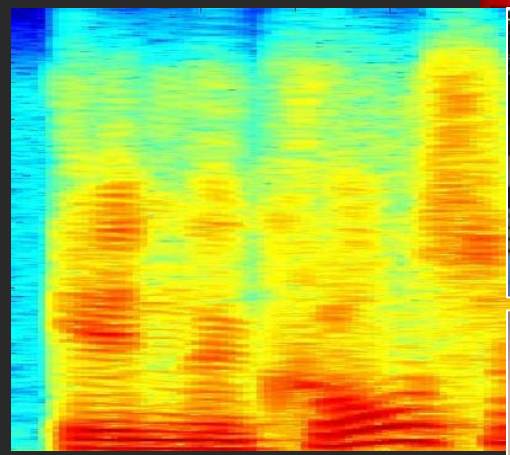
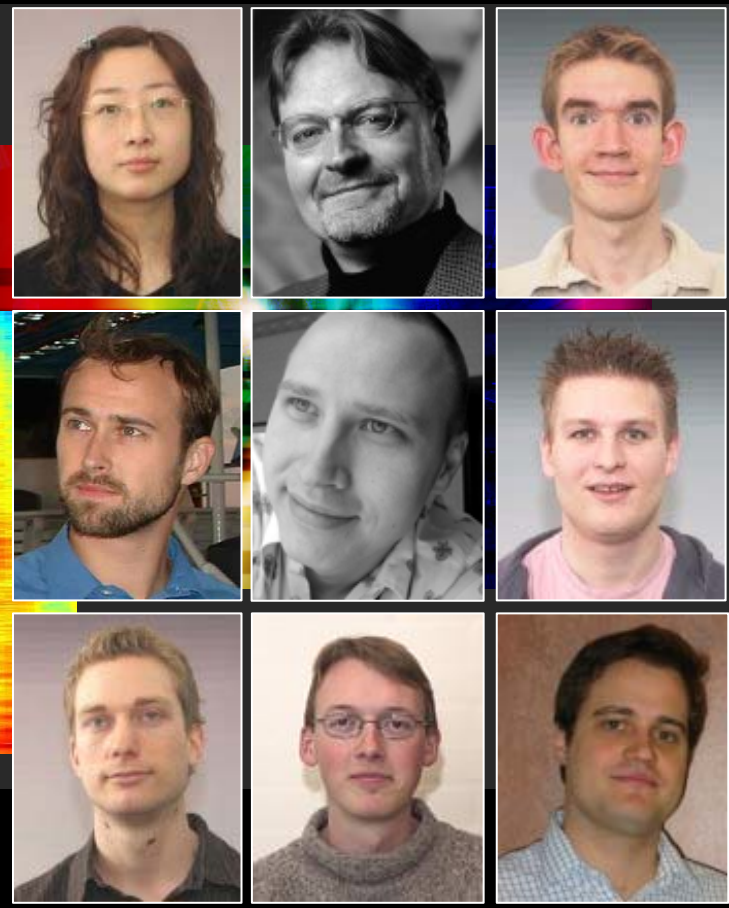
Search for sounds - a machine learning approach





www.intelligentsound.org

The digital music market

- **Wired, April 27, 2005:**

  "With the new Rhapsody, millions of people can now experience and share digital mu
  RealNetworks ch
  once consumers
  many people wil

- **Financial Time**

  LONDON - Visits
  on Christmas Da
  songs on to the

- **Wired, January 17, 2006:**

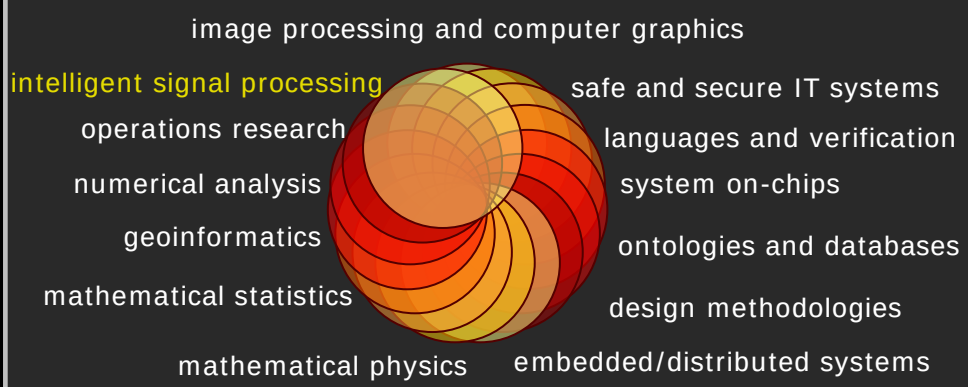
  Google said today it has offered to acquire digital radio advertising provider dMarc Broadcasting for $102 million in cash.

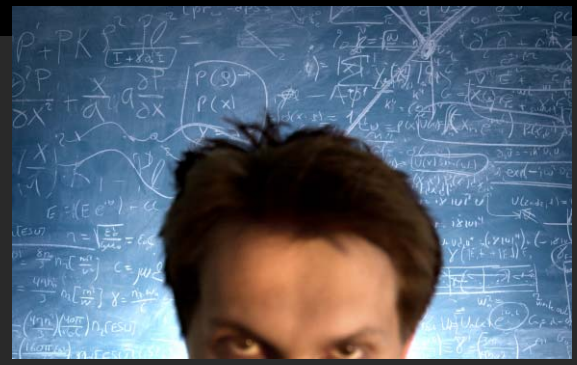
- Huge demand for tools: organization, search, retrieval
- Machine learning will play a key role in future systems

Oultine

- Machine leaning framework for sound search
- Genre classification
- Independent component analysis for music separation

Informatics and Mathematical Modelling, DTU

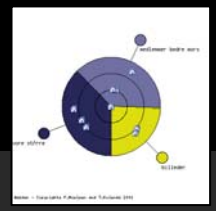




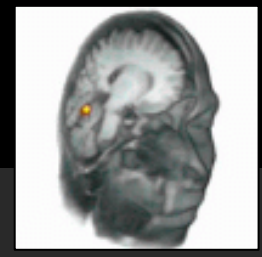
2003 figures

- 84 faculty members
- 28 administrative staff members
- 60 Ph.D. students
- 90 M.Sc. students annually
- 4000 students follow an IMM course annually

ISP Group



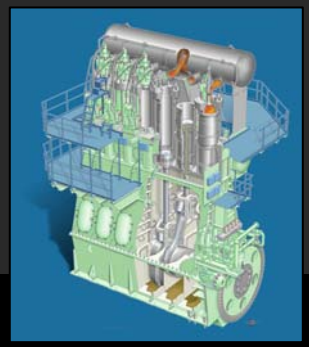
Multimedia



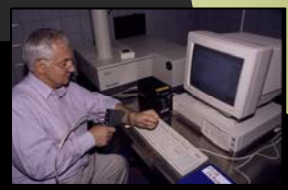
from processing to understanding

extraction of meaningful information by learning

tics



Monitor Systems



Biomedical

faculty

- 6+1 postdocs
- 20 Ph.D. students
- 10 M.Sc. students

Machine learning in sound information processing

audio data

Meta data
ID3 tags
context

ISO/IEC JTC1/SC29 WG11 MPEG MOVING PICTURE EXPERTS GROUP

Audioscrobbler.
The Social Music Technology Playground
Home · About · Data Access

User networks
co-play data
playlist
communities
user groups

MusicBrainz

machine learning model

Tasks
Grouping
Classification
Mapping to a structure
Prediction e.g. answer to query

Aspects of search

Specificity

- standard search engines
- indexing of deep content
- Objective: high retrieval performance

Similarity

- more like this
- similarity metrics
- Objective: high generalization and user acceptance

Specialized search and music organization

allmusic

Explore by Genre, mood, theme, country, instrument

Query by humming

IDMT

Fraunhofer Institut Digitale Medientechnologie

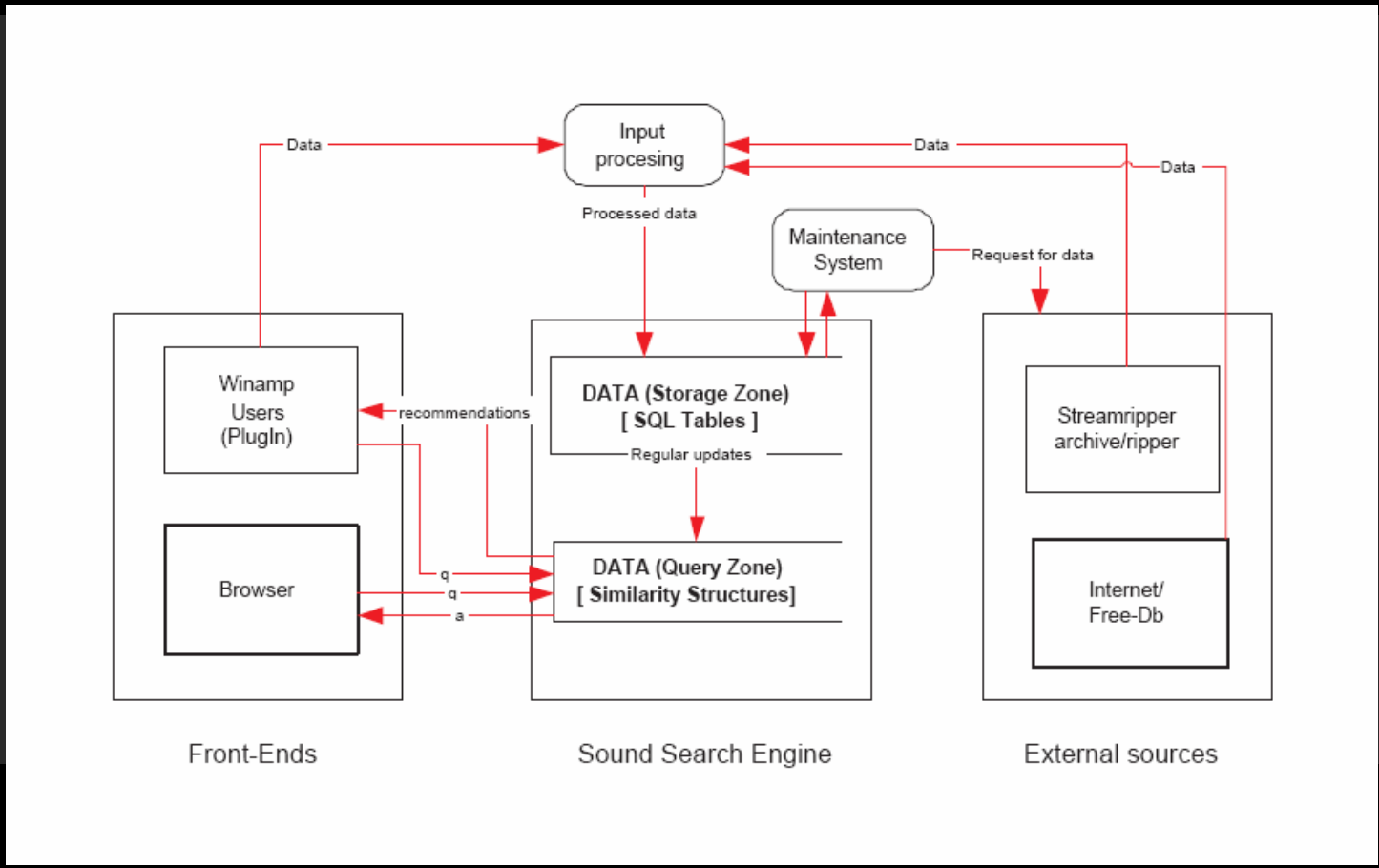
The NGSW is creating an online fully-searchable digital library of spoken word collections spanning the 20th century

