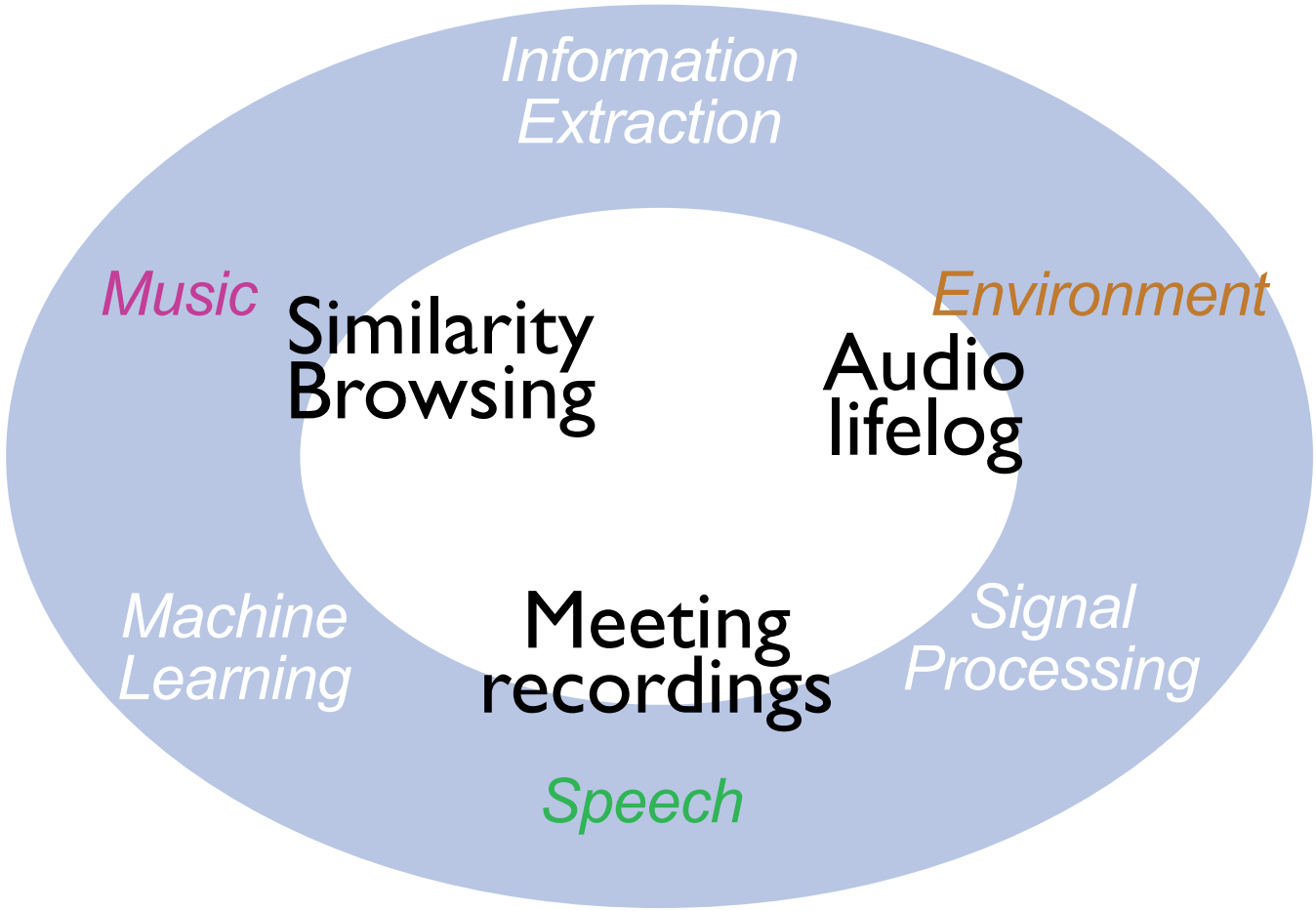
Searching and Describing Audio Databases

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

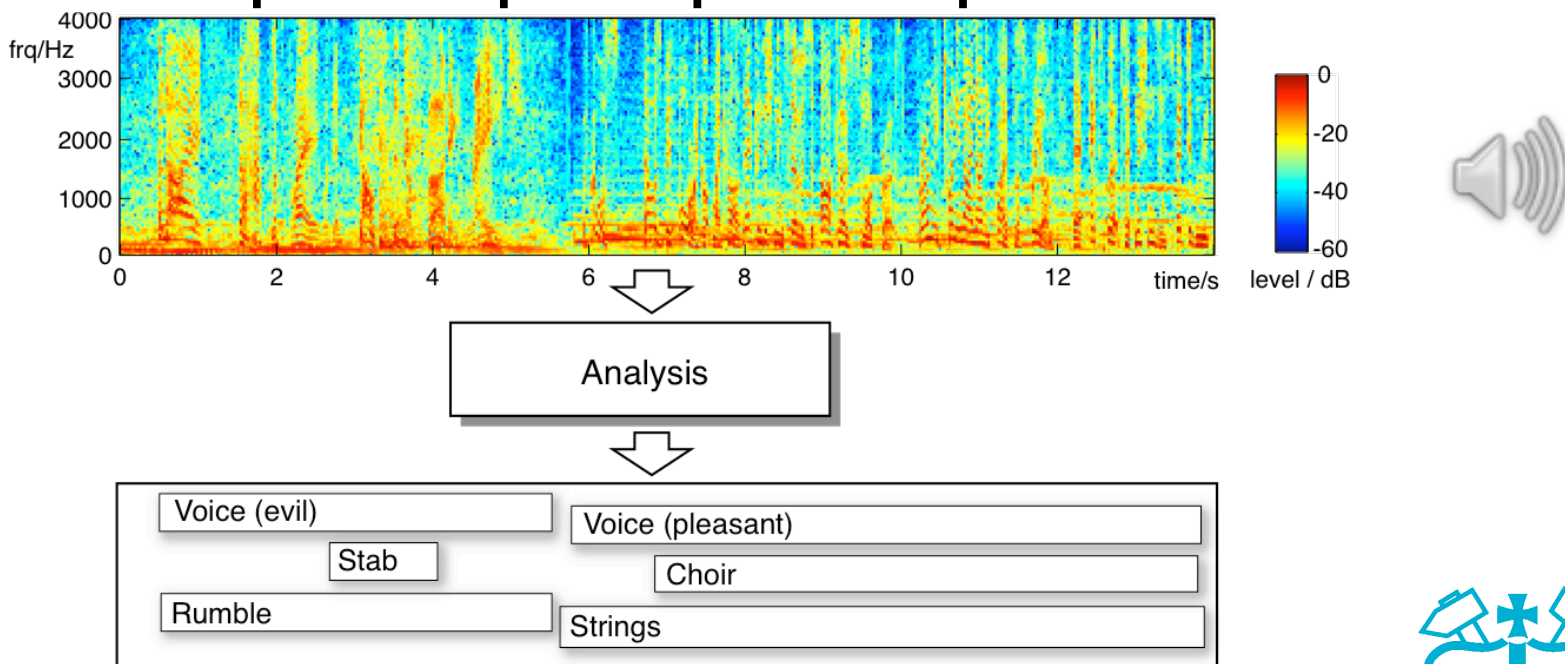
1. Music browsing
2. Meeting recordings
3. 'Personal Audio'

LabROSA Overview



Sound Description

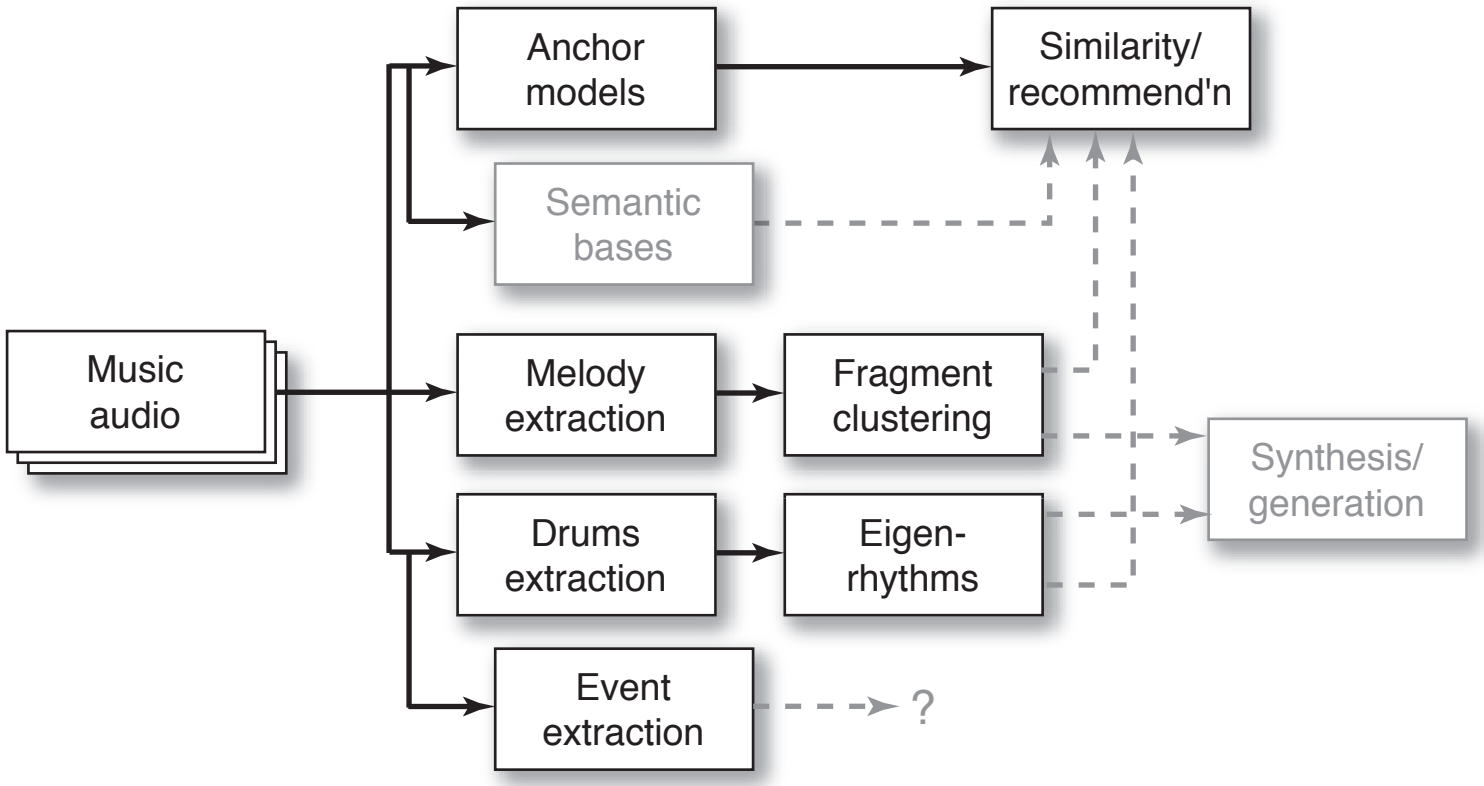
- **Challenge: Indexing continuous content**
  - what are the equivalents of words and phrases?
- **Perception:**
  - sounds as discrete events
- **Goal: Duplicate perceptual representation**



