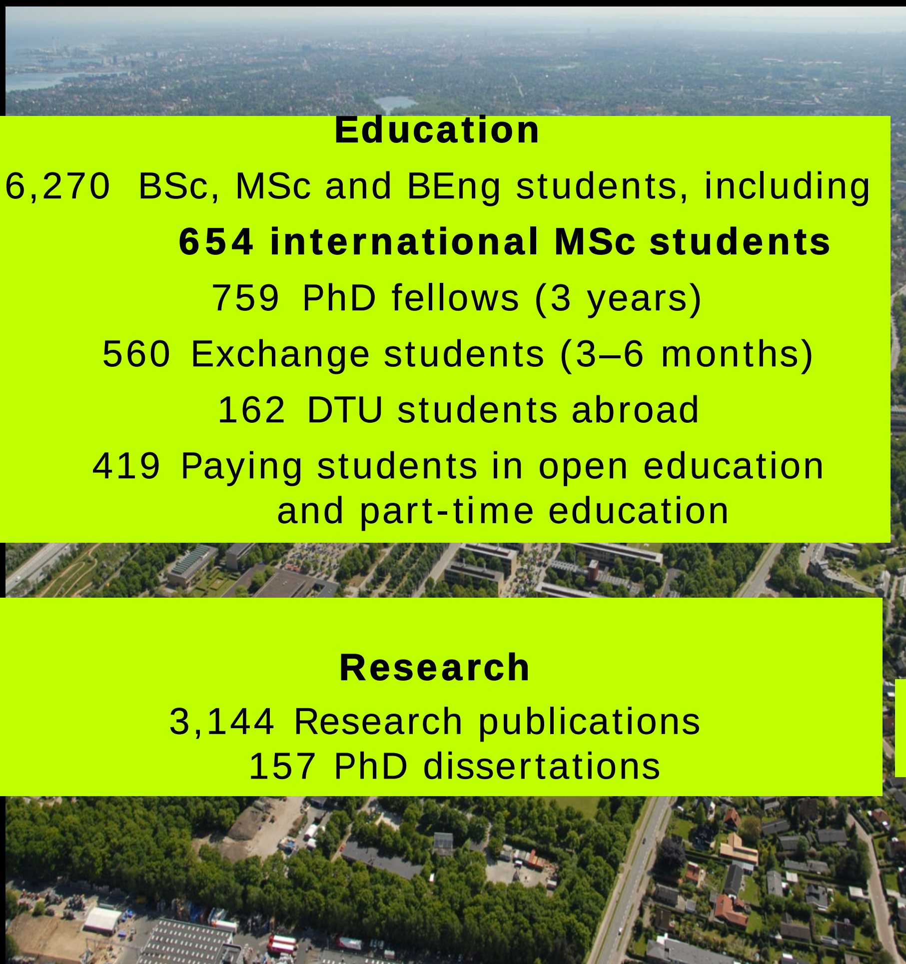
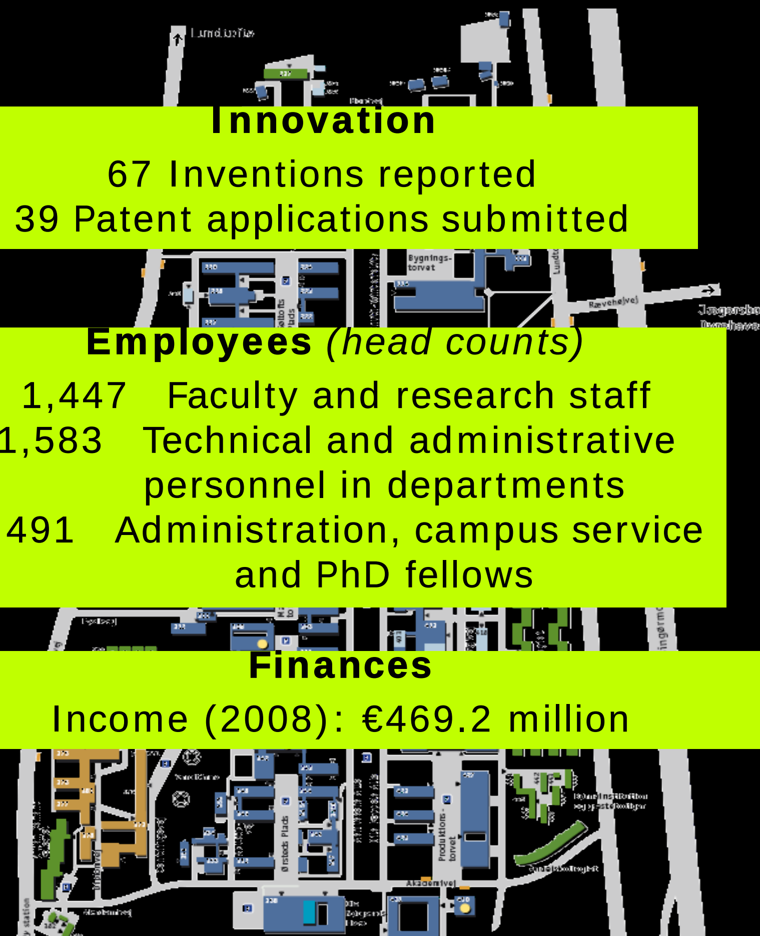
Extracting Meaning from Sound Signals

a machine learning approach

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

DTU, Lyngby Campus





Education

6,270 BSc, MSc and BEng students, including

654 international MSc students

759 PhD fellows (3 years)

560 Exchange students (3–6 months)

162 DTU students abroad

419 Paying students in open education and part-time education

Research

3,144 Research publications

157 PhD dissertations

Innovation

67 Inventions reported

39 Patent applications submitted

Employees *(head counts)*

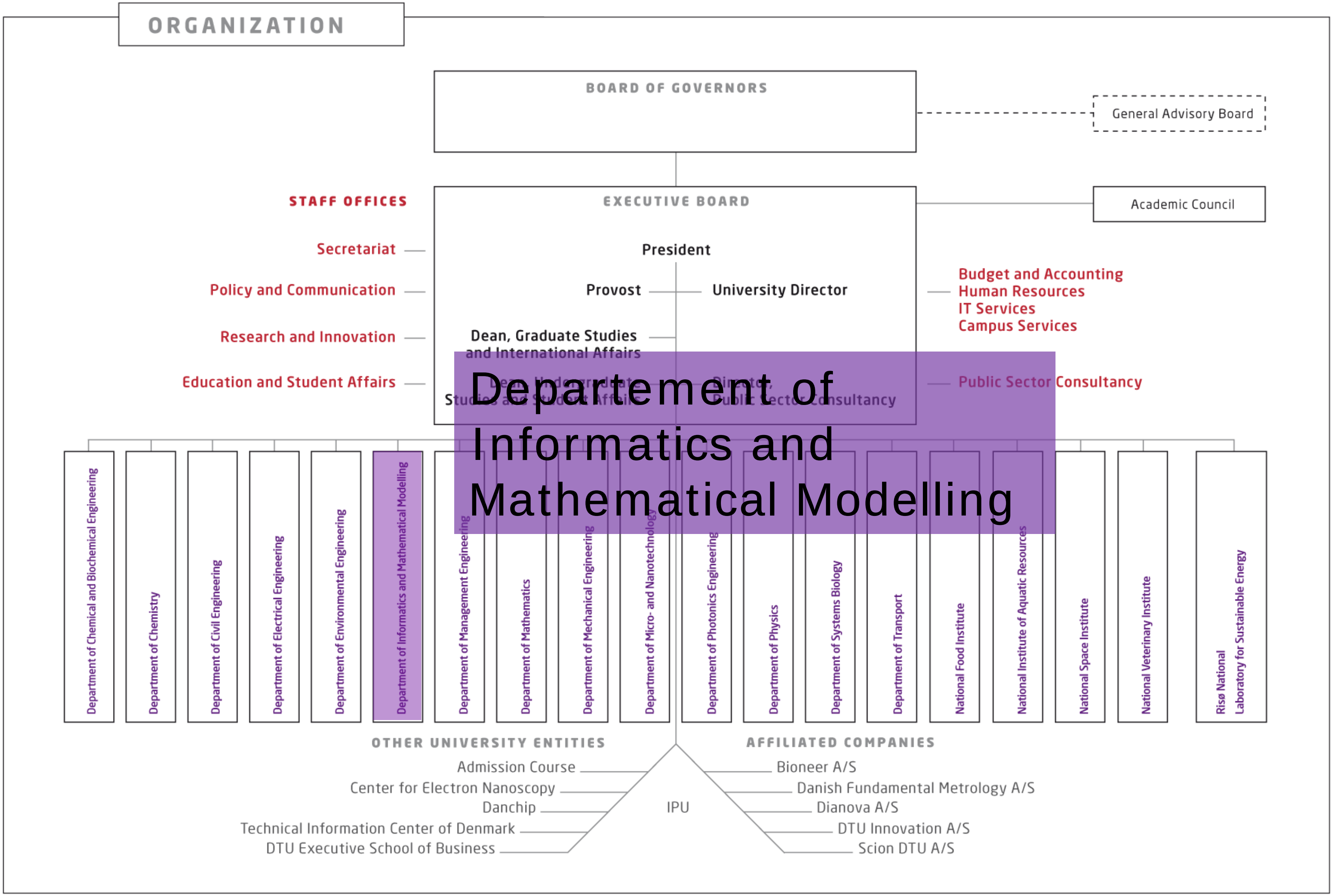
1,447 Faculty and research staff

1,583 Technical and administrative personnel in departments

1,491 Administration, campus service and PhD fellows

Finances

Income (2008): €469.2 million



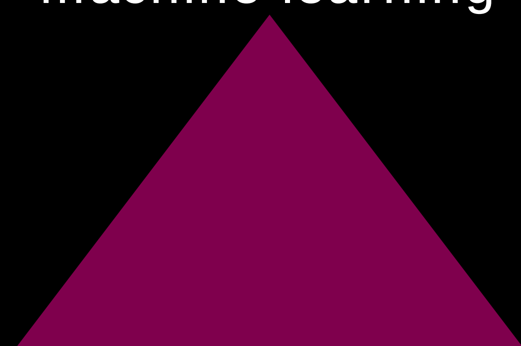
Departement of Informatics and Mathematical Modelling

