User Interfaces and Gamification: Design and Evaluation

Andreas Nürnberger and Sebastian Stober
Data & Knowledge Engineering Group, Faculty of Computer Science
Otto-von-Guericke-Universität Magdeburg, Germany

Email: andreas.nuernberger@ovgu.de, sebastian.stober@ovgu.de

Outline

- Day 1: Adaptation and Personalization: Concepts and Challenges
- Day 2: Adaptive Music Retrieval: An Overview
- Day 3: Adaptive Hierarchies: Constrained Clustering and Utility
- Day 4: Adaptive Music Similarity
- **Day 5: User Interfaces and Gamification: Design and Evaluation**

Definition

Gamification is the use of game thinking and game mechanics in a non-game context to engage users and solve problems.

[wikipedia]

CHALLENGE:

How can we use gamification for evaluation?

Use of Gamification in MIR

a) to collect ground truth

b) to give test users a concrete task

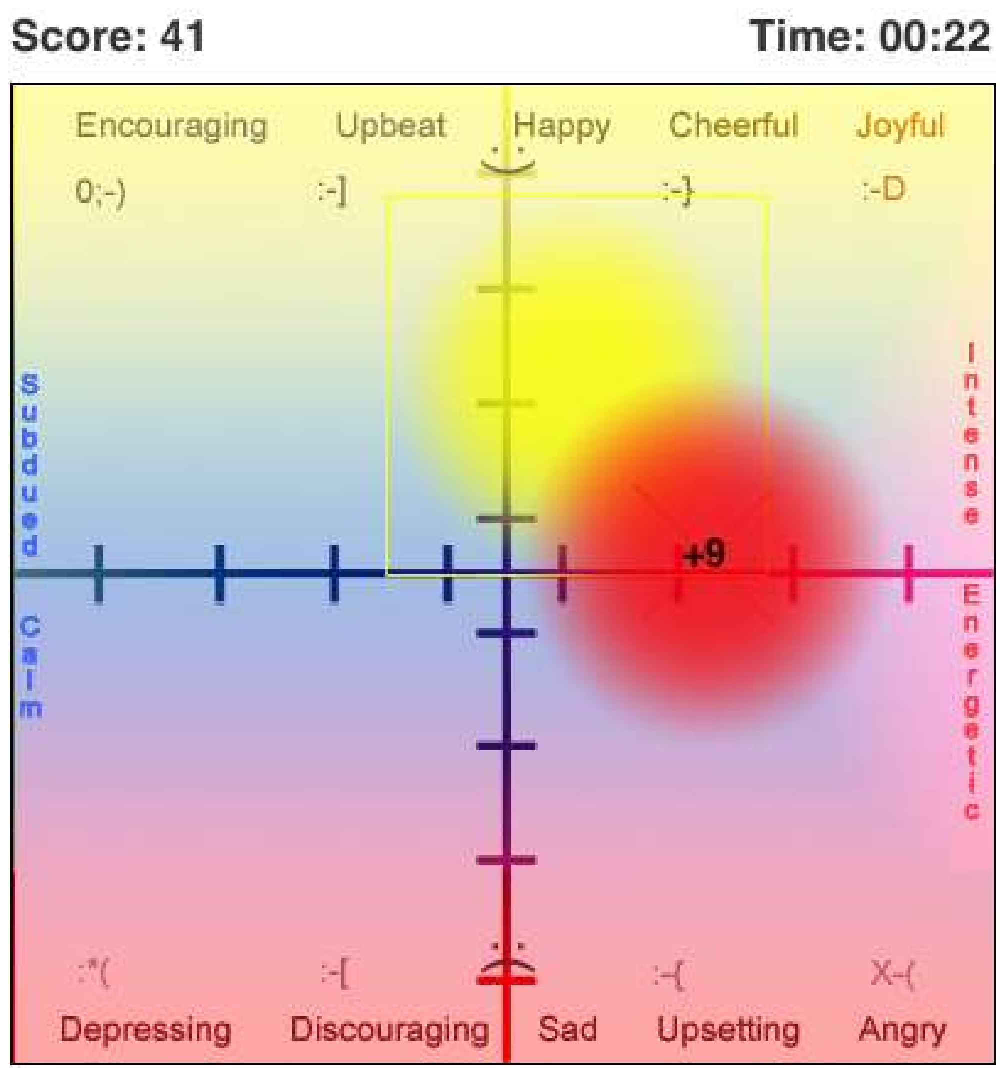
c) …

Input Agreement Games

MoodSwings:

A Collaborative Game For Music Mood Label Collection

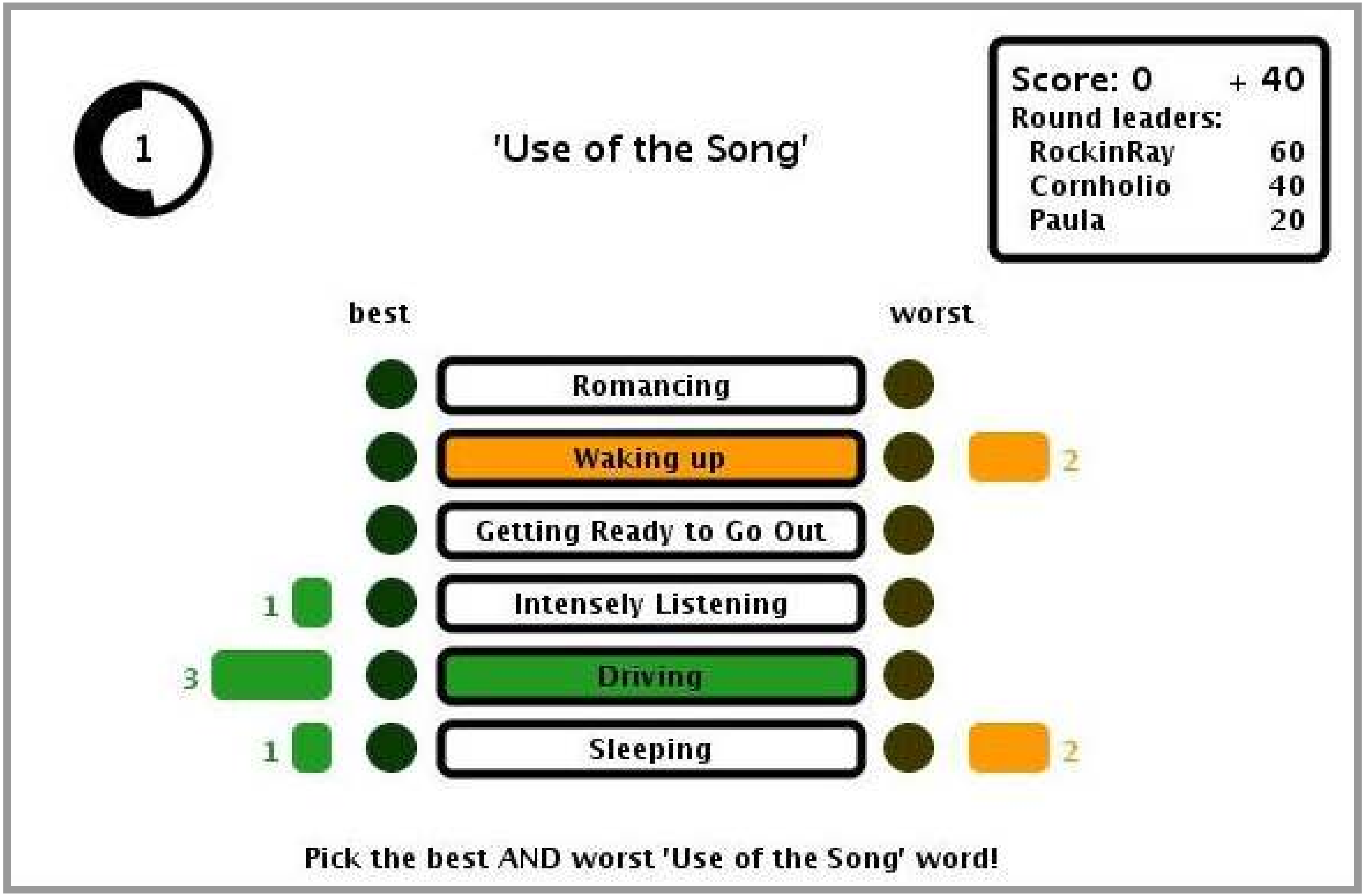
Kim et al., ISMIR 2008



Input Agreement Games

[Michael Mandel 2007, no longer available online]

