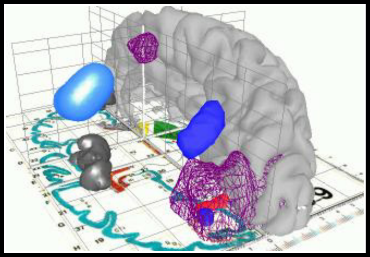
Extracting meaning from audio signals – a machine learning and signal processing approach

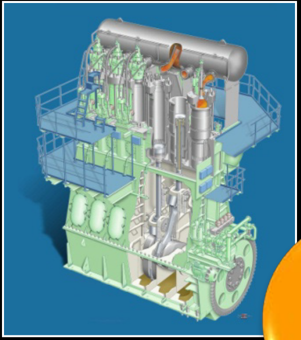
Jan Larsen
Cognitive Systems Section
Dept. of Informatics and Mathematical Modelling
Technical University of Denmark
jl@imm.dtu.dk, www.imm.dtu.dk/~jl

Potential of technological contributions

- Involvement of people and the inclusiveness goal
- Handling of massive amounts of often conflicting data
- Enabling user-centric crowd computing
- Context detection and adaptation
- New intelligent tools eliminating trival work - enhancing experience



Group profile



Systems neuro-science

Digital economy

Multimedia

Mobile services

Machine learning

Signal processing

Cognitive modeling

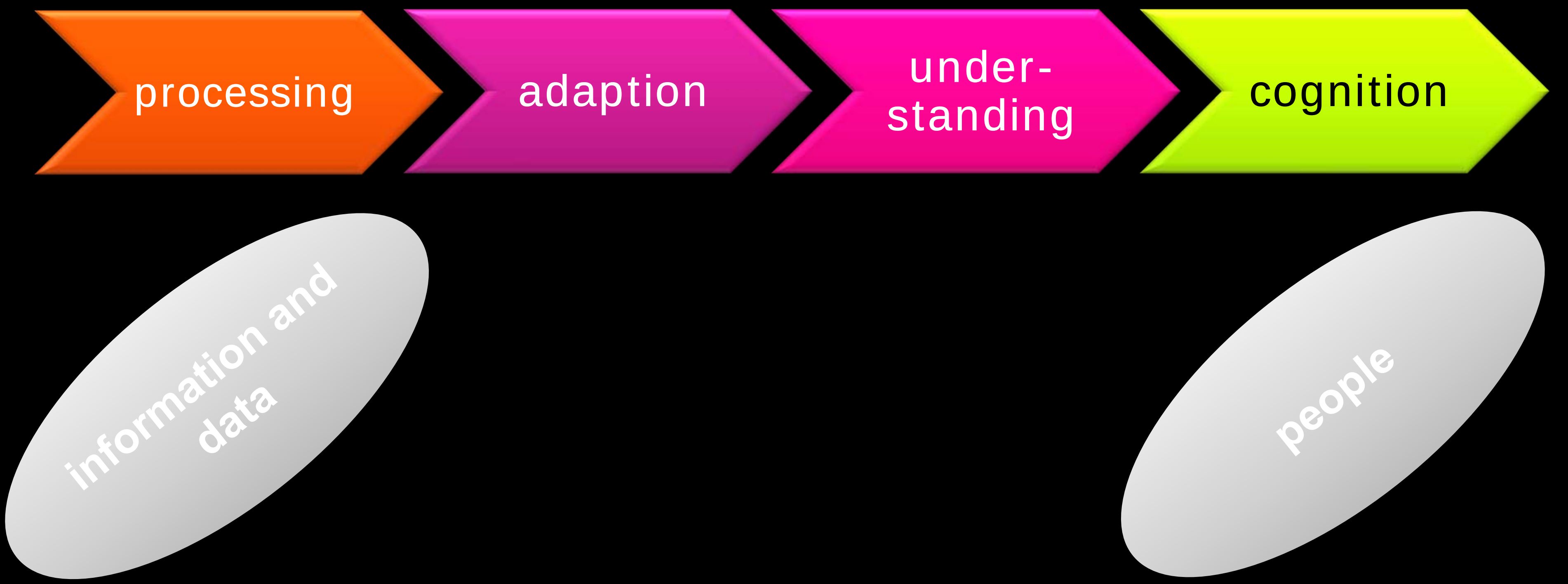
Biomedical

extraction of meaningful and actionable information by ubiquitous learning from data

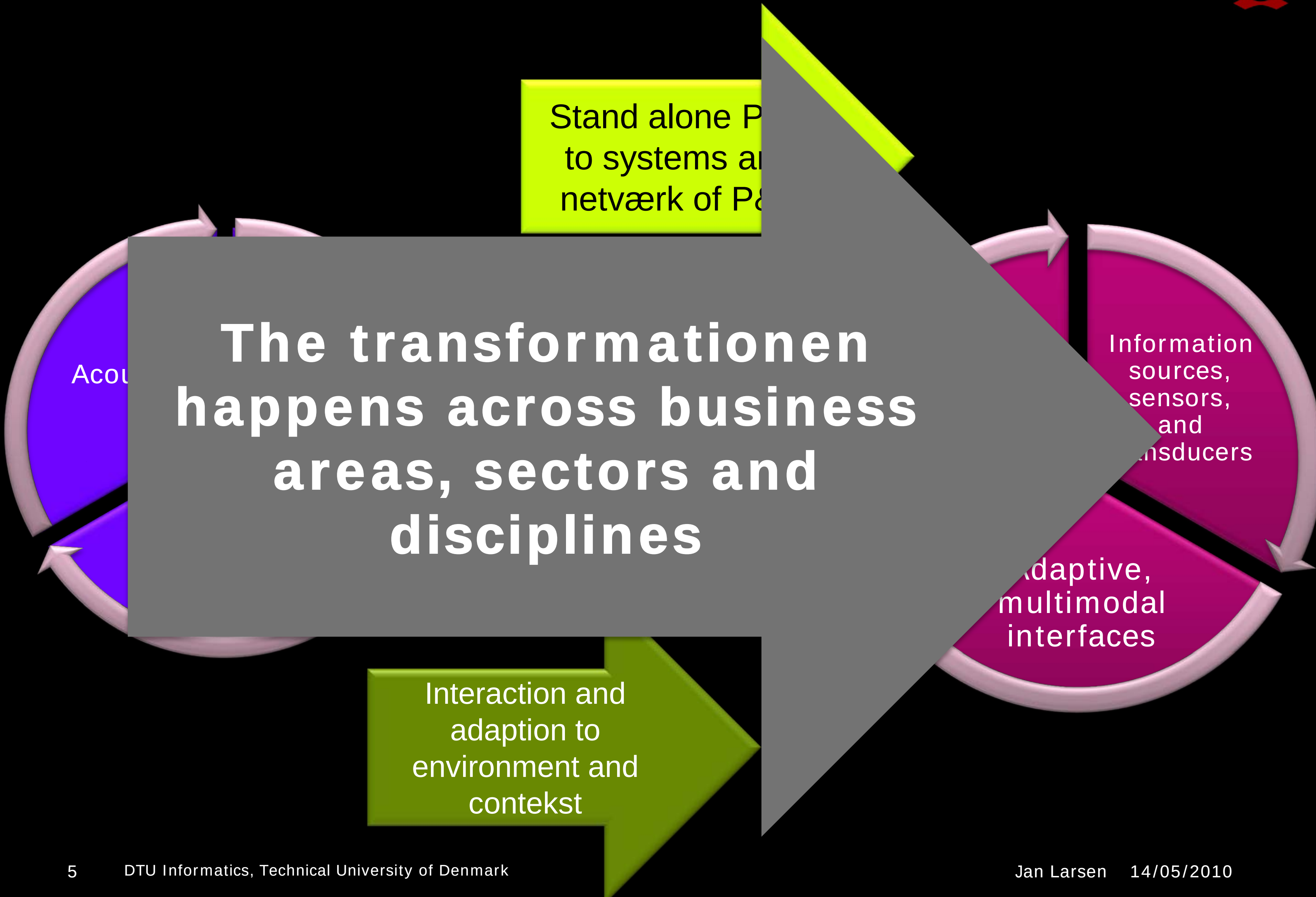
- 5 faculty
- 1 adj. prof.
- 3 postdocs
- 4 adm
- 20 Ph.D. students
- 10 M.Sc. students