Section for Cognitive Systems

Why do we do it? VISION

What do we do? MISSION

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



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

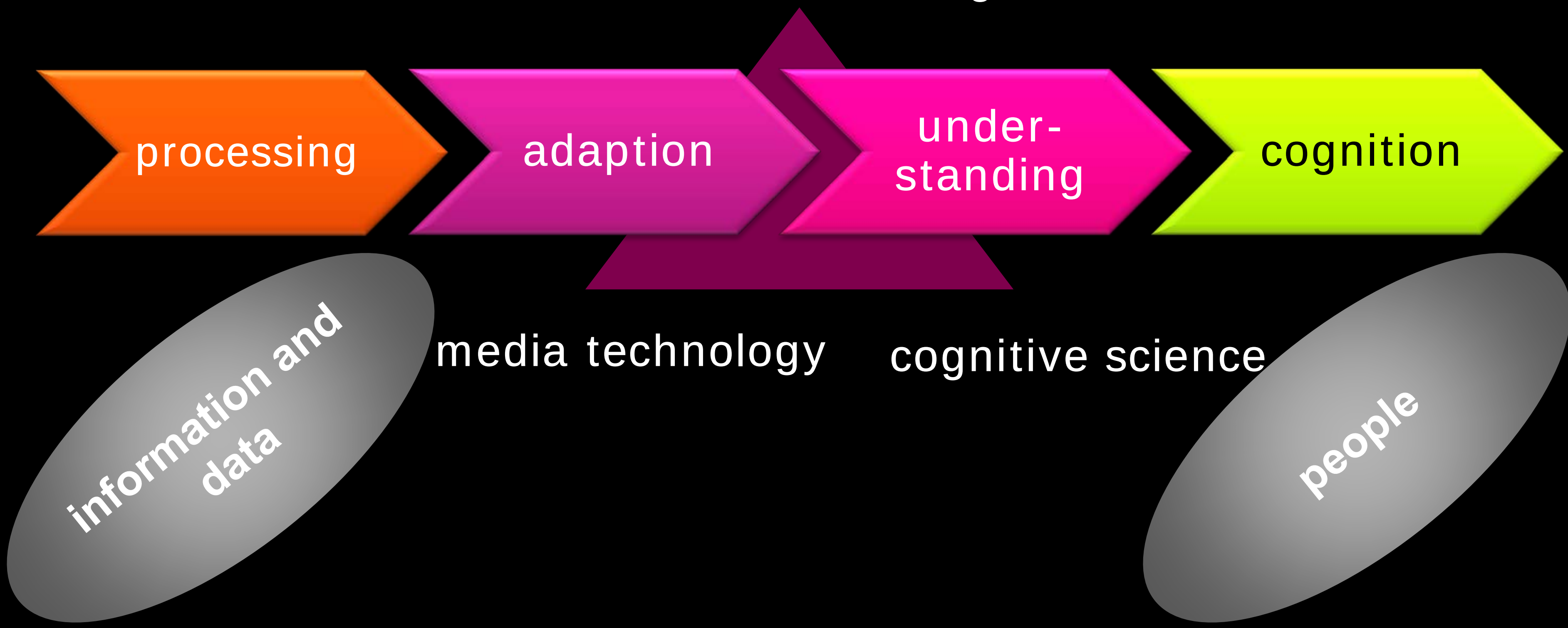
Allan Touring

Theory of computing 1940’es

Norbert Wiener

Cybernetics

1948



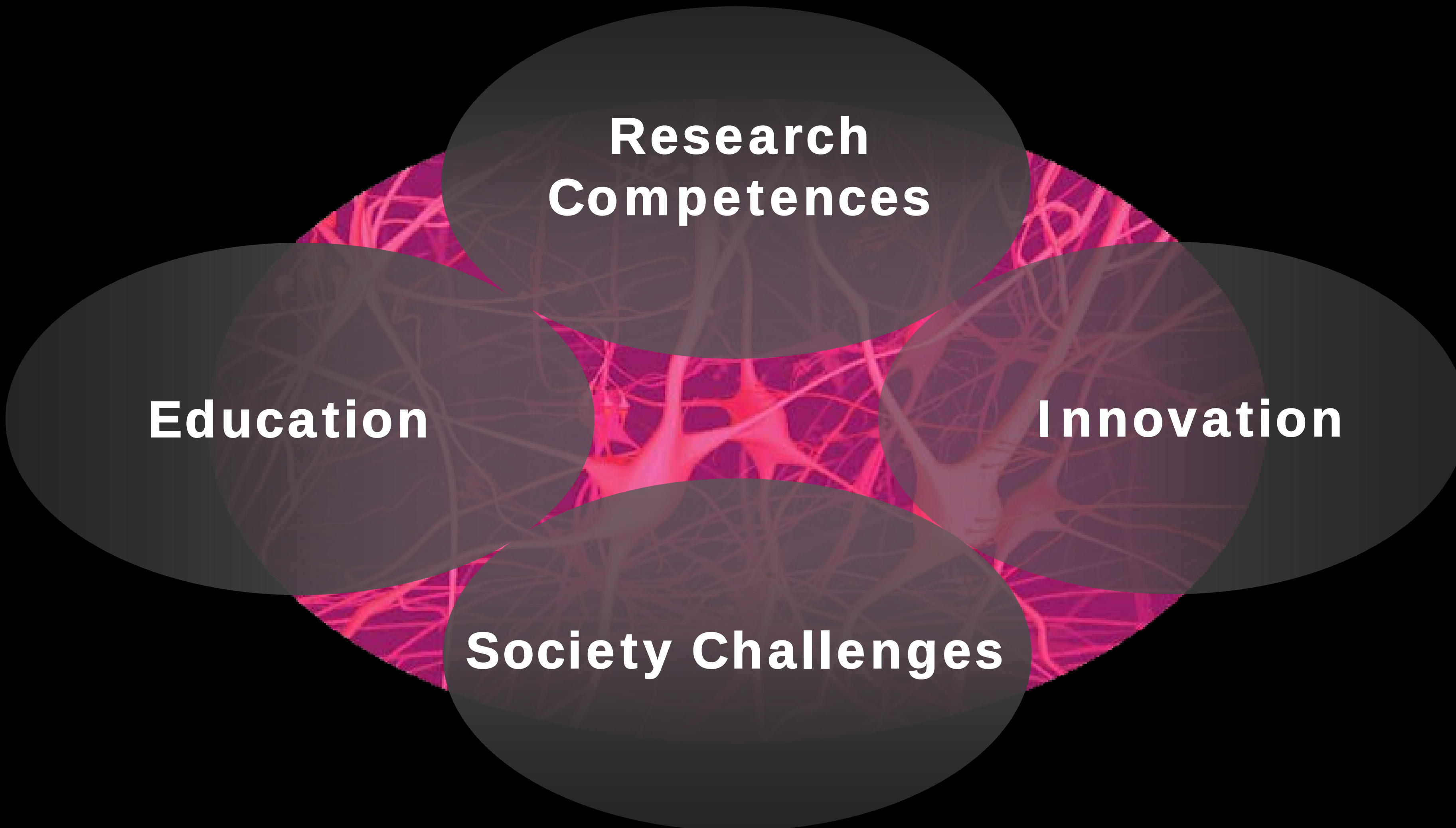
Mission

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



Society challenges

Future improvement in productivity and quality of life requires **organization and integration of internet-size data sets**

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

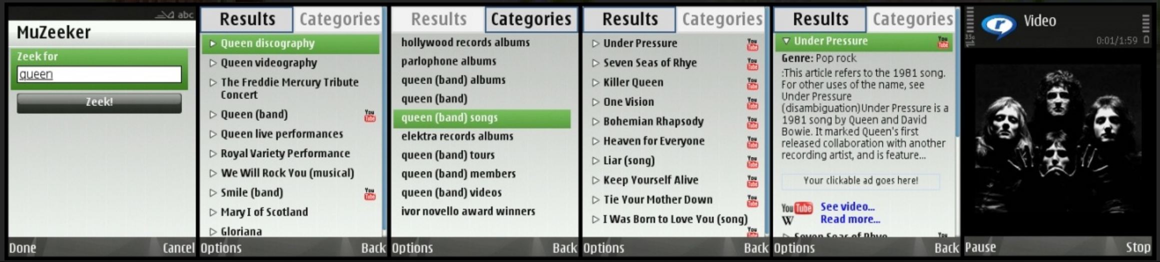
Brain modeling and mental decoding are crucial for augmented cognition, lifelong learning, and may revolutionize health services extraction of meaningful and actionable information from audio by ubiquitous learning from data

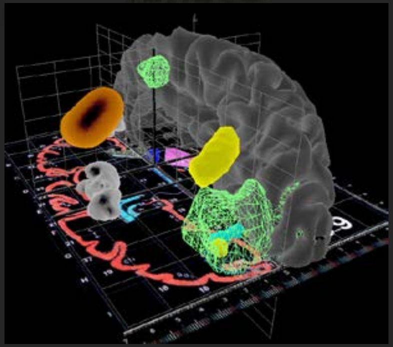
Research Competences

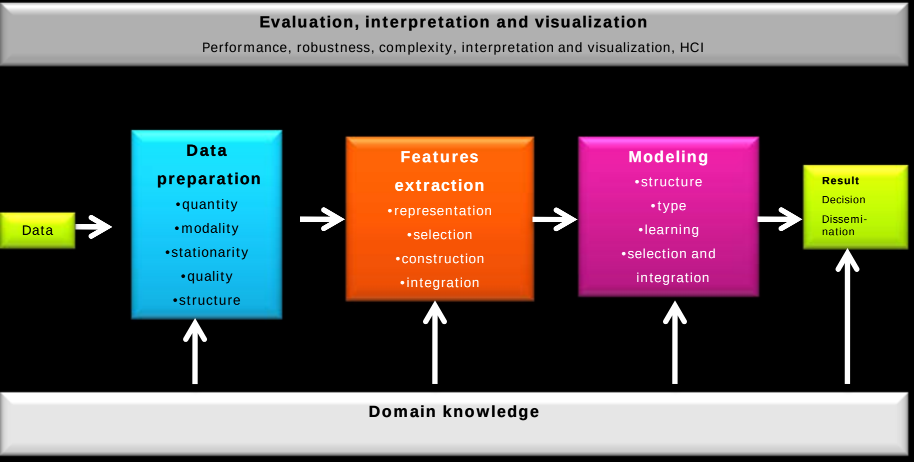
Media technology: mobile platforms, digital media, social networks, search, navigation, and semantics

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

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







Machine learning

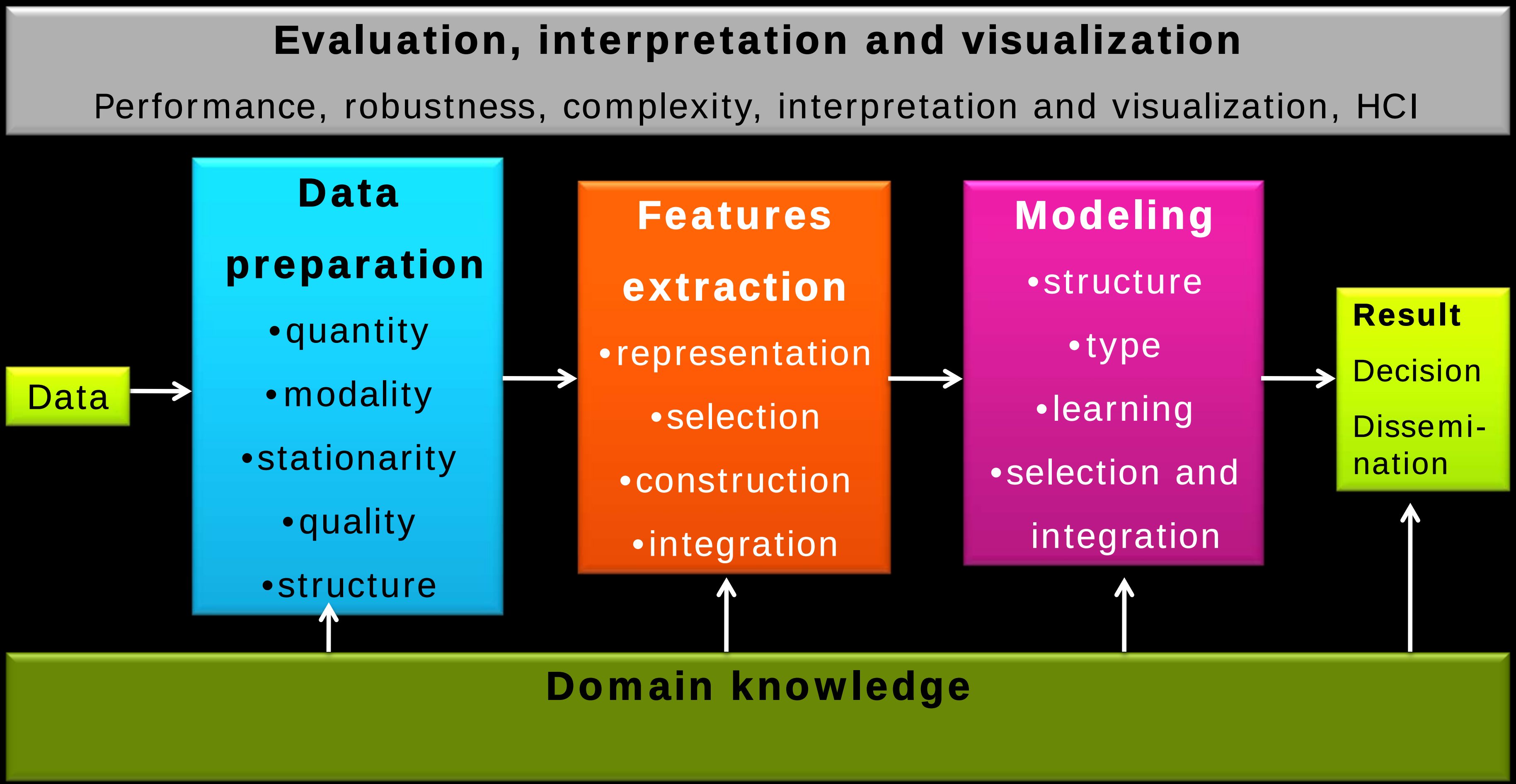
Statistical machine **learning abstracts data to active knowledge by identifying predictive relations** and has become a major driver of the knowledge society. Machine learning drives the Google economy, empowers bioinformatics, and enables mind reading in neuroimaging.

Our research in machine learning is rooted in statistics, including Bayesian and in resampling based methods, and has a strong algorithmic component. Past developments include ensembles, approximate inference, blind signal separation, and multi-way methods.

Current theoretical work concerns sparse representations, infinite models, multiway methods, and complex networks.

