SOUND AI

Professor, PhD Jan Larsen

Section for Cognitive Systems

DTU Compute, Technical University of Denmark

My dream related to sound...

To create better quality of life by providing augmented and immersive sound experiences for people in society 4.0 using AI technology

Industry 4.0 = Civilization 4.0

It is a cognitive revolution that could be even more disruptive than earlier as it concerns not only the industry but the whole way we live our lives.

AI - **Artificial Intelligence**

is a tool for

IA - **Intelligence Augmentation**

research focus

CoSound

Machine learning based processing of audio data and related information, such as context, users' states, interaction, intention, and goals with the purpose of providing innovative services related to societal challenges in

Transforming big audio data into semantically interoperable data assets and knowledge: enrichment and navigation in large sound archives such as broadcast

Experience economy and edutainment: new music services based on mood, optimization of sound systems

Healthcare: Music interventions to improve quality of life in relation to disorders such as e.g. stress, pain, and ADHD
User-driven optimization of hearing aids

SOUND IS AFFECTIVE


https://www.youtube.com/watch?v=to7uIG8KYhg
SpaceSaver

What are the mechanism? – the BRECVEM model

- **Brain stem reflexes** linked to acoustical properties, e.g. loudness
- **Evaluative conditioning** – association between music and emotion when they occur together
- **Emotional contagion** – emotion expressed in music, sad is e.g. linked low-pitches, slow, and quiet
- **Rhythmic entrainment** – movement synchronization with rhythm
- **Visual images** – creation of visual images
- **Episodic memories** – e.g. strong emotion when you hear a melody linked to an episode
- **Cognitive appraisal** - mental analysis of music an creation of analytic or aesthetic pleasure (hit-songs)
- **Musical expectancy** - balance between surprise and expectation

Ref: Juslin, P. N. and Västfäll, D. *Emotional response to music: The need to consider underlying mechanism. Behavioral and Brain Sciences, vol. 31, pp. 559–621, 2008*.
Line Gebauer & Peter Vuust, *Music interventions in Health Care*, *2014.*

AI IS EFFECTIVE

What is machine learning?

Learning structures and patterns form from historical data to reliably predict outcome for new data.

Computers only do what they are programmed to do. ML infers new relations and patterns, which were not programmed. They learn and adapt to changing environment.

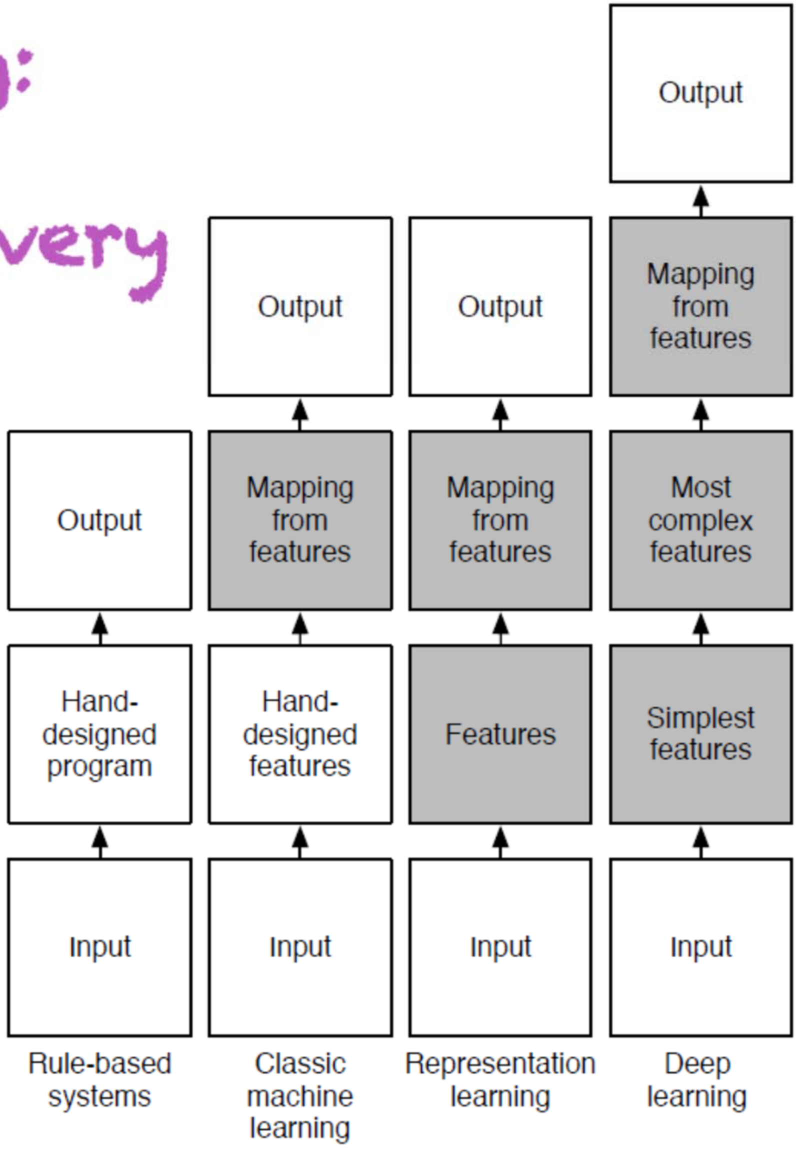
1. Computer systems that automatically improve through experience, or learns from data.
2. Inferential process that operate from representations that encode probabilistic dependencies among data variables capturing the likelihoods of relevant states in the world.
3. Development of fundamental statistical computational-information-theoretic laws that govern learning systems - including computers, humans, and other entities.

M. I. Jordan and T. M. Mitchell. *Machine learning: Trends, perspectives, and prospects*. Science, July 2015.
Samuel J. Gershman, Eric J. Horvitz, Joshua B. Tenenbaum. *Computational rationality: A converging paradigm for intelligence in brains, minds, and machines.* Science, July 2015.

