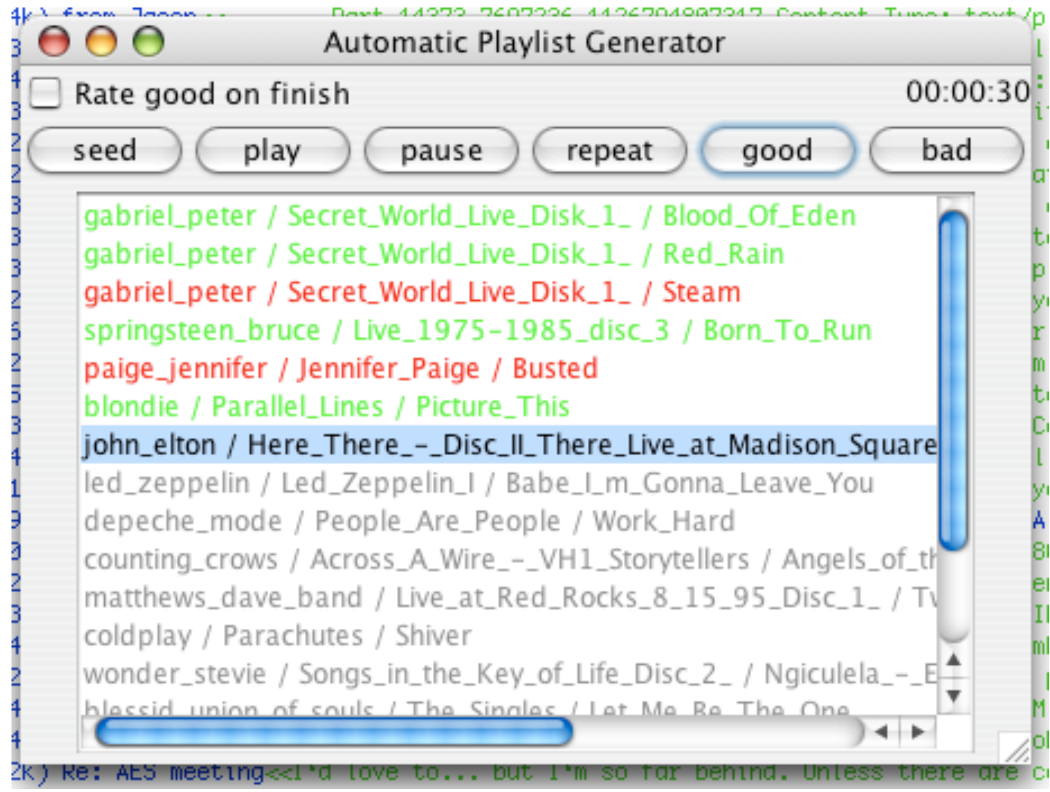


PS = Pohle, Schnitzer; GT = George Tzanetakis; LB = Barrington, Turnbull, Torres, Lanckriet; CB = Christoph Bastuck; TL = Lidy, Rauber, Pertusa, Iñesta; ME = Mandel, Ellis; BK = Bosteels, Kerre; PC = Paradzinets, Chen

IM = IMIRSEL M2K; ME = Mandel, Ellis; TL = Lidy, Rauber, Pertusa, Iñesta; GT = George Tzanetakis; KL = Kyogu Lee; CL = Laurier, Herrera; GH = Guaus, Herrera

Active-Learning Playlists

- 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

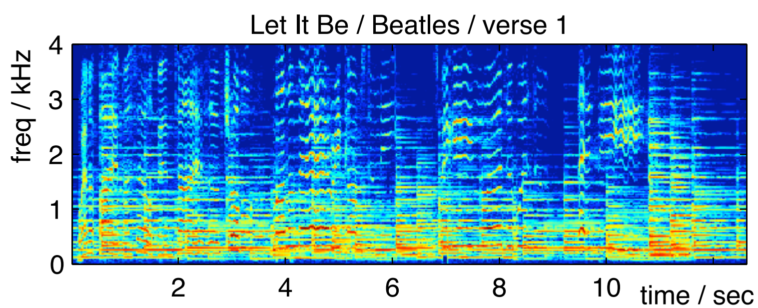


Cover Song Detection

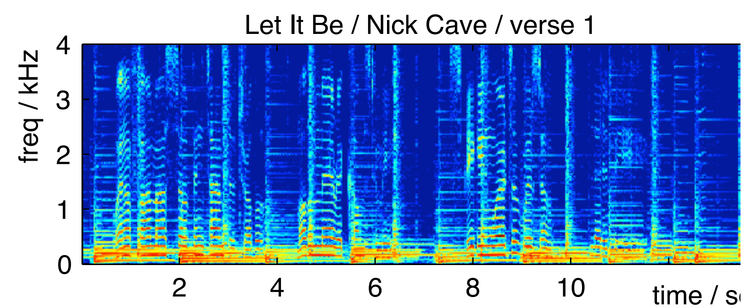
Ellis & Poliner '07

- "Cover Songs" = reinterpretation of a piece
  - different instrumentation, character
  - no match with "timbral" features