The legacy of
Allan Touring and Nobert Wiener

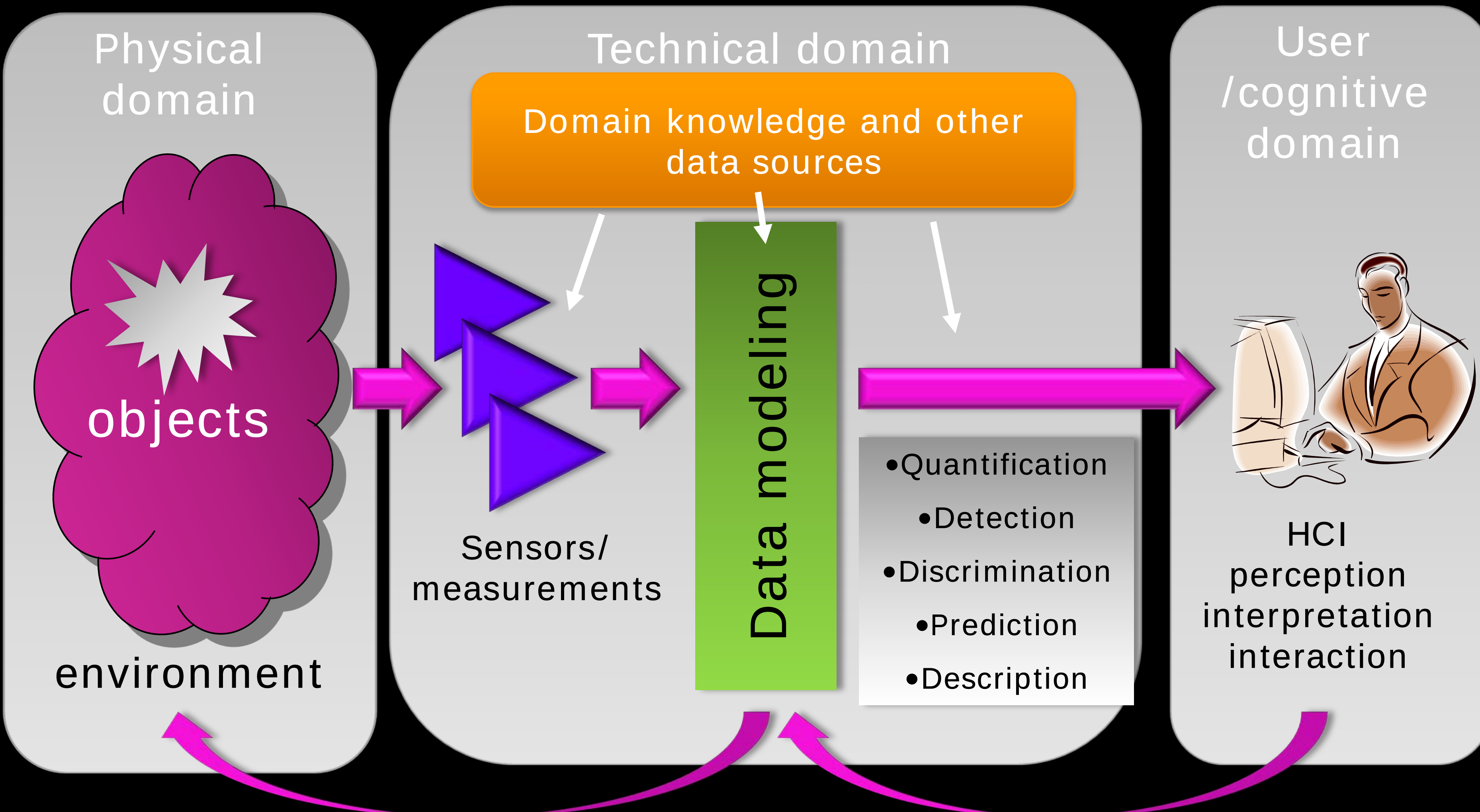
- theory of computing
- cybernetics



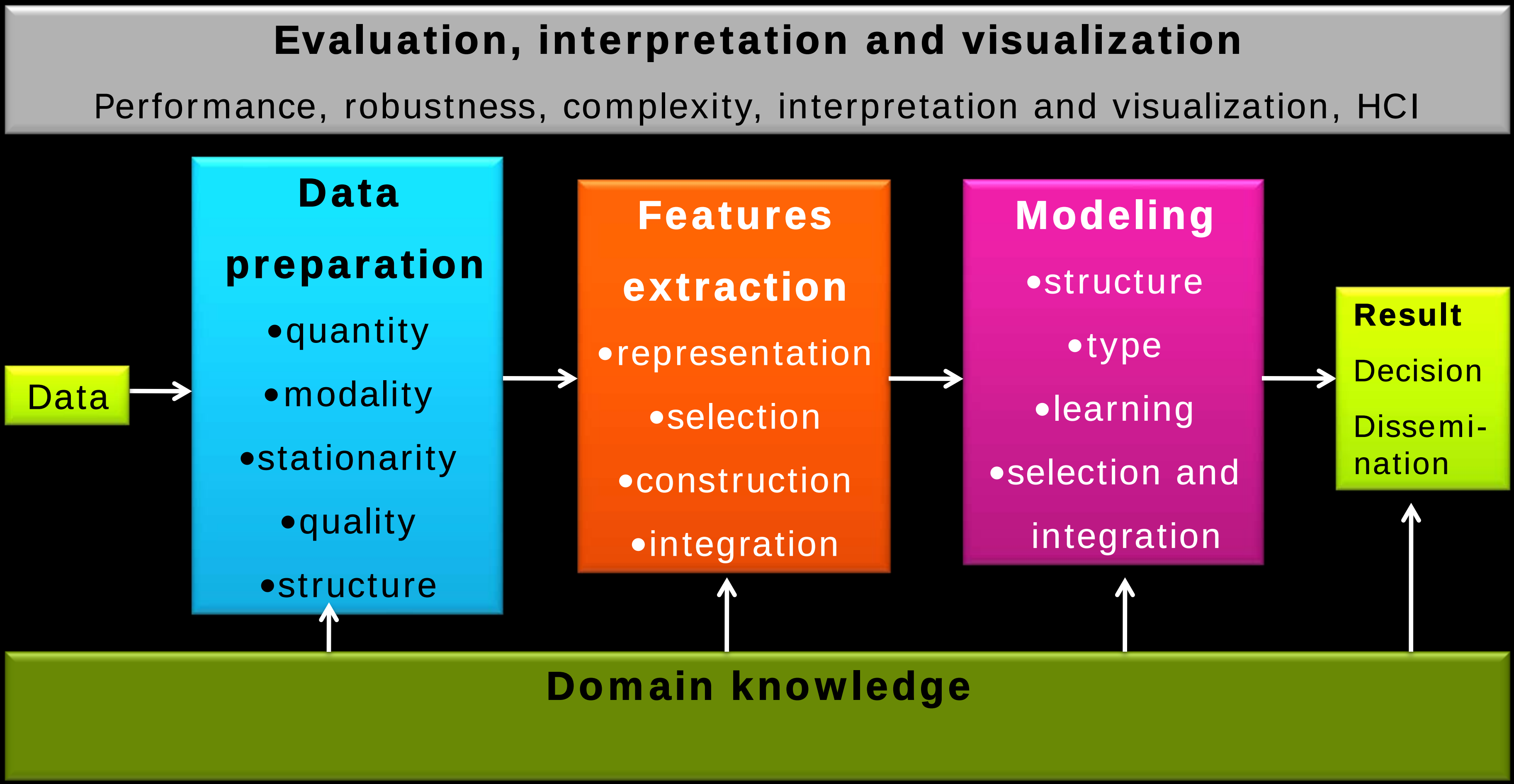
Transformation of sound technologies



Information processing pipeline



Technical data modeling framework



Learning from massive data sets

Disentanglement of confusing, ambiguous, conflicting and vast amounts of information

Perform specific tasks

– Exploration
– Retrieval
– Search
– Physical operatio manipulation
– Information enric
– Making informatic actionable
– Navigation and control

– Decision support

Examples

•Detecting topics in large text corpra

•Automatic annnotation/labeling of songs with genre, mood, etc.

•Speech and image recognition

The unreasonable effectiveness of data

- E. Wigner 1960: The unreasonable efffectiveness of mathematics in the natural sciences
- There is often a sufficient number of data such that simple methods performs better than complex methods
- The power of learning with from unlabeled data which are abundant
- The power of linking many different sources
- Bridging semantic gaps
  - The same meaning can be expressed in many ways – and the same expression can convey many different meanings
  - Shared cognitive and cultural contexts helps the disambiguation of meaning
  - Ontologies: a social construction among people with a common shared motive
  - Classical handcrafted ontology building is infeasible – crowd computing / crowd sourcing is possible!

Ref: A. Halevy, P. Norvig, F. Pereira: The unreasonbale effectiveness of data, IEEE Intelligent Systems, March/April, pp. 8-12, 2009.

The potential of learning machines

- Most real world problems are too complex to be handled by classical physical models and systems engineering approach
- In most real world situations there is access to data describing properties of the problem
- Learning machines can offer
  - Learning of optimal prediction/decision/action
  - Adaptation to the usage environment
  - Explorative analysis and new insights into the problem and suggestions for improvement

Intelligent Sound Project

- FTP project 2005-2009
- 14 mil DKK
- Participants: DTU and Aalborg University

www.intelligentsound.org

Huge demand for tools

Organization, search and retrieval

- Recommender systems (”taste prediction”)
- Playlist generation
- Finding similarity in music (e.g., genre classification, instrument classification, etc.)
- Hit prediction
- Newscast transcription/search
- Music transcription/search