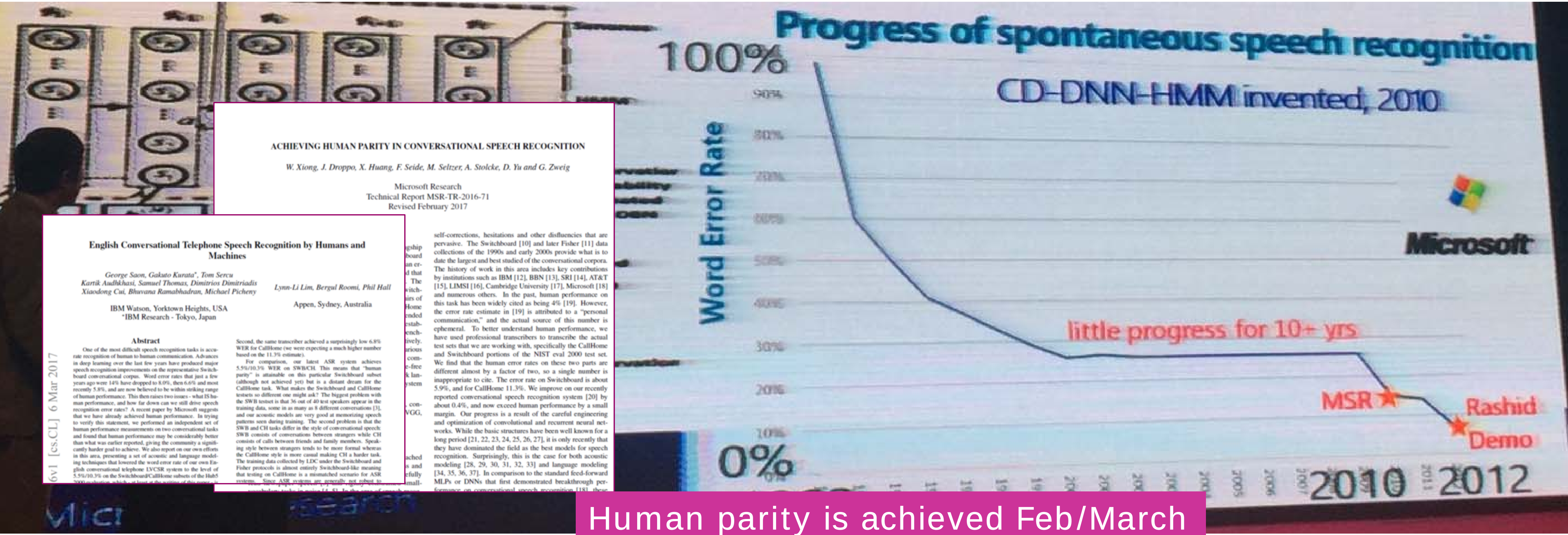
Deep Learning: Automating Feature Discovery

Geoff Hinton, Yoshua Bengio, Yann LeCun, Deep Learning Tutorial, NIPS 2015, Montreal.

Deep learning is a disruptive technology



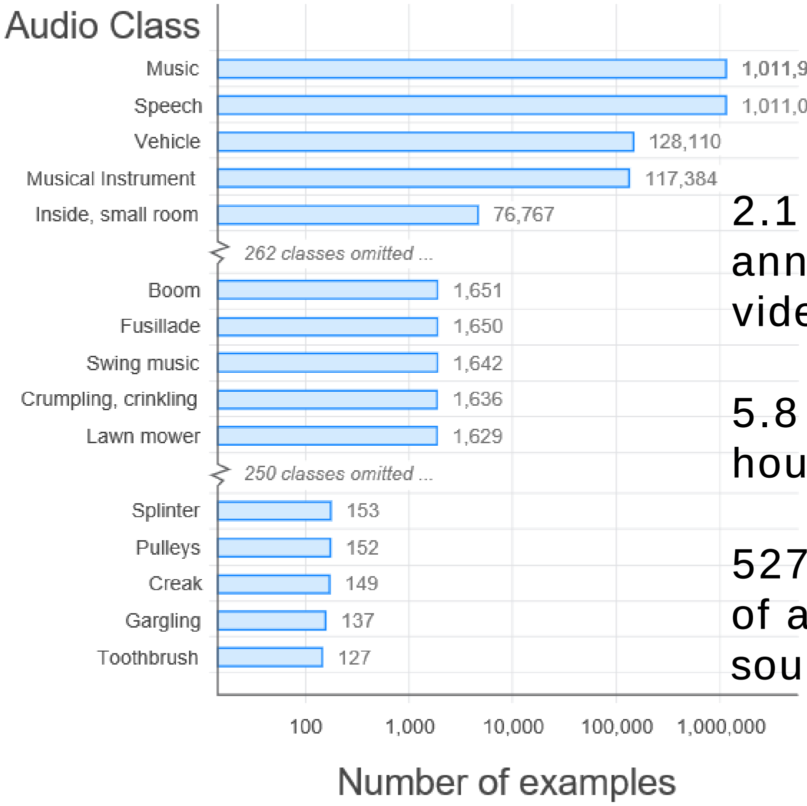
Machine learning is very successful for speech recognition and chat bots



Human parity is achieved Feb/March 2017

Geoffrey Hinton, Li Deng, Dong Yu, George Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara Sainath, and Brian Kingsbury. *Deep Neural Networks for Acoustic Modeling in Speech Recognition.* IEEE Signal Processing Magazine, 82, Nov. 2012.
George Saon, Gakuto Kurata, Tom Sercu, Kartik Audhkhasi, Samuel Thomas, Dimitrios Dimitriadis, Xiaodong Cui, Bhuvana Ramabhadran, Michael Picheny, Lynn-Li Lim, Bergul Roomi, Phil Hall. *English Conversational Telephone Speech Recognition by Humans and Machines, https://arxiv.org/abs/1703.02136, March 2017*
W. Xiong, J. Droppo, X. Huang, F. Seide, M. Seltzer, A. Stolcke, D. Yu, G. Zweig. *Achieving Human Parity in Conversational Speech Recognition, https://arxiv.org/abs/1610.05256, October 2016.*

Machine learning is very successful for audio classification



2.1 million annotated videos

5.8 thousand hours of audio

527 classes of annotated sounds

Table 2: Comparison of performance of several DNN architectures trained on 70M videos, each tagged with labels from a set of 3K. The last row contains results for a model that was trained much longer than the others, with a reduction in learning rate after 13 million steps.