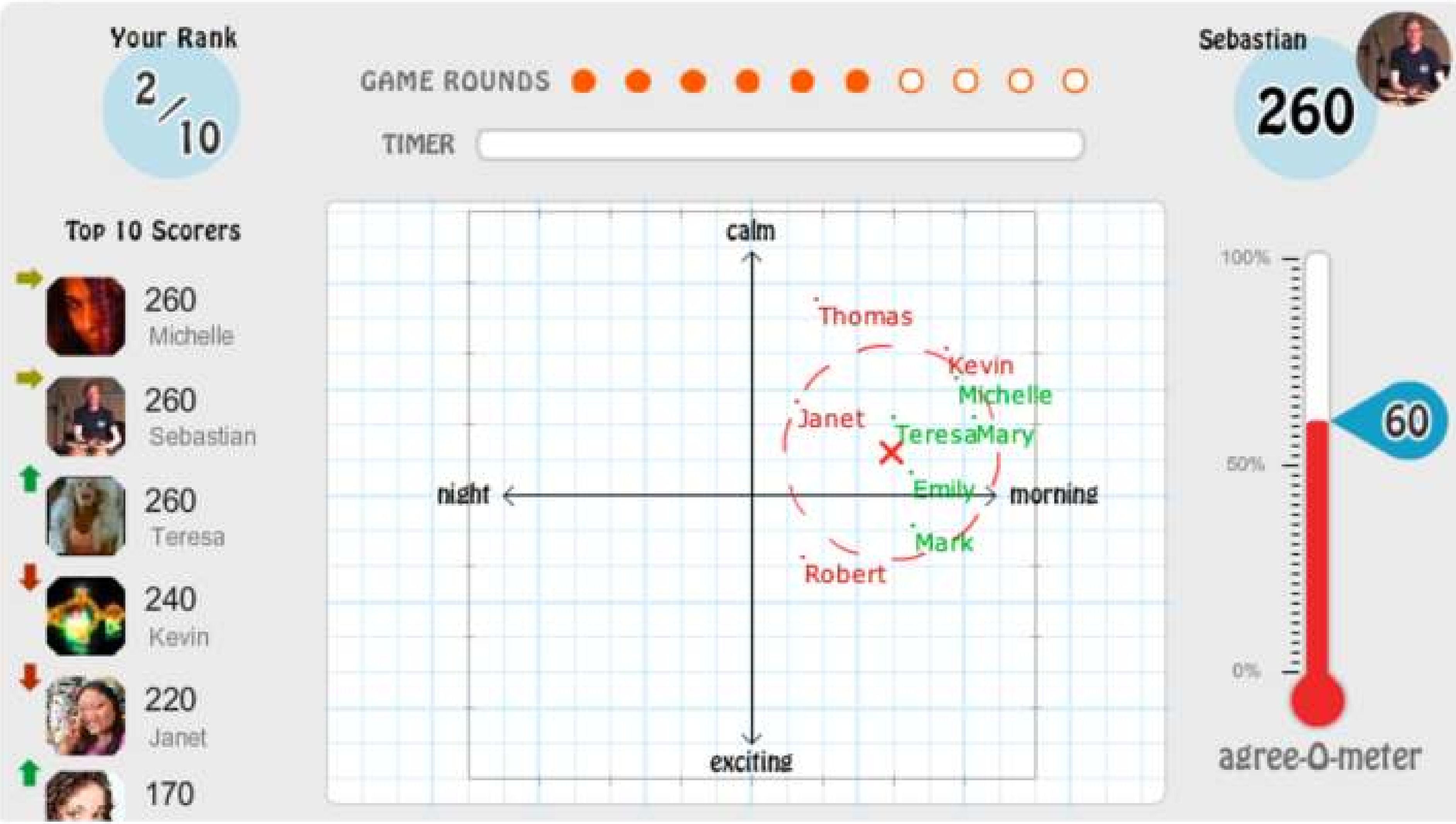
Input Agreement Games



[Douglas Turnbull 2007, no longer available online]

Input Agreement Games

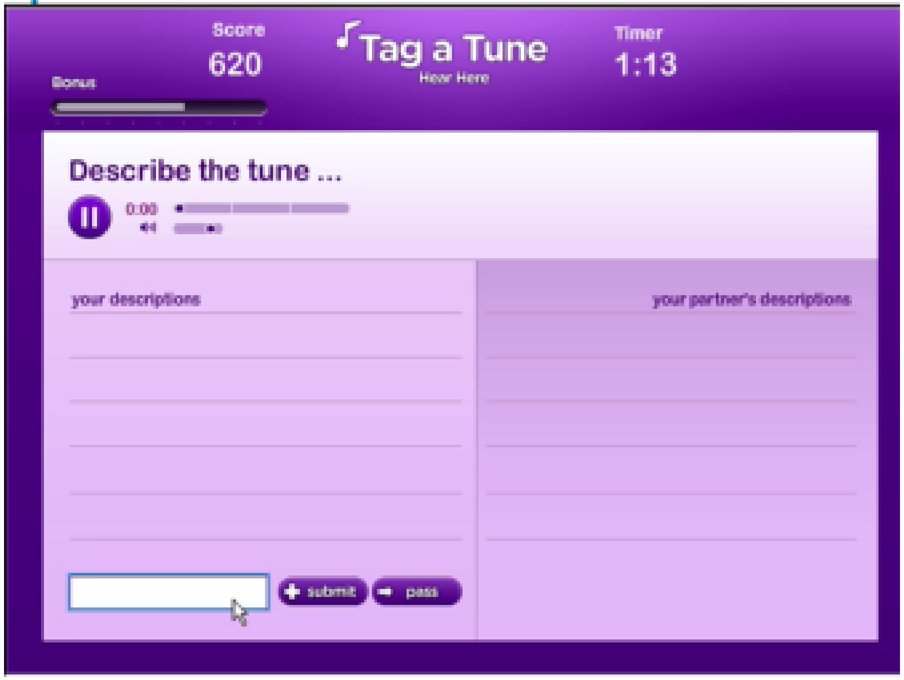
Herd It [UCSD, http://herdit.org/], play @ facebook



TagATune – Music Annotation Game

- http://tagatune.org/

play @ http://www.gwap.com/gwap/gamesPreview/tagatune/



- multiplayer
- describe music and find out whether same song
- bonus round for similarity judgments

Example 1:

COMPARING SIMILARITY ADAPTATION APPROACHES

Experimental Setup: Dataset

- Magnatagatune
  - 25863 clips from 5405 source MP3s (446 albums, 230 artists)
  - extracted features
  - tagged by players (188 unique tags)
  - similarity judgments (bonus round)
    - 533 different clip-triples
    - players vote for most different clip (7650 votes in total)
- notes:
  - used only global features and aggregated local ones
  - added new EchoNest features “dancability” and “energy”
  - added genre tags from Magnatagatune
  - preprocessed tags
  - similarity judgments inconsistent

Experimental Setup: Dataset – Tags

- tag preprocessing:
  1. merging of singular and plural forms
     e.g., "guitar" and "guitars"
  2. spelling correction
     e.g., "harpsicord" → "harpsichord"
  3. combination of semantically identical tags
     e.g., "funk" and "funky"
  4. creation of meta-tags with higher coverage for groups of tags that express the same concept
     e.g., "instrumental" = "instrumental" or "no vocal(s)" or "no voice(s)" or "no singer(s)" or "no singing"
  5. removal of unused tags
     (w.r.t. the relevant subset of Magnatagatune)

Experimental Setup: Features & Facets

<table>
<tr><th>feature</th><th>dim.</th><th>value description</th><th>#facets</th></tr>
<tr><td>key</td><td>1</td><td>0 to 11 (one of the 12 keys) or -1 (none)</td><td></td></tr>
<tr><td>mode</td><td>1</td><td>0 (minor), 1 (major) or -1 (none)</td><td></td></tr>
<tr><td>loudness</td><td>1</td><td>overall value in decibel (dB)</td><td></td></tr>
<tr><td>tempo</td><td>1</td><td>in beats per minute (bpm)</td><td>1 each</td></tr>
<tr><td>time signature</td><td>1</td><td>3 to 7 ($\frac{3}{4}$ to $\frac{7}{4}$), 1 (complex), or -1 (none)</td><td></td></tr>
<tr><td>danceability</td><td>1</td><td>between 0 (low) and 1 (high)</td><td></td></tr>
<tr><td>energy</td><td>1</td><td>between 0 (low) and 1 (high)</td><td></td></tr>
<tr><td>pitch mean</td><td>12</td><td>dimensions correspond to pitch classes</td><td>1 | 12</td></tr>
<tr><td>pitch std. dev.</td><td>12</td><td>dimensions correspond to pitch classes</td><td>1 | 12</td></tr>
<tr><td>timbre mean</td><td>12</td><td>normalized timbre PCA coefficients</td><td>1 | 12</td></tr>
<tr><td>timbre std. dev.</td><td>12</td><td>normalized timbre PCA coefficients</td><td>1 | 12</td></tr>
<tr><td>tags</td><td>99</td><td>binary vector (very sparse)</td><td>14 | 99</td></tr>
<tr><td>genres</td><td>44</td><td>binary vector (very sparse)</td><td>1</td></tr>
</table>

top: global features, middle: aggregated features, bottom: tags 26 | 155

Experimental Setup: Constraints

- "clip c is the most dissimilar of (a,b,c)" (1 vote)
  ➠ d(a,b) < d(a,c) & d(a,b) < d(b,c) (2 constraints)

- problem: contradictions

➠ graph-based constraints filtering [McFee et al. '09]:

1. construct directed multigraph
   - nodes = clip pairs
   - edges = relative distance constraints
   - (a,b)→(a,c) ⇔ constraint d(a,b) < d(a,c) exists

   ➠ 15300 edges (1598 unique)
2. remove length 2 cycles

   ➠ 6898 edges (860 unique)
3. construct directed acyclic graph (randomized, greedy)
   - start with no edges
   - add edges in random order
   - omit edges that introduce cycles

   ➠ no change

Experimental Setup: Algorithms

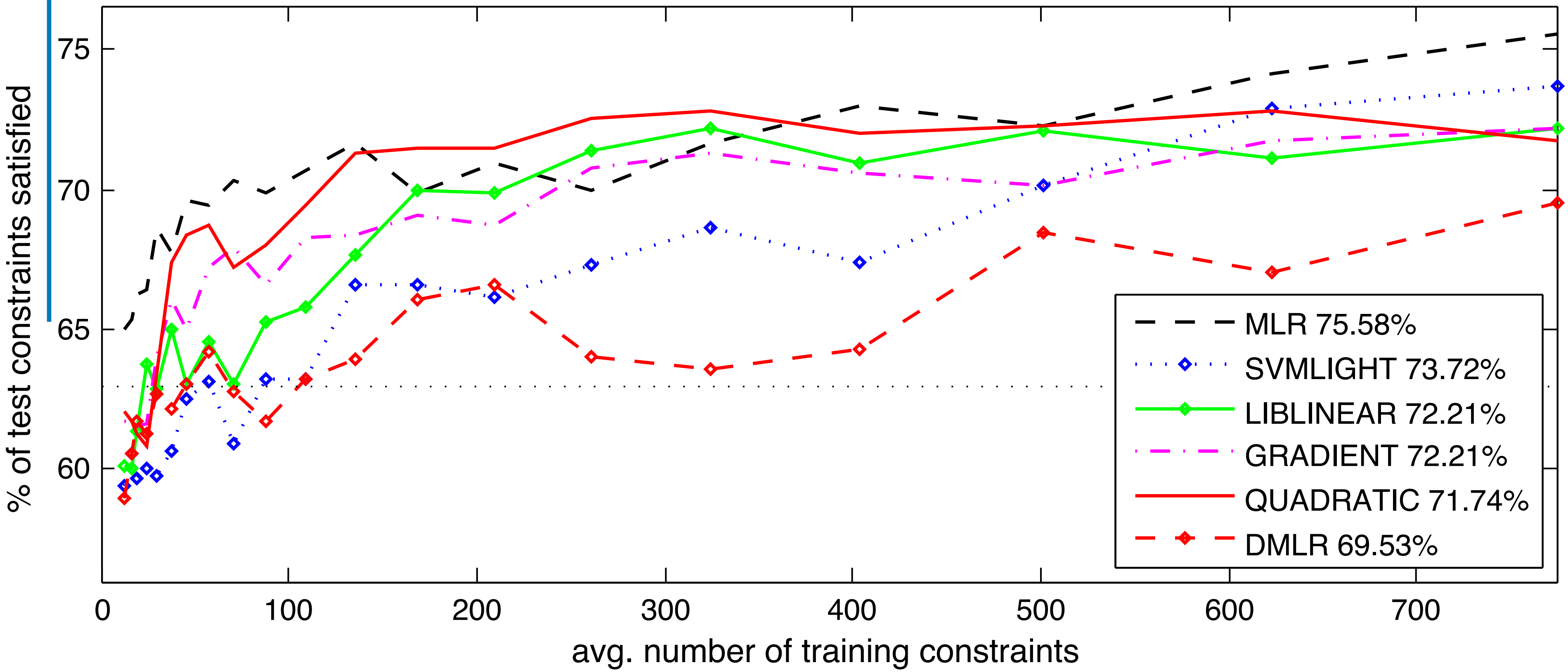
- linear facet-based approaches using 26 and 155 facets
  - Gradient Following
  - Quadratic Programming (sum(slack^2))
  - Linear SVM (LibLinear)*
- Mahalanobis distance learners using raw feature vectors
  - Linear SVM ($\text{SVM}^{\text{light}}$)*
    - restricted to diagonal *W*
    - much like LibLinear, but features are point-wise squared difference vectors, i.e. for constraint (s,a,b) : $x = (s-b)^2 - (s-a)^2$
  - Metric Learning to Rank (MLR)
  - diagonal Metric Learning to Rank (DMLR)