Machine learning in sound information processing

audio data

Meta data

ID3 tags

context

machine learning model

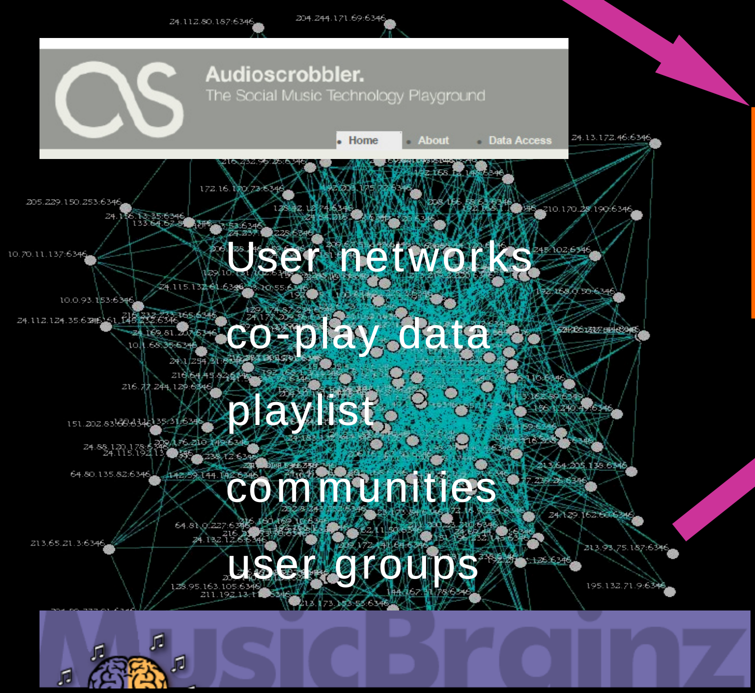
Tasks

Grouping

Classification

Mapping to a structure

Prediction e.g. answer to query



User networks

co-play data

playlist

communities

user groups

Specialized search and music organization

last.fm the social music revolution

Using social network analysis

allmusic AMG

Explore by Genre, mood, theme, country, instrument

Query by humming

Fraunhofer IDMT Institut Digitale Medientechnologie

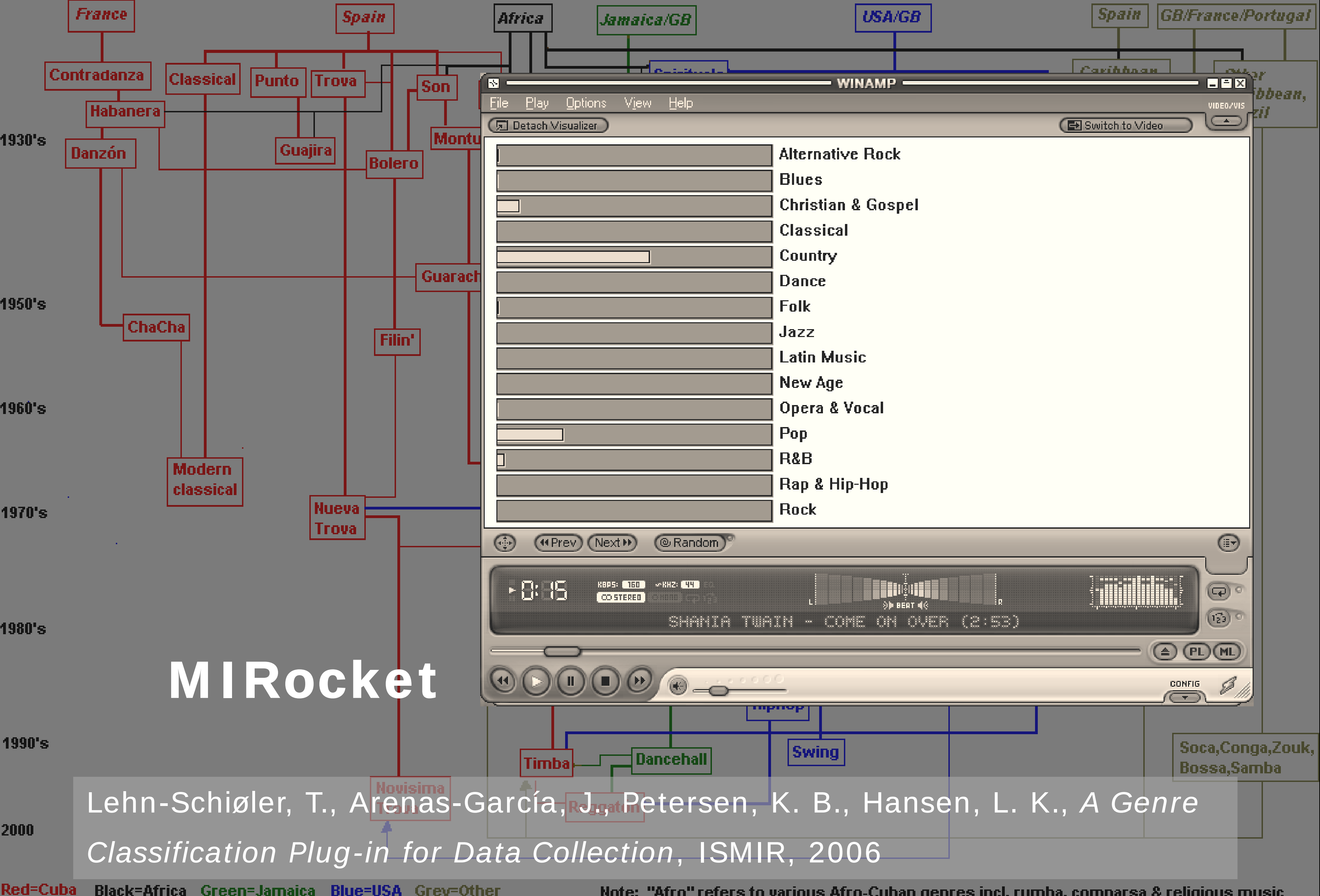
The NGSW is creating an online fully-searchable digital library of spoken word collections spanning the 20th century

Organize songs according to tempo, genre, mood

PANDORA

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

MIRocket



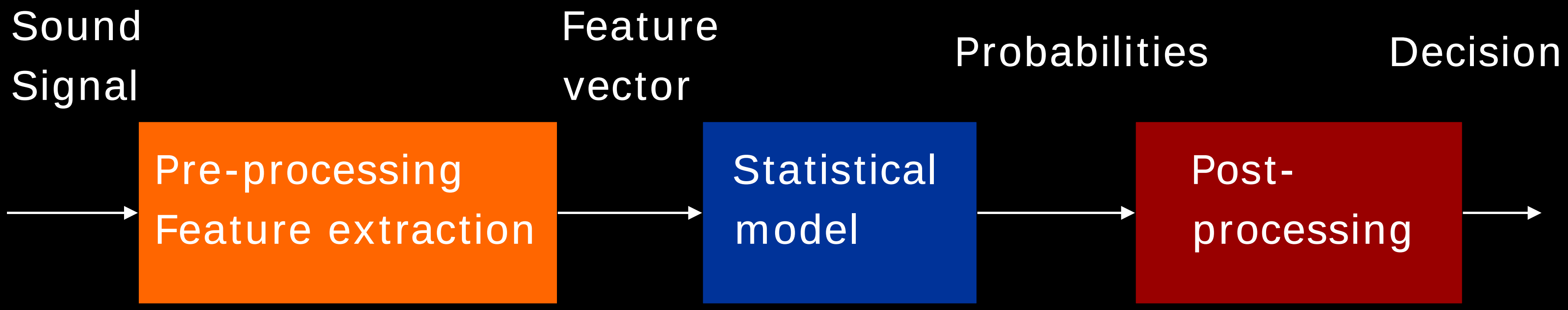
Lehn-Schiøler, T., Arenas-García, J., Petersen, K. B., Hansen, L. K., *A Genre Classification Plug-in for Data Collection*, ISMIR, 2006

Genre classification

- Prototypical example of predicting meta and high-level data
- The problem of interpretation of genres
- Can be used for other applications e.g. context detection in hearing aids

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

Features for genre classification

30s sound clip from the center of the song

6 MFCCs, 30ms frame

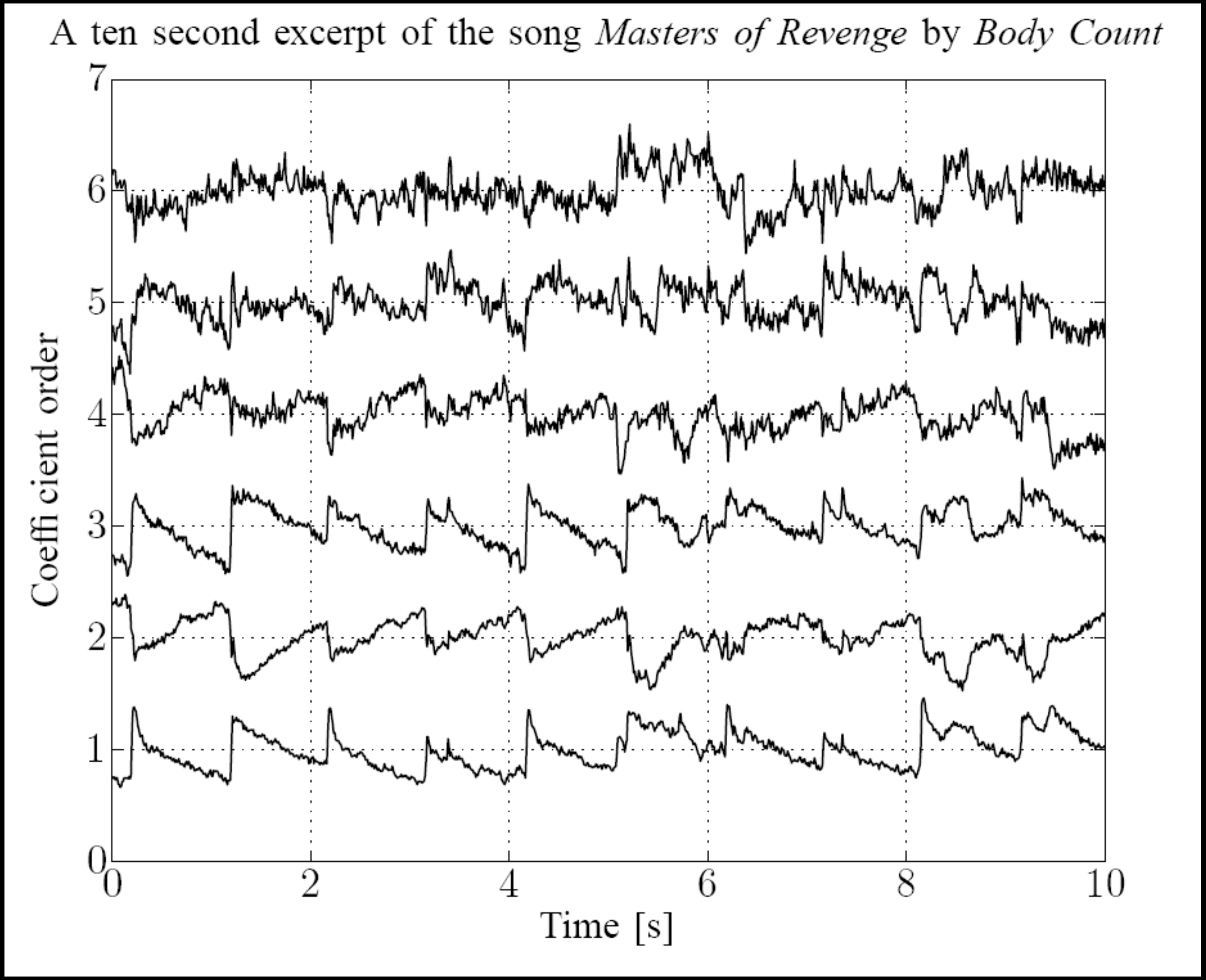
6 MFCCs, 30ms frame

6 MFCCs, 30ms frame

3 ARCs per MFCC, 760ms frame

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

Example of MFCC's



- Cross correlation
- Temporal correlation

Results reported in

- Meng, A., Ahrendt, P., Larsen, J., Hansen, L. K., Temporal Feature Integration for Music Genre Classification, IEEE Transactions on Speech and Audio Processing, 2007.
- 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.
- Ahrendt, P., Goutte, C., Larsen, J., *Co-occurrence Models in Music Genre Classification*, IEEE International workshop on Machine Learning for Signal Processing, pp. 247-252, 2005.
- Ahrendt, P., Meng, A., Larsen, J., *Decision Time Horizon for Music Genre Classification using Short Time Features*, EUSIPCO, pp. 1293--1296, 2004.
- Meng, A., Shawe-Taylor, J., *An Investigation of Feature Models for Music Genre Classification using the Support Vector Classifier*, International Conference on Music Information Retrieval, pp. 604-609, 2005

Best results

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

Best 11-genre confusion matrix

Alternative	41.8	6.4	4.5	3.6	3.6	2.7	8.2	2.7	4.5	3.6	18.2
Country	0.9	72.7	7.3	0.0	4.5	2.7	4.5	0.9	2.7	0.0	3.6
Easy-listening	1.8	11.8	61.8	2.7	4.5	2.7	2.7	0.0	2.7	3.6	5.5
Electronica	5.5	0.9	10.9	41.8	8.2	5.5	7.3	10.9	2.7	5.5	0.9
Jazz	0.9	4.5	8.2	10.9	50.0	2.7	3.6	2.7	7.3	6.4	2.7
Latin	3.6	8.2	2.7	4.5	3.6	37.3	8.2	8.2	4.5	11.8	7.3
Pop&Dance	6.4	9.1	6.4	9.1	0.9	11.8	43.6	2.7	3.6	2.7	3.6
Rap&Hiphop	0.0	0.0	0.9	7.3	0.9	4.5	3.6	62.7	1.8	17.3	0.9
RB&Soul	0.9	8.2	9.1	0.9	9.1	11.8	7.3	9.1	29.1	5.5	9.1
Reggae	0.9	0.9	0.0	3.6	4.5	5.5	1.8	17.3	3.6	61.8	0.0
Rock	25.5	16.4	5.5	0.9	5.5	2.7	6.4	0.0	6.4	1.8	29.1

Music separation

- A possible front end component for the music search framework
- Noise reduction
- Music transcription
- Instrument detection and separation
- Vocalist identification