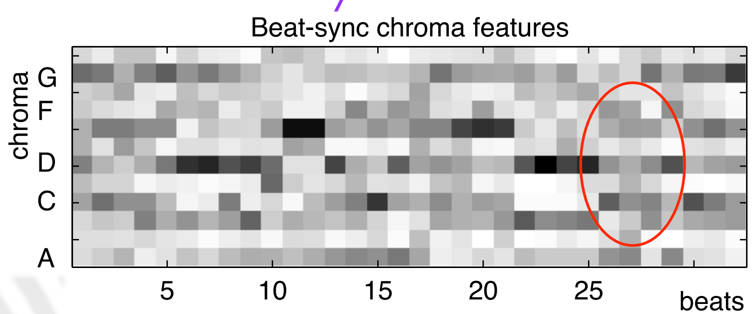
Let It Be - The Beatles

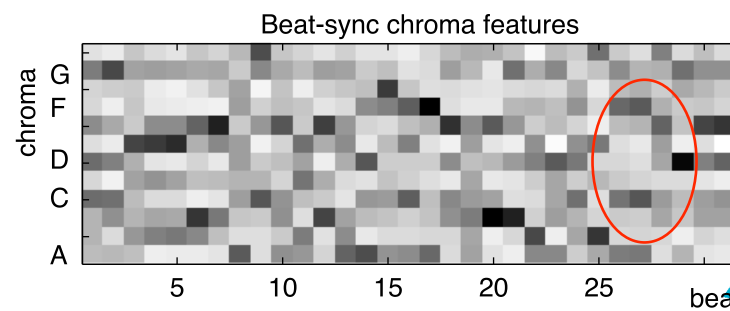


Let It Be - Nick Cave



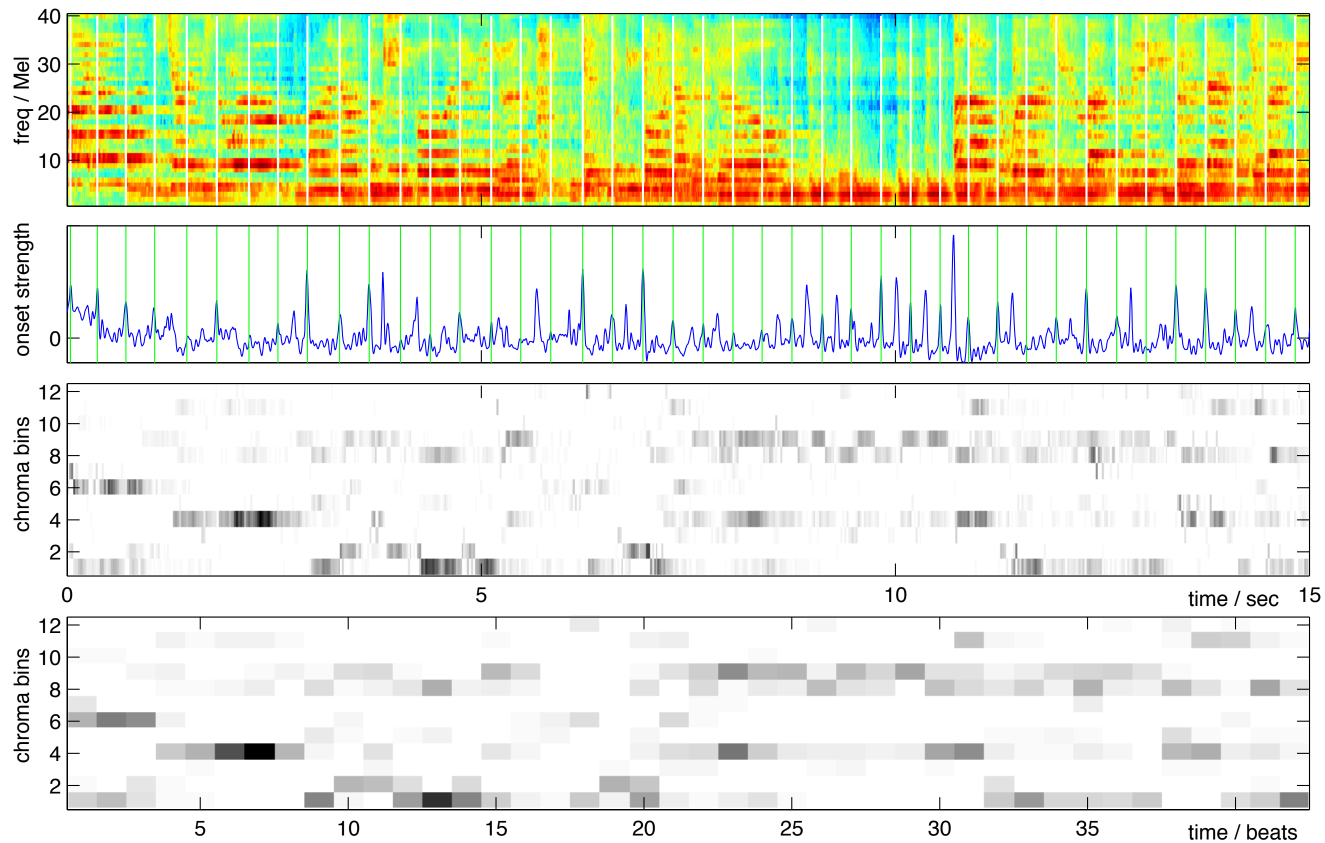
- Need a different representation!
  - beat-synchronous chroma features





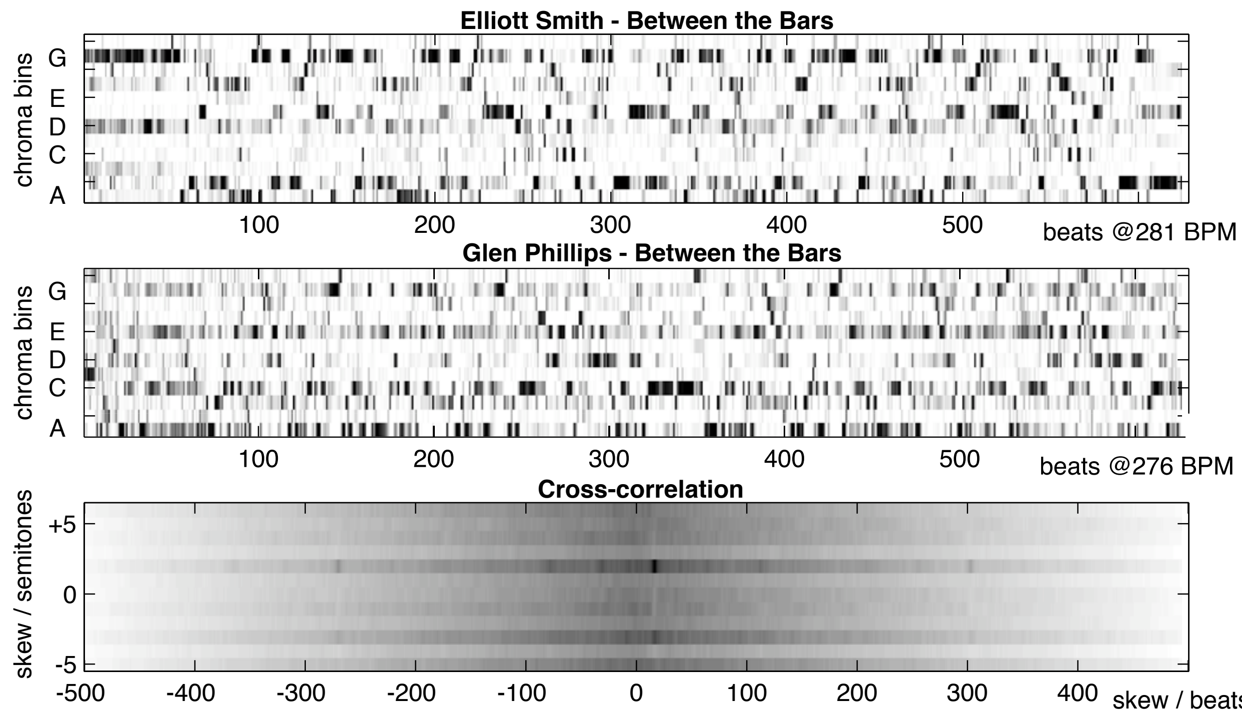
Beat-Synchronous Chroma Features

- Beat + chroma features / 30ms frames
  → average chroma within each beat
  - compact; sufficient?



Matching: Global Correlation

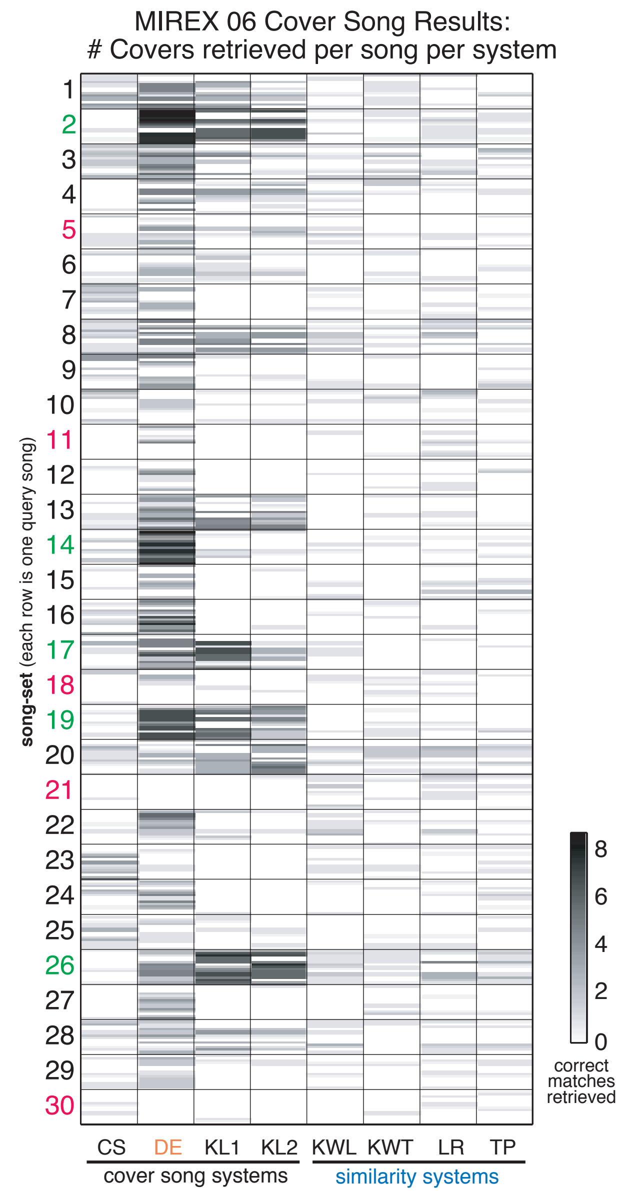
- Cross-correlate *entire* beat-chroma matrices
  - ... at all possible transpositions
  - implicit combination of match quality and duration



- One good matching fragment is sufficient...?