1. Music Signal Analysis

- A lot of music data available
  - e.g. 60G of MP3
    ≈ 1000 hr of audio/15k tracks
- What can we do with it?
  - identify implicit structure...
- Quality vs. quantity
  - Speech recognition lesson:
    10x data, 1/10th annotation, twice as useful
- Motivating Applications
  - music search / browsing by similarity
  - insight into music

Musical Information Extraction

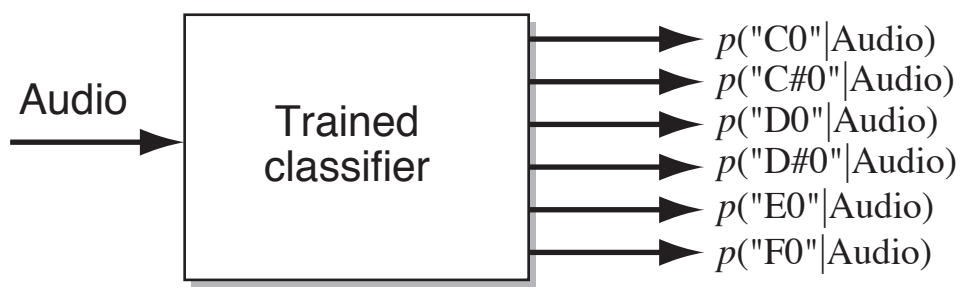


- Lots of data
  + noisy transcription
  + weak clustering
  ⇒ musical insights?

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:



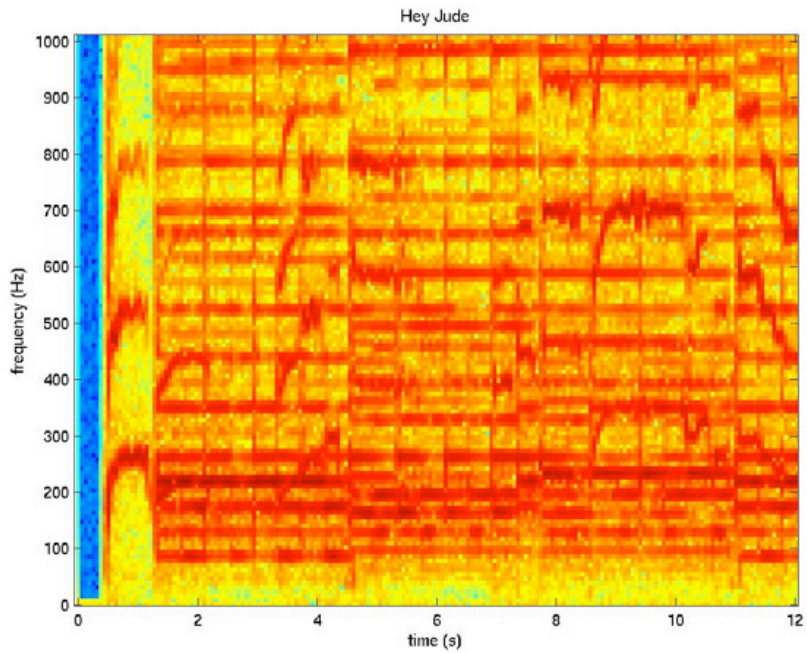
  - single N-way discrimination for "melody"
  - per-note classifiers for polyphonic transcription

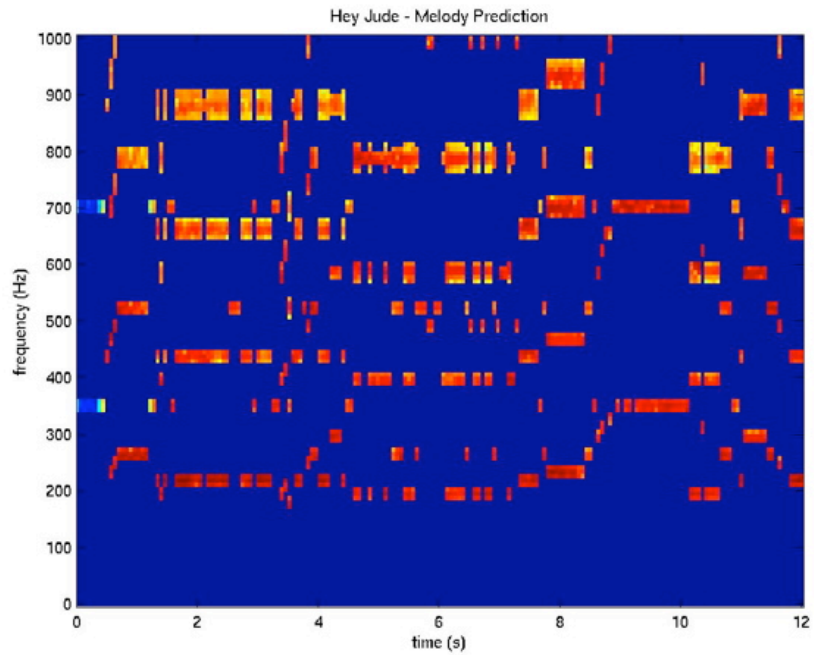
Classifier Transcription Results

- **Trained on MIDI syntheses (32 songs)**
  - SMO SVM (Weka)
- **Tested on MIREX set**
  - foreground/bg separation