Semi-supervised learning methods

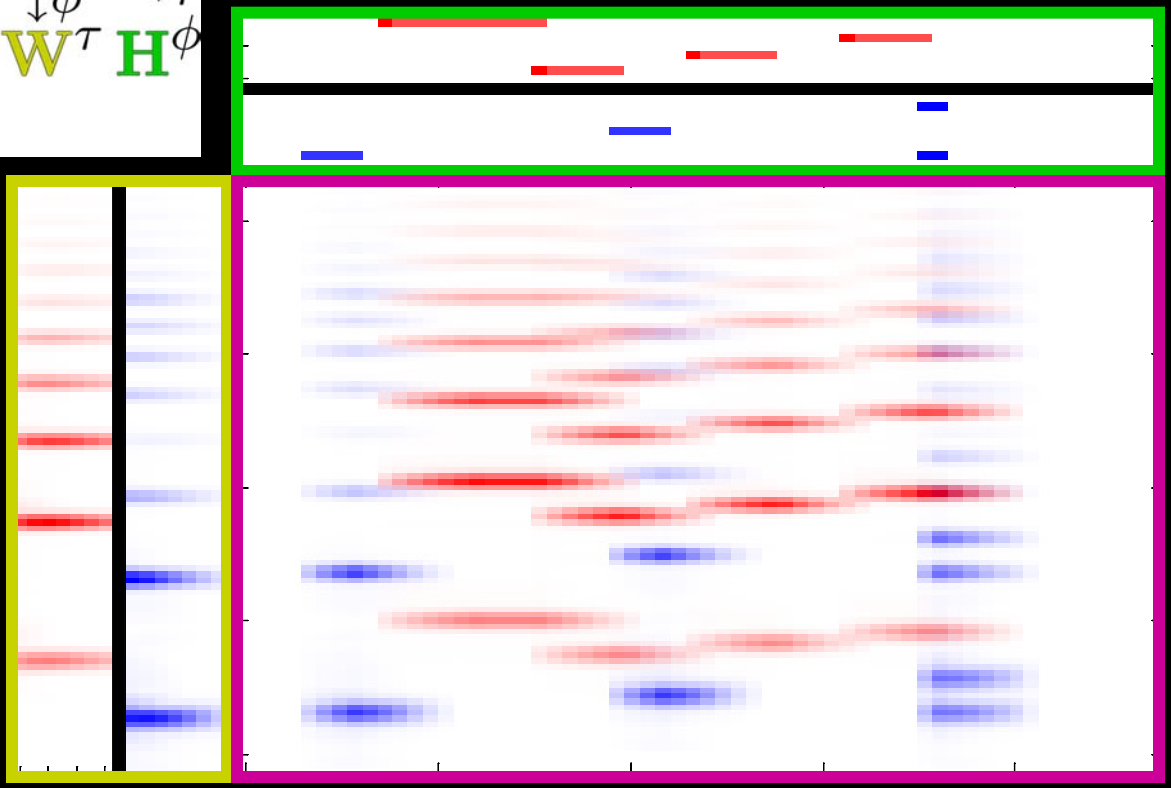
Pedersen, M. S., Larsen, J., Kjems, U., Parra, L. C., *A Survey of Convolutive Blind Source Separation Methods*, Springer Handbook of Speech, Springer Press, 2007

Nonnegative matrix factor 2D deconvolution

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

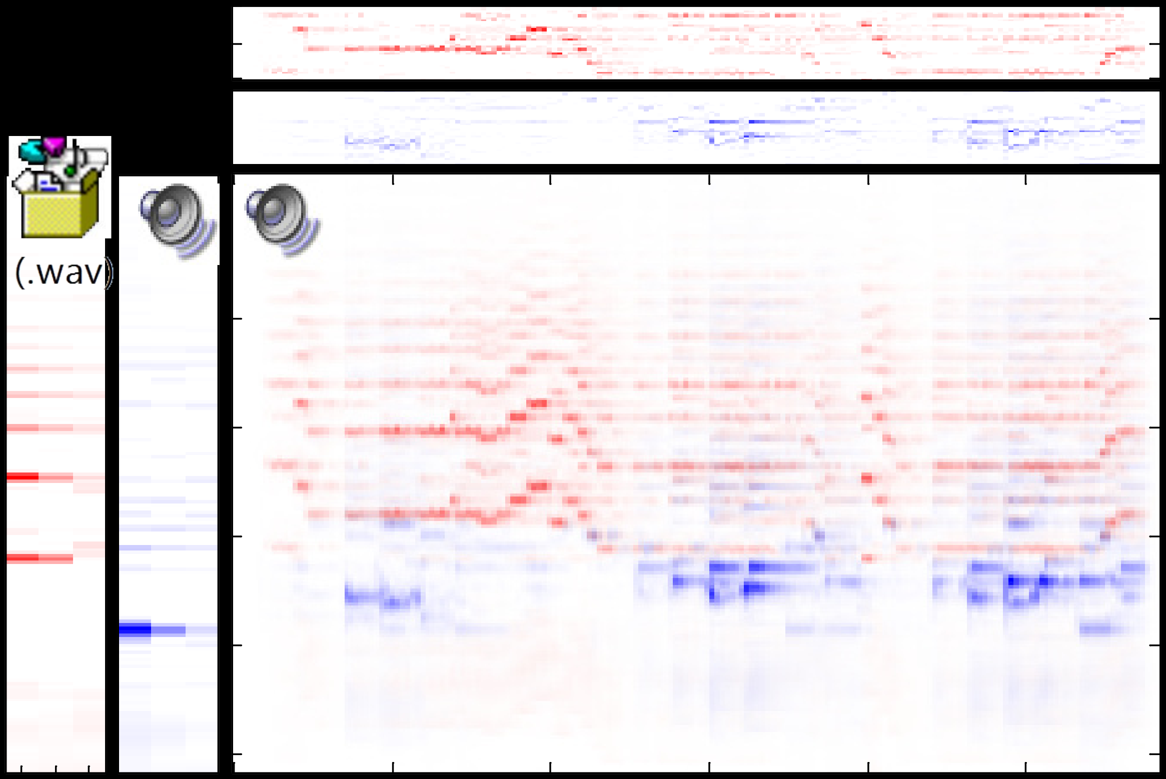
time

pitch

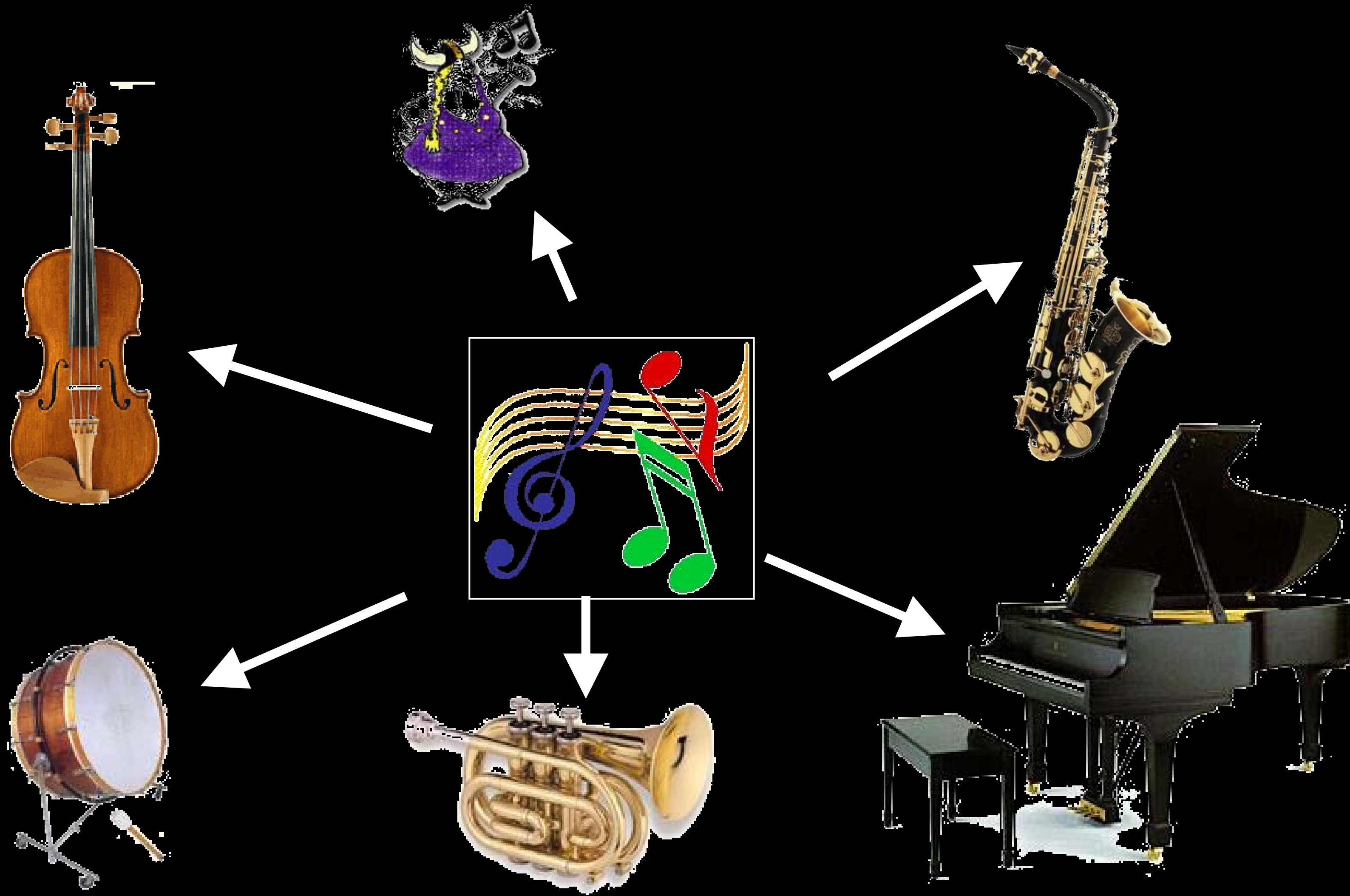


M. N. Schmidt, M. Mørup *Nonnegative Matrix Factor 2-D Deconvolution for Blind Single Channel Source Separation*, ICA2006, 2006. Demo also available.

