

Introduction to Music Information Retrieval Research: Past, Present and Future

J. Stephen Downie, Ph.D.

20 July 2007

Graduate School of Library & Information Science
University of Illinois at Urbana-Champaign
jdownie@uiuc.edu



National Science Foundation
WHERE DISCOVERIES BEGIN



THE ANDREW W. MELLON FOUNDATION



Agenda

- Why are we here?
- Brief History of ISMIR/MIREX
- Introduction to MIR
- HUMIRS Overview
- M2K Overview
- MIREX Overview
- Questions, Demos and Discussions

Why are we here?

- To provide a brief "big picture" overview of the exciting developments being made by MIR researchers from all around the world.
- To encourage the ongoing participation in MIR research and development of those with librarianship, business, computer science, engineering, musicological, etc. backgrounds.
- To outline the exciting work being done by the MIR/MDL communities to put MIR/MDL research on a strong scientific foundation

Why are we here? (Pt. 2)

- Demonstrate a broader world “vision” of what can be done to help MIR research move forward *scientifically*
- Build bridges across disciplines to forge stronger “real world” research, implementation and *evaluation* programmes
- Open communication channels for input from all interested parties

Why are we here? (Pt. 3)

- Post-Talk Brainstorming
 - To take stock of where we are right now and where to move next
 - To identify key areas where partnerships are both necessary and possible collaboration and funding scenarios
 - To discuss how to move forward on establishing “real world” collections, installations, and user feedback data to improve system design

What is MIR?

- Born ca. 1960's in IR research
- Major recent growth precipitated by advent of networked digital music collections
- Informed by multiple disciplines and literatures
- ISMIR started in 2000

ISMIR/MIREX Overview

- Conceived @ Berkeley 1999 with Dr. Don Byrd
- 1999 ACM SIGIR Workshop on MIR
 - First mention of “TREC-like” evaluations
- Plymouth, Bloomington, Paris, Baltimore
Barcelona, London, Victoria, Vienna
- Bloomington (2001) and Barcelona (2004)
important points in MIREX history

Overarching Problem

- No way single way to scientifically compare and contrast techniques
- ISMIR 2001 “Resolution” explicitly recognized this problem

ISMIR 2001 Resolution

There is a current need for metrics to evaluate the performance and accuracy of the various approaches and algorithms being developed within the Music Information Retrieval research community. A key component for the development of such metrics would be a corpus of electronic data consisting of both audio and structured music data. Such a corpus would need to be readily available to the research community with international clearance of all relevant copyright and other intellectual property rights necessary to use this data solely for the purpose of music information retrieval research.

2001-2003

- Secured funding to explore evaluation issues:
“MIR/MDL Evaluation Project”
- 3 Special Workshops:
 - Portland July 2002
 - Paris October 2002
 - Toronto August 2003
- Report compiled and begging for money began!
 - <http://www.music-ir.org/evaluation/wp.html>

Audio Description Contest

- Barcelona 2004
- Music Technology Group (Dr. Serra's Lab)
- Contest Categories
 - Genre Classification/Artist Identification
 - Melody Extraction
 - Tempo Induction
 - Rhythm Classification
- MIREX built upon the lesson learned by ADC

Defining Music Information Retrieval?

- Music Information Retrieval (MIR) is the process of searching for, and finding, music objects, or parts of music objects, *via* a query framed musically and/or in musical terms
- *Music Objects*: Scores, Parts, Recordings (WAV, MP3, etc.), etc.
- *Musically framed query*: Singing, Humming, Keyboard, Notation-based, MIDI file, Sound file, etc.
- *Musical terms*: Genre, Style, Tempo, etc.

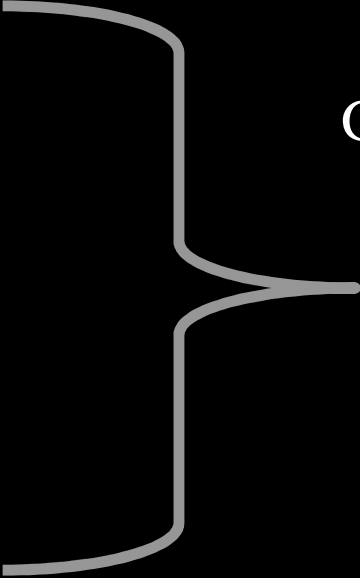
The “Brass Ring” MIR System

- Multimodal, Multirepresentational, Multicultural
- Has a meaningful abstracting/thumbnail feature for determinations and browsing
- Employs an intelligent, user-definable, *experientially-grounded*, relevance-feedback/classification mechanism:
 - User inputs “song” into system and can tell system which aspect(s) of the music (e.g., throbbing bass, sweet violins, tempo, rap-like vocals, etc.) is/are the key factor(s) that should be the basis for gathering similar items
 - Would overcome user input errors

What makes MIR so tricky?

Music information is:

- Multifaceted
- Multimodal
- Multirepresentational
- Multiexperiential
- Multicultural



Given the inherent complexities of music information, only a multidisciplinary research approach could possibly lead to the development of a robust MIR system.

Multifaceted (Pt. 1)

- Pitch
 - Pitch is “the perceived quality of a sound that is chiefly a function of its fundamental frequency in --the number of oscillations per second ” (Randel 1986)
 - Also, the distance between pitches: intervals
- Temporal
 - Meter, duration, rhythm, tempo, etc.
- Harmonic
 - When two or more pitches occur at the same time, a simultaneity, or harmony, occurs. Also known as polyphony, while absence of polyphony is called monophony.

Multifaceted (Pt. 2)

- Timbre
 - Tone-colour
 - Flute v. Kazoo v. Violin v. Bass Drum
- Editorial
 - Fingerings, Ornamentation, Dynamic instructions (e.g., *ppp*, *p*, *..f*, *fff*), Slurs, Articulations, *Stacatti*, Bowings, etc.
- Textual
 - Lyrics, *Libretti*
- Bibliographic
 - Title, Catalogue Num., Composer, Publisher, Lyricist, etc.

Multirepresentational (Pt. 1)

- Solfege
 - do, re, mi, fa, so, etc.
- Pitch names
 - A, B, C, D, E, F#, A^b, etc.
- Chord Names
 - Cmaj, Dmin, Am7, etc.
- Scale Degree
 - I, II, III, IV, V, VI, VII
- Interval
 - +1, 0, -3, -8, +6, etc.

Multirepresentational (Pt. 2)

- MIDI Events:

12:1:000	key	7	G6	100	190
12:2:000	key	7	G6	100	190
12:3:000	key	7	A6	100	190
12:4:000	key	7	F#6	100	286
13:1:096	key	7	G6	100	94
13:2:000	key	7	A6	100	190

- Graphic Score:



Multimodal

- Music as thought
 - Tune running through head
- Music as auditory events
 - Sound waves hitting eardrums
 - Sound in electromechanical formats
 - WAV, MP3, AU, CD, LPs and Tapes
- Music as graphic language
 - Symbolic representations
 - Scores
 - MIDI files and other discrete encodings
 - etc.

Multiexperiential (Pt. 1)

- Music as object of study
 - Perform, Analyze
- Music as foreground
 - Concert going, Deliberate audition
- Music as background
 - Movie scores, Shopping malls, Housecleaning
- Music social signifier
 - Protest, Peace, Group songs, “Brow-ness”: High, Middle, Low, etc.

Multiexperiential (Pt. 2)

- Music as *aide memoire*
 - Soundtrack recordings, Camp songs, War songs, Ballades, etc.
- Music as tradition
 - Hymns, Folksongs, Nursery songs, etc.
- Music as drug
 - Stimulation
 - Stay awake, Frenzied dancing, etc.
 - Relaxation
 - Stress relieve, Forgetfulness, Sleep, etc.
 - Seduction

Multicultural

- Different notation/representational schemes
 - E.g., Modern art music
- Lack of notation/representational schemes
 - E.g., Jazz (improvised), Aural and oral traditions
- Different scales and modes
 - E.g., Quartertone music, Gamelan music, Eastern music
- Different grammars of musical affect and gesture
 - E.g., Inuit throat music, Indian ragas
- Different accessibility to recordings and recording technologies

Research can be classified by music representation

Representation	Description	Research
Symbolic	Notation (scores, charts), Event-based recordings (MIDI), Hybrid representations	Matching, Theme/Melody Extraction, Voice Separation, Musical Analysis
Audio	Recordings, Streaming Audio, Instrument Libraries	Sound/Song Spotting, Transcription, Timbre Classification, Musical Analysis
Visual	Scores	Score Reading (“Optical Music Recognition”)
Metadata	Cataloging, Bibliography, Descriptions	Library Testbeds, Traditional IR, Interoperability

Research can be classified by intended use

- Locating MIR Systems
 - Designed to find music or music information
 - Tend toward broad collections (i.e., breadth)
 - Intended for general use (think google.com)
- Analytic/Production MIR Systems
 - Designed to help musicologists, theorists and music typesetters, etc.
 - Tend toward fine-grained access to smaller collections (i.e., depth)
 - Intended for expert use (assumes music skill and knowledge)

MIR Communities (Pt. 1)

Community	Type of Institution(s)	Research Areas
Computer Science, Information Retrieval	Academic, Commercial	Representation, Indexing, Retrieval, Machine Learning, User Interface Design
Audio Engineering, Digital Signal Processing	Academic, Commercial	Compression, Feature Detection, Pitch Tracking, Machine Learning, Classification, Musical Analysis
Musicology, Music Theory	Academic	Representation, Musical Analysis



MIR Communities (Pt. 2)

Community	Type of Institution(s)	Research Areas
Library Science	Libraries, Academic	Representation, Metadata, User Studies, Classification, Intellectual Property Rights, User Interface Design
Cognitive Science, Psychology, Philosophy	Academic	Representation, Perception, User Studies, Ontology
Law	Government, Legal Profession, Academic	Intellectual Property Rights

