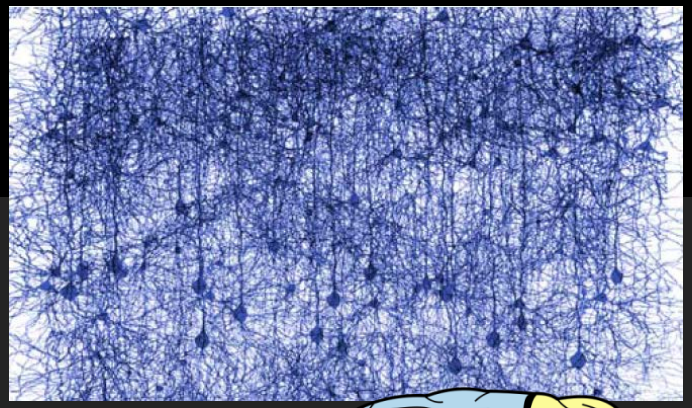
Lars Kai Hansen
lkai@dtu.dk

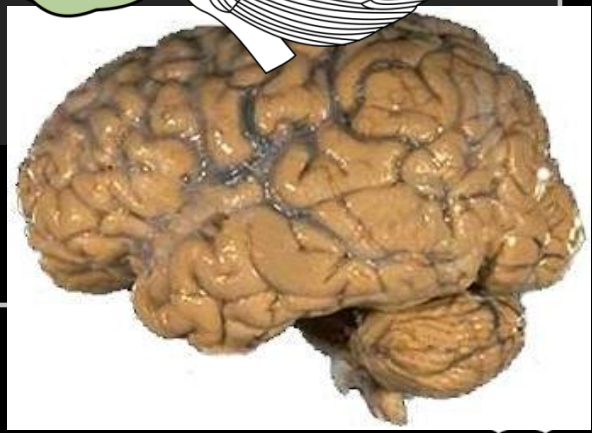
Sensing the deep structure of signals

"...from data to symbols"

the brain design

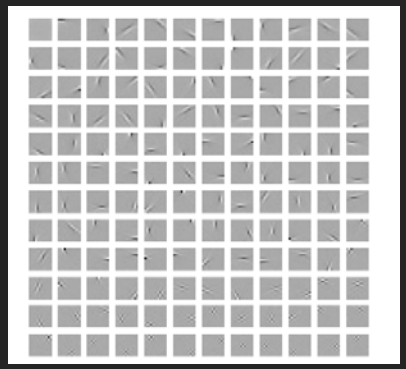


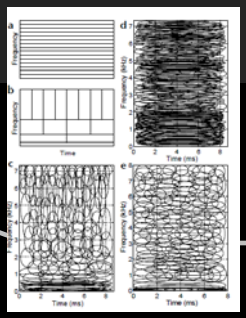
AI/ML has learned to use
three (at least) brain design principles
i) Division of labor: dedicated
”centers” for vision, hearing, smell etc
ii) Neural networks of simple computers
iii) Learning – adaptivity -plasticity



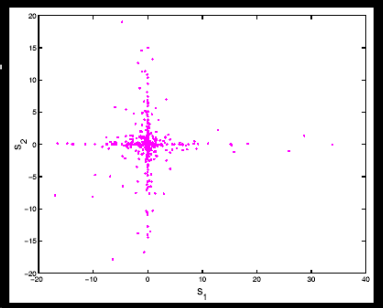
What’s next – to learn from the laws of cognition? how do brains represent / index the world, how do they rank importance? How do they pursue objetives - we need a new ***principia***
...

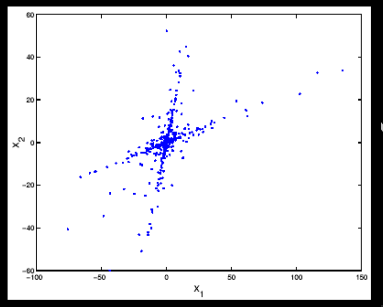
Optimal representations?





Important for engineering proxies for human information processing...
Cf. efficient coding of ”context-to-action” mapping





Outline

What is deep structure?

- Cognitive components and attention modeling
- Ecology of audio signals

Is structure is determined by the environment: Statistics/ physics / mechanisms?

Uniqueness of perception in the brain &
Uniqueness in deep neural networks

What about higher order cognition, social cognition?

things...?

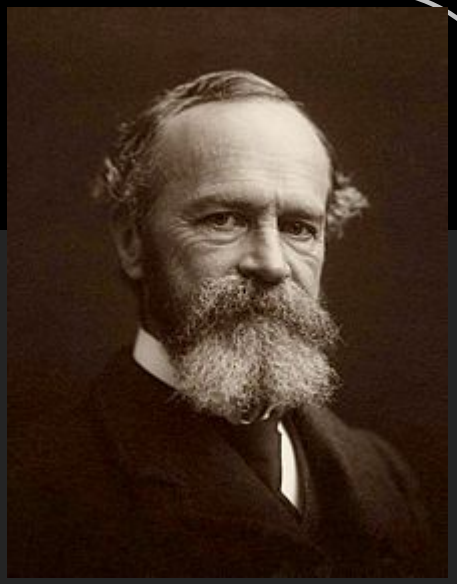
Attention & human optimality

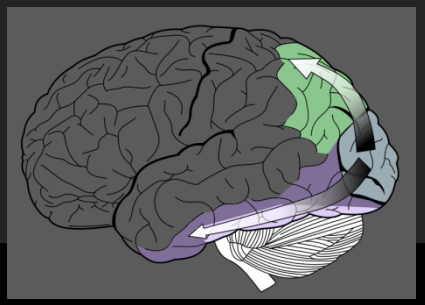
"... the withdrawal from some things
in order to deal effectively with others"

William James (1890)

"... To behave adaptively in a complex world, an animal must select, from the wealth of information available to it, the information that is most relevant at any point in time. This information is then evaluated in working memory, where it can be analyzed in detail, decisions about that information can be made, and plans for action can be elaborated. The mechanisms of attention are responsible for selecting the information that gains access to working memory."

Eric I. Knudsen (2007)





W. James, *The Principles of Psychology, Vol. 1, Dover Publications,* 1880/1950.

E.I. Knudsen, "Fundamental Components of Attention," *Annual Review of Neuroscience, vol. 30, no. 1, pp. 57–78, 2007.*

DTU

Deep structure needed to predict the future

Processing in the brain is based on extremely well-informed / optimized representations and mechanisms –

A key issue is selective attention, …before solving attention we to address a

Fundamental question:

What can you attend to?

or...

What is an object / chunk of information?

Cognitive component analysis ...what we can attend to

- The **object / chunk** is a key notion in cognitive psychology
  - ...number of objects in short time memory, objects ”race to short term memory”

Miller, G.A. (1956), The Magical Number Seven, Plus or Minus Two: Some Limits on our Capacity for Processing Information. Psychological Review, 63, 81-97

Bundesen, C., Habekost, T. and Kyllingsbæk, S., 2005. A neural theory of visual attention: bridging cognition and neurophysiology. Psychological review, 112(2), p.291).

  - Miller: “..we are not very definite about what constitutes a chunk of information.“
  - A pragmatic definition of an object could be: **An object is a signal source with independent behavior in a given environment** (..imagined?)

- Theoretical issues: The relation between supervised and un-supervised learning. Related to the discussion of the utility of unlabeled examples in supervised learning and fast/one sample learning...

- Practical Issues: Can we predict which digital media components a user will pay attention to? -a key challenge for cognitive systems.

”Cognitive compatibility”

Unsupervised Learning

Hidden variable

$$p(\mathbf{s} \mid \mathbf{x}, \mathbf{w}_u) \propto p(\mathbf{x} \mid \mathbf{s}, \mathbf{w}_u) p(\mathbf{s} \mid \mathbf{w}_u)$$

”Cognitive event”: Data, sound, image, behavior

Supervised learning

$$p(\mathbf{y} \mid \mathbf{x}, \mathbf{w}_s)$$

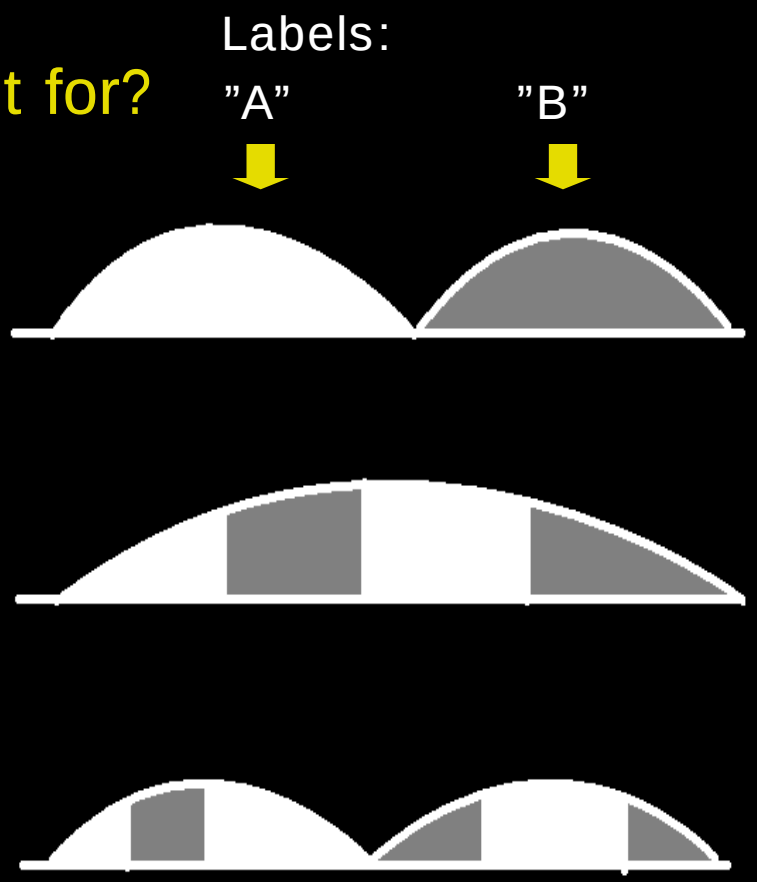
”Cognitive” label, i.e. provided by a human observer

How well do these learned representations match: s = y ?