Organize songs according to tempo, genre, mood

PANDORA™

search for related songs using the “genes of music”

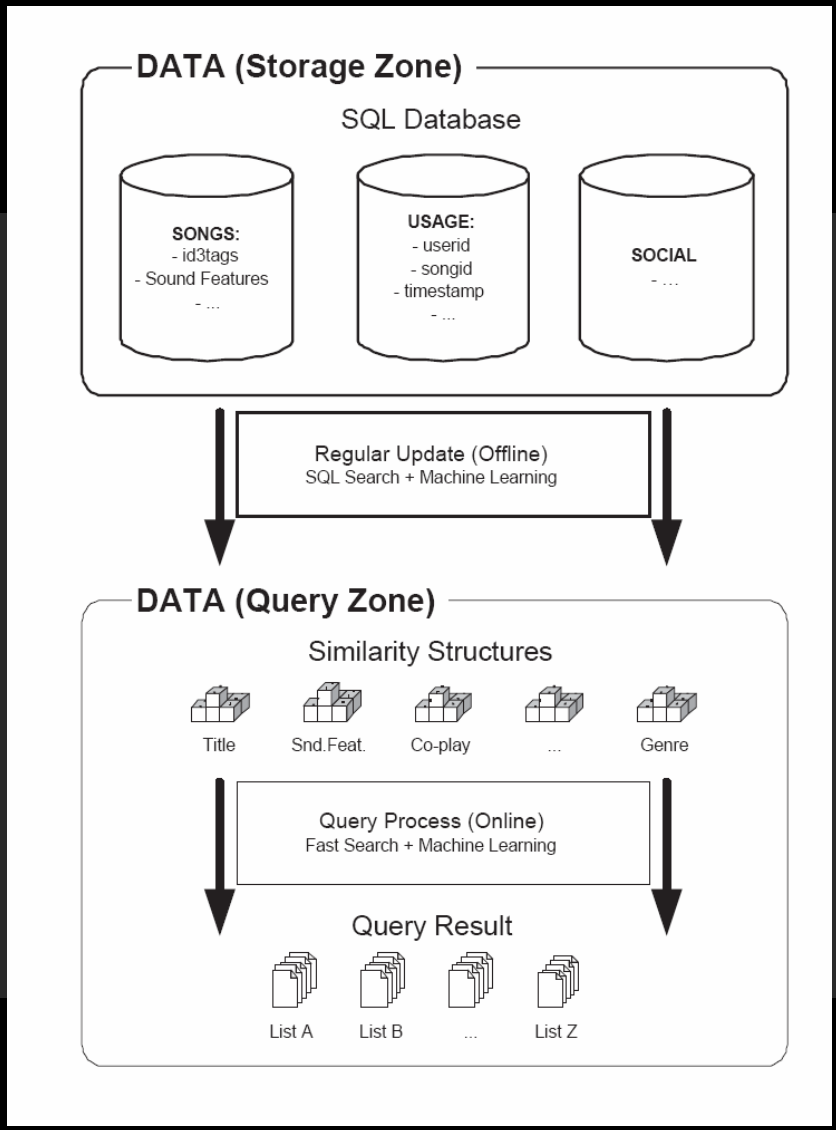
System overview




SoundSearch 0.1
Introduction
Financial Times (ft.com) 12:46 p.m. ET Dec. 28, 2005:
"LONDON - Visits to music downloading Web sites saw a 50 percent rise on Christmas Day as hundreds of thousands of people began loading songs on to the iPods they received as presents."
SoundSearch 0.1 combines co-play patterns, expert evaluations and music features to help you retrieve the music you like.
Use these music features to organize your search:
Co-play
Beat
Expert
Sound
Start the Music:
Now Playing
This field displays information about the artist currently playing. The information is retrieved from *text mining* of public domain internet sites.

WINAMP demo June 2006

Storage and query



Similarity structures

- Low le
  - Ad h
    RCC
- High le
  - Basi
    Kme
    MoH
- Metric
  - Eucli
    eartl
    Bour

- loudness
- zero-crossing energy
- log-energy
- down sampling
- autocorrelation
- peak detection,
- delta-log-loudness

- pitch
- brightness
- bandwidth
- harmonicity
- spectrum power
- subband power
- centroid
- roll-off
- low-pass filtering
- spectral flatness
- spectral tilt
- shaprness
- roughness

,

ine,
om

Predicting the answer from query

$$p(s_a | s_q, u)$$

- s_a: index for answer song
- s_q: index for query song
- u : user (group index)
- c_i : hidden cluster index of similarity i

Intelligent Sound Project
IMM (DTU) – CS, CT (AaU)

- Signal processing
- Databases
- Machine learning

Demo: Sound search engine

Demo: Matlab toolbox

Phd projects

Group publications

Joint publications

Workshops/ Phd-courses

Research ”tasks”

AaU Communication Technology:

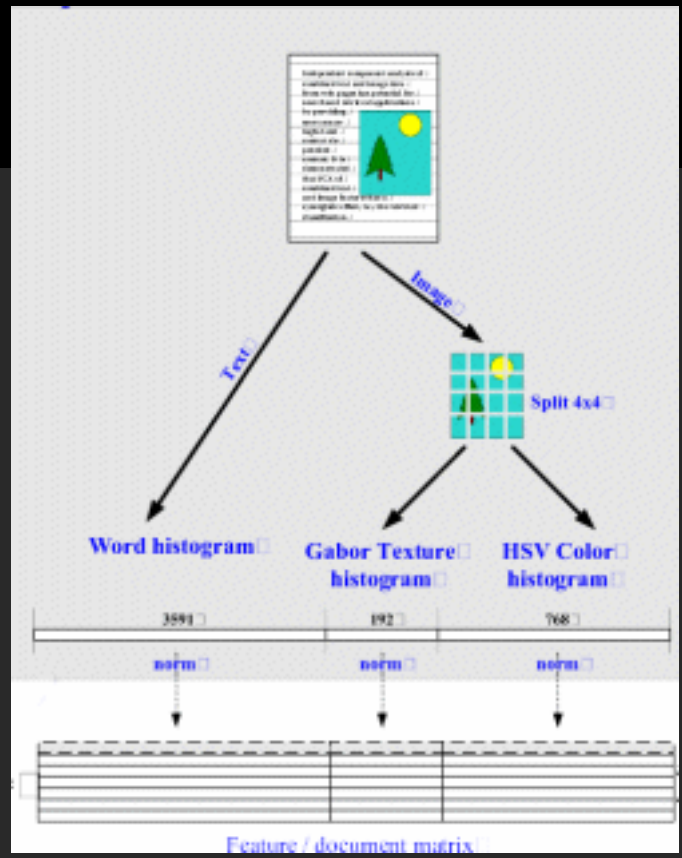
TASK i): Features for sound based context modelling - MPEG and beyond
TASK ii): Signal separation in noisy environments: ICA and noise reduction

AaU Computer Science/Database Management:

TASK iii): Multidimensional management of sound as context
TASK iv): Advanced Query Processing for Sound Feature Streams

DTU IMM-ISP

TASK v): Context detection in sound streams
TASK vi): Webmining for sound



ISOUND PUBLICATIONS 2005-2006

•L. Feng, L. K. Hansen, *On low level cognitive components of speech*, International Conference on Computational Intelligence for Modelling (CIMCA'05), 2005

•A. B. Nielsen, L. K. Hansen, U. Kjems, *Pitch Based Sound Classification*, Informatics and Mathematical Modelling, Technical University of Denmark, DTU, 2005

•L. K. Hansen, P. Ahrendt, J. Larsen, *Towards Cognitive Component Analysis*, AKRR'05 - International and Interdisciplinary Conference on Adaptive Knowledge Representation and Reasoning, Pattern Recognition Society of Finland, Finnish Artificial Intelligence Society, Finnish Cognitive Linguistics Society, 2005

•A. Meng, P. Ahrendt, J. Larsen, *Improving Music Genre Classification by Short-Time Feature Integration*, IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. V, pp. 497-500, 2005

•L. Feng, L. K. Hansen, *PHONEMES AS SHORT TIME COGNITIVE COMPONENTS*, International Conference on Acoustics, Speech and Signal Processing (ICASSP'06), 2005

•M. S. Pedersen, T. Lehn-Schiøler J. Larsen, *BLUES from Music: BLind Underdetermined Extraction of Sources from Music*, ICA2006, 2006