Frame-level pitch concordance

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

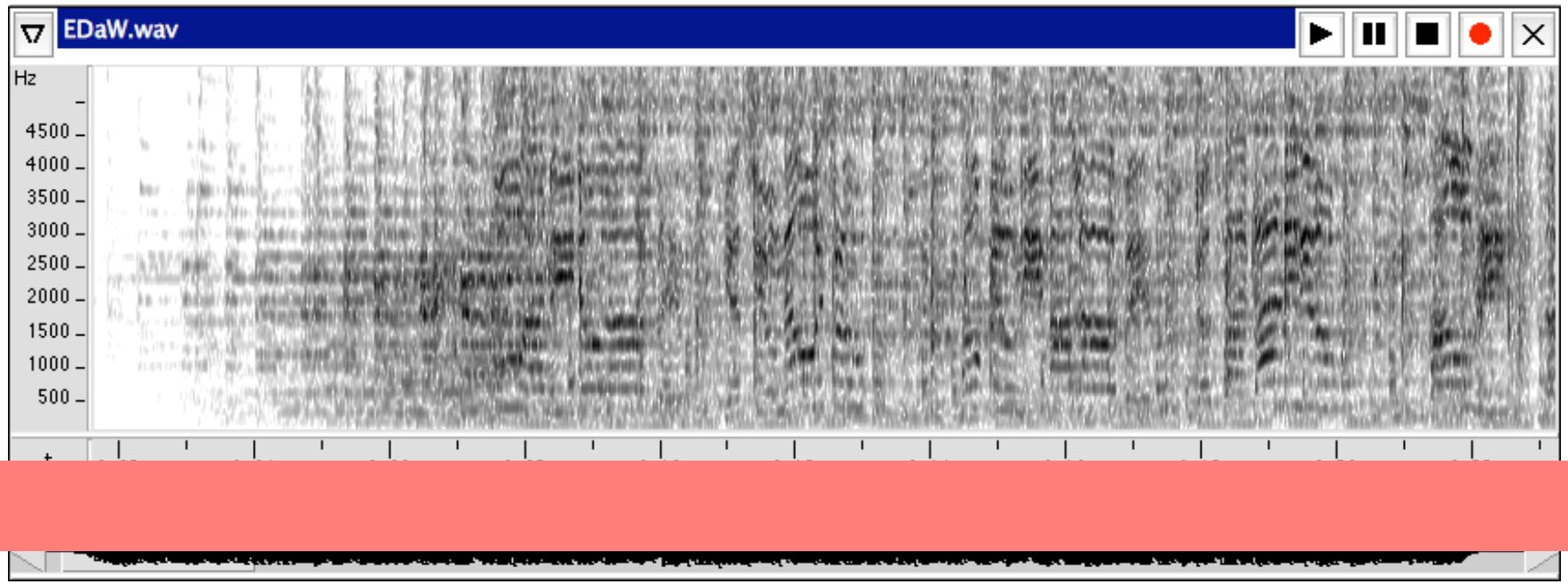




Chord Transcription

with Alex Sheh

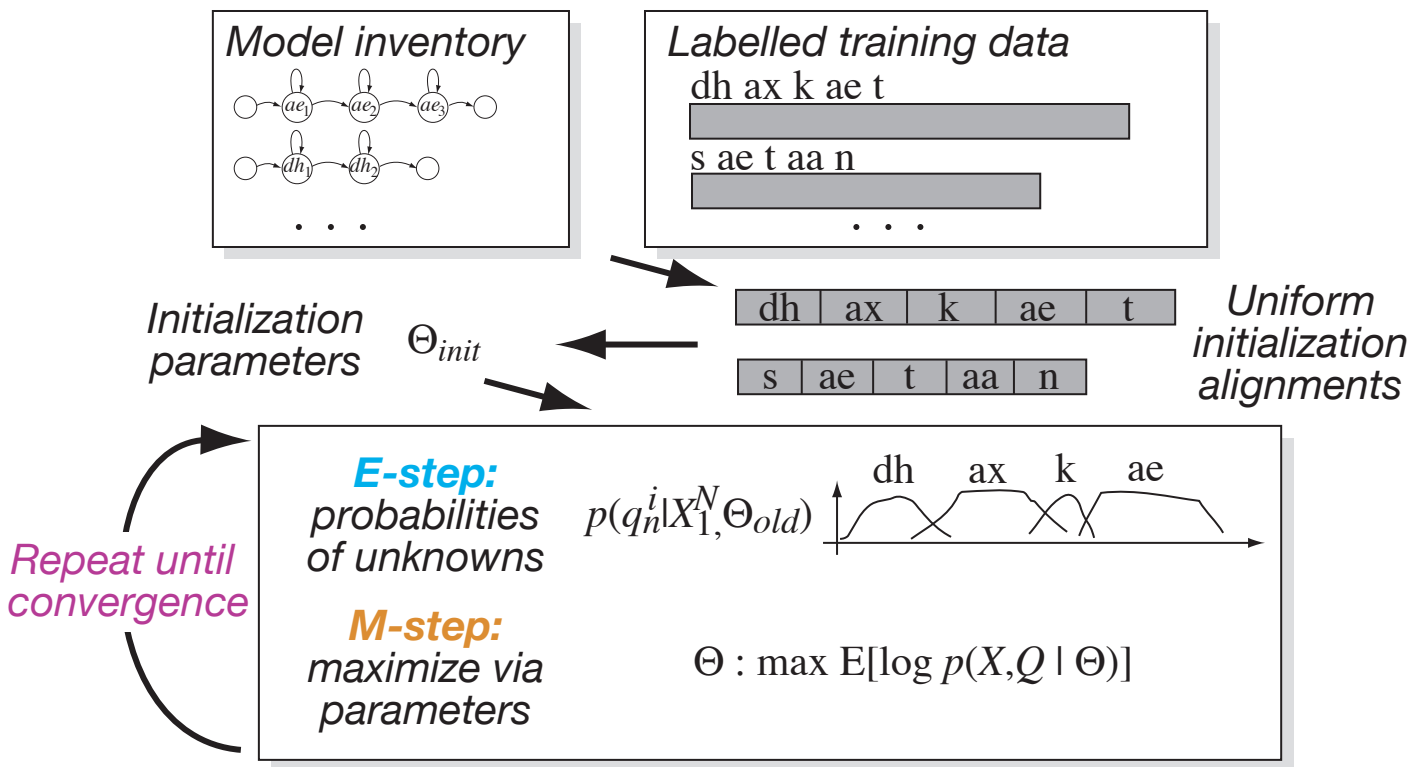
- Basic problem:
  Recover chord sequence labels from audio



  - Easier than note transcription ?
  - More relevant to listener perception ?

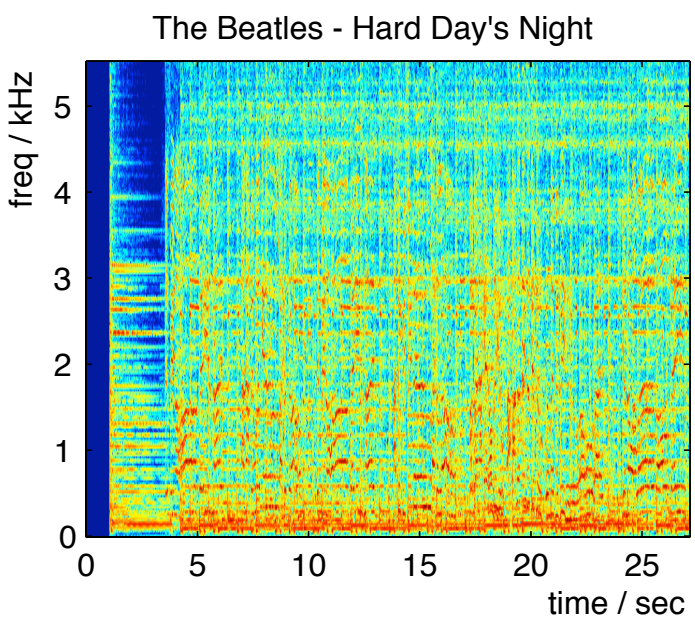
EM HMM Re-Estimation

- Estimate 'soft' labels using current models
- Update model parameters from new labels
- Repeat until convergence to local maximum



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

Chord Results

- Recognition weak, but forced-alignment OK

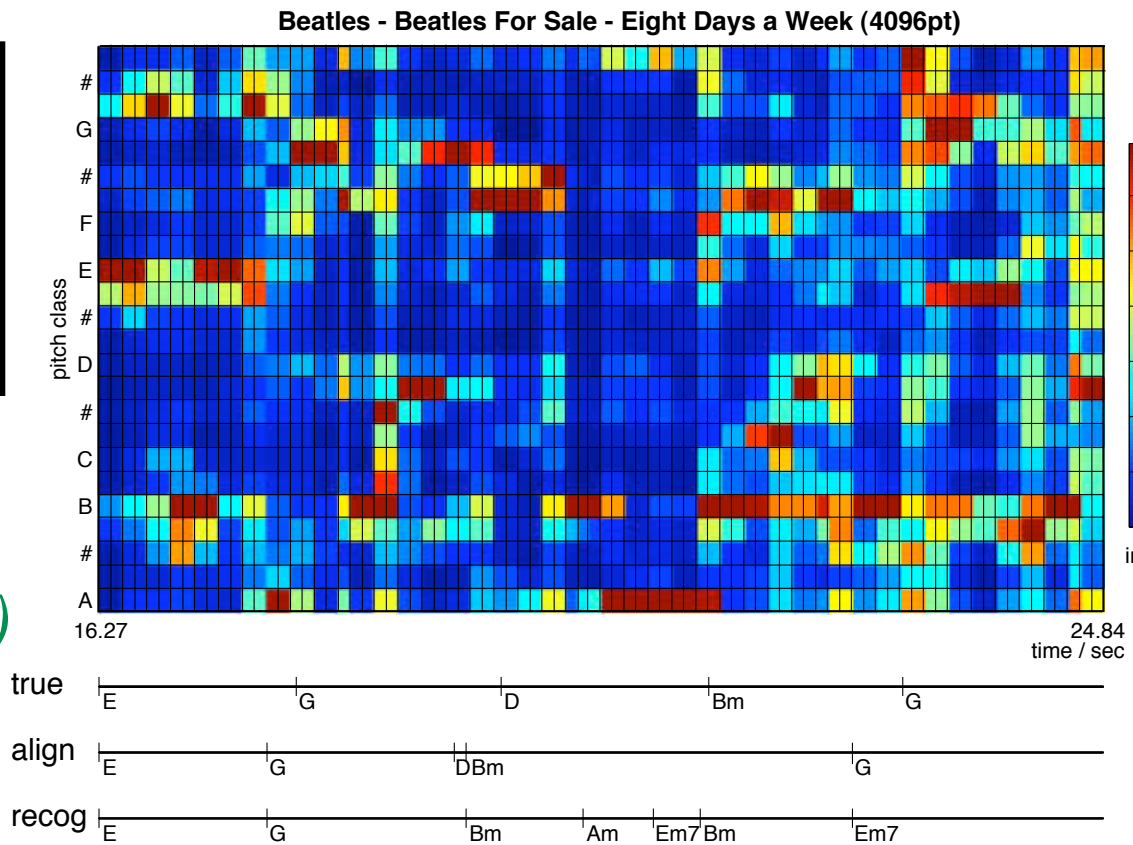
Frame-level Accuracy

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

(random ~3%)

MFCCs are poor (can overtrain)