Overview of MIR communities

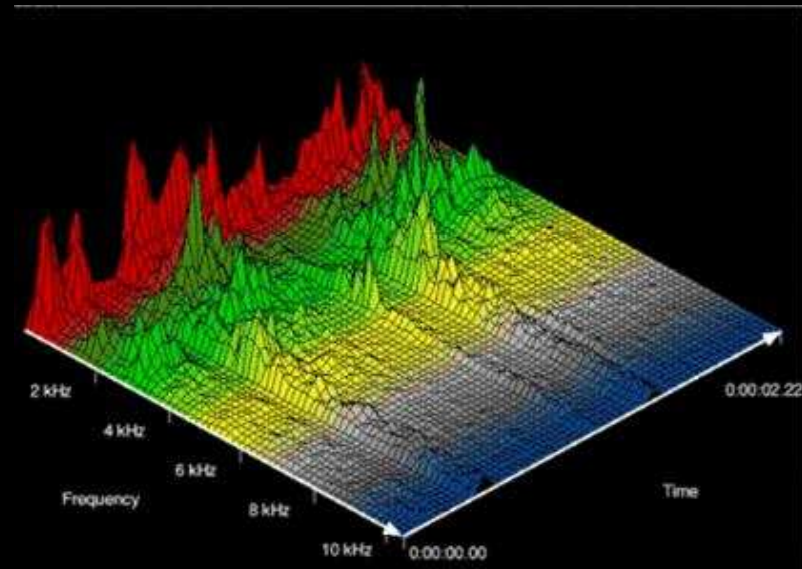
- Background
- Research issues
- Methodology
- Emphases

Computer science, Information retrieval

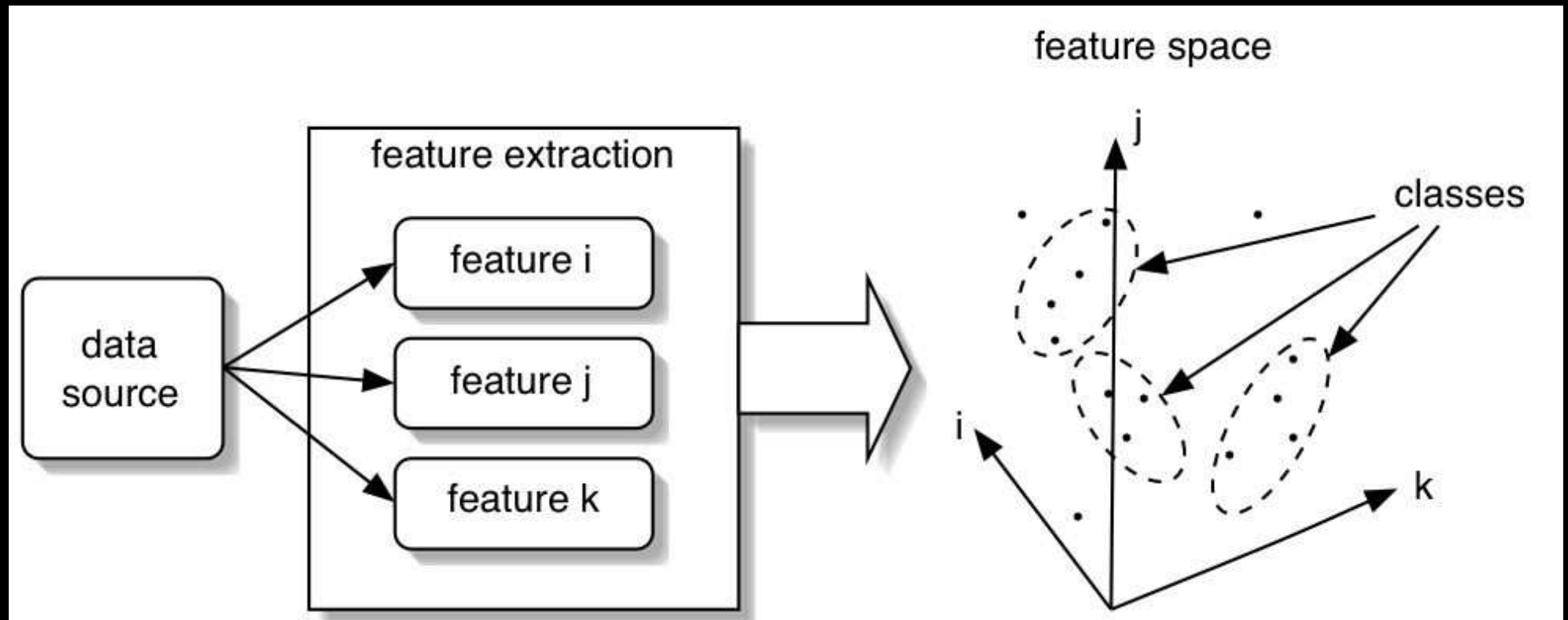
- Originated in text IR research in the 1960's
- Emphasizes retrieval modality
- Text methods employed
 - Reduce music to text
 - Employ text IR methods
- Emphasis on QBH

Audio engineering, Digital signal processing

- Grows out of DSP and speech recognition research
- Feature detection mostly from FFT and MFCC
- Analysis
 - Clustering
 - HMM's
 - Neural networks

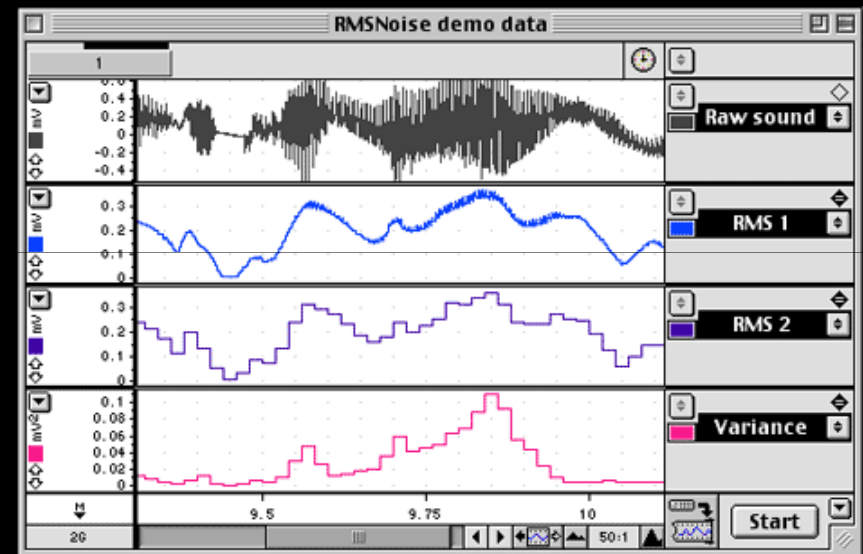
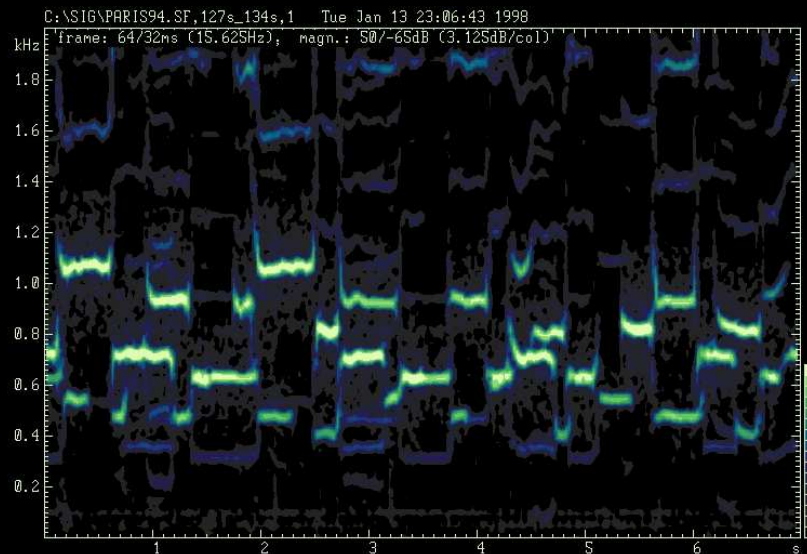


Classification (Pt. 1)



Classification: audio features (Pt. 2)

- Frequency domain (e.g., FFT, MFCC's)
- Time domain (e.g., RMS, autocorrelation)



Classification: partitioning feature space (Pt. 3)

- Unsupervised
 - Choose distance metric (e.g., Euclidian distance)
 - Choose (or don't) number of classes
 - Partition into classes that minimize within-class distance and maximize between-class distance
- Supervised
 - Manually classify a subset of the data objects
 - Use those class labels to “train” a machine learning algorithm
 - Have it generate class labels for unclassified objects
 - “Reward” algorithm for correct labels, “punish” it for incorrect labels
 - Repeat

Musicology, Music theory

- Recently transformed by computational techniques, will be further transformed by access to music collections
- Computational musical analysis
- MIR systems specialized for musicologists

Library and information science

- Has been dealing with music manually since antiquity
- Covers broad range of issues
 - Scale
 - IP rights
 - Usability
- Focus on testbeds



*The
Lester S. Levy
Collection of Sheet Music*

Cognitive science, Psychology, Philosophy

- Music perception
 - Theoretical models
 - Experimental results
- Epistemology and Philosophy
 - Cultural criticism
 - Ethnomusicology
- MIR researchers have given little attention to this body of work

Law/Business

- IP rights/Market strategies
- Large music collections (e.g. Napster) have great cultural importance but uncertain legal implications
- Focus on harnessing potential markets but limiting piracy

Summary: Music Retrieval as....(good points)

- Text retrieval (Symbol-based)
 - Apply text retrieval techniques
 - Builds on known IR techniques and research
- String retrieval (Symbol-based)
 - Sees music as long strings
 - Approximate string match techniques; Fault tolerant
 - Similar to gene sequence retrieval
- Speech retrieval (Audio-based)
 - Builds upon large body of speech recognition research
 - Gives hope with regard to processing large volumes of recorded CDs and MP3s, etc.