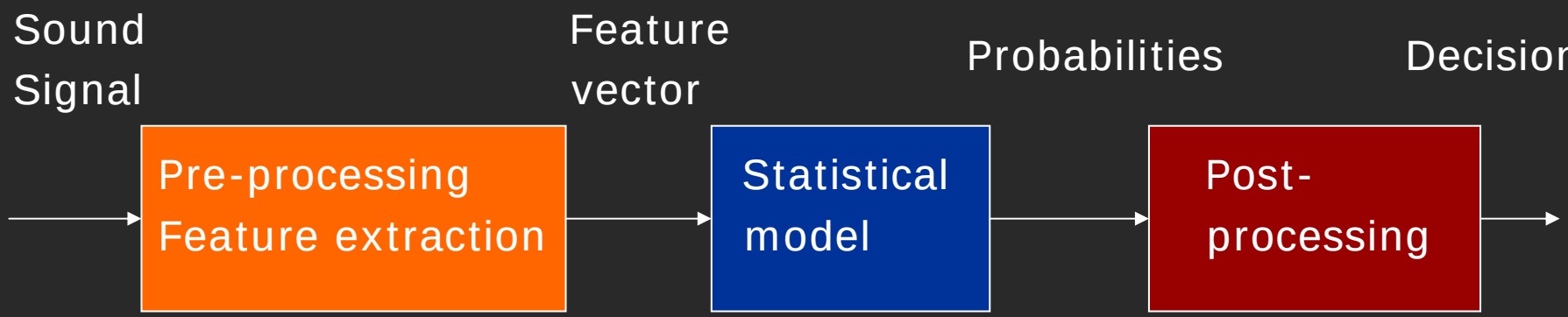
•M. N. Schmidt, M. Mørup *Nonnegative Matrix Factor 2-D Deconvolution for Blind Single Channel Source Separation*, ICA2006, 2006

Genre classification

- Prototypical example of predicting meta data
- The problem of interpretation of genres
- Can be used for other applications e.g. hearing aids
- Models

Model

- Making the computer classify a sound piece into musical genres such as jazz, techno and blues.



How do humans do?

- Sounds – loudness, pitch, duration and timbre
- Music – mixed streams of sounds
- Recognizing musical genre
  - physical and perceptual: instrument recognition, rhythm, roughness, vocal sound and content
  - cultural effects

How well do humans do?

- Data set with 11 genres
- 25 people assessing 33 random 30s clips

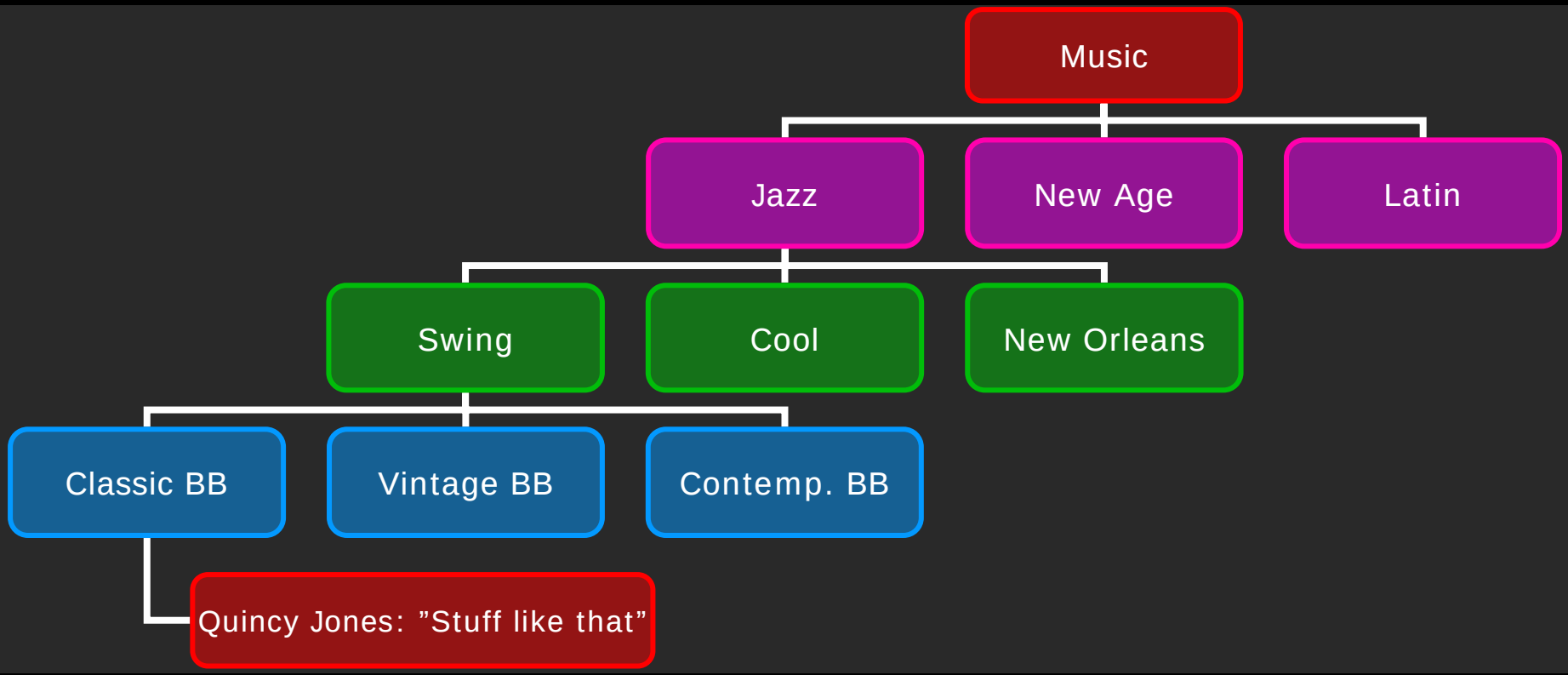
accuracy
54 - 61 %

Baseline: 9.1%

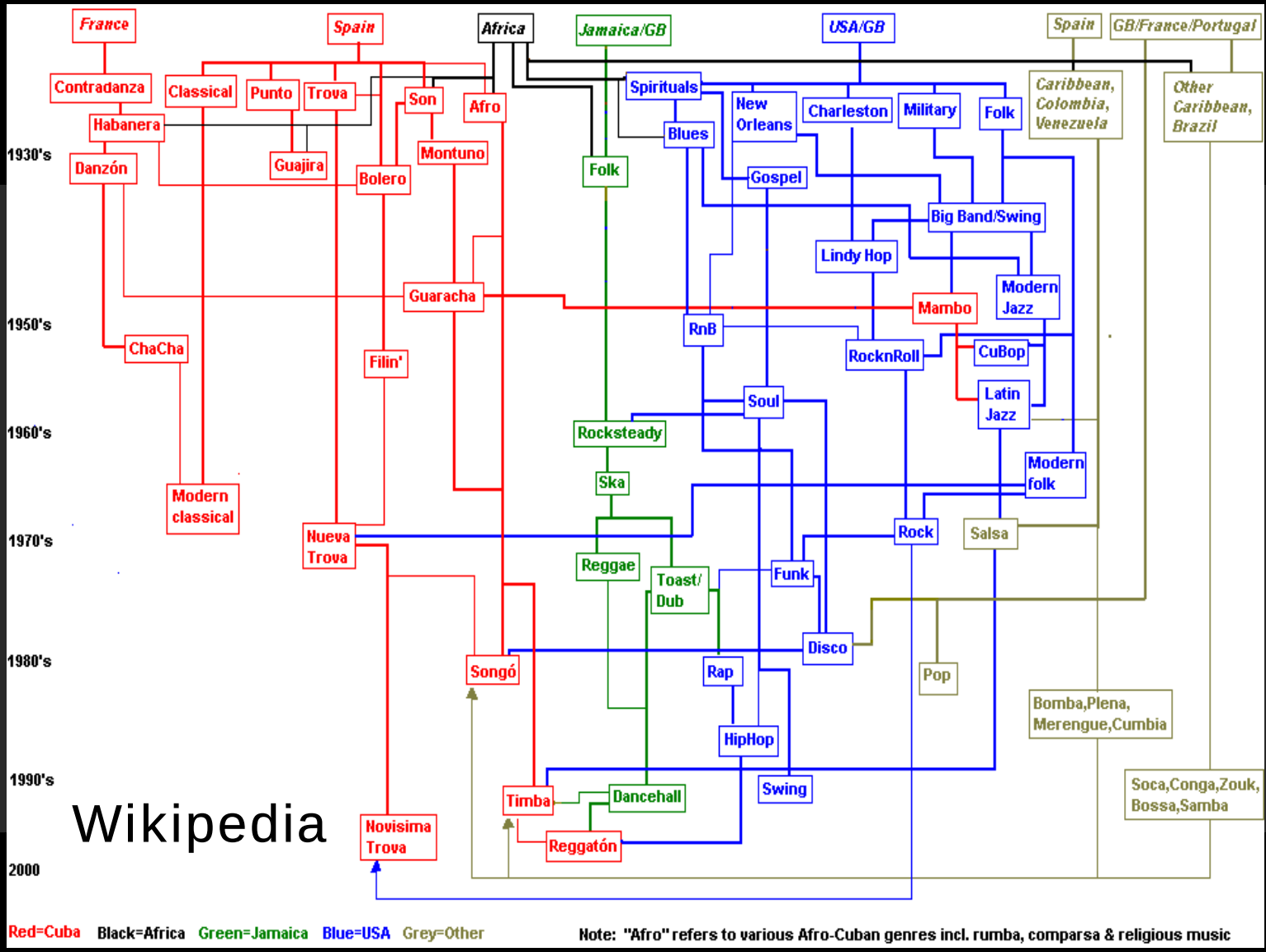
What's the problem ?