*soft weight constraints (may be violated)

Experimental Setup: Data Partitioning

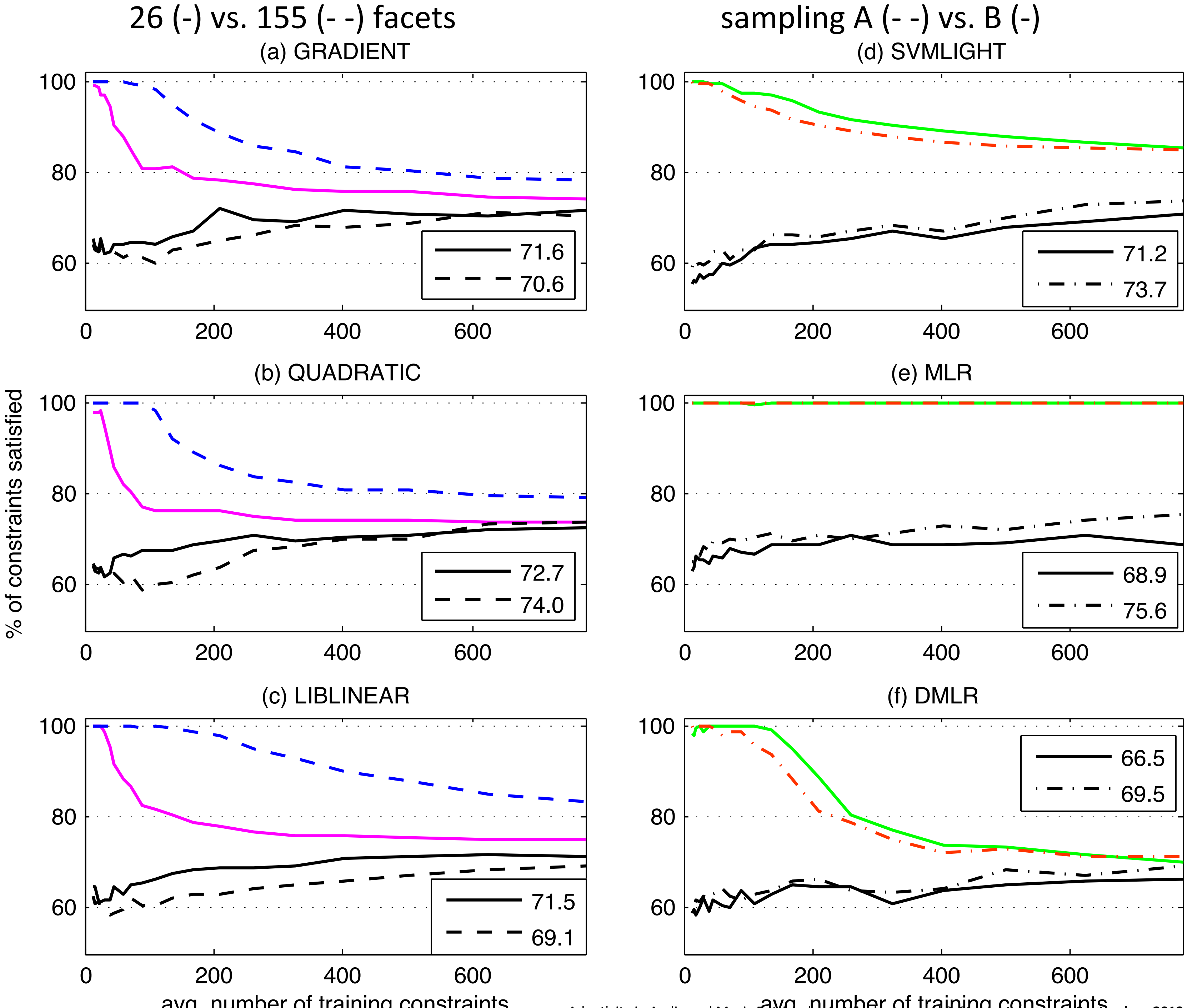
- generally: 10-fold cross-validation
- sampling variants:
  A. random sampling of constraints
     - 774 constraints for training, 86 for testing
  B. random sampling of clips/triplets
     - all constraints refering to the same clip belong to same bin
     - effectively: sampling 337 graph components (triplets)
     - bins of 33 or 34 triplets with 2 or 3 constraints per triplets
     - 770-779 constraints for training, 81-90 for testing
- training sets are expanded exponentially starting with 13 constraints (A) or 5 triples (B)

Results – 26 Facets vs. Metric Learning



- averaged over 20 folds on sampling A
- baseline (random facet weights, n=1000) @ 63%

Results – Details



Observations

- effect of #facets:
  - 155 facets much better on train but worse on test
  - performance match only with many constraints
  - classical over-fitting (simpler model generalizes quicker)
  - for 26 facets, QP almost meets upper bound (train performance)
  - 155 facets increase upper bound for QP by 5%
- effect of sampling:
  - MLR preformance drops by 6% on sampling B!
    - seems to be sensitive to sampling method
- MLR maintains 100% on training data
- QP copes best with constraint sets it cannot fulfil
  (quick adaptation to a good trade-off)

Future Directions

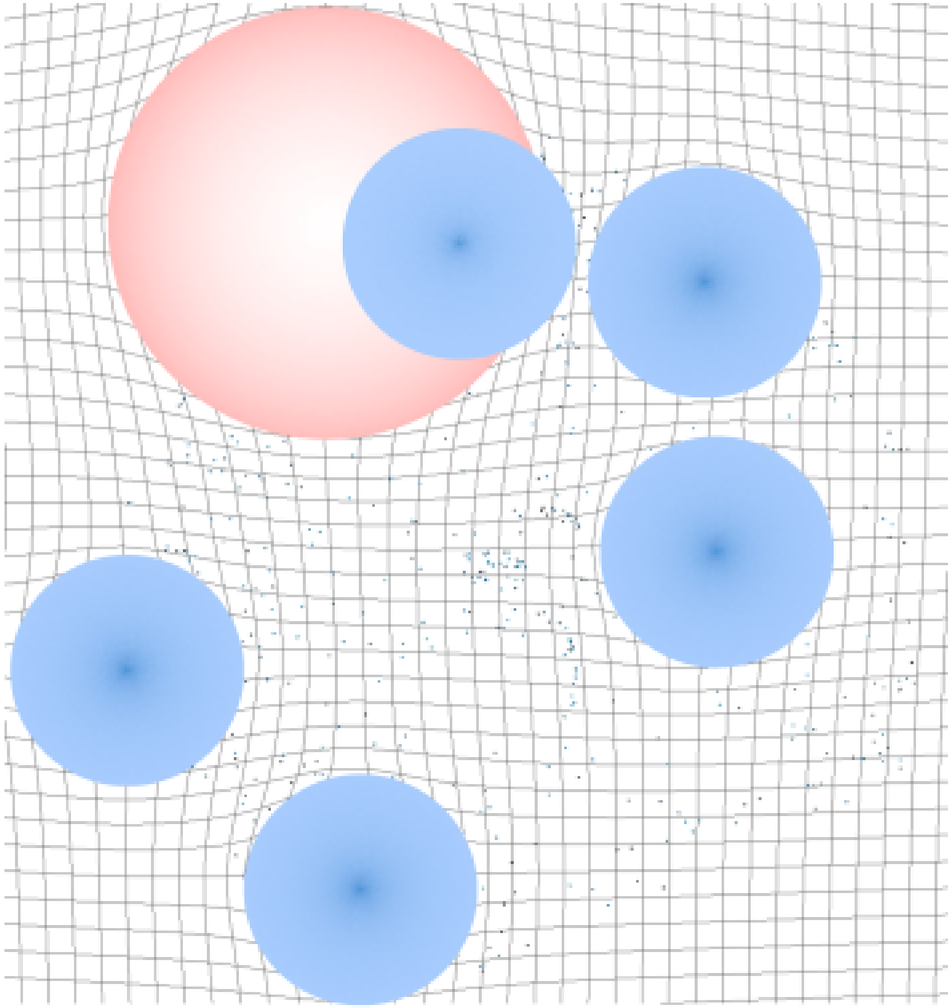
- How can we combine the ability of simple models to quickly generalize with superior adaptability of more complex ones?
  - regularization
  - model blending
- How can we support long-term (possibly life-long) adaptations?
  - change of preferences
  - decay of constraint importance
- How can we build better benchmarks?
  - collect more and better groud truth data
  - measure real user satisfaction

Example 2:

EVALUATING THE ADAPTIVE SPRINGLENS

Focus-Adaptive SpringLens*

- multi-focus fish-eye distortion highlights nearest neighbors



- primary lens
  - controlled by user
  - enlarges region of interest
  - more space for details
  - preserves context
- secondary lenses
  - data-driven
  - highlight nearest neighbors
  - show "wormholes"
  - neighbors come closer

*based on SpringLens non-linear distortion technique [Germer et al. '06]

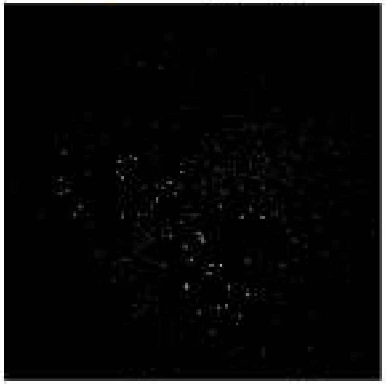
PhotoGalaxy (inverted color scheme)

Variants of User Input Controls

1. Panning & Zooming (P&Z)
   - left mouse (drag/pan), wheel (zoom)
   - cursor (pan), +/- (zoom)
2. Adaptive SpringLens (SL)
   - right mouse (click / hold&move), wheel (lens zoom)

common functions:

- change thumbnail size
- apply display filter:


collapse all


focus


sparse


expand all

Research Questions

1. How does the lens-based user-interface compare in terms of usability to common panning & zooming techniques that are very popular in interfaces using a map metaphor (such as Google Maps)?