Which domains are COCA relevant for?

- If ”statistical structure” in the relevant feature space is well aligned with the label structure we expect high **cognitive compatibility**
- Unsupervised-then-supervised learning –aka *pre-training*- can explain “learning from one example”

The Good,
the Bad,
and the Ugly…



Vector space representation

- Abstract representation - can be used for all digital media
- A “cognitive event” is represented as a point in a high-dimensional ”feature space” – document similarity ~ spatial proximity in a given metric

- Text: Term/keyword histogram, N-grams
- Image: Color histogram, texture measures
- Video: Object coordinates (tracking), active appearance models
- Sound: Spectral coefficients, mel cepstral coefficients, gamma tone filters

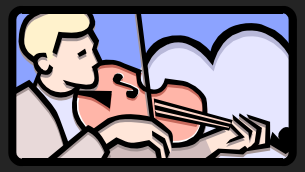
Contexts can be identified by their feature associations (= Latent semantics)

S. Deerwester et al. *Indexing by latent semantic analysis*.
Journal of the American Society for Information Science, 41(6), 391-407, (1990)

The independent component hypothesis

- Challenge: Presence of multiple agents/contexts
- Need to ”blindly” separate source signals = learn contexts
- ICA, NMF, tensor factorization provides (almost) unique solutions to...

$$x(\mathit{feature}, \mathit{time}) = \sum_k A(\mathit{feature}, k)\ s(k, \mathit{time})$$



Linear mixing generative model ICA - "Synthesis" simplistic model incorporating sparsity and independence

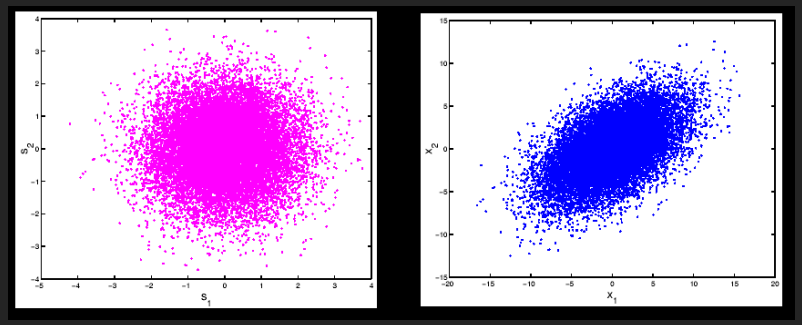
Space-time matrix

Component's "where"

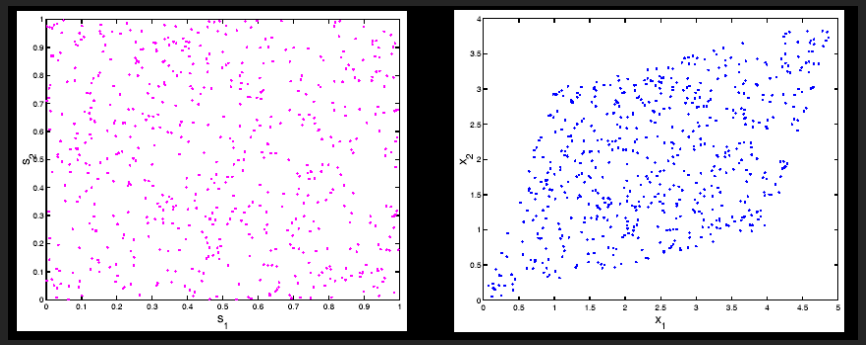
Vector of "what"

$$x(loc, time) = \sum_k A(loc, k)\ s(k, time)$$

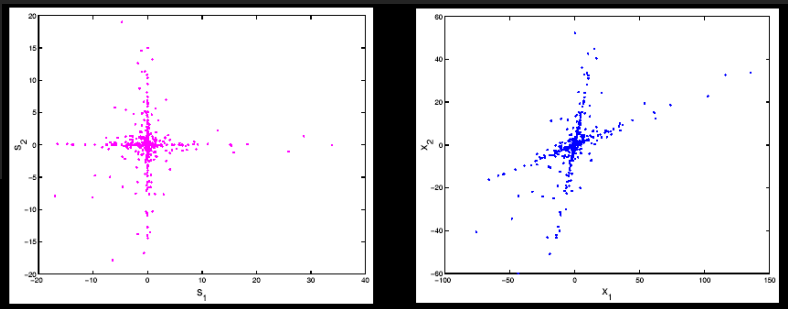
Normal sources



Dense sources



Sparse sources

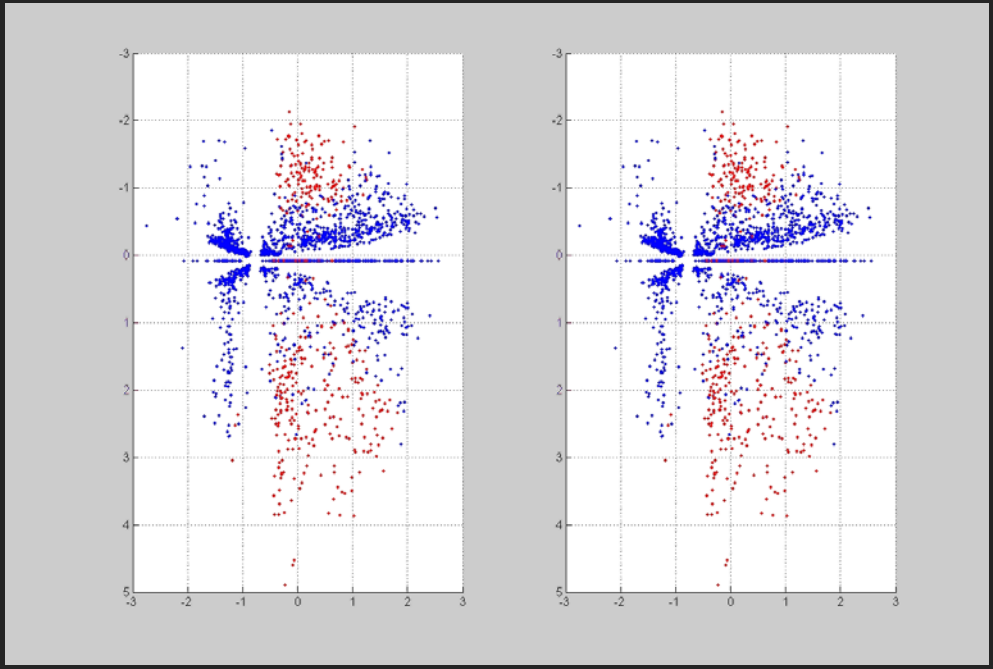


Protocol for comparing supervised and unsupervised learning

- Use the “unsupervised-then-supervised” scheme to implement a classifier:
  - Train the unsupervised scheme, eg., ICA
  - Freeze the ICA representation (A matrix)
  - Train a simple (e.g. Naïve Bayes) classifier using the features obtained in unsupervised learning Use
- Compare with supervised classifier
  - Error rates of the two systems
  - Compare posterior probabilities

Phoneme classification

Nasal vs oral: ”Esprit project ROARS” (Alinat et al., 1993)

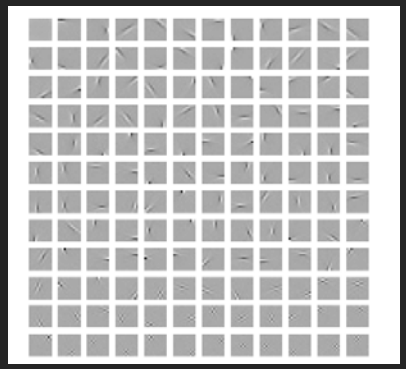


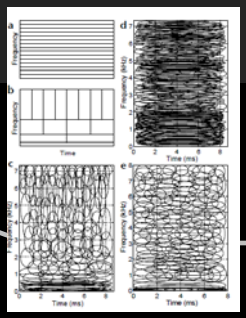
Binary classification

Error rates: **0.23** (sup.), **0.22** (unsup.)

Bitrates: **0.48** (sup.), **0.39** (unsup.)

Optimal representations?