Summary: Music Retrieval as....(weak points)

- Text retrieval (Symbol-based)
 - Symbol dependent (only as rich as the representation)
 - Reductionist approach; Polyphony difficult in IR model
- String retrieval (Symbol-based)
 - Same as above
 - Can be inefficient
- Speech retrieval (Audio-based)
 - Huge datasets; Computationally expensive
 - Part extraction/searching next to impossible

Summary: World Views (Pt. 1)

- **Library Science**
 - Free access; Bibliographic control; Standards; Metadata; Human intervention; Name ->Locate ->Retrieve; OPAC model
- **Computer Science**
 - Centrality of the algorithm; Algorithmic efficiency; Pattern matching; Pragmatic; Reductionist; Search engine model
- **Information Science/Information Retrieval**
 - Tradition of evaluation; Centrality of relevance; Precision and recall

Summary: World Views (Pt. 2)

- Audio Engineering
 - Signal processing; Centrality of the recording; Holistic
- Business/Law
 - Paid access; Fault tolerance; Centrality of the market; Music as property
- Musicology
 - Fine-grained access; Analysis; Connections between works; Representation development; Assumptions of musical knowledge

MIR/MDL Evaluation Project

- Funded by Andrew W. Mellon Foundation and the National Science Foundation (NSF)
- To examine, via interaction the MIR/MDL community, what would be needed to fulfill the requirements of the 2001 Resolution
- Began July 2002 – morphed into several research streams

Prompting Questions

1. How do we determine, and then appropriately classify, the tasks that should make up the legitimate purviews of the MIR/MDL domains?
2. What do we mean by “success”? What do we mean by “failure”?
3. How will we decide that one MIR/MDL approach works better than another?
4. How do we best decide which MIR/MDL approach is best suited for a particular task?

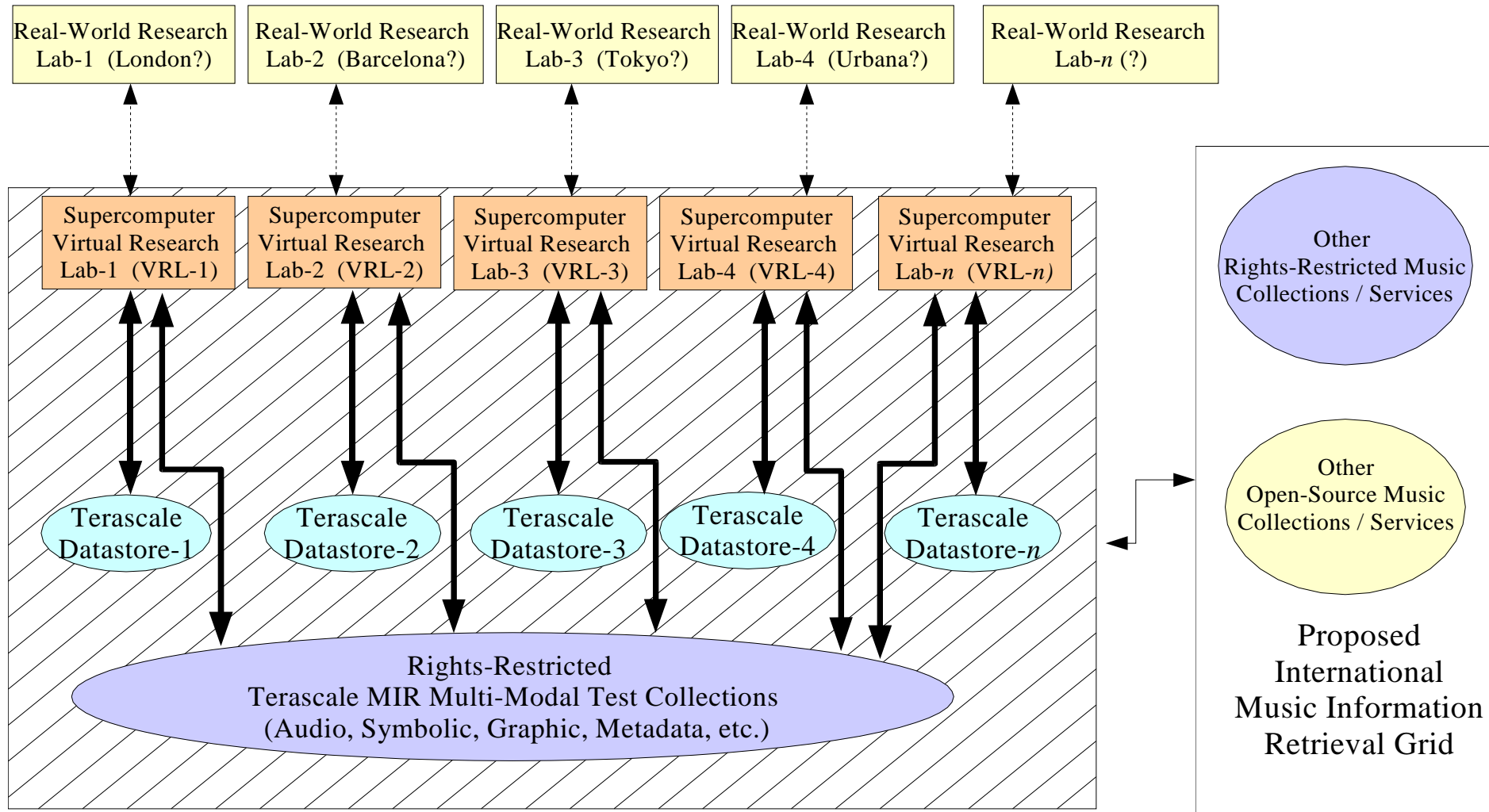
Four Paths to Making Real Evaluations a Reality

- **IMIRSEL**: International Music Information Retrieval System Evaluation Laboratory
- **HUMIRS**: Human Use of Music Information Retrieval Systems
- **M2K**: Music-to Knowledge
- **MIREX**: Music Information Retrieval Evaluation eXchange

IMIRSEL: First Principles

1. Security for the music materials
2. Accessibility for international, domestic and internal researchers
3. Sufficient computing and storage infrastructure for the computationally- and data-intensive MIR/MDL techniques examined

Virtual Research Labs Model



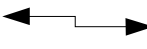
Legend:




 Super-Bandwidth I/O Channel NCSA Music Data **Secure Zone**



 Command/Control/Derived Data traffic via Internet



 Connection to International MIR Grid

A Important Distinction

- We must emphasize here the distinction between "queries" and "search statements".
- “**Queries**” are the verbalized expressions of a user's information need; whereas,
- “**Search statements**” are the expression(s) of the queries in the syntax of particular retrieval engines.

Nature of Music Queries

- Semantically-rich
- Syntactically-undetermined
- Structurally-independent of any particular search system
- Content-variable
 - (i.e., can contain singing, text, recorded examples, notation, etc.)

In Other Words....

- Simply put we need to know:
 - What real users want
 - How real users will interact with the systems
- To determine:
 - Which MIR tool best matches:
 - The type(s) of music sought
 - The type of query submitted
 - The ultimate use of the desired music

HUMIRS: Query Record Criteria

The query records developed must :

1. be grounded in real-world needs and uses;
2. be representative of the complexity of real-world queries;
3. be neutral with regard to the retrieval method employed; and,
4. be data-rich so realistic and meaningful “relevance” judgments can be made

A TREC Query

<num> Number: 409

<title> legal, Pan Am, 103

<desc>Description:

What legal actions have resulted from the destruction of Pan Am Flight 102 over Lockerbie, Scotland, on December 21, 1988?

<narr> Narrative:

Documents describing any charges, claims, or fines presented to or imposed by any court or tribunal are relevant, but documents that discuss charges made in diplomatic jousting are not relevant.

Figure 1. TREC topic statement (from Voorhees, 2002).

Ugly Ducklings: MIR Queries

From: XXXXXXXXXXXX

Subject: Early 80's - Please identify this song! (it's *very* difficult, though)

Newsgroups: alt.music.lyrics

Date: 2000-12-14 09:42:24 PST

Hi, this is so difficult because I only remember those damn FRAGMENTS of it, which can (in combination with possible errors) make it VERY difficult to identify this song!

But I'll try my best to make myself clear as possible.

This song MUST be from the period 1979-1984, most likely 1981 or 1982.

Tempo: about 120 bpm

Sounds VERY close to a SAGA or Asia tune (maybe it is SAGA even! ;)

OK here I go...(gonna add the chords for you guitarists out there ;)

Ugly Ducklings: MIR Queries (Pt. II)

[verse 1]

F C Bb Bb C

Crazy onto the café

F C Bb

I'm drinking coffee, she came away

F C Bb Bb C

She ordered precious sum of money ???

F C Bb

deedeedeedeedeedeedeede....

C

Ohohohoo

[(instrumental) F C Bb Bb C F C Bb]

[verse 2] [...]

[chorus]

Dm Bb

Another da-----y, in the afternoon

Dm Bb(7)

Principal Features of a MIR TREC-like Query Record

1. High quality audio representation(s)
2. Verbose Metadata:
 1. About the “user”
 2. About the “need”
 3. About the “use”
3. Symbolic representation(s) of the music presented

Comment: Less like a traditional TREC topic statement and more like the kind of information garnered in a traditional, well-conducted, reference interview. This suggests that the involvement of professional music librarians in the development of the TREC-like music query records is very important — perhaps even critical.

Three Informative User Studies

- Downie, J. Stephen and Sally Jo Cunningham. 2002. Toward a theory of music information retrieval queries: System design implications. In *Proceeding of the 3rd International Conference on Music Information Retrieval (ISMIR 2002)*.
- Bainbridge, David, Cunningham, Sally Jo and J. Stephen Downie. 2003. How people describe their music information needs: A grounded theory analysis of music queries. In *Proceedings of the 4th International Conference on Music Information Retrieval (ISMIR 2004)*.
- Lee, Jin Ha and J. Stephen Downie. 2004. Survey of music information needs, uses, and seeking behaviours: Preliminary findings. In *Proceeding of the 5th International Conference on Music Information Retrieval (ISMIR 2004)*.

Categories of Expression

Information need description	Percentage of queries
• BIBLIOGRAPHIC	• 75.2%
• LYRICS	• 14.3%
• GENRE	• 9.9%
• SIMILAR WORKS	• 9.9%
• AFFECT (i.e., description of mood)	• 7.5%
• LYRIC STORY	• 6.8%
• TEMPO	• 2.5%
• EXAMPLE	• 1.8%

Categories of Intended Use

Category of intended use	Percentage
• Locate (e.g., “Where can I find...”)	• 49.7%
• Research (i.e., background information, etc.)	• 19.3%
• Perform (i.e., play piece(s) on instrument)	• 18.6%
• Collection Building (e.g., add to pre-existing collection one or more similar items)	• 18.0%
• Listen (i.e., as opposed to perform)	• 6.8%

Types Of Metadata (Pt. I)

- **Content Metadata**

- ***Musical metadata***: data derived directly from-- or directly descriptive of--the music itself
 - (e.g., lyrics, melody, tempo,mood, etc.)
- ***Bibliographic metadata***: traditionally-used metadata that describes the item
 - (e.g., title, author, publisher, etc.)

