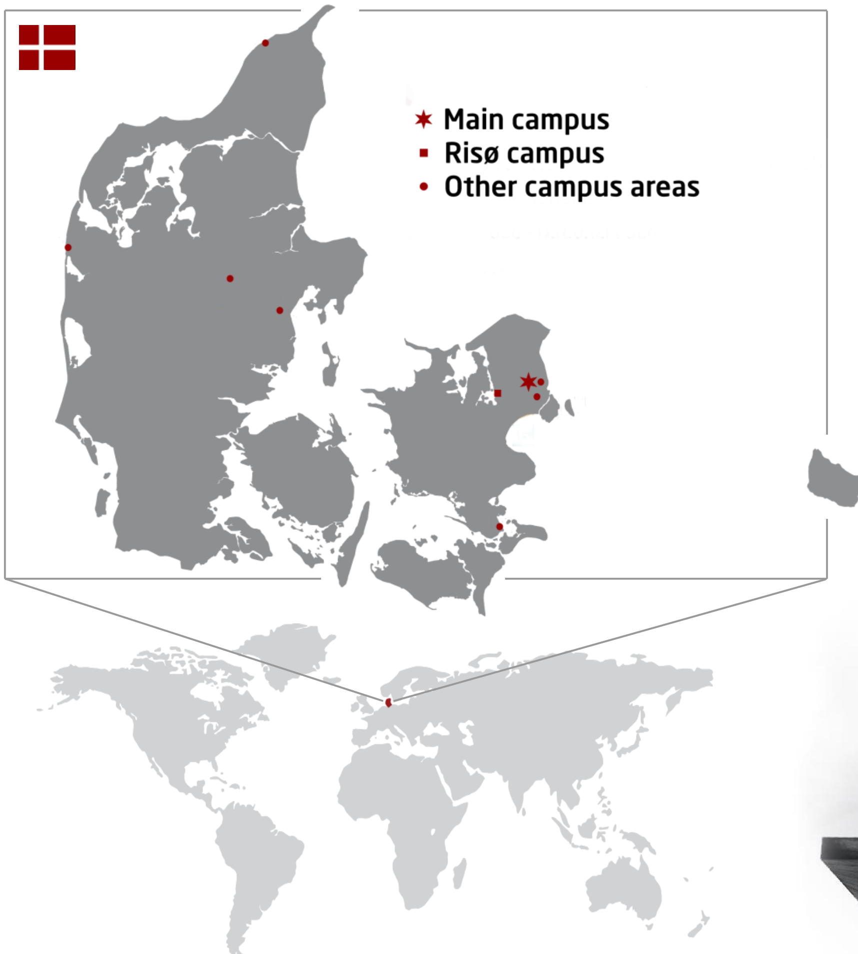
Creating meaning in audio and music signals

Jan Larsen, Associate Professor PhD
Cognitive Systems Section
Dept. of Applied Mathematics and Computer Science
Technical University of Denmark
janla@dtu.dk, www.compute.dtu.dk/~jl

DTU COMPUTE

Technical University of Denmark

(founded 1829; first rector H.C. Ørsted)



Ranking

Leiden *Crown Indicator* 2010

no. 1 in Scandinavia

no. 7 in Europe

DTU facts and figures

Education

7072 BSc, MSc og Beng students

incl. 626 international MSc students

1197 PhD students

626 exchange studens

296 DTU students at exhange programs

Innovation

87 registered IPR

46 submitted patent applications

Research

3648 research publications

241 PhD theses

Personel

31 DVIP

2657 VIP

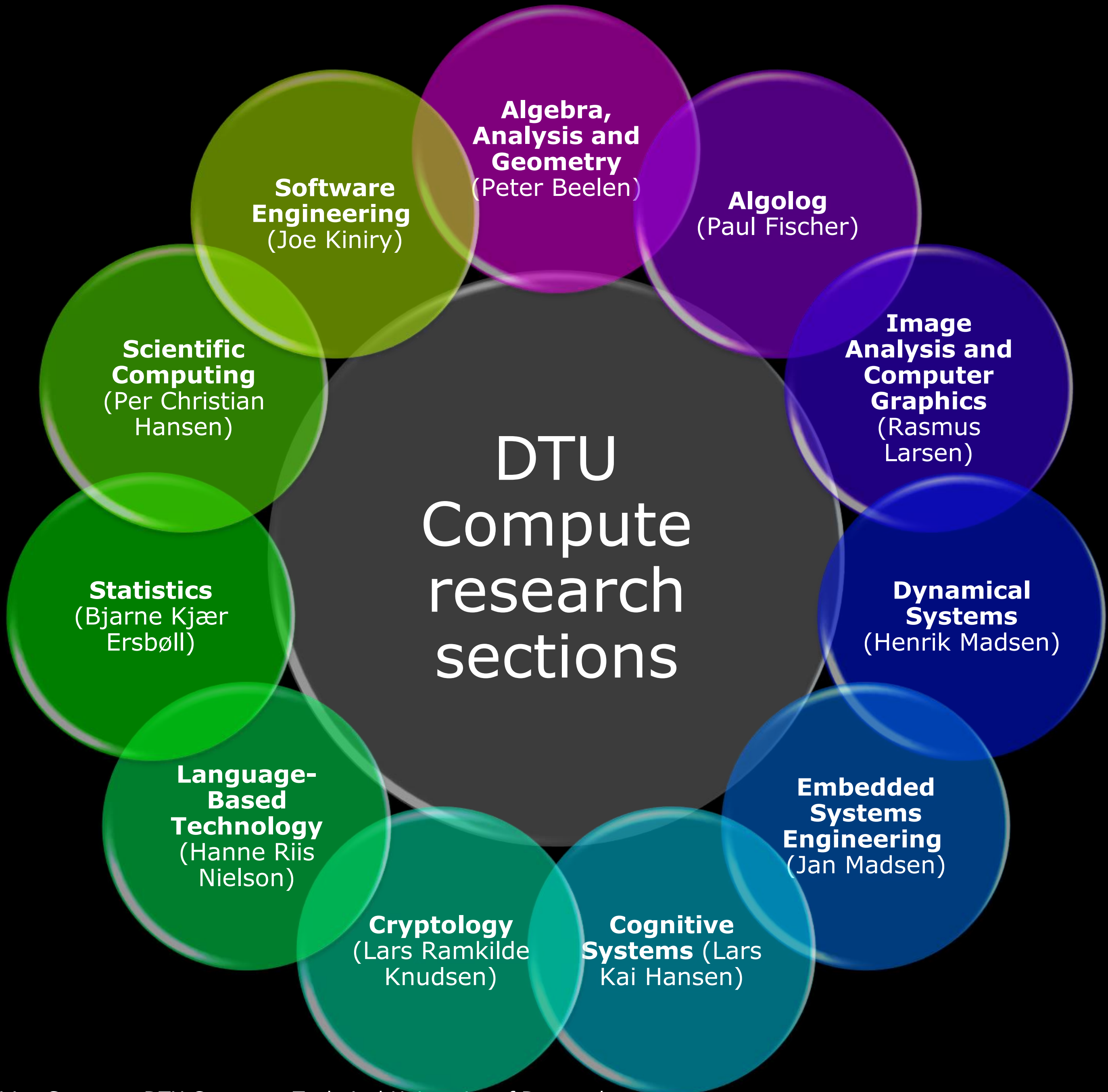
2221 TAP

1007 PhD students

Economy 5.8 BDKK

Public sector consultancy
Strategic contract with Danish ministries 338 MDKK

Buildings 454.420 m²


DTU Compute research sections
Algebra, Analysis and Geometry (Peter Beelen)
Algolog (Paul Fischer)
Image Analysis and Computer Graphics (Rasmus Larsen)
Dynamical Systems (Henrik Madsen)
Embedded Systems Engineering (Jan Madsen)
Cognitive Systems (Lars Kai Hansen)
Cryptology (Lars Ramkilde Knudsen)
Language-Based Technology (Hanne Riis Nielson)
Statistics (Bjarne Kjær Ersbøll)
Scientific Computing (Per Christian Hansen)
Software Engineering (Joe Kiniry)

Cognitive Systems Section

Why do we do it? VISION

What do we do? MISSION

- 2 professors
- 7 associate prof.
- 1 assistant prof.
- 1 senior researcher
- 5 postdocs
- 17 Ph.D. students
- 5 project coordinators
- 2 programmers
- 1 admin assistant
- 10 M.Sc. students

machine learning

media technology

cognitive science

Vision

Cognition refers to the representations and processes involved in thinking and decision making. Cognitive systems integrate information processing in brains and computers for collaborative problem solving.

Our vision is to design and implement profound cognitive systems for augmented human cognition in real-life environments

Our research is driven both by curiosity and by an engineering desire to do good: To better understand human behaviors and to create engineering solutions with a positive impact on human well-being and productivity.

We will contribute to DTU's vision of excellence and strive to be a highly valued partner for our national and international networks.

Legacy of cognitive systems

Allan Turing

Theory of computing 1940’es

Norbert Wiener

Cybernetics

1948

processing

adaption

under-standing

cognition

information and data

people

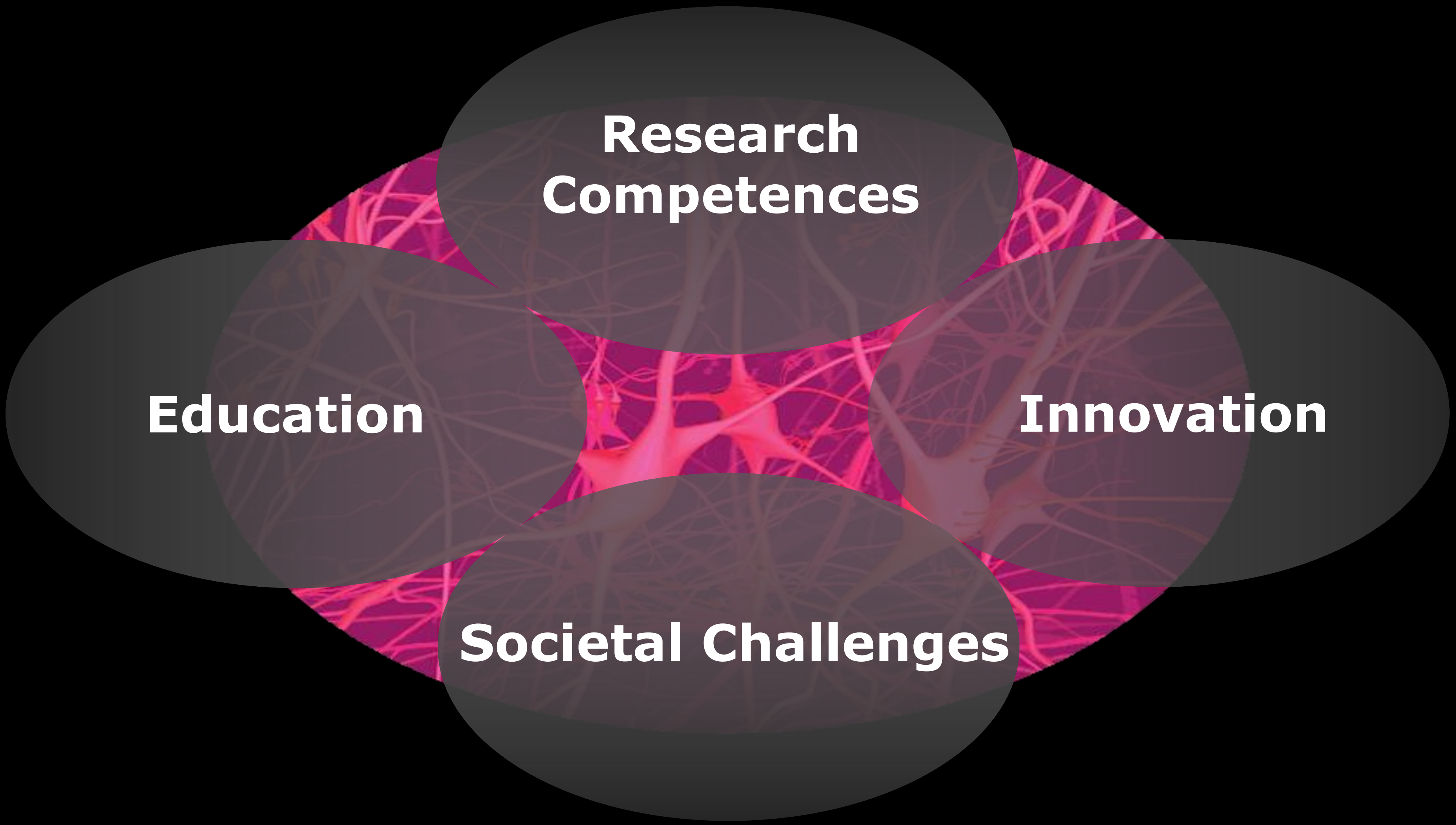
Mission

To measure, model, and augment cognition from neuron to internet scale systems

A cognitive system should optimize itself according to:

The statistical model of the domain, the psycho-physical model of the users, the social context, and the computational resources in time and space

Interplay and Synergy



Research

Machine Learning
Neuroinformatics
Human computer interaction
Cognitive Psychology

Education

Machine learning
Signal processing
Cognitive engineering
Digital media personalization, meta data, and web2.0
HCI and user experience modeling
Mobile technologies and modeling

Innovation

Danish Sound Technology Network
Professional Networks
Industrial PhD and Master Students
Commissioned Industrial Research

Future improvement in productivity and quality of life requires organization and integration of **Web-scale data sets**

Digital media modeling enables ubiquitous access to actionable information for personal development and organization of interpersonal relations

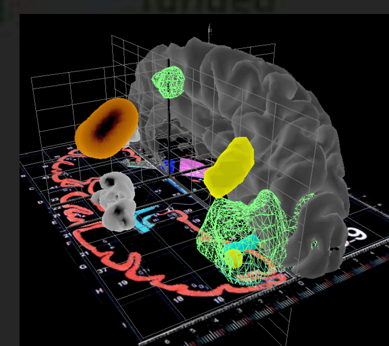
Brain modeling and mental decoding are crucial for augmented cognition, lifelong learning, and may revolutionize health services

Research Competences

Media technology: mobile platforms, digital media, social networks, search, navigation, and semantics

Machine learning: statistical modeling, signal processing, and complex networks

Cognitive science: perception, cognition, psycho-physics, and human computer interfacing



Bjørn Sand Jensen

Jens Brehm Nielsen

Jens Madsen

Rasmus Troelsgaard

Lars Kai Hansen

Mikkel N. Schmidt

Jerónimo Arenas-García

Ling Feng

Anders Meng

Seliz Karadogan

Letizia Marchegiani

Peter Ahrendt

Michael Kai Petersen

Michael Syskind Pedersen

CREATING MEANING IN AUDIO

Lasse Lohilahti Mølgaard

Tue Lehn-Schiøler

Kaare Brandt Petersen

Mission

Measure, model, extract, and augment meaningful and actionable information from audio and related information, social context, psycho-physical model of the users by ubiquitous learning from data and optimizing the computational resources

Specific research competences in audio

- Audio segmentation
- Genre, mood and metadata prediction
- Cognitive components
- Source separation
- Context based spoken document retrieval
- Preference elicitation

Specialized search and music organization

Search using mood

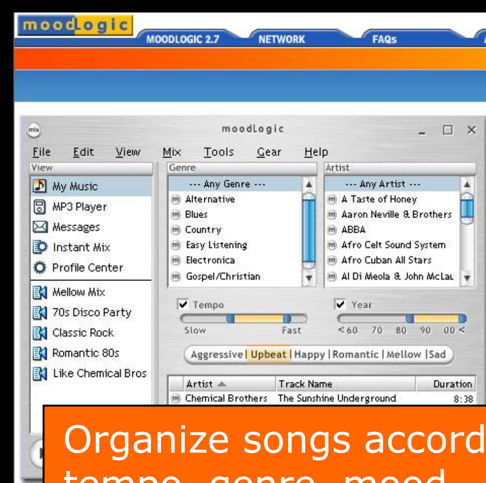
Using social network analysis

Listen and identify music

Explore by genre, mood, theme, country, instrument

Query by humming

The NGSW is creating an online fully-searchable digital library of spoken word collections spanning the 20th century