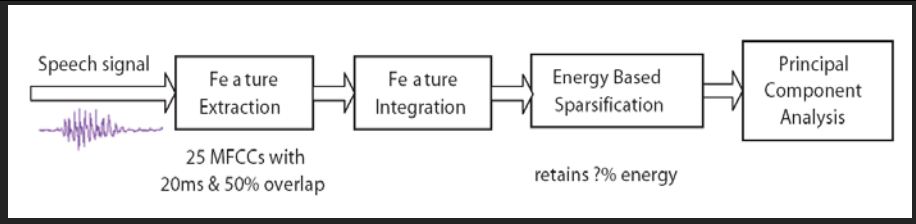




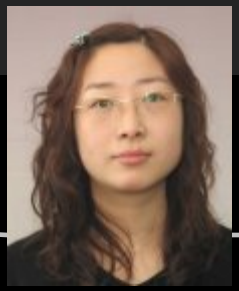
Important for engineering proxies for human information processing...
Cf. efficient coding of ”context-to-action” mapping

Cognitive components of speech

- Basic representation: Mel weigthed cepstral coefficients (MFCCs)
- Modeling at different time scales 20 msec – 1000 msec



- Phonemes
- Gender
- Speaker identity



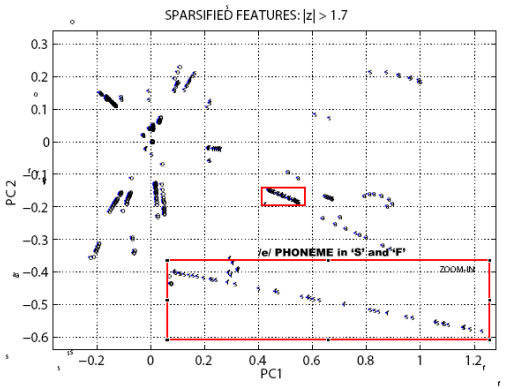
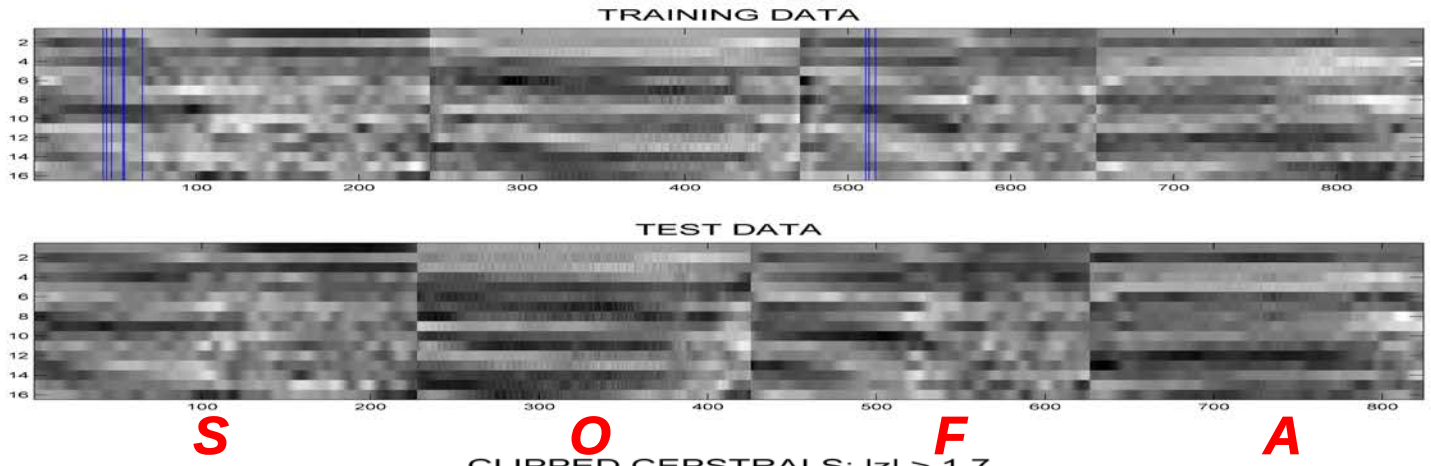
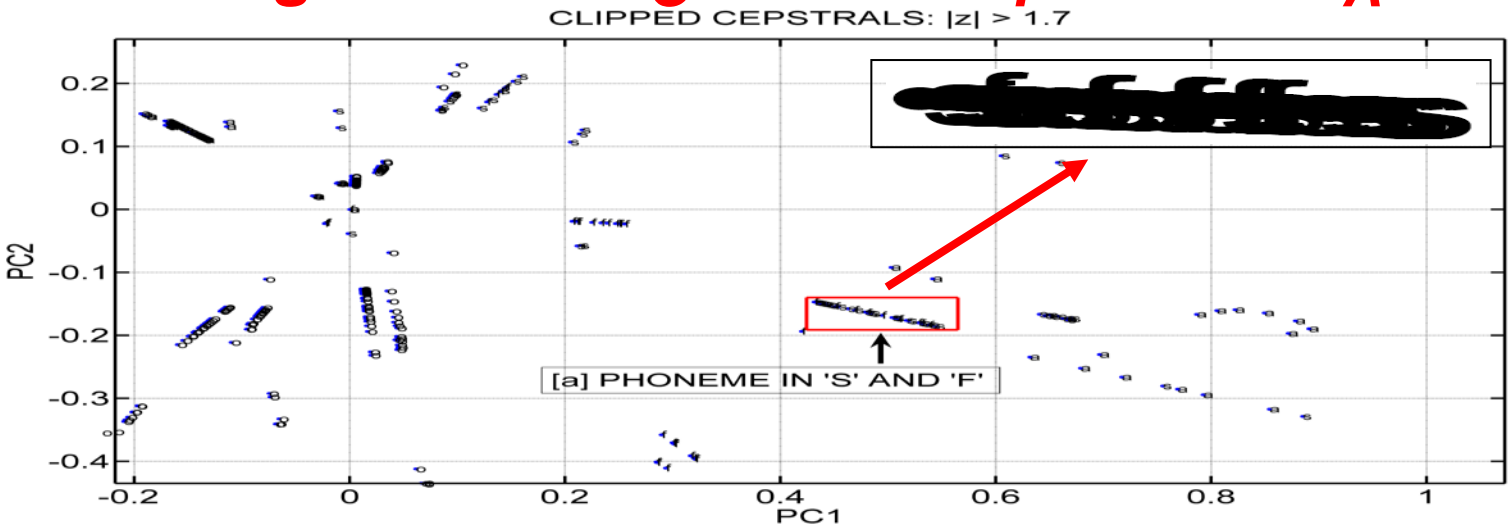


Figure 3: The latent space is formed by the two first principal components of data consisting of four separate utterances representing the sounds 's', 'o', 'f', 'a'. The structure clearly shows the sparse component mixture, with 'rays' emanating from the origin (0,0). The ray embraced in a rectangle contains a mixture of 's' and 'f' features, a cognitive component associated with the vowel /e/ sound.

Co-worker: Ling Feng

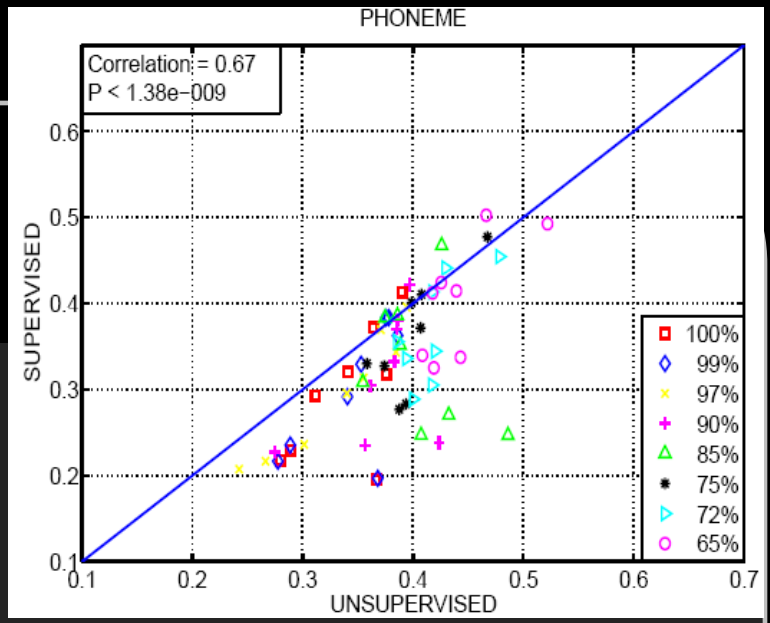
Mel weighted cepstral coeff. (MFCC)