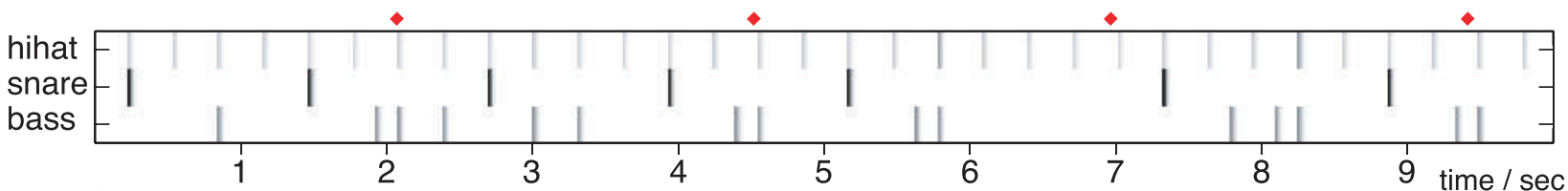
PCPs better (ROT helps generalization)



Eigenrhythms: Drum Pattern Space

with John Arroyo

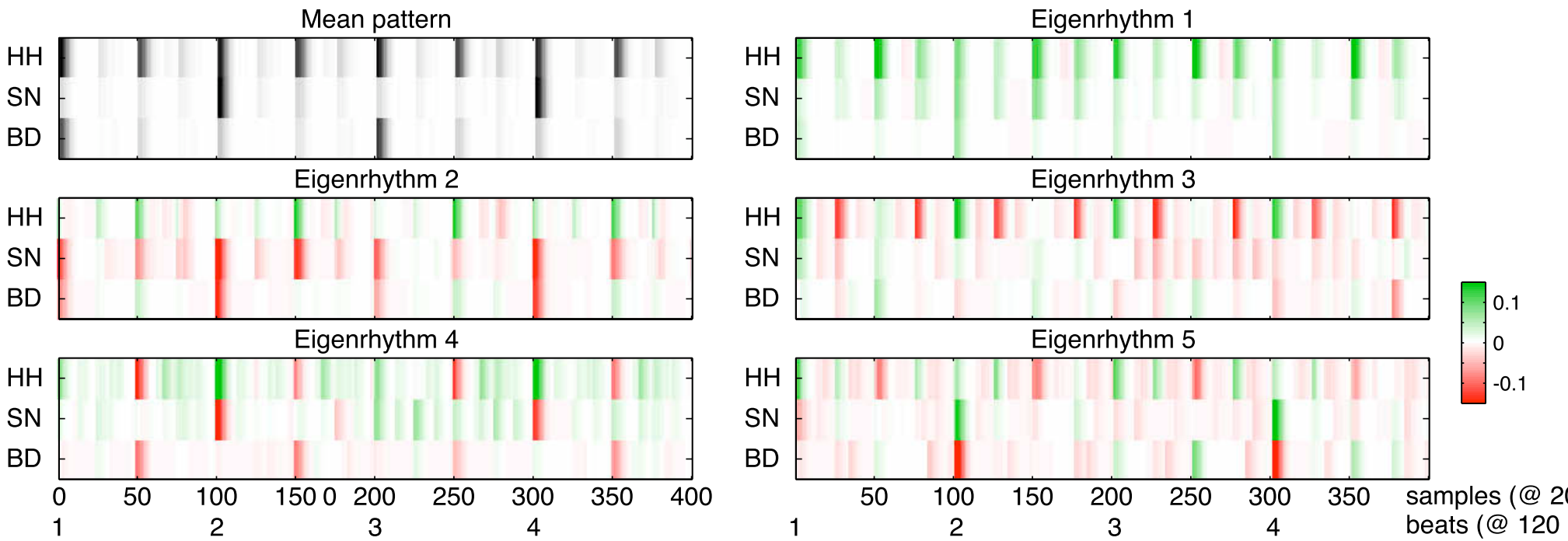
- Pop songs built on repeating "drum loop"
  - bass drum, snare, hi-hat
  - small variations on a few basic patterns



- Eigen-analysis (PCA) to capture variations?
  - by analyzing lots of (MIDI) data
- Applications
  - music categorization
  - "beat box" synthesis

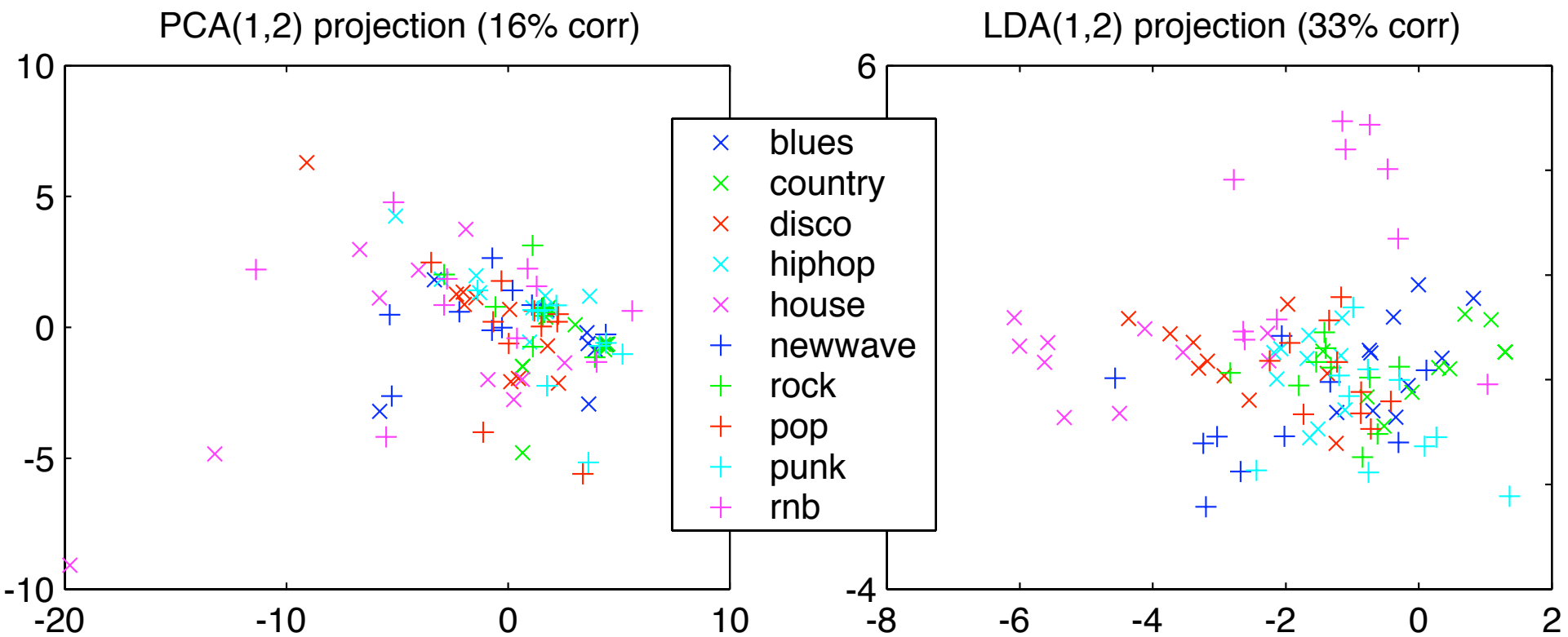
Eigenrhythms

- Need 20+ Eigenvectors for good coverage of 100 training patterns (1200 dims)
- Top patterns:



Eigenrhythms for Classification

- Projections in Eigenspace / LDA space

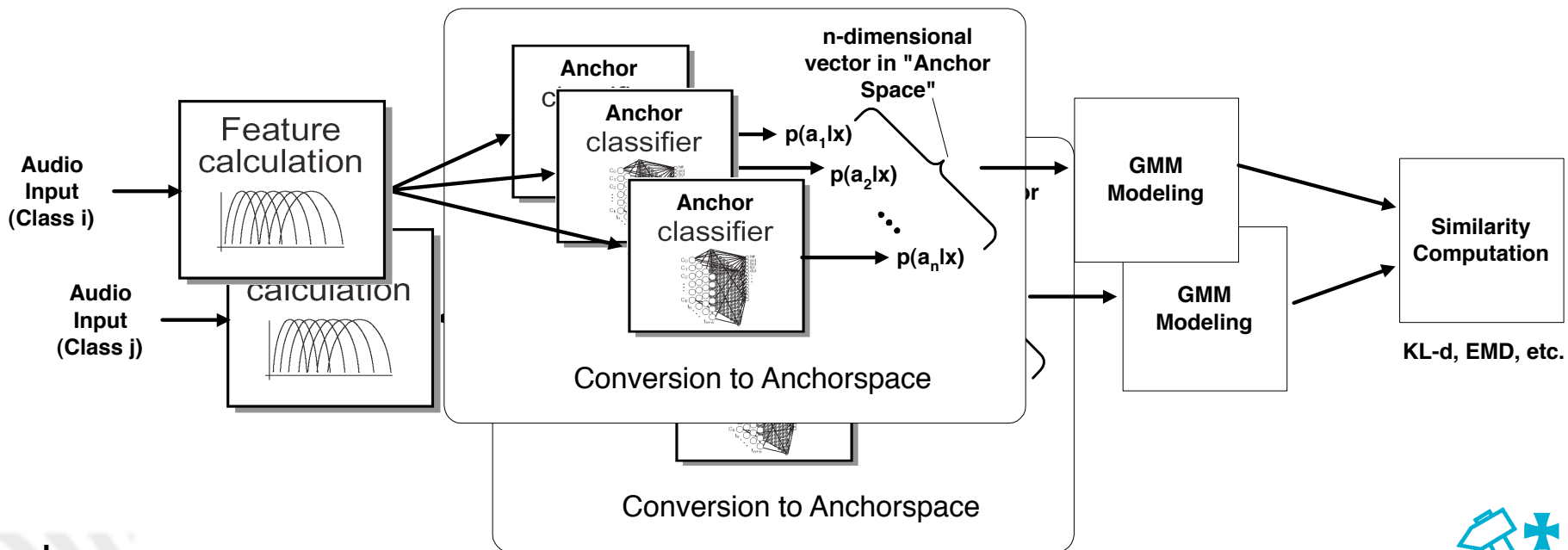


- 10-way Genre classification (nearest nbr):
  - PCA3: 20% correct
  - LDA4: 36% correct