2. How much do users actually use the secondary focus or would a common fish-eye distortion (i.e. only the primary focus) be sufficient?

3. What interaction patterns do emerge?

4. What can be improved to further support the user and increase user satisfaction?

Experiment Outline

- pre-experiment questionnaire
  - general background of participants
- training under supervision until familiar with user-interface
- solving a retrieval task with different input controls:

	group A:	group B:
1.	only P&Z	only SL
2.	only SL	only P&Z
3.	combination	

recorded:
- screen & control actions
- audio (think aloud protocol)
- webcam video
- gaze (Tobii T60 eye tracker)

- post-experiment questionnaire
  - usability judgments
  - feedback for improvements

Retrieval Task (Tagging Game)

- given
  - an image collection
  - with 5 topics, described by
    - a short text and
    - 2-3 representative images
  - find at least 5 images belonging to each topic

- notes:
  - topics are non-overlapping
  - relevance judgments fully up to the user’s point of view
  - handouts for guidance
  - no time limit
  - 5 minutes of interaction sufficient

Image Collections

- 4 image collections from a personal collection*
  - fixed order of presentation
  - collection #1 for training (250 images)
  - collections #2-4 labeled for evaluations (each 350 images)
- image resized to fit 600x600
- ground truth labels for collections #2-4
  - 5 non-overlapping topics each
- all images unknown to the participants (no bias)

* dataset can be provided under Creative Commons Attribution-Noncommercial-Share Alike License

Collection 2: Barcelona (350 images)

1. Tibidabo

2. Sagrada Família

3. Stone Hallway in Park Güell

4. Beach & Sea

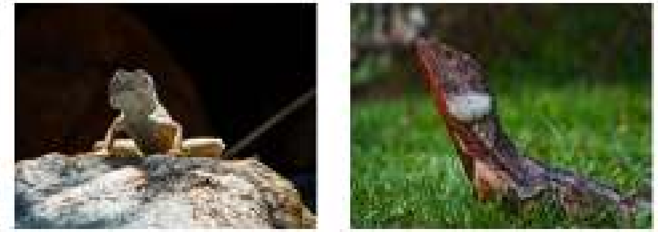
5. Casa Milà

Collection 3: Japan (350 images)

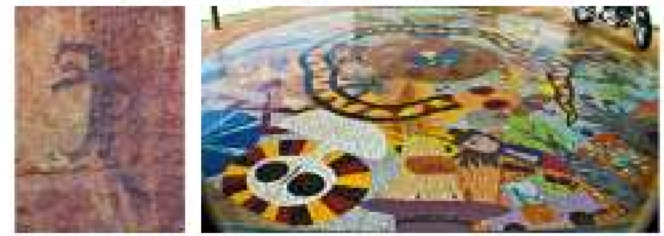
1. Owls
2. Torii
3. Paintings
4. Osaka Aquarium
5. Traditional Clothing

Collection 4: Western Australia (350 images)

1. Lizards

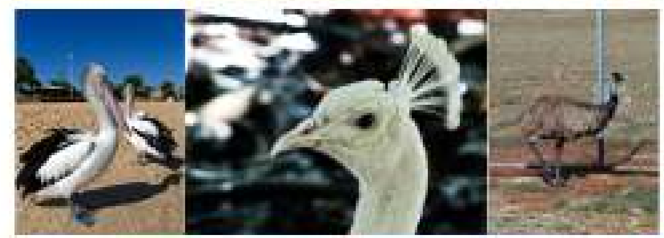


2. Aboriginal Art



3. Plants (Macro)

4. Birds



5. Ningaloo Reef

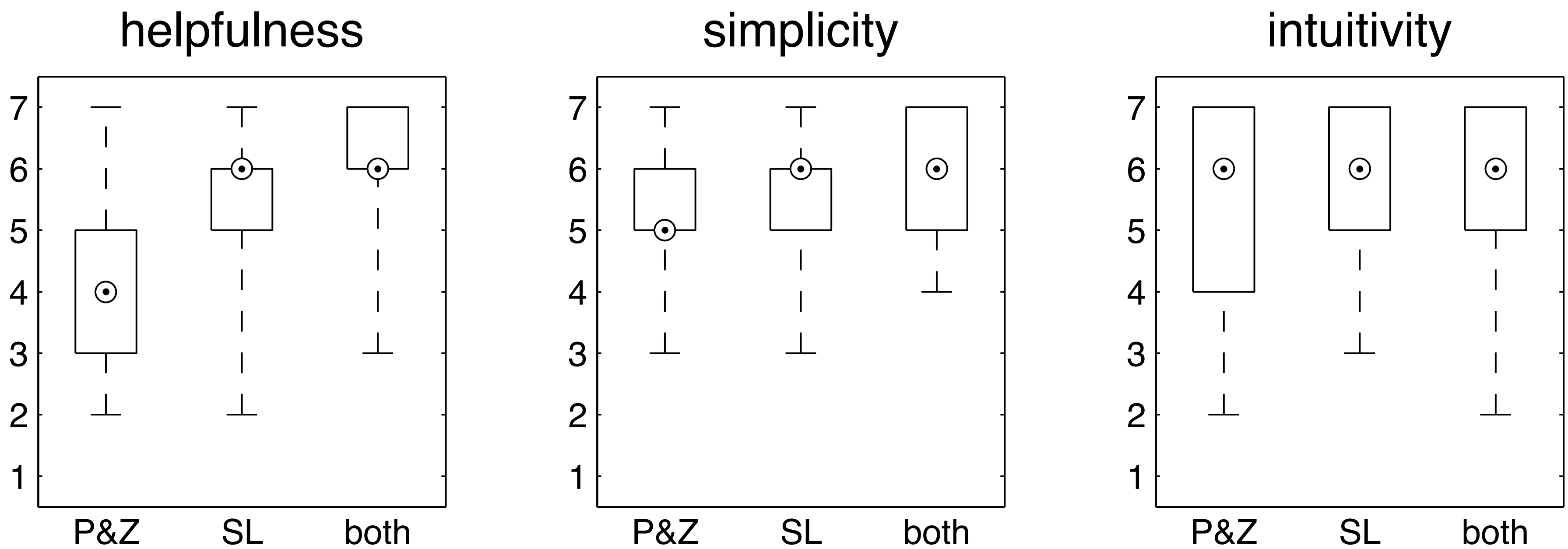
Participants

- 30 graduate and post-graduate students
- between 19 and 32 years old (mean = 25.5)
- 40% female

- 70% studied computer science
- 35% had background in computer vision or UI design
- 43% took photos on a regular basis
- 30% use software to manage their photo collection
- 77% were open to new user-interface concepts

- between 30 and 60 minutes per session

Results: Usability Comparison



- note: combined interface out of competition!

- 50% rated SL as significantly more helpful than P&Z while equally complicated in use
- intuitivity of SL slightly higher than P&Z (unexpected!)
- simplicity of BOTH highest (learning effect?)

Results: Usefulness of Secondary Focus

- analysis of recorded information for collection #4 (BOTH)
- 914 image-label events
- classification of events by:
  1. location of image when last spotted before labeling
  2. topic w.r.t. to topic of image in primary focus

focus region	primary	ext. primary	**secondary**	none
same topic	37.75	4.27	**30.74**	4.38
other topic		4.49	13.24	2.08
no focus				3.06
total	37.75	8.75	43.98	9.52

(some combinations are impossible)

Results: Search Strategies (collection #4)

- type 1: excessive P&Z
  - larger thumbnail size, deeper zoom level, a lot of panning
  - gaze: sequential / zigzag scans
- type 2: “eagle eye”
  - spot relevant images at high zoom level (dominant color?)
  - w/o focus
- type 3: continuous PF = quick scan with lens
  - no or little zoom, small thumbnails
  - main attention on (extended) PF (eyes guide lens)
  - moderate attention on SF
  - occasional “freezes” to scan whole region
- type 4: “jumping” focus (one SF becomes PF)
  - like navigating an invisible neighborhood graph
  - main attention on SF

Results: User Feedback

- overcrowded PF in dense regions
  - workaround: temporarily zoom into the region which lets the images drift further apart
  - possible solution: force-based spreading on hover
- SF mostly useless at deep zoom levels (off-screen)
  - off-screen visualization, navigation shortcuts
- avoid increasing “empty space” at deep zoom levels
  - automatically increase thumbnail size
- optional (temporary) re-arrangement into grid layout

→ better integrate P&Z and SL

Results: User Feedback (2)

feature requests:

- visualize already explored regions ("fog of war")
- undo / reverse "playback"
- advanced filters
  - e.g. by dominant color
- generate SF for a <u>set of images</u>
  - goal: query with already labeled images to find more relevant ones (bootstrapping classifier)

⟹ tested in simulation experiment
published at the 8th Int. Workshop on Adaptive Multimedia Retrieval (AMR'10), Linz, Austria, Aug. 2010.