Organize songs according to tempo, genre, mood

search for related songs using the “400 genes of music”

Aspects of search and navigation

Specificity

- standard search engines
- indexing of deep content

Objective: high retrieval performance

Similarity

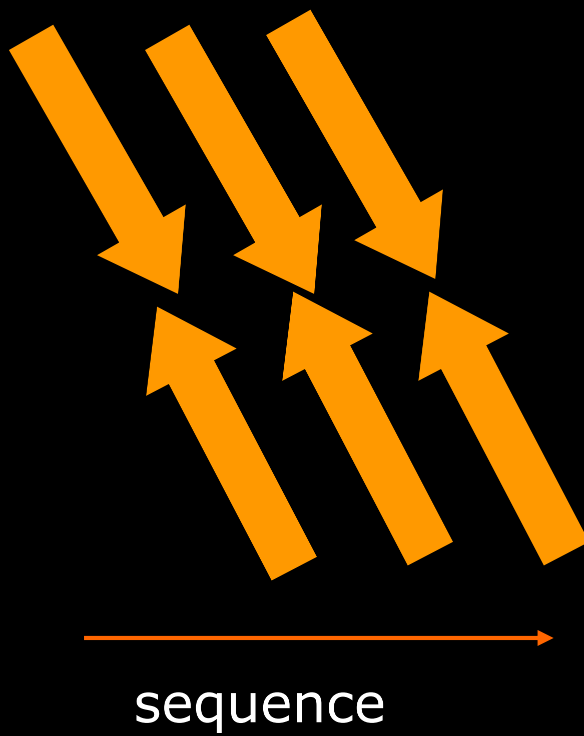
- more like this
- serendipity
- similarity metrics

Objective: high generalization and user acceptance

A cognitive architecture

Combine bottom-up and top-down processing

- Top-down user feedback
  - High specificity
  - Time scales: long, slowly adapting
- Bottom-up data modeling
  - High sensitivity
  - Time scales: short, fast adaptation



Courtesey of Lars Kai Hansen, DTU

DTU

DTU

DR

Syntonetic

Musikzonen

Geckon

COSOUND

Royal School of Library and Information Science

Hindenburg Systems

UCL

Queen Mary University of London

B&O

Danish Council for Strategic Research Project 2012-2015

Copenhagen University

Aalborg University

State and University Library

University of Glasgow

Vision

The overall vision is to foster truly participatory, collaborative, and cross-cultural tools for enrichment of audio streams which can improve interactivity, findability, experienced quality, ability to co-create, and boost productivity in a broad sense.

Mission

We have establish a multi-disciplinary strategic research activity to build a flexible modular audio data processing platform which enables new products and services for the

- commercial sector
- public service sector
- education and cultural research

Hypothesis

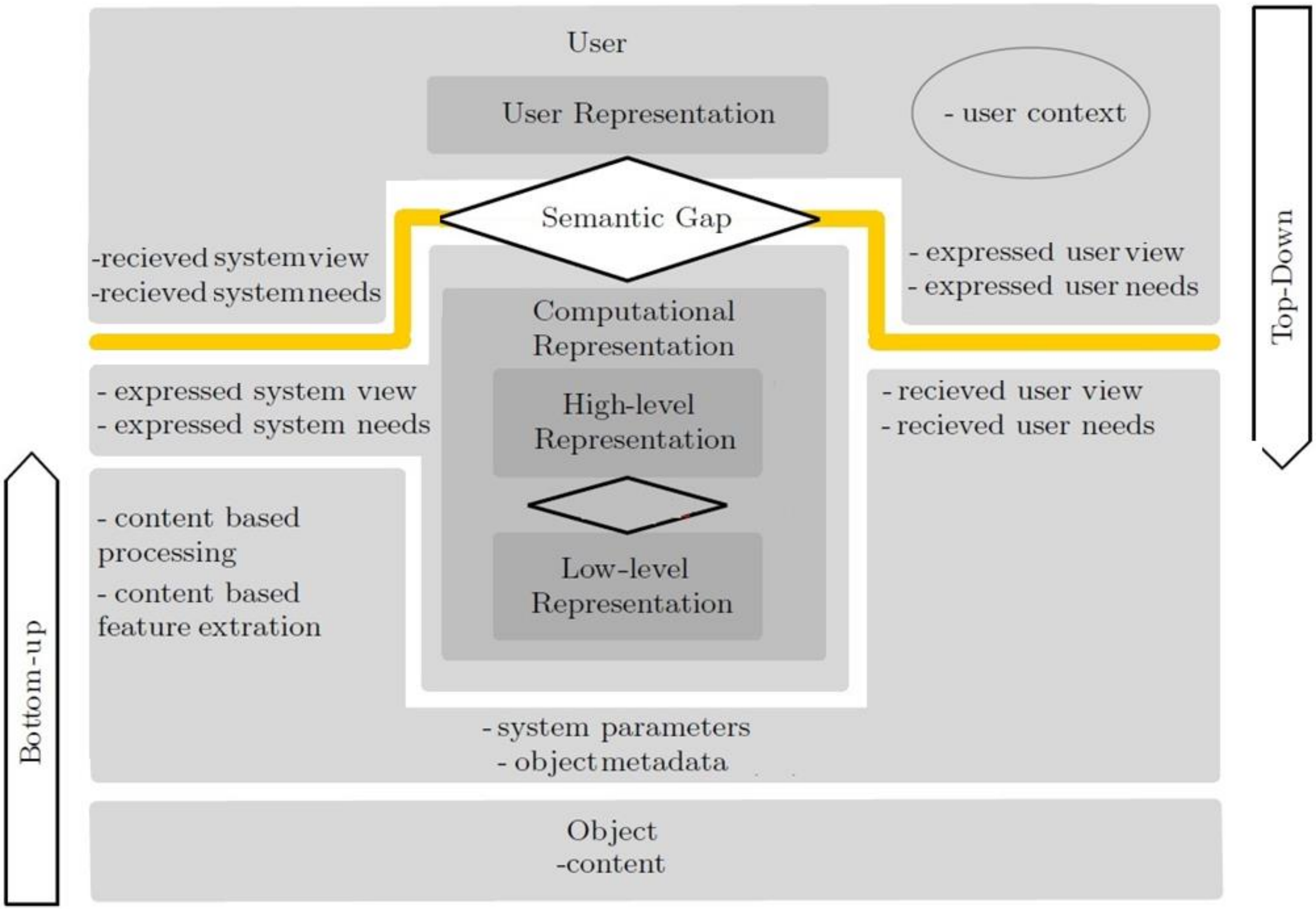
Top-down user streams

Learning cognitive representations and interaction

The main hypothesis is that the integration of bottom-up data derived from audio streams and top-down data streams from users can enable actionable cognitive representations, which will positively impact and enrich user interaction with massive audio archives, as well as facilitating new commercial success in the Danish sound technology sector.

Buttom up audio streams

Framework


User
User Representation
- user context
Semantic Gap
-recieved system view
-recieved system needs
- expressed user view
- expressed user needs
Computational Representation
- expressed system view
- expressed system needs
High-level Representation
- recieved user view
- recieved user needs
Top-Down
Bottom-up
- content based processing
- content based feature extration
Low-level Representation
- system parameters
- object metadata
Object
-content

Aspects of users

Content preference

State of mind

Objective/task

Context

Top-down view - *user driven*

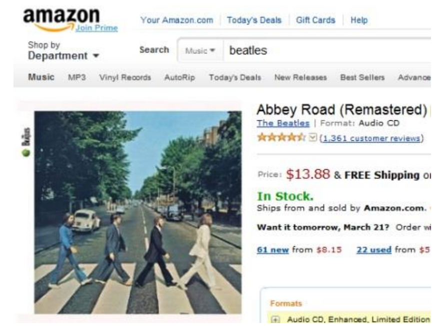
Preference

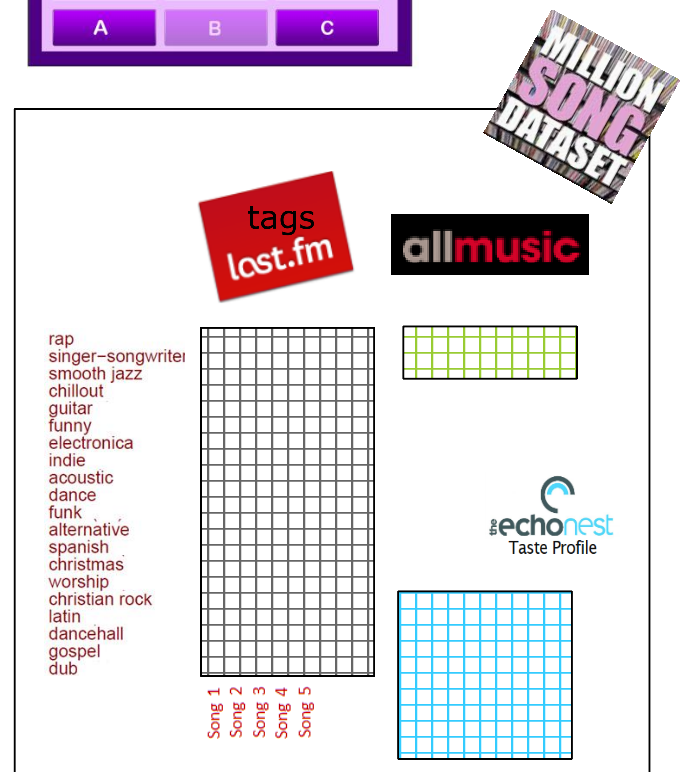
”I'll give *Abby Road* album 4/5 stars”

“I prefer *Yesterday* over *How do you sleep?*”

“I'll rate *Yesterday* as 0.7 on a 0-1 scale”

“I don't like jazz today”





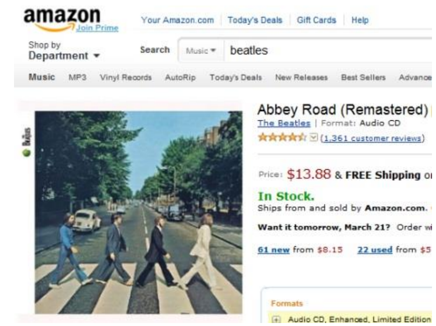
Top-down view - *user driven*

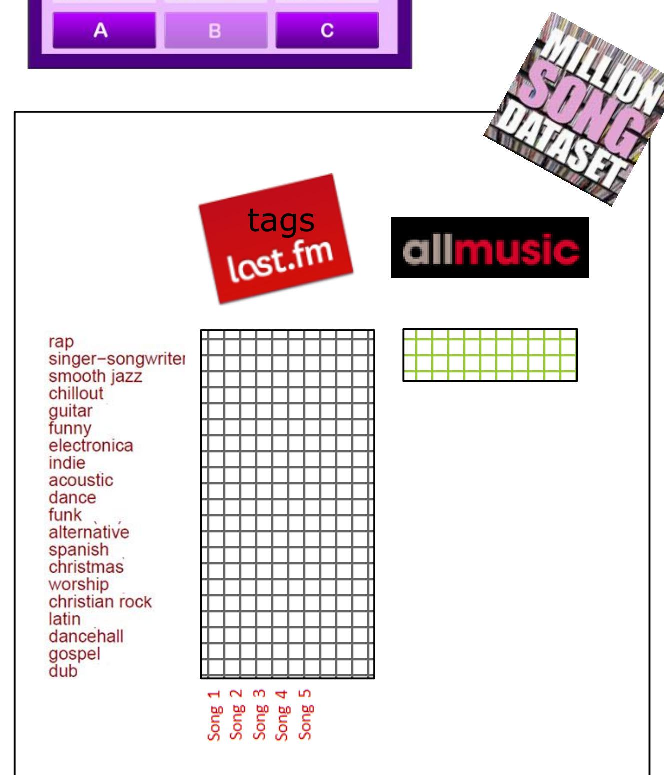
Listening patterns (indirect preference)

You listened to *Helter Skelter* 666 times…

so did a guy named Charles.

You listen to heavy metal in your car



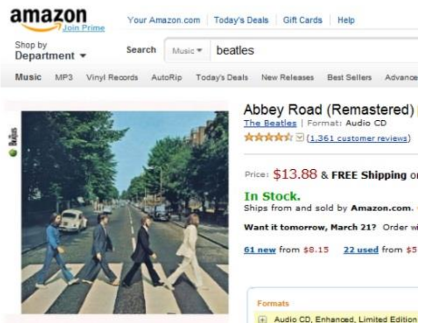


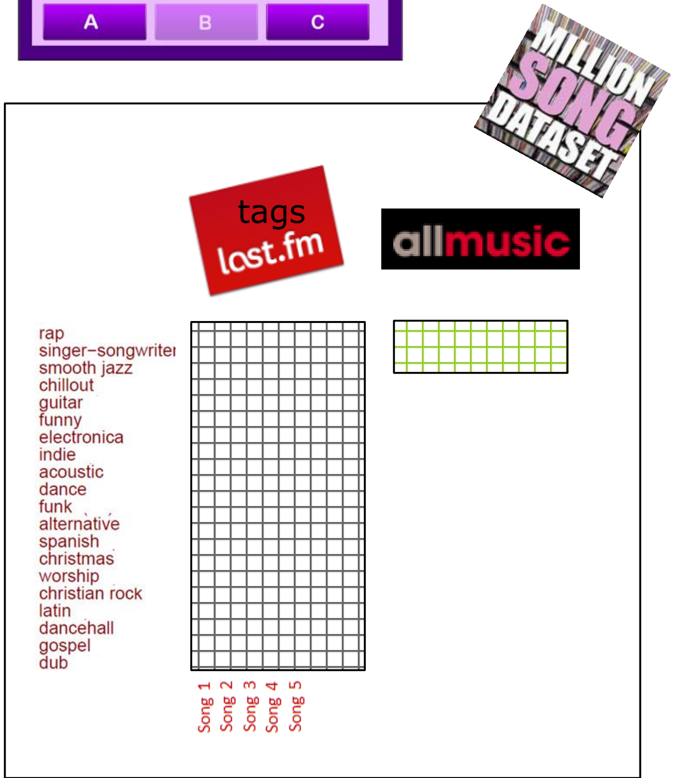
Top-down view - *user driven*

Music similarity/relations

"Out of the three: *Helter Skelter*, *Yesterday*, *When I'm Sixty-Four* - *Helter Skelter* is the odd-one out" (e.g. Magna-tag-a-tune)

Yesterday is from the same album as the band *Dizzy Miss Lizzy.*





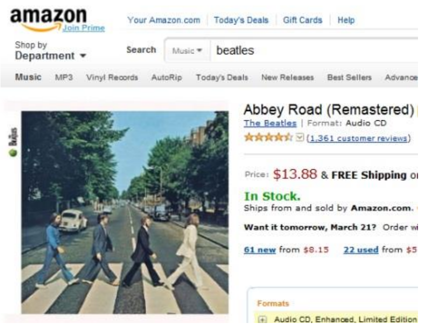
Top-down view - *user driven*

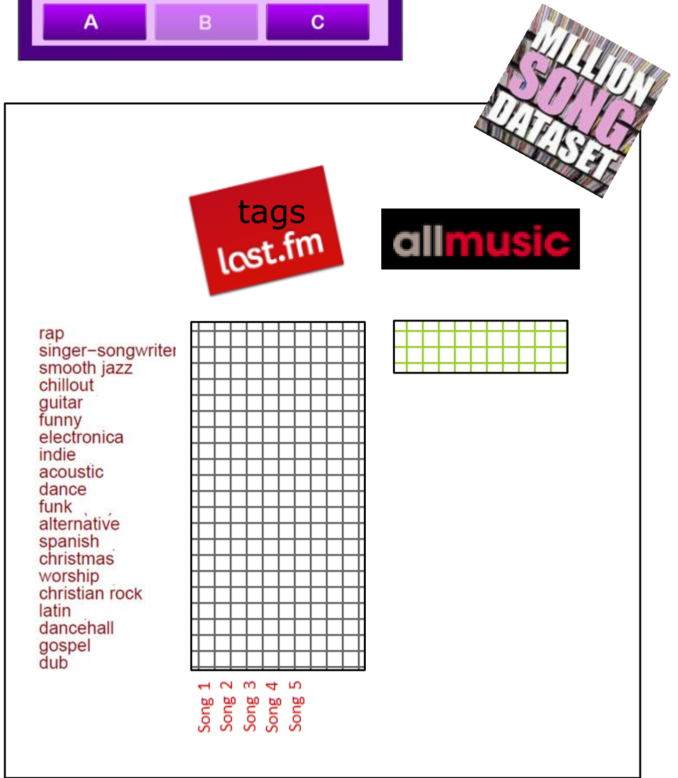
Music emotion/mood

"*When I'm Sixty-Four* is happier than *Helter Skelter*"

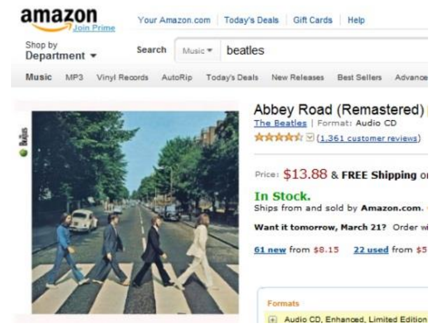
How happy is *When I'm Sixty-Four* – from 1-5?