Music Similarity Browsing

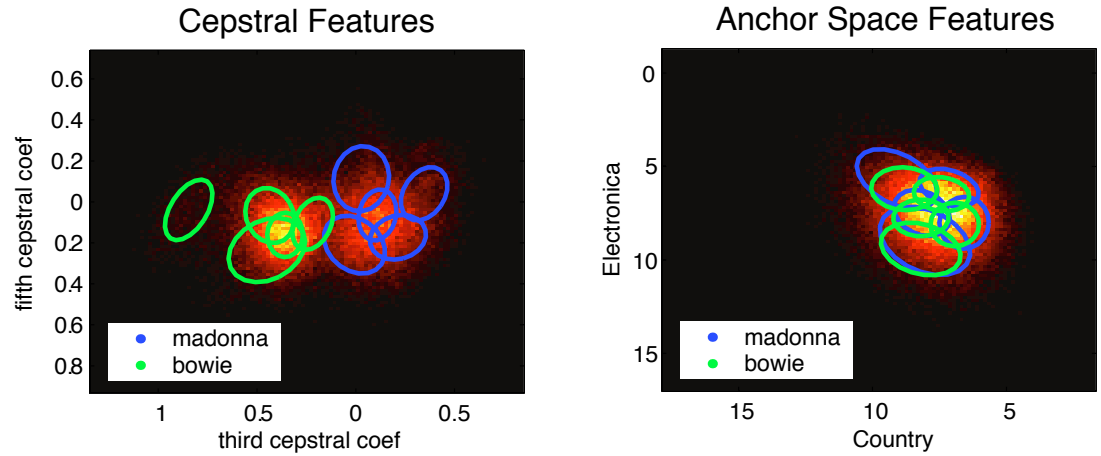
with Adam Berenzweig

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

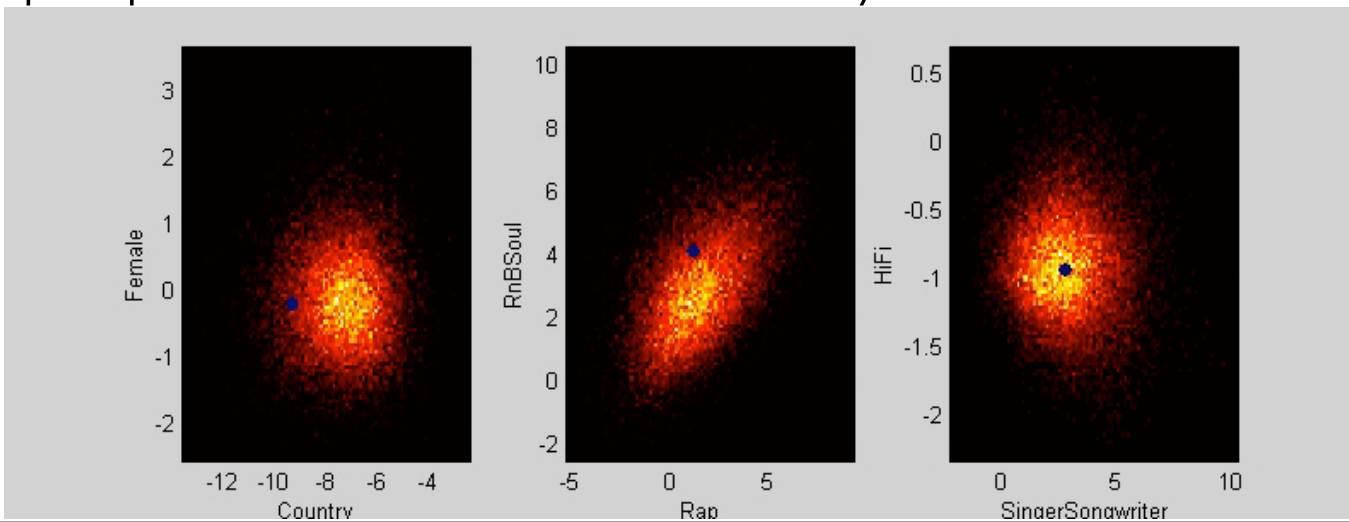


Anchor Space

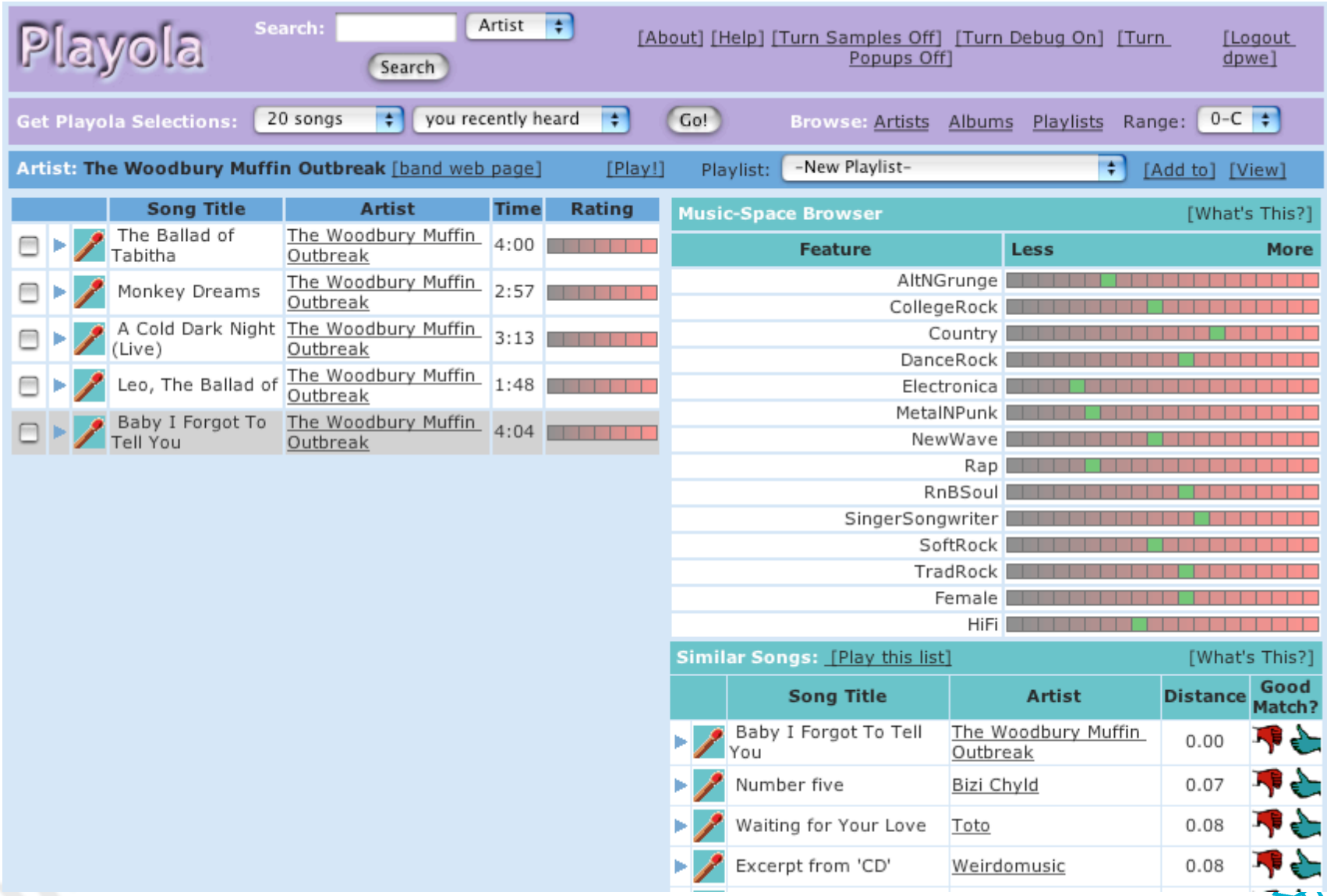
- Frame-by-frame high-level categorizations
  - compare to raw features?



  - properties in distributions? dynamics?