Types Of Metadata (Pt. II)

- **Context Metadata**
 - ***Relational metadata***: data about the item's relationships (artificially created or socially constructed) with other music related items (e.g., genre; indications of similarity, etc.)
 - ***Associative metadata***: data indicating associated use in other works, media or events (e.g., sampling; use in TV show, movies or commercials; use at special events, etc.)

Customer Reviews

- Published on www.epinions.com
- Focused on the **book, movie and music**
- Each review associated with:
 - a genre label
 - a numerical quality rating

Read Review of Pavarotti - Italian Wedding Favorites

Review Summary

Goes good with fondue, wine, and a summer sunset...

Aug 04 '05

Author's Product Rating



numerical rating
associated

Pros

Very romantic and soulful: a collection of Pavarotti's best

Cons

None, unless you are biased against Pavarotti

The Bottom Line

Excellent choice for a romantic dinner, or for Pavarotti lovers.

used in our
experiments

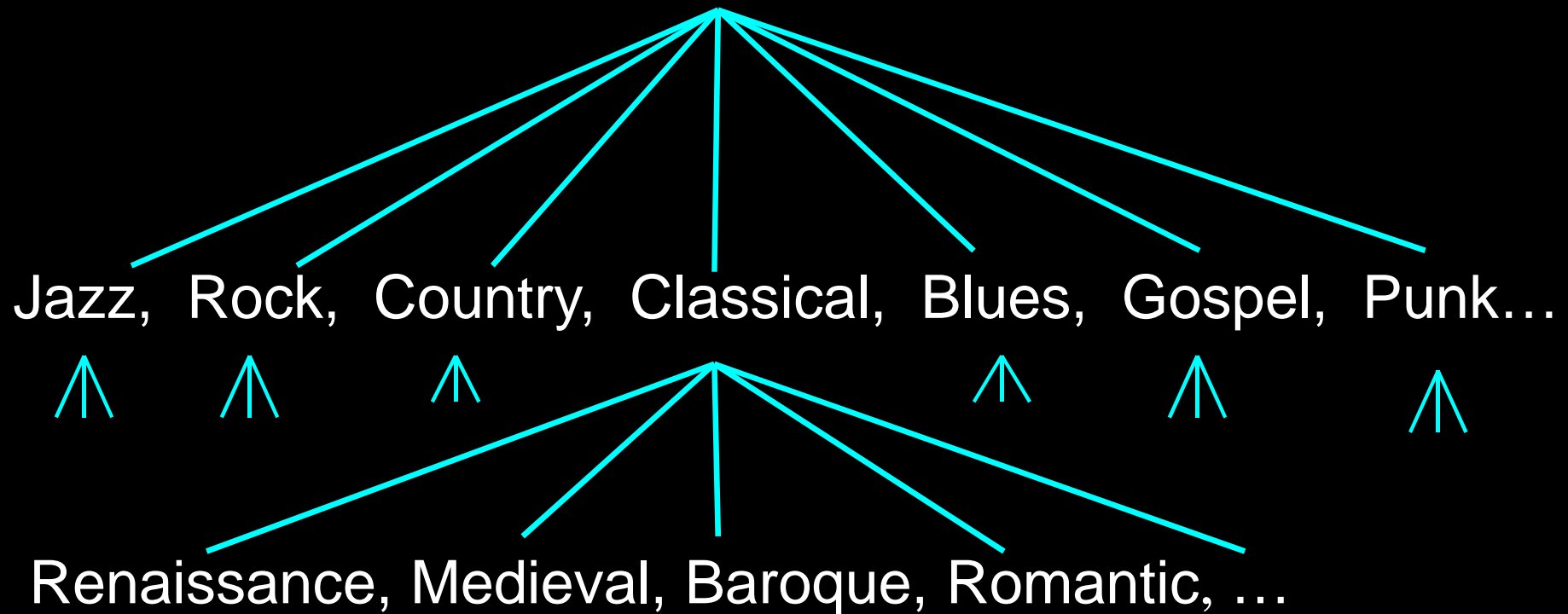
Full Review

I suppose what this CD should be called is "Favorite Italian Love Songs", because the other reviewer of this (quick review) was apparently a bit annoyed that these songs are not what one would sing at an Italian wedding. They are all love songs, however, which I suspect is where that wedding part comes in. On top of that all of the songs are sung in Italian, which is probably where the "Italian" part comes in. Hence, Italian Love Songs may have been a better choice for a title. I suppose that Italian Love Songs had already been taken by another Pavarotti CD, however, so they went with "Wedding Favorites", knowing that most people like myself wouldn't notice.

I listen to every music genre depending upon my mood. One has Metallica and Rob Zombie for some harsh, driving, rhythmic music; Britney Spears and the Spice Girls for synthesizer-driven Bee-Bop type music; Micheal Jackson for synthesizer-driven popular music; Celine Dion for slow romantic type music; and Beethoven for

Music Genres

28 Major Genre Categories



Experimental Setup

- to build and evaluate a prototype criticism mining system that could automatically :
 - predict the **genre** of the work being reviewed
 - predict the **quality rating** assigned to the reviewed item
 - differentiate **book reviews and movie reviews**, especially for items in the same genre
 - differentiate **fiction and non-fiction book reviews**

Data set

Reviews on	Book	Movie	Music
#. Of reviews	1800	1650	1800
#. Of genres	9	11	12
Mean of review length	1,095 words	1,514 words	1,547 words
Std. Dev. of review length	446 words	672 words	784 words
Term list size	41,060	47,015	47,864

Genre Taxonomy : Music

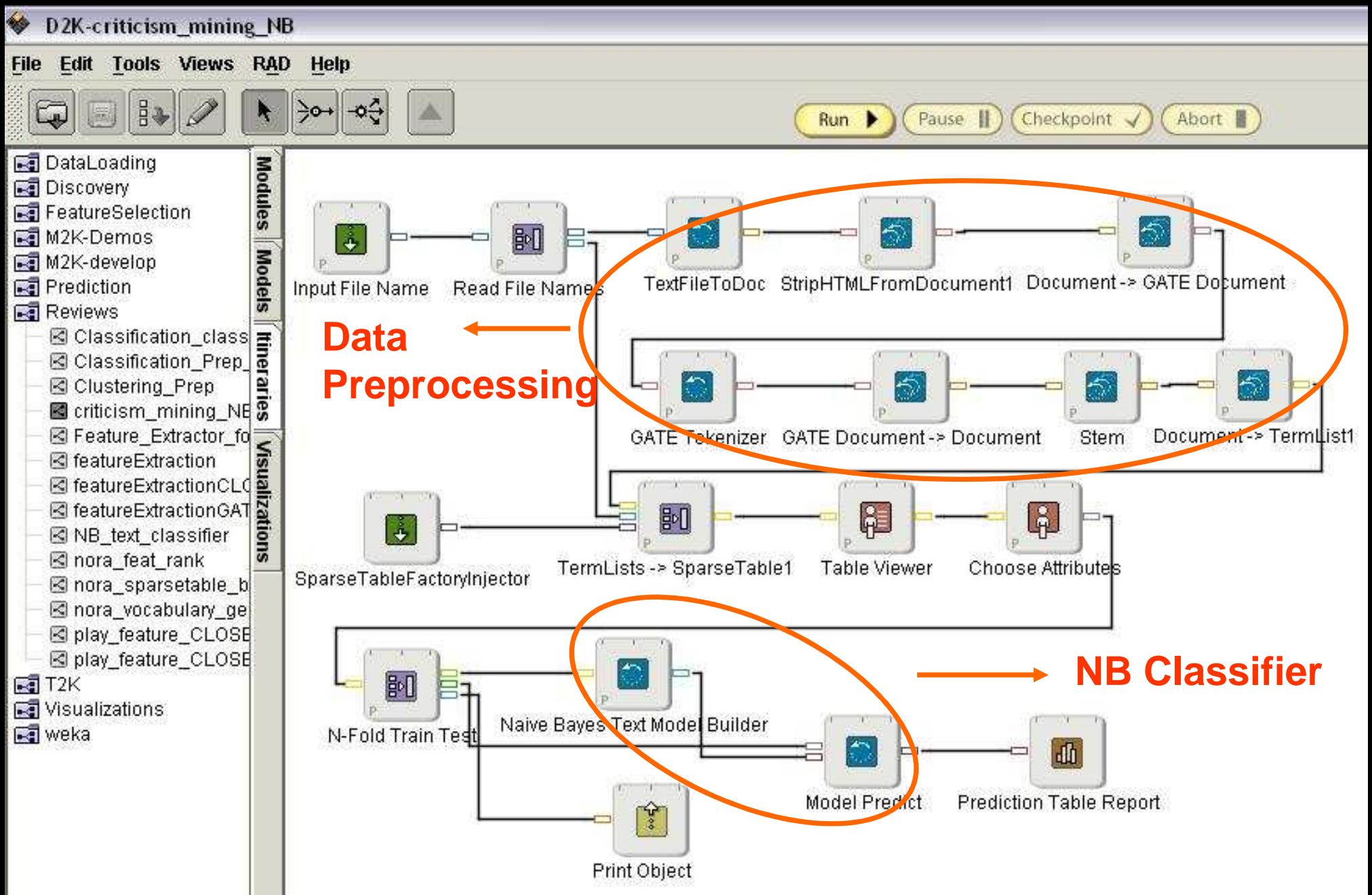
Blues	Heavy Metal
Classical	International
Country	Jazz Instrument
Electronic	Pop Vocal
Gospel	R&B
Hardcore/Punk	Rock & Pop

- The genre labels and the rating information provided the ground truth for experiments

Categorization Model & Implementation

- Naïve Bayesian (NB) Classifier
 - Computationally efficient
 - Empirically effective
- Text-to-Knowledge (T2K) Toolkit
 - A text mining framework
 - Ready-to-use modules and itineraries
 - Natural Language Processing tools integrated
 - Supporting fast prototyping of text mining

NB itinerary in T2K



Genre Classification

Reviews on	Book	Movie	Music
Number of genres	9	11	12
Reviews in each genre	200	150	150
Term list size (terms)	41,060	47,015	47,864
Mean of review length (words)	1,095	1,514	1,547
Std Dev of review length (words)	446	672	784
Mean of precision	72.18%	67.70%	78.89%
Std Dev of precision	1.89%	3.51%	4.11%