(1 being sad, 5 being happy).





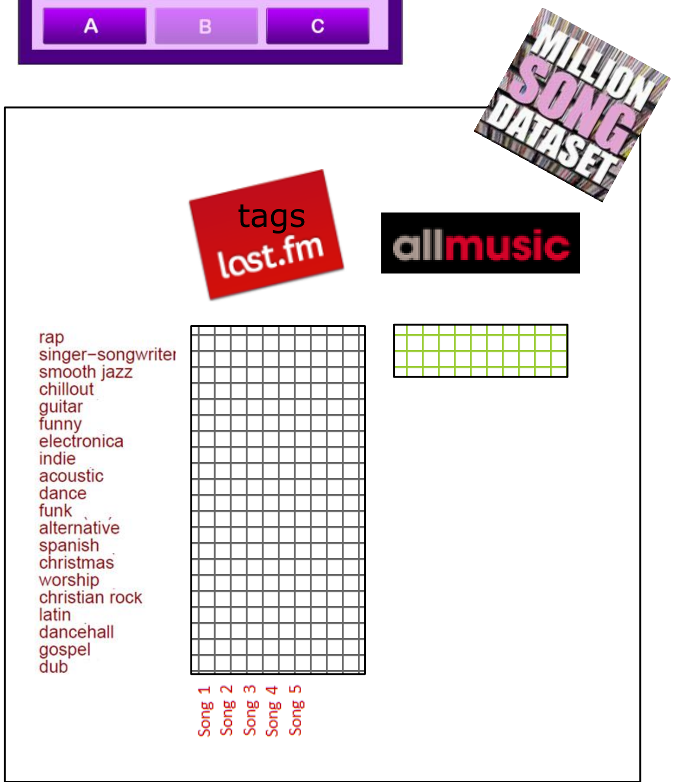
Top-down view - *user driven*



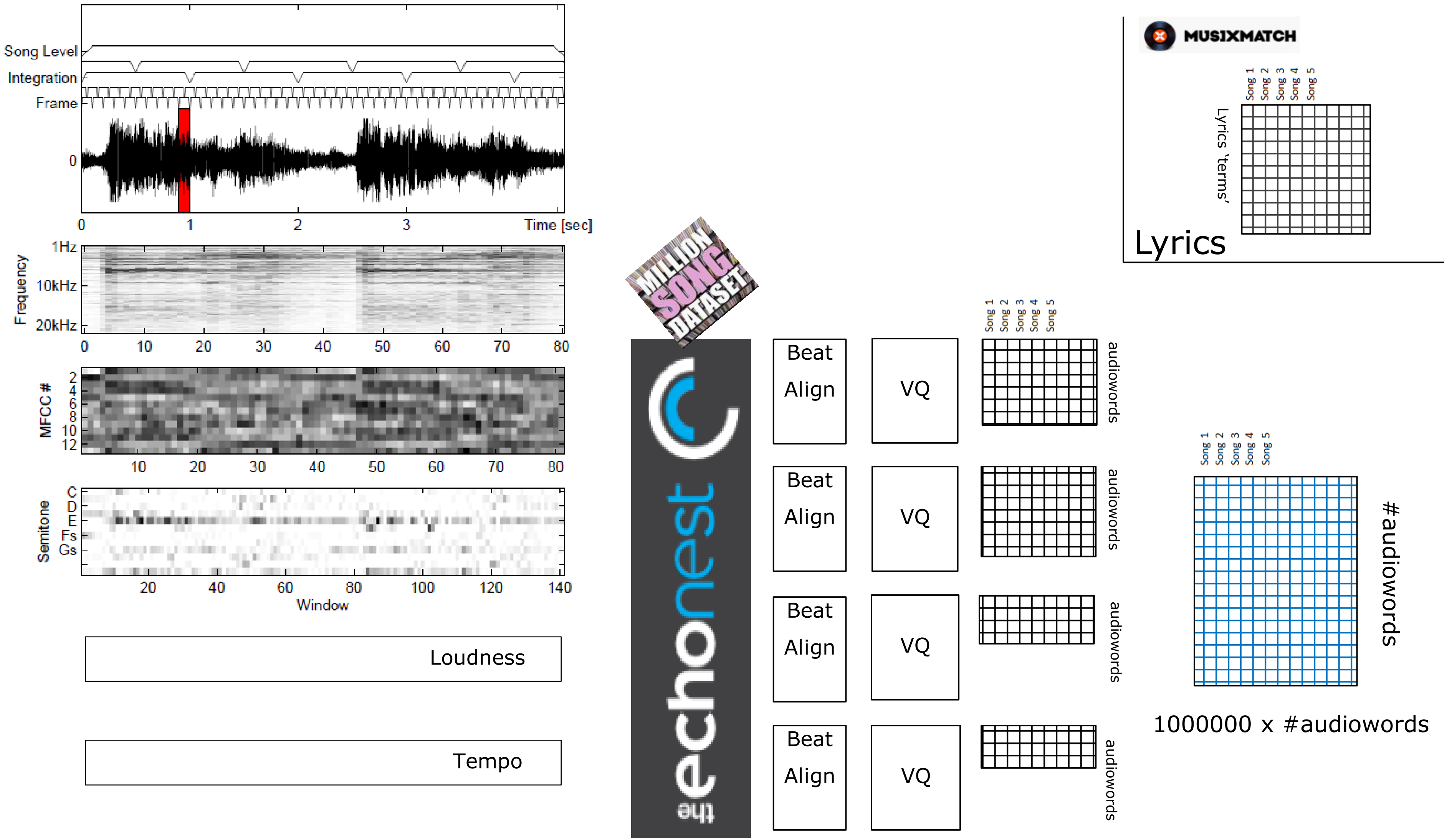
Annotation - categories and tags

Genre/style

Open vocabulary tags



Bottom-up view – *content driven*



Two elements of the framework

Computational representation of audio

- Goal is to construct a scalable a universal representation/model which supports many of the defined tasks – and preferably inline with the users representation

Elicitation of user preferences in audio

- Goal is to efficiently and robustly to elicit, model and predict top-down aspects such as preference and other perceptual and cognitive aspects

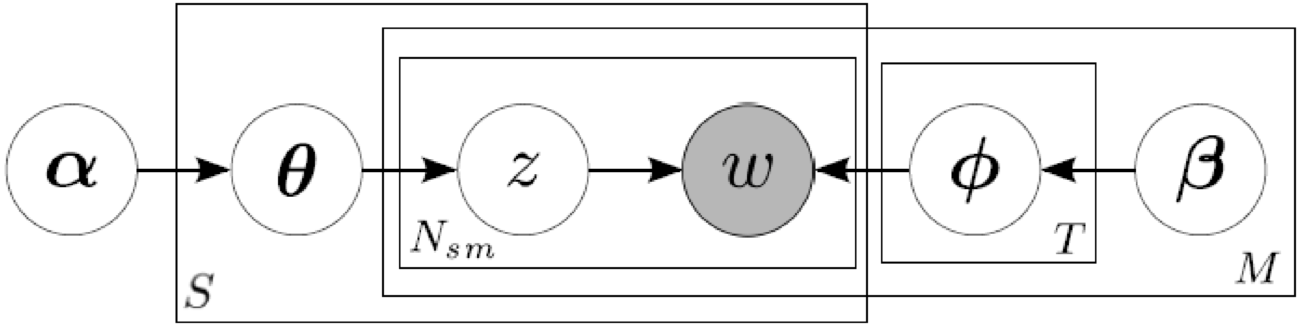
Multi-modal Latent Dirichelt Allocation model

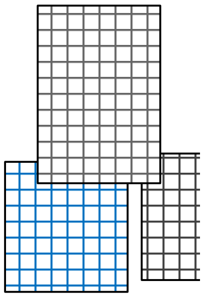
Bjørn Sand Jensen, Rasmus Troelsgaard, Jan Larsen and Lars Kai Hansen, *Towards a universal representation for audio information retrieval and analysis*, International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2013.

Is latent representation obtained by considering the audio and lyrics modalities is well aligned -in an unsupervised manner – with ’cognitive’ variables ?

Is it possible to predict evaluate human categories and metadata information from latent representation?

mmLDA model





- For each topic $z \in [1;T]$ in each modality $m \in [1;M]$
  Draw $\phi_z^{(m)} \sim \mathit{Dirichlet}(\boldsymbol{\beta}^{(m)})$.
  This is the parameters of the z^{th} topic's distribution over vocabulary $[1;V^{(m)}]$ of modality m.
- For each song $s \in [1;S]$
  - Draw $\boldsymbol{\theta}_s \sim \mathit{Dirichlet}(\boldsymbol{\alpha})$.
    This is the parameters of the s^{th} song's distribution over topics $[1;T]$.
  - For each modality $m \in [1;M]$
    * For each word $w \in [1;N_{sm}]$
      · Draw a specific topic $z^{(m)} \sim \mathit{Categorical}(\boldsymbol{\theta}_s)$
      · Draw a word $w^{(m)} \sim \mathit{Categorical}(\boldsymbol{\phi}_{z^{(m)}}^{(m)})$

Elements of the inference

- Collapsed Gibbs sampling
- Each Gibbs sampler is run for a limited number of completesweeps through the training songs
- The model state with the highest model evidence within the last 50 iterations is regarded as a MAP estimate from which point estimates of the
  - topic-song, $p(z|s)$
  - and the modality specific word-topic $p(w^{(m)}|z)$

  and distributions are taken using the expectations of the corresponding Dirichlet distributions.
- Evaluation of model performance on unknown test songs, s, is performed using the procedure of fold-in by estimating the topic distribution, $p(z|s)$ for the new song, by keeping the all the word-topic counts fixed during a number of new Gibbs sweeps.
- Testing on a modality not included in the training phase requires an estimate of the word-topic distribution, $p(w(m)|z)$, of the held out modality, m. This is obtained by keeping the song-topic counts fixed while only updating the word-topic counts for that specific modality.

Million Song Dataset

- Music Data
- Tags
- Lyrics
- Audio features
- Vector quantisation → Audio words
- Genre and Style labels

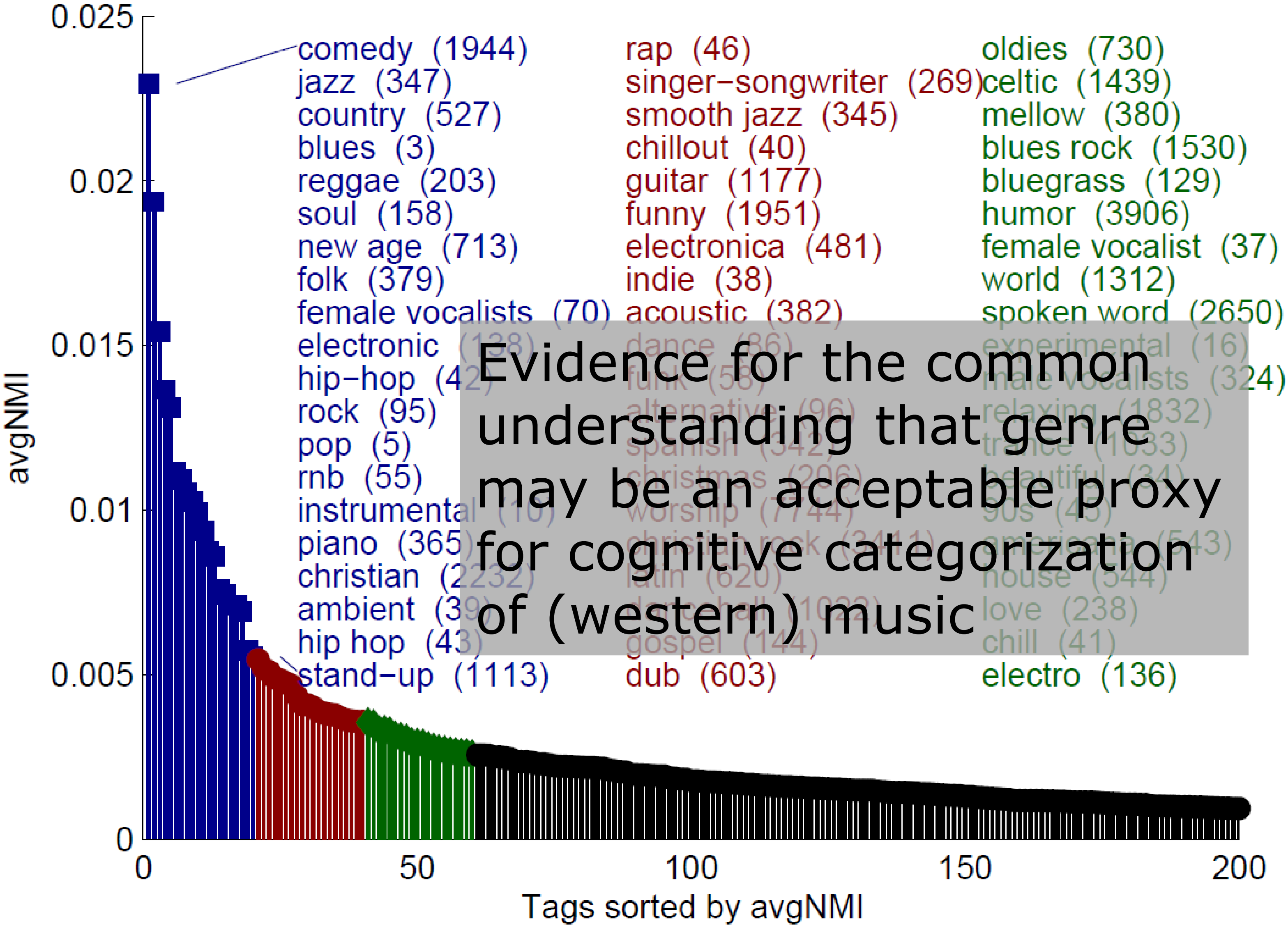
Normalized mutual information between a single tag and the latent topic representations

$$\mathrm{MI}\left(w_i{}^{(tag)}, z|s\right)$$

$$= \mathrm{KL}\left(\hat{p}\left(w_i{}^{(tag)}, z|s\right) || \hat{p}\left(w_i{}^{(tag)}|s\right)\hat{p}\left(z|s\right)\right),$$

$$\mathrm{NMI}\left(w_i{}^{(tag)}, z|s\right) = 2\frac{\mathrm{MI}\left(w_i{}^{(tag)}, z|s\right)}{H\left(w_i{}^{(tag)}|s\right) + H\left(z|s\right)}$$