Architectures	Steps	Time	AUC	d-prime	mAP
Fully Connected	5M	35h	0.851	1.471	0.058
AlexNet	5M	82h	0.894	1.764	0.115
VGG	5M	184h	0.911	1.909	0.161
Inception V3	5M	137h	**0.918**	**1.969**	0.181
ResNet-50	5M	119h	0.916	1.952	**0.182**
ResNet-50	17M	356h	**0.926**	**2.041**	**0.212**

Mean average precision mAP is low because of low class prior $<10^{-4}$.

AUC is the area under curve of true positive rate vs. false positive rate.

Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, Marvin Ritter. *Audio Set: An ontology and human-labeled dataset for audio events*, IEEE ICASSP 2017, New Orleans, LA, March 2017.

Shawn Hershey, Sourish Chaudhuri, Daniel P. W. Ellis, Jort F. Gemmeke, Aren Jansen, Channing Moore, Manoj Plakal, Devin Platt, Rif A. Saurous, Bryan Seybold, Malcolm Slaney, Ron Weiss, Kevin Wilson. *CNN Architectures for Large-Scale Audio Classification*, ICASSP 2017, New Orleans, LA, March 2017.

Machine learning is very successful for speech generation

WaveNet is a deep generative model of raw audio waveforms

WaveNets are able to generate speech which mimics any human voice and which sounds more natural than the best existing Text-to-Speech systems, reducing the gap with human performance by over 50%.

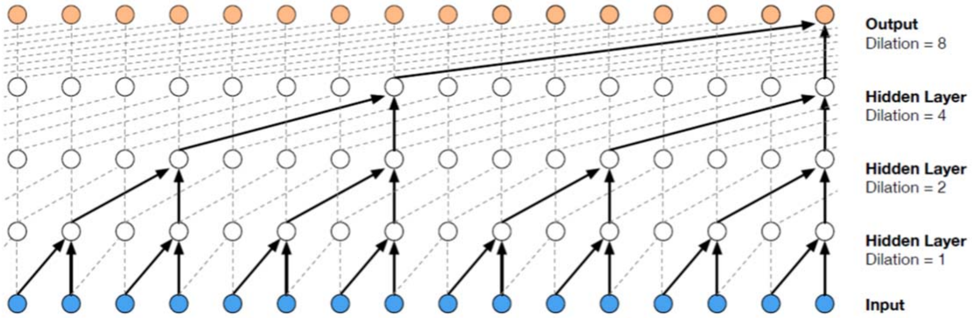
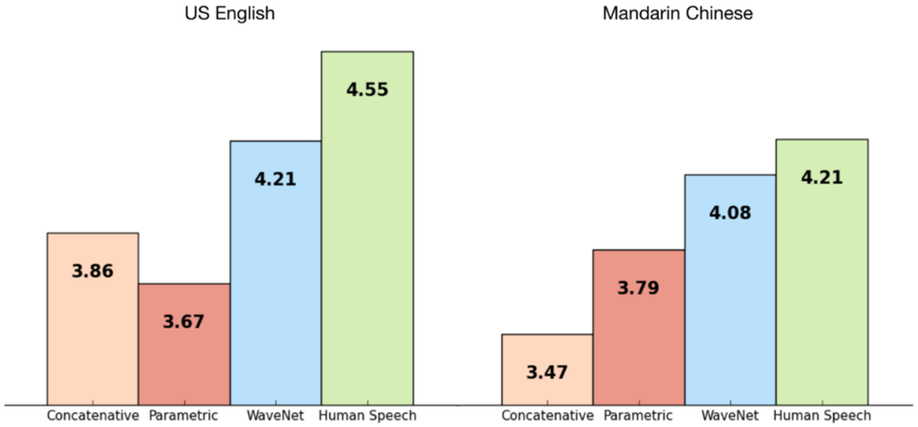


Figure 3: Visualization of a stack of *dilated* causal convolutional layers.



Äron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, Koray Kavukcuoglu. *WAWENET: A Generative Model for Raw Audio,* https://arxiv.org/pdf/1609.03499.pdf, Sept 2016, https://deepmind.com/blog/wavenet-generative-model-raw-audio/

BLACK BOX OF AI

Objectives:

Trust
Causality
Transferability
Decomposability
Informativeness

Legal issues: European Union regulations on algorithmic decision-making and a “right to explanation”

Davide Castelvecchi: http://www.nature.com/polopoly_fs/1.20731!/menu/main/topColumns/topLeftColumn/pdf/538020a.pdf, Nature, Vol. 538, 6 Oct. 2016
Z.C. Lipton: *The mythos of model interpretability*, arXiv:1606.03490, 2016.
Bryce Goodman, Seth Flaxman: *European Union regulations on algorithmic decision-making and a “right to explanation”*, https://arxiv.org/pdf/1606.08813v3.pdf

What defines simple and complex problems and how do we solve them them?