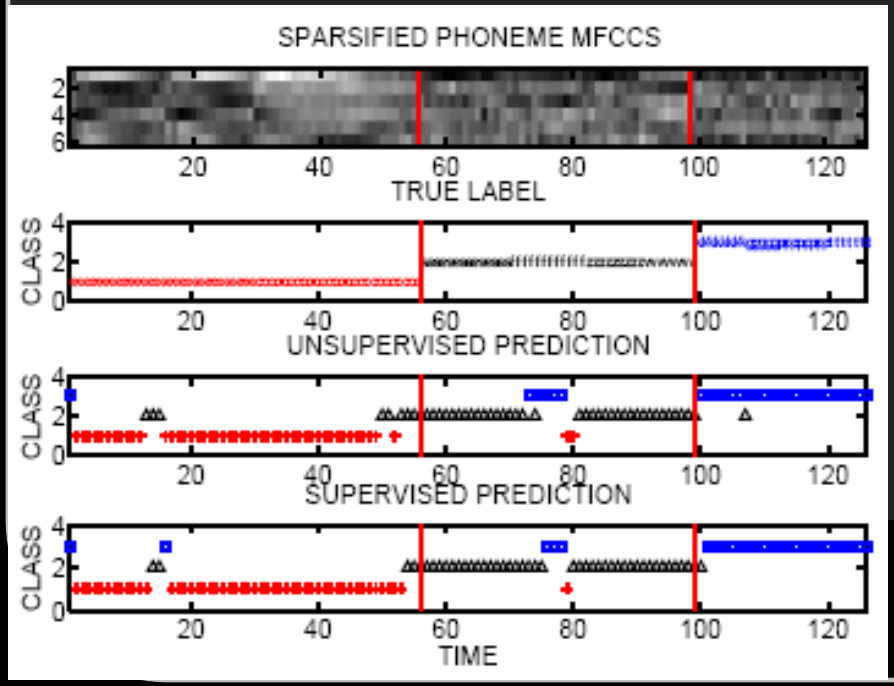




Error rate comparison

For the given time scales and thresholds, data locate around $y = x$, and the correlation coefficient $\rho = 0.67$, $p < 1.38e{-}09$.

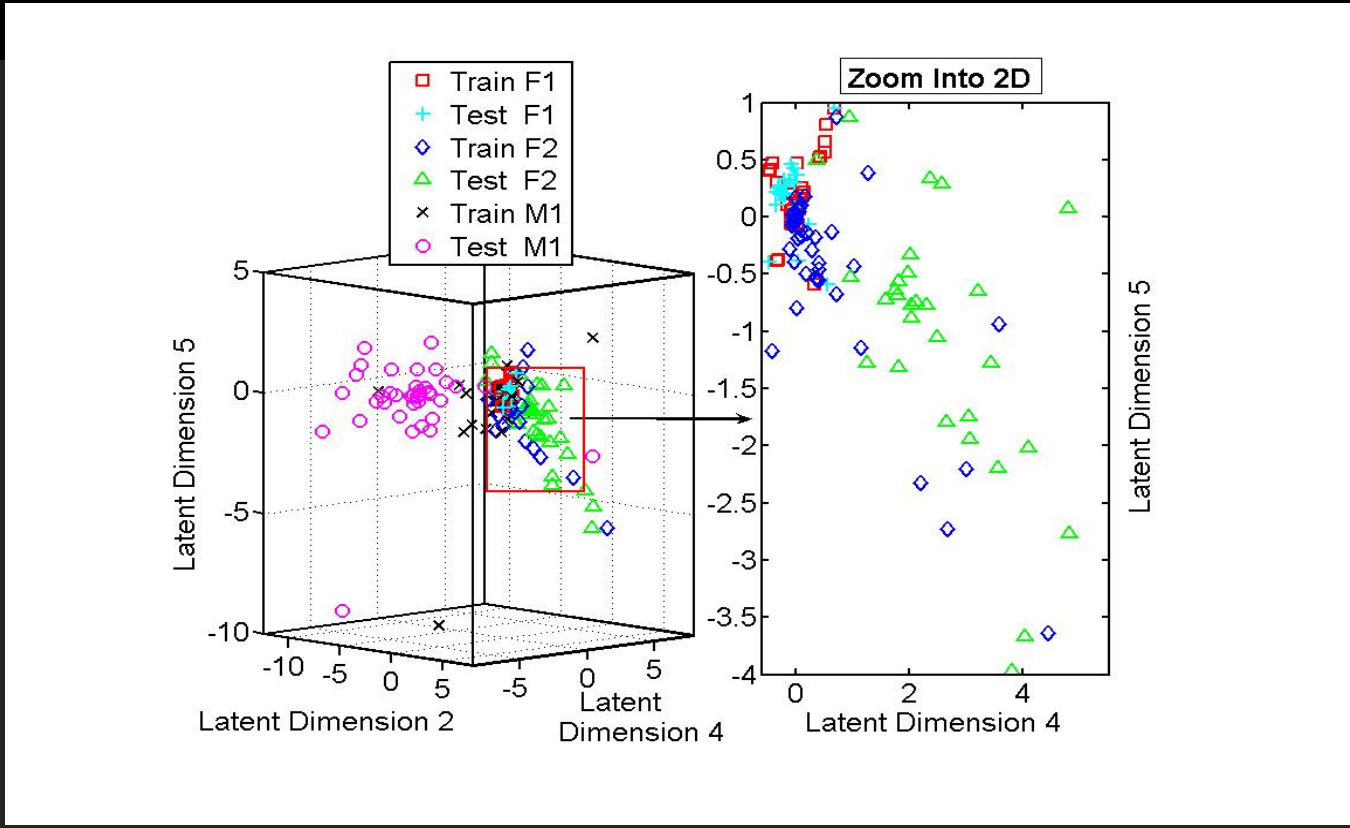




Sample-to-sample correlation

- Three groups: vowels eh, ow; fricatives s, z, f, v; and stops k, g, p, t.
- 25-d MFCCs; EBS to keep 99% energy; PCA reduces dimension to 6.
- Two models had a similar pattern of making correct predictions and mistakes, and the percentage of matching between supervised and unsupervised learning was 91%.

Longer time scales



Time integrated (1000ms) MFCC's: text independent speaker recognition....

Feng & Hansen (CIMCA, 2005)

Error rate correlations for super/unsupervised learning for different cognitive time scales and events

Challenged by degree of sparsity and time averaging

Gender, Identity, Height etc are the Audio Gist variables

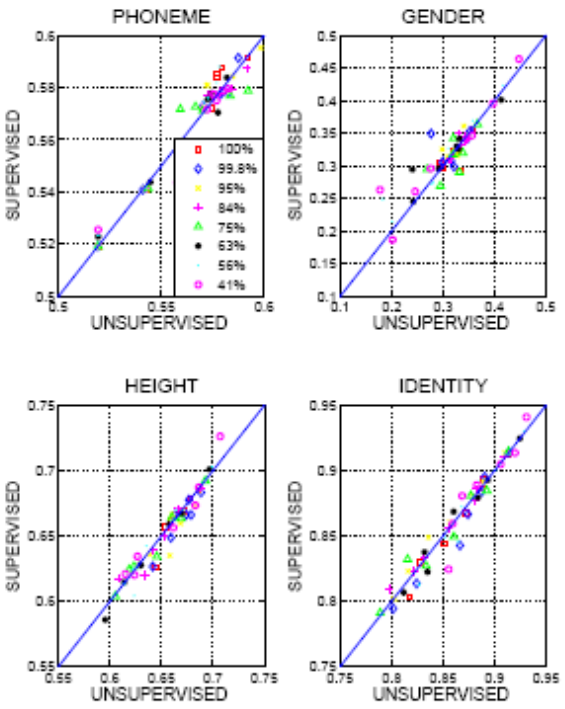


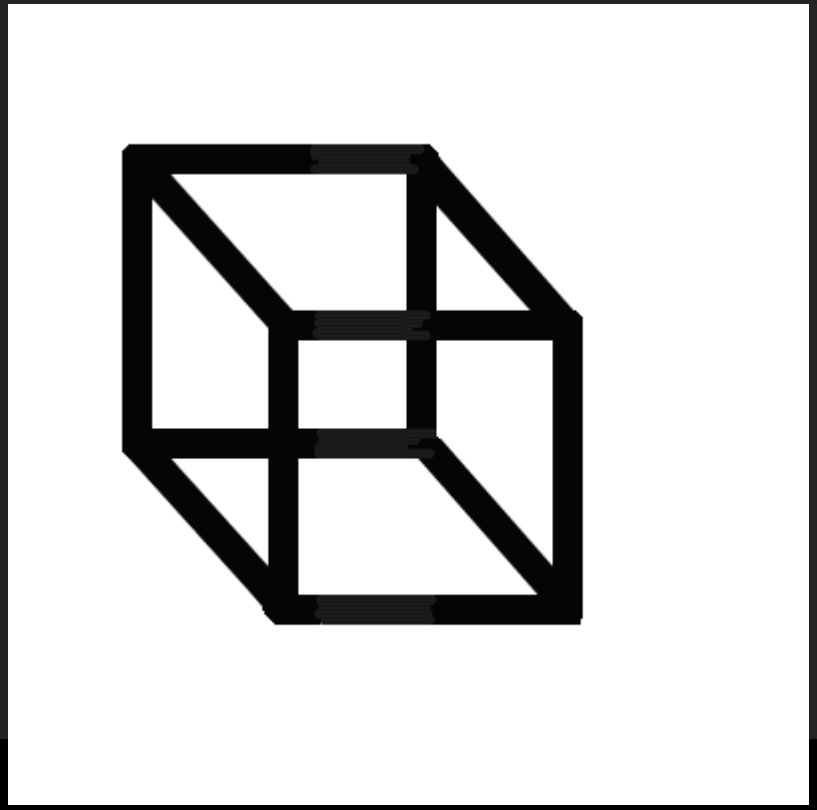
Fig. 4. Figure shows test error rates of both supervised and unsupervised learning on four topics: phonemes, gender, height and identity. Solid lines indicate $y = x$ in the coordinate systems. All data located along this line, meaning high correlation between supervised and unsupervised learning.