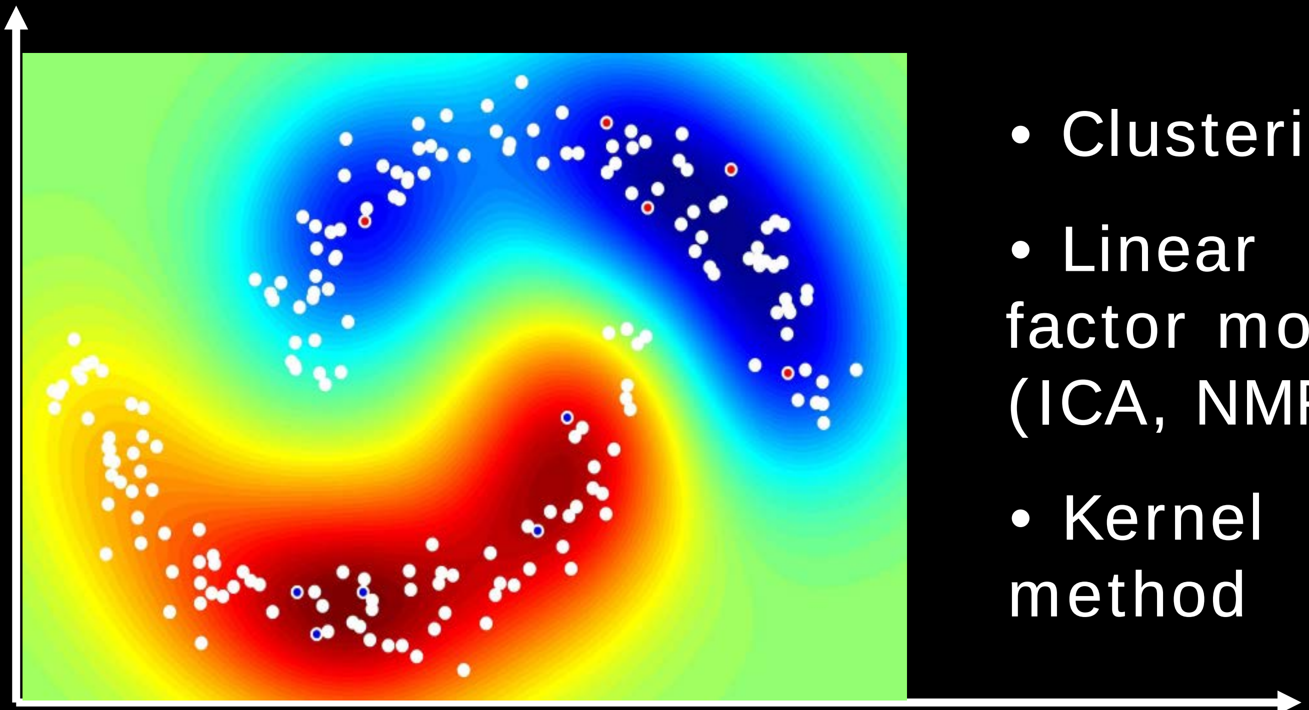
https://ml.imm.dtu.dk/

Data modeling framework



Unsupervised learning

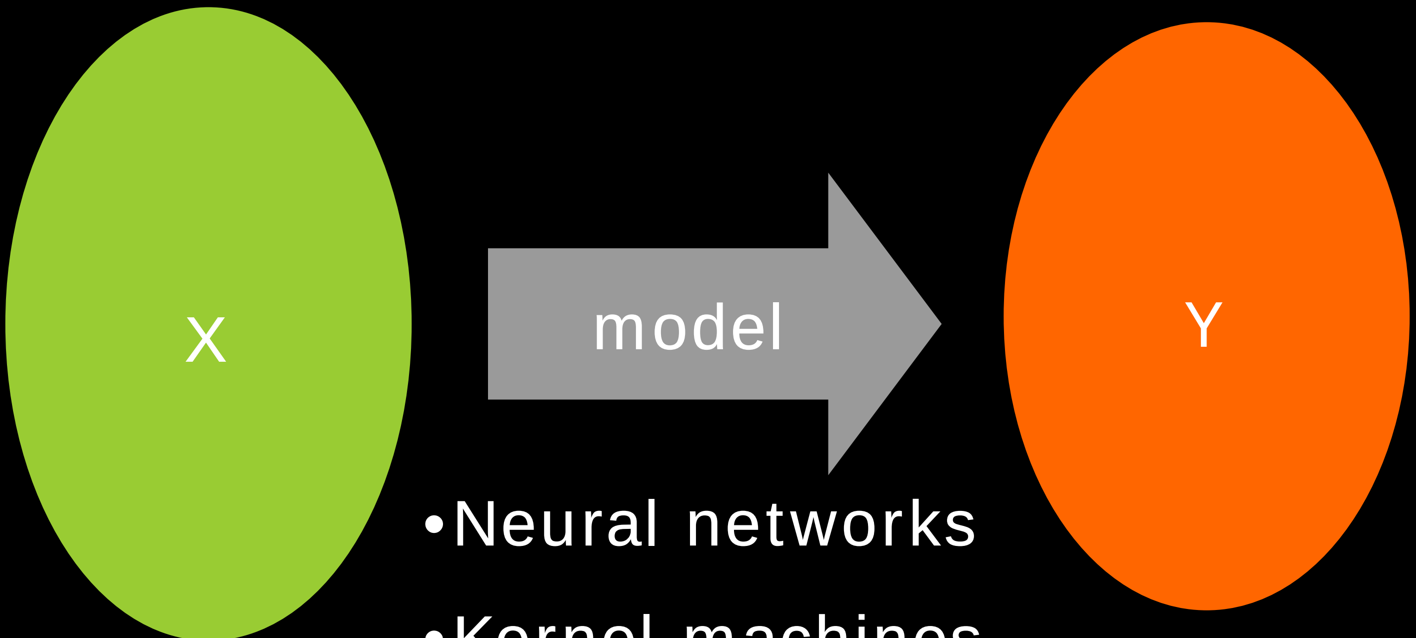
- Probabilistic modeling of structure in multivariate data
- Preprocessing, data reduction, outlier detection



- Clustering
- Linear factor models (ICA, NMF)
- Kernel method

Supervised learning

- Mapping between domains – from features to decision
- Based on a data set of simultaneous observations of X and Y



- Neural networks
- Kernel machines
- Bayesian learning

Semi-supervised learning

- Learning from labeled and unlabeled data
- Optimal use of inexpensive unlabeled data
- Quantification of robustness

Active learning

- Active learning - related method in which samples are initially unknown
  - Labelling may be expensive or laborsome
  - Methods should decide which samples help learning most

Huge demand for tools: organization, search, information enrichment

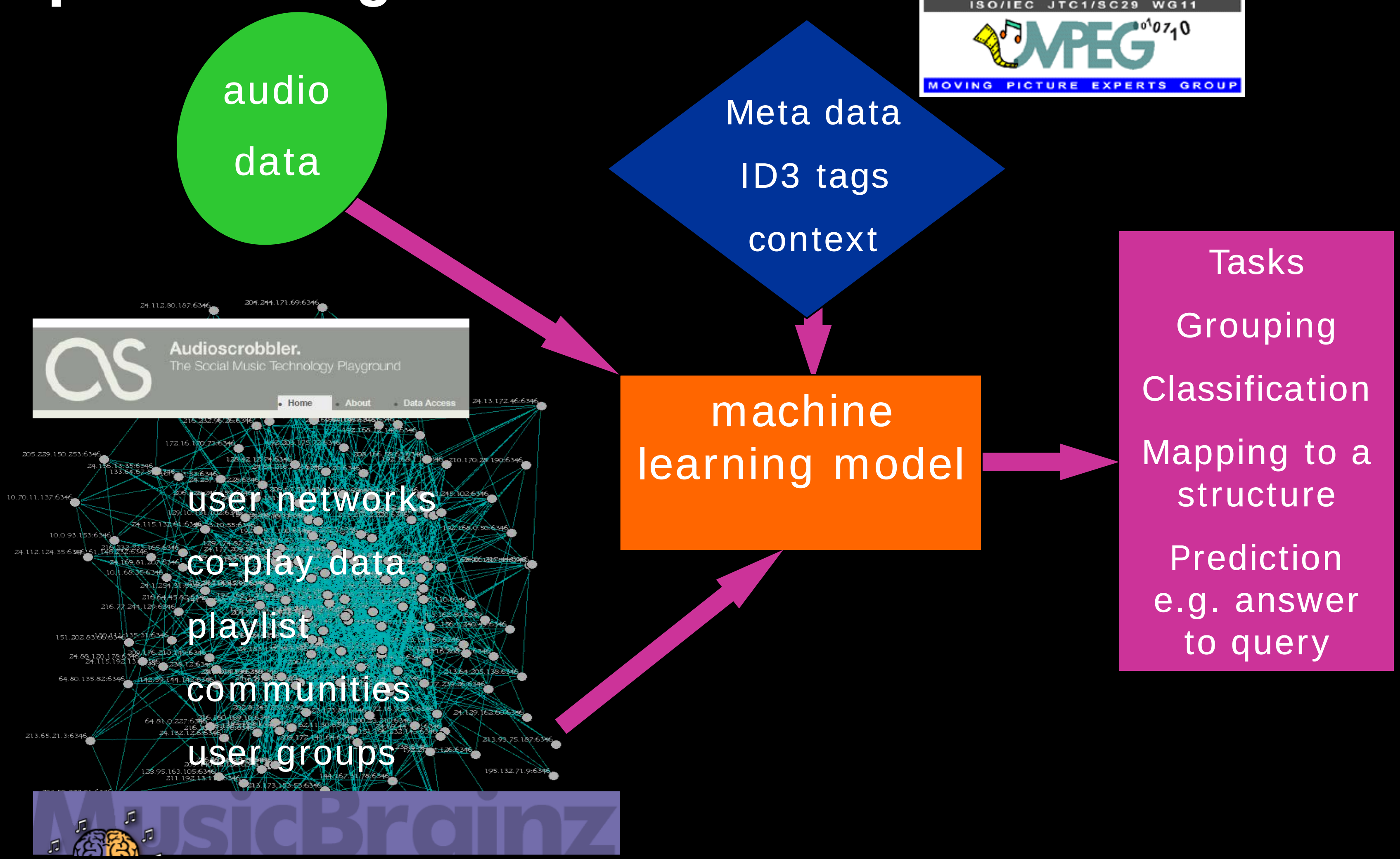
- Recommender systems (”taste prediction”)
- Playlist generation
- Finding similarity in music (e.g., genre classification, instrument classification, etc.)
- Meta data generation (emotional tags, labels)
- Newscast transcription/search
- Music transcription/search
- Audio separation

Intelligent Sound Project

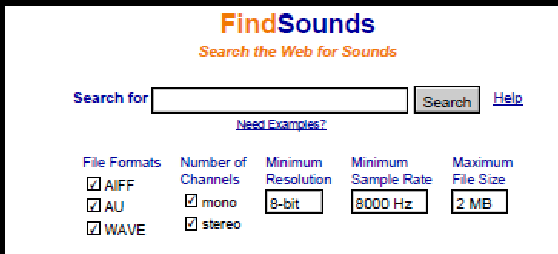
- FTP project 2005-2009
- 14 mil DKK
- Participants: DTU and Aalborg University

www.intelligentsound.org

Machine learning in sound information processing



Specialized search and music organization



last.fm the social music revolution

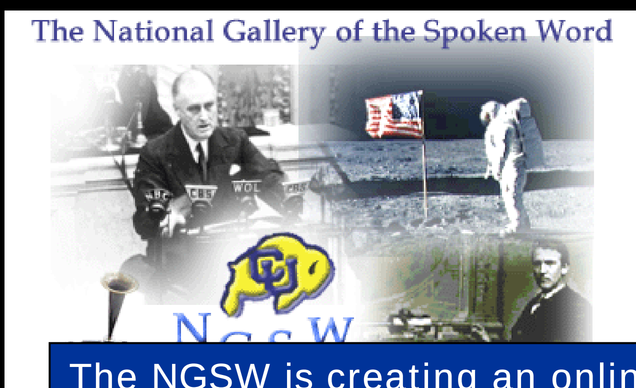
Using social network analysis

allmusic AMG

Explore by genre, mood, theme, country, instrument

Query by humming

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"

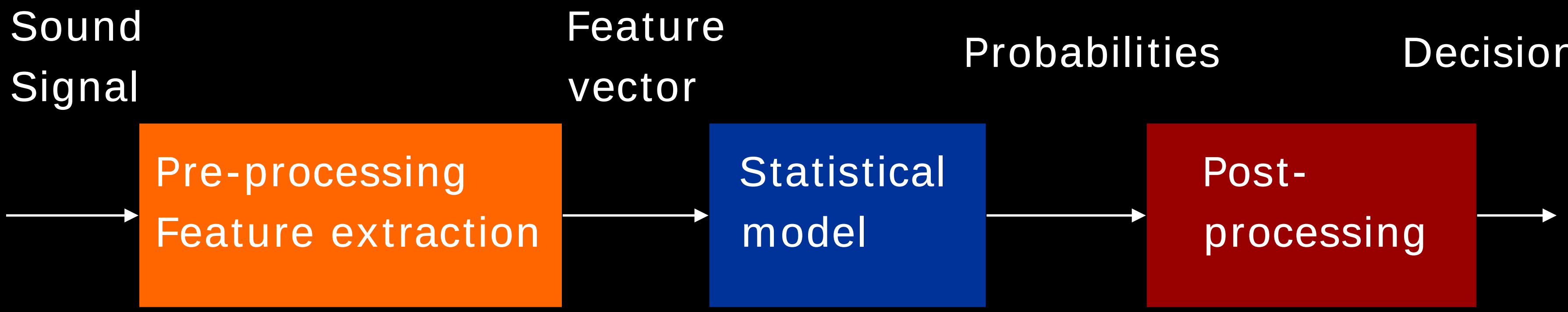

WINAMP
File
Play
Options
View
Help
Detach Visualizer
Alternative Rock
Blues
Classical
Country
Dance
Folk
Jazz
Opera & Vocal
Pop
R&B
Rap & Hip-Hop
Rock
Prev
Next
Random
0:39
MUDDY WATERS - GOT MY MOJO WORKING (3:00)

Meta data generation: 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 or blues.



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, $r=1,..,80$

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 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

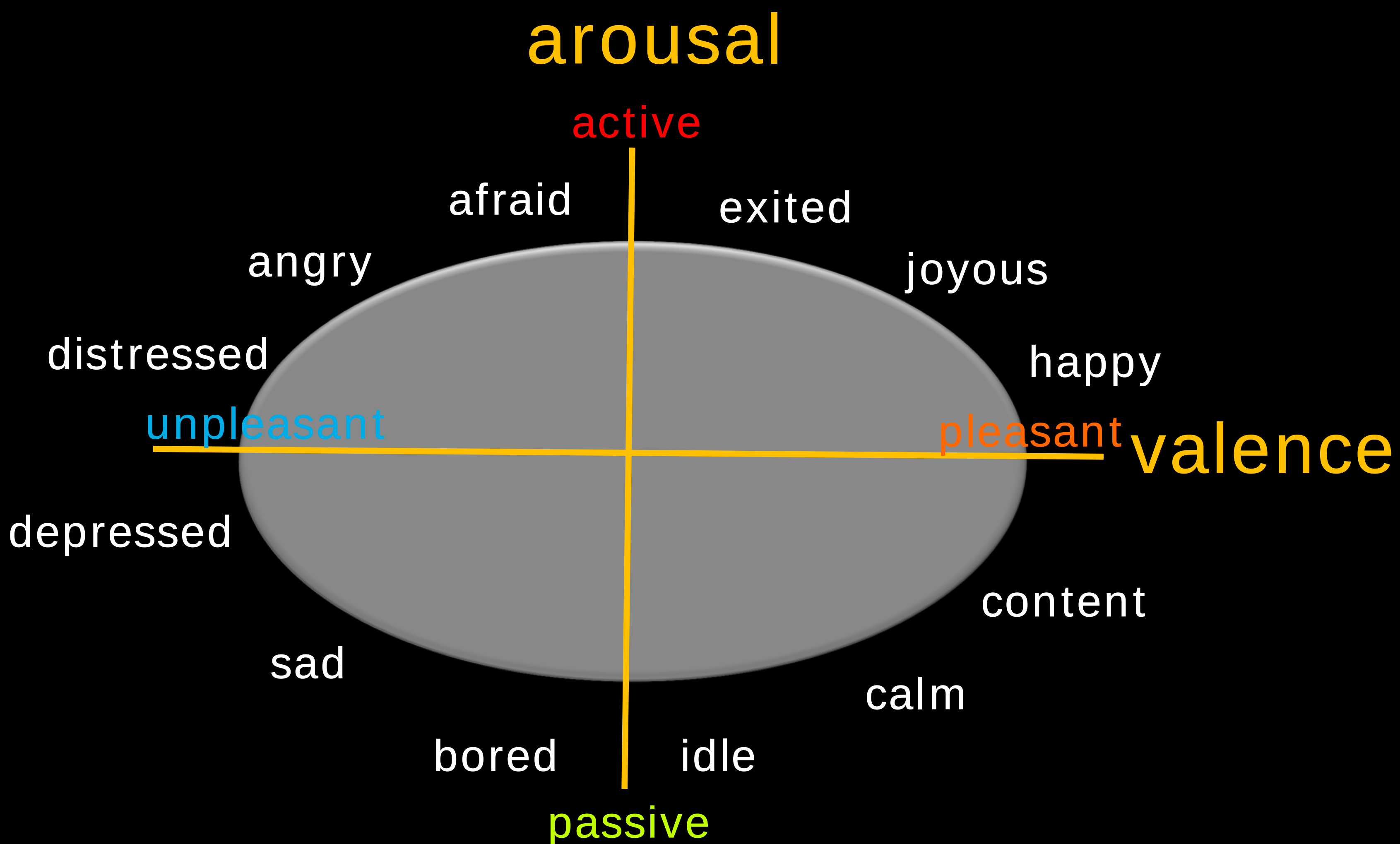
Best 11-genre confusion matrix

	Alternative	Country	Easy-listening	Electronica	Jazz	Latin	Pop&Dance	Rap&Hiphop	RB&Soul	Reggae	Rock
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	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	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