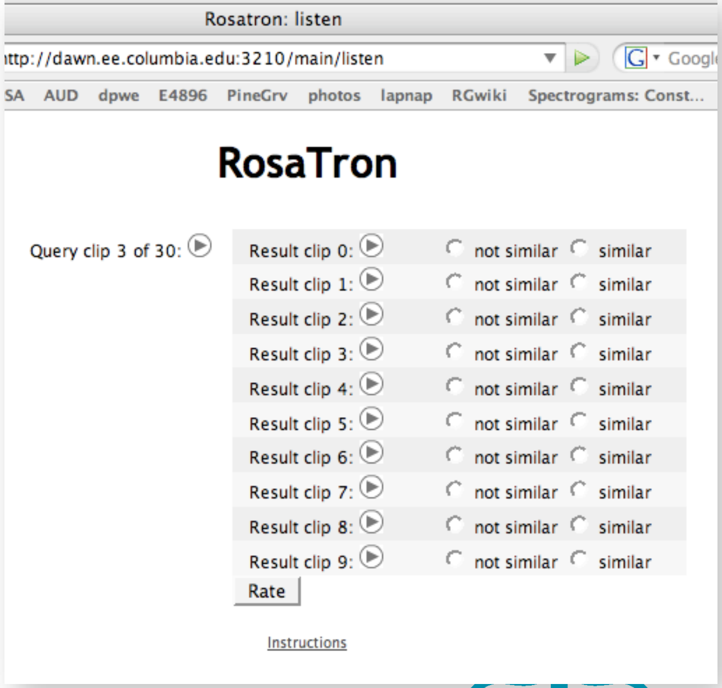
MIRROR 06 Results

- **Cover song contest**
  - 30 songs x 11 versions of each (!)
  - (data has not been disclosed)
  - # true covers in top 10
  - 8 systems compared (4 cover song + 4 similarity)
- **Found 761/3300 = 23% recall**
  - next best: 11% guess: 3%



Cross-Correlation Similarity

- **Use cover-song approach to find similarity**
  - e.g. similar note/instrumentation sequence
  - may sound very similar to judges
- **Numerous variants**
  - try on chroma (melody/harmony) and MFCCs (timbre)
  - try full search (xcorr) or landmarks (indexable)
  - compare to random, segment-level stats
- **Evaluate by subjective tests**
  - modeled after MIREX similarity



Cross-Correlation Similarity

- Human web-based judgments
  - binary judgments for speed
  - 6 users x 30 queries x 10 candidate returns

Algorithm	Similar count
(1) Xcorr, chroma	48/180 = 27%
(2) Xcorr, MFCC	48/180 = 27%
(3) Xcorr, combo	55/180 = 31%
(4) Xcorr, combo + tempo	34/180 = 19%
(5) Xcorr, combo at boundary	49/180 = 27%
(6) Baseline, MFCC	81/180 = 45%
(7) Baseline, rhythmic	49/180 = 27%
(8) Baseline, combo	**88/180 = 49%**
Random choice 1	22/180 = 12%
Random choice 2	28/180 = 16%

- Cross-correlation inferior to baseline...
  - ... but is getting somewhere, even with 'landmark'

Cross-Correlation Similarity

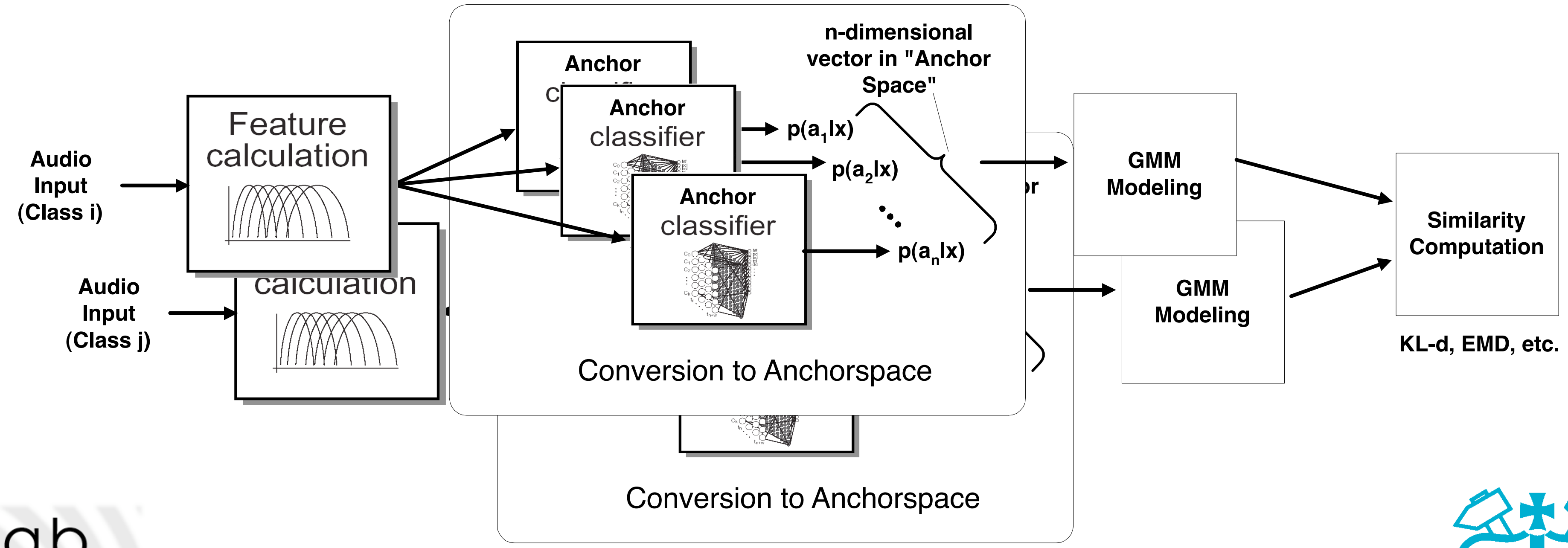
- Results are not overwhelming
  - .. but database is only a few thousand clips

Too Much Dave Matthews Band	Too Much Dave Matthews Band -0.00	Erotica Madonna -0.00	Don t Tell Me Madonna -0.00	Waiting Madonna -0.00	Where Life Begins Madonna -0.00	Did You Do Madonna -0.00
Hey Nineteen Steely Dan	Hey Nineteen Steely Dan -0.00	Where Life Begins Madonna -0.00	Erotica Madonna -0.00	Don t Tell Me Madonna -0.00	Now I m Following You Part II Madonna -0.00	Too Much Dave Matthews -0.00
Little 15 Depeche Mode	Little 15 Depeche Mode -0.00	Don t Tell Me Madonna -0.00	Lolita Suzanne Vega -0.00	Where Life Begins Madonna -0.00	Macy s Day Parade Green Day -0.00	Seconds U2 -0.00
The Same Deep Water As You Cure	The Same Deep Water As You Cure -0.00	Scarlet U2 -0.00	Breathing in fumes Depeche Mode -0.00	Where Life Begins Madonna -0.00	Erotica Madonna -0.00	Try Just A Li Harder Roxette -0.00
Scarlet U2	Scarlet U2 -0.00	The Same Deep Water As You Cure -0.00	Rollin Garth Brooks -0.00	small talk Roxette -0.01	In the Light Led Zeppelin -0.01	I m Sorry Roxette -0.01
Flying Beatles	Flying Beatles -0.00	Breathing in fumes Depeche Mode -0.00	Keep It Together Madonna -0.00	Where Life Begins Madonna -0.00	Erotica Madonna -0.00	Let s Preten Married Prince -0.00
Breathing in fumes Depeche Mode	Breathing in fumes Depeche Mode -0.00	Flying Beatles -0.00	Where Life Begins Madonna -0.00	Erotica Madonna -0.00	I Wish U Heaven Prince -0.00	Dragon Atta Bonus Remi Queen -0.00
Bad Moon Rising Creedence Clearwater Revival	Bad Moon Rising Creedence Clearwater Revival -0.00	Let s Pretend We re Married Prince -0.00	Don t look now Creedence Clearwater Revival -0.00	Cry Baby Madonna -0.00	Fashion Victim Green Day -0.00	Shiver And Cure -0.00