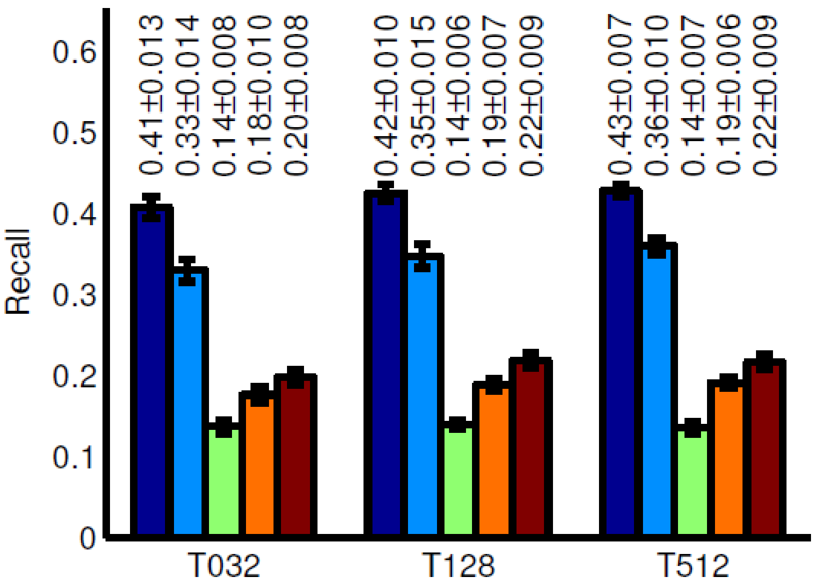
$$\mathrm{avgNMI}(w_i{}^{(tag)}) = \frac{1}{N_s}\sum_{s=1}^{Ns}\mathrm{NMI}\left(w_i{}^{(tag)}, z|s\right)$$



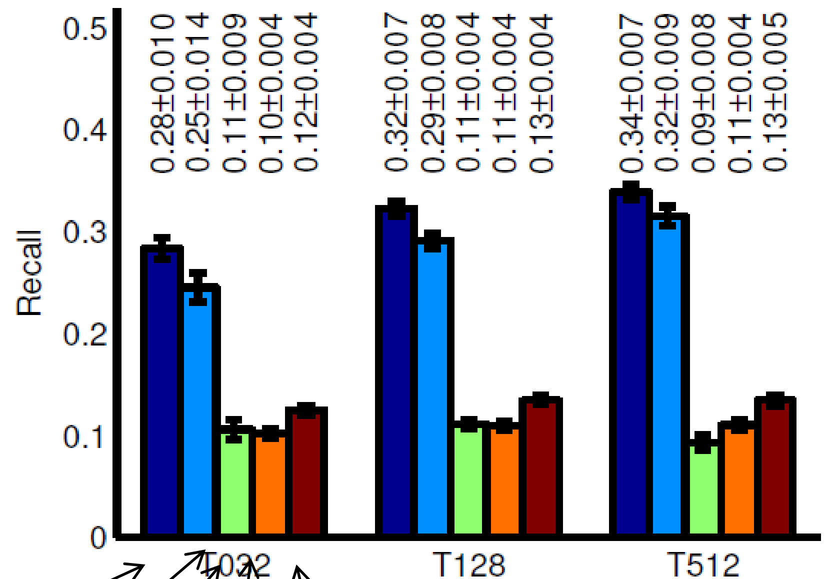
Evidence for the common understanding that genre may be an acceptable proxy for cognitive categorization of (western) music

128 topics using audio and lyrics modalities

Genre and style prediction



(a) Genre



(b) Style

$$\text{Precision} = \frac{tp}{tp + fp}$$

$$\text{Recall} = \frac{tp}{tp + fn}$$

Combined

Tags

Lyrics

Audio

Audio+lyrics

Genre specific classification error

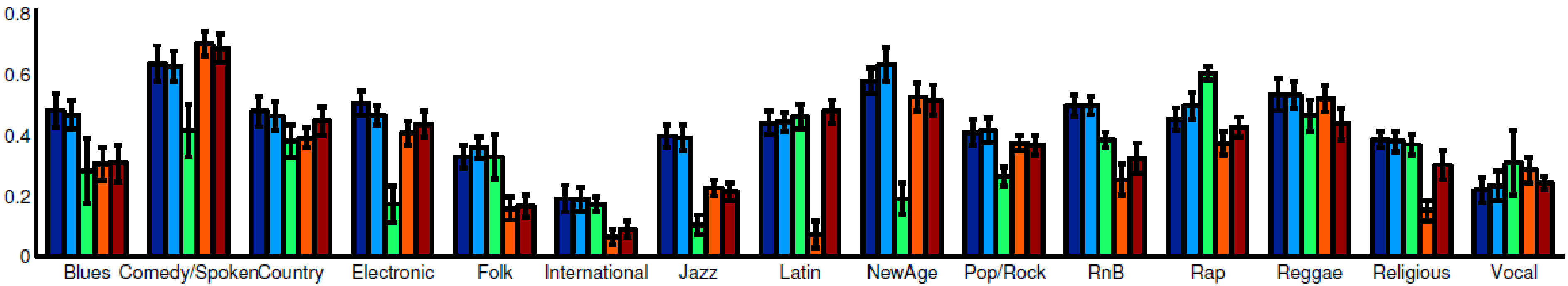
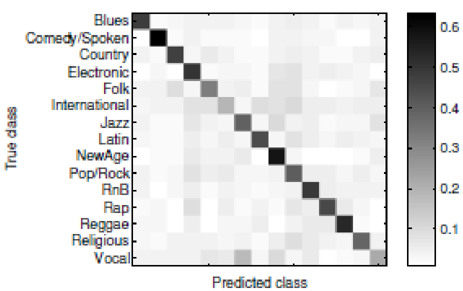
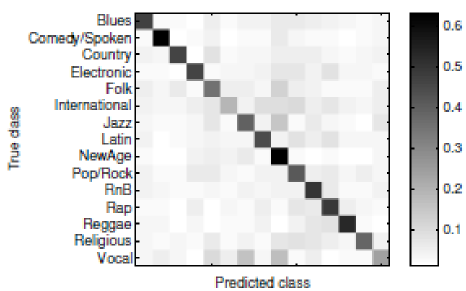


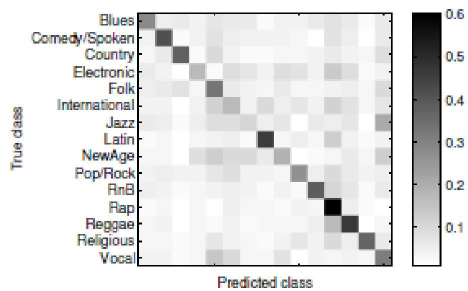
Fig. 4: Dark blue: Combined model, Light Blue: Tags, Green: Lyrics, Orange: Audio, Red: Audio+Lyrics, genre, $T = 128$.



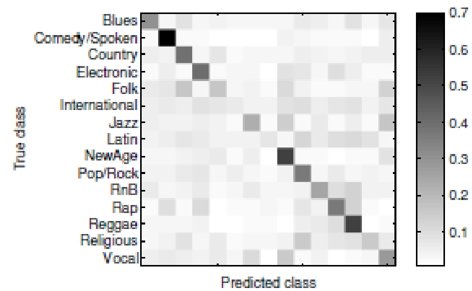
(a) Combined Model



(b) Tag Model



(c) Lyrics Model



(d) Audio Model

Preference eliciation

- Bjørn Sand Jensen, Jens Brehm Nielsen, and Jan Larsen. *Efficient Preference Learning with Pairwise Continuous Observations and Gaussian Processes*, IEEE International Workshop on Machine Learning for Signal Processing, 2011.
- Bjørn Sand Jensen, Javier Saez Gallego and Jan Larsen. *A Predictive model of music preference using pairwise comparisons*. International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2012.
- Jens Madsen, Bjørn Sand Jensen, Jan Larsen and Jens Brehm Nielsen. *Towards Predicting Expressed Emotion in Music from Pairwise Comparisons*, 9th Sound and Music Computing Conference, 2012.
- Jens Madsen, Jens Brehm Nielsen, Bjørn Sand Jensen and Jan Larsen. *Modeling Expressed Emotions in Music using Pairwise Comparisons*. 9th International Symposium on Computer Music Modeling and Retrieval (CMMR) 2012.
- Jens Brehm Nielsen, Bjørn Sand Jensen and Jan Larsen, *Pseudo Inputs For Pairwise Learning With Gaussian Processes*, IEEE International Workshop on Machine Learning for Signal Processing, 2012.
- Jens Brehm Nielsen, Jakob Nielsen: Efficient Individualization of Hearing and Processers Sound, ICASSP2013.

Preference elicitation refers to the problem of developing a decision support system capable of generating recommendations to a user, thus assisting him in decision making. It is important for such a system to model user's preferences accurately, find hidden preferences and avoid redundancy. This problem is sometimes studied as a **computational learning theory** problem

Ref: Wikipedia

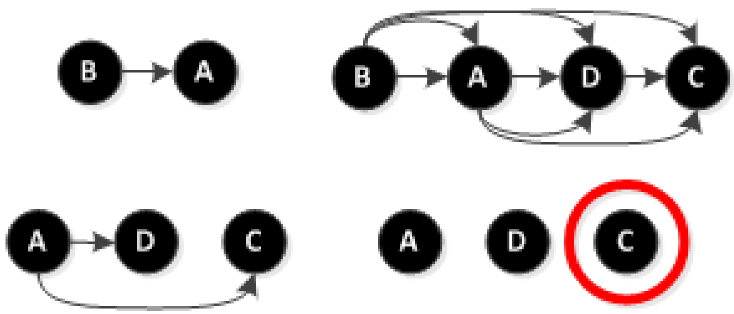
Main assumption

User preference recorded from behavior and interactions is a proxy for aspects of human cognition

Indirect or relative scaling

- Task is comparing a set of objects and rank them in order or assign a value to the similarity between them.
- Elicitation by relative comparisons eliminates the need for absolute references and explanation - less why questions!
- Difficult to articulate experience/opinion
- Issues related to learning from limited number of songs

2AFC (Pairwise), k-AFC, ranking, odd-one out.



Similarity / Continuous (degree of preference/ confidence)

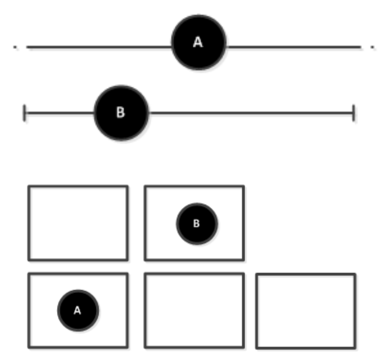
Direct or absolute sacling

- Elicitates a specific aspect
- Learning from few songs might by complex due to perceptual and cognitive processes
- Difficult to understand/explain scale
- Difficult to consistently rate music/settings/emotions on direct scales (dimensional or categorical)
  - communication biases due to uncertainties in scales, anchors or labels
  - lack of references causes drift and inconsistencies

Infinite, ordinal, bounded, continuous scale

Categorical (classification):

Binary / multi-class



The background: Weber's law

'Just noticable difference' is relative to stimuli strength

$$dp = k\ dS/S$$

Perception

Stimuli, e.g. weight

prop. constant

$$p = k\ \ln\left(\frac{S}{S_0}\right)$$

"Weber's Law", *Encyclopedia Americana*, 1920.

Pairwise comparison versus direct scaling

- Thurnstones ”Priciple of comparative judments”
  - ”The discrimal process” – the total process of discrimating stimuli
  - Assumptions
    1. preference (utility function, or in Thurstone's terminology, *discriminal process*) for each stimulus
    2. The stimulus whose value is larger at the moment of the comparison will be preferred by the subject
    3. These unobserved preferences are normally distributed in the population
- The “phsycological scale is at best an artificial construct” (Thurnstone)
- Lockhead claims that everything is relative……

G. R. Lockhead, “Absolute Judgments Are Relative: A Reinterpretation of Some Psychophysical Ideas.,” Review of General Psychology, vol. 8, no. 4, pp. 265–272, 2004.

L. L. Thurstone, “A law of comparative judgement.,” Psychological Review, vol. 34, 1927.

A. Maydeu-Olivares: ”On Thutstone’s Model For Paired Comparisons and Ranking Data”, Barcelona Univ.

A non-parametric approach

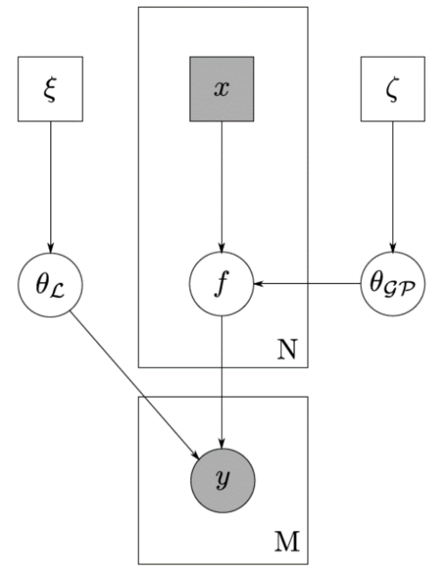
$$p\left(y_k|\mathbf{f}_k,\sigma\right) = \Phi\left(y_k \frac{f\left(\mathbf{x}_{u_k}\right) - f\left(\mathbf{x}_{v_k}\right)}{\sqrt{2}\sigma_{\mathcal{L}}}\right)$$

$$p\left(\mathcal{Y}|\mathcal{X}\right) = \prod_{k=1}^{K} p\left(y_k|\mathbf{f}_k,\boldsymbol{\theta}_{\mathcal{L}}\right)$$

$$\mathbf{f}\ |\sigma_s,\sigma_\ell \sim \mathcal{GP}\left(\mathrm{m}\left(\mathbf{x}\right),\mathrm{k}\left(\mathbf{x},\cdot\right)_{\sigma_s,\sigma_\ell}\right)$$

$$p\left(\mathbf{f},\boldsymbol{\theta}|\mathcal{Y},\mathcal{X}\right) = \frac{p\left(\boldsymbol{\theta}_{\mathcal{GP}}\right)p\left(\mathbf{f}|\boldsymbol{\theta}_{\mathcal{GP}},\mathcal{X}\right)p\left(\boldsymbol{\theta}_{\mathcal{L}}\right)p\left(\mathcal{Y}|\mathbf{f},\boldsymbol{\theta}_{\mathcal{L}}\right)}{p\left(\mathcal{Y}|\mathcal{X}\right)}$$

$$p\left(\mathcal{Y}|\mathcal{X}\right) = \int\int\int p\left(\theta_{\mathcal{GP}}\right)p\left(\mathrm{f}|\theta_{\mathcal{GP}},\mathcal{X}\right)p\left(\theta_{\mathcal{L}}\right)p\left(\mathcal{Y}|\mathrm{f},\theta_{\mathcal{L}}\right)d\theta_{\mathcal{GP}}d\theta_{\mathcal{L}}d\mathrm{f}.$$



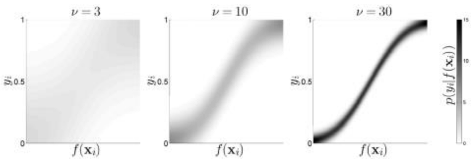
C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning*, MIT Press, 2006.

W. Chu and Z. Ghahramani, "Preference learning with Gaussian Processes," *ICML 2005 - Proceedings of the 22nd International Conference on Machine Learning*, pp. 137–144, 2005.