11-genre problem (some overlap) : 50% error

human error about 43%

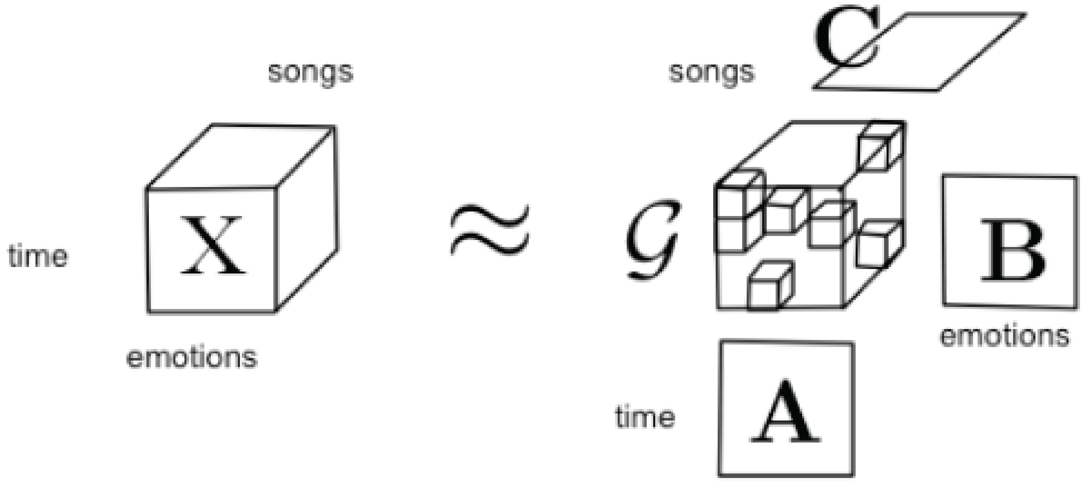
Emotional spaces



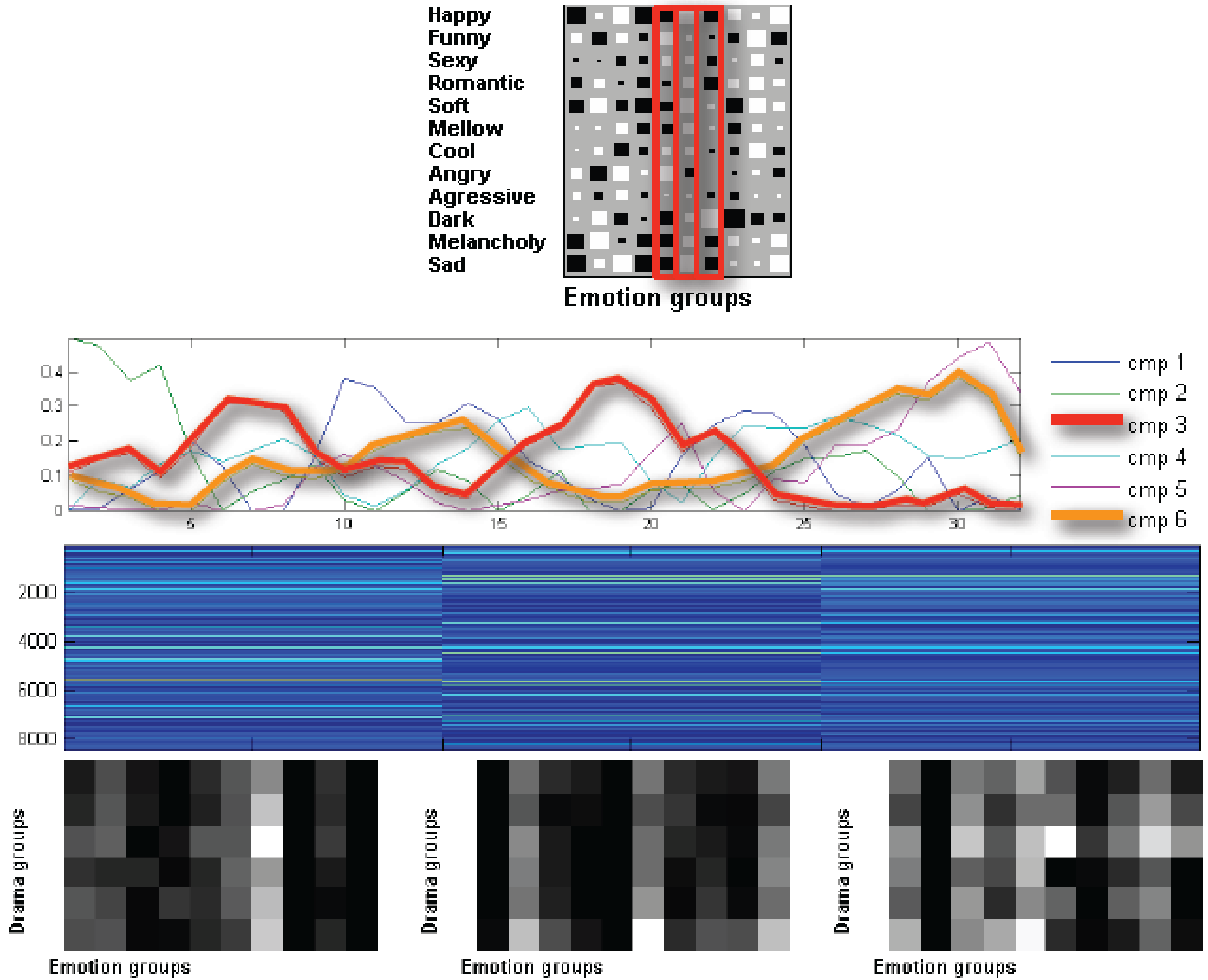
J. A. Russel: "A Circumplex Model of Affect," *Journal of Personality and Social Psychology*, 39(6):1161, 1980

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

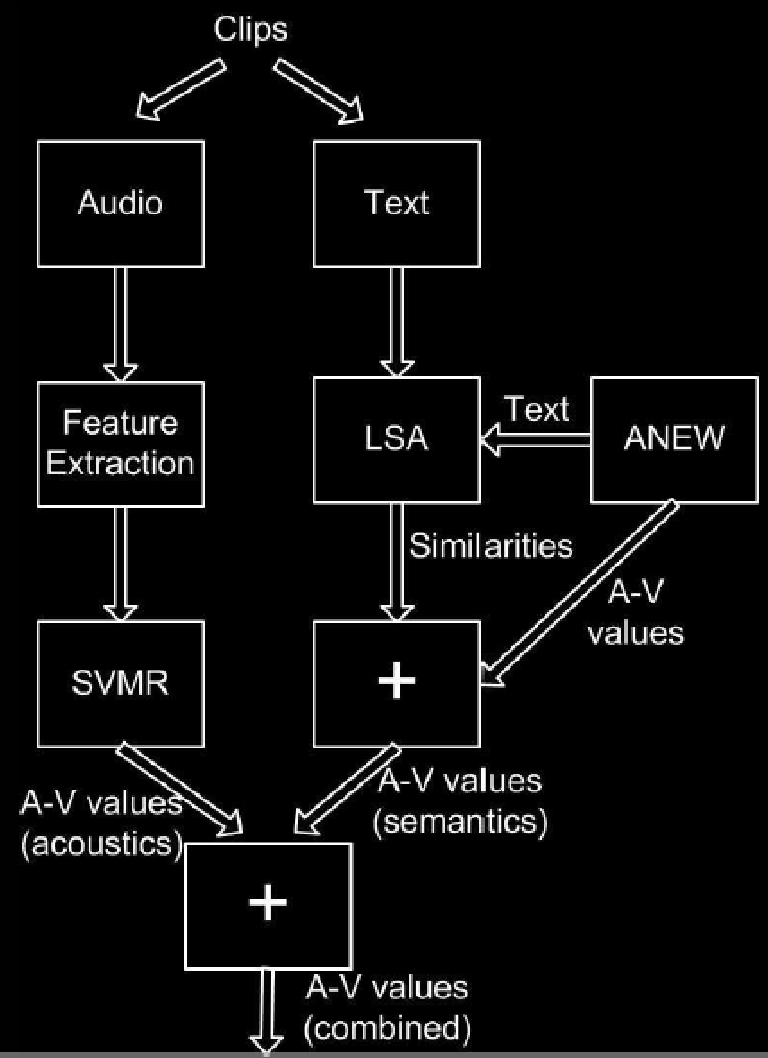
Emotion modelling



$$x_{ijk} \approx \mathcal{G} \times \mathbf{A} \times \mathbf{B} \times \mathbf{C} = \sum_{lmn}^{LMN} g_{lmn} a_{il} b_{jm} c_{kn}$$

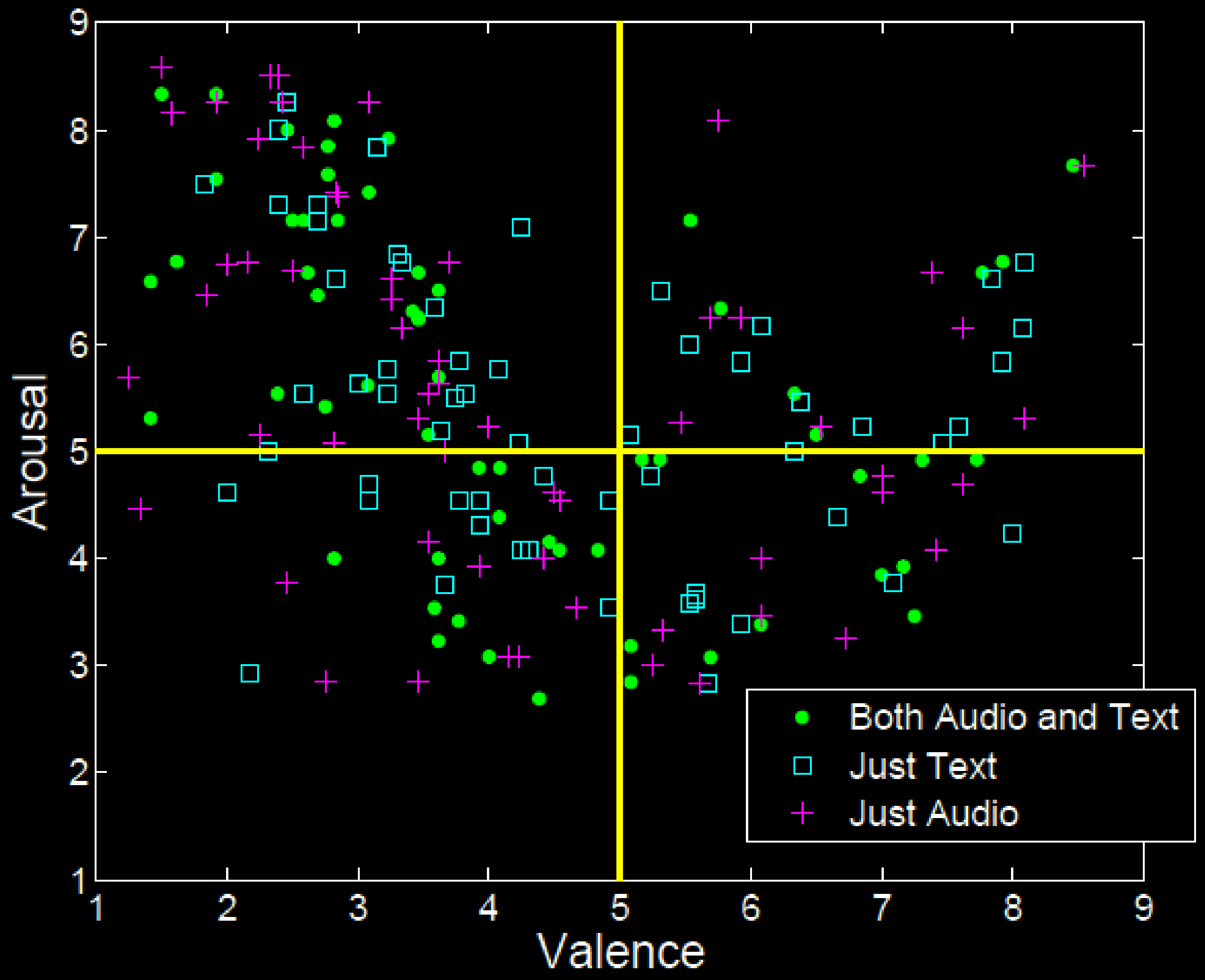

Happy
Funny
Sexy
Romantic
Soft
Mellow
Cool
Angry
Agressive
Dark
Melancholy
Sad
Emotion groups
0.4
0.3
0.2
0.1
0
5
10
15
20
25
30
cmp 1
cmp 2
cmp 3
cmp 4
cmp 5
cmp 6
2000
4000
6000
8000
Drama groups
Emotion groups
Drama groups
Emotion groups
Drama groups
Emotion groups

Semantics and Acoustics Features for Emotional Recognition in Speech



S. Karadogan, J. Larsen, Combining Semantics and Acoustics Features for Valence and Arousal Recognition in Speech, *CIP 2012*.