‘Playola’ Similarity Browser



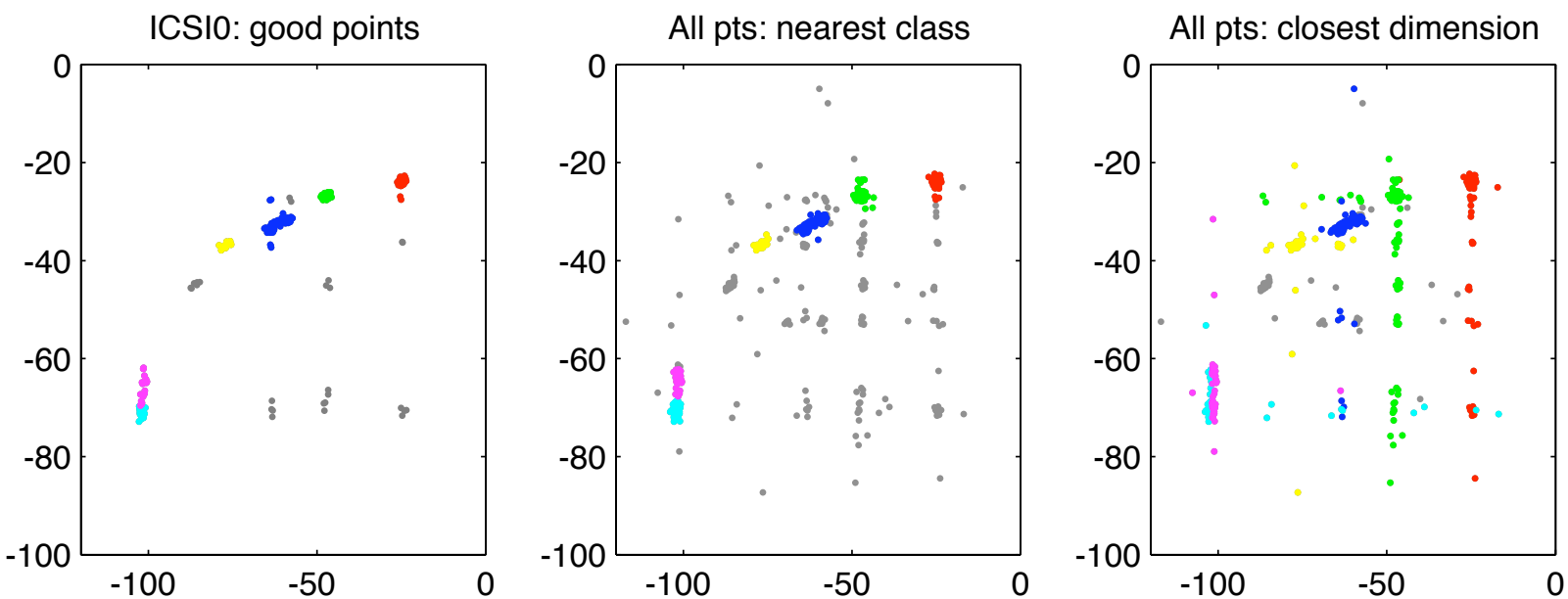
2. Meeting Recordings

with Jerry Liu and ICSI

- **Novel application for speech processing**
  - summarize & review content of meetings
- **What really happens in meetings?**
  - analysis of decision making, participant roles
- **Multi-mic recordings for speaker turns**
  - e.g. ad-hoc sensor setups (multiple PDAs)

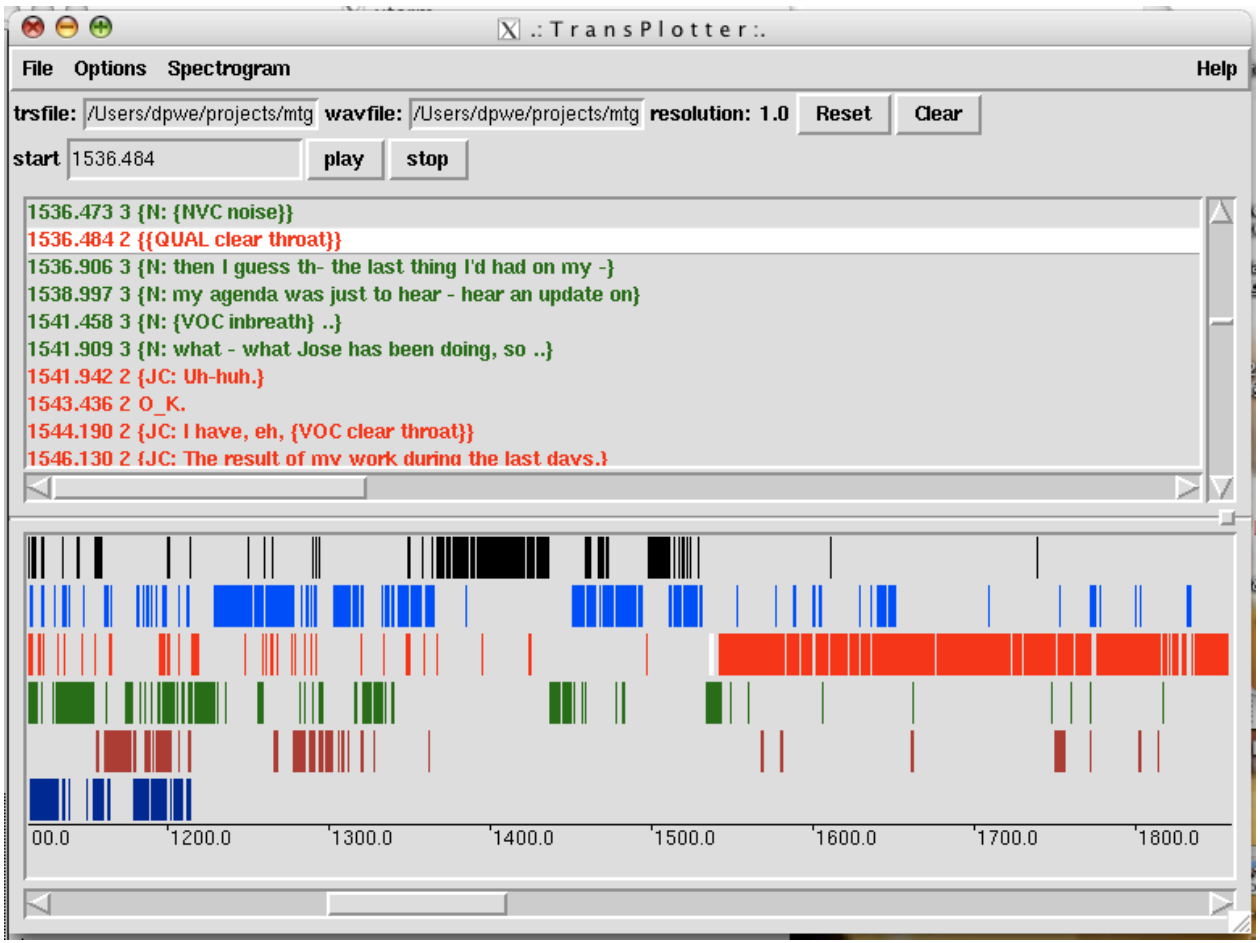
Speaker Turns from Timing Diffs

- Find best timing skew between mic pairs
- Find clusters in high-confidence points
- Fit Gaussians to each cluster,
  assign that class to all frames within radius



Browsing Meeting Recordings

- Information in patterns of speaker turns



3. "Personal Audio"

with Keansub Lee

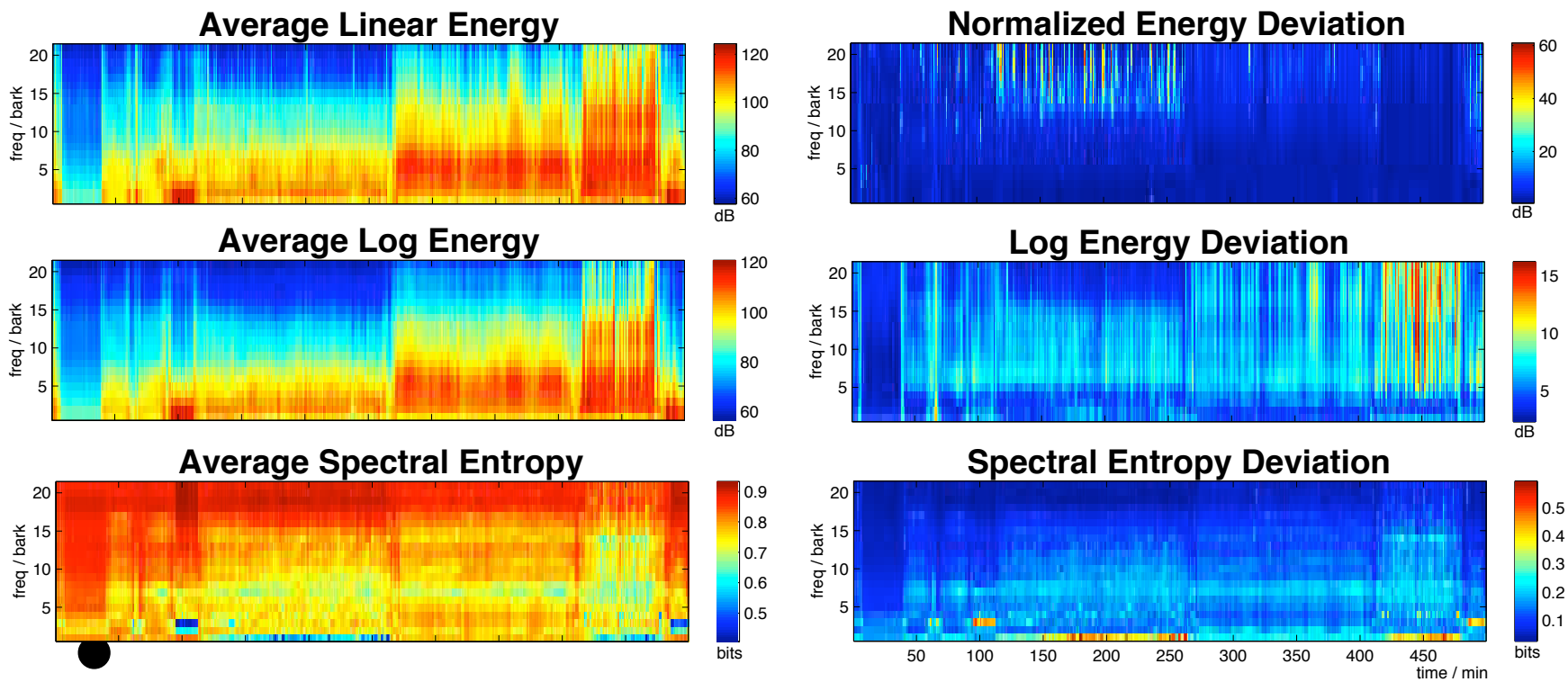
- Easy to record everything you hear
  - ~100GB / year @ 64 kbps
- Very hard to find anything
  - how to scan?
  - how to visualize?
  - how to index?
- Starting point: Collect data
  - ~ 60 hours (8 days, ~7.5 hr/day)
  - hand-mark 139 segments (26 min/seg avg.)
  - assign to 16 classes (8 have multiple instances)