5 fold random cross validation for book and movie reviews
3 fold random cross validation for music reviews

Confusion : Music

Classified As ▶	Blu.	Cla.	Cou	Ele.	Gos.	Pun.	Met.	Int'l	Jazz	Pop.	RB	Roc.
Blues	0.61	0	0.10	0	0	0	0	0	0	0	0	0.29
Classical	0	0.94	0	0.03	0	0	0	0	0	0	0	0.03
Country	0	0	0.92	0	0.03	0	0	0	0	0	0	0.06
Electr.	0	0	0	0.92	0	0	0.06	0	0	0	0	0.03
Gospel	0	0	0.05	0	0.80	0	0	0	0	0	0.05	0.10
Punk	0	0	0	0.05	0	0.71	0.05	0	0	0	0	0.19
Metal	0	0	0	0	0	0	0.89	0	0	0	0	0.11
Int'l	0	0.04	0.00	0.04	0	0	0	0.81	0	0	0	0.04
Jazz	0	0	0	0.04	0	0	0	0	0.89	0.04	0	0.04
Pop Vo.	0	0	0.04	0.07	0	0	0	0.04	0.07	0.68	0	0.11
R&B	0	0	0	0	0	0	0	0	0	0.06	0.88	0.06
Rock	0.03	0	0.03	0	0	0	0.03	0	0	0.03	0	0.89

Rating Classification

- Five-class classification
 - 1 star vs. 2 stars vs. 3 stars vs. 4 stars vs 5 stars
- Binary Group classification
 - 1 star + 2 stars vs. 4 stars + 5 stars
- *ad extremis* classification
 - 1 star vs. 5 stars

5 fold random cross validation for all experiments

Rating : Music Reviews

Experiments	5 classes	Binary Group	<i>Ad extremis</i>
Number of classes	5	2	2
Reviews in each class	200	400	400
Term list size (terms)	35,600	33,084	32,563
Mean of review length (words)	1,875	2,032	1,842
Std Dev of review length (words)	913	912	956
Mean of precision	44.25%	79.75%	85.94%
Std Dev of precision	2.63%	3.59%	3.58%

Confusion : Music Reviews

Classified As ▶	1 star	2 stars	3 stars	4 stars	5 stars
1 star	0.61	0.24	0.07	0.05	0.02
2 stars	0.24	0.15	0.36	0.15	0.09
3 stars	0.11	0.13	0.41	0.20	0.15
4 stars	0.03	0.06	0.10	0.32	0.48
5 stars	0	0	0.09	0.11	0.80

Conclusions

- Customer reviews are an excellent resource for studying humanities materials
- Successful experiments:
 - High classification precisions:
 - Genres; Ratings; Book vs. movie reviews
 - Fiction vs. non-fiction book reviews
 - Reasonable confusions
- Text mining techniques can help find important information about the materials being reviewed

Criticism Mining : make the ever-growing consumer-generated review resources useful to humanities scholars.

HUMIRS Research Summary

1. Explicit capturing and analysis of a wide variety real-world music queries upon which to base the creation of the query records.
2. Development of formal requirements for the necessary elements (and their constituent data types) to be used in the query records.

Research Summary (Cont.)

3. Validation of the “reasonable person” relevance judgment assumption through inter-rater reliability studies.
4. Continued acquisition of more music information (audio, symbolic, and metadata) with a special effort to acquire “top hits” popular music and more non-Western musics to make real-world, real-time, user studies a possibility.

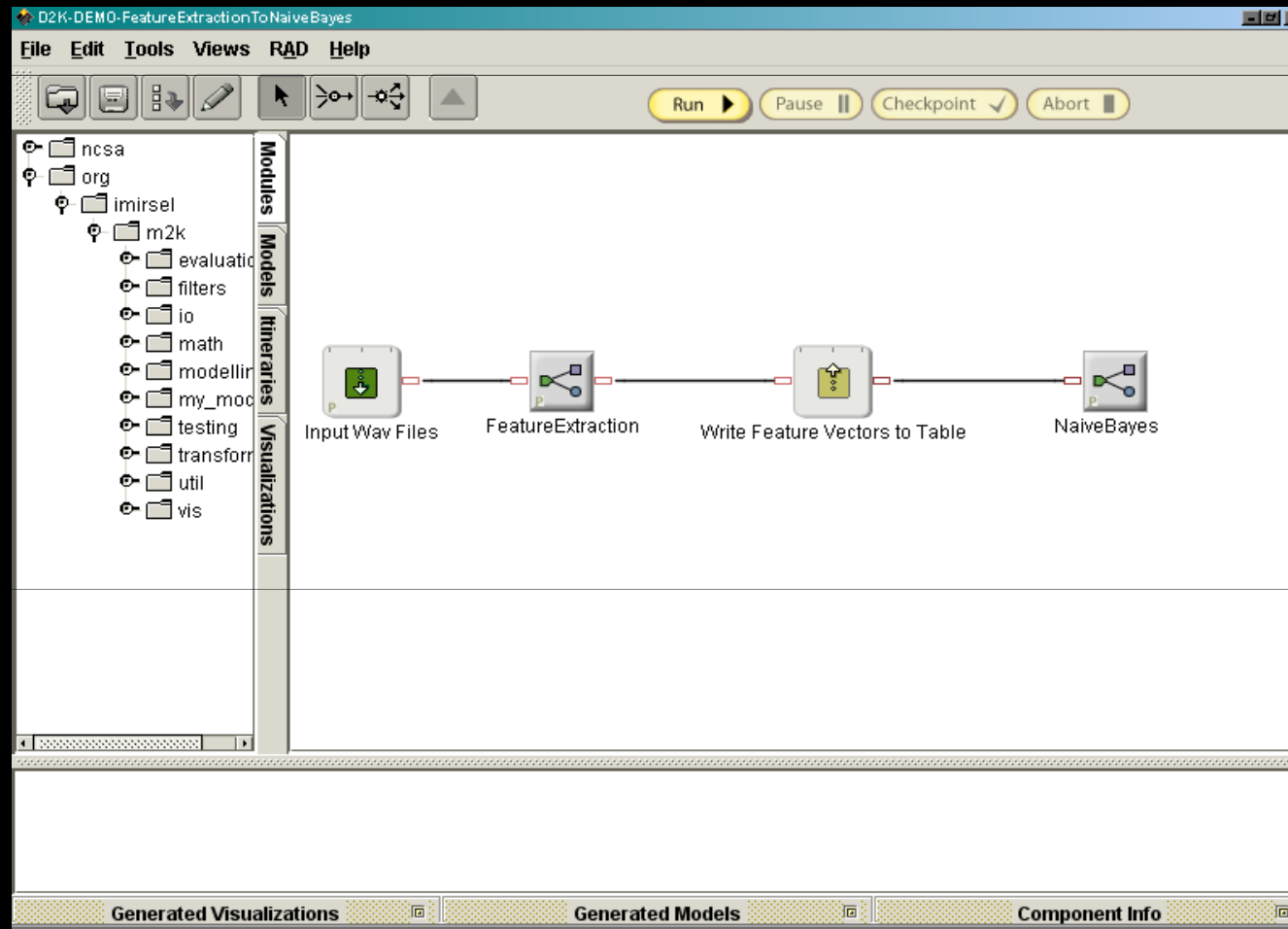
Music-to-Knowledge (M2K)

- Goal: Have both a toolset and the evaluation environment **available to researchers**
- Visual data flow programming built upon NCSA's Data to Knowledge (D2K) machine learning environment
- Java based, easily portable
- Supports distributed computing

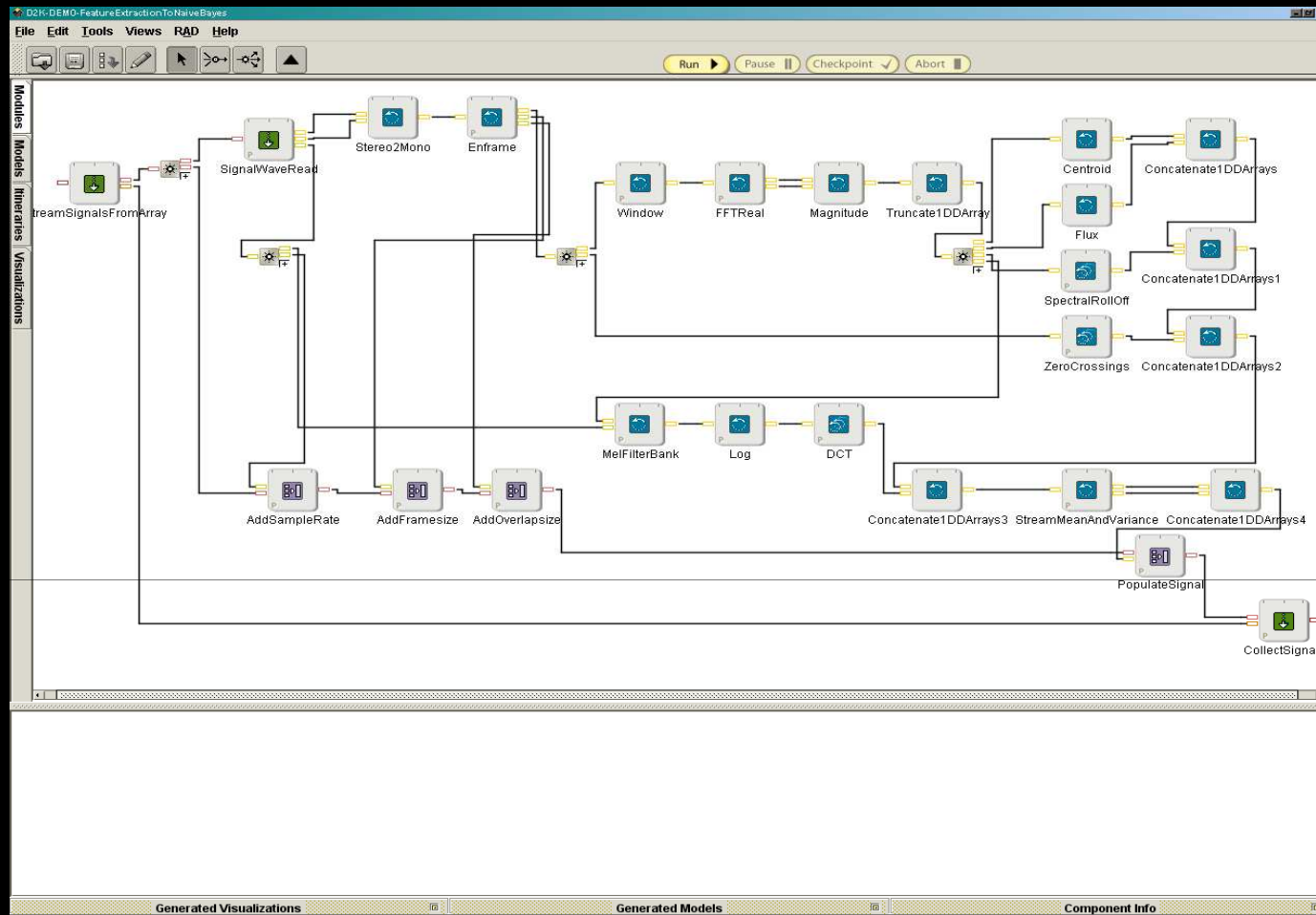
How M2K/D2K Works

- Signal processing and machine learning code is written into *modules*
- Modules are ‘wired’ together to produce more complicated programs called *itineraries*
- Itineraries can then be run or used themselves as modules allowing *nesting* of programs
- Individual modules and nested itineraries can be assigned to be *parallelized* across all machines in a network, or to individual machines in a network