Semantics and Acoustics Features for Emotional Recognition in Speech



The valence dimension is more about what we say, while the arousal dimension is more about how we say it

		Weights (we_sem / we_ac)	Combined Result
Valence	MAE	0.80 / 0.20	1.40
	RMSE	0.85 / 0.15	1.77
Arousal	MAE	0 / 1	1.28
	RMSE	0.20 / 0.80	1.52

Audio separation

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

Semi-supervised learning methods

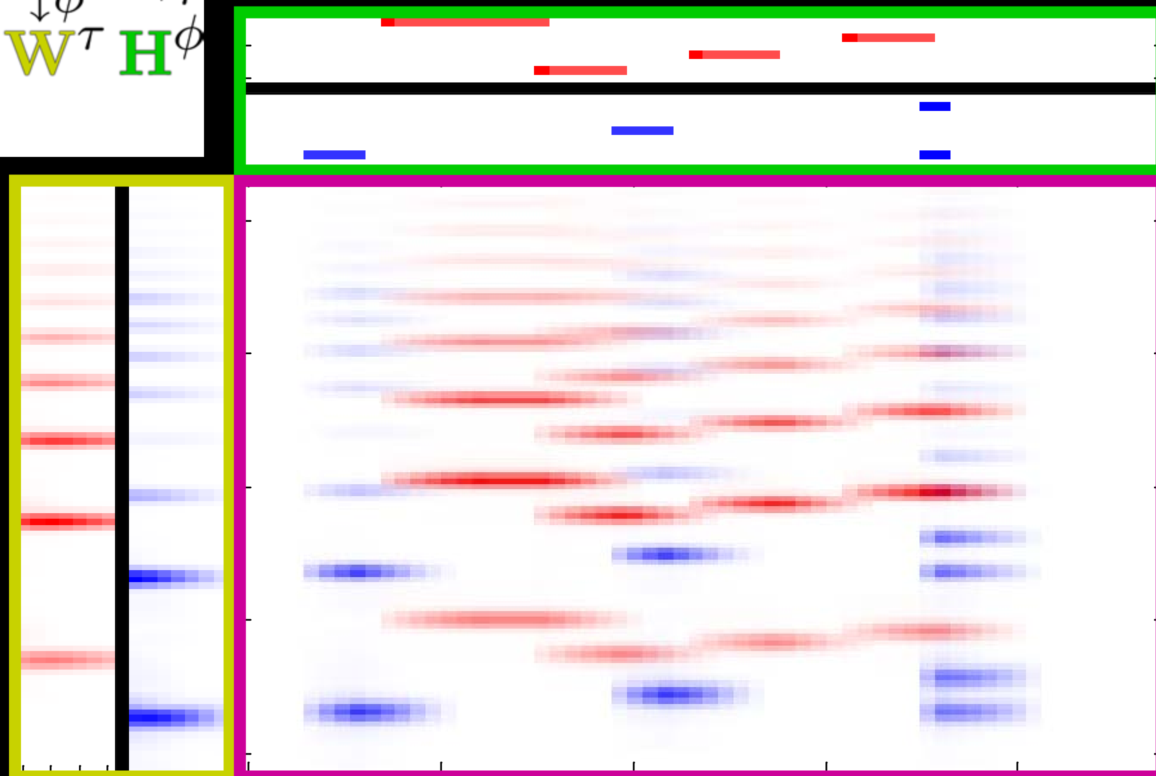
Pedersen, M. S., Larsen, J., Kjems, U., Parra, L. C., *A Survey of Convolutive Blind Source Separation Methods*, Springer Handbook of Speech, Springer Press, 2007

Nonnegative matrix factor 2D deconvolution

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

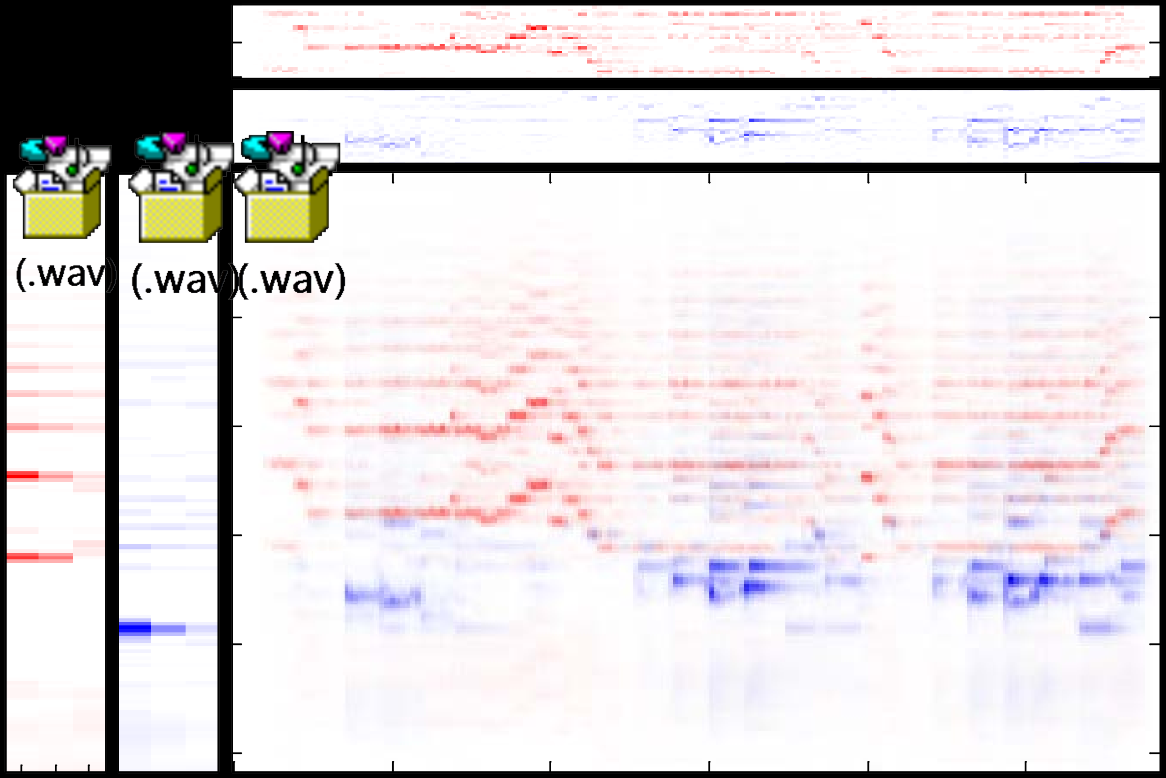
time

pitch

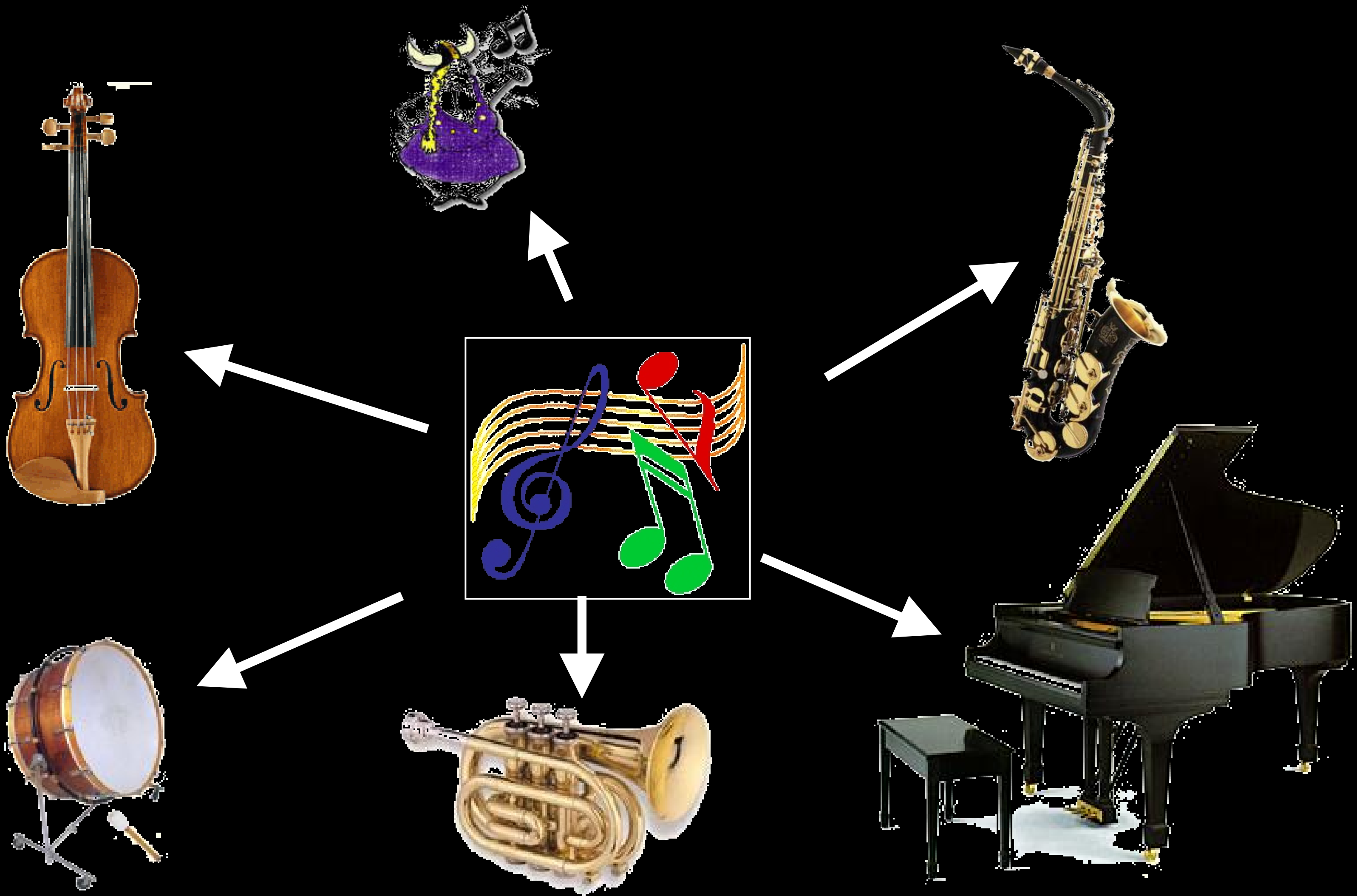


M. N. Schmidt, M. Mørup *Nonnegative Matrix Factor 2-D Deconvolution for Blind Single Channel Source Separation*, ICA2006, 2006. Demo also available.