Framework

$$\mathbf{f}_k|\sigma_s,\sigma_\ell \sim \mathcal{GP}\left(\mathrm{m}\left(\mathbf{x}_\mathrm{k}\right), \mathrm{k}\left(\mathbf{x}_\mathrm{k},\cdot\right)_{\sigma_s,\sigma_\ell}\right)$$

$$\mathrm{k}\left(p\left(\mathbf{x}|\boldsymbol{\theta}\right), p\left(\mathbf{x}|\boldsymbol{\theta}'\right)\right) = \int \left(p\left(\mathbf{x}|\theta\right) p\left(\mathbf{x}|\boldsymbol{\theta}'\right)\right)^{1/q} d\mathbf{x}$$



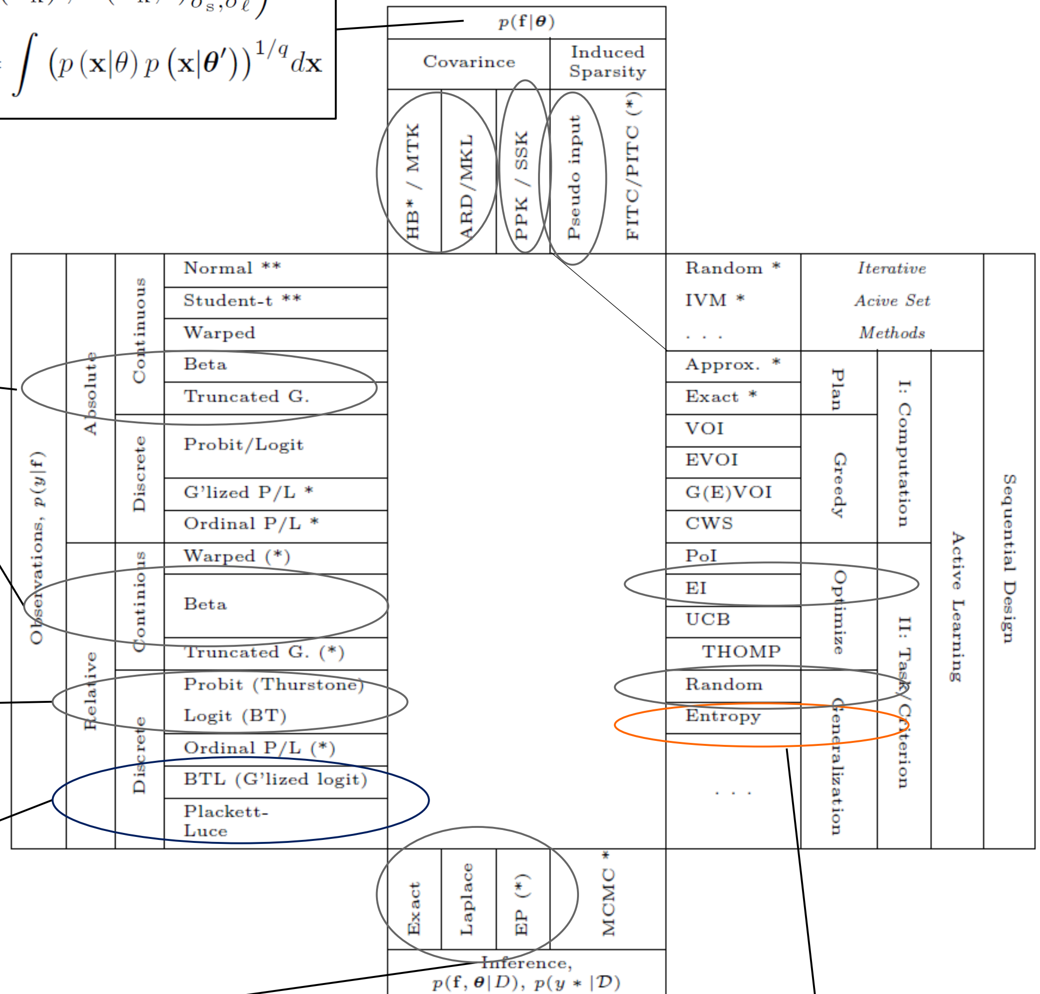
$$\beta(\mathbf{f}_k) = \nu(1-\mu(\mathbf{f}_k)), \alpha(\mathbf{f}_k) = \nu\mu(\mathbf{f}_k)$$

$$\mu\left(\mathbf{f}_k,\sigma\right) = \Phi\left(\frac{f(v_k)-f(u_k)}{\sqrt{2}\sigma}\right)$$

$$y_k \sim \mathrm{Beta}\left(\alpha(\mathbf{f}_k), \beta(\mathbf{f}_k)\right)$$

$$p\left(y_k|\mathbf{f}_k,\sigma\right) = \Phi\left(y_k \frac{f\left(\mathbf{x}_{u_k}\right) - f\left(\mathbf{x}_{v_k}\right)}{\sqrt{2}\sigma_{\mathcal{L}}}\right)$$

$$p\left(\mathbf{y}_k|\mathbf{f}_k\right) = \prod_{j=1}^{C-1} \frac{e^{f\left(\mathbf{x}_{\mathbf{y}_{k(j)}}\right)}}{\sum_{i=j}^{C} e^{f\left(\mathbf{x}_{\mathbf{y}_{k(i)}}\right)}}$$



I Approximate first level posterior, $p(\mathbf{f}|\boldsymbol{\theta},\mathcal{X},\mathcal{Y})$ using Laplace or EP with $\boldsymbol{\theta}$ fixed.

II Find ML/MAP-II point-estimates of the hyperparemetrs $\hat{\boldsymbol{\theta}}$ based on marginal likelihood approximation, provided by the first level approximation.

... iterate until convergence of $\hat{\boldsymbol{\theta}}$ or the marginal likelihood / evidence.

$$EVOI\ \left(\mathcal{E}_k\right) \equiv \iint p\left(\mathbf{f}_k|\mathcal{E}_k,\mathcal{D}\right) p\left(y_k|\mathbf{f}_k,\mathcal{D}\right) \log p\left(y_k|\mathbf{f}_k,\mathcal{D}\right) dy d\mathbf{f} - \int p\left(y_k|\mathcal{E}_k,\mathcal{D}\right) \log p\left(y_k|\mathcal{E}_k,\mathcal{D}\right) dy$$

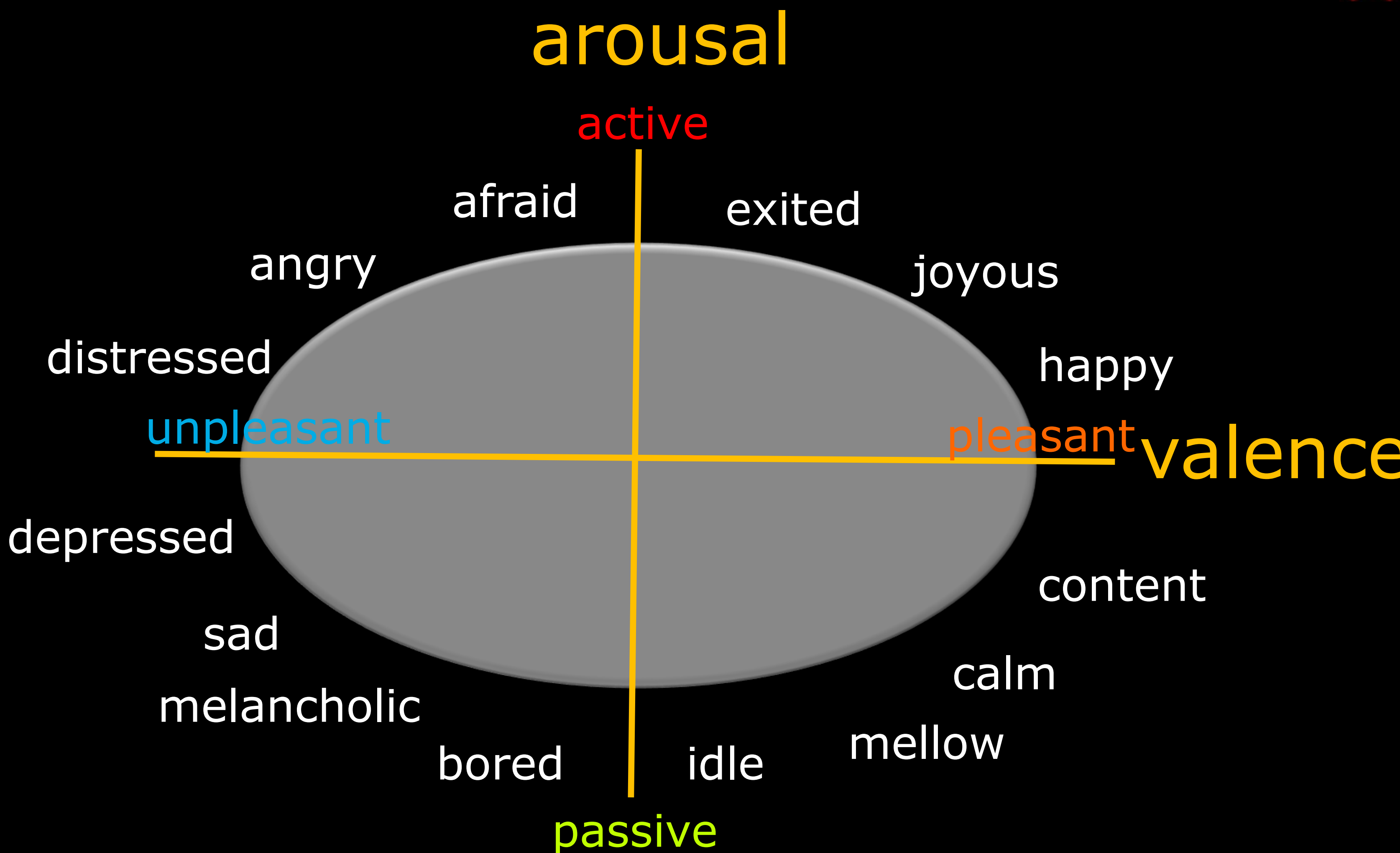
Expressed emotions

- Jens Madsen, Bjørn Sand Jensen, Jan Larsen and Jens Brehm Nielsen. *Towards Predicting Expressed Emotion in Music from Pairwise Comparisons*, 9th Sound and Music Computing Conference, 2012.
- Jens Madsen, Jens Brehm Nielsen, Bjørn Sand Jensen and Jan Larsen. *Modeling Expressed Emotions in Music using Pairwise Comparisons*. 9th International Symposium on Computer Music Modeling and Retrieval (CMMR) 2012.
- Madsen, J., Jensen, B.S., Larsen, J., Predictive modeling of expressed emotions in music using pairwise comparisons. M. Aramaki et al. (Eds.): CMMR 2012, LNCS 7900, pp. 253–277, 2013. Springer-Verlag Berlin Heidelberg 2013.

Is it possible to model the users representation of expressed emotion using pairwise comparisons?

Which scaling method should we use?

Emotional spaces



J. A. Russel: "A Circumplex Model of Affect," *Journal of Personality and Social Psychology*, 39(6):1161, 1980

J. A. Russel, M. Lewicka, and T. Niit, "A Cross-Cultural Study of a Circumplex Model of Affect," *Journal of Personality and Social Psychology*, vol. 57, pp. 848-856, 1989

Experimental setup

- **20 excerpts** of **15 second** length were chosen to be evenly distributed in the AV space using a linear regression model and subjective evaluation.
- **8 participants** each evaluated all **190 unique pairwise comparisons**.
- **Question to participants:** Which sound clip was the most

(Arousal) ***excited, active, awake?*** and (Valence) ***positive, glad, happy?***

Audio representation

- 30 dimensions of Mel-frequency cepstral coefficients (MFCC).
- Spectral- flux, roll-off, slope and variation (SSD).
- Zero crossing rate and statistical shape descriptors (TSS).

Features extracted by YAAFE (Yet-Another-Audio-Feature-Extraction) Toolbox

Performance using different audio features

Training size	5%	7%	10%	20%	40%	60%	80%	100%
MFCC	0.3402	0.2860	0.2455	0.2243	0.2092	0.2030	0.1990	0.1949
Envelope	0.4110*	0.4032	0.3911	0.3745	0.3183	0.2847	0.2780	0.2761
Chroma	0.3598	0.3460	0.3227	0.2832	0.2510	0.2403	0.2360	0.2346
CENS	0.3942	0.3735	0.3422	0.2994	0.2760	0.2676	0.2640	0.2621
CRP	0.4475	0.4336	0.4115	0.3581	0.2997	0.2790	0.2735	0.2729
Sonogram	0.3325	0.2824	0.2476	0.2244	0.2118	0.2061	0.2033	0.2026
Pulse clarity	0.4620	0.4129	0.3698	0.3281	0.2964	0.2831	0.2767*	0.2725
Loudness	0.3261	0.2708	**0.2334**	**0.2118**	**0.1996**	**0.1944**	**0.1907**	**0.1862**
Spec. disc.	0.2909	0.2684	0.2476	0.2261	0.2033	0.1948	0.1931	0.1951
Spec. disc. 2	0.3566	0.3223	0.2928	0.2593	0.2313	0.2212	0.2172	0.2138
Key	0.5078	0.4557	0.4059	0.3450	0.3073*	0.2959	0.2926	0.2953
Tempo	0.4416	0.4286	0.4159	0.3804	0.3270	0.3043	0.2953	0.2955
Fluctuations	0.4750	0.4247	0.3688	0.3117	0.2835	0.2731	0.2672	0.2644*
Pitch	0.3173	0.2950	0.2668	0.2453	0.2301	0.2254	0.2230	0.2202
Roughness	**0.2541**	**0.2444**	0.2367	0.2304	0.2236	0.2190	0.2168	0.2170
Spectral crest	0.4645	0.4165	0.3717	0.3285	0.2979	0.2866*	0.2828	0.2838
Echo. timbre	0.3726	0.3203	0.2797	0.2524	0.2366	0.2292	0.2258	0.2219
Echo. pitch	0.3776	0.3264	0.2822	0.2492	0.2249	0.2151	0.2089	0.2059
$Base_{low}$	0.4122	0.3954	0.3956	0.3517	0.3087	0.2879	0.2768	0.2702

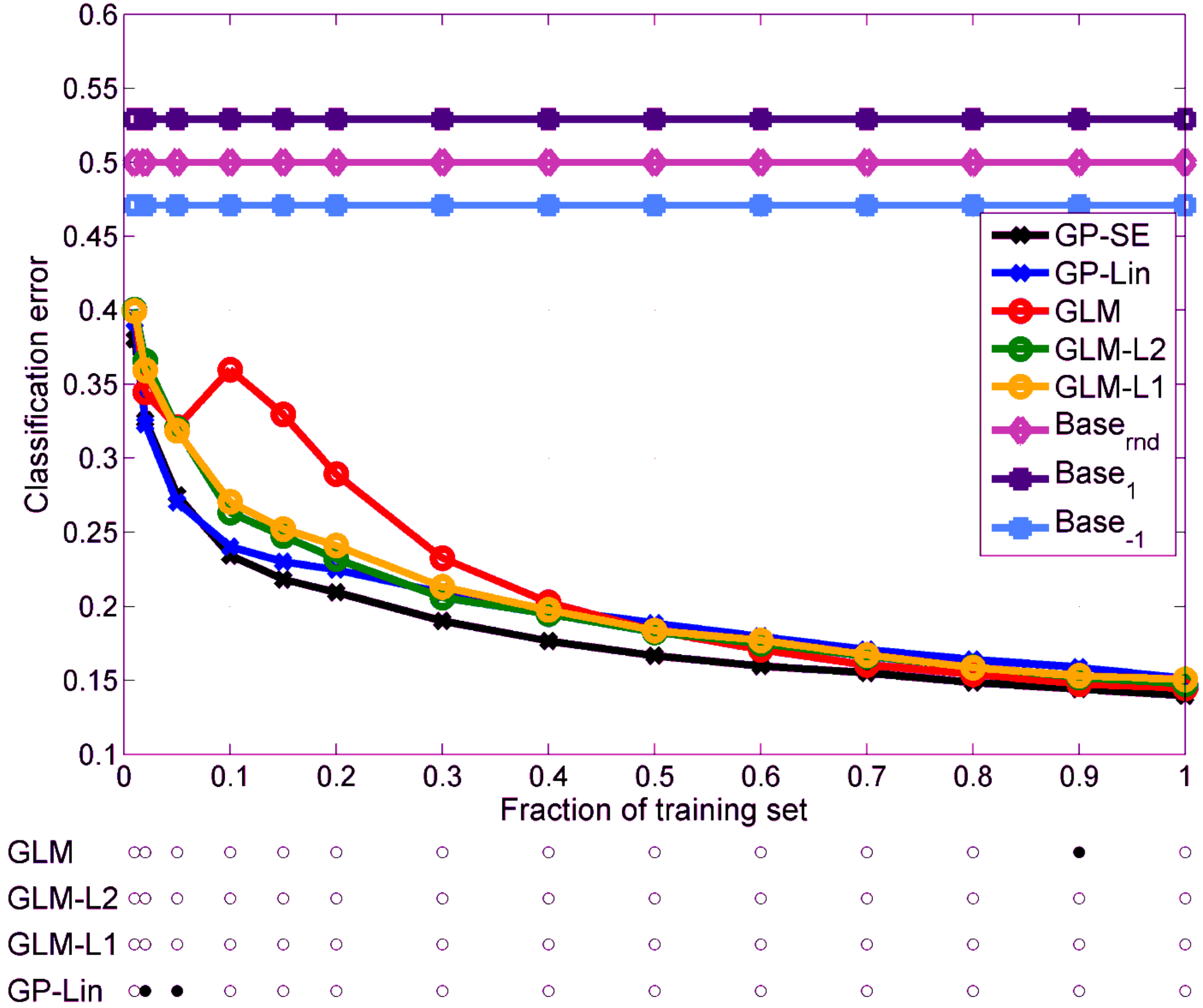
Table 4.2. Arousal: Classification error learning curves as an average of 50 repetitions and 13 individual user models, using only the mean of the features. McNemar test between all points on the learning curve and $Base_{low}$ resulted in $p < 0.05$ for all models except results marked with *, with a sample size of 12.350