Uniqueness of representations?

Modern society's deep specialization requires efficient shared representations

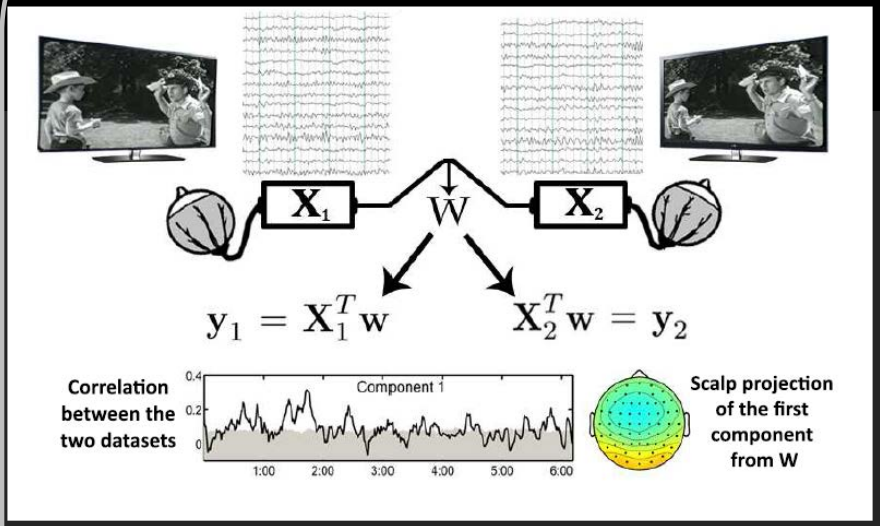
You know what I mean - right?

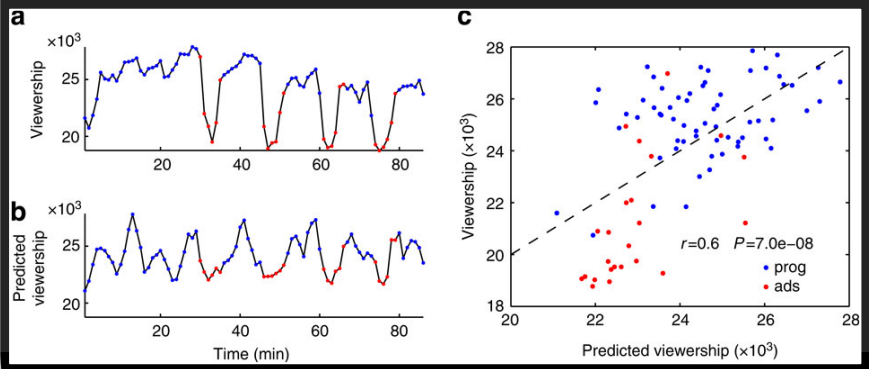
Does machine learning also develop shared representations and if so - why?

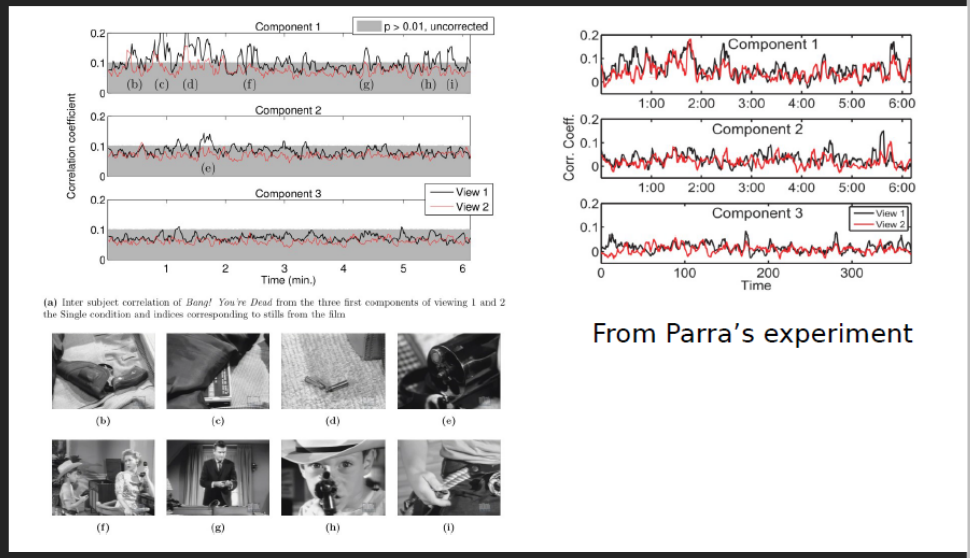


http://wisebytes.net/illusions/necker.php

Universality of attention to digital media



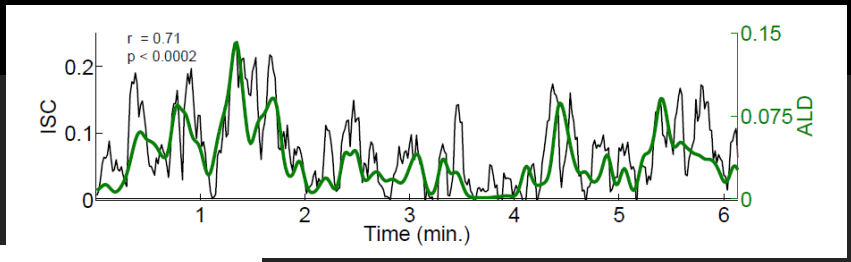




JP Dmochowski, P Sajda, J Dias, LC Parra, "Correlated components of ongoing EEG point to emotionally laden attention -a possible marker of engagement?" Frontiers of Human Neuroscience, 6:112, 2012.
JP Dmochowski, MA. Bezdek, BP. Abelson, JS. Johnson, EH Schumacher, LC Parra, "Audience preferences are predicted by temporal reliability of neural processing", Nature Communications 5:4567, 2014.
AT Poulsen, S, Kamronn, J Dmochowski, LC Parra, LK, Hansen:. "Measuring engagement in a classroom: Synchronised neural recordings during a video presentation". *arXiv preprint arXiv:1604.03019 (2016)*.

What is the joint attention signal?

Driven by early visual response
hich is modulated by attention...



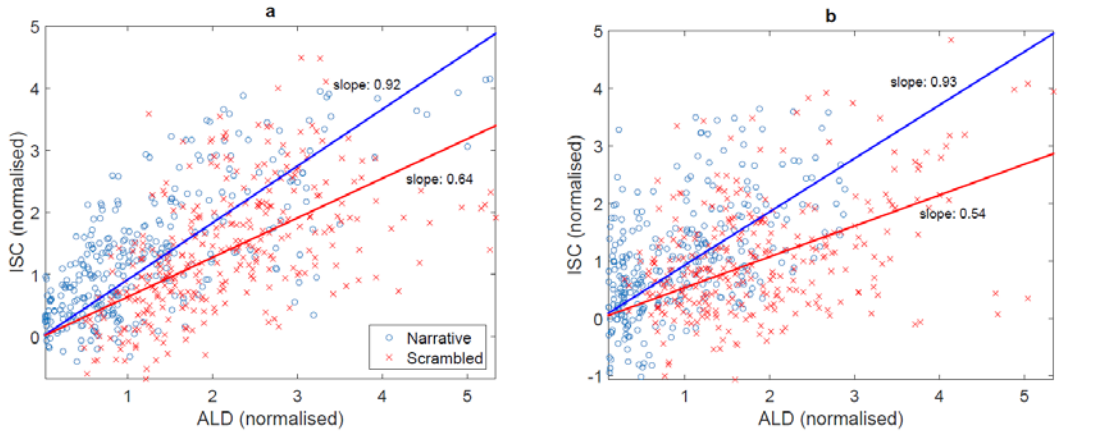


Figure 3. Assuming that the relation between the ISC obtained from the first CorrCA component while subjects where watching *Bang You're Dead* (a) and the ALD calculated from the film can be modeled as a simple proportionality with a factor ('gain') dependent on attention and similarly for *Sophies Choice* (b). A significant drop in the gain factors are found between the narrative sequence (blue) and the scrambled versions (red) ($p < 0.01$: block permutation test with block size $B = 25$sec).

Real-time feasible in (sub)-groups,
correlate with computed saliency...