Demonstration of the 2D convolutive NMF model



Separating music into basic components



Separating music into basic components

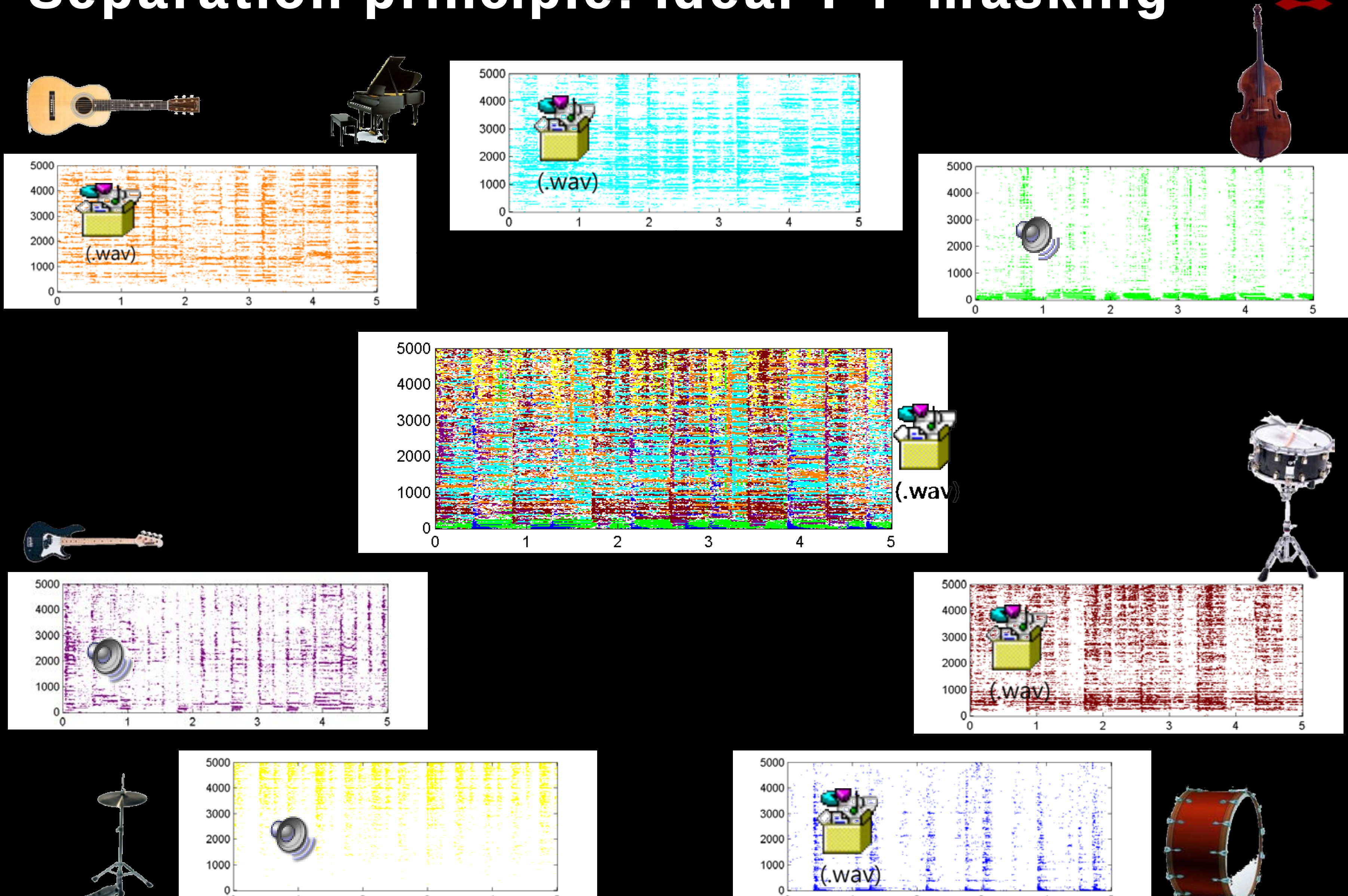
- Combined ICA and masking

- 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

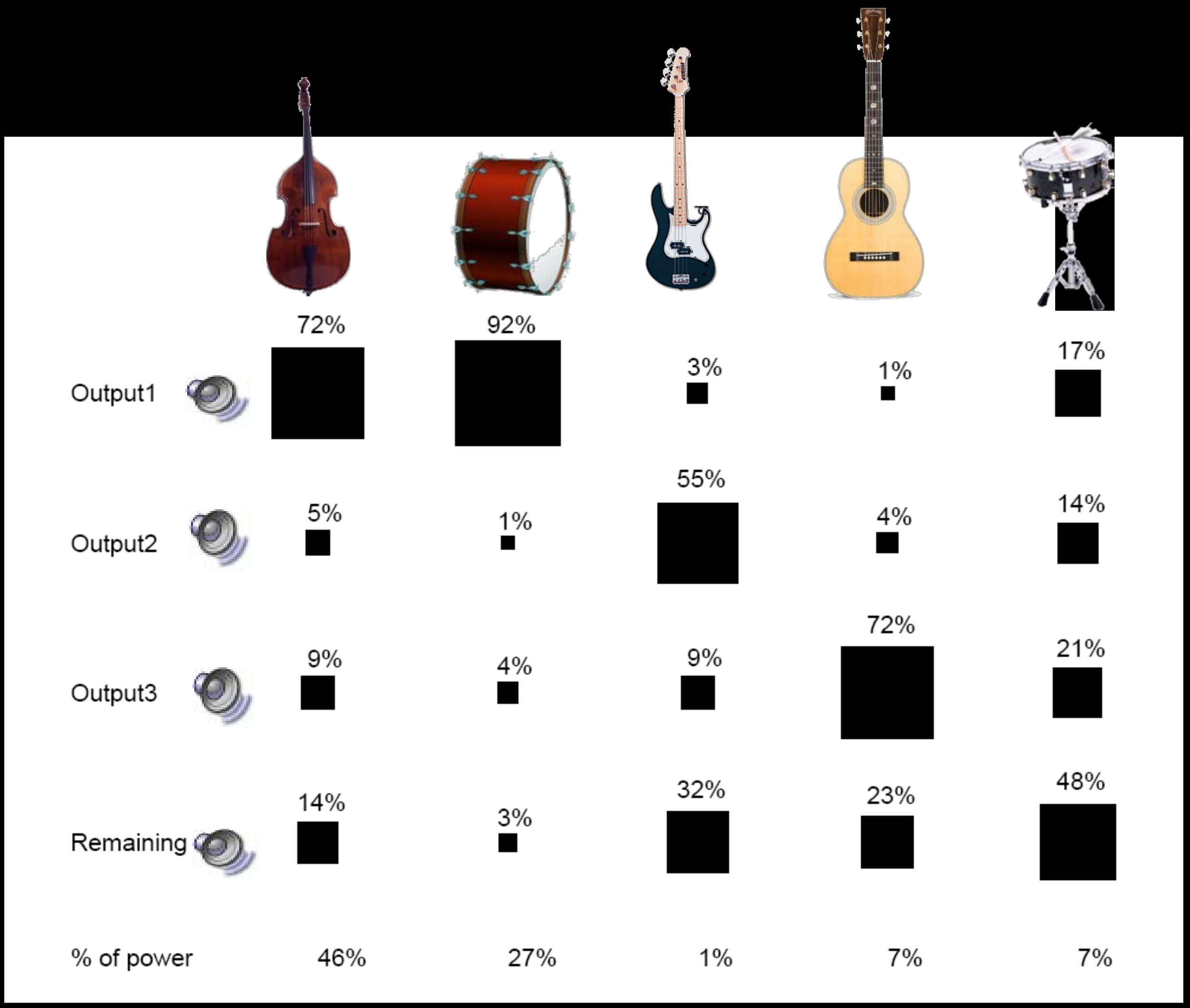
- Stereo recording of the music piece is available.
- The instruments are separated to some extent in time and in frequency, i.e., the instruments are sparse in the time-frequency (T-F) domain.
- The different instruments originate from spatially different directions.

Separation principle: ideal T-F masking

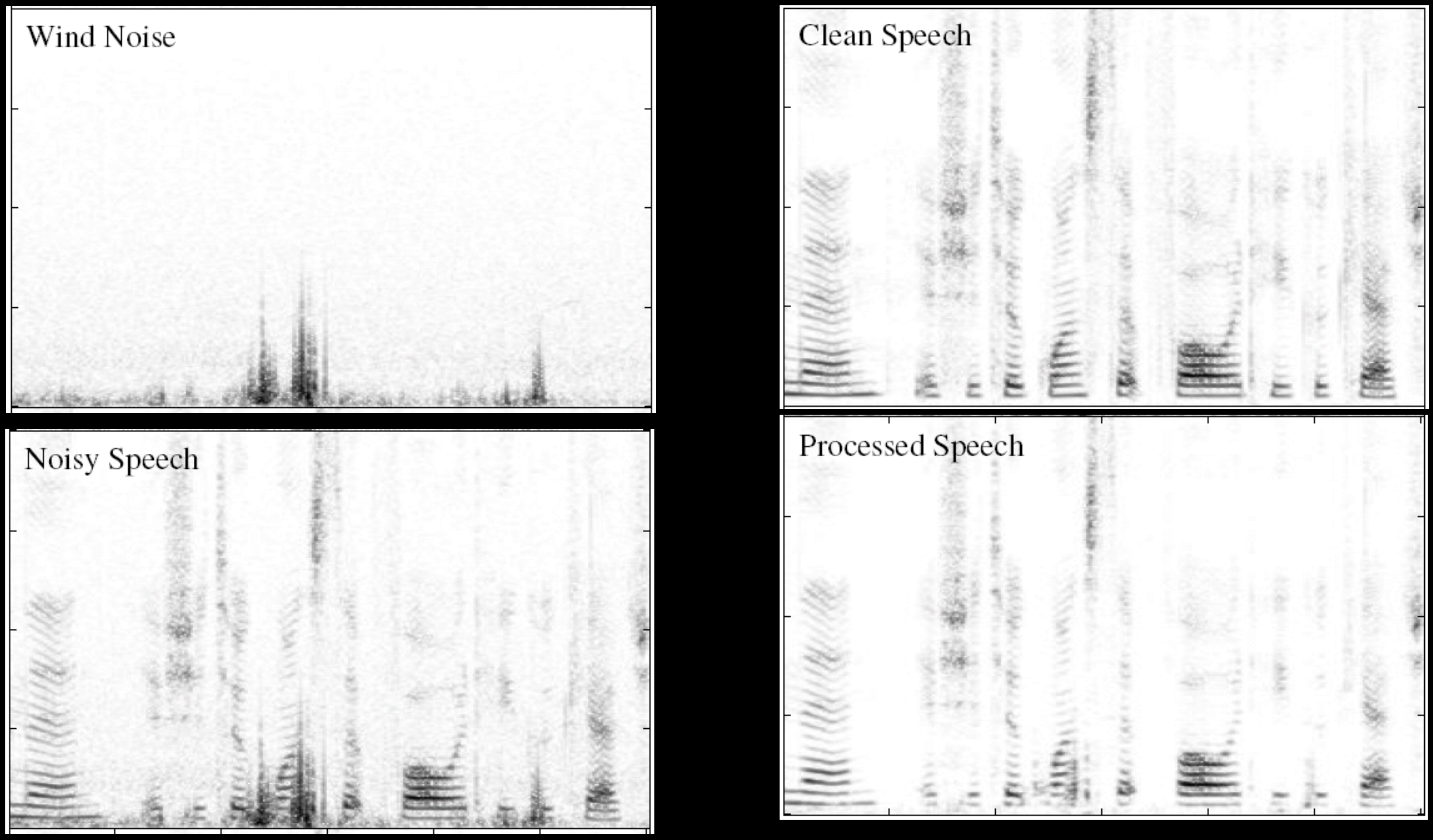


Results

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



Wind noise reduction



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

Single channel separation: Sparse NMF decomposition

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

original

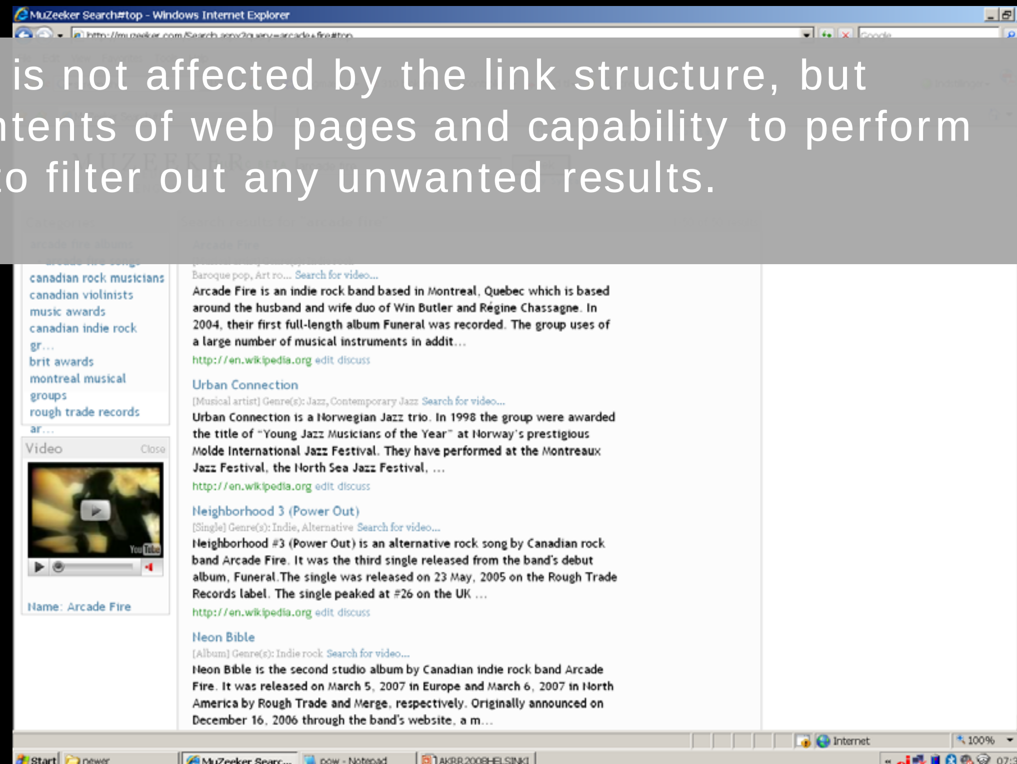
cleaned

alternative method (qualcom)

A cognitive search engine - MuZeeker

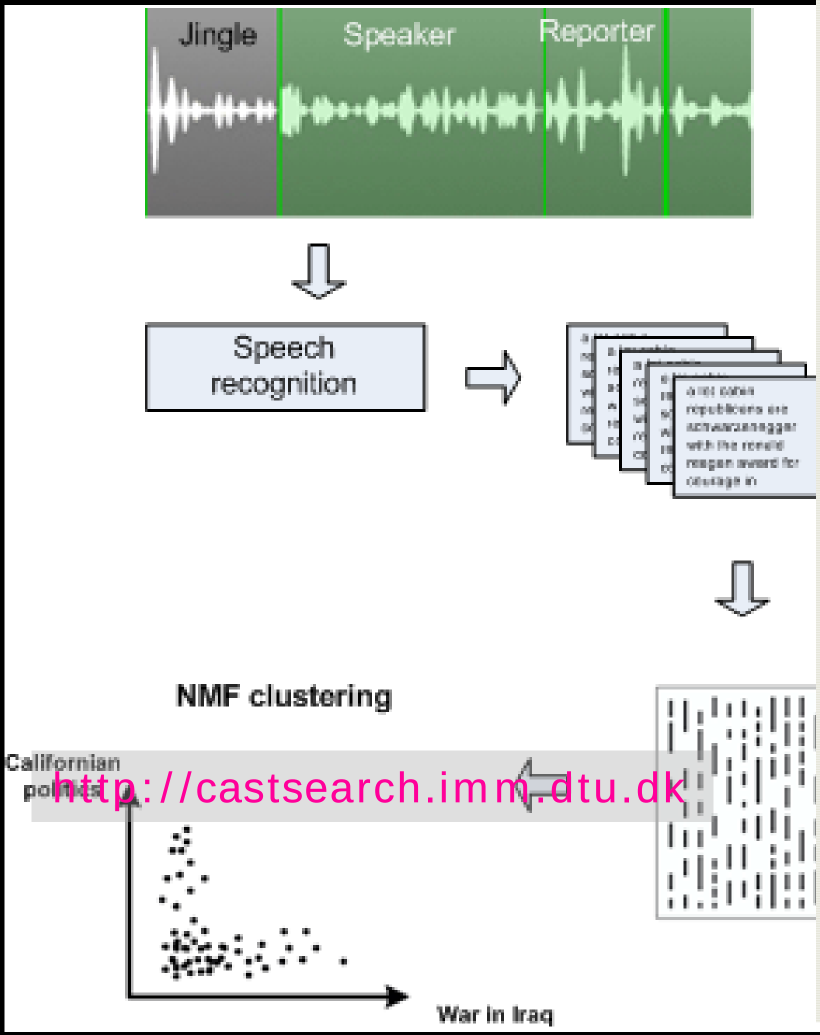
Idea is to create a search engine that is not affected by the link structure, but instead based solely on the actual contents of web pages and capability to perform categorizing. This making it possible to filter out any unwanted results.