Demonstration of the 2D convolutive NMF model



Separating music into basic components



Separating music into basic components

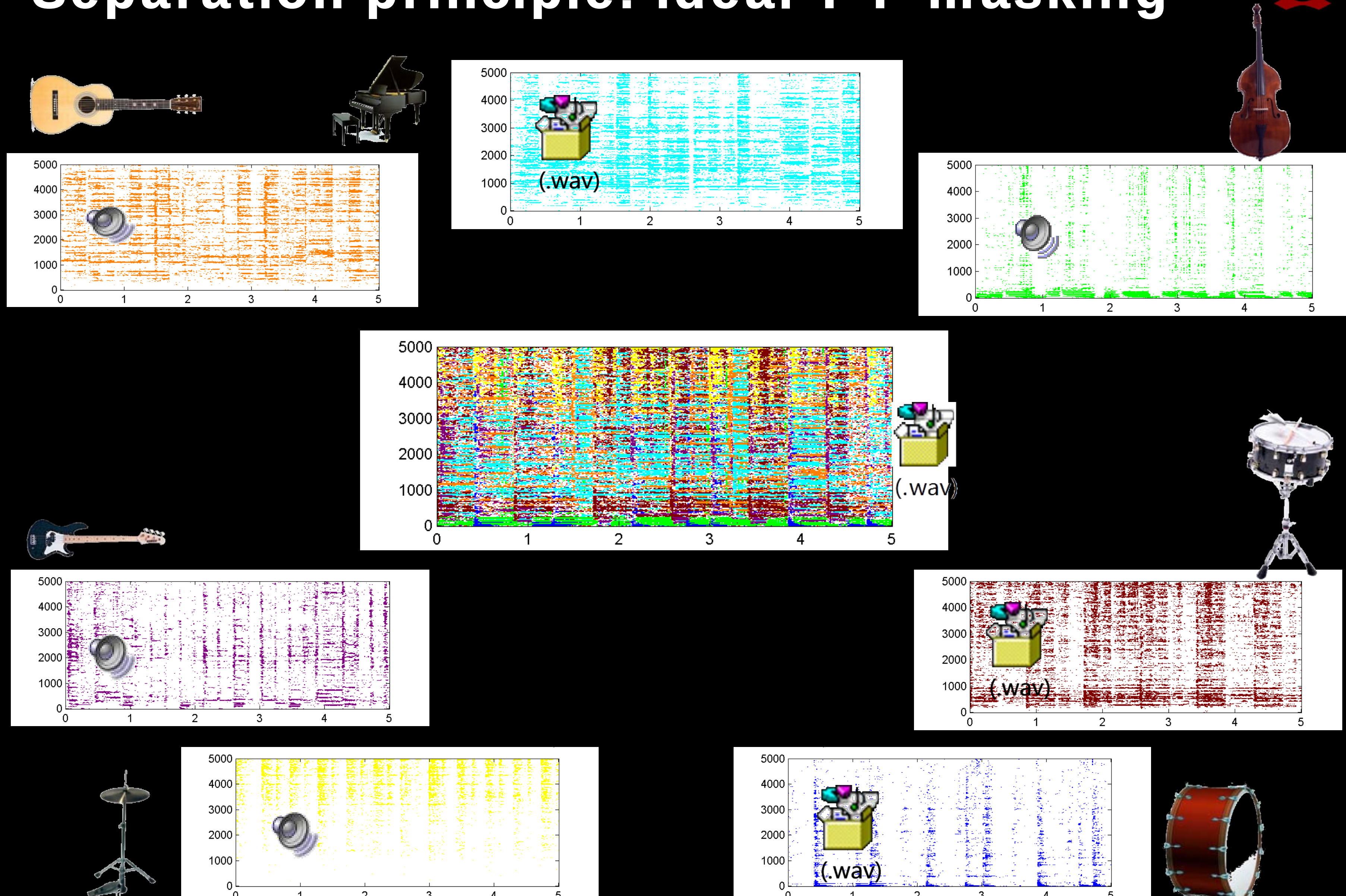
- Combined ICA and masking

  - Pedersen, M. S., Wang, D., Larsen, J., Kjems, U., Two-microphone Separation of Speech Mixtures, IEEE Transactions on Neural Networks, 2007
  - Pedersen, M. S., Lehn-Schiøler, T., Larsen, J., *BLUES from Music: BLind Underdetermined Extraction of Sources from Music*, ICA2006, vol. 3889, pp. 392-399, Springer Berlin / Heidelberg, 2006
  - Pedersen, M. S., Wang, D., Larsen, J., Kjems, U., *Separating Underdetermined Convolutive Speech Mixtures*, ICA 2006, vol. 3889, pp. 674-681, Springer Berlin / Heidelberg, 2006
  - Pedersen, M. S., Wang, D., Larsen, J., Kjems, U., *Overcomplete Blind Source Separation by Combining ICA and Binary Time-Frequency Masking*, IEEE International workshop on Machine Learning for Signal Processing, pp. 15-20, 2005

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

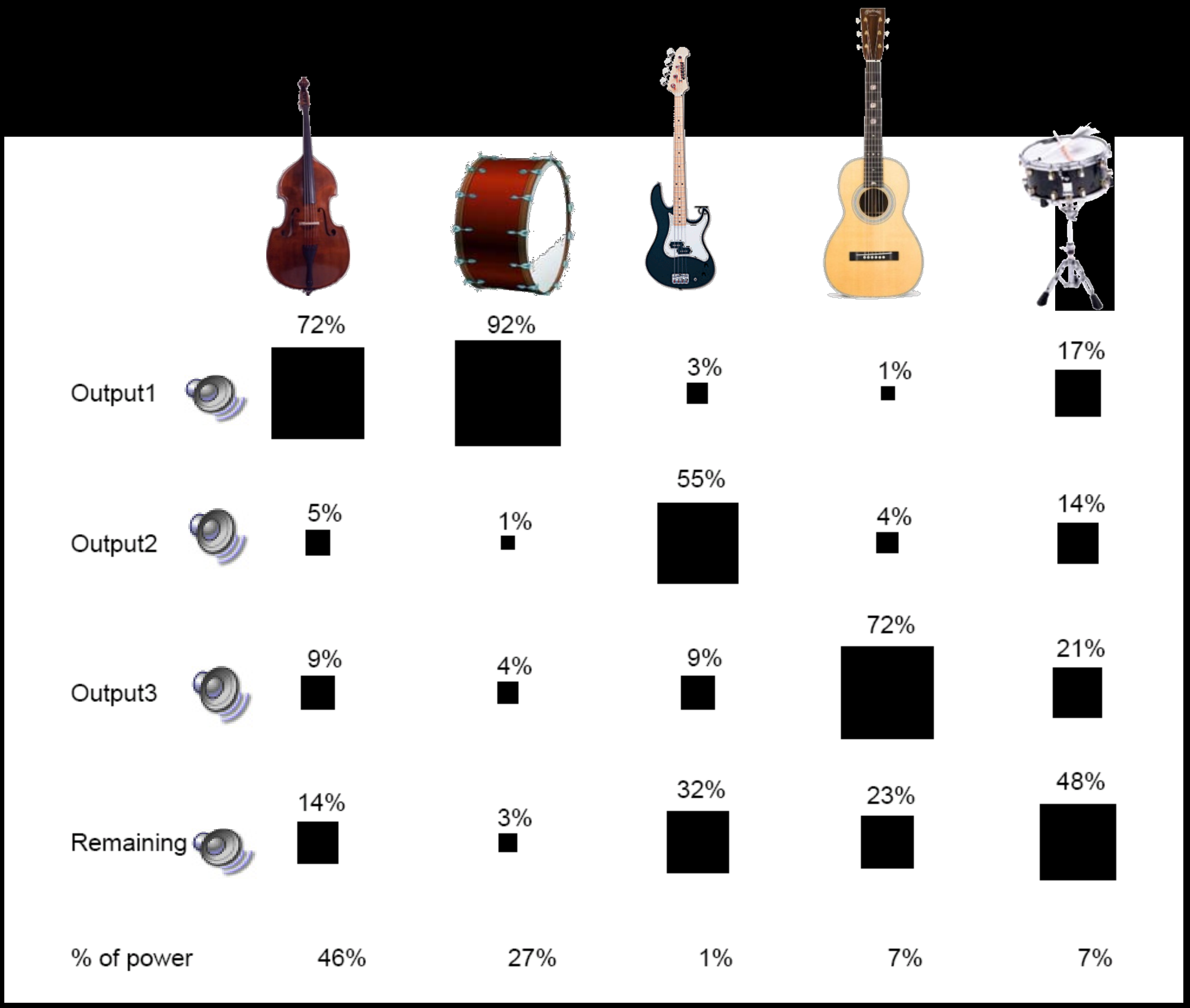


Results

- 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 and code are available online via http://www.imm.dtu.dk

Results

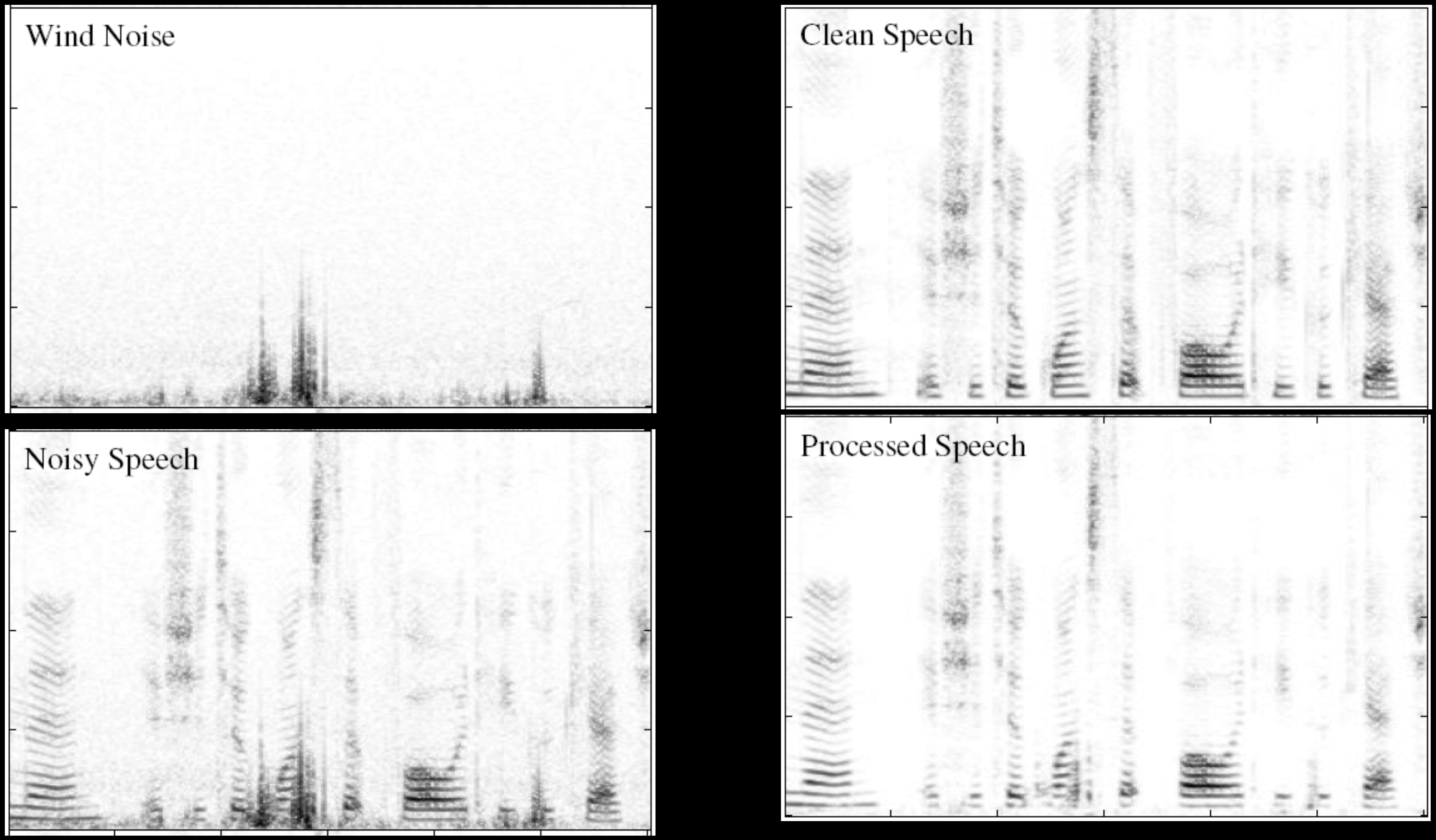
- The segregated outputs are dominated by individual instruments
- Some instruments cannot be segregated by this method, because they are not spatially different.