Hiliard et al. Sensory gain control (amplification) as a mechanism of selective attention
Phil.Trans. R. Soc. Lond. B (1998) 353, 1257^1270

Ok, Deep network

Reducing the Dimensionality of Data with Neural Networks

G. E. Hinton* and R. R. Salakhutdinov

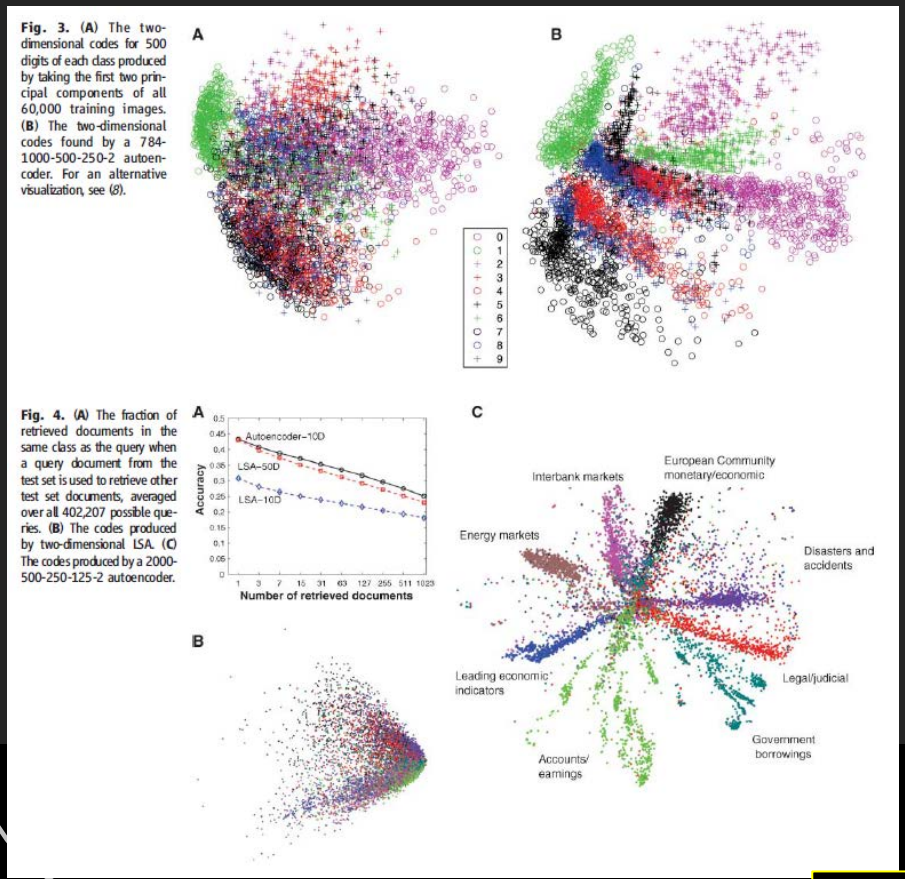


Fig. 3. (A) The two-dimensional codes for 500 digits of each class produced by taking the first two principal components of all 60,000 training images. (B) The two-dimensional codes found by a 784-1000-500-250-2 autoencoder. For an alternative visualization, see (8).

Fig. 4. (A) The fraction of retrieved documents in the same class as the query when a query document from the test set is used to retrieve other test set documents, averaged over all 402,207 possible queries. (B) The codes produced by two-dimensional LSA. (C) The codes produced by a 2000-500-250-125-2 autoencoder.

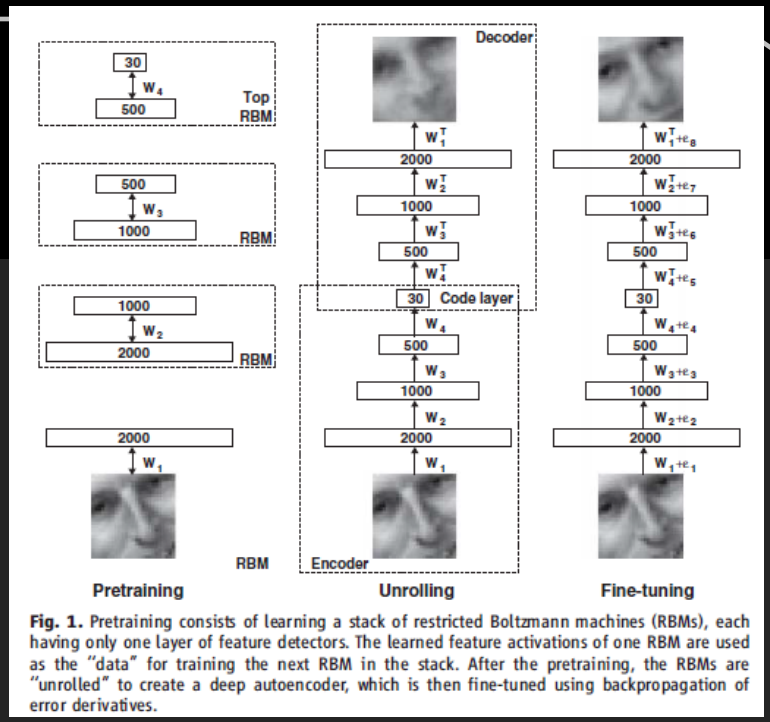


Fig. 1. Pretraining consists of learning a stack of restricted Boltzmann machines (RBMs), each having only one layer of feature detectors. The learned feature activations of one RBM are used as the "data" for training the next RBM in the stack. After the pretraining, the RBMs are "unrolled" to create a deep autoencoder, which is then fine-tuned using backpropagation of error derivatives.

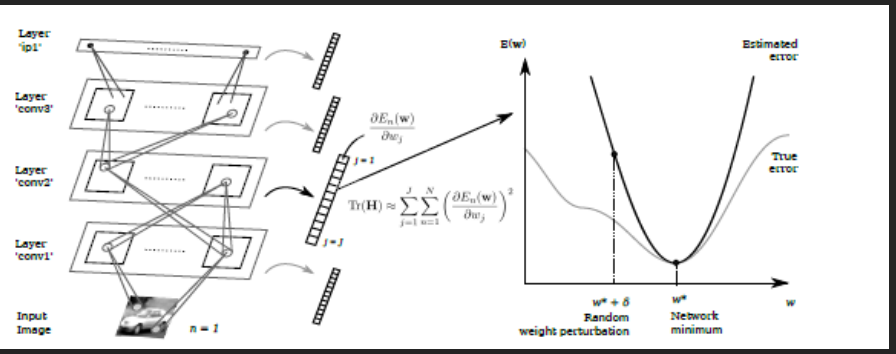
Hinton, Geoffrey E., and Ruslan R. Salakhutdinov. "Reducing the dimensionality of data with neural networks." *Science* 313, no. 5786 (2006): 504-507.

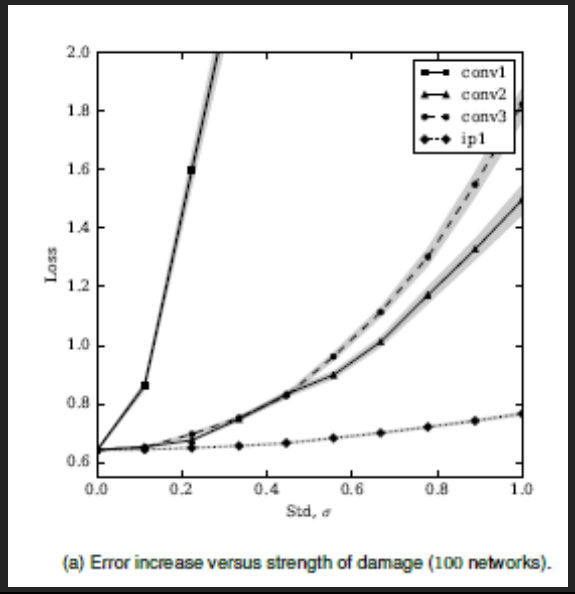
How 'well-determined' are the representations by ecology… sensitivity analysis

On the resilience of deep neural networks to weight damage

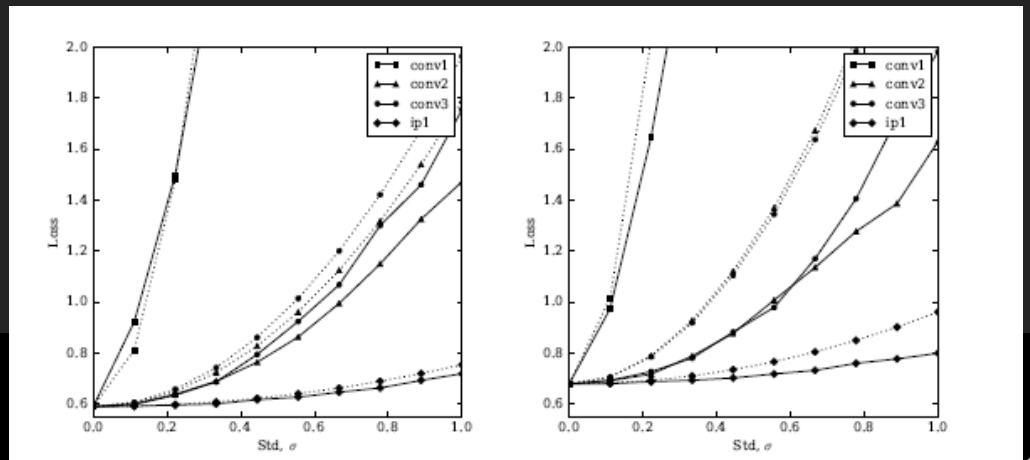
Rasmus Jessen Aaskov
Dept. of Applied Mathematics and Computer Science
Technical University of Denmark

Lars Kai Hansen
Dept. of Applied Mathematics and Computer Science
Technical University of Denmark





(a) Error increase versus strength of damage (100 networks).