- Technical problem: Hierarchical, multi-labels
- Real problems: Musical genre is not an intrinsic property of music
  - A subjective measure
  - Historical and sociological context is important
  - No Ground-Truth

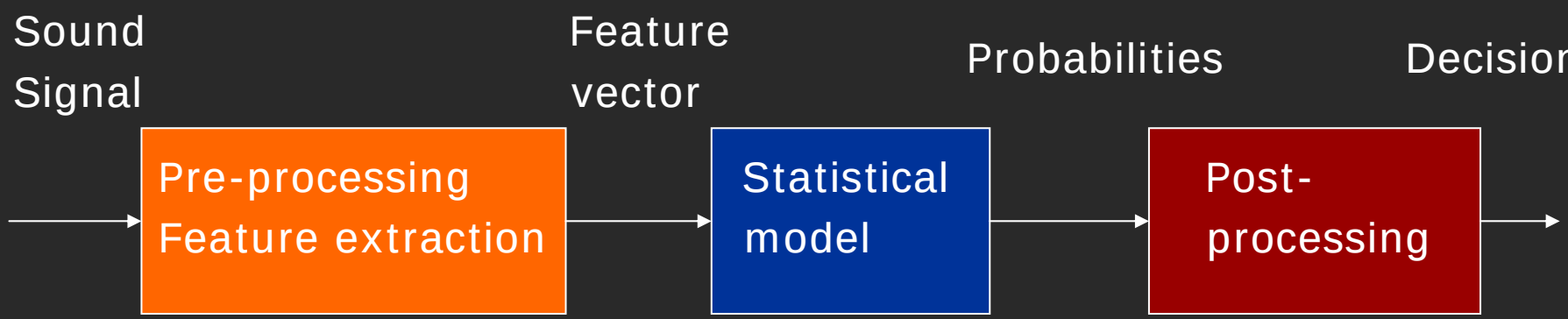
Music genres form a hierarchy



(according to Amazon.com)


France
Spain
Africa
Jamaica/GB
USA/GB
Spain
GB/France/Portugal
Contradanza
Classical
Punto
Trova
Son
Afro
Spirituals
New Orleans
Charleston
Military
Folk
Caribbean, Colombia, Venezuela
Other Caribbean, Brazil
Habanera
Blues
1930's
Danzón
Guajira
Bolero
Montuno
Folk
Gospel
Big Band/Swing
Lindy Hop
Modern Jazz
Guaracha
Mambo
1950's
RnB
ChaCha
Filin'
RocknRoll
CuBop
Latin Jazz
Soul
1960's
Rocksteady
Modern folk
Modern classical
Ska
Nueva Trova
Rock
Salsa
1970's
Reggae
Toast/ Dub
Funk
Disco
1980's
Songó
Rap
Pop
Bomba,Plena, Merengue,Cumbia
HipHop
1990's
Swing
Soca,Conga,Zouk, Bossa,Samba
Timba
Dancehall
Reggatón
Novisima Trova
2000
Wikipedia
Red=Cuba
Black=Africa
Green=Jamaica
Blue=USA
Grey=Other
Note: "Afro" refers to various Afro-Cuban genres incl. rumba, comparsa & religious music

Music Genre Classification Systems



Features

- Short time features (10-30 ms)
  - MFCC and LPC
  - Zero-Crossing Rate (ZCR), Short-time Energy (STE)
  - MPEG-7 Features (Spread, Centroid and Flatness Measure)
- Medium time features (around 1000 ms)
  - Mean and Variance of short-time features
  - Multivariate Autoregressive features (DAR and MAR)
- Long time features (several seconds)
  - Beat Histogram

Features for genre classification

30s sound clip from the center of the song

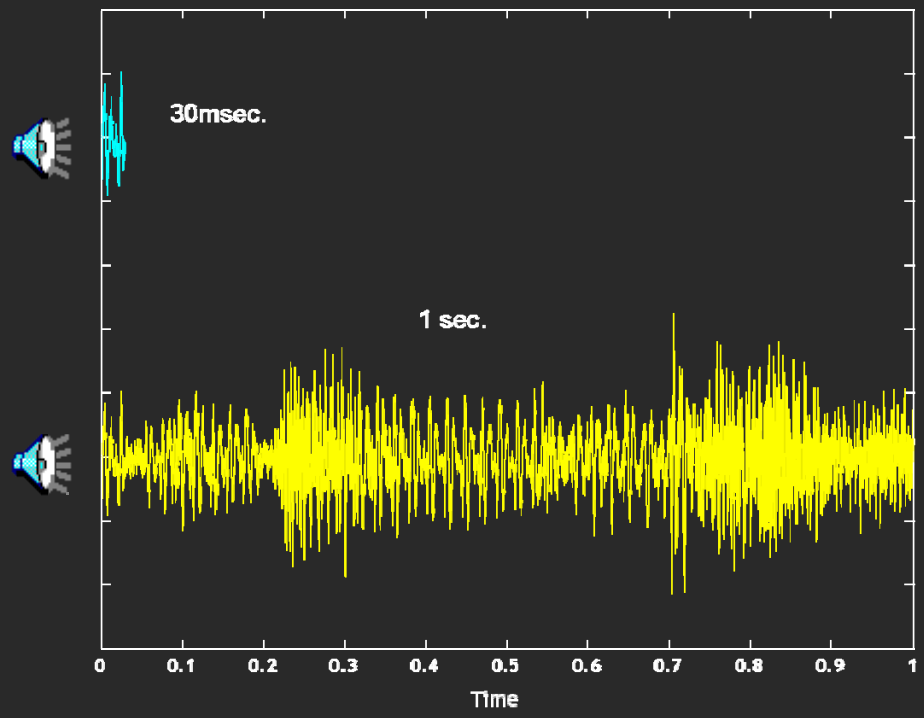
6 MFCCs, 30ms frame

6 MFCCs, 30ms frame

6 MFCCs, 30ms frame

3 ARCs per MFCC, 760ms frame

30-dimensional AR features, $x_r, r=1,..,80$


30msec.
1 sec.
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
Time

Statistical models

- Desired: $p(c|s)$ (class c and song s)
- Used models :
  - Intregration of MFCCs
  - Linear and non-linear neural networks
  - Gaussian classifier
  - Gaussian Mixture Model
  - Co-occurrence models

Best results

- 5-class problem (with little class overlap) : 2% error
  - Comparable to human classification on this database
- Amazon.com 6-class problem (some overlap) : 30% error
- 11-class problem (some overlap) : 50% error
  - human error about 43%

The Clever Jukebox

The Art of Automated Genre Classification

examples:







Theory:

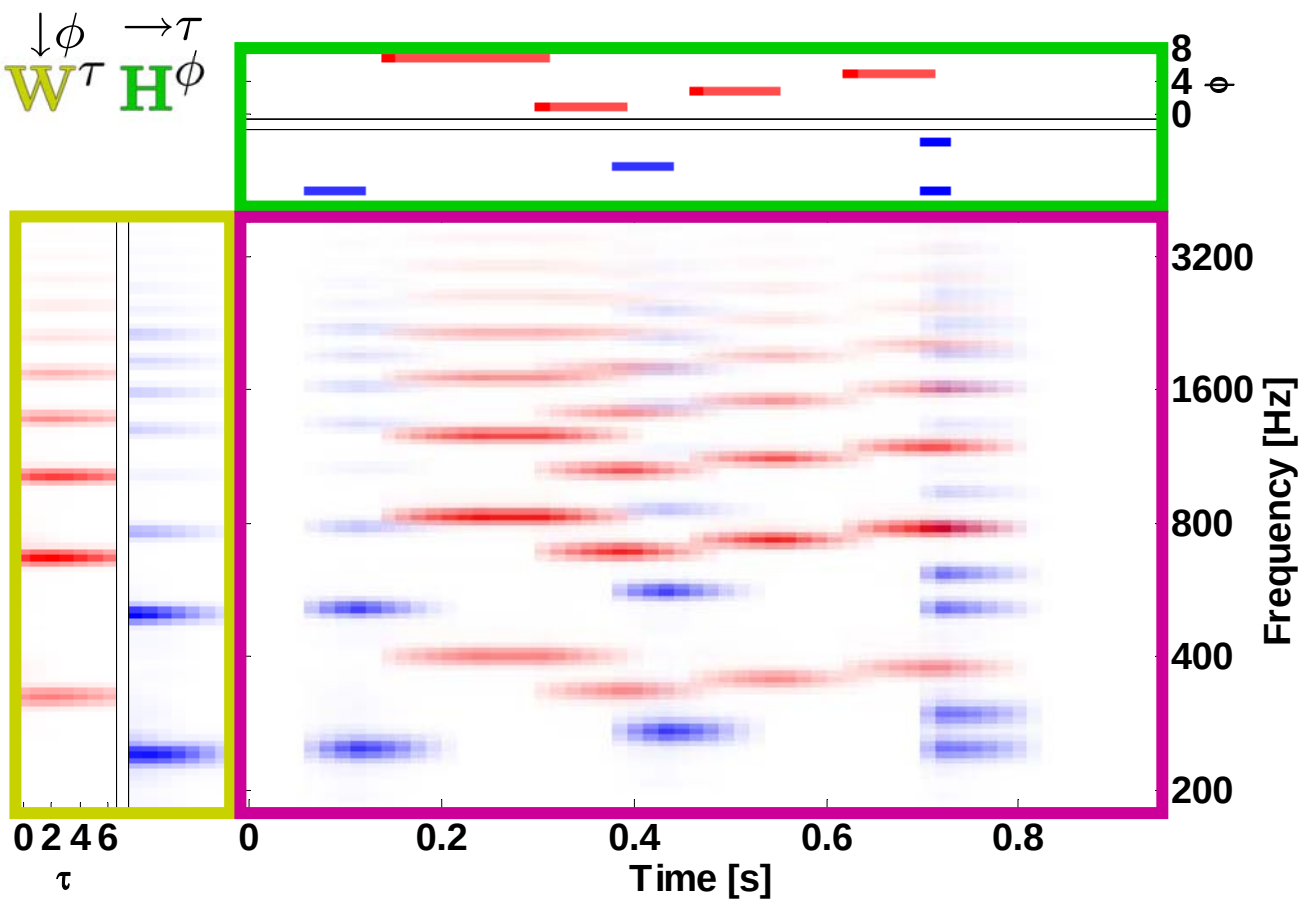
Automatic musical genre classification can be defined as the science (or art) of finding computer algorithms that take a (digitized) sound clip as input and yield a musical genre as output. The goal of automated genre classification is, of course, that the musical genre which is output should agree with the human classification of the sound into genre.