Conclusion on combined ICA T-F separation

- An unsupervised method for segregation of single instruments or vocal sound from stereo music.
- 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.

Wind noise reduction



M.N Schmidt, J. Larsen, F.T. Hsiao: Wind noise reduction using non-negative sparse coding, 2007.

Sparse NMF decomposition

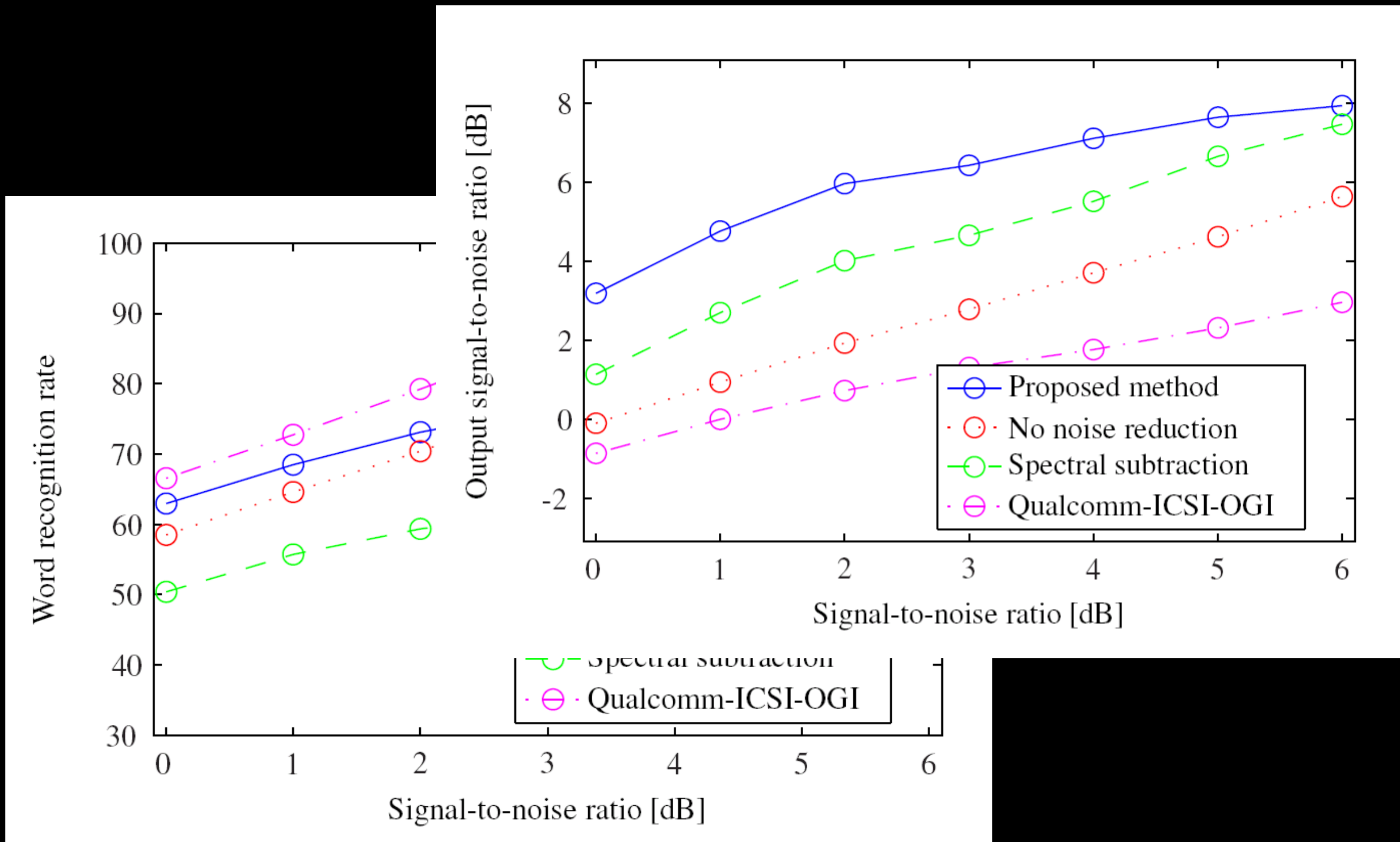
- Code-book (dictionary) of noise spectra is learned
- Can be interpreted as an advanced spectral subtraction technique

original

cleaned

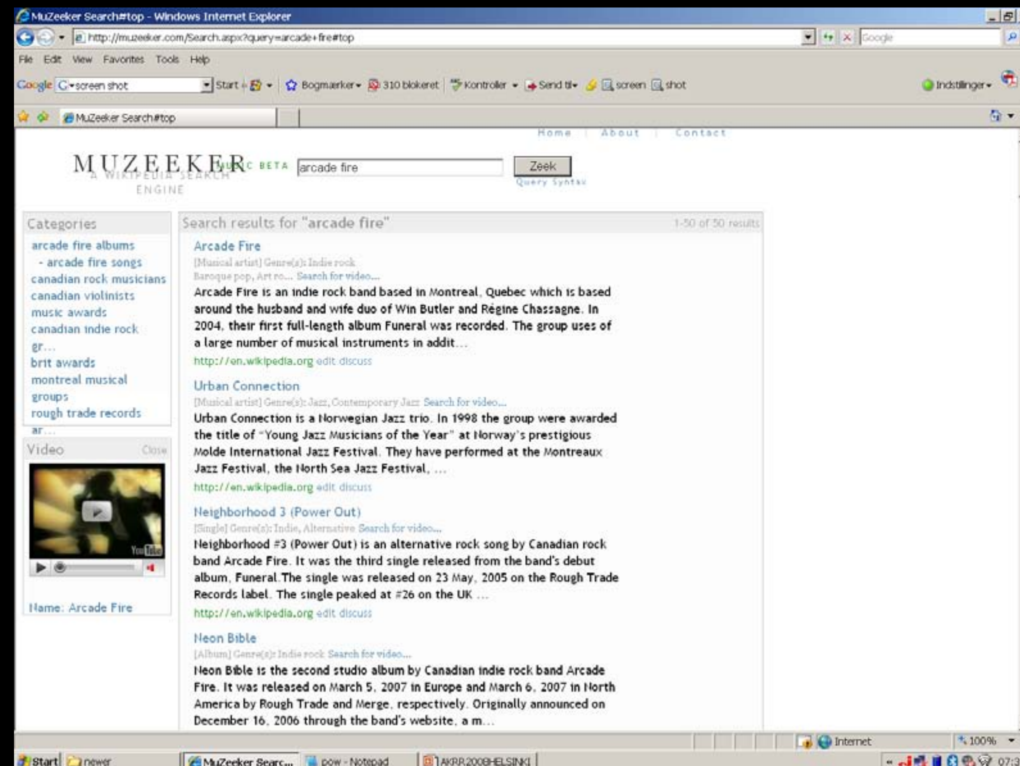
alternative method (qualcom)

Objective performance



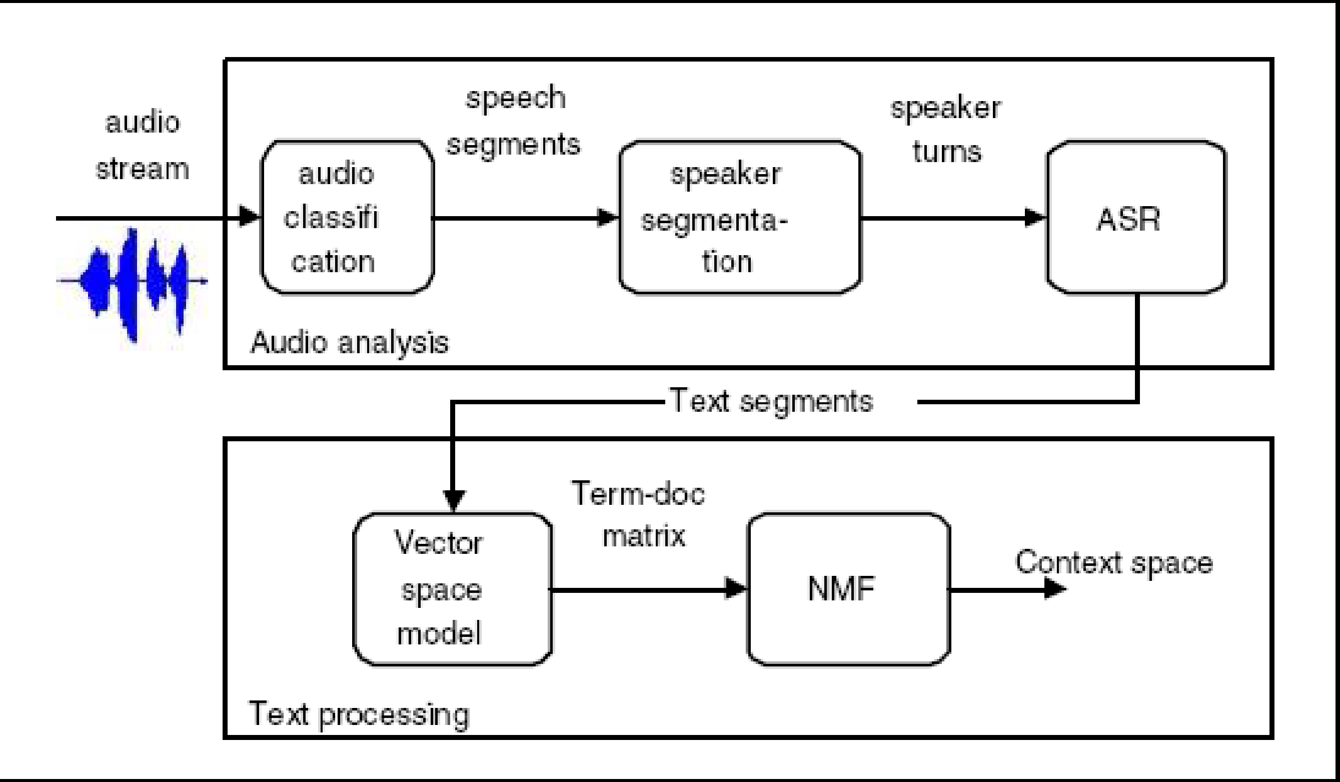
A cognitive search engine - Muzeeker

- Wikipedia based common sense
- Wikipedia used as a proxy for the music users mental model
- Implementation: Filter retrieval using Wikipedia's article/ categories