Example 3:

DYNAMIC VISUALIZATIONS FOR EVOLVING MUSIC COLLECTIONS

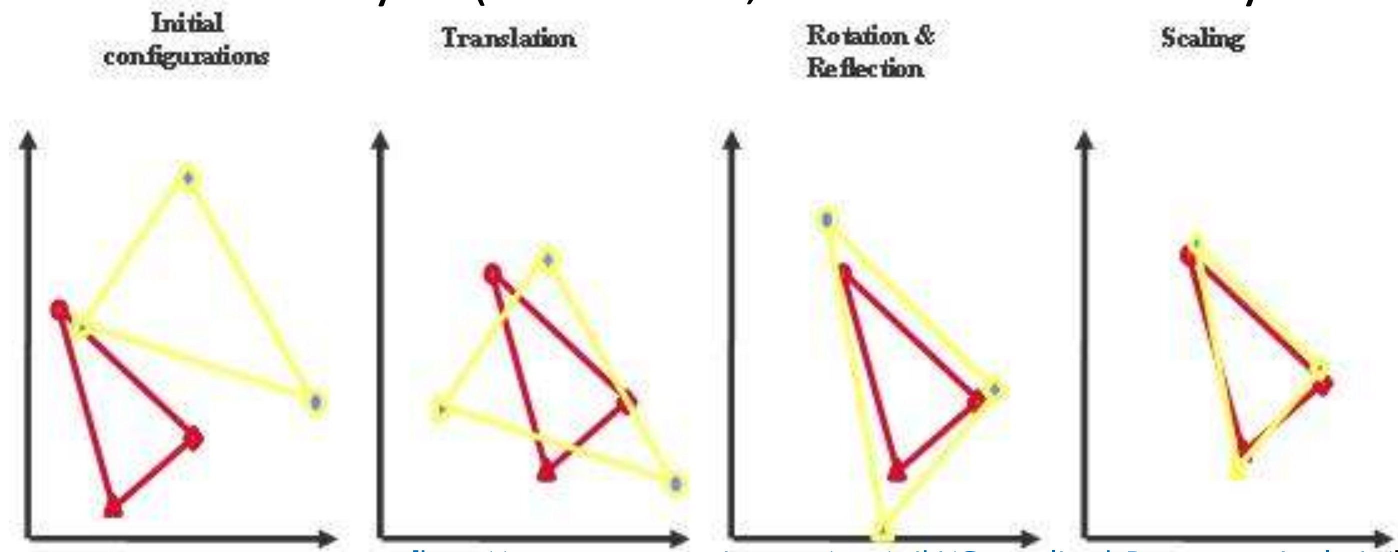
Challenge

- so far:
  - static music collection (dataset)
- in reality:
  - collections change (mostly grow)
  - maps may quickly become outdated
- problem:
  - re-computing a map from scratch may confuse the user
  - try to modify the existing map a little as possible to accommodate changes

The Candidates

- Multidimensional Scaling (MDS)
  - compute a new map and try to align it with the previous one

    -> Procrustes Analysis (translation, rotation & uniformly scaling)



[http://www.sensorysociety.org/ssp/wiki/Generalized_Procrustes_Analysis/]

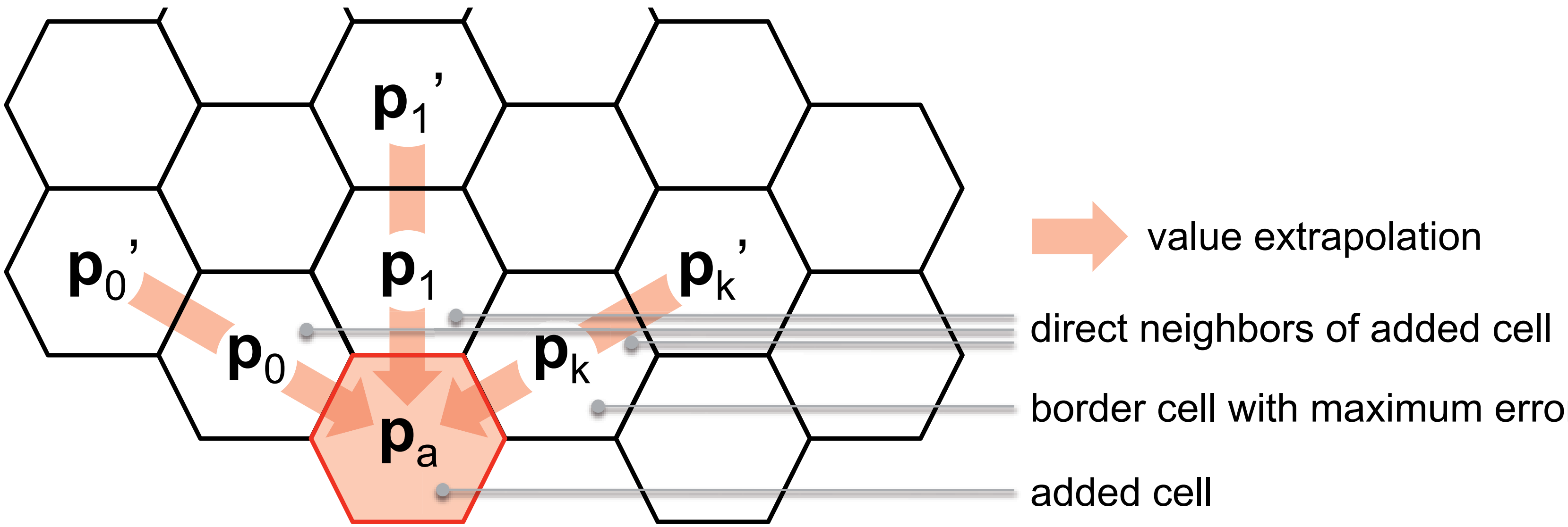
- Landmark Multidimensional Scaling (LMDS)
  - use only a sample of all points (“landmarks”) to compute the map

    = initial songs
  - place all other points w.r.t. their distances to the landmarks

    = new songs

The Candidates (2)

- Growing Self-Organizing Maps (GSOM)
  - SOM structure adapts to accommodate new data
  - new cells may be added as needed at the boundary



  - problem: requires vector space representation of data (e.g., through MDS vectorization)

The Candidates (3)

- Stochastic Neighbor Embedding (SNE)
  - goal: preserve the <u>probabilities</u> of points being neighbors
  - use Kullback-Leibler divergence as cost function
    (compares probability distributions)

$$D_{KL}(p_i, q_i) = \sum_{j \neq i} p_{j|i} \log \frac{p_{j|\text{[redacted]}}}{q_{j|\text{[redacted]}}}$$

input space probabilities

output space probabilities

How to support change?

- use current map as initial solution
  (with random positions for new songs)

The Candidates (4)

- Neighbor Retrieval Visualizer (NeRV)
  - goal: consider both, visualization precision and recall
  - use Kullback-Leibler divergence both ways for cost function:

cost of missing a neighbor

$$E = \lambda \sum_i D_{KL}(p_i, q_i) + (1 - \lambda) \sum_i D_{KL}(q_i, p_i)$$

cost of retrieving dissimilar objects

  - with λ ∈ [0, 1] as trade-off control
  - reduces to SNE for λ = 1

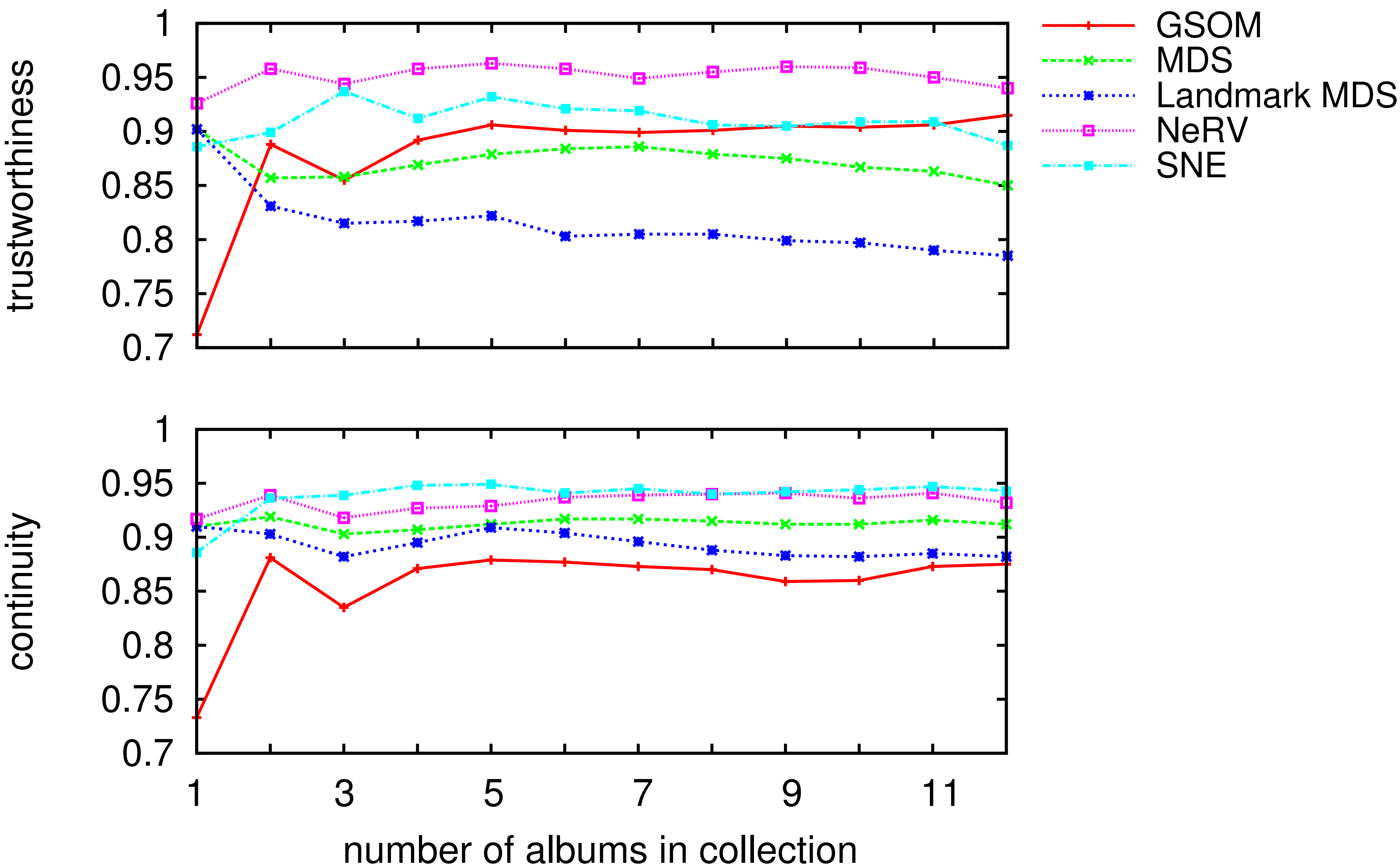
How to support change?