This demo illustrates an approach to the problem that first extract frequency-based sound features followed by a "linear regression" classifier. The basic features are the so-called mel-frequency cepstral coefficients (MFCCs), which are extracted on a time-scale of 30 msec. From these MFCC features, auto-regressive coefficients (ARs) are extracted along with the mean and gain to get a single (30 dimensional) feature vector on the time-scale of 1 second. These features have been used because they have performed well in a previous study (Meng, Ahrendt, Larsen (2005)). Linear regression (or single-layer linear NN) is subsequently used for classification. This classifier is rather simple; current research investigates more advanced methods of classification.

Research: Peter Ahrendt. Design: Sune Lehmann.

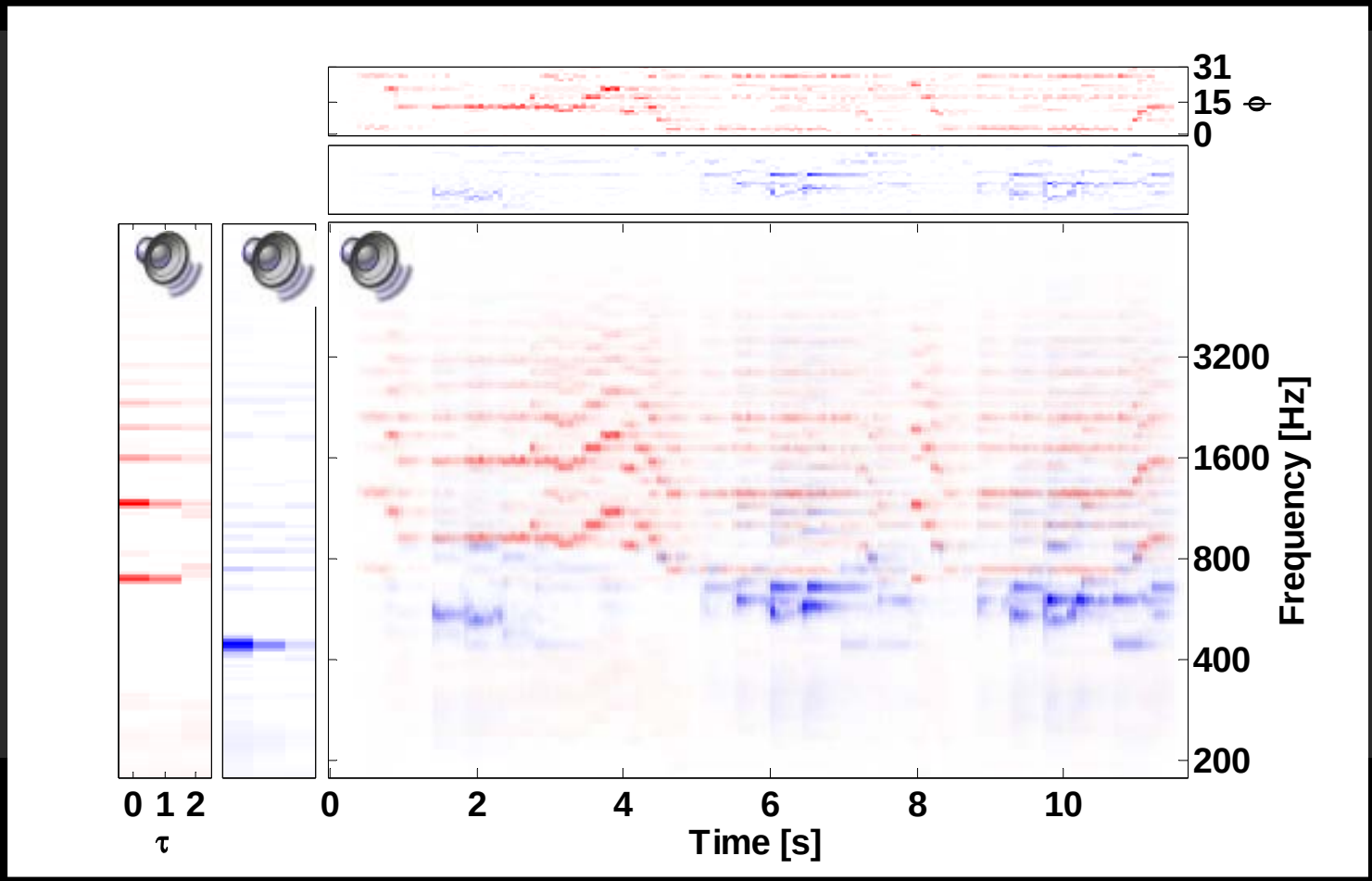
Nonnegative matrix factor 2D deconvolution

$$\mathbf{V} \approx \mathbf{\Lambda} = \sum_{\tau,\phi} \overset{\downarrow\phi}{\mathbf{W}^{\tau}} \overset{\rightarrow\tau}{\mathbf{H}^{\phi}}$$

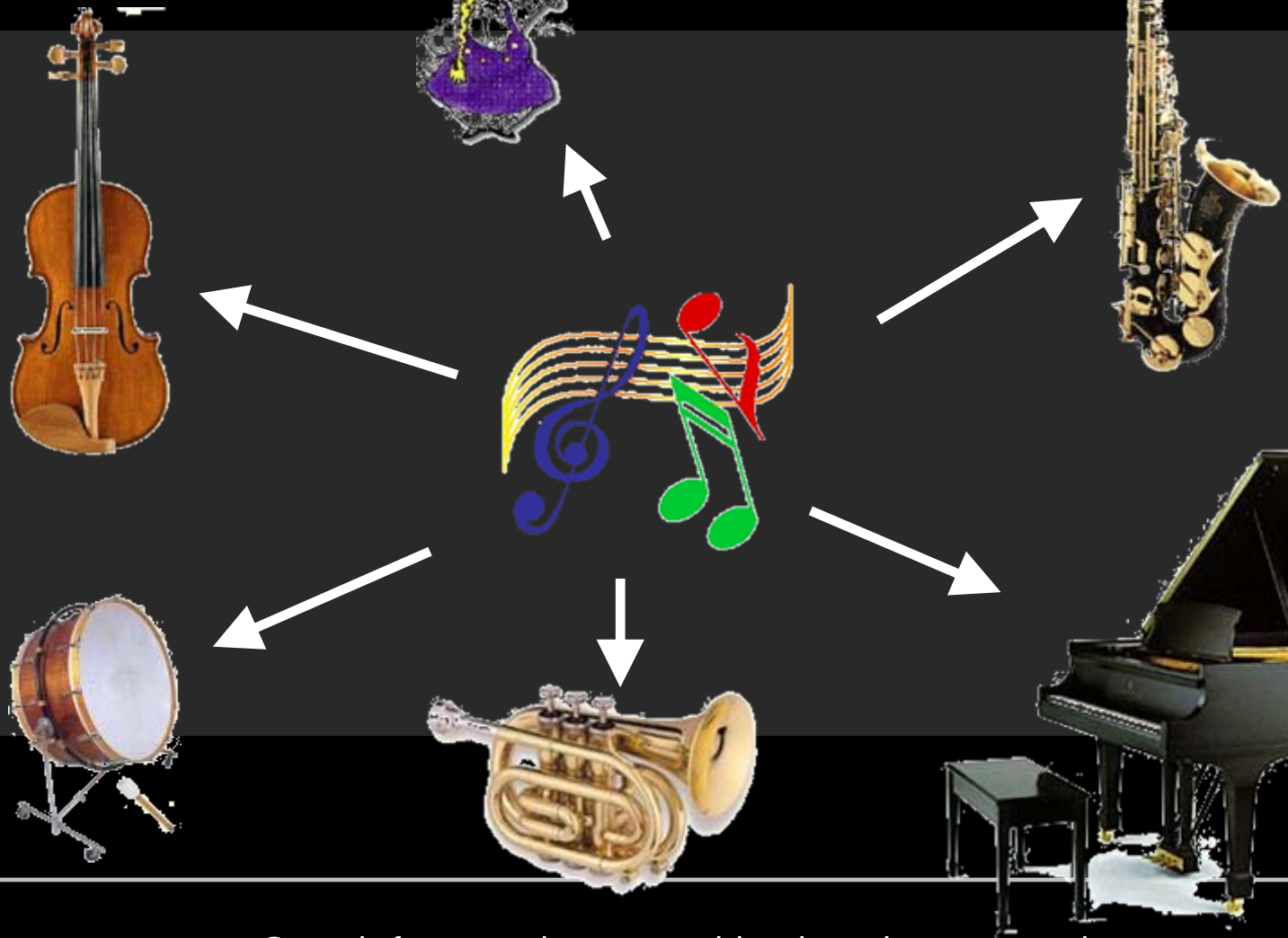


Ref: Mikkel Schmidt and Morten Mørup, ICA2006

Demonstration of the 2D convolutive NMF model

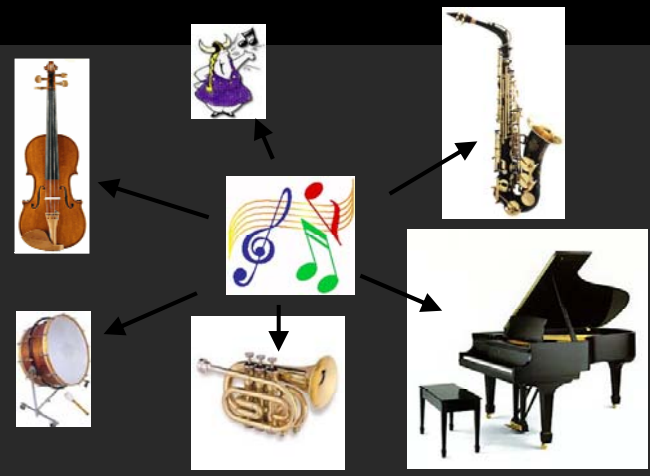


Separating music into basic components



Motivation: Why separating music?