passive

exploration and summarization

prediction

continuous learning
reflection
pro-activeness
engagement
experimentation
creativity

active and autonomous

Unreasonable effectiveness of

Mathematics E. Wigner, 1960

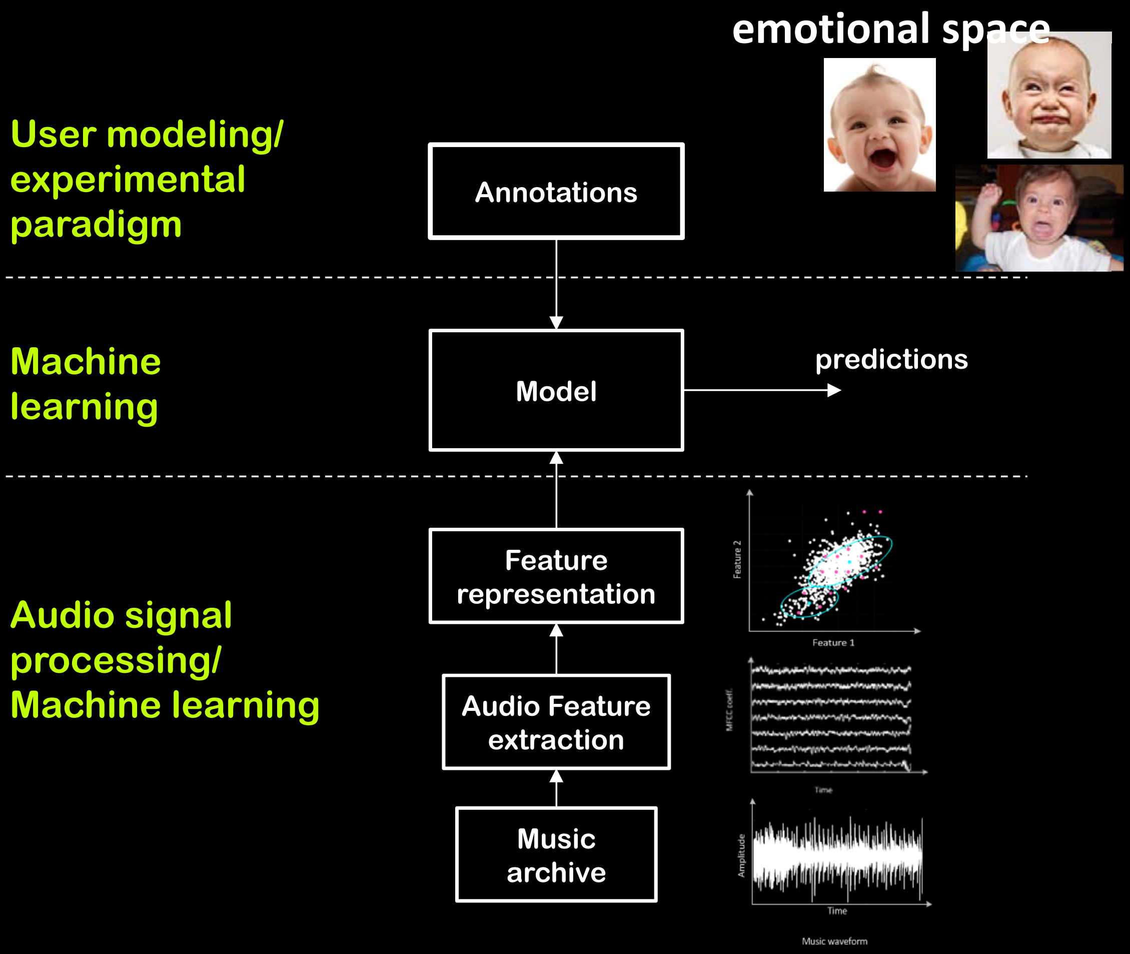
Data Halevy, Norvig, Pereira, 2009

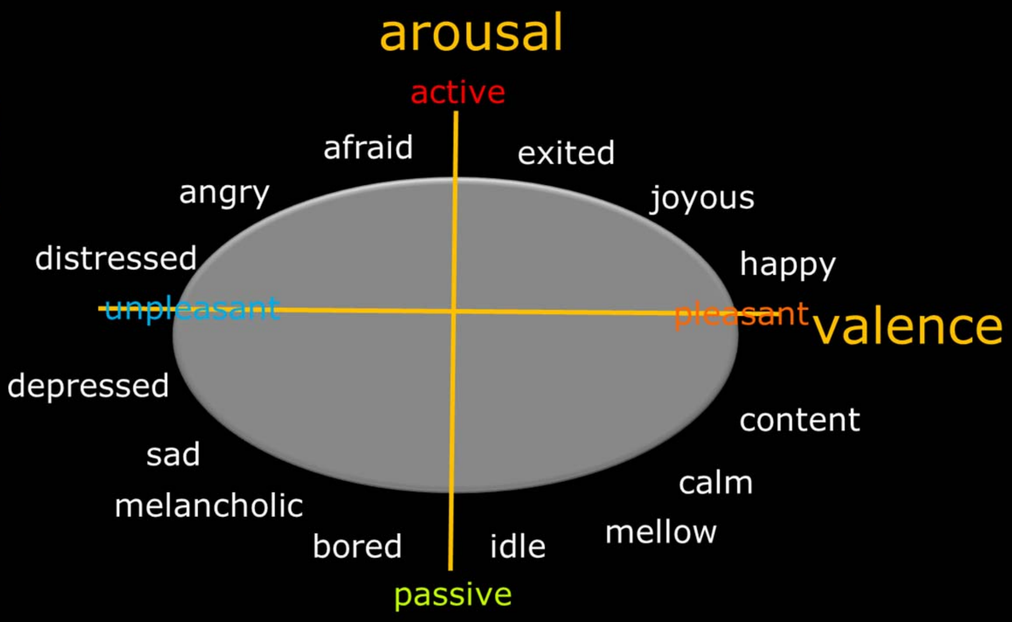
RNNs Karpathy, 2015

Experimentation and interaction users-in-the-loop