"Anchor Space"

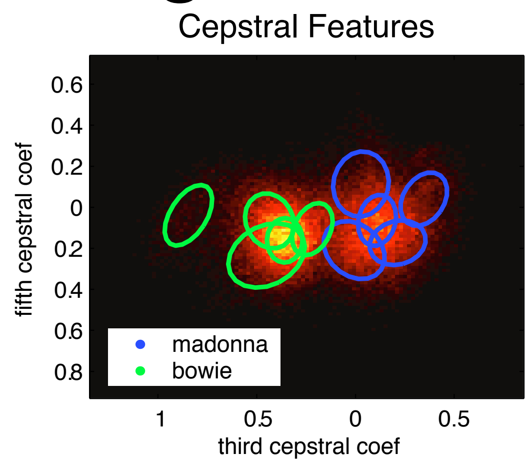
Berenzweig & Ellis '03

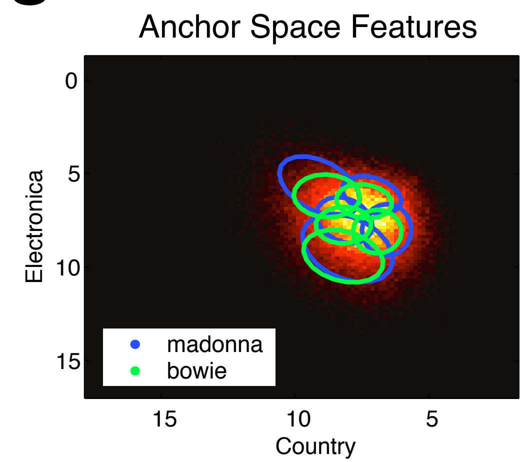
- Acoustic features describe each song
  - .. but from a signal, not a perceptual, perspective
  - .. and not the differences between songs
- Use genre classifiers to define new space
  - prototype genres are "anchors"



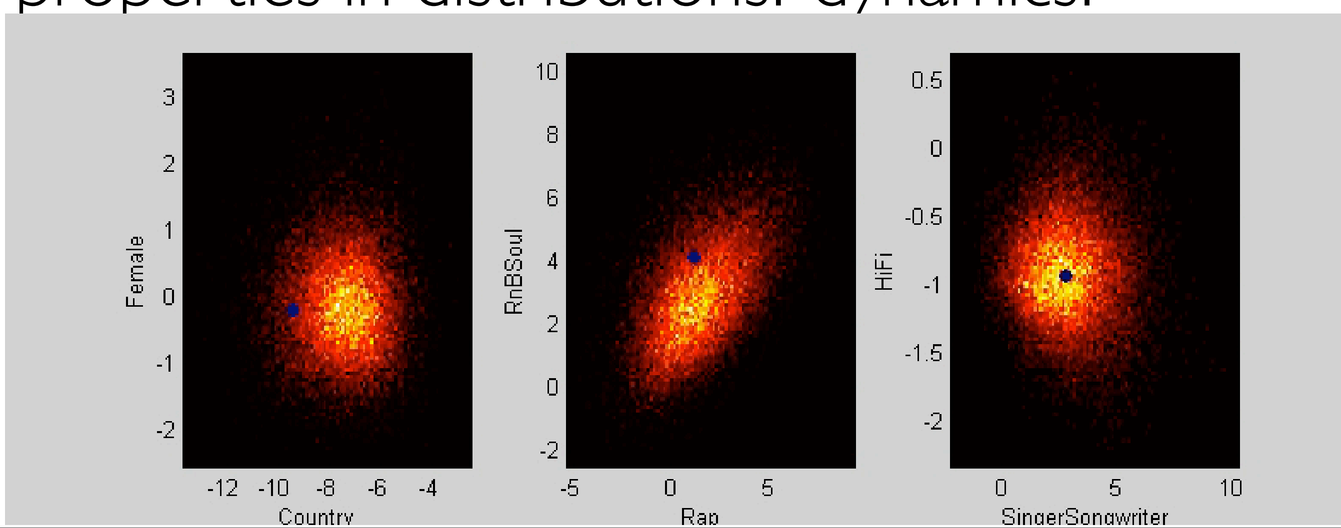
"Anchor Space"

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

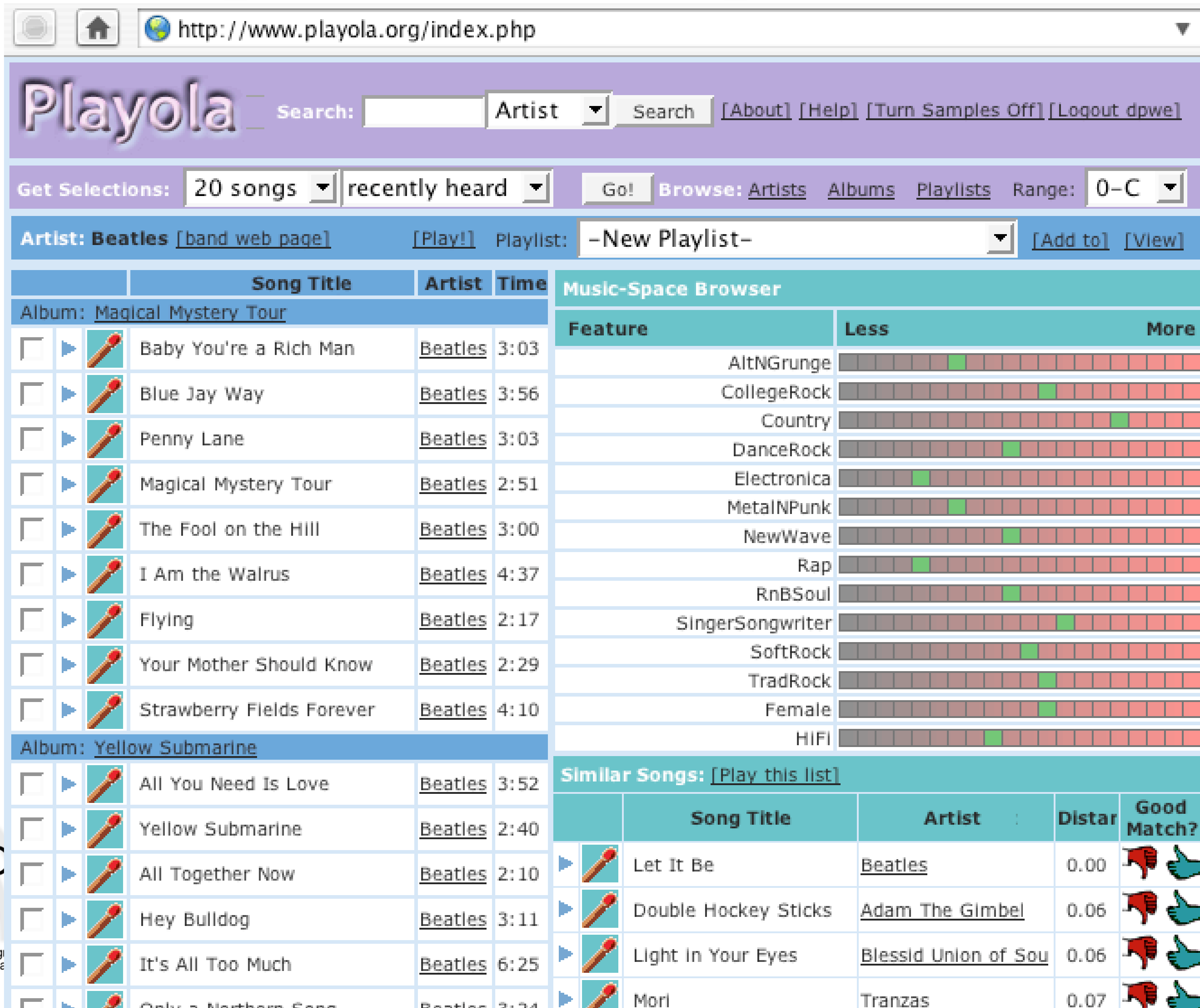




  - properties in distributions? dynamics?