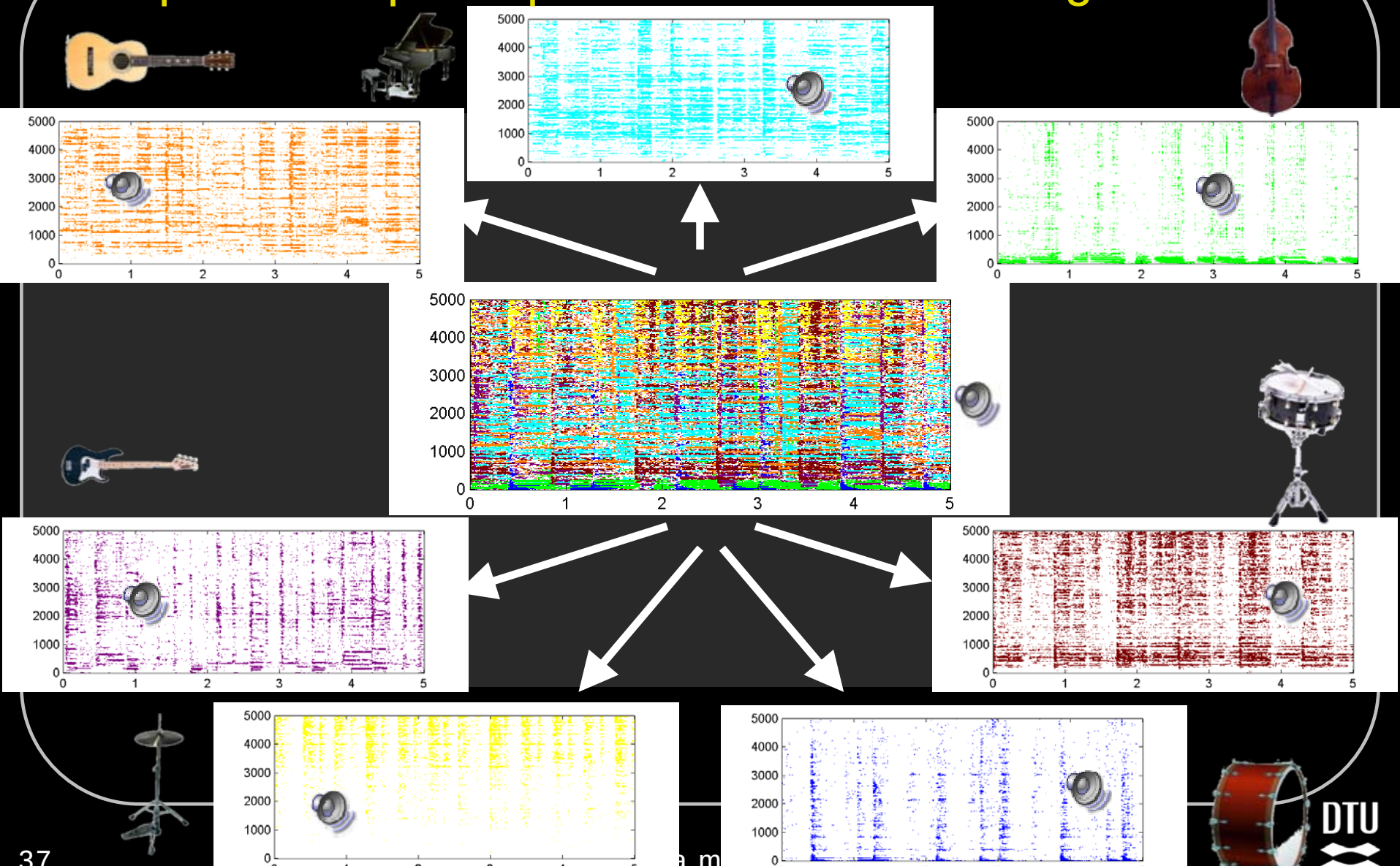
- Music Transcription
- Identifying instruments
- Identify vocalist
- Front end to search engine



Assumptions

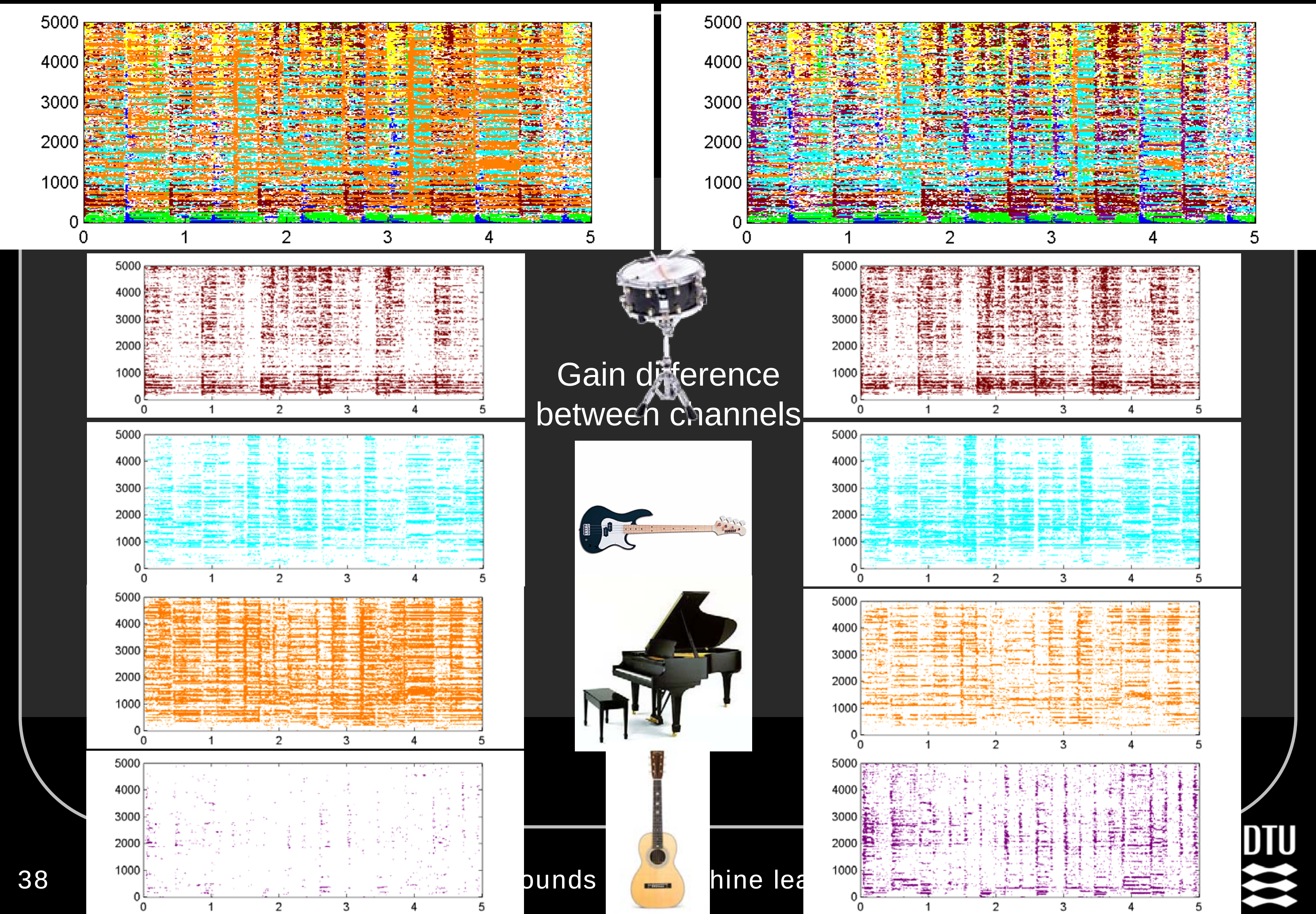
- Stereo recording of the music piece is available.
- The instruments are separated to some extent in time and in frequency, i.e. the instruments are sparse in the time-frequency (T-F) domain.
- The different instruments originate from spatially different directions.

Separation principle: ideal T-F masking



Stereo channel 1

Stereo channel 2



Separation principle 2: ICA

What happens if a 2-by-2 separation matrix **W** is applied to a 2-by-N mixing system?

ICA on stereo signals

- We assume that the mixture can be modeled as an instantaneous mixture, i.e.

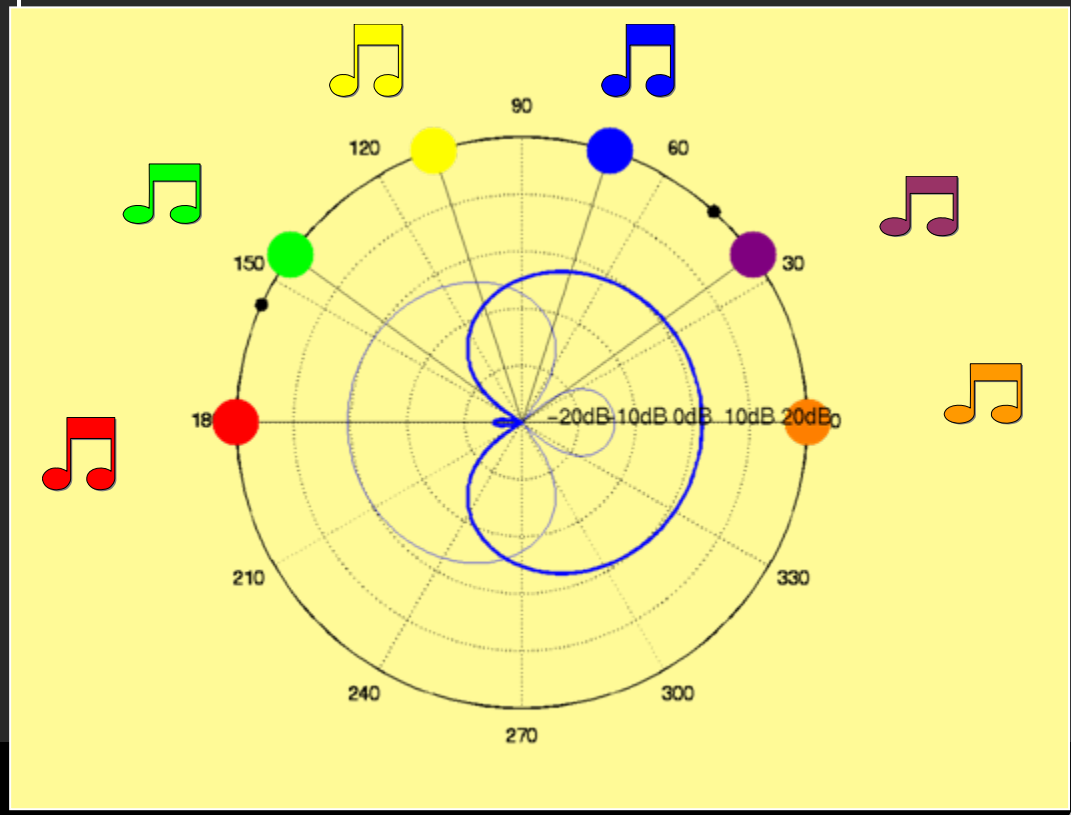
$$x = A(\theta_1, \ \ldots \ , \theta_N)s \qquad A(\theta) = \begin{bmatrix} r_1(\theta_1) & \cdots & r_1(\theta_N) \\ r_2(\theta_1) & \cdots & r_2(\theta_N) \end{bmatrix}$$

- The ratio between the gains in each column in the mixing matrix corresponds to a certain direction.