INTERACTIVE MACHINE LEARNING IN SOUND

Music Emotion Modeling

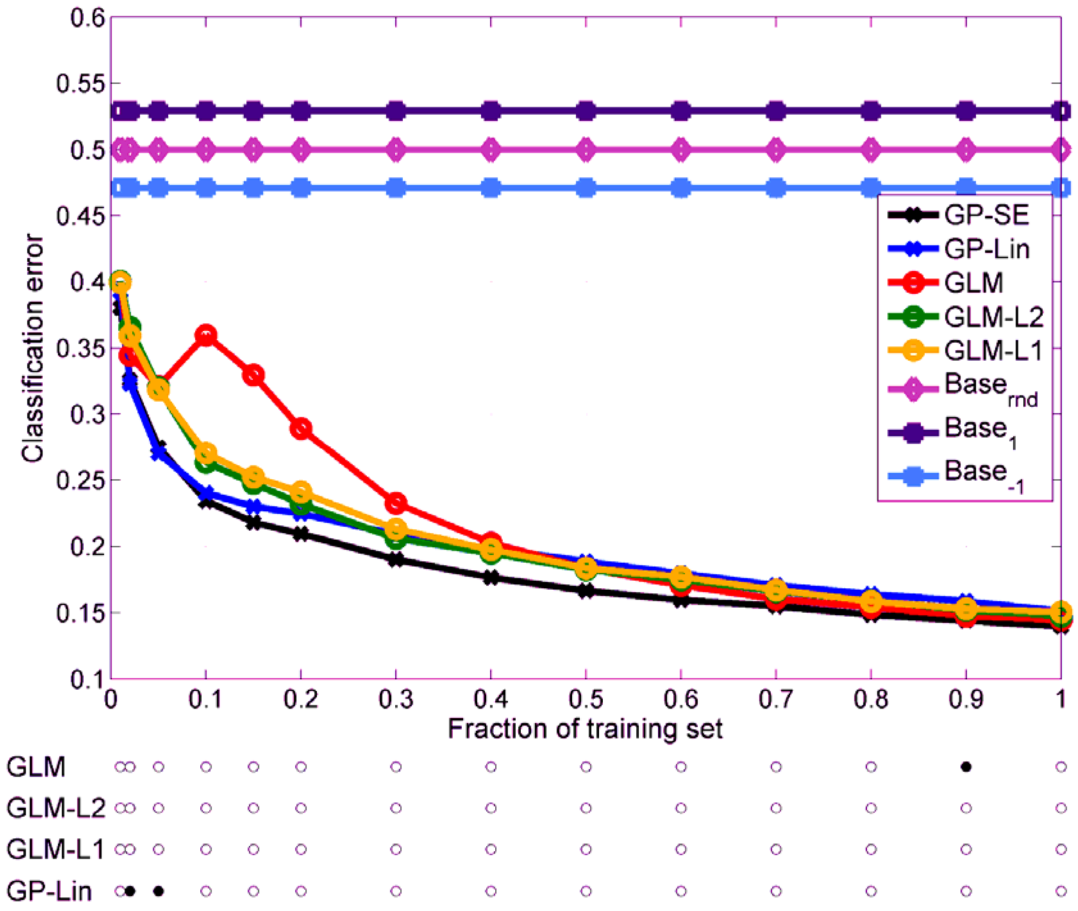




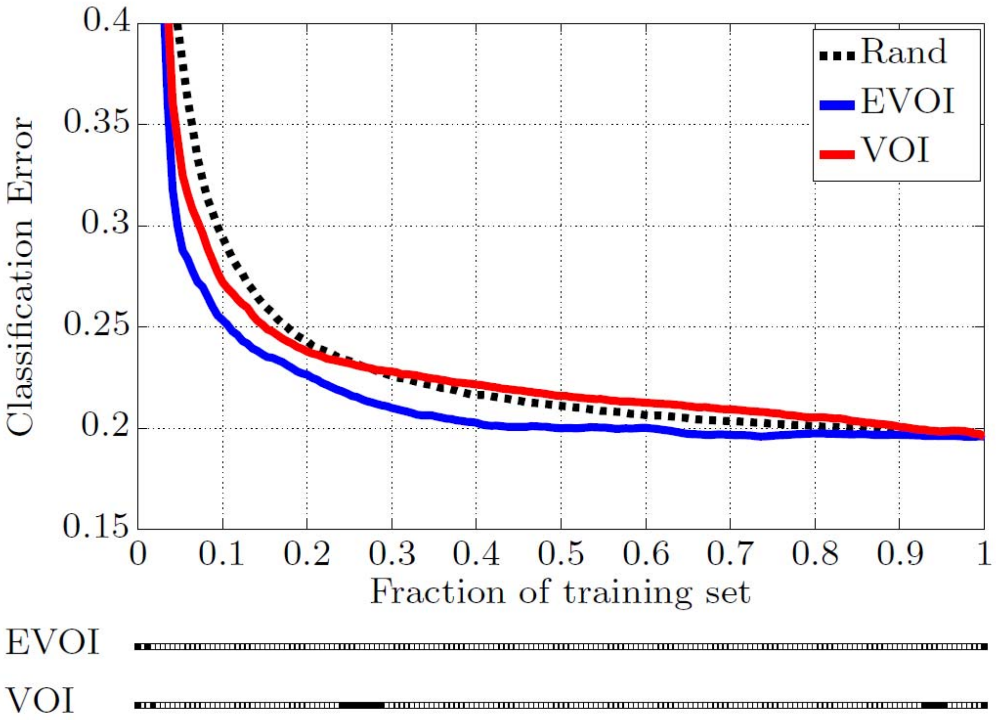
J. A. Russel: "A Circumplex Model of Affect," *Journal of Personality and Social Psychology*, 39(6):1161, 1980

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

Learning curve modeling arousal shows nonlinear modelling is best



How many pairwise comparisons do we need to model emotions?