‘Playola’ Similarity Browser

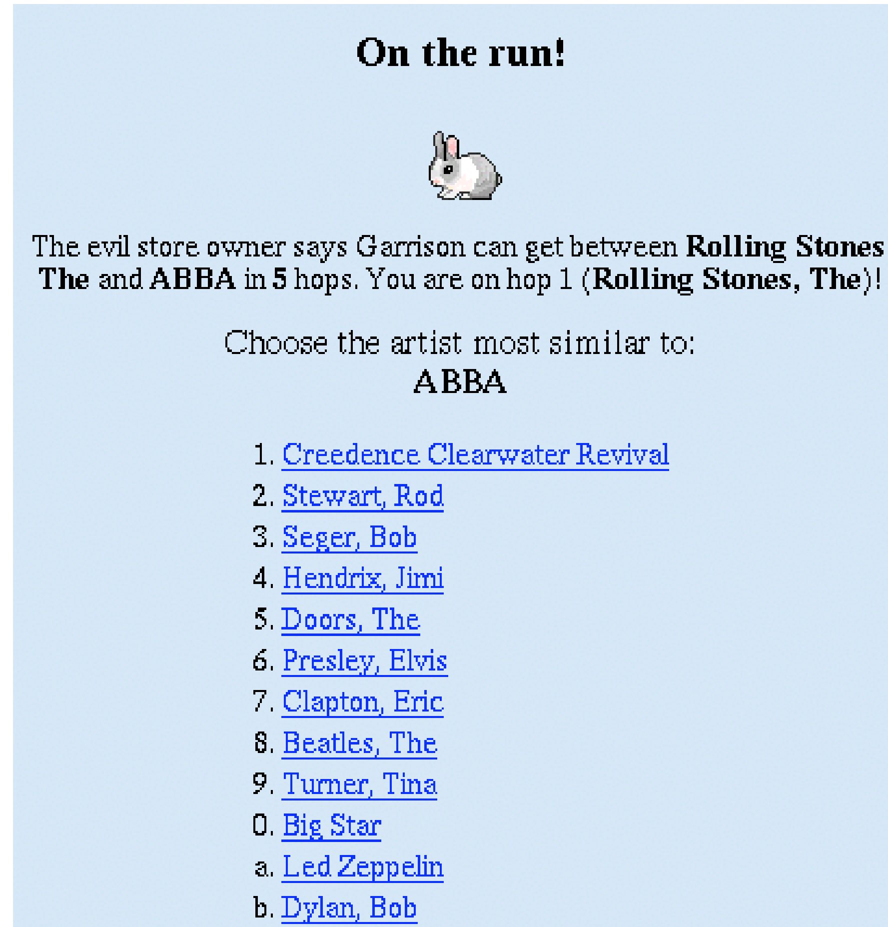


Ground-truth data

Ellis et al, '02

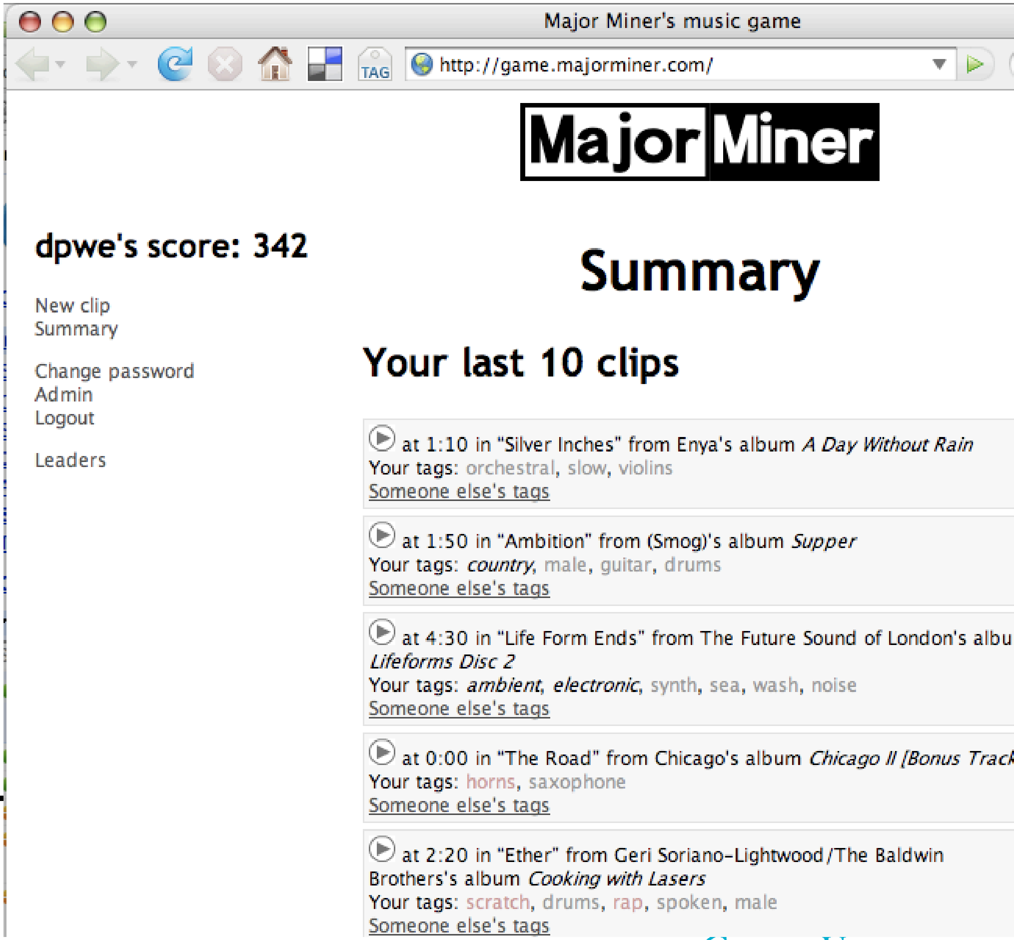
- Hard to evaluate Playola's 'accuracy'
  - user tests...
  - ground truth?

- "Musicseer" online survey/game:
  - ran for 9 months in 2002
  - > 1,000 users,
    > 20k judgments
  - http://labrosa.ee.columbia.edu/projects/musicsim/



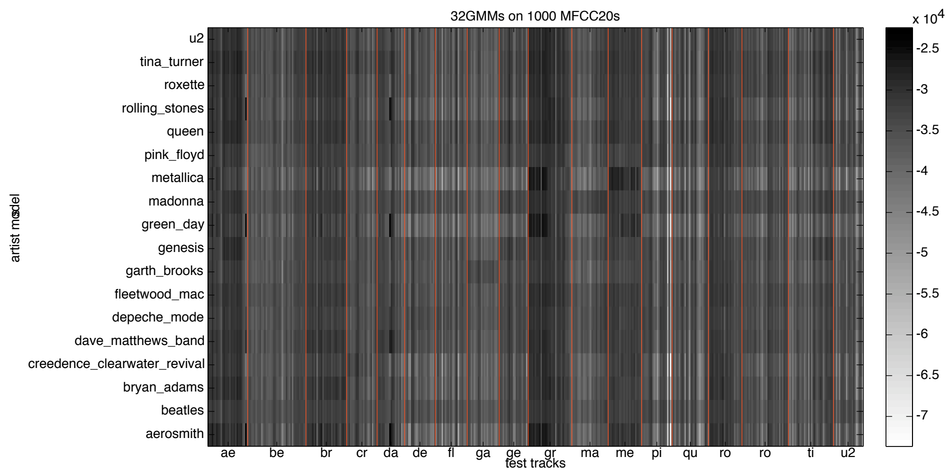
"Semantic Bases"

- Describe segment in human-relevant terms
  - e.g. anchor space, but more so
- Need ground truth...
  - what words to people use?
- MajorMiner game:
  - 400 users
  - 7500 unique tags
  - 70,000 taggings
  - 2200 10-sec clips used
- Train classifiers...