A Picture is Worth 1000 Words: Music Classifier Example



Music Classifier Example: Feature Extraction Nested Itinerary



Editing Parameters and Component Documentation

The screenshot displays a software interface for editing parameters and component documentation. The main window shows a flowchart of components connected in a pipeline. The components include: Window, FFTReal, Magnitude, Truncate1DDArray, Centroid, Flux, SpectralRollOff, Concatenate1DDArrays, MelFilterBank, Log, and DCT. A 'Properties' window is open for the MelFilterBank component, showing the following parameters:

- Number of filters in the Mel-scale filter bank: Value: 32
- Low frequency edge as ratio of sample rate: Value: 0.0
- High frequency edge as ratio of sample rate: Value: 0.5

A 'Component Info' window is also open, providing details about the MelFilterBank class:

- Alias: FeatureExtraction_MelFilterBank ; sings ; Concatenate1DDArrays?
- Common Name: MelFilterBank
- Class: org. imiset. m2k. filters. MelFilterBank
- Overview: This module takes as input, a magnitude or powerspectrum in the range from $\omega = [0, \pi]$ and filters the spectrum through Mel-spaced triangular overlapping filters.
- Detailed Description: A spectrum on the range $\omega \in [0, \pi]$ represented as a 1 dimensional double array is taken as input, as well as the sample rate. Low and high frequency bounds are set as parameters as a ratio to the sample rate. Therefore, these values should be in the range of $[0.0, 0.5]$, to represent DC and the half sample rate, respectively. A Mel filterbank with a settable number of filters is constructed and applied to the inputs, producing a 1-dimensional double array output, whose dimensionality is the number of filters used.
- Data Handling: The input data is not modified.
- Inputs: Input: Magnitude-Spectrum; ID; Magnitude Spectrum (1-D double array).
- Sample Rate: java.lang.Integer; Sample Rate (int)
- Outputs: Mel Filter Outputs; ID; The bands of the Filter bank (1-D double array).
- Properties: Number of filters in the Mel-scale filter bank; Sets the number of triangular Mel spaced filters to be used in the filterbank; Low frequency edge as ratio of sample rate

M2K: Main Goals

- Promote **collaboration and sharing** through a common, modular toolset
- A ‘black box’ approach to provide commonly needed algorithms for fast prototyping
- Alleviate the ‘reinventing the wheel’ problem

TREC-Like MIR Evaluation

- Consensus holds that work should continue on developing TREC-like evaluations with the provisos that:
 1. any TREC-like approach developed be centered on the unique nature of music information and not “artificially imposed” on MIR/MDL systems simply because of the perceived “convenience” of the approach;
 2. the integration of music metadata not be overlooked; and,
 3. the TREC-like approach not become the sole means of evaluating the performance of MIR/MDL systems.

The TREC Approach

- National Institute of Standards and Technology developed a testing and evaluation paradigm for the text retrieval community, called TREC (*Text REtrieval Conference*)
- Under this paradigm, each text retrieval team is given access to:
 1. a standardized, large-scale test collection of text;
 2. a standardized set of test queries; and,
 3. a standardized evaluation of the results each team generates.

TREC Metrics

- Cranfield experiments of the early 1960's
- Two metrics have proved themselves to be particularly useful and important:
 - *precision* (i.e., the ratio of relevant documents retrieved to the number of documents retrieved)
 - *recall* (i.e., the ratio of relevant documents retrieved to the number of relevant documents present in the system)..

MIREX Overview

- Began in 2005
- Tasks defined by community debate
- Data sets collected and/or donated
- Participants submit code to IMIRSEL
- Code rarely works first try 😊
- Data collections tend to contain corrupt files
- Meet at ISMIR to discuss results

MIREX 2005 Overview

Contest Name	Submissions	Countries	Individuals
Audio Artist Identification	8	5	13
Audio Drum Detection	7	7	10
Audio Genre Classification	13	11	21
Audio Key Detection	5	3	6
Audio Melody Extraction	8	7	12
Audio Onset Detection	7	5	11
Audio Tempo Detection	8	6	12
Symbolic Genre Classification	5	4	9
Symbolic Key Detection	5	3	6
Symbolic Melodic Similarity	6	6	15

MIREX - Music Classification Tasks

- Audio genre classification
 - Classify music into one of approximately 10 genres
 - Top accuracy: 82.34%
- Audio artist classification
 - Classify music into one of 73 artists
 - Top accuracy: 72.45%
- Symbolic genre classification
 - Classify MIDI files into one of 38 genres in a genre taxonomy
 - Top accuracy: 64.33%

MIREX - Transcription Tasks

- Note onset detection
 - Determine onset times of all musical events
 - Top F1-Measure: 80.07%
- Audio melody extraction
 - Transcribe the fundamental frequency of the main melodic voice in polyphonic mixtures
 - Top accuracy: 71.40%
- Audio drum transcription
 - Transcribe bass drum, snare drum, and hihat in music
 - Top F1-Measure: 67.00%

MIREX - Other Audio Tasks

- Perceptual tempo induction
 - Determine the two dominant tempi in a piece, as well as their salience and alignment in time
 - Top weighted accuracy: 68.9%
- Audio key finding
 - Determine the key of a 1252 pieces, tonic and mode
 - Top accuracy: 89.55%

MIREX - Symbolic Tasks

- Symbolic key finding
 - Determine the key of 1252 pieces, tonic and mode
 - Top accuracy: **91.40%**
- Symbolic melodic similarity
 - Produce a ranked list of *incipits* with similar melody to 11 queries
 - Top average dynamic recall: **65.98%**

Lessons from MIREX 2005

- Standardize evaluation metrics
 - Necessary for comparisons over time
- Significance tests and measures
 - Huge problems with underlying assumptions
 - Can yield insight into how algorithms can be combined
- Annotation **communities** and toolsets
 - Ground truth is a most precious resource
- Resolve dataset issues
 - Development vs. testing sets

Other Consensus Items

- Four ongoing problem areas identified:
 1. The complexity of music information
 2. The complexity of music queries
 3. The nature of “relevance” in the MIR context
 4. The lack of access to music collections caused by the current IP legal environment

2006 General Statistics

- 13 tasks
- 46 teams
- 50 individuals
- 14 different countries
- 10 different programming languages and execution environments
- 92 individual runs
- 98 result data matrices on the wiki

MIREX 2006 Tasks

- Audio Beat Tracking
- Audio Cover Song Identification
- Audio Melody Extraction (2 subtasks)
- Audio Music Similarity and Retrieval
- Audio Onset Detection
- Audio Tempo Extraction
- Query-by-Singing or Humming (2 subtasks)
- Score Following
- Symbolic Melodic Similarity (3 subtasks)