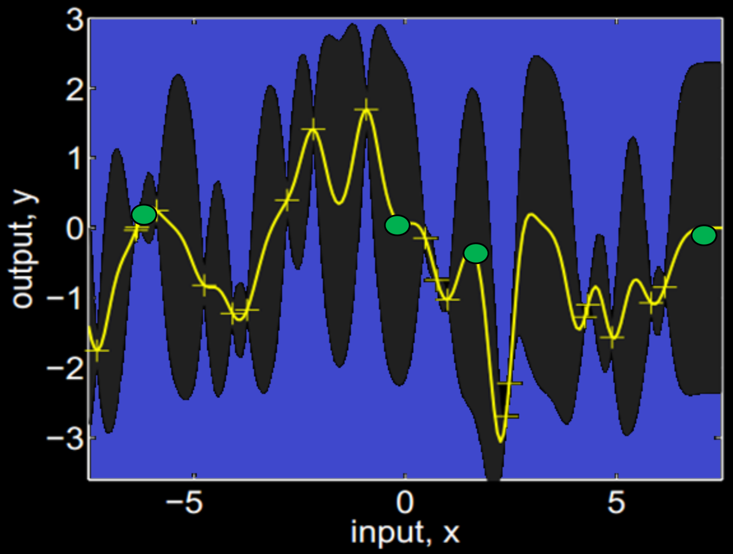
Using active learning
15% for valence
9% for arousal

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

Interactive Learning / Sequential Experimental Design

Generalization objective
Eliciting and learning the entire model / objective function.
Expected change in relative entropy is derived from the posterior and predictive distribution.

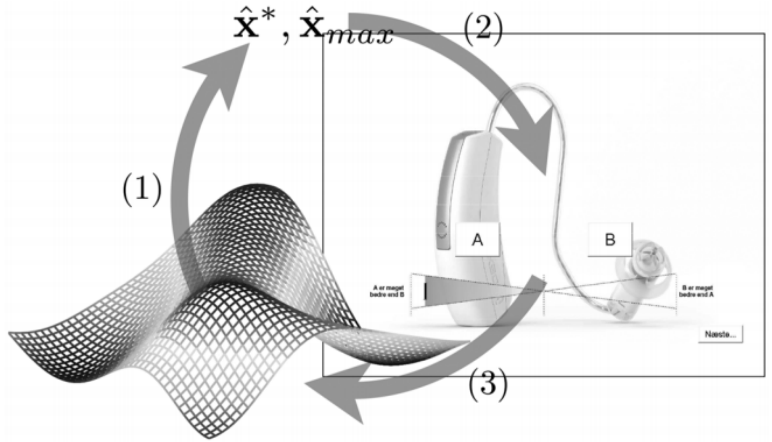
Optimization objective
Learning and identifying optimum
The Expected Improvement of a new candidate sample (green points) is derived from the predictive distribution.



Which of the four green parameters settings/products/interface, x, should the user assess (rate/annotate/see/hear), or where do we need tp evaluate objective performance measurements