3. Music Structure Discovery

- Use the many examples to map out the "manifold" of music audio
    - ... and hence define the subset that is music

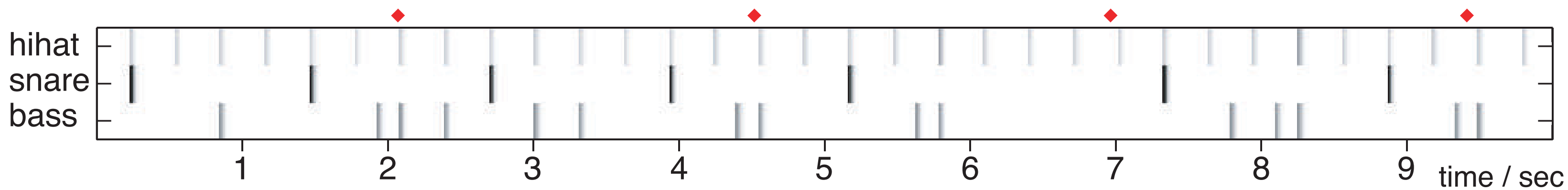


- Problems
    - alignment/registration of data
    - factoring & abstraction
    - separating parts?

Eigenrhythms: Drum Pattern Space

Ellis & Arroyo '04

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



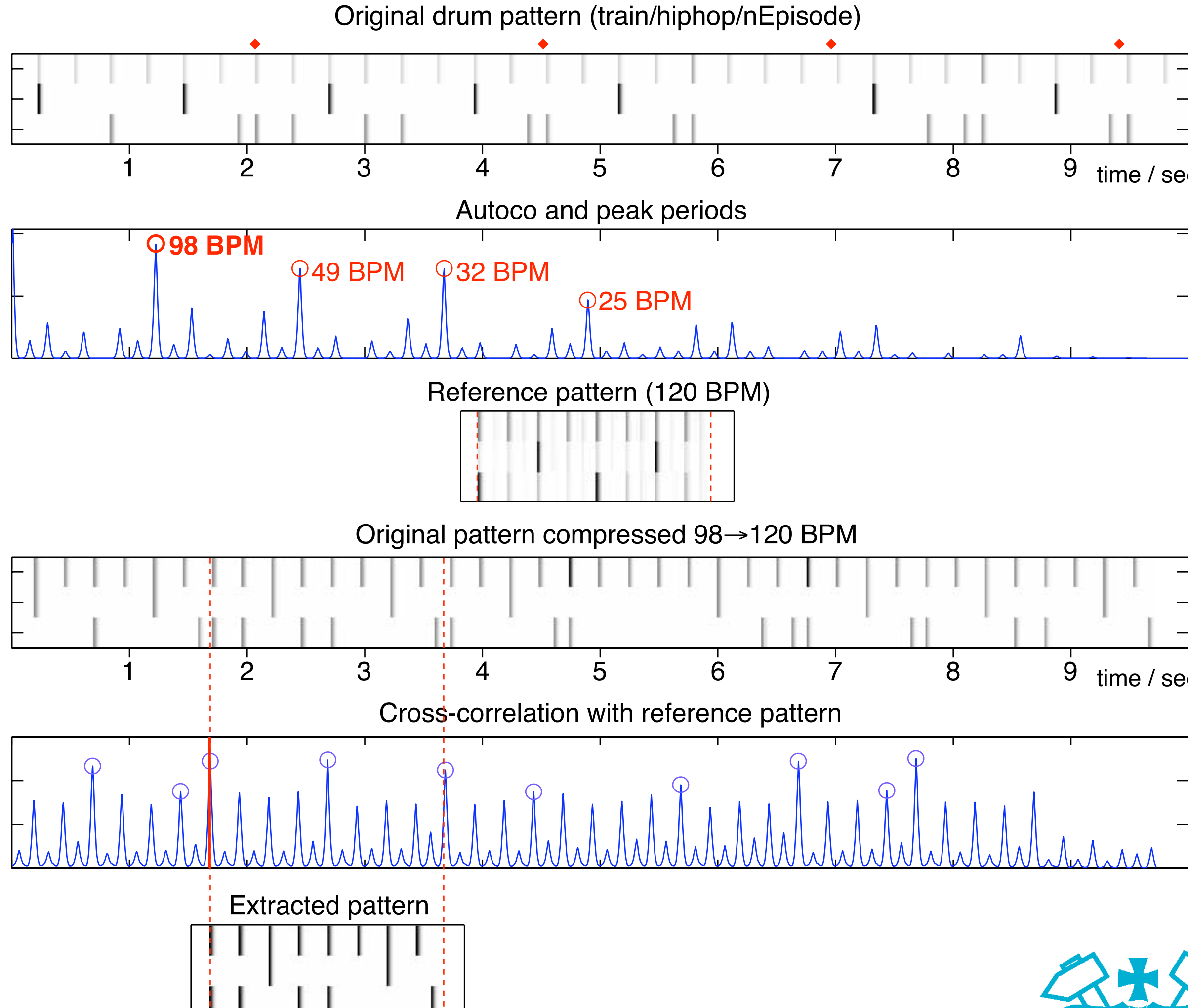
- Eigen-analysis (or ...) to capture variations?
  - by analyzing lots of (MIDI) data, or from audio
- Applications
  - music categorization
  - "beat box" synthesis
  - insight

Aligning the Data

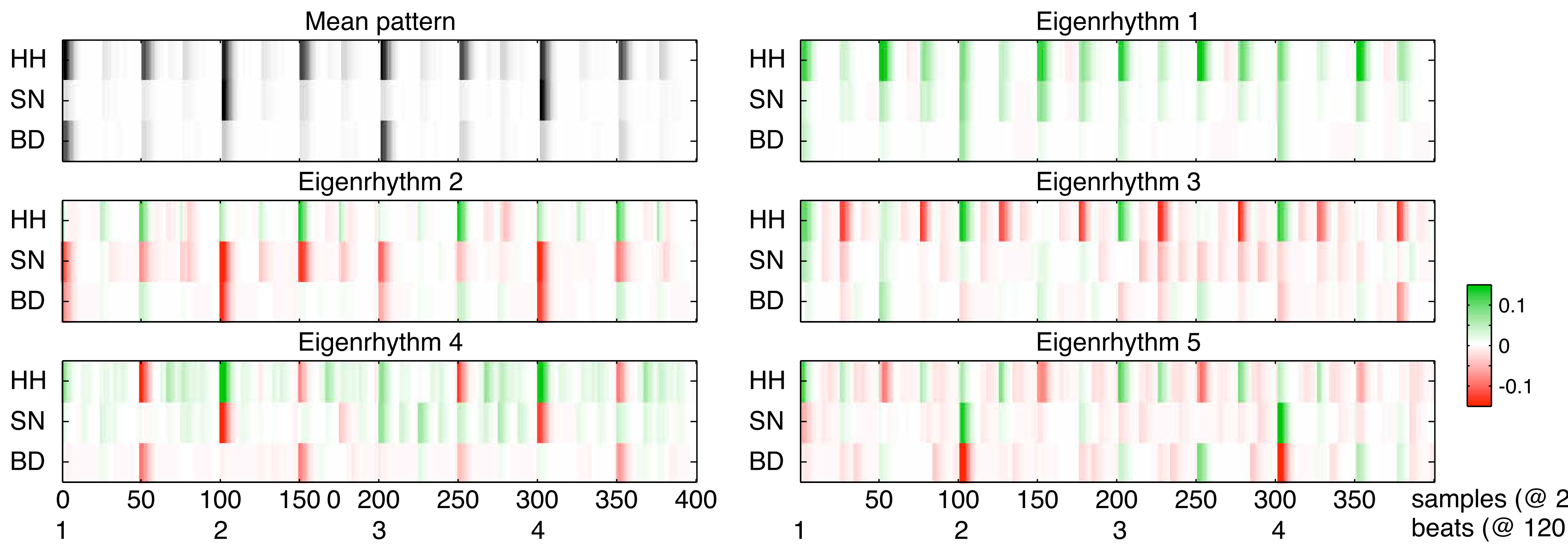
- Need to align patterns prior to modeling...