Direction dependent gain

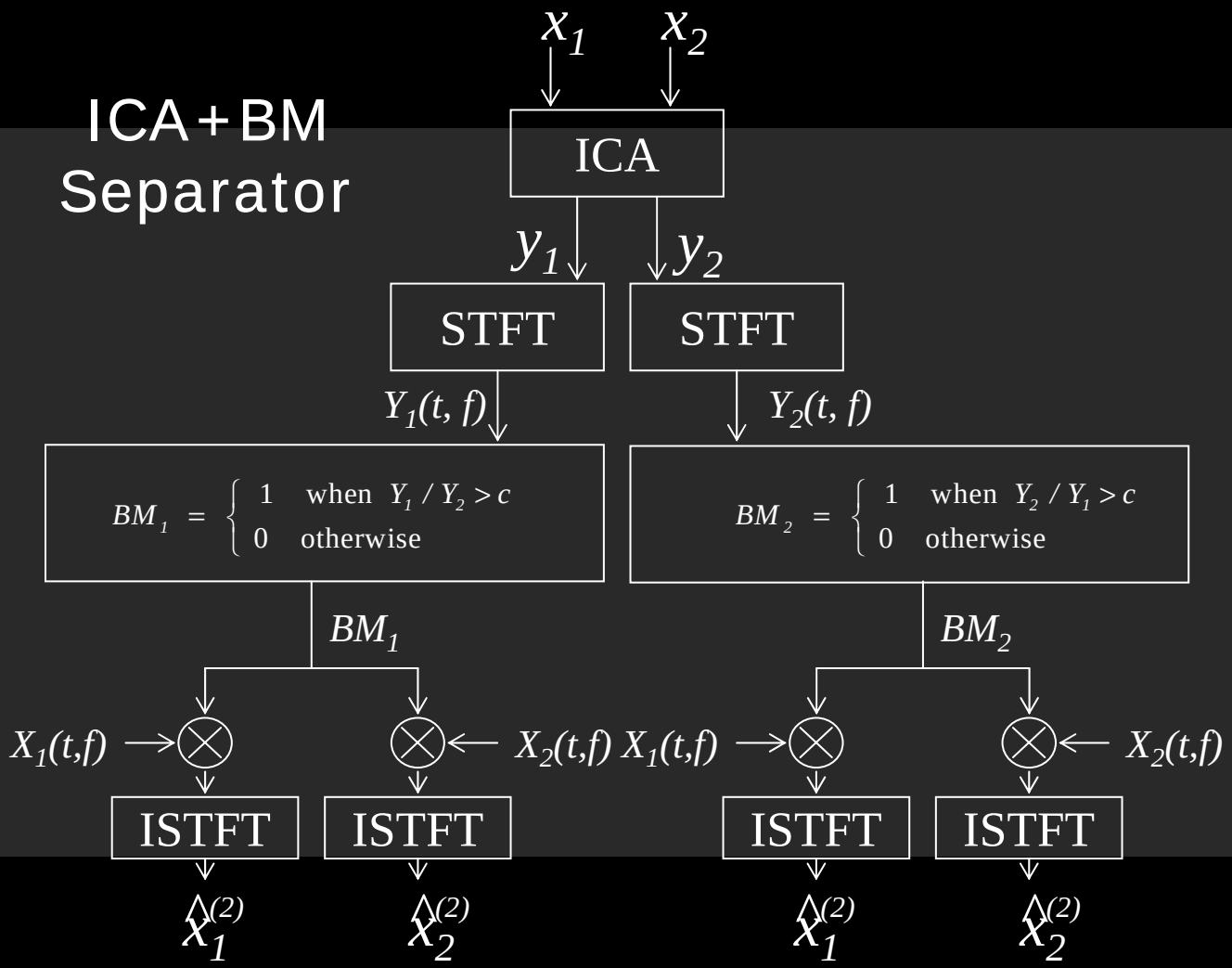
$$\mathbf{r}(\boldsymbol{\theta}) = 20\log|\mathbf{WA}(\boldsymbol{\theta})|$$

When **W** is applied, the two separated channels each contain a *group* of sources, which is as independent as possible from the other channel.