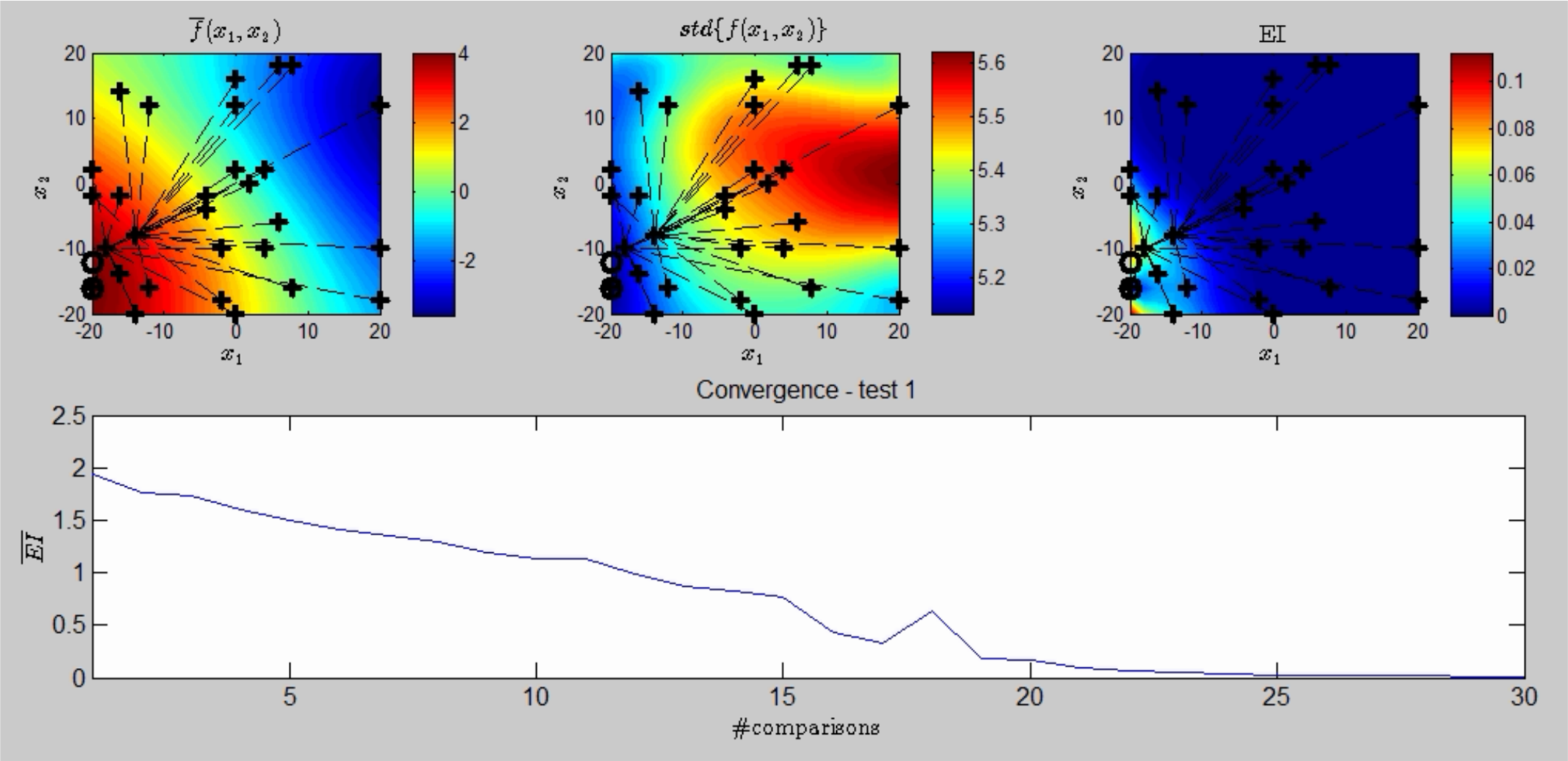
Hearing Aids

- Highly personal needs
- Dynamic environment and use with different needs.
- Latent, convoluted object functions which are difficult to express though verbal and motor actions.
- Users with disabilities – and often elderly people - with inconsistent and noisy interactions.



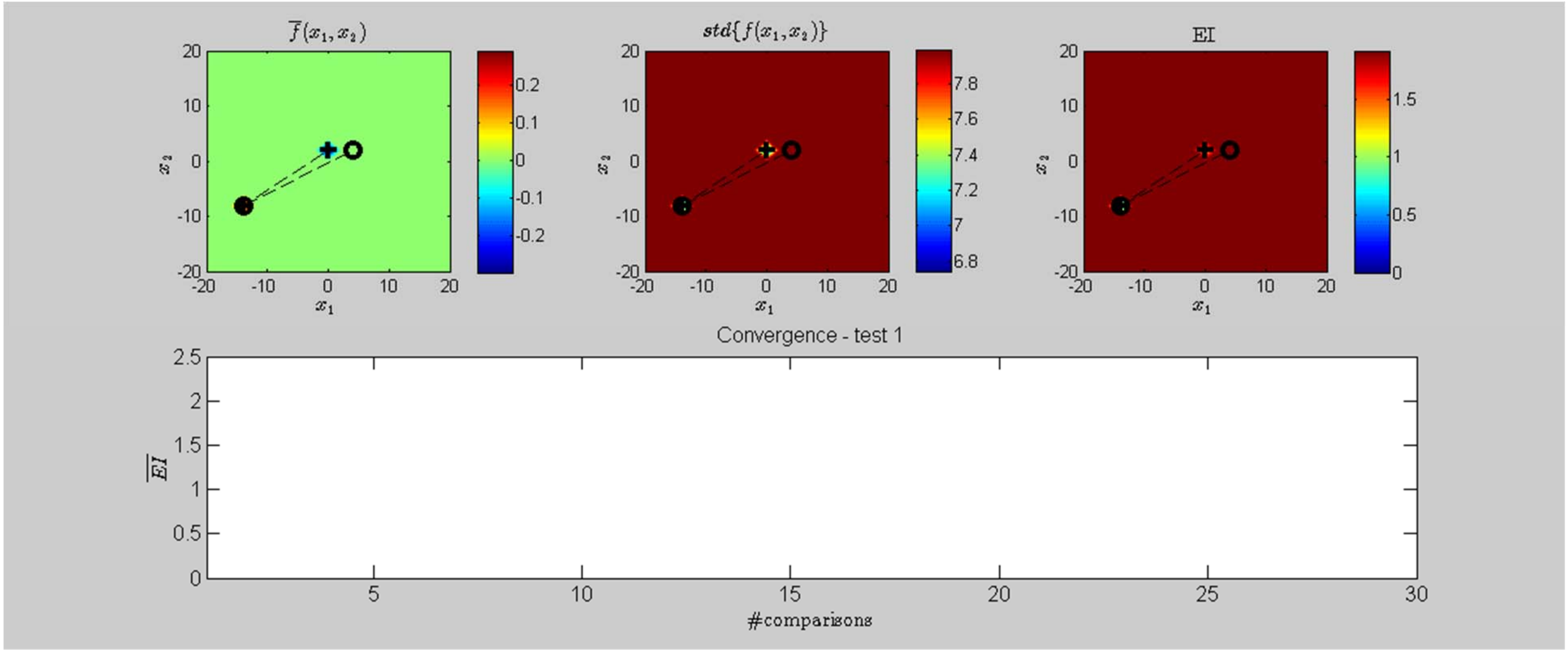
Jens Brehm Nielsen, Jakob Nielsen: Efficient Individualization of Hearing and Processers Sound, ICASSP2013.
Jens Brehm Nielsen, Jakob Nielsen, Jan Larsen: Perception based Personalization of Hearing Aids using Gaussian Process and Active Learning, IEEE Trans. ASLP, vol. 23, no. 1, pp. 162 – 173, Jan 2015.