tempo (stretch):
by inferring BPM &
normalizing

downbeat (shift):
correlate against
'mean' template

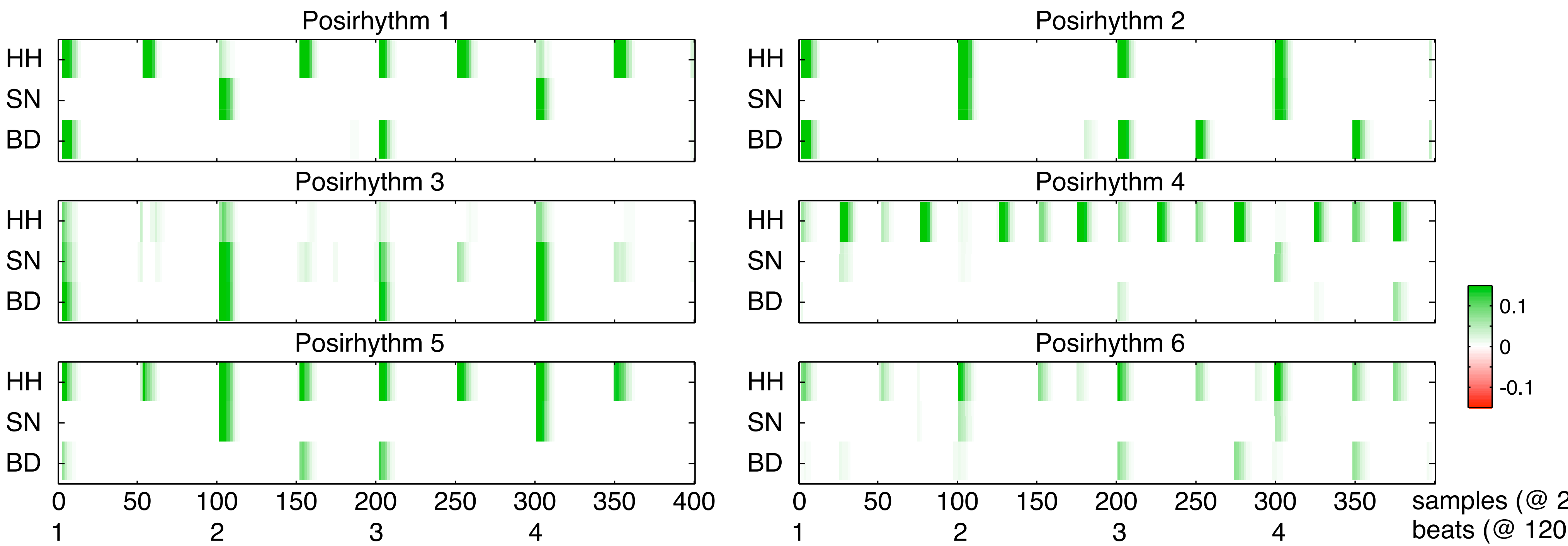


Eigenrhythms (PCA)



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

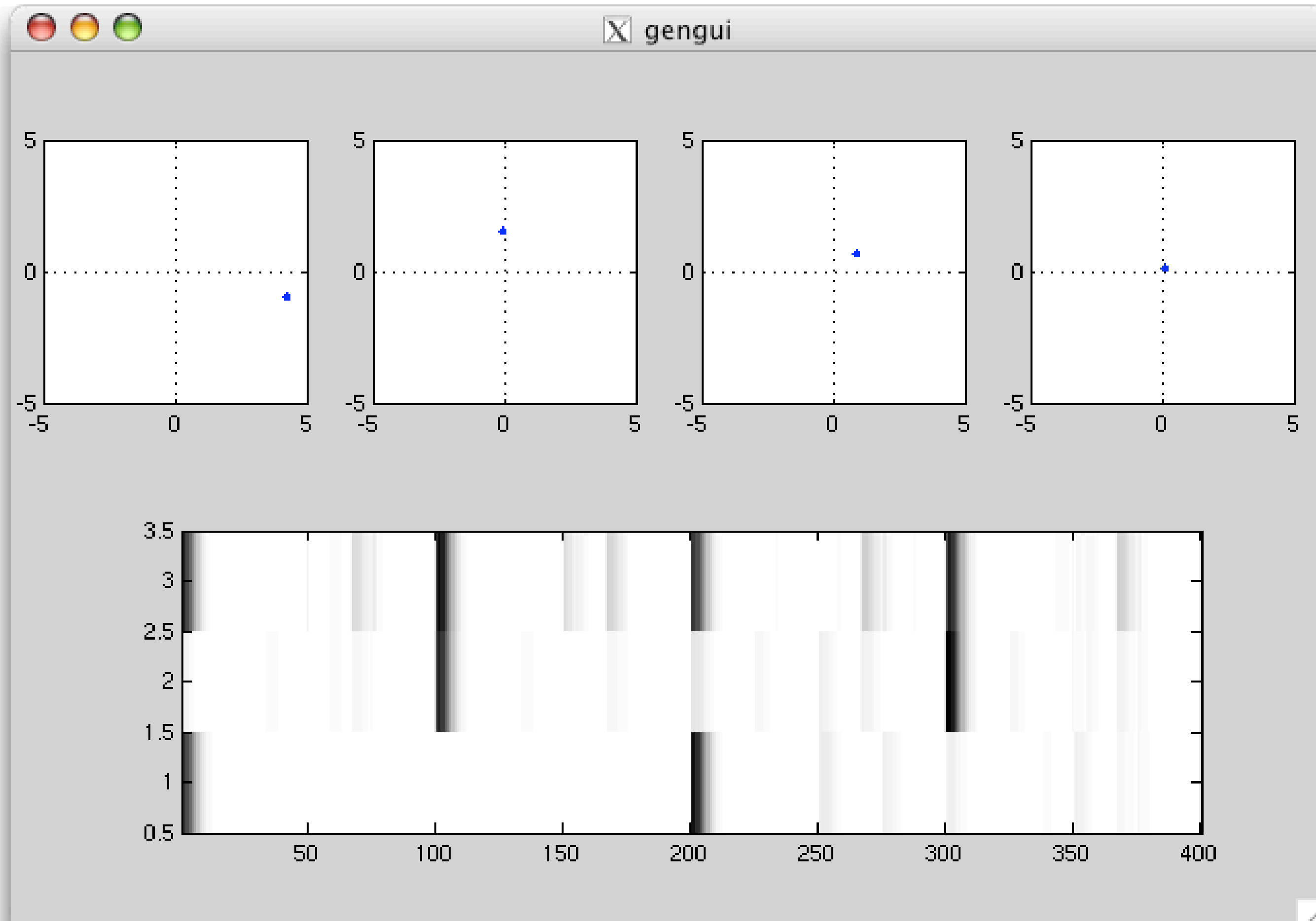
Posirhythms (NMF)