- Wikipedia based common
- Wikipedia used as a proxy for the music users mental model
- Implementation: Filter retrieval using Wikipedia's article/ categories
- Prefernce to MuZeeker over Google in task solvingf

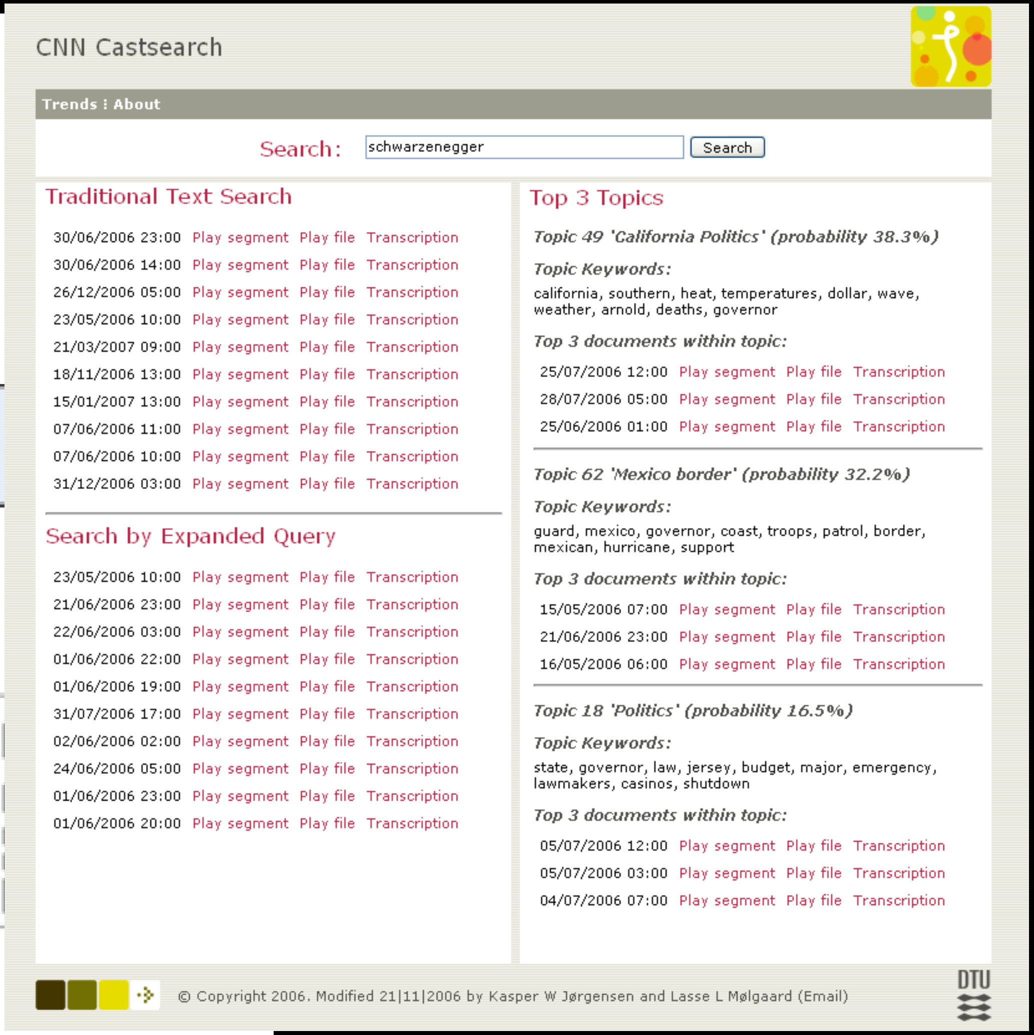


-

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

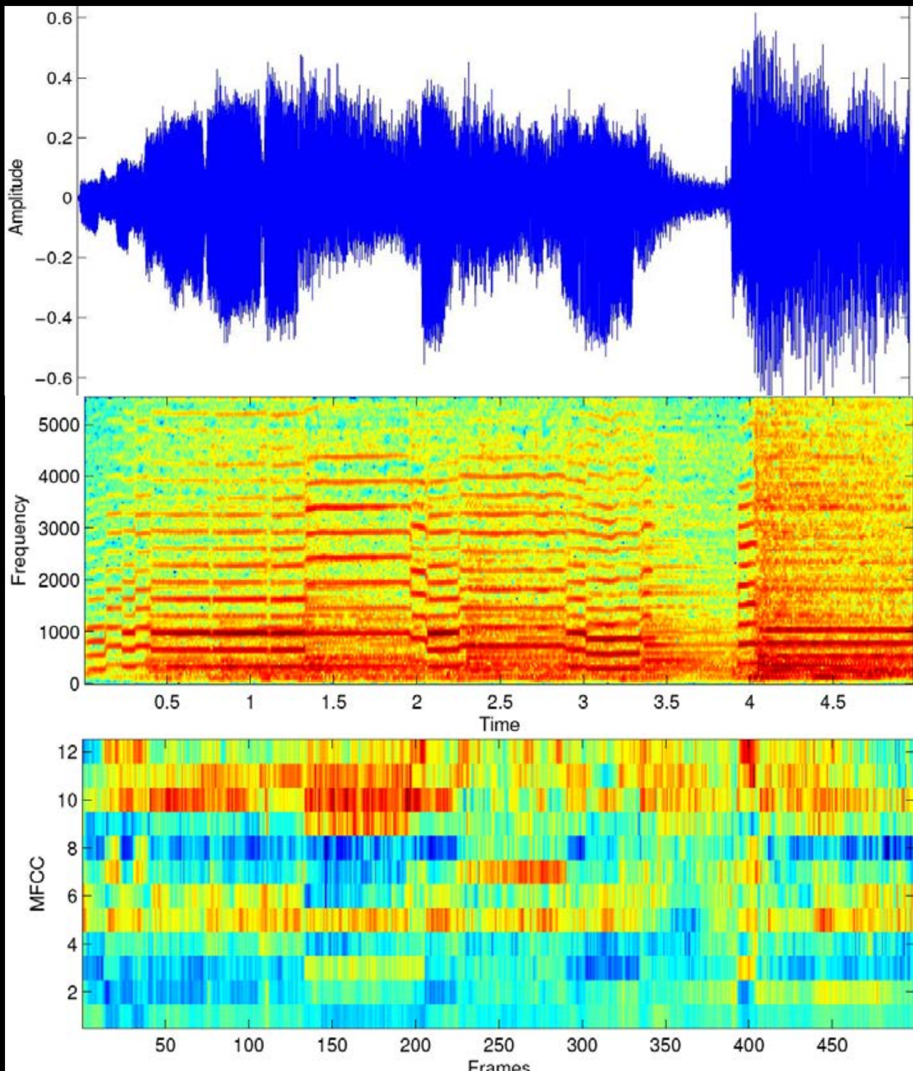


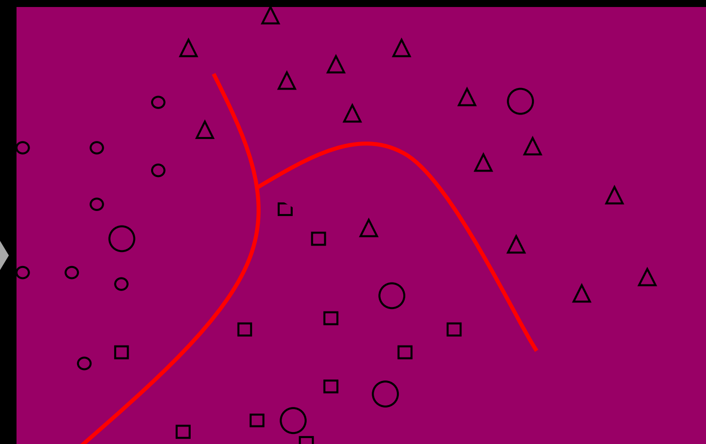
http://castsearch.imm.dtu.dk

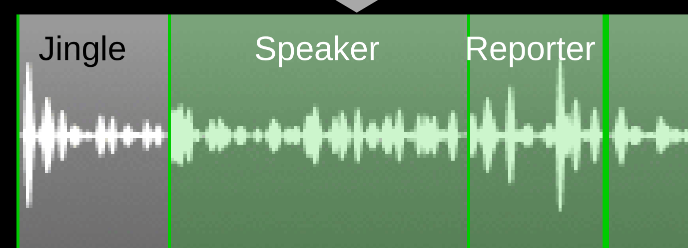


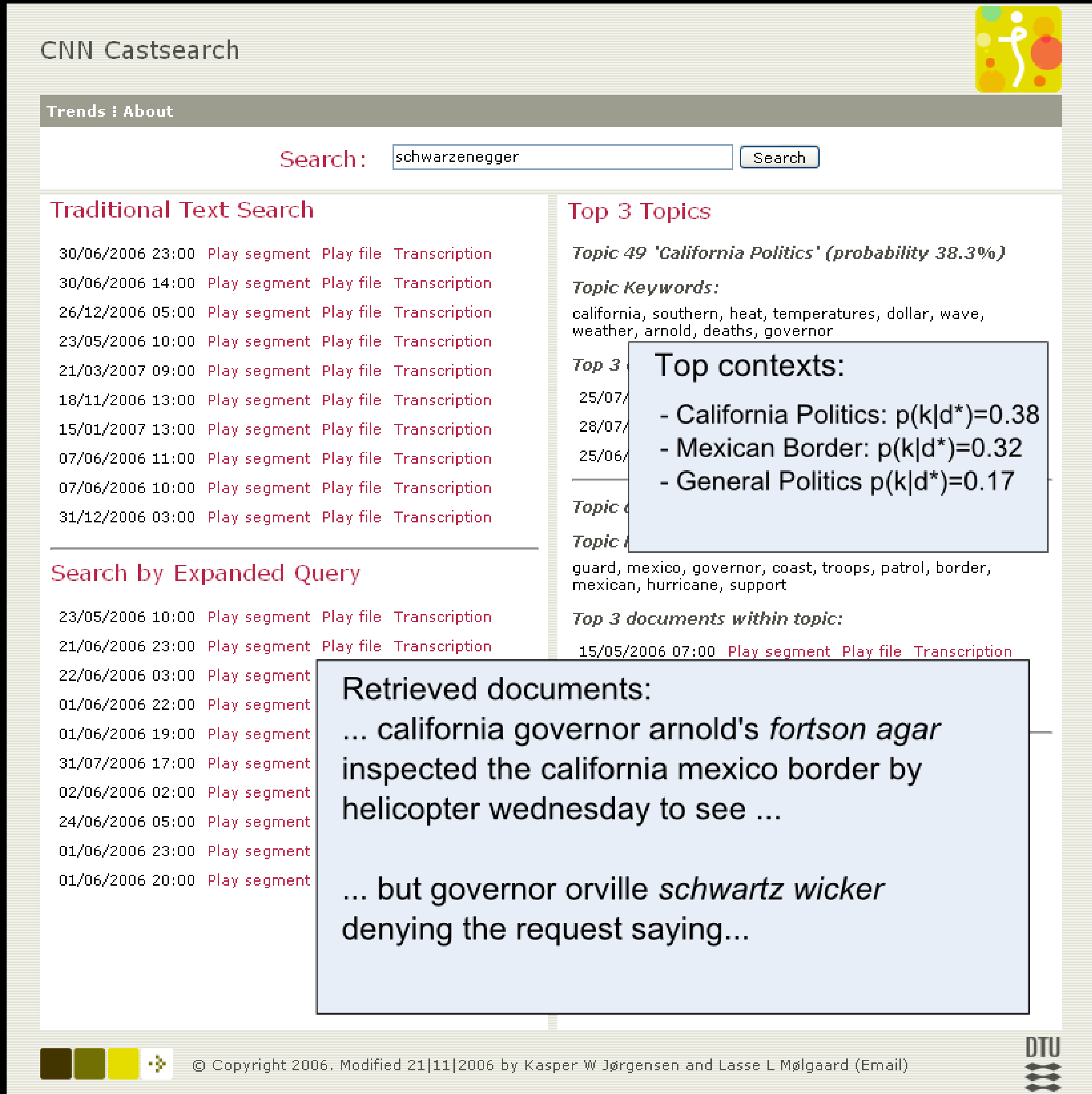
Ref: Lasse Mølgaard, Kasper Jørgensen, Lars Kai Hansen: "CASTSEARCH: Context based Spoken Document Retrieval," ICASSP2007

Sound segmentation








CNN Castsearch
Trends : About
Search: schwarzenegger Search
Traditional Text Search
30/06/2006 23:00 Play segment Play file Transcription
30/06/2006 14:00 Play segment Play file Transcription
26/12/2006 05:00 Play segment Play file Transcription
23/05/2006 10:00 Play segment Play file Transcription
21/03/2007 09:00 Play segment Play file Transcription
18/11/2006 13:00 Play segment Play file Transcription
15/01/2007 13:00 Play segment Play file Transcription
07/06/2006 11:00 Play segment Play file Transcription
07/06/2006 10:00 Play segment Play file Transcription
31/12/2006 03:00 Play segment Play file Transcription
Search by Expanded Query
23/05/2006 10:00 Play segment Play file Transcription
21/06/2006 23:00 Play segment Play file Transcription
22/06/2006 03:00 Play segment
01/06/2006 22:00 Play segment
01/06/2006 19:00 Play segment
31/07/2006 17:00 Play segment
02/06/2006 02:00 Play segment
24/06/2006 05:00 Play segment
01/06/2006 23:00 Play segment
01/06/2006 20:00 Play segment
Top 3 Topics
Topic 49 'California Politics' (probability 38.3%)
Topic Keywords:
california, southern, heat, temperatures, dollar, wave, weather, arnold, deaths, governor
guard, mexico, governor, coast, troops, patrol, border, mexican, hurricane, support
Top 3 documents within topic:
15/05/2006 07:00 Play segment Play file Transcription
Top contexts:
- California Politics: p(k|d*)=0.38
- Mexican Border: p(k|d*)=0.32
- General Politics p(k|d*)=0.17
Retrieved documents:
... california governor arnold's *fortson agar* inspected the california mexico border by helicopter wednesday to see ...
... but governor orville *schwartz wicker* denying the request saying...
© Copyright 2006. Modified 21|11|2006 by Kasper W Jørgensen and Lasse L Mølgaard (Email)
DTU

AV integration

Acoustic epe

+ Visual ete

= perceptual eke / ete

Vision influences auditory perception!