- use current map as initial solution
  (with random positions for new songs)

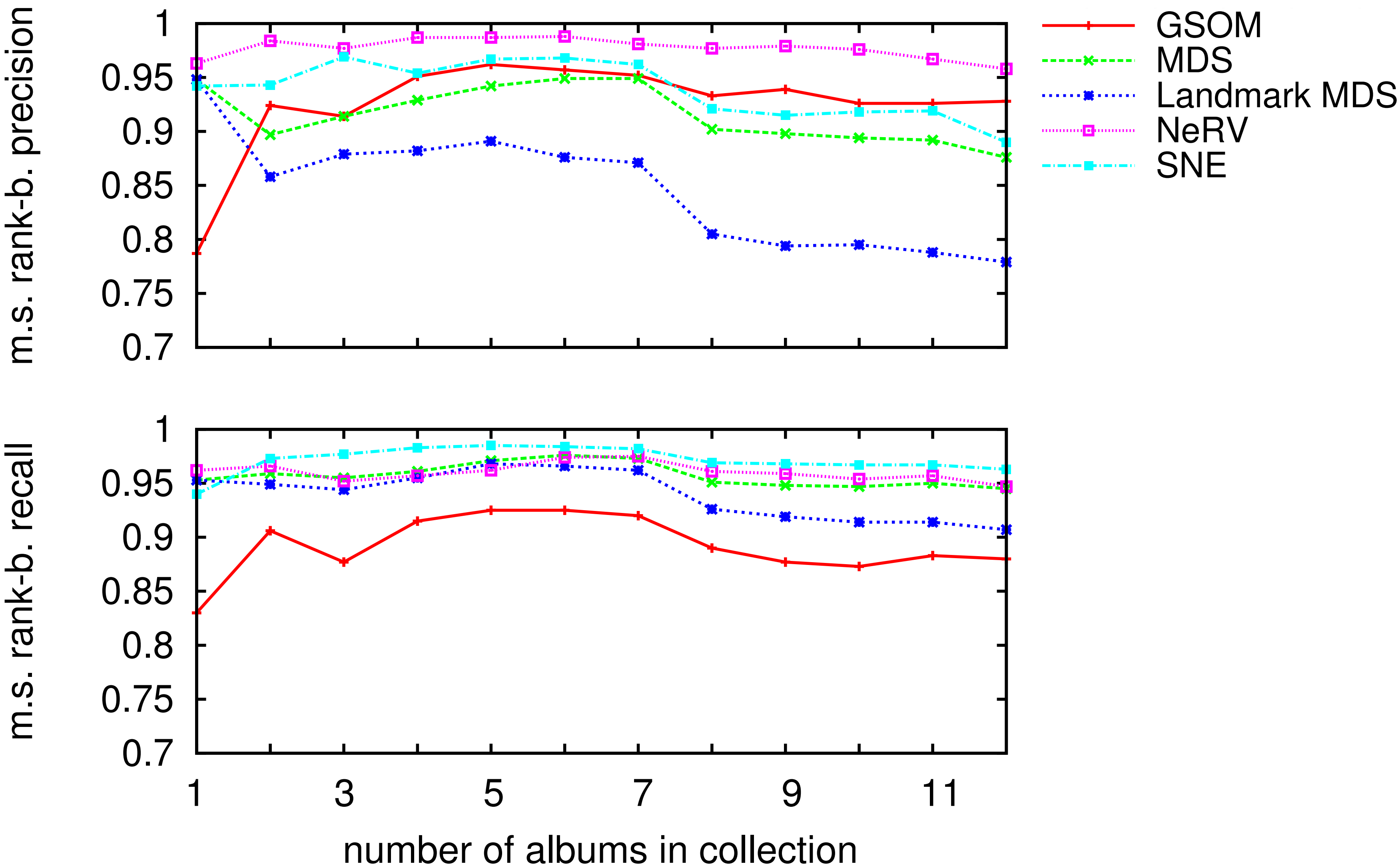
Two-fold Evaluation

- compare performance measures:
  - continuity
  - trustworthiness
  - (mean smoothed) precision & recall
  - mean position change
- ask users
  - … to play a memory game
  - … to rate the different visualizations
- benchmark dataset:
  - 12 official albums of The Beatles, added in order of relase

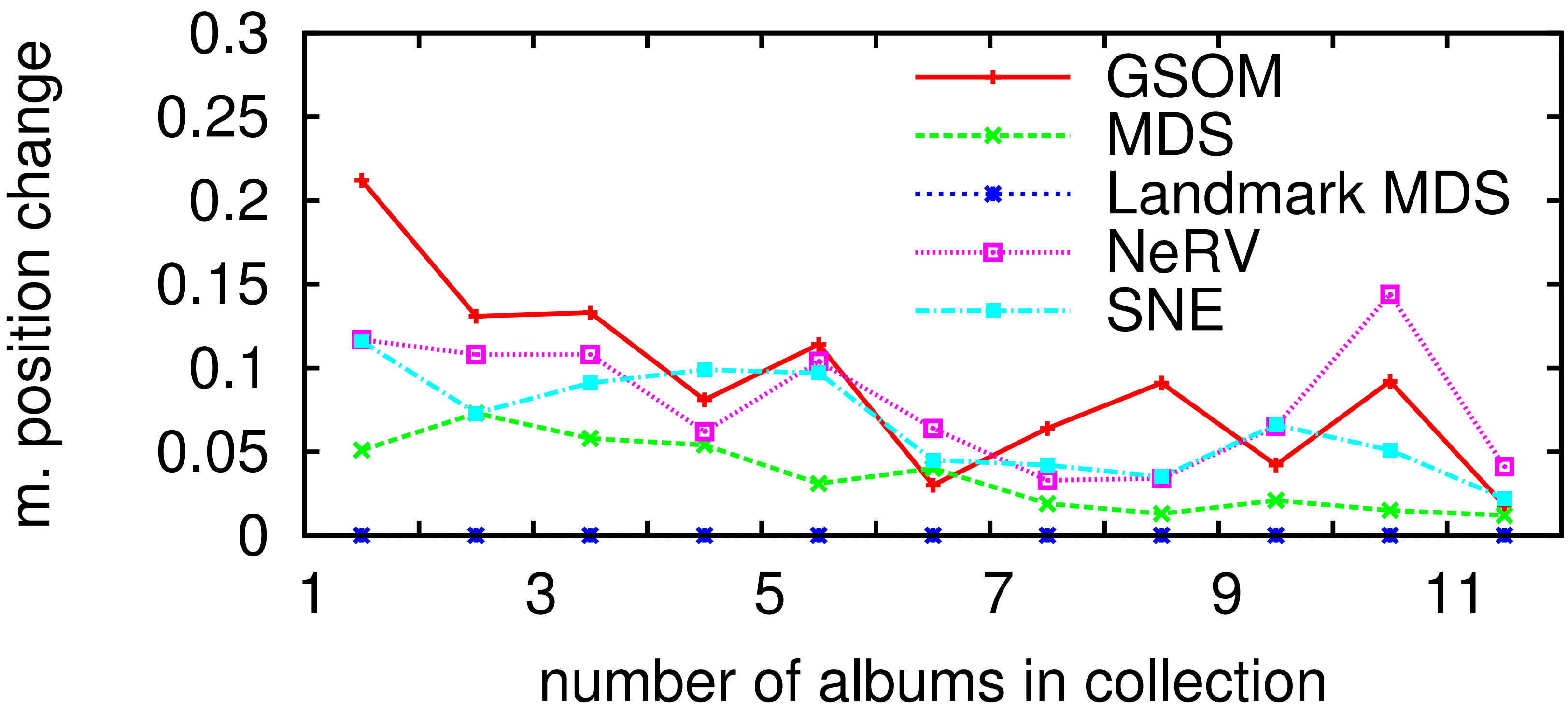
Performance Measure Comparison



Performance Measure Comparison (2)



Performance Measure Comparison (3)



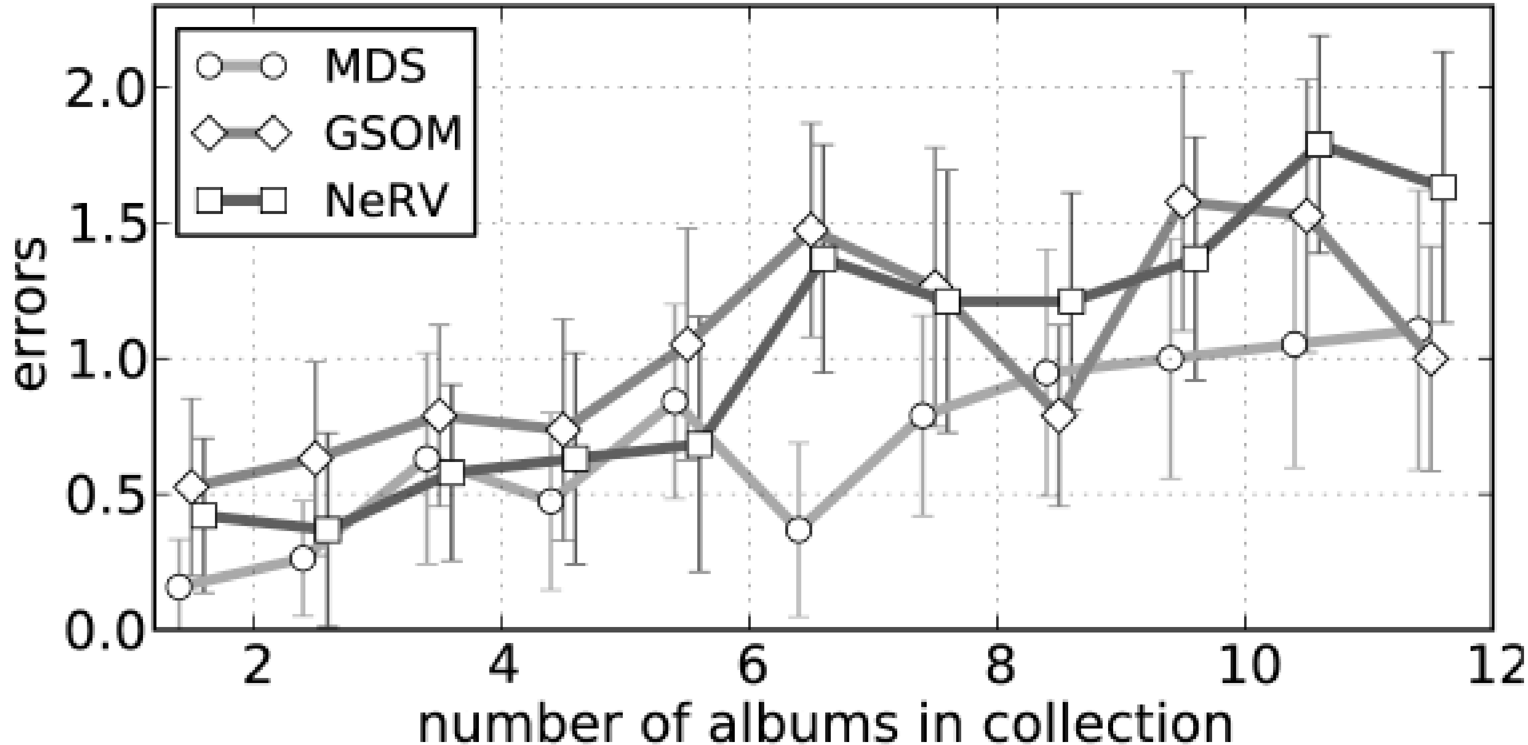
Memory Game

- n=19 participants
- 12 albums (11 steps)