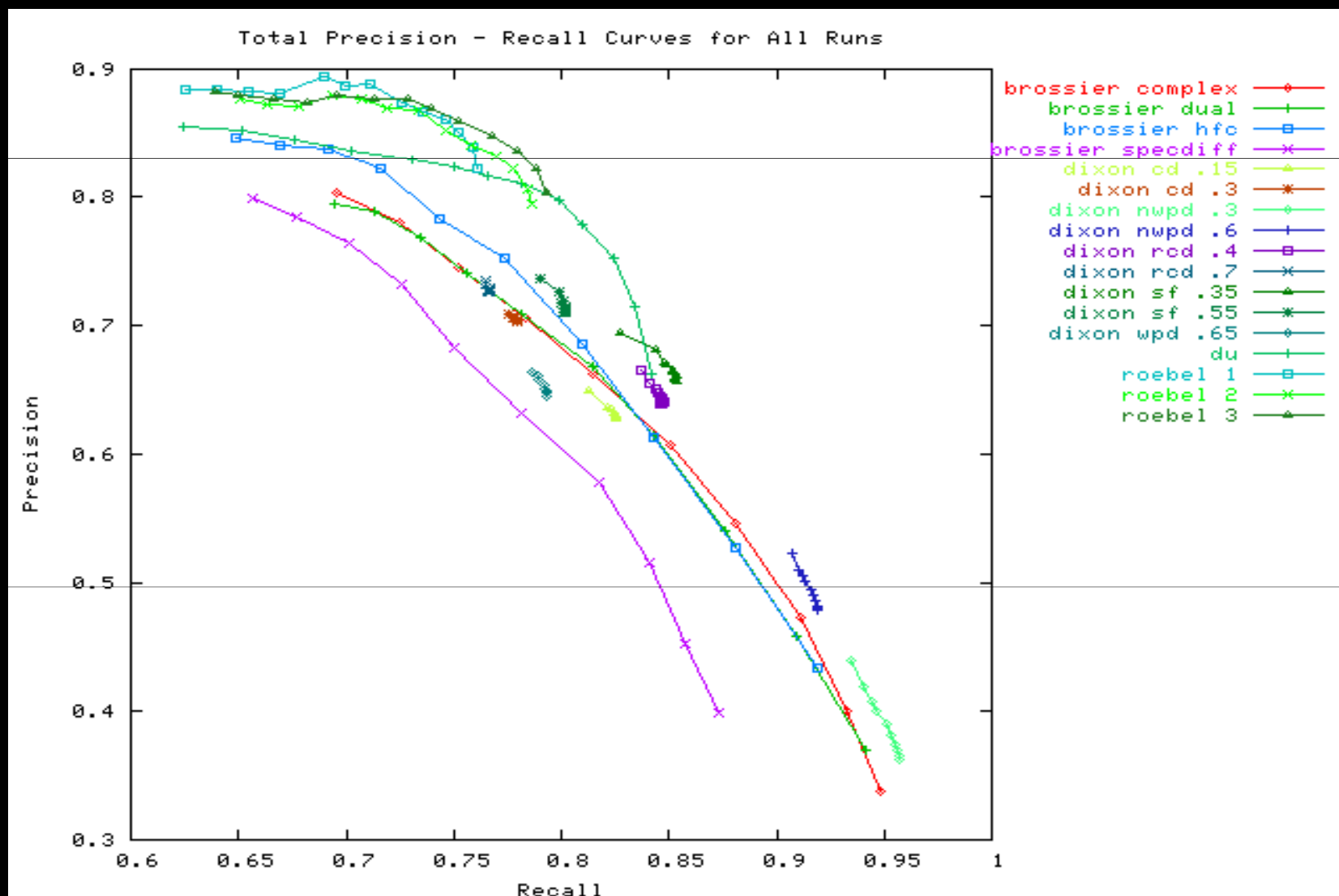
MIREX 2006 Tasks

- Audio Beat Tracking
- **Audio Cover Song Identification**
- Audio Melody Extraction (2 subtasks)
- **Audio Music Similarity and Retrieval**
- Audio Onset Detection
- Audio Tempo Extraction
- Query-by-Singing or Humming (2 subtasks)
- Score Following
- Symbolic Melodic Similarity (3 subtasks)

New for MIREX 2006

- New Tasks
 - Audio Cover Song
 - Score Following
 - QBSH
- New Evaluations
 - Multiple parameters in Onset Detection
 - Evalutron 6000: Human similarity judgments
 - Friedman tests

Onset Detection



MIREX 2006 Tasks

- Audio Beat Tracking
- Audio Cover Song Identification
- Audio Melody Extraction (2 subtasks)
- **Audio Music Similarity and Retrieval (AMS)**
- Audio Onset Detection
- Audio Tempo Extraction
- Query-by-Singing or Humming (2 subtasks)
- Score Following
- **Symbolic Melodic Similarity (3 subtasks) (SMS)**

Scoring Scheme

- Three **BROAD** categories:
 - Not Similar (NS)
 - Somewhat Similar (SS)
 - Very Similar (VS)
- Continuous **FINE** score range:
 - 0 (no similarity at all)
 - 10 (absolute similarity)

Differences in 2006

Evaluation Task Descriptions

- Evaluate how well various algorithms retrieve results that are...
- **SMS**...MELODICALLY similar to a given query. You will find in the candidate files a variety of different instrumentations as set by the creators of the MIDI files. We need you to look beyond the differences in timbre and instrumentation in assigning your grading scores.
- **AMS**...MUSICALLY similar to a given query. You will be presented with files from a number of different music genres. Please assign the scores according to what you find 'sounds' similar and do not take into account whether you like the music or not.

Evalutron 6000

mirex EVALUTRON 6000 EVALUTRON 6000 SANDBOX VERSION

Welcome sandbox1 [Sign out](#) [Change My Settings](#)

Home Audio Player Selection My Assignment Instructions

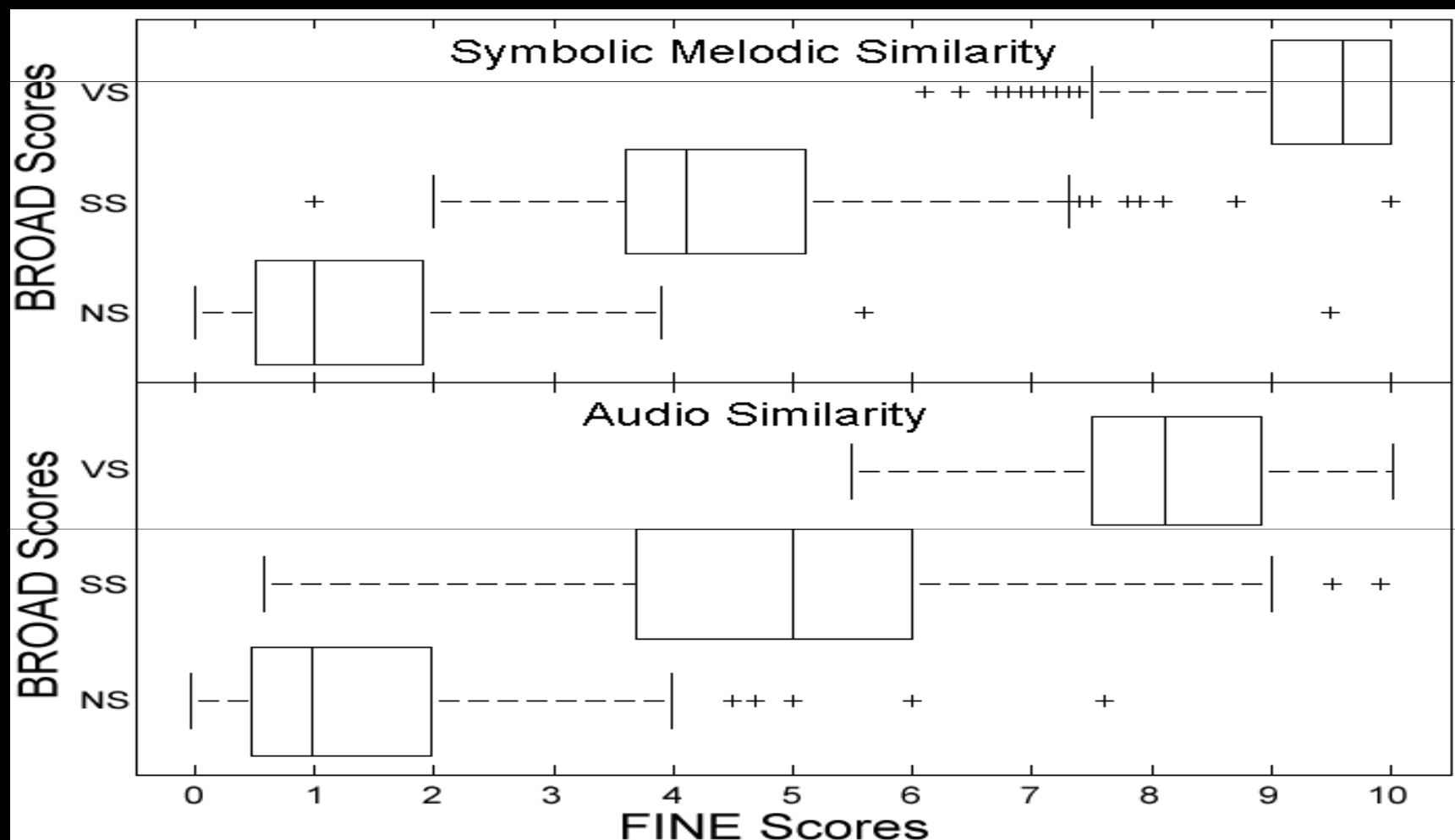
THIS PAGE CONTAINS 10 CANDIDATES FOR QUERY ID # 2 < Previous Query Next Query >

<p>Query ID#2</p> <p>Audio Player</p> <p>First Mid Last</p> <p>Align Player</p>	<p>Listen to Candidate #b011638</p> <p>Audio Player</p> <p>First Mid Last</p>	<p>Select Broad Category</p> <p><input checked="" type="radio"/> NOT Similar <input type="radio"/> Somewhat Similar <input type="radio"/> VERY Similar</p> <p>[SAVED]</p>	<p>Select Fine Score</p> <p>0 10</p> <p>0 [SAVED]</p>
<p>Align Player</p>	<p>Listen to Candidate #b011614</p> <p>Audio Player</p> <p>First Mid Last</p>	<p>Select Broad Category</p> <p><input checked="" type="radio"/> NOT Similar <input type="radio"/> Somewhat Similar <input type="radio"/> VERY Similar</p> <p>[SAVED]</p>	<p>Select Fine Score</p> <p>0 10</p> <p>0 [SAVED]</p>
<p>Align Player</p>	<p>Listen to Candidate #b011644</p> <p>Audio Player</p> <p>First Mid Last</p>	<p>Select Broad Category</p> <p><input checked="" type="radio"/> NOT Similar <input type="radio"/> Somewhat Similar <input type="radio"/> VERY Similar</p> <p>[SAVED]</p>	<p>Select Fine Score</p> <p>0 10</p> <p>0 [SAVED]</p>
<p>Align Player</p>	<p>Listen to Candidate #b011624</p> <p>Audio Player</p> <p>First Mid Last</p>	<p>Select Broad Category</p> <p><input checked="" type="radio"/> NOT Similar <input type="radio"/> Somewhat Similar <input type="radio"/> VERY Similar</p> <p>[SAVED]</p>	<p>Select Fine Score</p> <p>0 10</p> <p>0 [SAVED]</p>
<p>Align Player</p>	<p>Listen to Candidate #b011647</p> <p>Audio Player</p> <p>First Mid Last</p>	<p>Select Broad Category</p> <p><input checked="" type="radio"/> NOT Similar <input type="radio"/> Somewhat Similar <input type="radio"/> VERY Similar</p> <p>[SAVED]</p>	<p>Select Fine Score</p> <p>0 10</p> <p>[] [Waiting]</p>

Evalutron 6000 Data

	SMS	AMS
No. of events logged	23,491	46,254
No. of submitted algorithms	8	6
Total no. of queries	17	60
Total no. of query-candidate pairs	905	1,629
Total no. of evaluations	2,715	4,887
No. of graders	21	24
No. of queries per grader	15	7-8
Avg. size of candidate lists	15	27
Avg. no. of evaluations per grader	225	205