Applications

- Automatic appointment-book history
  - fills in when & where of movements
- "Life statistics"
  - how long did I spend in meetings this week vs. last
  - most frequent conversations
  - favorite phrases??
- Retrieving details
  - what exactly did I promise?
  - privacy issues...
- Nostalgia?

Features for Long Recordings

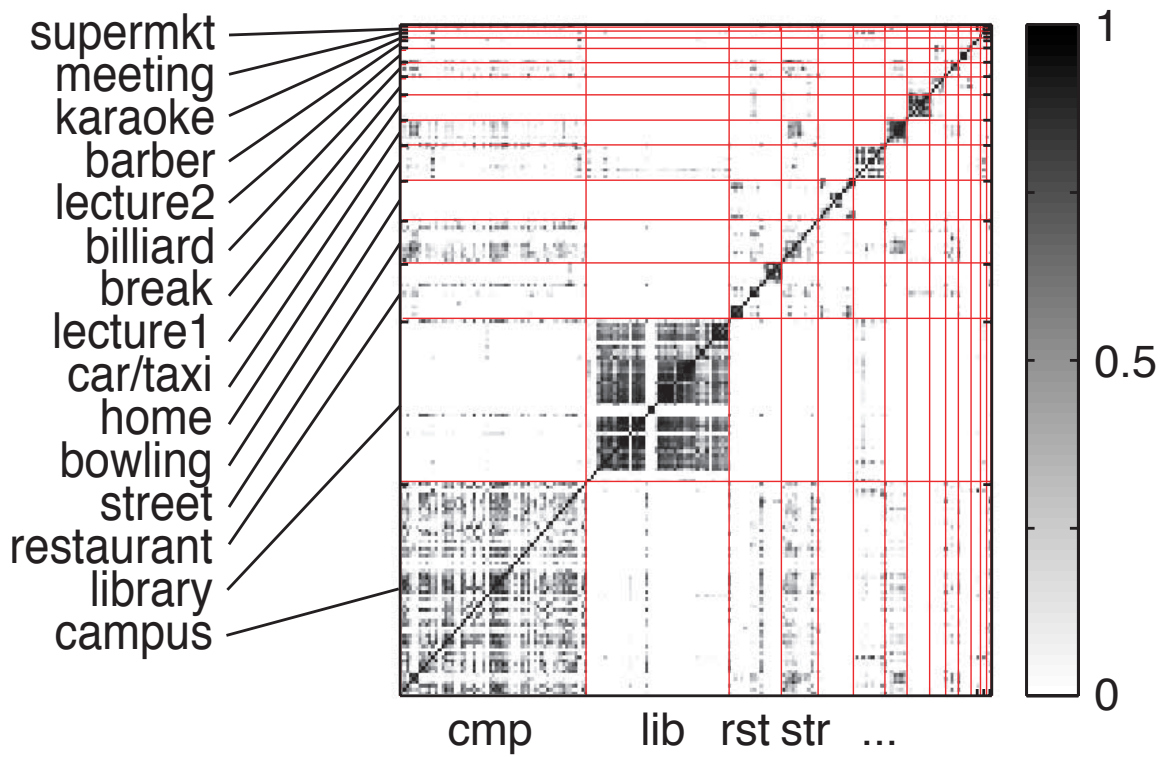
- Feature frames = 1 min (not 25 ms!)
- Characterize variation within each frame...



- and structure within coarse auditory bands

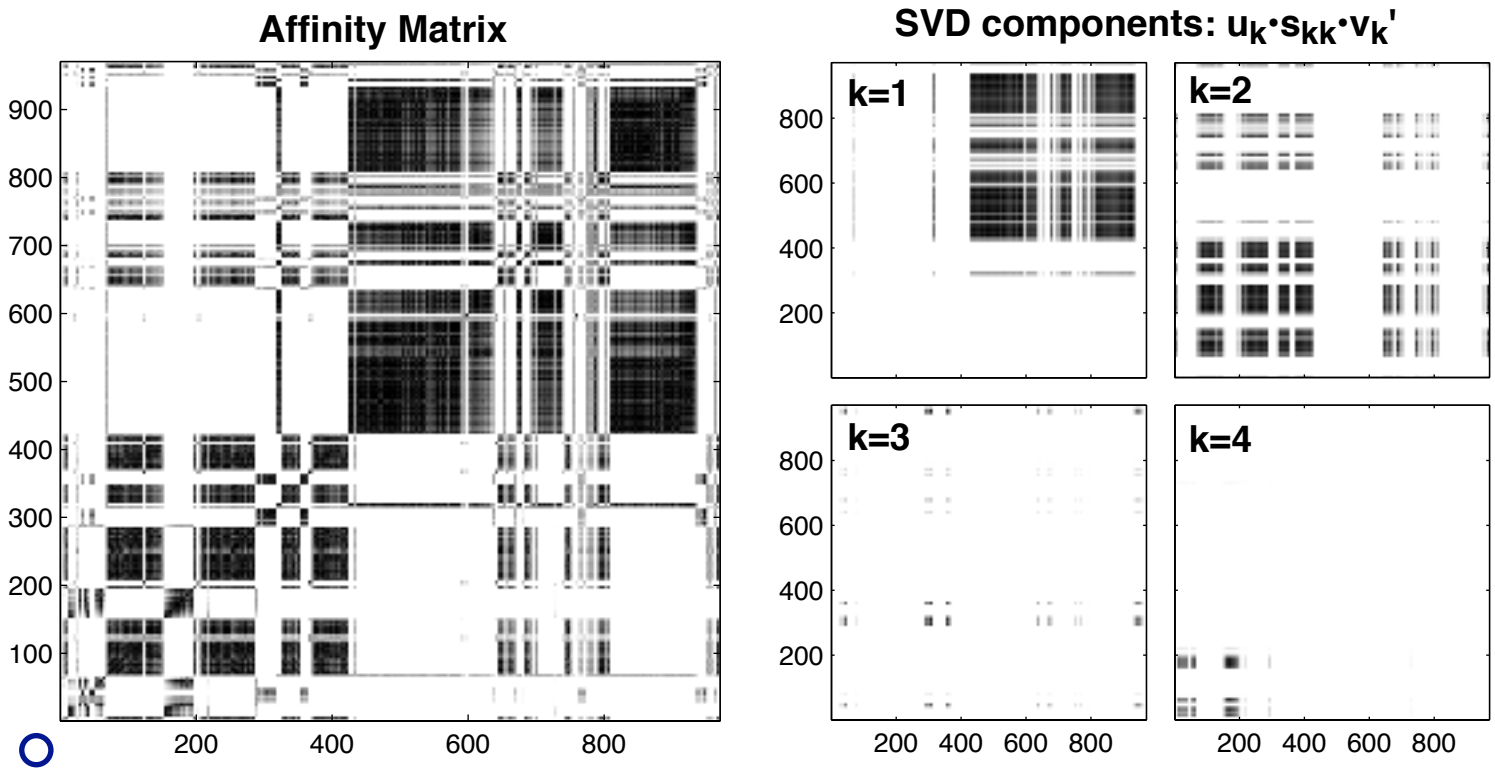
Segment clustering

- Daily activity has lots of repetition: Automatically cluster similar segments
  - 'affinity' of segments as KL2 distances



Spectral Clustering

- Eigenanalysis of affinity matrix: A = U•S•V'