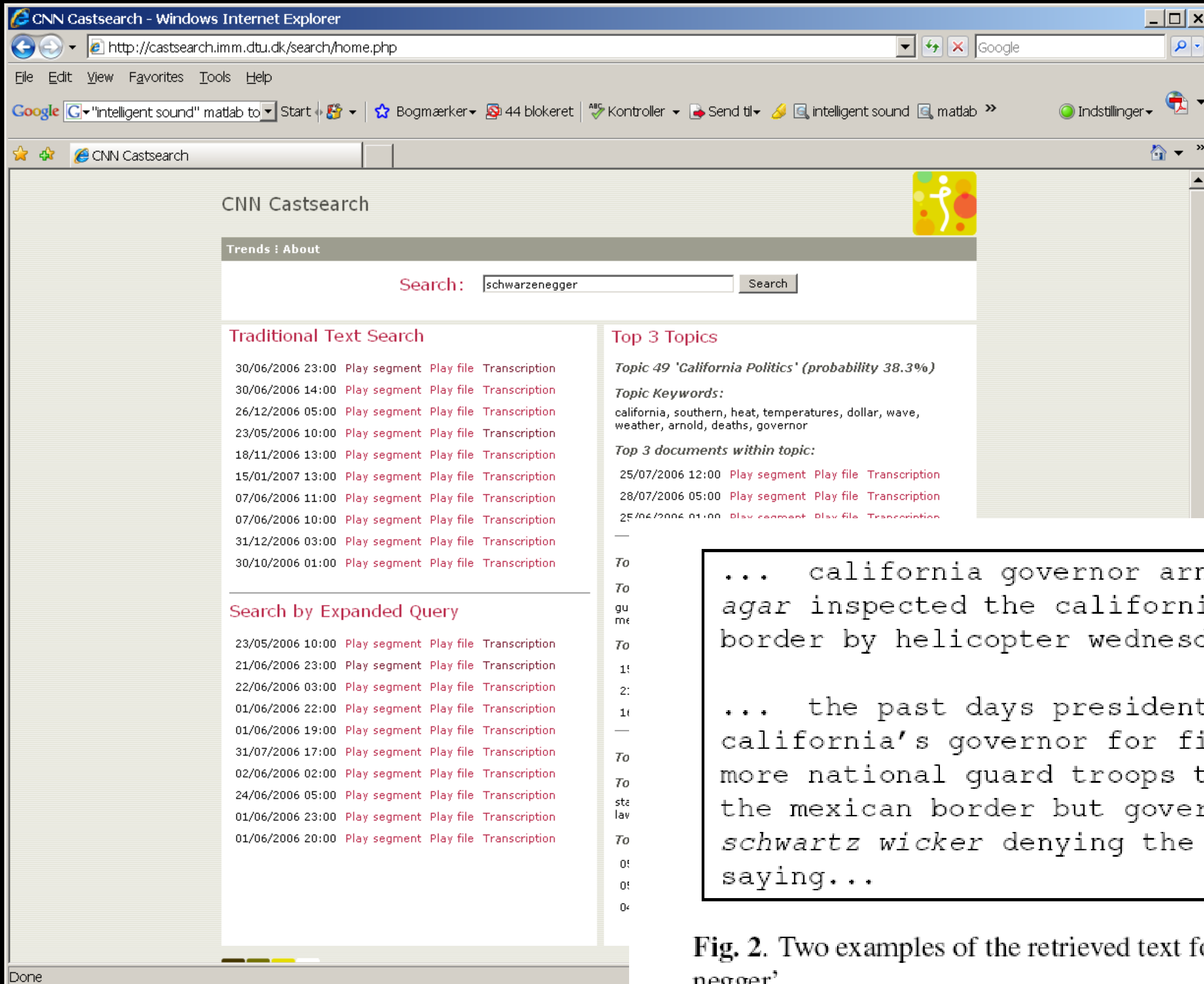


- Muzeeker.com

A cognitive search engine – CASTSEARCH: Context based Spoken Document Retrieval



Ref: Lasse Mølgaard, Kasper Jørgensen, Lars Kai Hansen: "CASTSEARCH: Context based Spoken Document Retrieval," ICASSP2007



... california governor arnold's *fortson agar* inspected the california mexico border by helicopter wednesday to see ...

... the past days president bush asking california's governor for fifteen hundred more national guard troops to help patrol the mexican border but governor orville *schwartz wicker* denying the request saying...

Fig. 2. Two examples of the retrieved text for a query on 'schwarzenegger'.

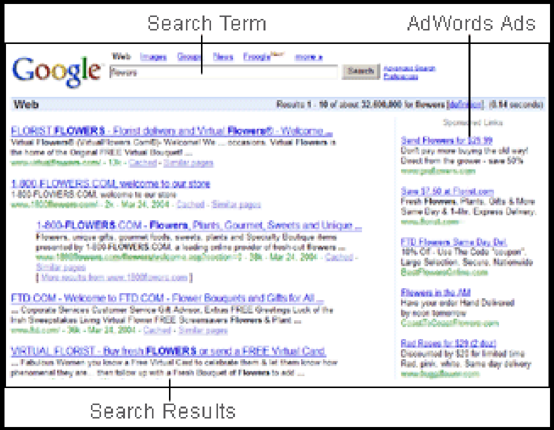
Ref: http://castsearch.imm.dtu.dk

Vertical search

- Deep web databases
  - Digital media
  - For profit: DMR issues
- Specialized search engines
  - Professional users
  - Modeling deep structure
- Key role in Web 2.0
  - User generated content
  - Bioinformatics
  - Neuroinformatics:
    - BrainMap, Brede search engine

Horizontal search

- Google
  - Volume
  - Ranking
  - Explorative vs retrieval
  - Adword business model
- Semantic web
  - Wikipedia
  - User generated content



Courtesey of Lars Kai Hansen, DTU

Crowd computing and user involvement

Challenges: There is a social/phychological interia towards traditional solutions

1. The Retarding Power (or Inertia) of a Word
2. A Partial Res
3. Tradition Ca
4. Words and T
5. Inadmissible Range of Data
6. Association of Objects with Senses
7. All Information Given is Valid

Users' engagement and motivation through relevance, surprice and precision of results

Ref: James Kowalick http://www.triz-journal.com/archives/1998/08/c/default.asp

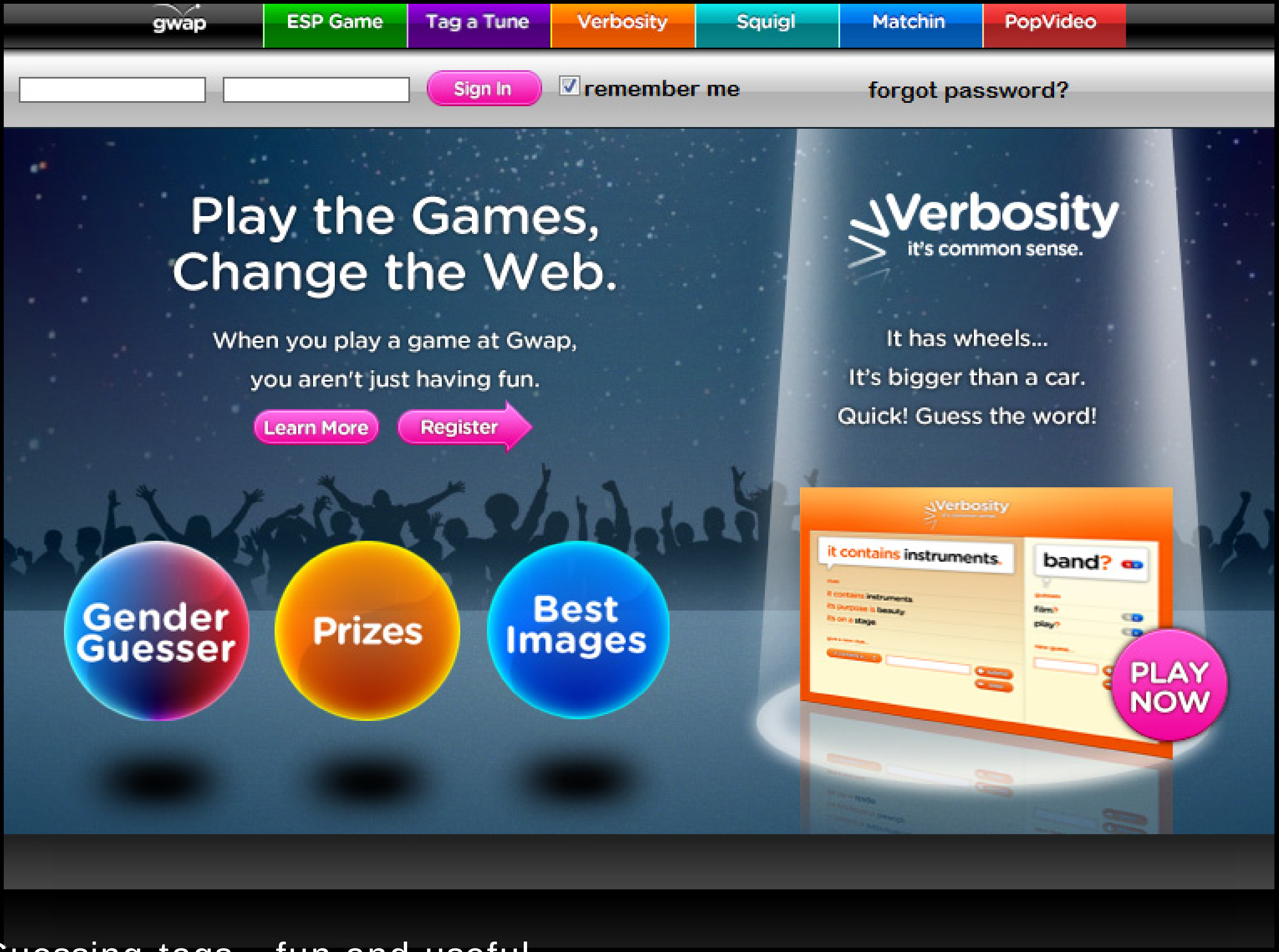
Voictor Fey and Eugene Rivin: Innovation on Demand, 2005.