Scoring Distributions



Grader Agreement

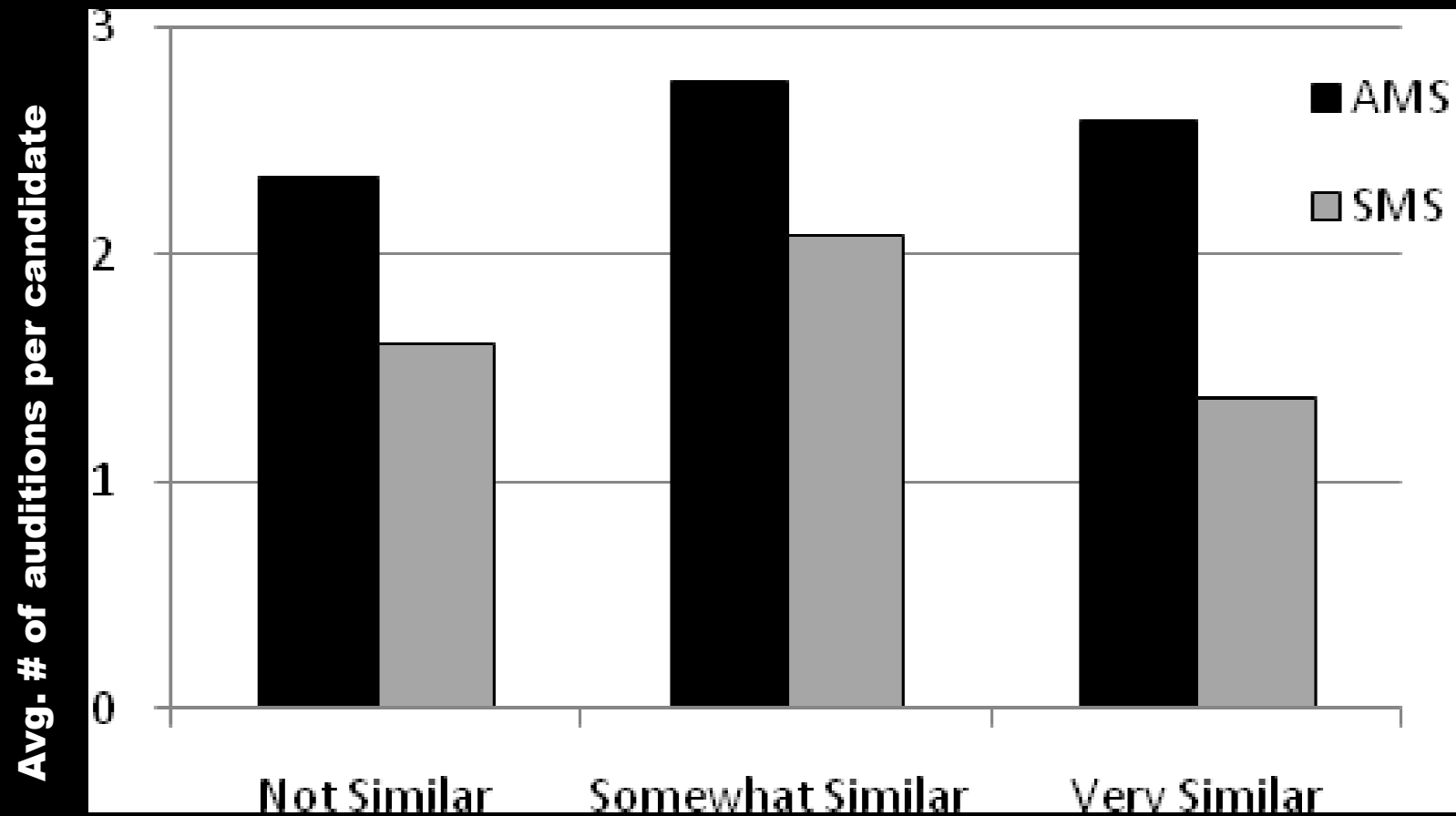
	3-Level SMS		3-Level AMS	
VS,VS,VS	114	12.6%	61	3.7%
SS,SS,SS	38	4.2%	137	8.4%
NS,NS,NS	263	29.1%	293	18.0%
Total triples	415	45.9%	491	30.1%
VS,VS	24	2.7%	150	9.2%
SS,SS	158	17.5%	469	28.8%
NS,NS	288	31.8%	404	24.8%
Total doubles	470	51.9%	1023	62.8%
VS,SS,NS	20	2.2%	115	7.1%
Total	905	100.0%	1629	100.0%

Fleiss's Kappa Agreement Metrics

	3-Level (NS, SS, VS)	2-Level (S, NS)
SMS	0.3664	0.3201
AMS	0.2141	0.2989

Each value falls in the “FAIR” range (Landis and Koch, 1977)

Differences in Evaluator Effort



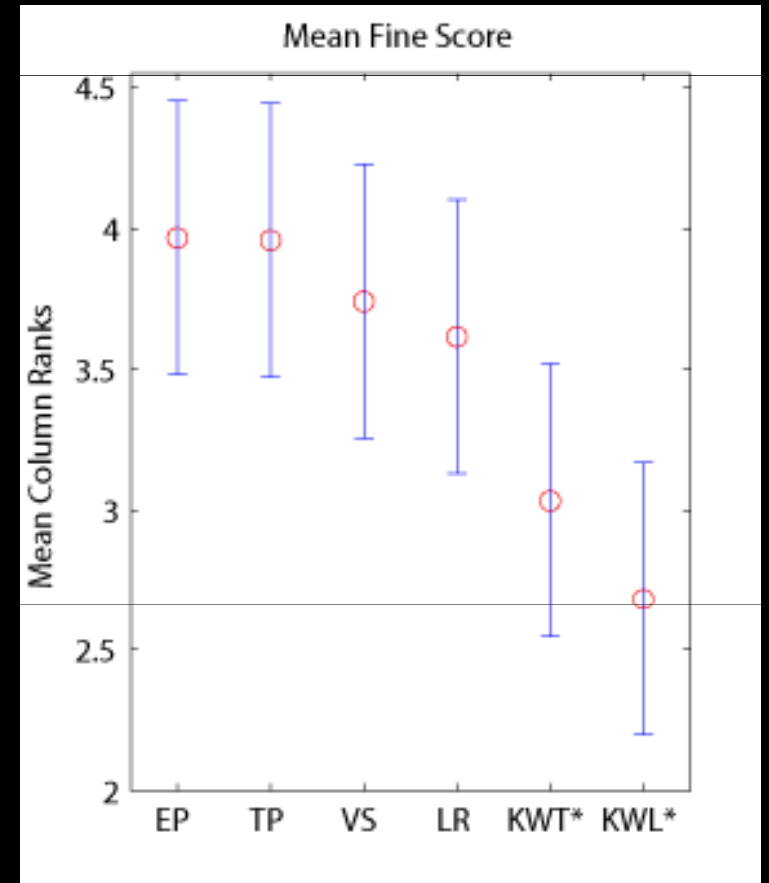
Closing Comments and Suggestions

- Task definitions appear significant
 - Press the community to clarify (esp. AMS)
- Drop the three grader per query-candidate pair in favour of more queries examined
- Keep the 3-Level Broad Score (for now)
- Continue more in-depth examination of the Broad Score/Fine Score relationship

Friedman Tests

Audio Music Similarity and Retrieval

Friedman's ANOVA Table					
Source	SS	df	MS	Chi-Sq	Prob>Chi-Sq
Columns	84.733	5	16.947	24.291	0.000
Error	961.767	295	3.260		
Total	1046.50	359			



Friedman's Test: Multiple Comparisons

TeamID	TeamID	Lowerbound	Mean	Upperbound	Significance
EP	TP	-0.963	0.008	0.980	FALSE
EP	VS	-0.755	0.217	1.188	FALSE
EP	LR	-0.630	0.342	1.313	FALSE
EP	KWT	-0.030	0.942	1.913	FALSE
EP	KWL	0.320	1.292	2.263	TRUE
TP	VS	-0.763	0.208	1.180	FALSE
TP	LR	-0.638	0.333	1.305	FALSE
TP	KWT	-0.038	0.933	1.905	FALSE
TP	KWL	0.312	1.283	2.255	TRUE
VS	LR	-0.847	0.125	1.097	FALSE
VS	KWT	-0.247	0.725	1.697	FALSE
VS	KWL	0.103	1.075	2.047	TRUE
LR	KWT	-0.372	0.600	1.572	FALSE
LR	KWL	-0.022	0.950	1.922	FALSE
KWT	KWL	-0.622	0.350	1.322	FALSE

IMIRSEL's Future MIREX Plans

- Continue to explore more statistical significance testing procedures beyond Friedman's ANOVA test
- Continue refinement of Evalutron 6000 technology and procedures
- Continue to establish a more formal organizational structure for future MIREX contests
- Continue to develop the evaluator software and establish an open-source evaluation API
- Make useful evaluation data publicly available year round
- Establish a webservices-based IMIRSEL/M2K online system prototype (Audio Cover Song Identification?)

Ongoing Discussion Items

- **General Discussion Items**
 - **Task vs. Contest issues**
 - **Rewarding strong submissions**
 - **Standardization of I/O formats (across tasks & years)**
 - **Communication issues**
 - **Scheduling issues**
 - **Data provider/participant issue**
- **Data Issues**
 - **Testing-data shortages**
 - **Metadata availability and quality**
 - **Locating and generating new datasets**
 - **Re-using datasets**

Discussion

- **Submission Issues**
 - **Robustness & dirty data**
 - **Platform independence, static linking, “hard-coded” paths and values**
 - **Scalability issues – 5000+ files**
 - **Parallelizability (esp. feature extraction)**
- **MIREX 2007 Planning**
 - **Mid-year planning session (Vienna, April 2007)**
 - **New task ideas**
 - **Tweaks to current tasks**
 - **Possible new data set**

Introducing.....

mirex web service



But wait, it gets better.....

mirex web service



The Amazing.....



Extra-Special Thanks to....

- Andrew W. Mellon Foundation
- National Science Foundation
- The MIR/MDL community
- All the fine folks of SSMS!

Thanks Again!

And Some Links of Course!

- Main IMIRSEL page
 - <http://music-ir.org/evaluation/>
- M2K Video Demo
 - <http://music-ir.org/m2kvid/>
- MIREX DIY Demo
 - <http://cluster3.lis.uiuc.edu:8080/mirexdyidemo>
- MIREX Wiki
 - <http://music-ir.org/mirexwiki>