(a) Example of a well performing network (20.25% final test error rate). (b) Example of a network with poor performance (23.71% final test error rate).

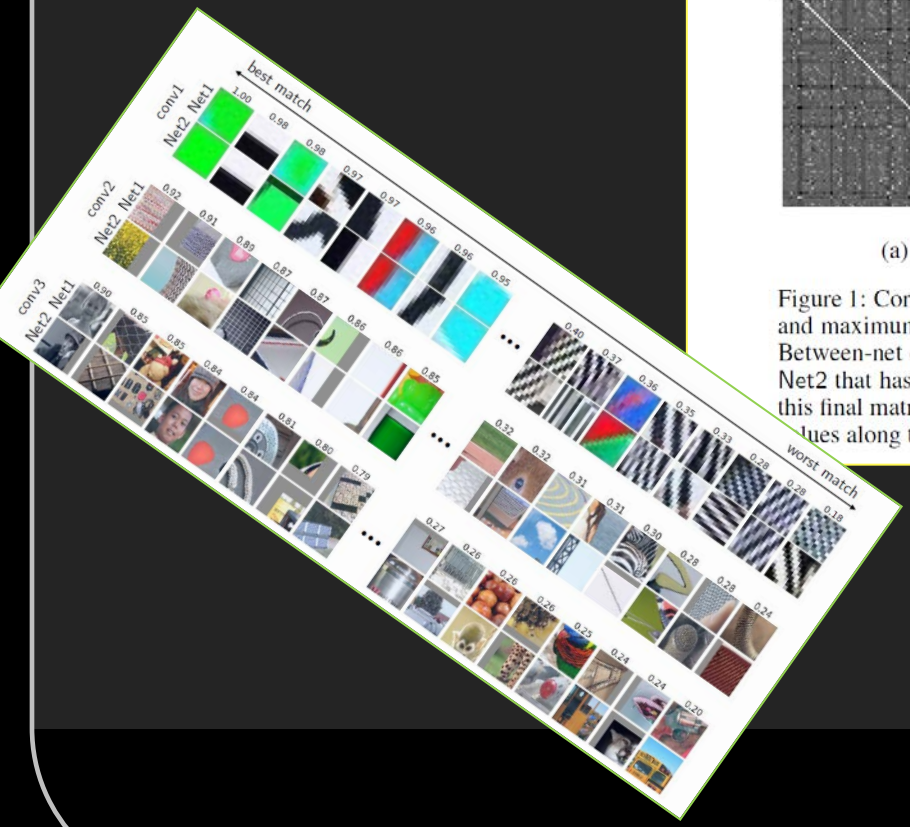
Figure 3. Two examples of how our method for estimating loss behaves in different cases. In both figures, the average loss of a repeated damage experiment and our estimated expected loss is shown for each of the individual layers. The expected loss is illustrated as a dotted line with distinctive markers for each layer.

RJ Aaskov, LK Hansen "On the resilience of deep neural networks to weight damage." *In review* (2016)

Published as a conference paper at ICLR 2016

CONVERGENT LEARNING: DO DIFFERENT NEURAL NETWORKS LEARN THE SAME REPRESENTATIONS?

Yixuan Li[1]*, Jason Yosinski[1]*, Jeff Clune[2], Hod Lipson[3], & John Hopcroft[1]



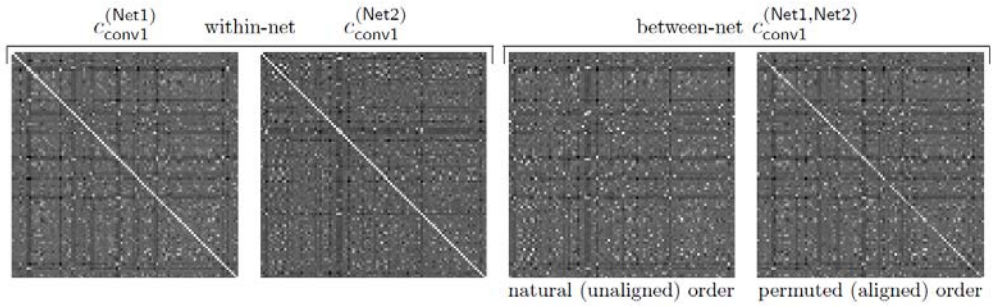
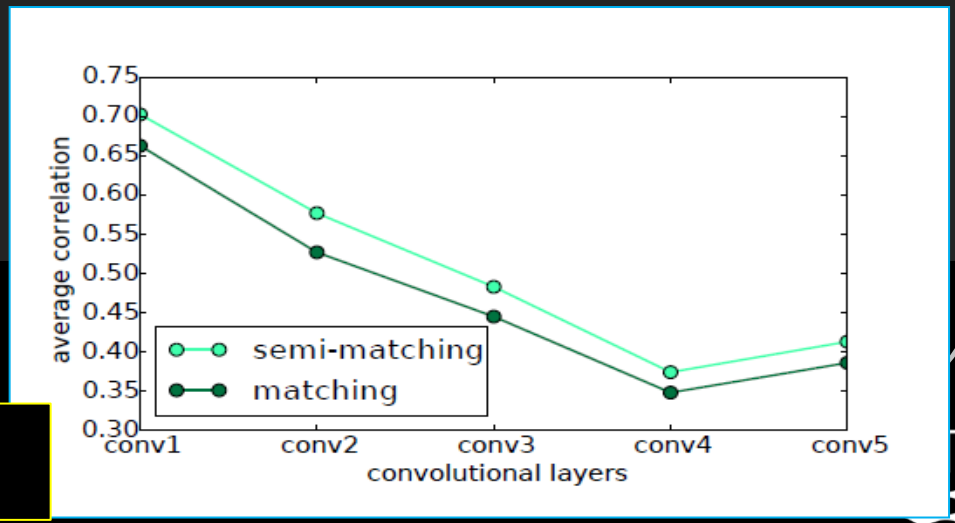
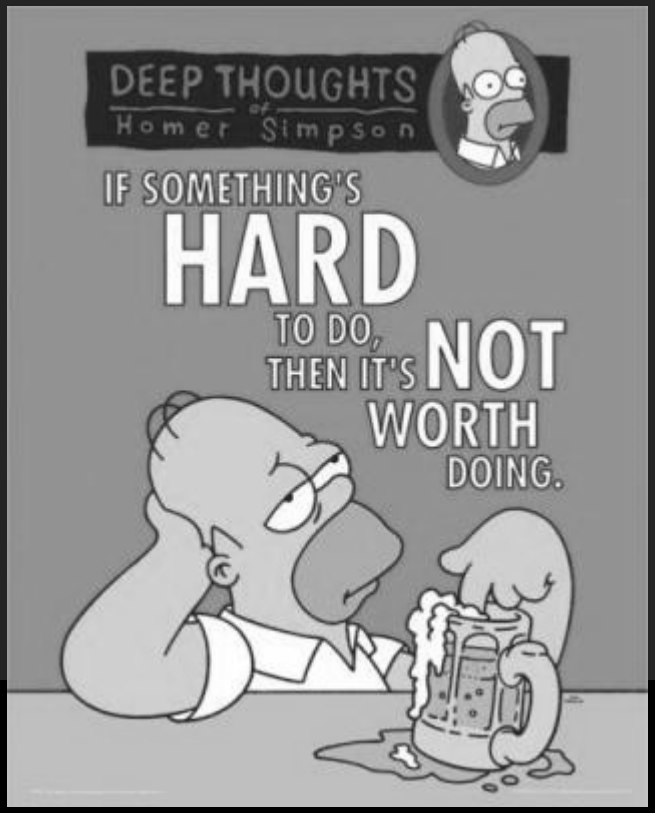


Figure 1: Correlation matrices for the conv1 layer, displayed as images with minimum value at black and maximum at white. **(a,b)** Within-net correlation matrices for Net1 and Net2, respectively. **(c)** Between-net correlation for Net1 vs. Net2. **(d)** Between-net correlation for Net1 vs. a version of Net2 that has been permuted to approximate Net1's feature order. The partially white diagonal of this final matrix shows the extent to which the alignment is successful; see Figure 3 for a plot of the values along this diagonal and further discussion.



L, Yixuan, J Yosinski, J Clune, H Lipson, J Hopcroft. "Convergent Learning: Do different neural networks learn the same representations?." *arXiv preprint arXiv:1511.07543* (2015)

What about “higher order” cognition?