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

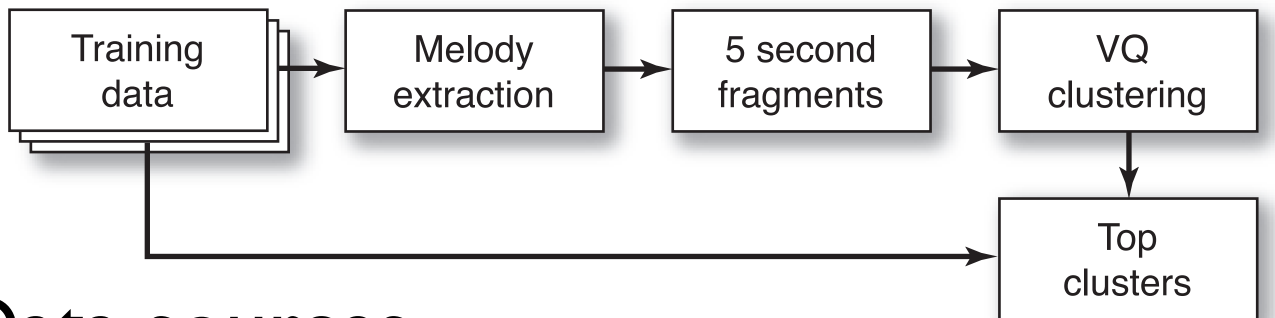
Eigenrhythm BeatBox

- Resynthesize rhythms from eigen-space



Melody Clustering

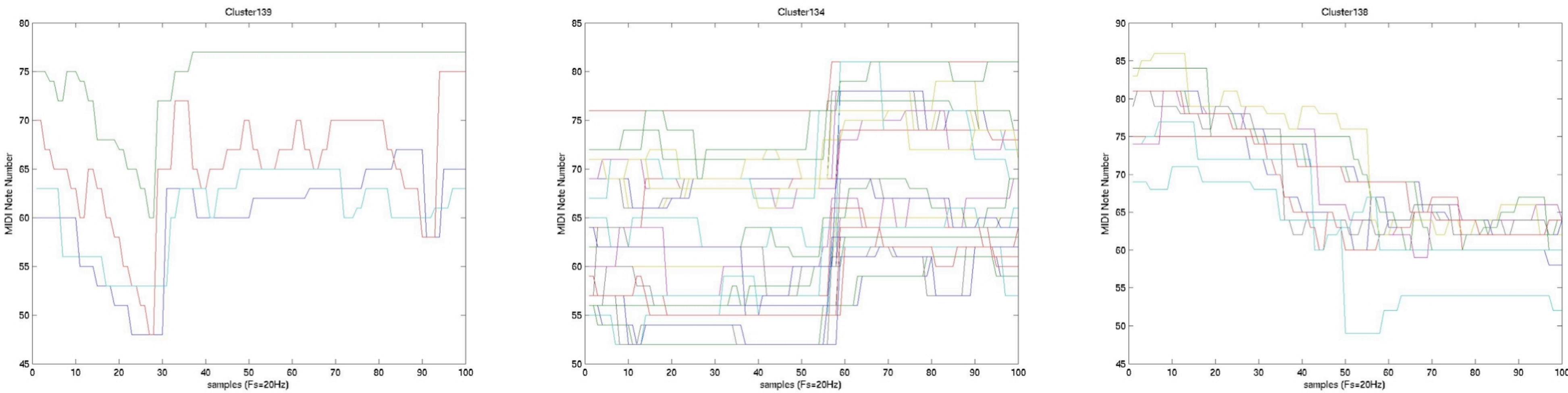
- Goal: Find 'fragments' that recur in melodies
  - .. across large music database
  - .. trade data for model sophistication



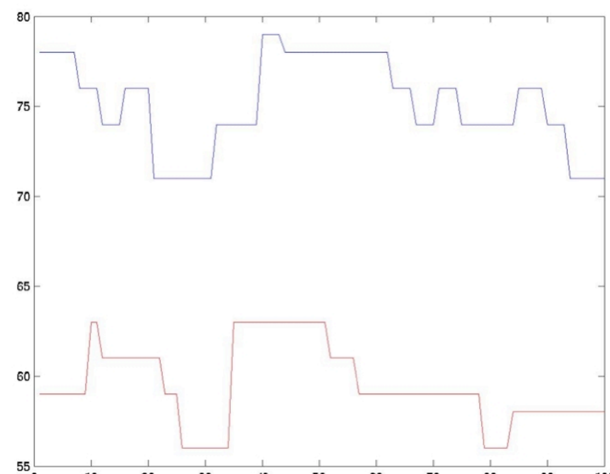
- Data sources
  - pitch tracker, or MIDI training data
- Melody fragment representation
  - DCT(1:20) - removes average, smoothes detail

Melody Clustering

- Clusters match underlying contour:

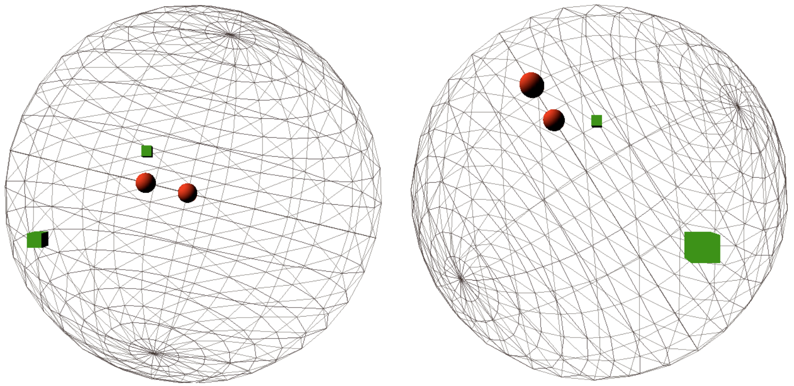


- Some interesting matches:
  - e.g. Pink + Nsync

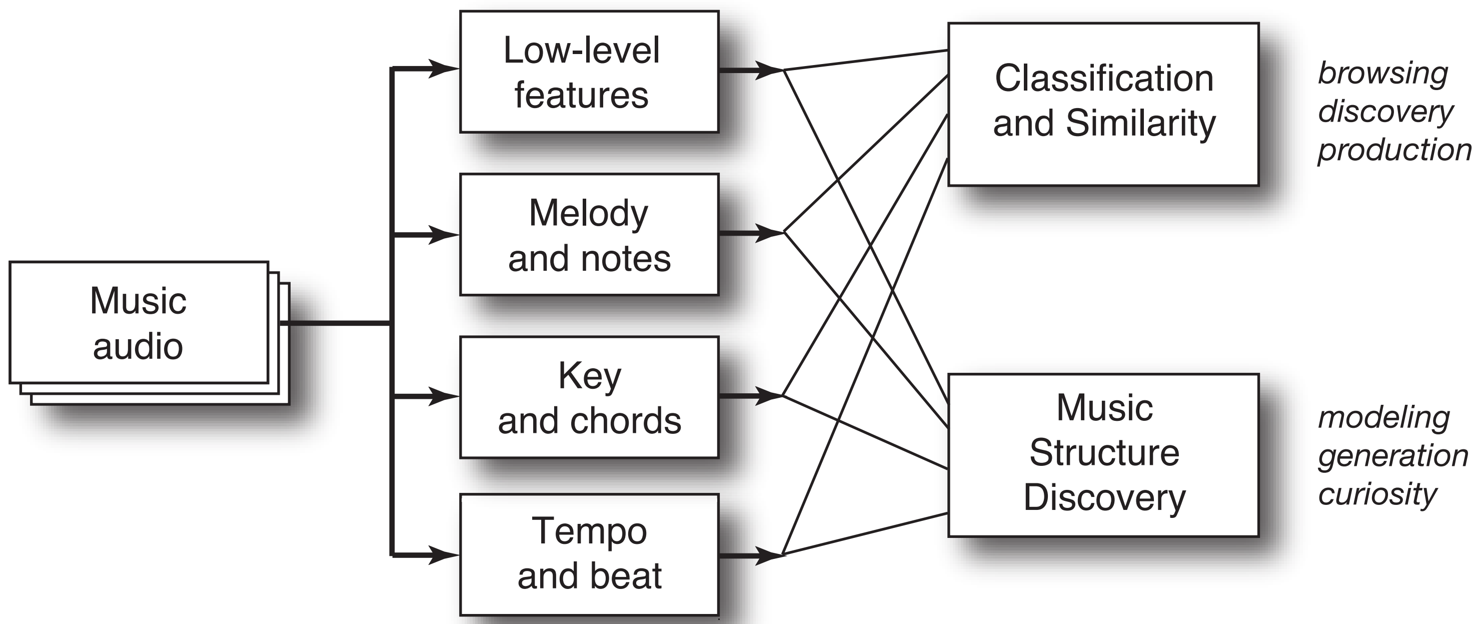


Beat-Chroma Fragment Codebook

- **Idea: Find the very popular music fragments**
  - e.g. perfect cadence, rising melody, ...?
- **Clustering a large enough database should reveal these**
  - but: registration of phrase boundaries, transposition
- **Need to deal with really large datasets**
  - e.g. 100k+ tracks, multiple landmarks in each
  - but: Locality Sensitive Hashing can help - quickly finds 'most' points in a certain radius
- **Experiments in progress...**



Conclusions



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