Pairwise (2AFC) personalization of HA



Hearing Aids

A real interactive optimization sequence in 30 iterations



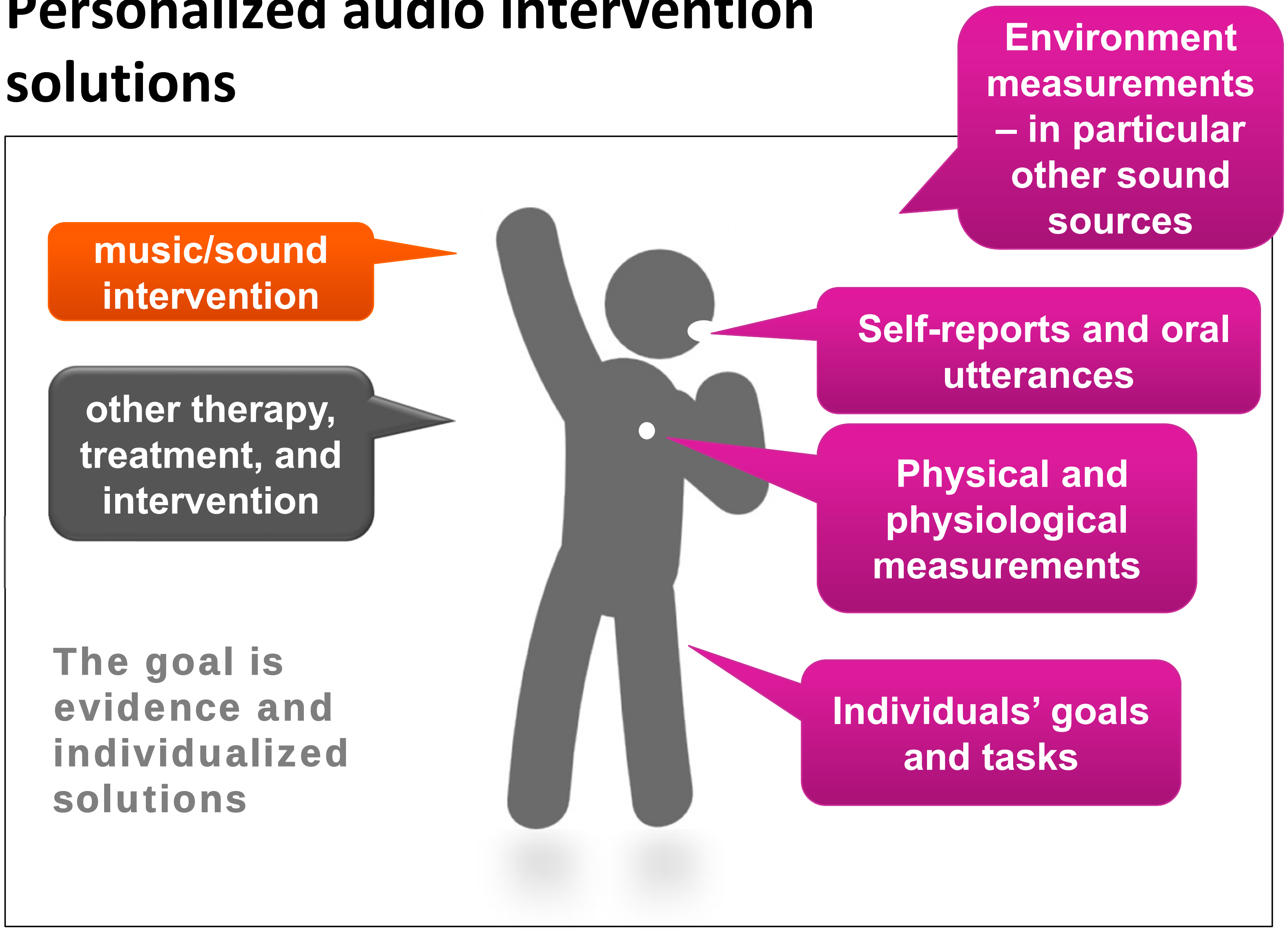
MUSIC AND SOUND INTERVENTION FOR IMPROVING SLEEP IN DEMENTIA PATIENTS

- Anecdotal reports
- Preserved ability to engage in musical activities
- Reduce social isolation
- Improve cognitive symptoms
- Reduce aggression
- More research needed
- Effects might not be specific to music

People highly absorbed in music (AIMS) listening to unfamiliar, but preferred music has higher recovery from a stress situation

S.L. Carstensen, J. Madsen, J. Larsen. *The Influence of Familiarity and Absorption on the Effectiveness of Music in Stress Reduction, in submission 2017.*

Personalized audio intervention solutions



FUTURE

Cognizant audio systems

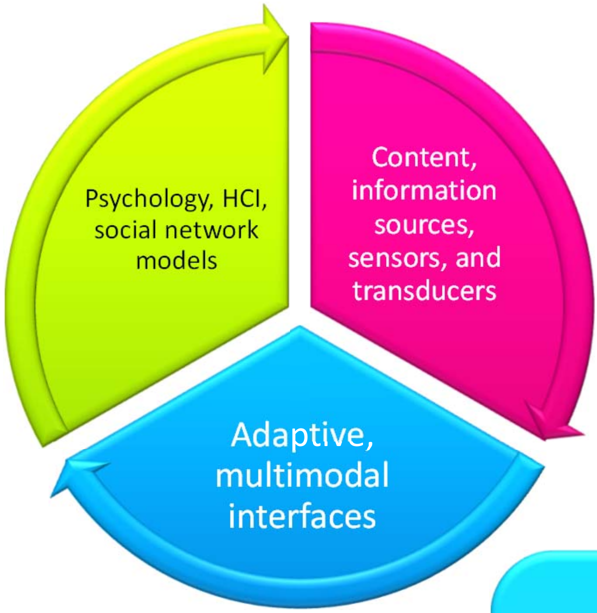
fully informed and aware systems

Context:

who, where, what

Users in the loop:

direct and indirect



Listen in on audio and other sensor streams to segment, identify and understand

Interactive dialog with the user enables long term/continuous behavior tracking, personalization, elicitation of perceptual and affective preferences, as well as adaptation

Flexible integration with other media modalities

Mixed modality experience: Use other modalities to enhance, substitute or provide complementary information

THE WAYS AHEAD

- Need for possibility to include co-creation and production.
- Need for more data across domains and situations.
- Need for systems and platforms that enables experimentation and direct user interaction.
- Need for better AI and machine learning methodology that can provides robust, interpretable, interactive learning from few examples.