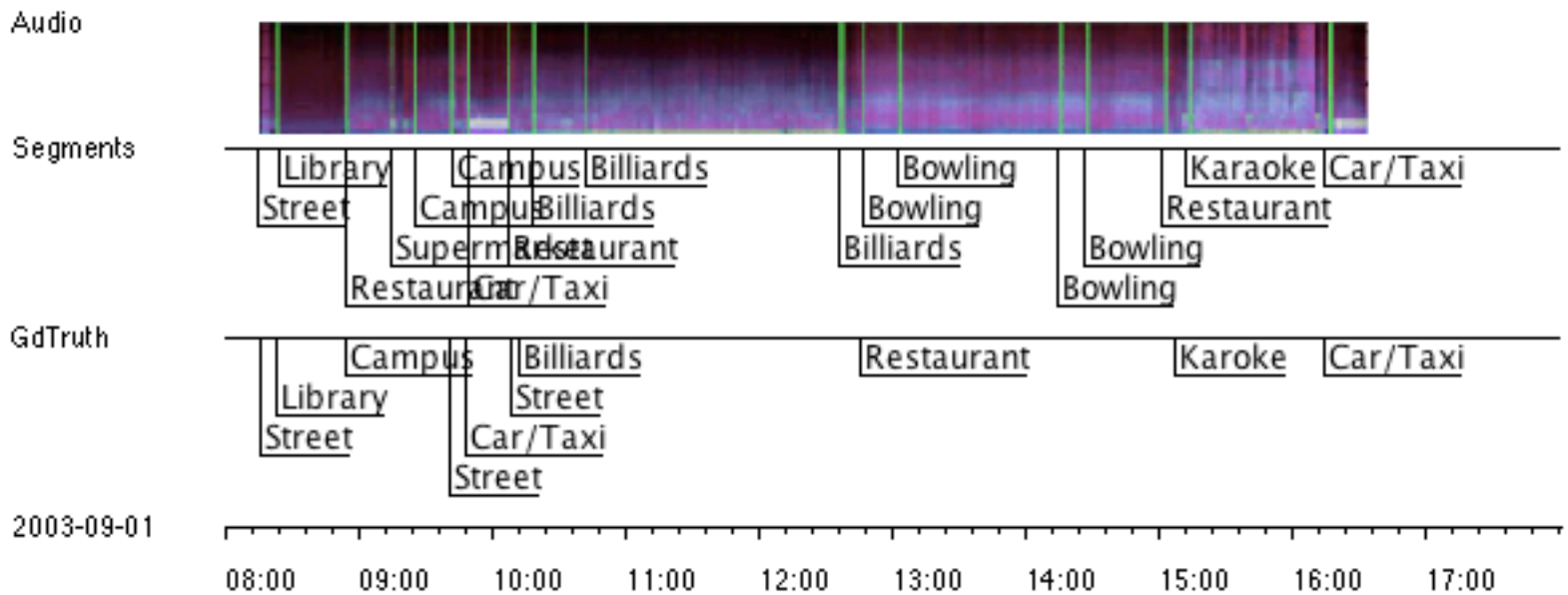


  - eigenvectors v_k give cluster memberships

- Number of clusters?

Clustering Results

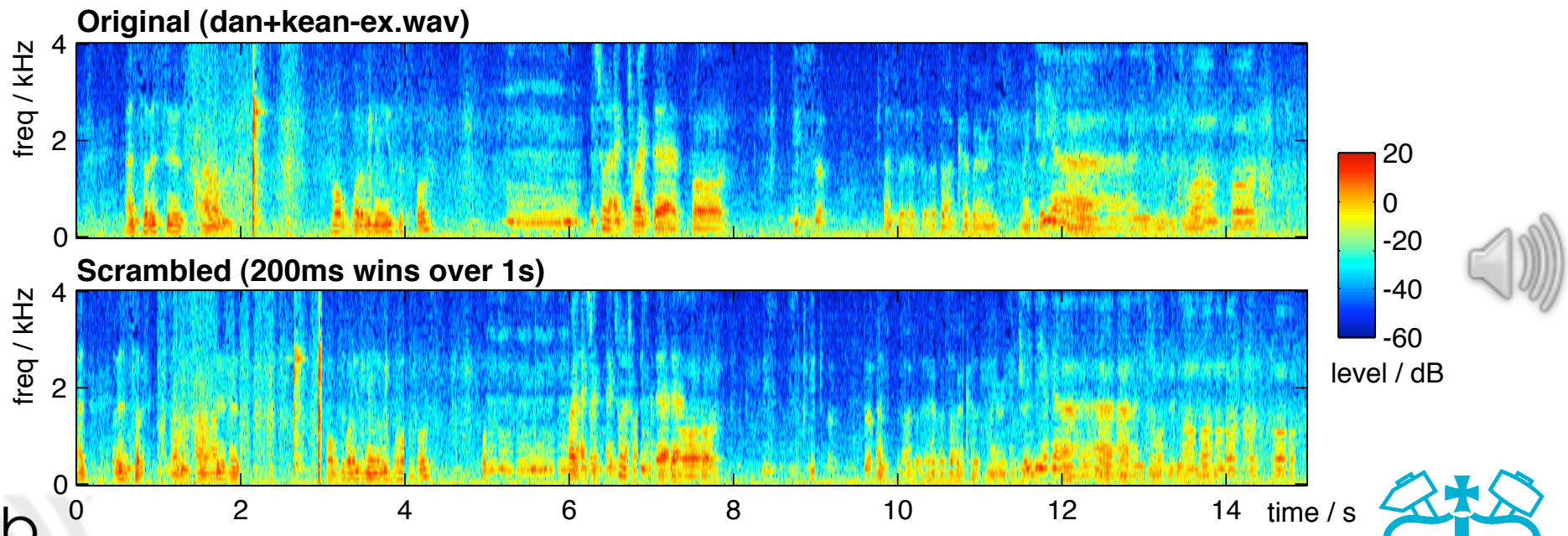
- Clustering of automatic segments gives 'anonymous classes'
  - BIC criterion to choose number of clusters



- Frame-level scoring gives ~70% correct
  - errors when same 'place' has multiple ambiences

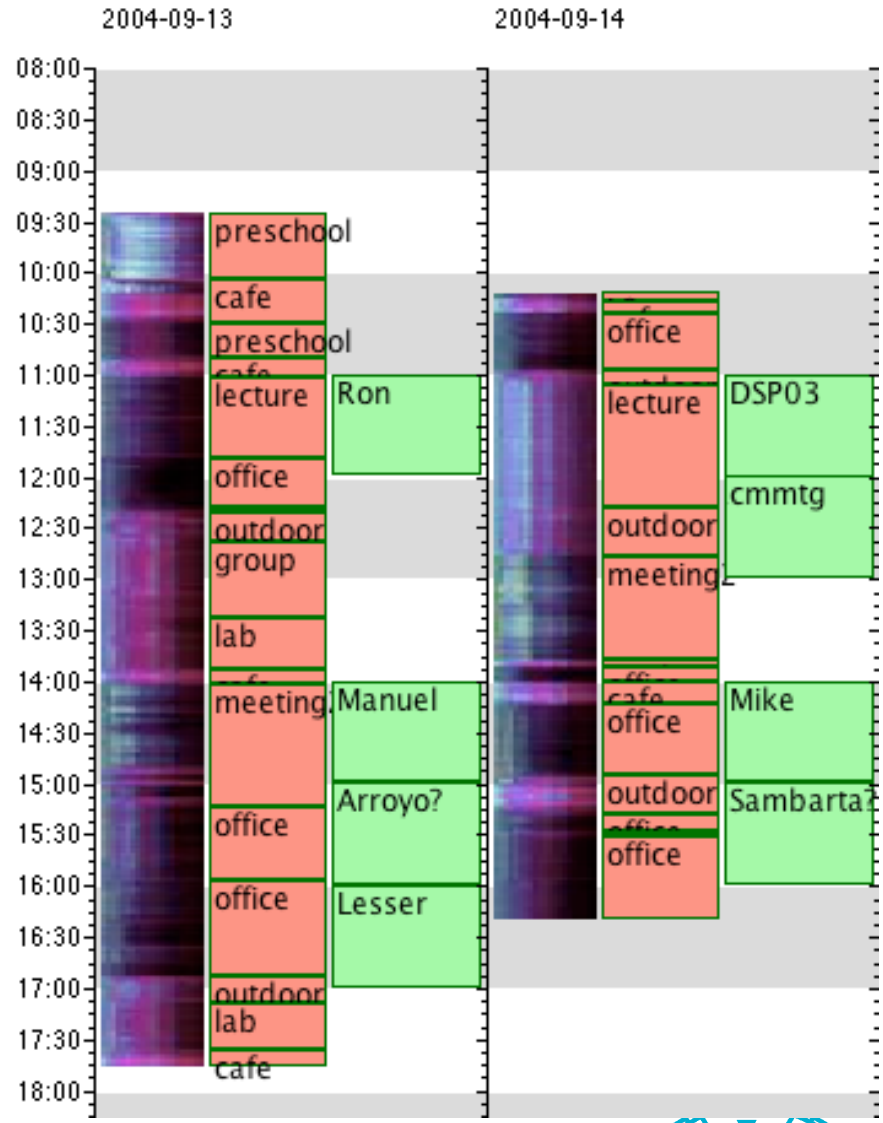
Privacy

- Recordings affront expectations of privacy
  - critical barrier to progress
- Technical solutions? Speech Scrambling
  - scramble 200ms segs of speech (long term OK)
  - high-confidence speaker ID to bypass



Visualization and Browsing

- Visualization / browsing / diary inference
  - link in other information sources
    - diary
    - email
- NoteTaker interface:
  - "what was I hearing?"



Personal Audio: Current Work

- **Voice segment detection / identification**
  - poor and variable SNR
  - channel variation
- **Novelty detection**
  - segments that *don't* cluster against archive
- **Spatial information**
  - identifying relative motion of head/sources
  - .. for motion detection
  - .. for source separation

LabROSA Summary

- LabROSA
  - signal processing
    + machine learning
    + information extraction
- Applications
  - Music: Transcription, Recommendation
  - Speech: Recognition, Organization
  - Environment: Detection, Description
- Also...
  - signal separation, compression, dolphins...