Performance using different audio features

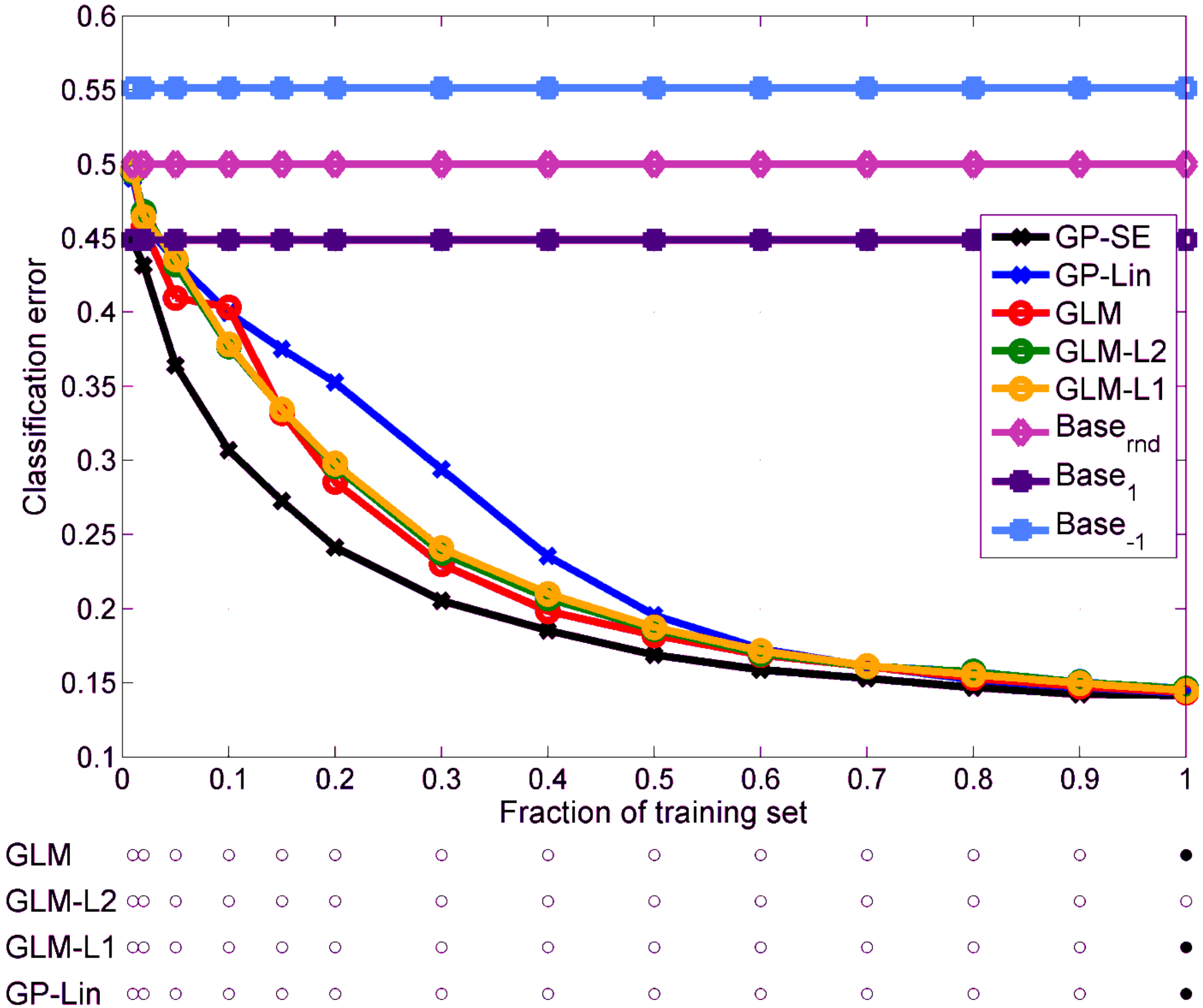
Training size	5%	7%	10%	20%	40%	60%	80%	100%
MFCC	0.4904	0.4354	0.3726	0.3143	0.2856	0.2770	0.2719	0.2650
Envelope	**0.3733**	**0.3545**	0.3336	0.3104	0.2920	0.2842	0.2810	0.2755
Chroma	0.4114*	0.3966*	0.3740	0.3262	0.2862	0.2748	0.2695	0.2658
CENS	0.4353	0.4139	0.3881	0.3471	0.3065	0.2948	0.2901*	0.2824
CRP	0.4466	0.4310	0.4111	0.3656	0.3066	0.2925	0.2876	0.2826
Sonogram	0.4954	0.4360	0.3749	0.3163	0.2884	0.2787	0.2747	0.2704
Pulse clarity	0.4866	0.4357	0.3856	0.3336	0.3026	0.2930	0.2879	0.2810
Loudness	0.4898	0.4310	0.3684	0.3117	0.2854	0.2768	0.2712	0.2664
Spec. disc.	0.4443	0.4151	0.3753	0.3263	0.2939	0.2857	0.2827	0.2794
Spec. disc. 2	0.4516	0.4084	0.3668	0.3209	0.2916	0.2830	0.2781	0.2751
Key	0.5303	0.4752	0.4104	0.3370	0.2998	0.2918	0.2879	0.2830*
Tempo	0.4440	0.4244	0.3956	0.3559*	0.3158	0.2985	0.2933	0.2883
Fluctuations	0.4015	0.3584	**0.3141**	**0.2730**	**0.2507**	**0.2433**	**0.2386**	**0.2340**
Pitch	0.4022	0.3844	0.3602	0.3204	0.2926	0.2831	0.2786	0.2737
Roughness	0.4078	0.3974	0.3783	0.3313	0.2832	0.2695	0.2660	0.2605
Spec. crest	0.4829	0.4289	0.3764	0.3227	0.2994	0.2942	0.2933	0.2923
Echo. timbre	0.4859	0.4297	0.3692	0.3127	0.2859	0.2767	0.2732	0.2672
Echo. pitch	0.5244	0.4643	0.3991*	0.3275	0.2942	0.2841	0.2790	0.2743
$Base_{low}$	0.4096	0.3951	0.3987	0.3552	0.3184	0.2969	0.2893	0.2850

Table 4.1. Valence: Classification error learning curves as an average of 50 repetitions and 13 individual user models, using both mean and standard deviation of the features. McNemar test between all points on the learning curve and $Base_{low}$ resulted in $p < 0.05$ for all models except results marked with *, with a sample size of 12.350

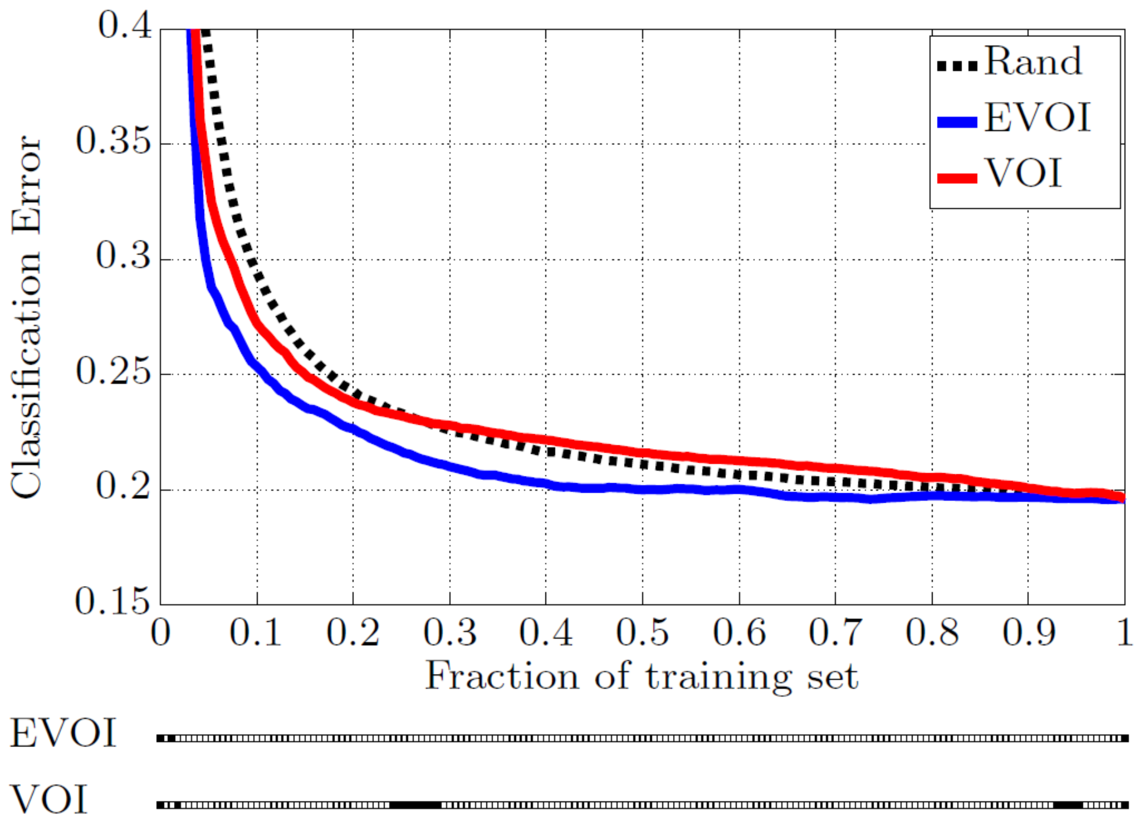
Learning Curve (Arousal)



Learning Curve (Valence)



How many pairwise comparisons do we need to model emotions?