try the demo at:
http://demos.dke-research.de/beatles-history-explorer/

Memory Game – Results

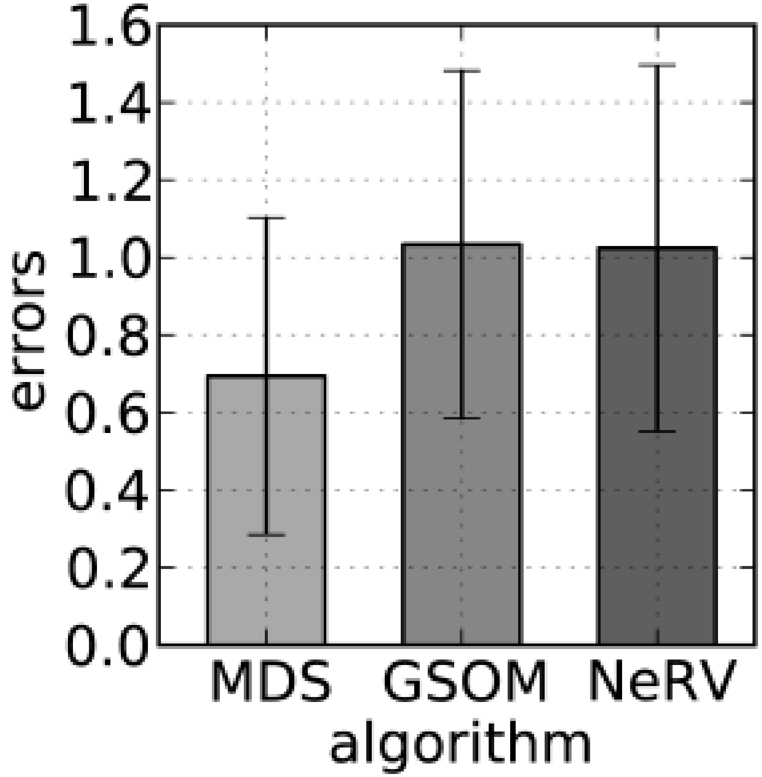
- errors per round:



- task gets harder
- MDS visualization appears to be easiest to follow

Memory Game – Results

- errors accumulated:



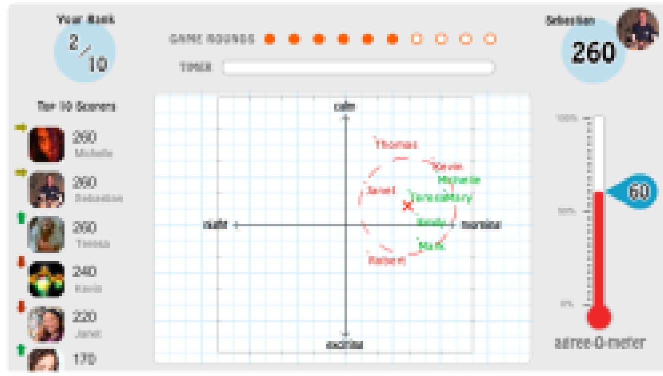
- mean memorization errors over all transitions and confidence intervals ($\alpha = 0.05$)
- MDS visualization appears to be easiest to follow

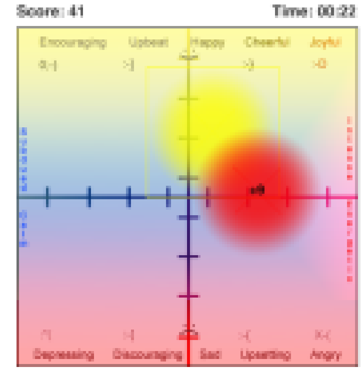
Future Directions

- test with other datasets
- test more algorithms
- modify NeRV to better support incremental collection changes
  - add another term to the cost function

Wrap-up: Use of Gamification in MIR

a) to collect ground truth

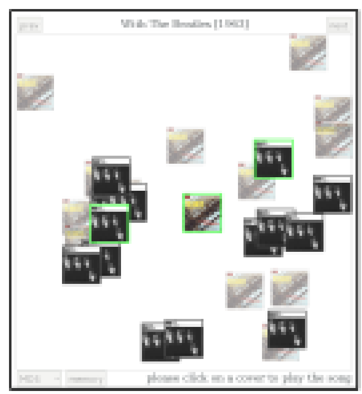




issues: may require further processing!

Do not blindly trust your data!

b) to give test users a concrete task



issues: game task may differ from real-world scenario

Part 4:

FROM USER-ADAPTIVE ORGANIZATION OF MUSIC COLLECTIONS TO BISOCIATIVE MUSIC DISCOVERY

Challenge

How can we make music recommendations more interesting?

➡ increase serendipity

Approach

leverage the effect of bisociations

➠ create an environment where serendipitous recommendations become more likely

Bisociations

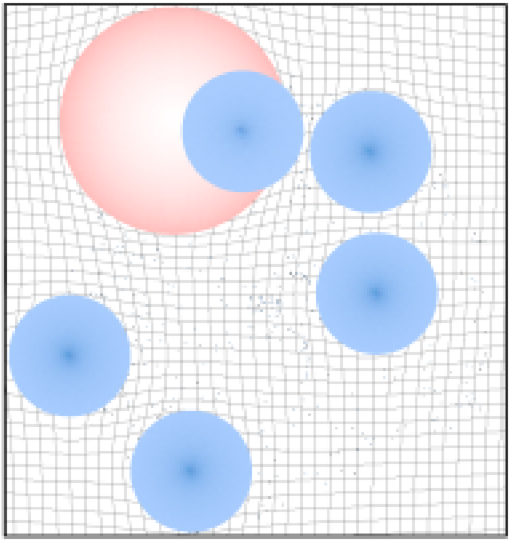
- Arthur Köstler: *The Act of Creation* (1964)

"the perceiving of a situation or idea, L, in two self-consistent but habitually incompatible frames of reference, M_1 and M_2.
The event L, in which the two intersect, is made to vibrate simultaneously on two different wavelengths, as it were. While this unusual situation lasts, L is not merely linked to one associative context but bisociated with two."

➡ simultaneous mental association of an idea or object with two fields / frames of reference ordinarily not regarded as related

➡ combine two different views on a music collection

Combining Orthogonal Similarity Spaces



projection weights

dynamics	0.0
rhythm	1.0
timbre	0.0

distortion weights

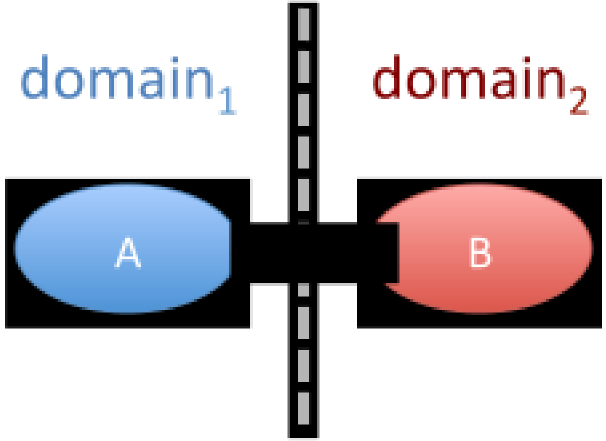
dynamics	1.0
rhythm	0.0
timbre	1.0

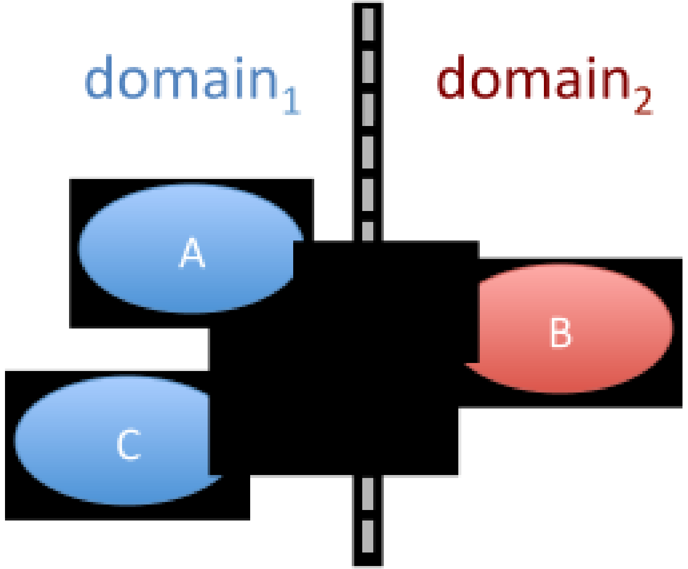
Bisociations in Graphs

- bridging concepts
  - established by ambiguous terms or metaphors
  - word-plays (context switching leads to a surprising outcome often perceived as joke)
- bridging graphs
  - connect concepts from different domains by inducing one or multiple paths between those concepts.
  - either the two concepts must lie in different domains or the path must contain at least one vertex in a different domain
- structural similarity
  - common structures in the context of each concept, i.e., similar subgraphs
  - may lead to same / very similar abstraction of both concepts

Bisociations by Bridging Graphs

= path that connects ideas or objects
  a) of different domains (ordinarily not regarded as related)
  b) by incorporating another domain