Cognitive AV integration

Purpose

To study AV integration and how it is influenced by physical and cognitive factors

- Behavioral experiments
  - Reveal the subjective audiovisual percept
- EEG
  - reveals the electro-physiological correlates of AV integration
- Mathematical modeling
  - Reveals the brain’s assumptions, goals and flaws in the integration of information across the senses

Research and innovation projects

2009 2014

Danish Sound Technology Network. Supported by DASTI. 14 MDKK + 8 MDKK (15 MDKK)

2012 2015

CoSound - a cognitive systems approach to enriched and actionable information from audio streams. Supported by the Danish Council for Strategic Research. 17.5 MDKK (6 MDKK)

CoSound

CoSound is a multi-discipilnary strategic research project addressing societal challenges related to **productivity, communication and well-being**

Productivity, communication and well-being depends on digital media and the delivery of multimodal media information on many different platforms including TV, social, and mobile media.

Music and media consumption is in a revolution

Traditional business models in the music, audio and broadcast sectors are challenged; however, the ubiquitous digitalization of media, localization information, and human behaviors has a huge and disruptive potential to be explored in strategic research.

Audio information represents a separate challenge over other modalities (e.g. text or visual information) since it can be sensed and perceived as an abstract, emotional stream.

CoSound

DTU Informatics

Syntonetic

DR

Musikzonen

B&O

Queen Mary University of London

Geckon

Royal School of Library and Information Science

Hindenburg Systems

UCL

Aalborg University

Department of Arts and Cultural Studies, Copenhagen University

State and University Library

University of Glasgow

CoSound

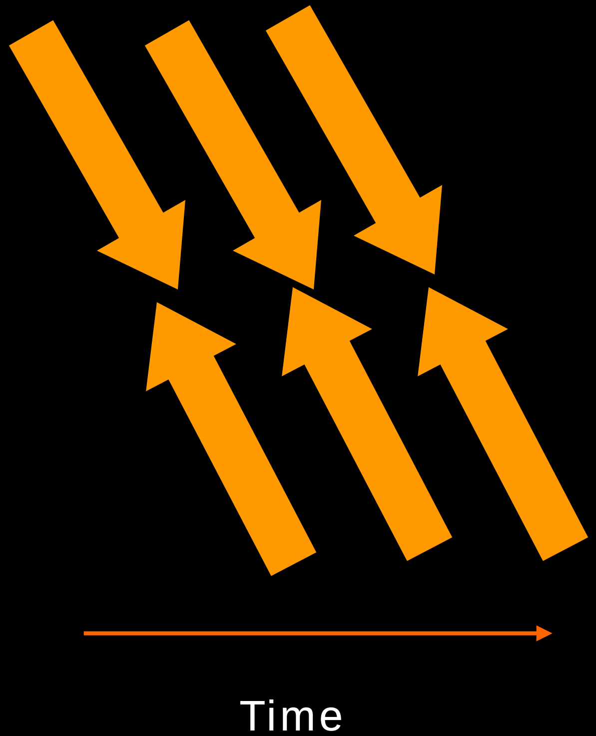
VISION

to develop a flexible modular audio data processing platform for new products and services in the commercial sector; the public service sector; and in educational and cultural research. We will prototype and evaluate solutions in all these areas.

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

CoSound

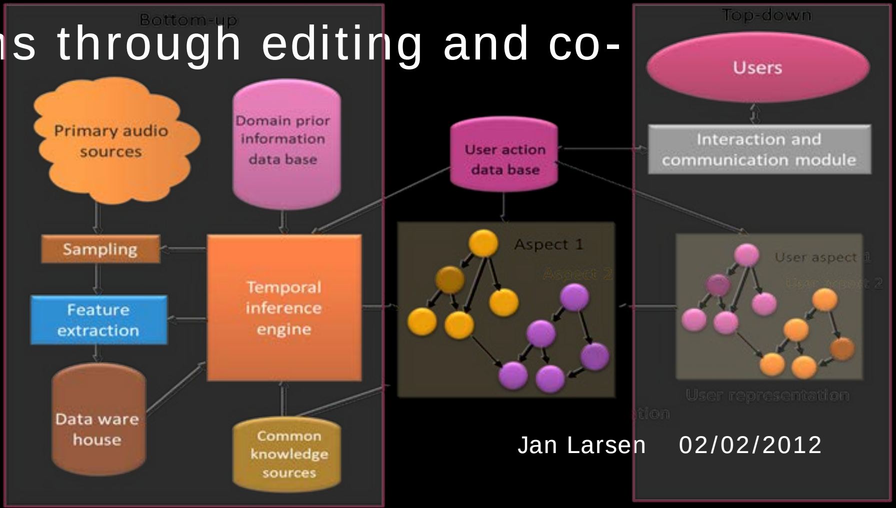
The main hypothesis is that the integration of bottom-up data derived from audio streams and top-down data streams from users can enable actionable cognitive representations, which will positively impact and enrich user interaction with massive audio archives, as well as facilitating new commercial success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

We will test the hypothesis at three different functionality levels:

1) personalized audio streams;

2) task driven navigation and organization;

3) sharing of enriched audio streams through editing and co-creation.



Danish Sound Technology Network

What is it?

What do we do?

VISION

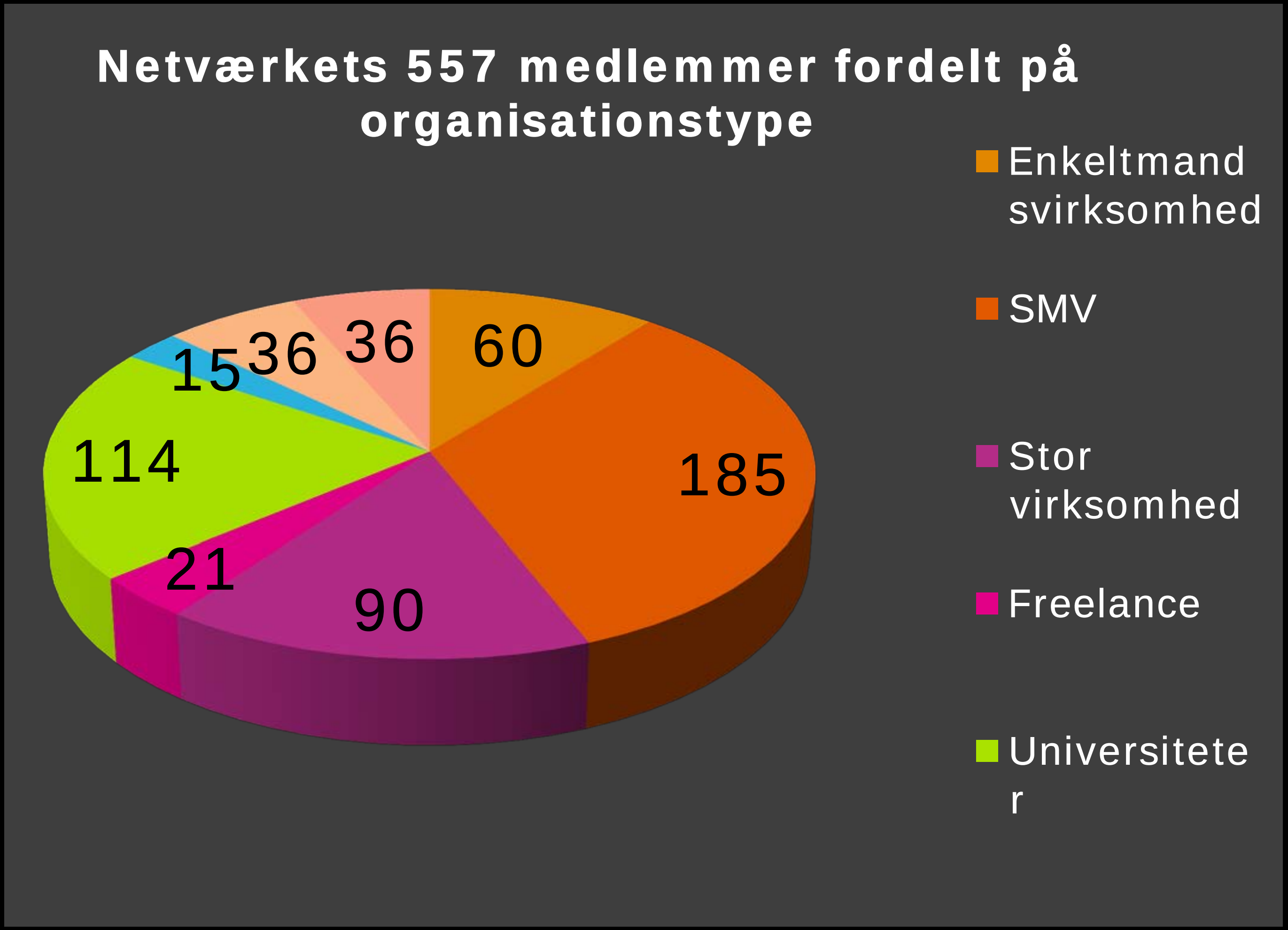
The vision of the Danish Sound Technology network is that Denmark is a leading country with regards to sound technology in terms of knowledge, research and education. Danish sound technology will be the epitome of high quality in products and services as well as in physical rooms and social contexts.

MISSION

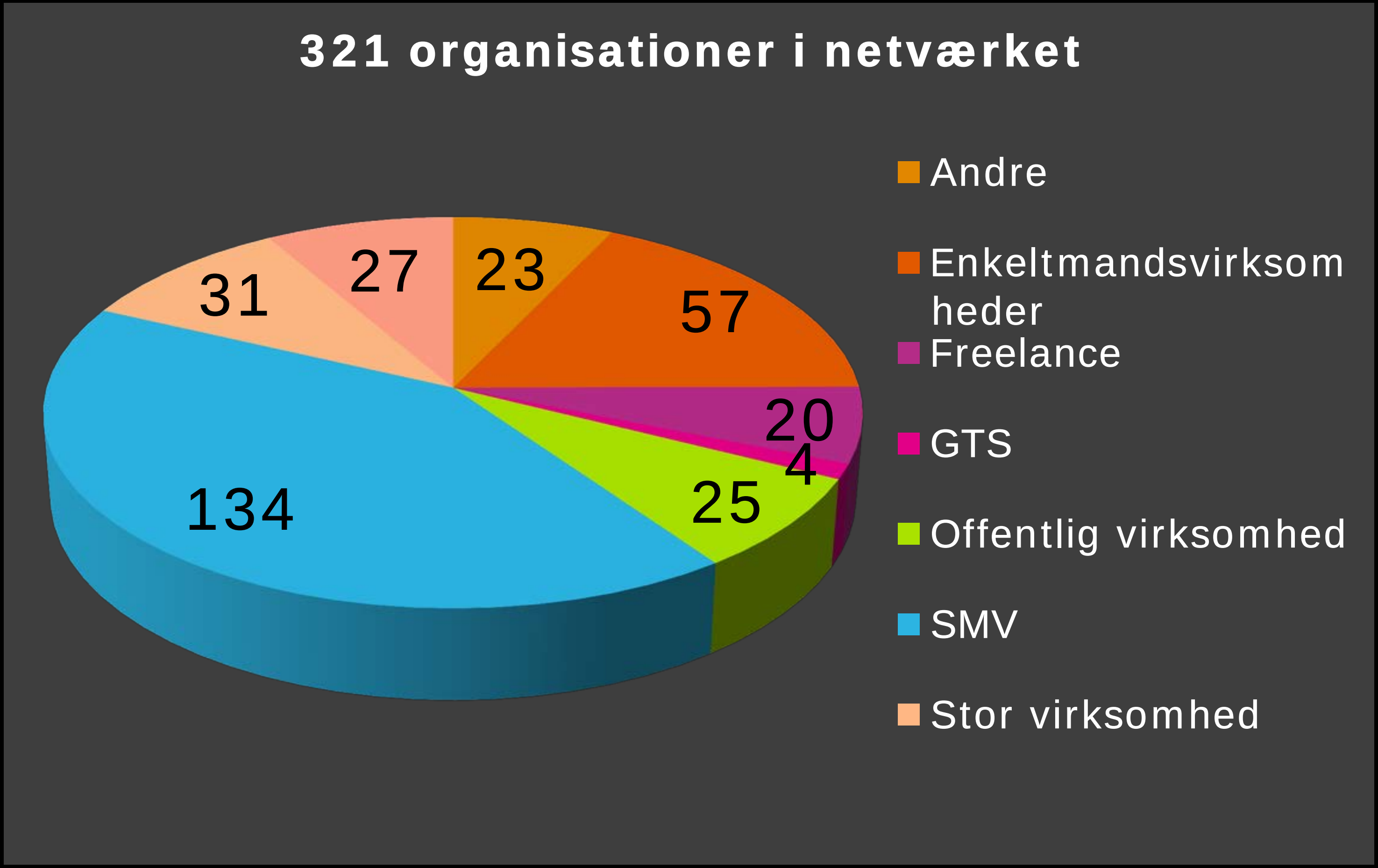
Danish Sound Technology Network embraces all individuals, organizations and businesses in Denmark in the area of sound technology. We create a new space for innovation, collaboration and dissemination of knowledge across

http://www.lydteknologi.dk/pa2011/

557 members in 321 companies and organizations



321 companies and organizations



Consortium partners in Danish Sound Technology Network

More than 100 researchers at

- Sections for Acoustics and Multimedia Information and Signal Processing, Electronics Systems, AAU
- Section for Media Technology, Dept. of Architecture, Design and Media Technology, AAU
- Acoustics Technology and Hearing Systems groups at Dept. of Electrical Engineering, DTU
- Section for Cognitive Systems at Dept. of Informatics and Mathematical Modelling, DTU
- Institute of Sensors, Signals and Electrotechnics, SDU
- DELTA

Danish positions of strength critical mass and visibility

Sound recording and reproduction	• Professional live sound systems • HiFi systems • Class D amplifier systems
Diagnostic and monitoring systems	• Environmental sound analysis • Forensics and surveillance • Measurement systems
Digital media systems	• Organization and retrieval of music and sound and semantic audio • Professional broadcast production systems • Home entertainment systems incl. gaming
Designed sound scapes and sound branding	• Sound communication • Sound for electric cars
Assistive technology and medical devices	• Hearing instruments • Assistive sound in the medical care sector