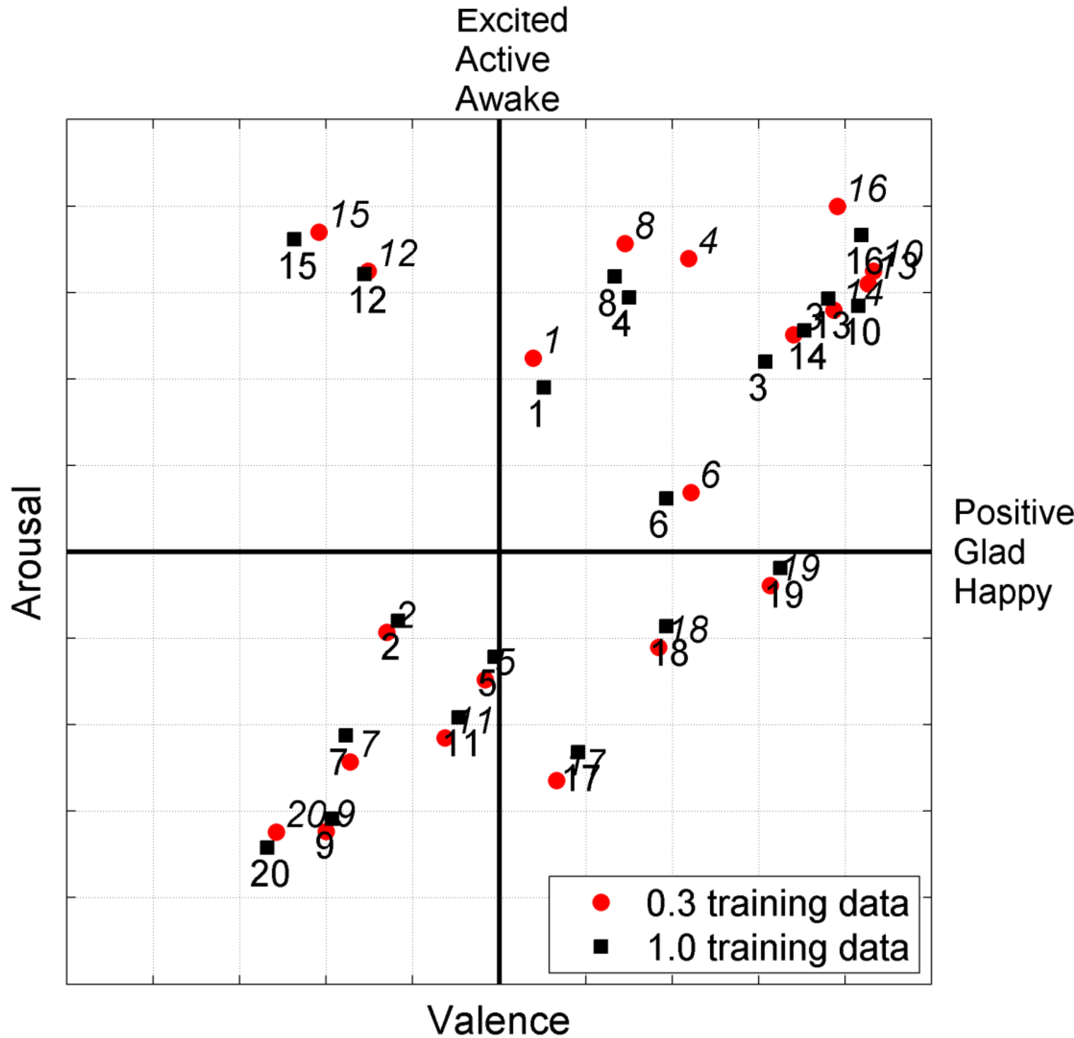
Using active learning

15% for valence

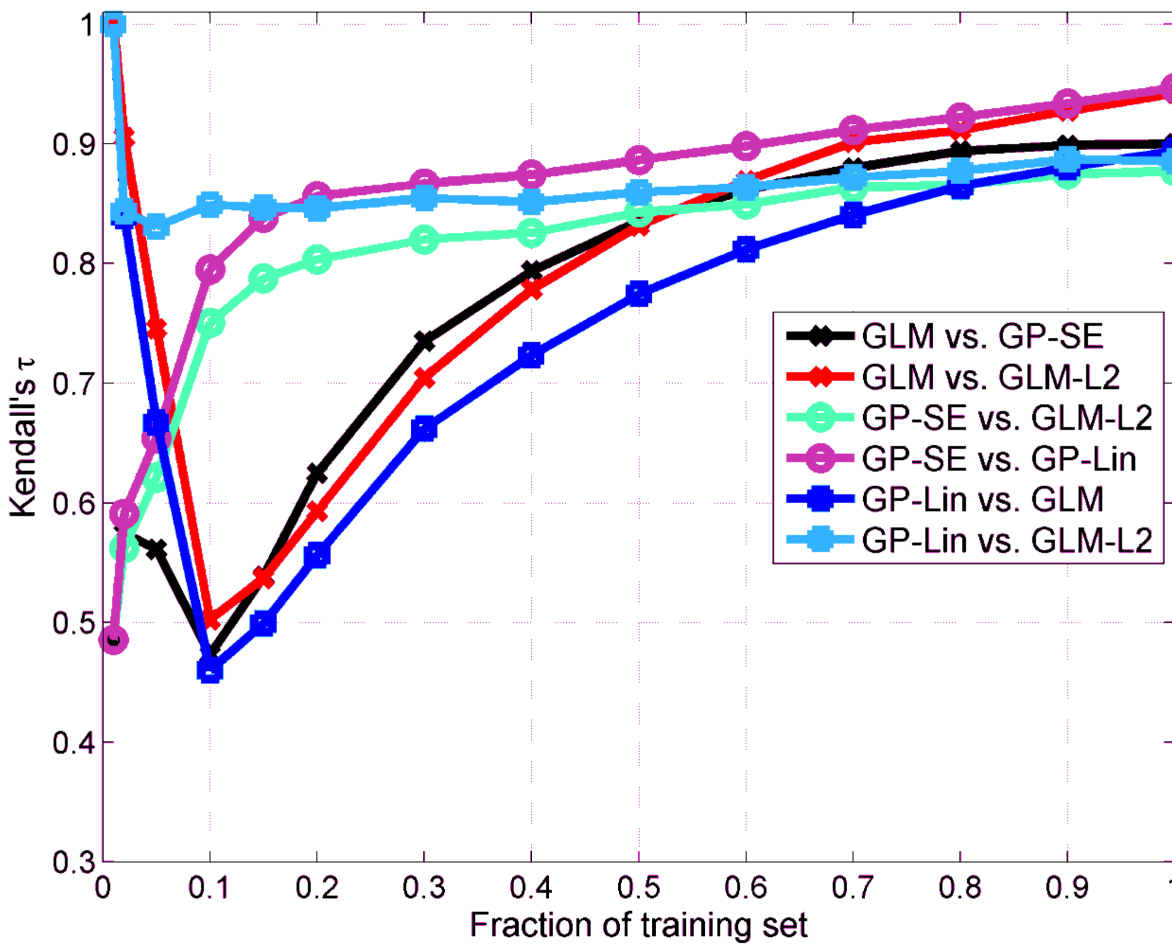
9% for arousal

AV-space

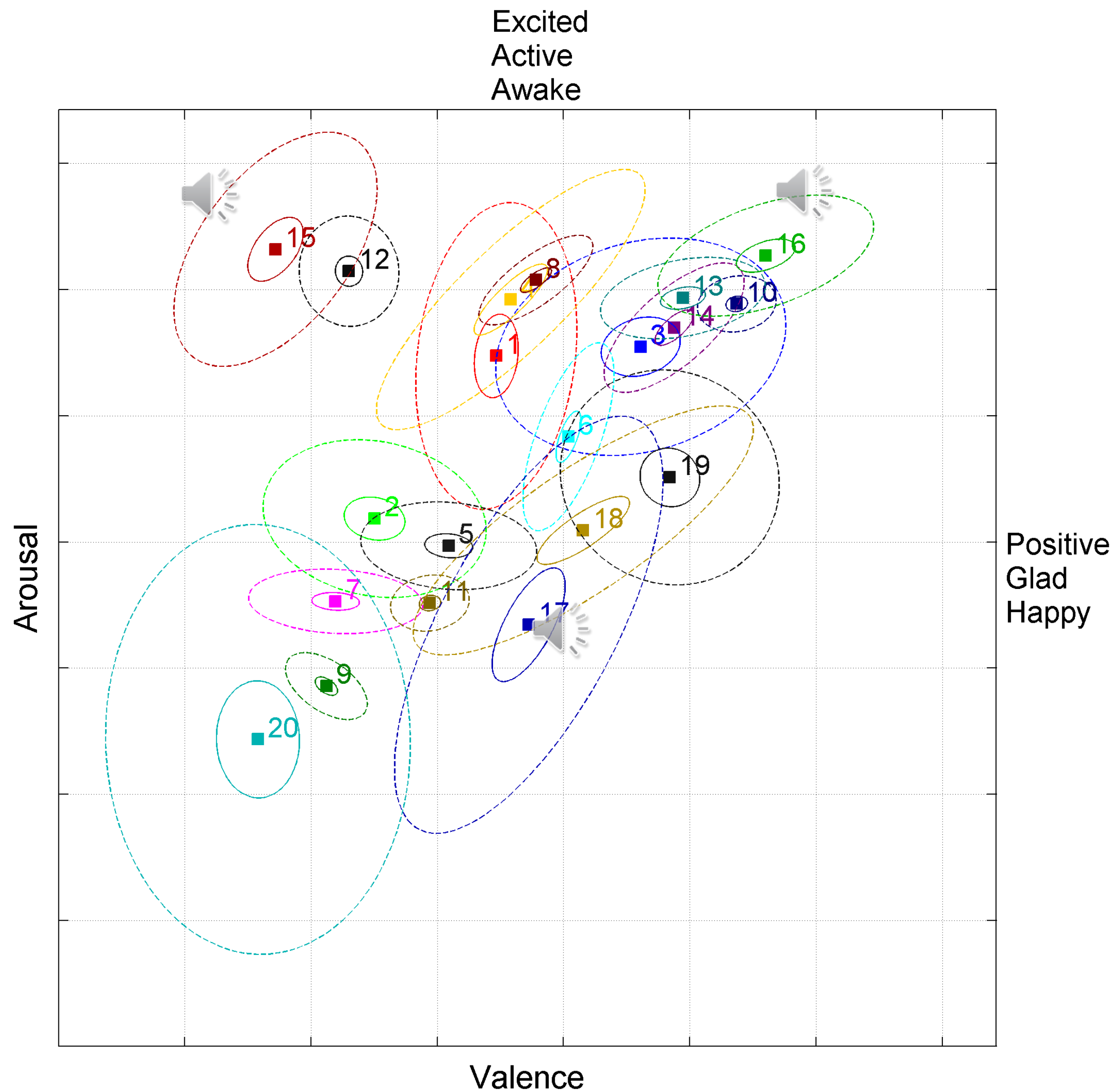
- No. Song name
- 1 311 - T and p combo
- 2 A-Ha - Living a boys adventure
- 3 Abba – That’s me
- 4 ACDC - What do you do for money hone
- 5 Aaliyah - The one I gave my heart to
- 6 Aerosmith - Mother popcorn
- 7 Alanis Morissette - These r the thoughts
- 8 Alice Cooper – I’m your gun
- 9 Alice in Chains - Killer is me
- 10 Aretha Franklin - A change
- 11 Moby – Everloving
- 12 Rammstein - Feuer frei
- 13 Santana - Maria caracoles
- 14 Stevie Wonder - Another star
- 15 Tool - Hooker with a pen..
- 16 Toto - We made it
- 17 Tricky - Your name
- 18 U2 - Babyface
- 19 UB40 - Version girl
- 20 ZZ top - Hot blue and righteous



Are rankings dependent on model choice? Ranking difference (Arousal)

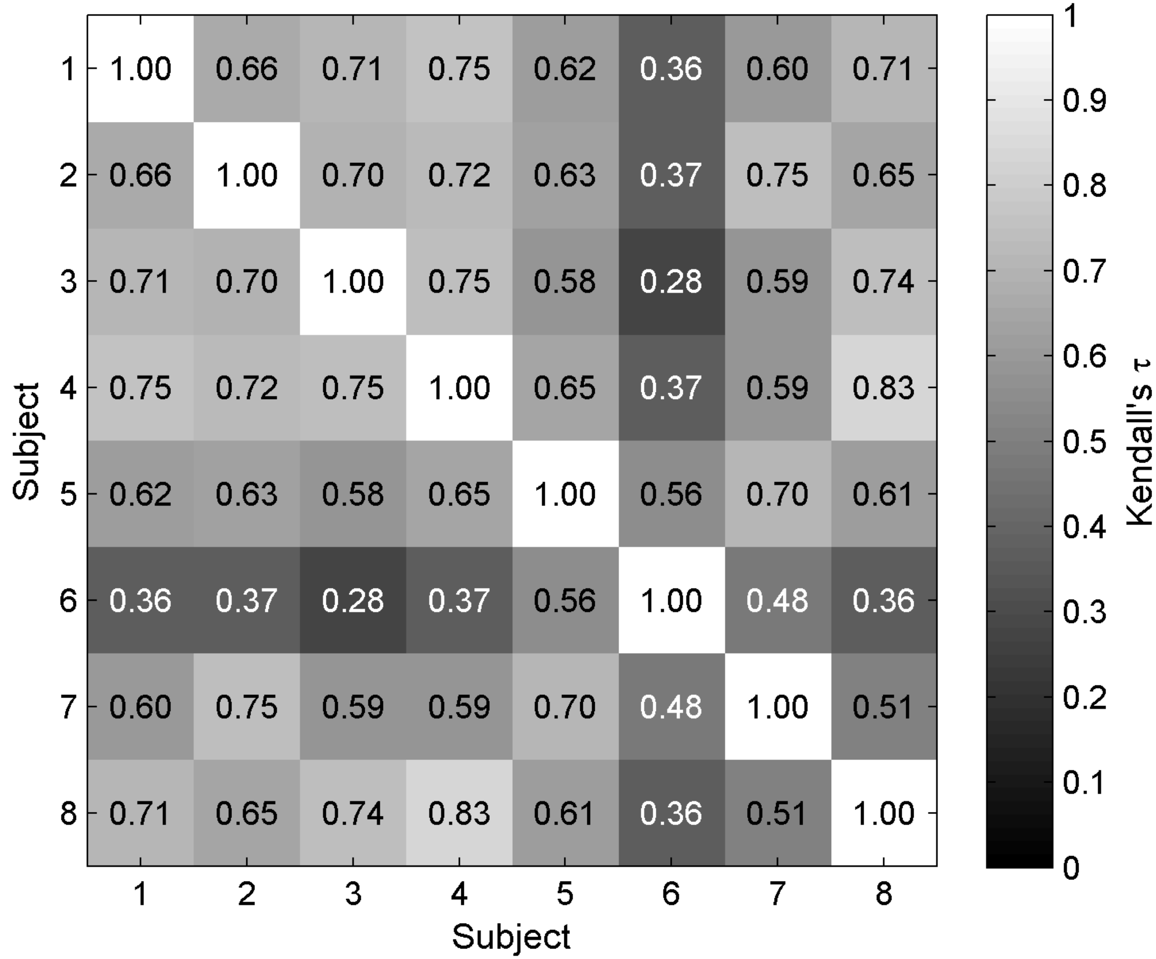


Is ranking of music subject dependent?



Valence / Arousal Space for GP model

Subjective difference in ranking (Arousal)



Main conclusion on eliciting emotions

- Models produce similar results using a learning curve
- Models produce different rankings specially when using a fraction of comparisons
- Large individual differences between the ranking of music expressed in music on dimensions of Valence and Arousal
- Promising error rates for both arousal and valence using as little as 30% of the training set corresponding to 2.5 comparisons per excerpt.
- Pairwise comparisons (2AFC) can scale when using active learning.

Music preference

- Bjørn Sand Jensen, Jens Brehm Nielsen, and Jan Larsen. *Efficient Preference Learning with Pairwise Continuous Observations and Gaussian Processes*, IEEE International Workshop on Machine Learning for Signal Processing, 2011.

Is it possible to model, interpret and predict individual music preference based on low-level audio features and pairwise comparisons?

Music Preference

Pilot study with:

…Classical, Rock/Pop, Heavy)

…30 sec) in each Genre

…s (students, 23-31 years, evaluation

…at home)

… 420 unique comparison based on a

…ign

…esenting two tracks: *Which song do*

$$\mathbf{f}_k | \sigma_s, \sigma_\ell \sim \mathcal{GP}\left(\mathrm{m}\left(\mathbf{x}_\mathrm{k}\right), \mathrm{k}\left(\mathbf{x}_\mathrm{k}, \cdot\right)_{\sigma_s, \sigma_\ell}\right)$$

$$\pi_k | \mathbf{f}_k, \sigma_\mathcal{L} = \Phi\left(y_k \frac{f\left(\mathbf{x}_{u_k}\right) - f\left(\mathbf{x}_{v_k}\right)}{\sqrt{2}\sigma}\right)$$

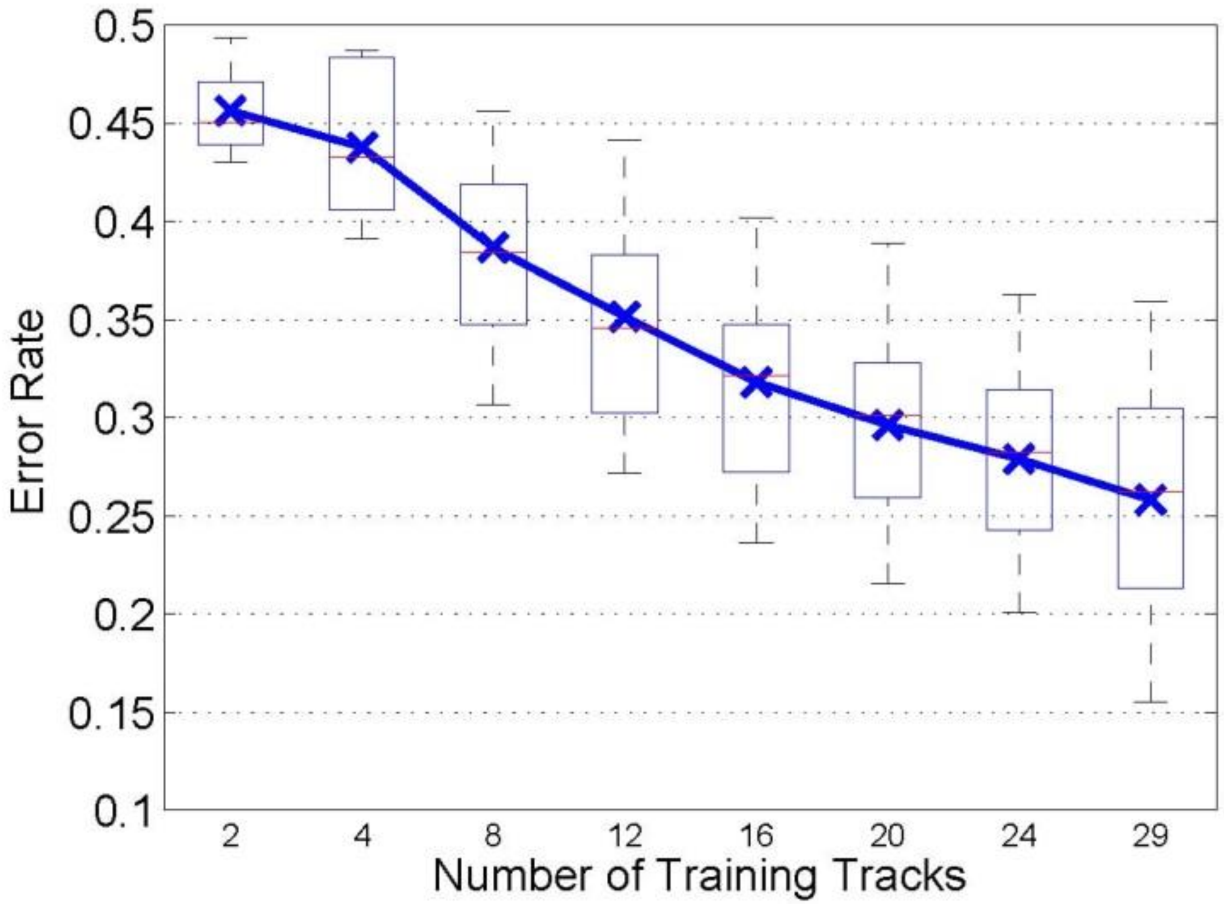
$$y_k \sim \mathrm{Bernoulli}\left(\pi_k\right)$$

Instances / tracks:

- Standard Audio Features using the Intelligent Sound Processing Toolbox `http://kom.aau.dk/project/isound/`
- MFCCs (26 dimensions, 1999 frames, incl. delta coefficients)
- $p(x)$ modeled with a two component Gaussian Mixture Model (GMM) for track: $p(x) = \sum p(z)p(x|z)$
- GMM fitted using K-means initialized EM