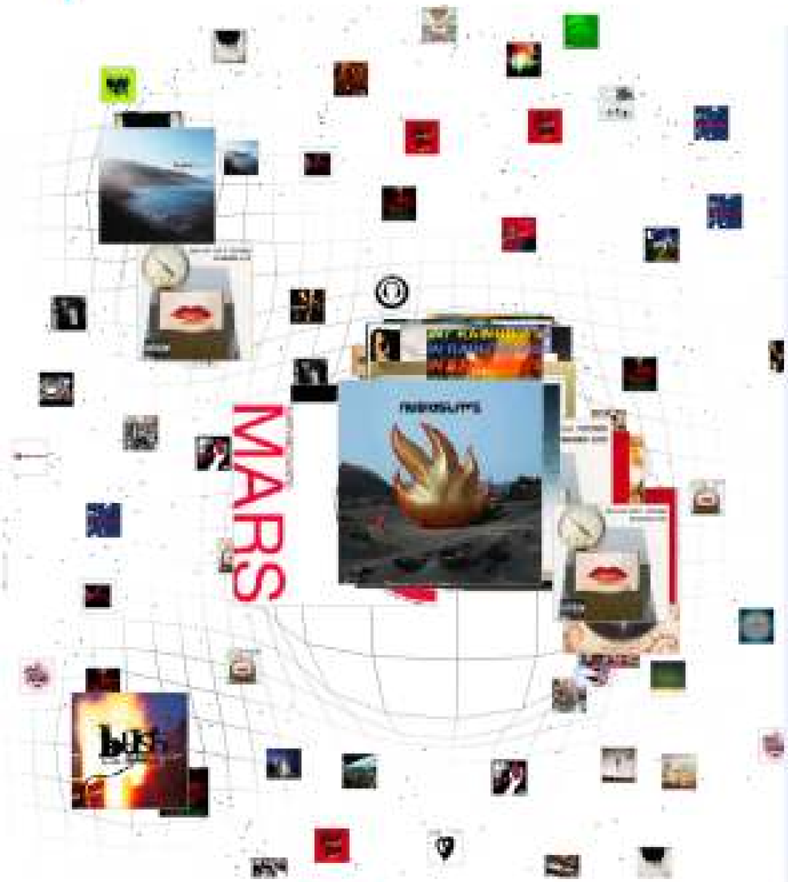


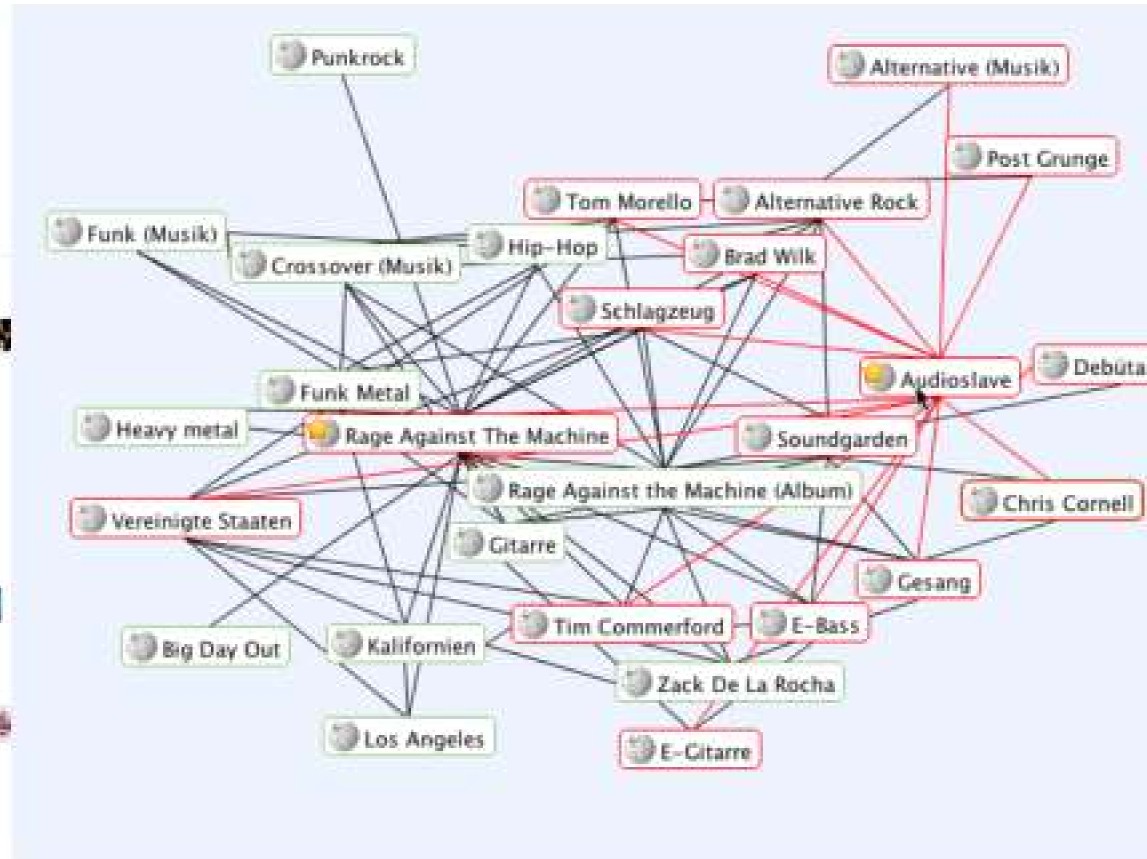


Similarity Space + Linked Data (Graph)

projection:
content-based similarity

nearest neighbors:
graph traversal





Linking Tracks and Metadata

audio recording

[redacted]

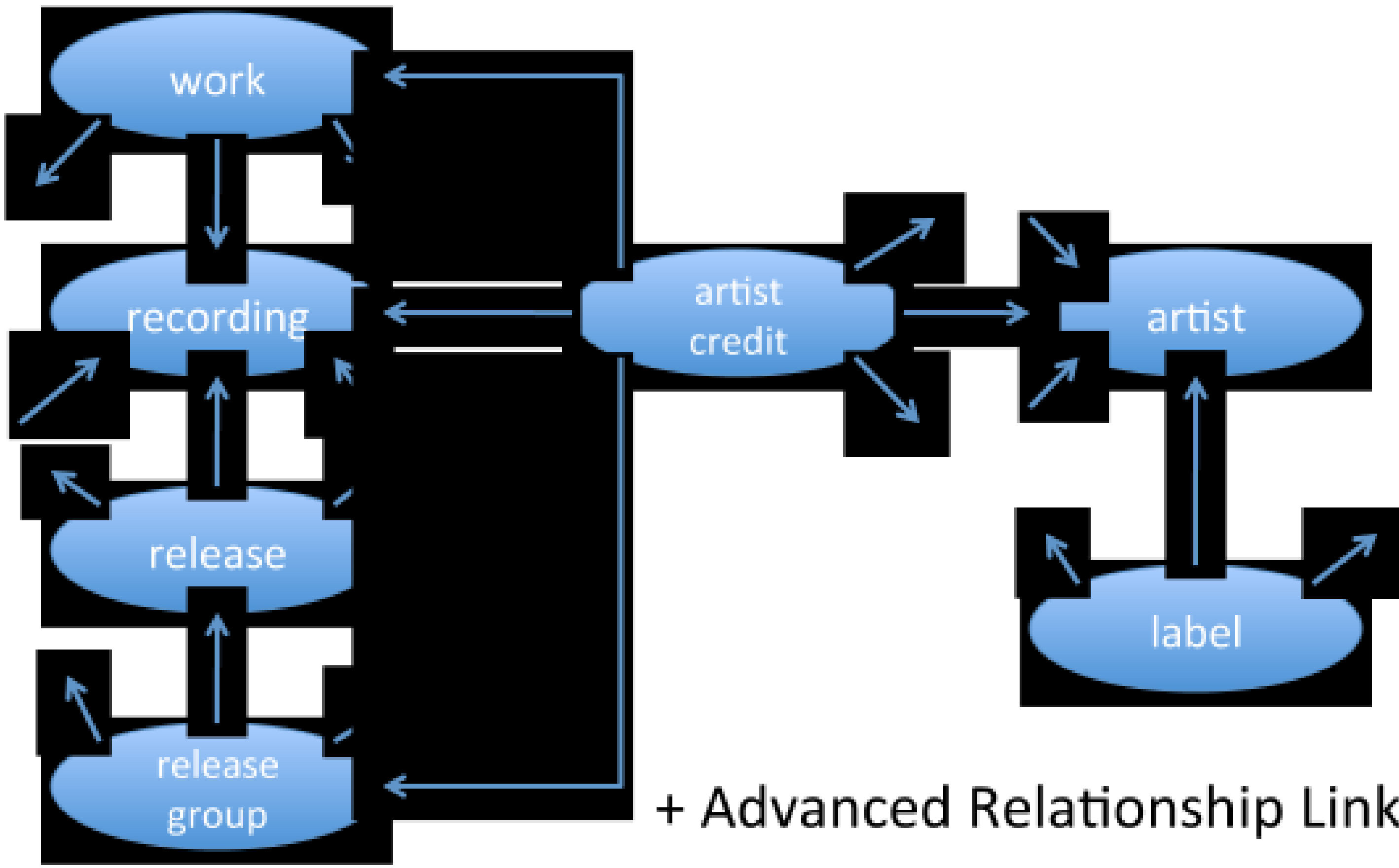
acoustica[redacted]erprint [redacted]mpliFIND Music Service

[redacted]

MusicBrainz

Neo4j
the graph database

The MusicBrainz Universe



+ Advanced Relationship Links

Examples:

- The song The Rockafeller Skank by Fatboy Slim includes a sample from the Just Brothers song Sliced Tomato.
- Paul Di'Anno was a member of Iron Maiden from 1977 until 1981.
- The Metallica album St. Anger was produced by Bob Rock & Metallica.

Relevance Measure for Traversal?

- should capture likelihood of serendipity
- possible simple heuristics:
  - prefer tracks that are projected far away from the primary focus (and thus most likely sound very different)
  - prefer tracks that the user has not listened to a lot or for a long time (and probably is no longer aware of)
  - prefer tracks of different artists and/or albums
- edge weights
  - inverse frequency weighting
    - similar to idf weights
    - favors rare ARLs
  - learn weights from feedback
- multiple paths ➟ aggregation method needed

The End

THANKS A LOT FOR LISTENING!

References for Further Reading

- Sebastian Stober; Thomas Low; Tatiana Gossen & Andreas Nürnberger. **Map-Based Exploration of Growing Music Collections**. In: *14th International Conference on Music Information Retrieval (ISMIR'13)*, 2013. (to appear)
- Daniel Wolff; Sebastian Stober; Andreas Nürnberger & Tillman Weyde. **A Systematic Comparison of Music Similarity Adaptation Approaches**. In: *13th International Conference on Music Information Retrieval (ISMIR'12)*, Pages 103-108, 2012.
- Sebastian Stober; Stefan Haun & Andreas Nürnberger. **Creating an Environment for Bisociative Music Discovery and Recommendation**. In: *Proceedings of Audio Mostly 2011 -- 6th Conference on Interaction with Sound -- Extended Abstracts*, Pages 1-6, Coimbra, Portugal, Sep 2011.
- Sebastian Stober; Christian Hentschel & Andreas Nürnberger. **Evaluation of Adaptive SpringLens - A Multi-focus Interface for Exploring Multimedia Collections**. In: *Proceedings of 6th Nordic Conference on Human-Computer Interaction (NordiCHI'10)*, Pages 785-788, Reykjavik, Iceland, Oct 2010.

all papers can be downloaded from http://wwwiti.cs.uni-magdeburg.de/~stober/publ/