TRIZ The theory of solving inventor's problems, http://en.wikipedia.org/wiki/TRIZ

M.S. Gazzaniga *et al.:* The Cognitive Neurosciences, 1994.

Samer Abdallah, Mark Plumbley: Information dynamics: patterns of expectation and surprise in the perception of music , Connection Science , vol. 21, issue 2, p. 89, 2009

ESP game



- Guessing tags - fun and useful
- Conceived by Luis von Ahn of Carnegie Mellon University

reCAPTCHA™
Digitizing Books One Word at a Time
HOME
WHAT IS reCAPTCHA
DIGITIZATION ACCURACY
WHAT IS A CAPTCHA
SECURITY
GET reCAPTCHA
MY ACCOUNT
EMAIL PROTECTION
RESOURCES
Type the two words:
reCAPTCHA™ stop spam. read books.
Submit
The words above come from scanned books.
By typing them, you help to digitize old texts.
reCAPTCHA is a free CAPTCHA service that helps to digitize bo
shows. Check out our paper in Science about it (or read more b
A CAPTCHA is a program that can tell whether its user is a hum
seen them — colorful images with distorted text at the bottom o
are used by many websites to prevent abuse from "bots," or aut
generate spam. No computer program can read distorted text a
cannot navigate sites protected by CAPTCHAs.

Research based vs user-driven knowledge and folksonomy

SØNDAG 23. AUGUST 2009 **POLITIKEN**

Når man holder op med at tro på forsknings-baseret viden og bare lader, som om det er en holdning som alle andre, så bliver vi me-re og mere bare overladt til, hvad folk me-ner, uafhængigt af fakta

Maja Horst

Assoc.Prof. CBS

- user driven knowledge is often inaccurate and misleading
- how do we avoid dominance by the popular (music recommendation systems)
- sufficient amount of contributions ensures the quality (wikipedia)

Measurement systems for ethical capital in the experience economy socio-economic value of online communication

- New research 3-year research project starting Aug. 2009 (CBS,DTU,Univ. Milan)
- Forrester Research Report shows web2.0 marked grows enormeously
- The assumption is that on-line spontaneous communication processes are predictible as they appear in networks and patterns which can be revealed by combining socio-economic studies, linguistics, text and network modeling

Responsible Business in the Blogosphere

Kulturarven kan ende i digitalt hul

Men hvis brugerne involveres bredt kan vi sammen skabe en levende digital kulturarv, der kan bidrage til sammenhæng i det danske samfund – hvis ikke, er der fare for at arven forsvinder i et digitalt sort hul, utilgængelig og død.

STIFINDERE
LARS KAI HANSEN
PROFESSOR, DTU INFORMATIK

peana. Effektiv eksponering kræver sandsynligvis, at der også laves en struktur for indsamling af metadata. Med metadata menes beskrivelser af indholdet: hvad betyder det?, hvem indgår?, hvor stammer det fra? Uden metadata er digitalt indhold utilgængeligt og impotent.

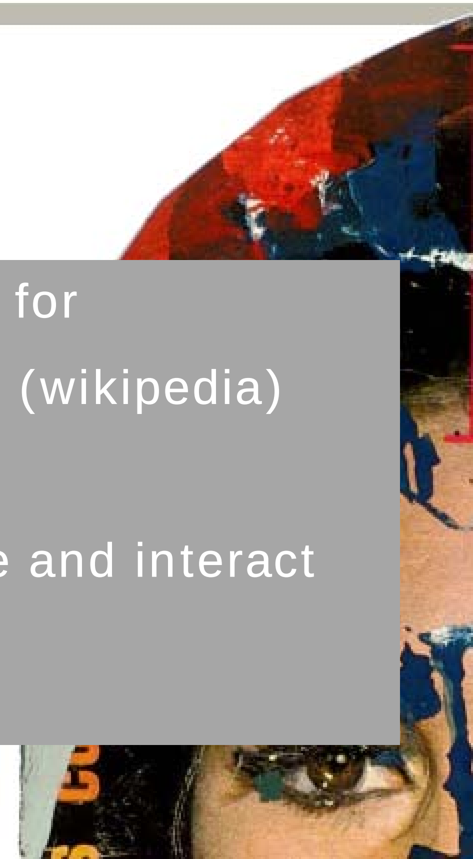
Rapporten er desværre noget uambitiøs, når det kommer til involvering af brugerne i skabelse af metadata, og især når det kommer til anvendelse af avanceret data-analyse. Det virker, som om udvalget i høj grad vil forlade sig på traditionelle metadata-kilder, niche-eksperter og bibliotekarer. Men det kan blive dyrt og svært at vedligeholde.

KULTURARVEN SKAL kunne tilgås a Google, det er en nødvendighed. Me Google forstår ikke sine data. Søgema skinen Google er uden konkurrence når man ved, hvad man leder efter. Hvi

POLITIKEN | ONSDAG 13. MAJ 2009

POLITIKEN | MANDAG 24. AUGUST 2009

Kulturarven skal gøres



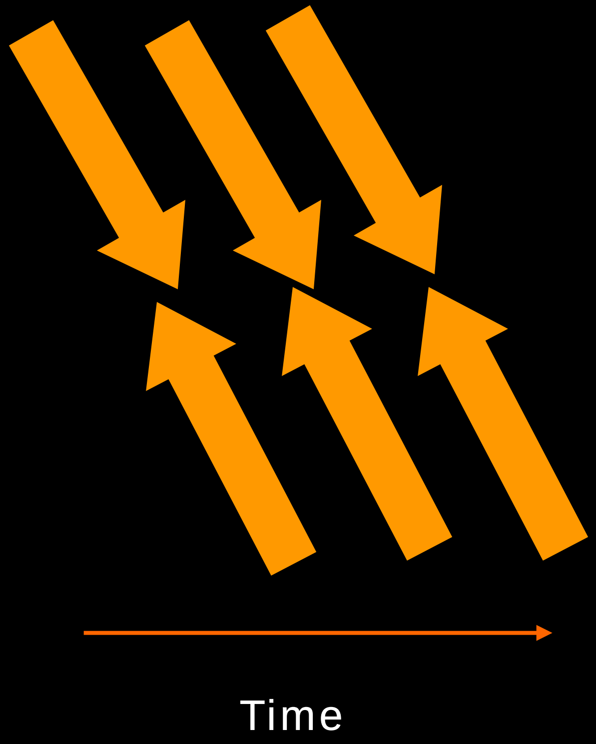
- Google only works if you know what you are searching for
- We need to integrate with common knowledge sources (wikipedia)
- We need to use learning to annotate meta data
- We need users to create additional content, collaborate and interact with data

objekter på danske museer, og de fleste vil også være fotograferet og frit tilgængelige for brugerne.

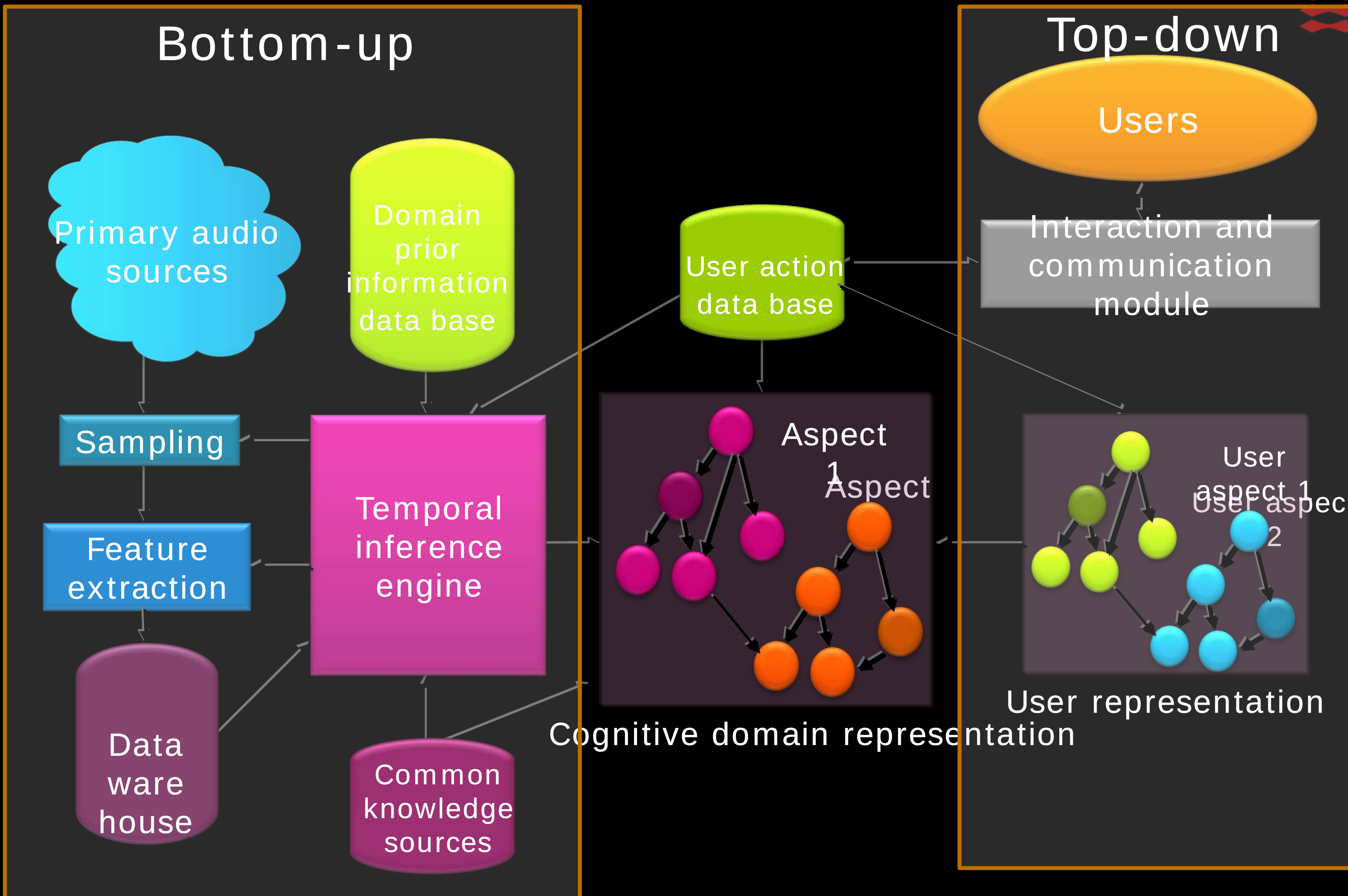
A cognitive architecture for search

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



CoSound architecture



Summary

- A cross-disciplinary effort is required to make research, innovation and commercial products and services
- Massiveness of data requires learning and cognitive modeling but has huge potential for new capabilities
- Integration of multiple information sources helps context detection and adaptation
- Internet penetration makes crowd sourcing possible and ensures inclusiveness
  - a window for the creative common
  - a way to bridging the semantic gap