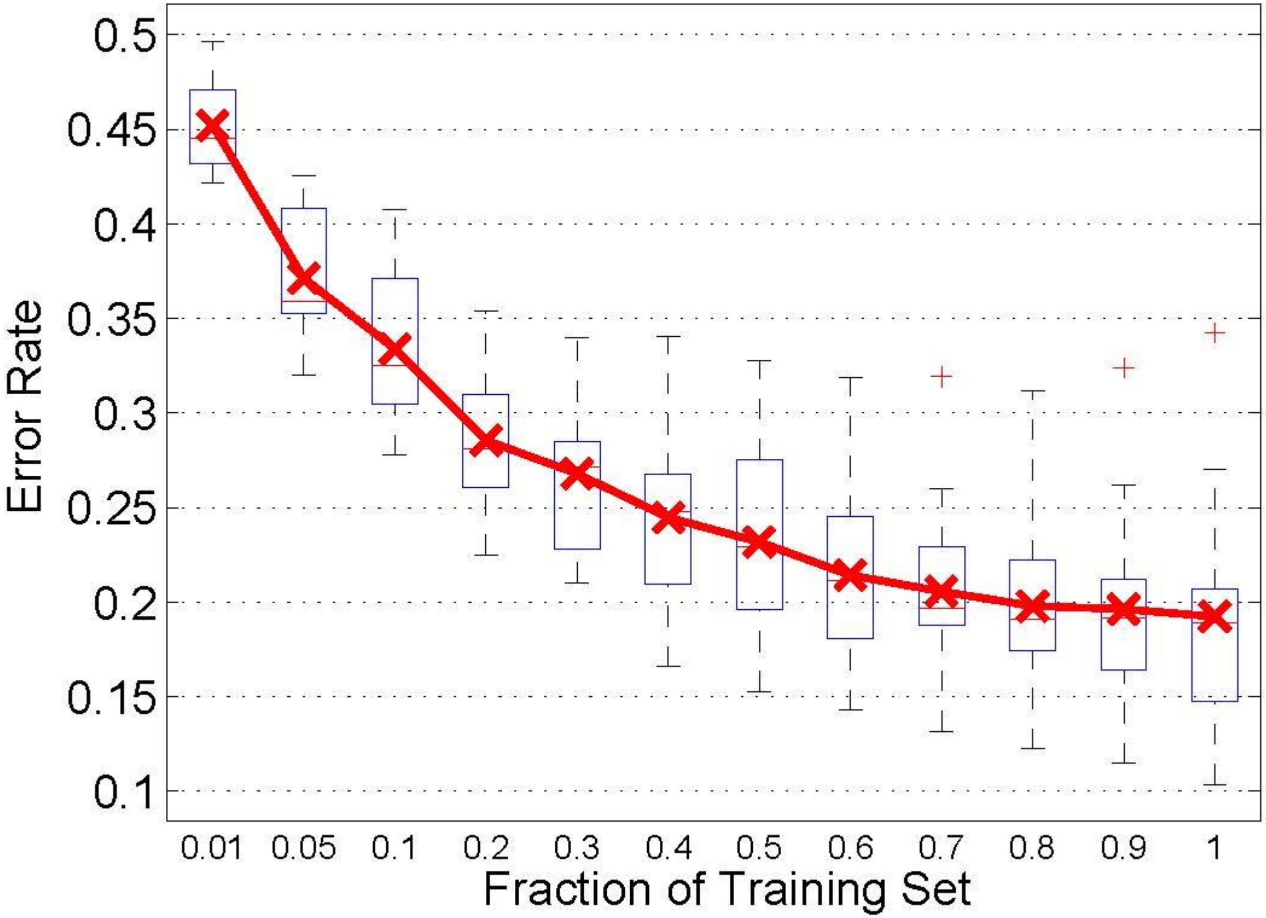
Music Preference

Leave one song out



Music Preference

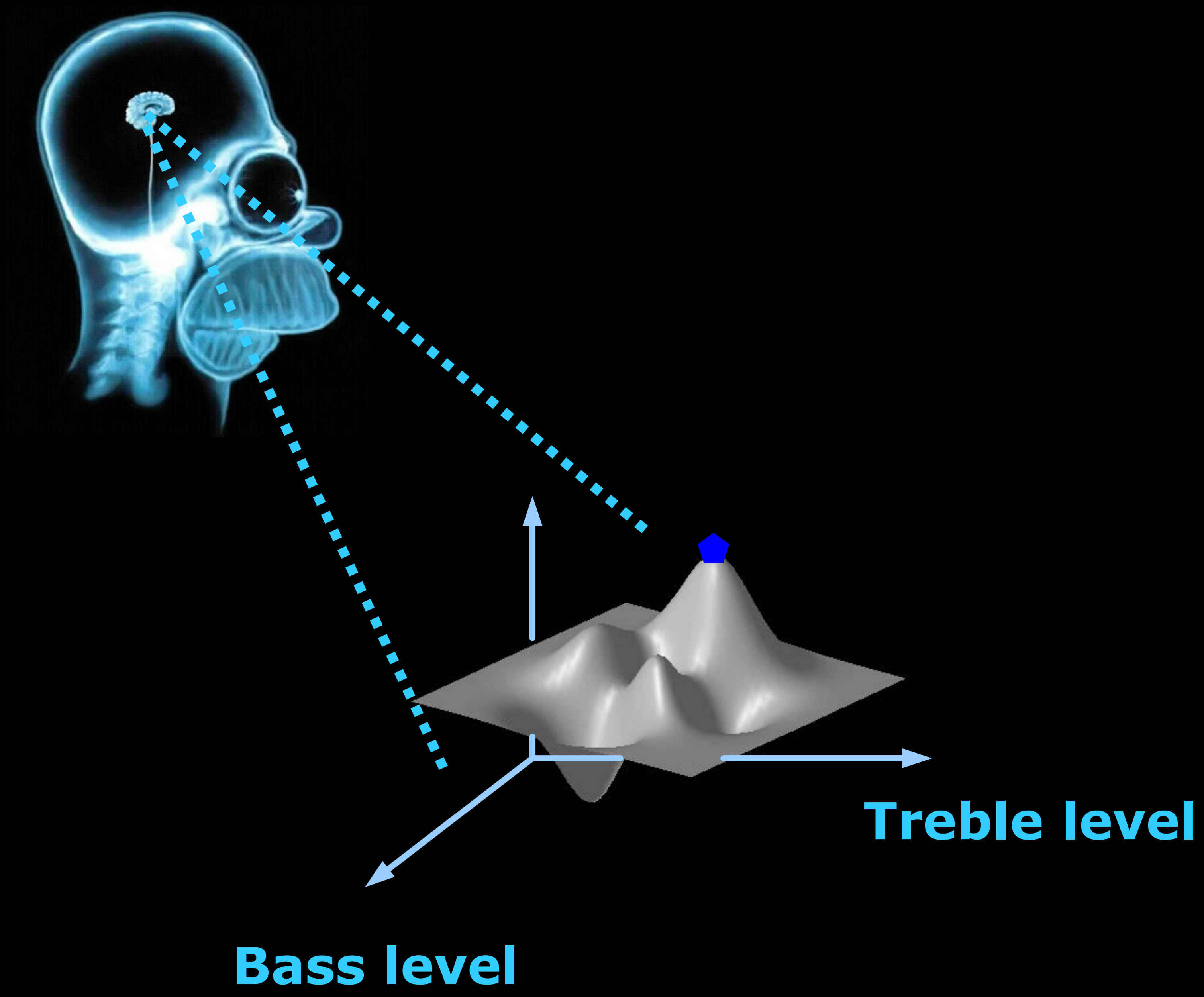
10 fold CV



Personalized Audio Systems – a Bayesian Approach

Jens Brehm Nielsen,
Bjørn Sand Jensen,
Toke Jansen Hansen,
Jan Larsen

AES Convention 135, New York, 17-20 October 2013


Treble level
Bass level

Personalizing an audio system

Learning

(1) A setting is selected in a clever way based on the model of the user's *internal representation*

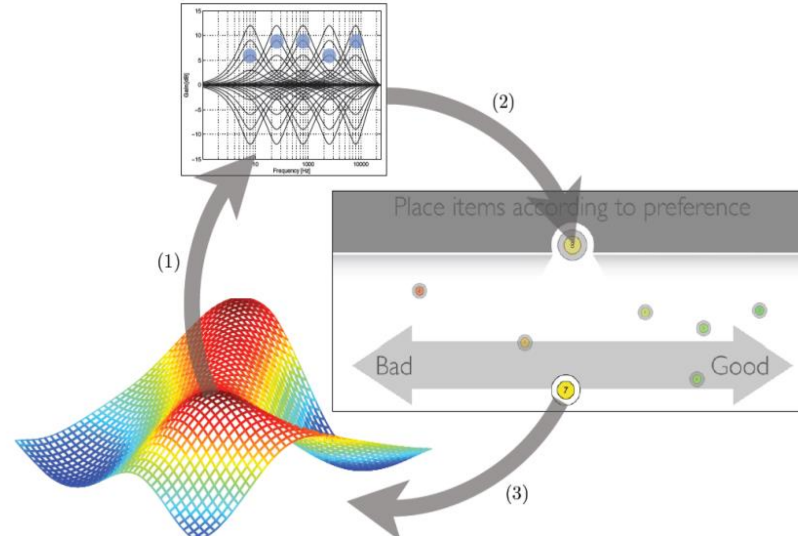
- which is a function, $f(x)$, (modeled by the Gaussian process) over device parameters, x.

DSP

(2) The new setting is *presented* to the user by processing the audio accordingly (standard DSP).

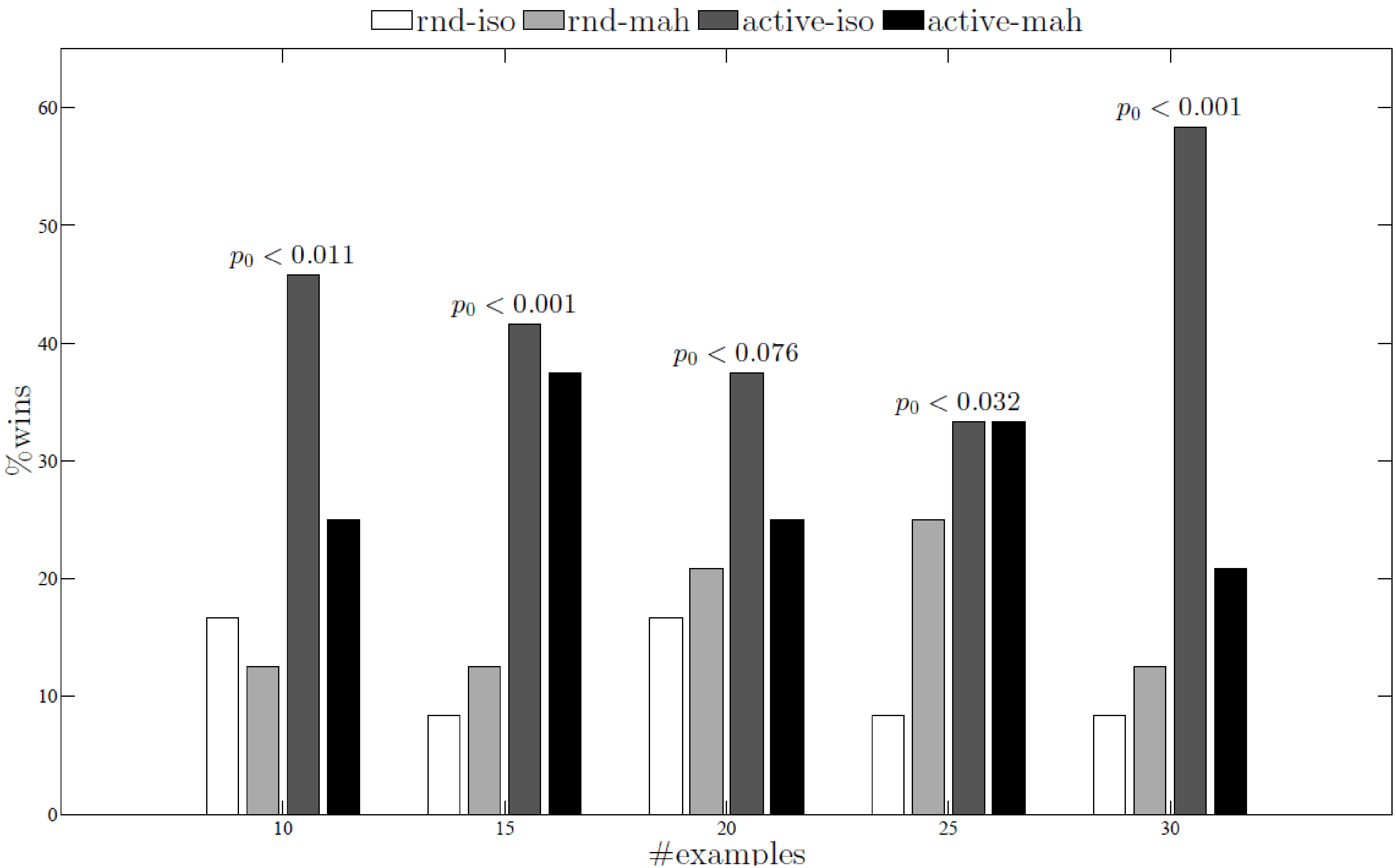
HCI

(3) The users listens to a stimuli and indicates his/her preferences in a simple interfaces with anchors



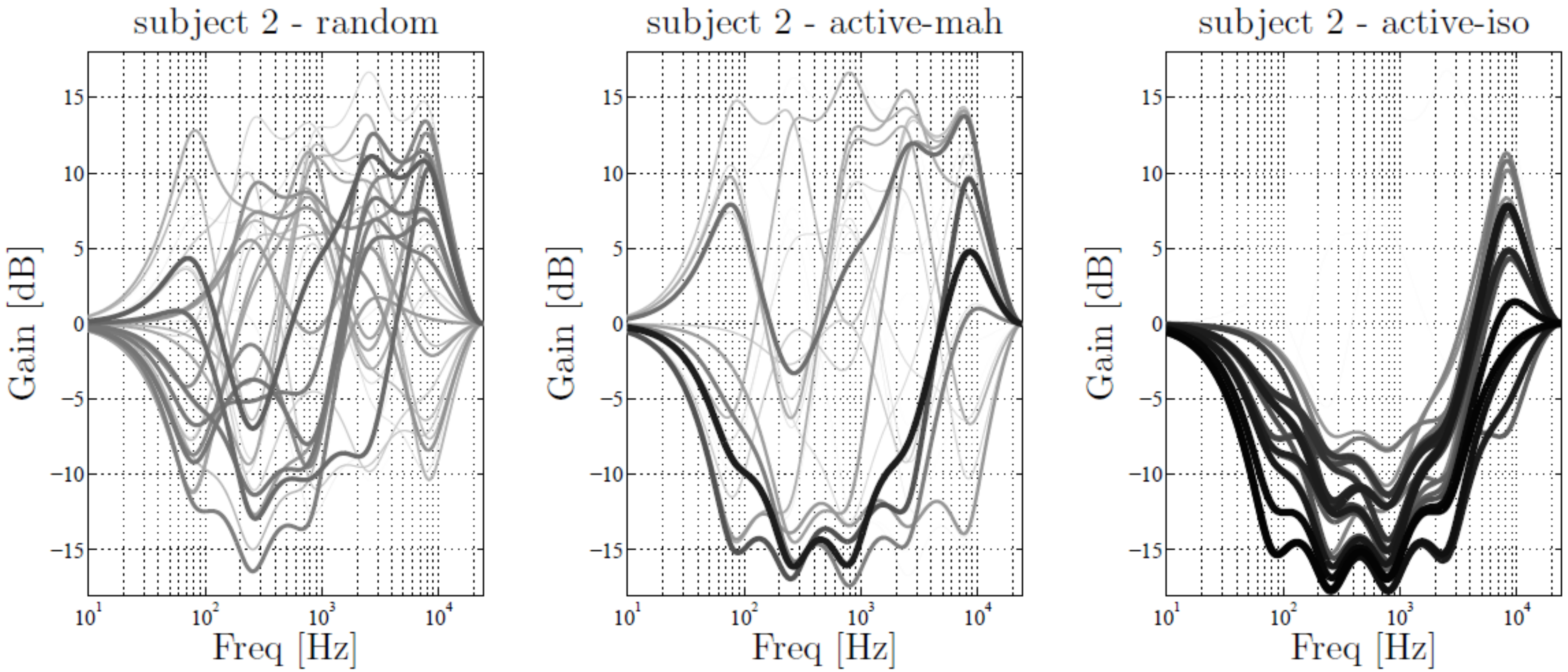
$$\sigma_s|\theta_s \sim \text{half student - t}$$
$$\sigma_\ell|\theta_\ell \sim \text{half student - t}$$
$$f_i|\sigma_s,\sigma_\ell \sim \mathcal{GP}\left(m\left(\mathbf{x}_i\right), \mathrm{k}\left(\mathbf{x}_i,\cdot\right)_{\sigma_s,\sigma_\ell}\right)$$
$$z_i|f_i \sim \mathcal{N}\left(f_i,\sigma_i\right)$$
$$y_i = \Phi^{-1}\left(z_i\right),$$

Results



(a) Learning Curve

Some Results



(b) Ratings