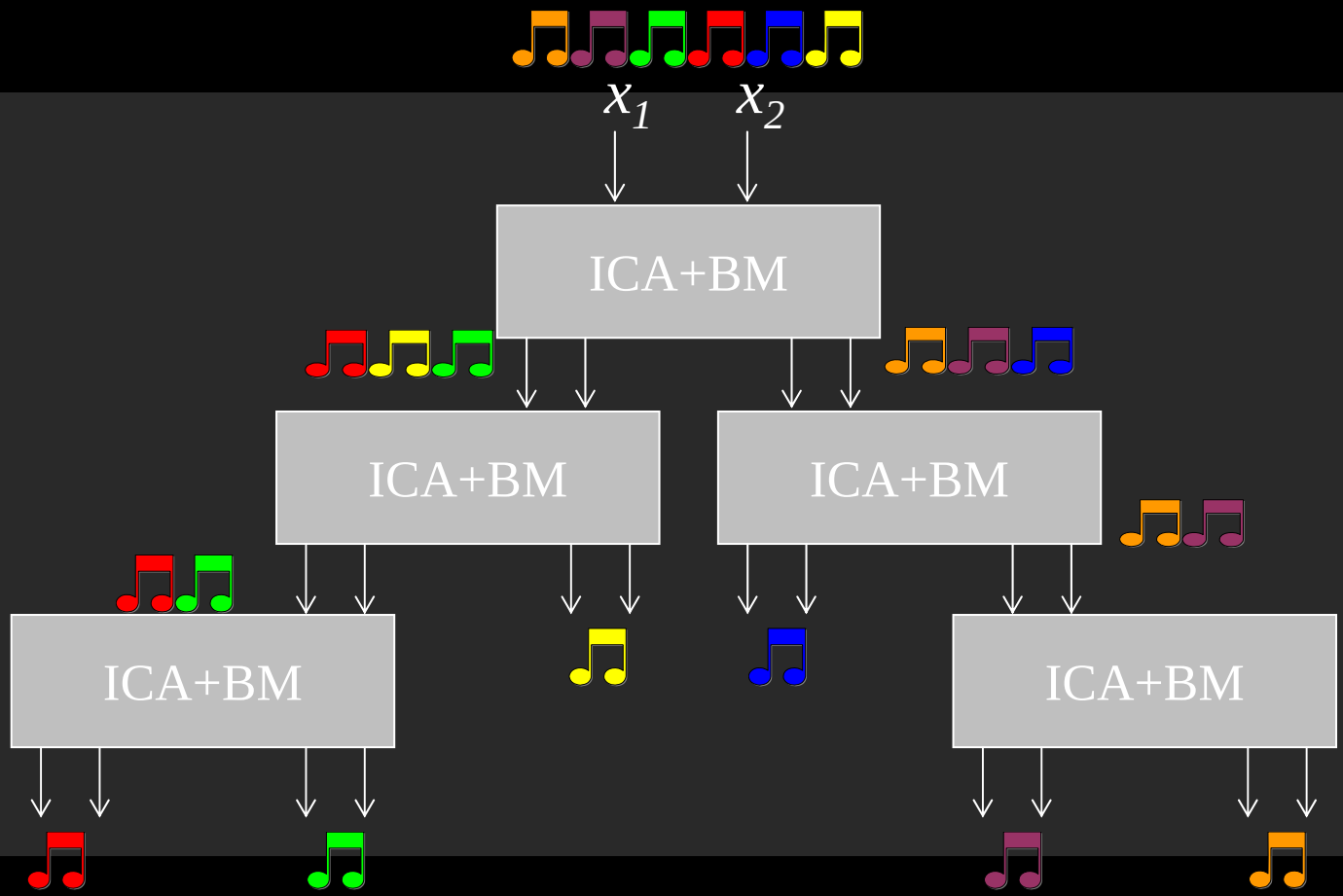


Combining ICA and T-F masking

ICA+BM Separator

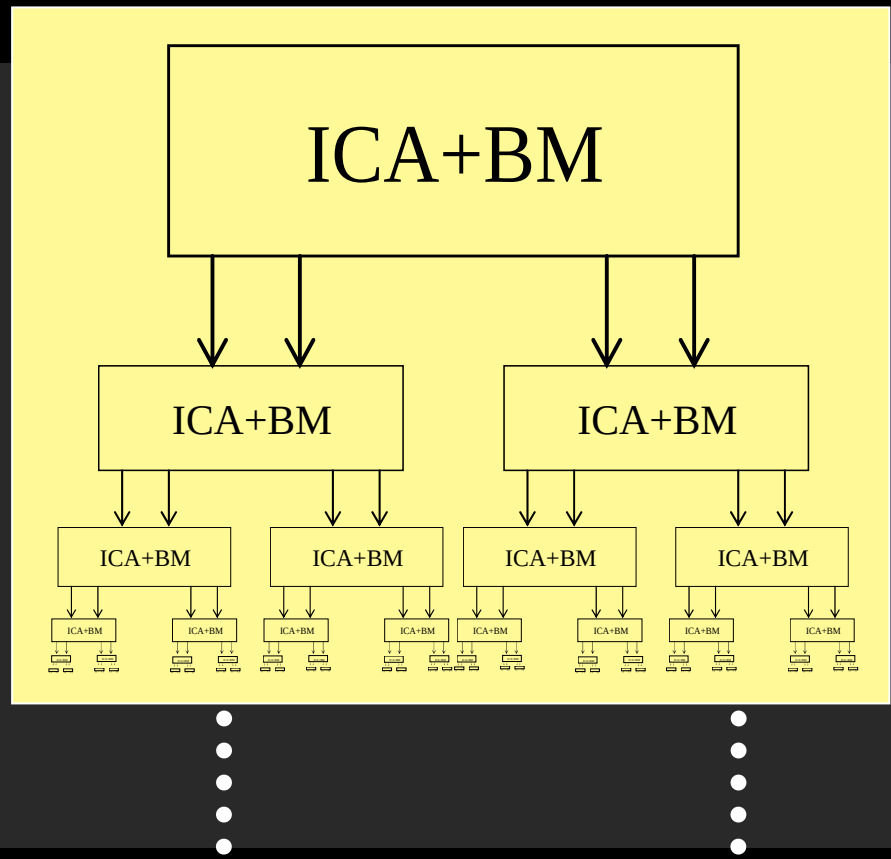


Method applied iteratively

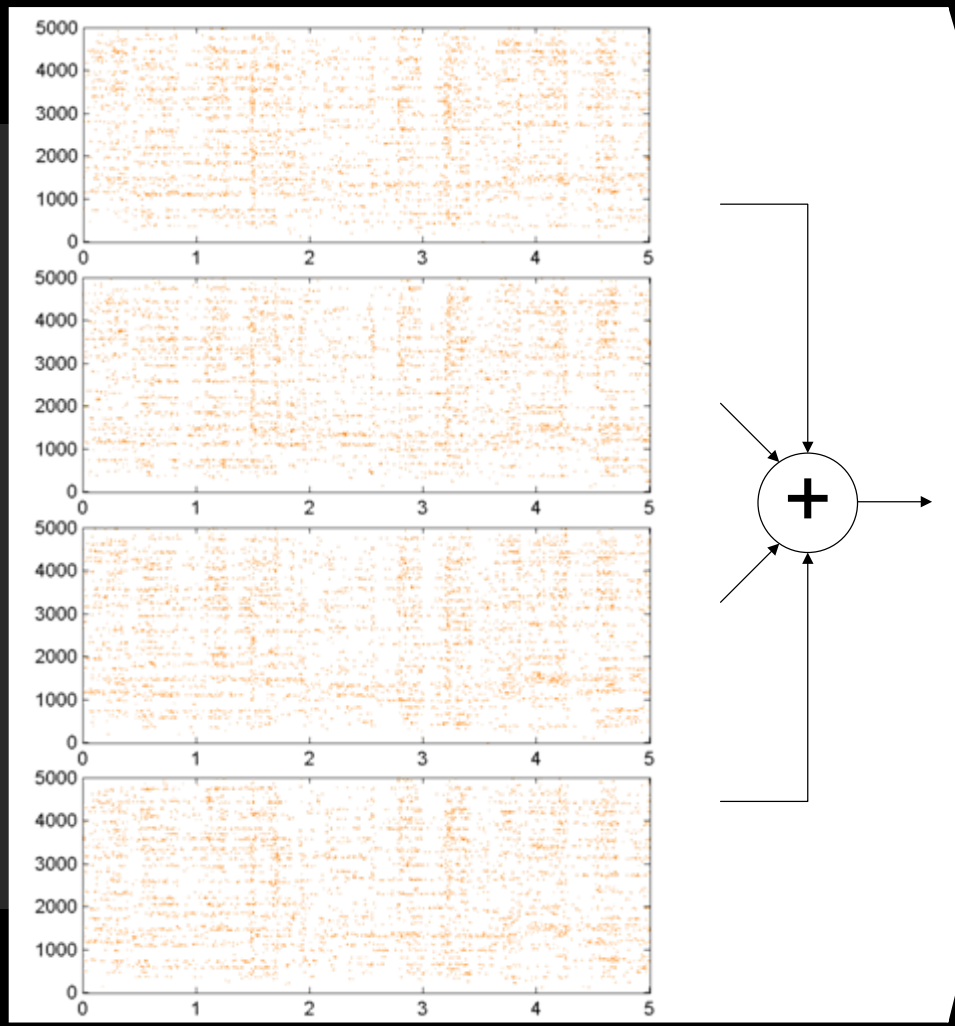


Improved method

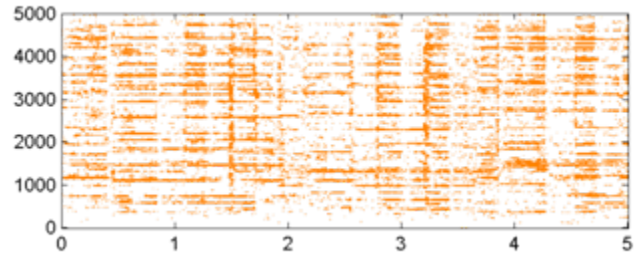
- The assumption of instantaneous mixing may not always hold.
- Assumption can be relaxed.
- Separation procedure is continued until very sparse masks are obtained.
- Masks that mainly contain the same source are afterwards merged.



Mask merging



If the signals in the time domain are correlated, their corresponding masks are merged.



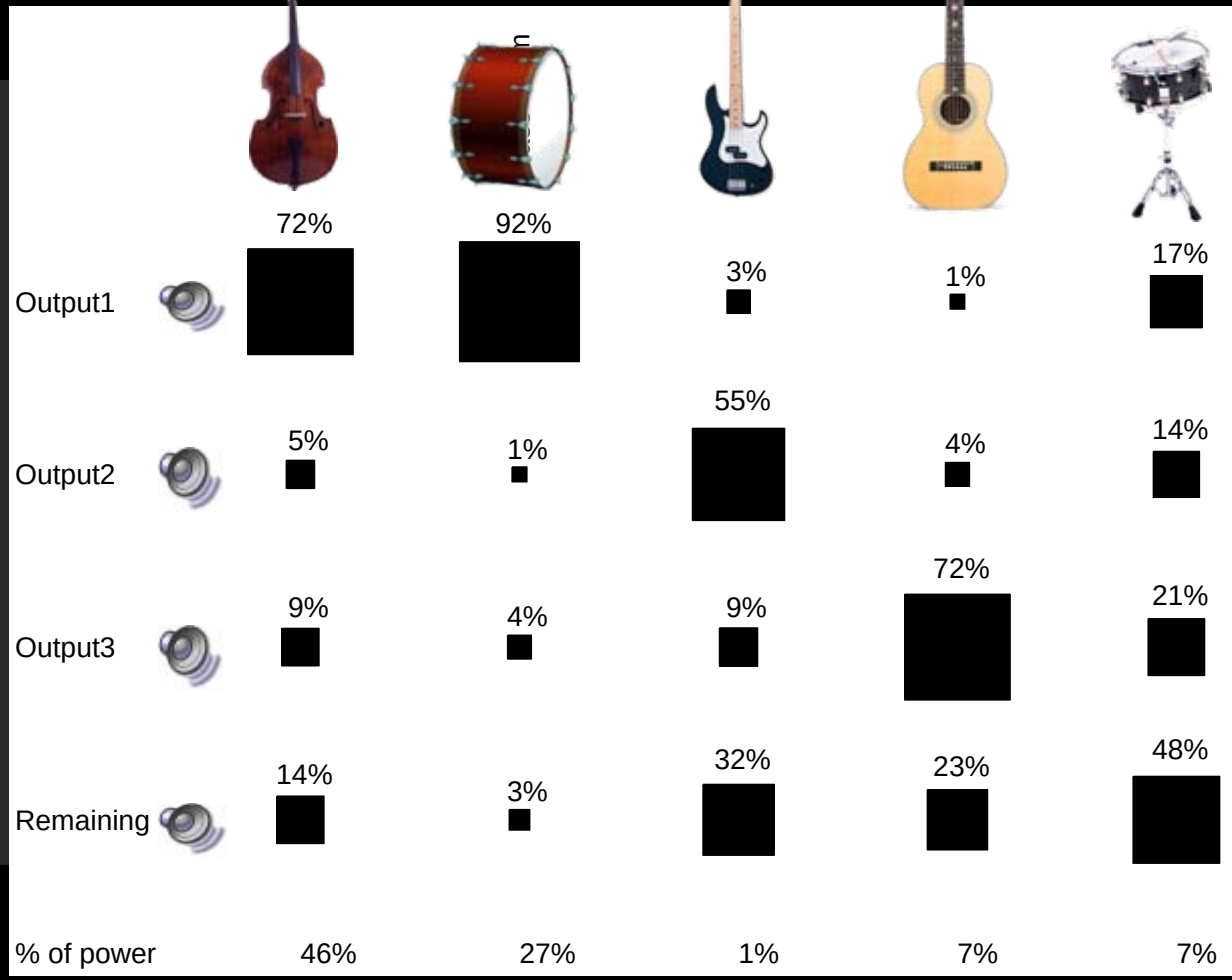
The resulting signal from the merged mask is of higher quality.

Results

- Evaluation on real stereo music recordings, with the stereo recording of each instrument available, before mixing.
- We find the correlation between the obtained sources and the by the ideal binary mask obtained sources.
- Other segregated music examples are available online.

Results

- The segregated outputs are dominated by individual instruments
- Some instruments cannot be segregated by this method, because they are not spatially different.



Conclusion on ICA separation

- We have presented an unsupervised method for segregation of single instruments or vocal sound from stereo music.
- Our method is based on combining ICA and T-F masking.
- The segregated signals are maintained in stereo.
- Only spatially different signals can be segregated from each other.
- The proposed framework may be improved by combining the method with single channel separation methods.

Conclusions

- Search is a “productivity engine” simply important to ….quality of life…
- Generic and specialized search engines: different criteria and challenges
- Machine learning is essential for search!
- Music search based on musical features, meta data, and social network information