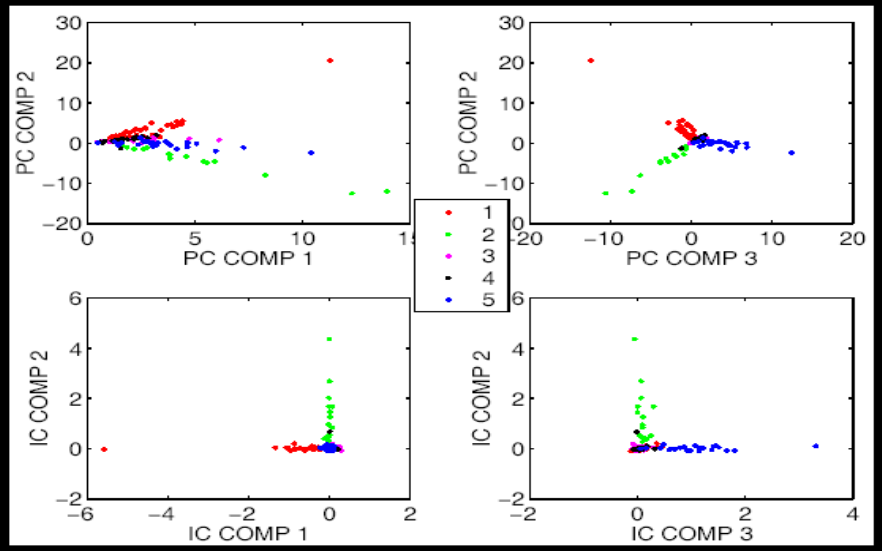


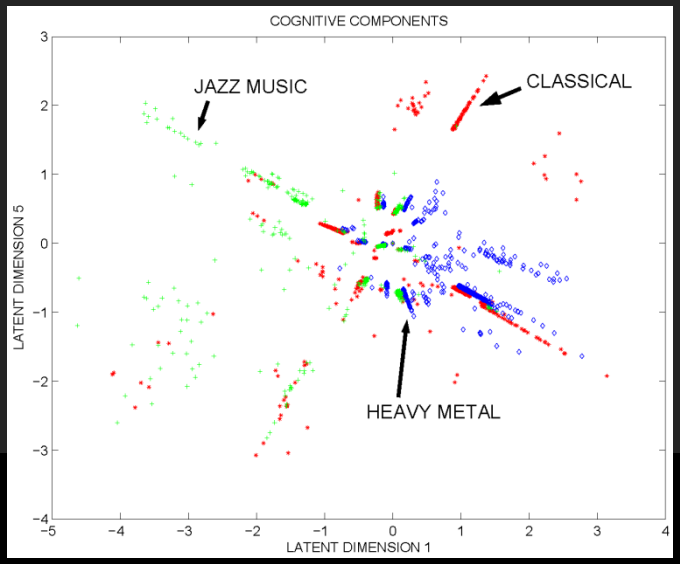
Found in translation...

Google's AI translation tool seems to have invented its own secret internal language

Johnson, M, et al: Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation. arXiv:1611.04558.

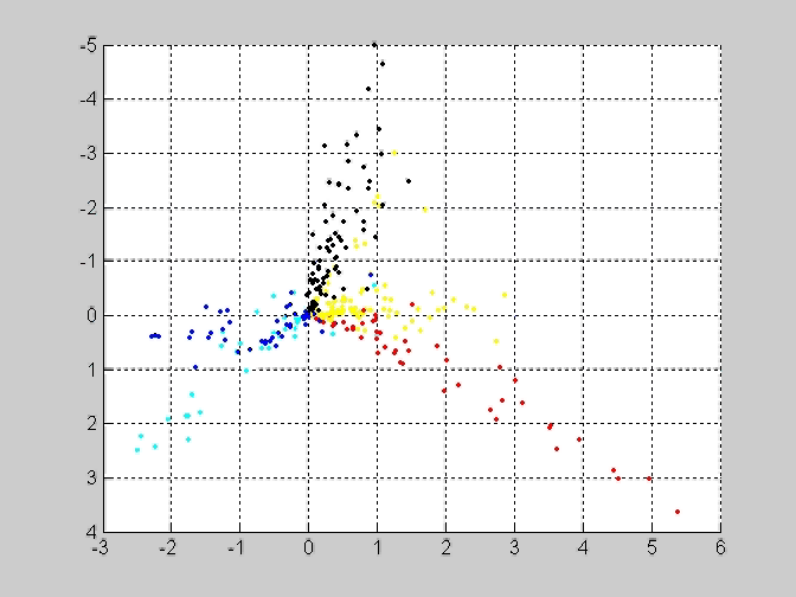
Linear mixture of independent contexts in term-document scatterplots





Linear mixture of independent contexts observed in short time features (mel-ceptrum) in a music database.

Social networks:
Linear mixtures of independent communities?

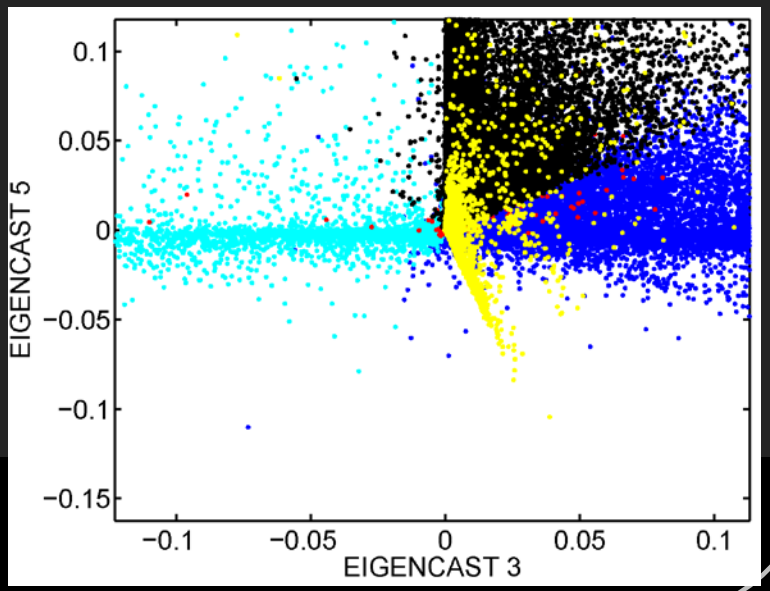


Genre patterns in expert's opinion on similar music artists

(AMG400, Courtesy D. Ellis)

"Movie actor network"
- A collaborative small world network

128.000 movies
380.000 actors



Hansen, L.K. and Feng, L., 2006. Cogito componentiter ergo sum.
In *Intl Conf on Independent Component Analysis and Signal Separation* (pp. 446-453). Springer Berlin.

DTU

CASTSEARCH - CONTEXT BASED SPEECH DOCUMENT RETRIEVAL

Lasse Lohilahti Mølgaard, Kasper Winther Jørgensen, and Lars Kai Hansen

Informatics and Mathematical Modelling
Technical University of Denmark Richard Petersens Plads
Building 321, DK-2800 Kongens Lyngby, Denmark

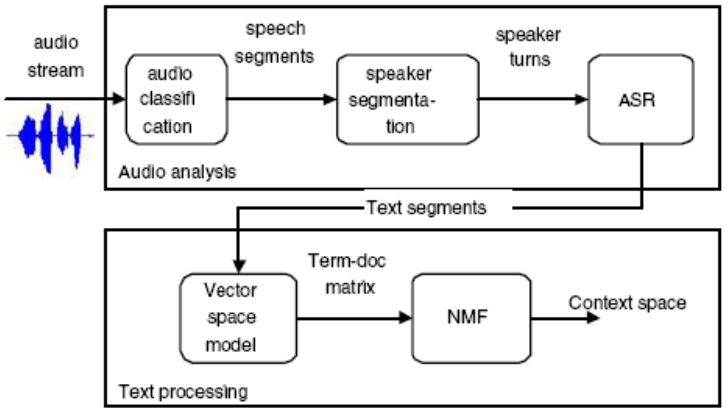
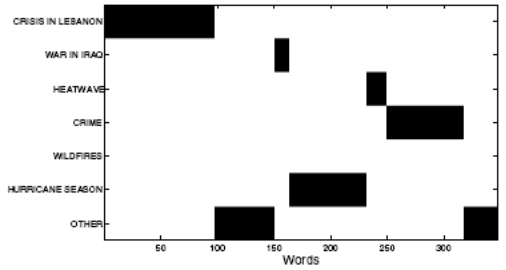
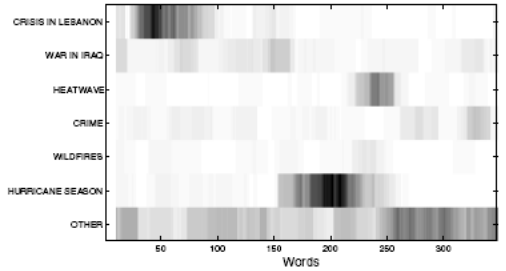


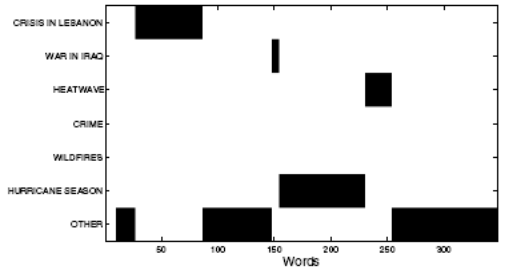
Fig. 1. The system setup. The audio stream is first processed using audio segmentation. Segments are then using an automatic speech recognition (ASR) system to produce text segments. The text is then processed using a vector representation of text and apply non-negative matrix factorization (NMF) to find a topic space.



(a) Manual segmentation.



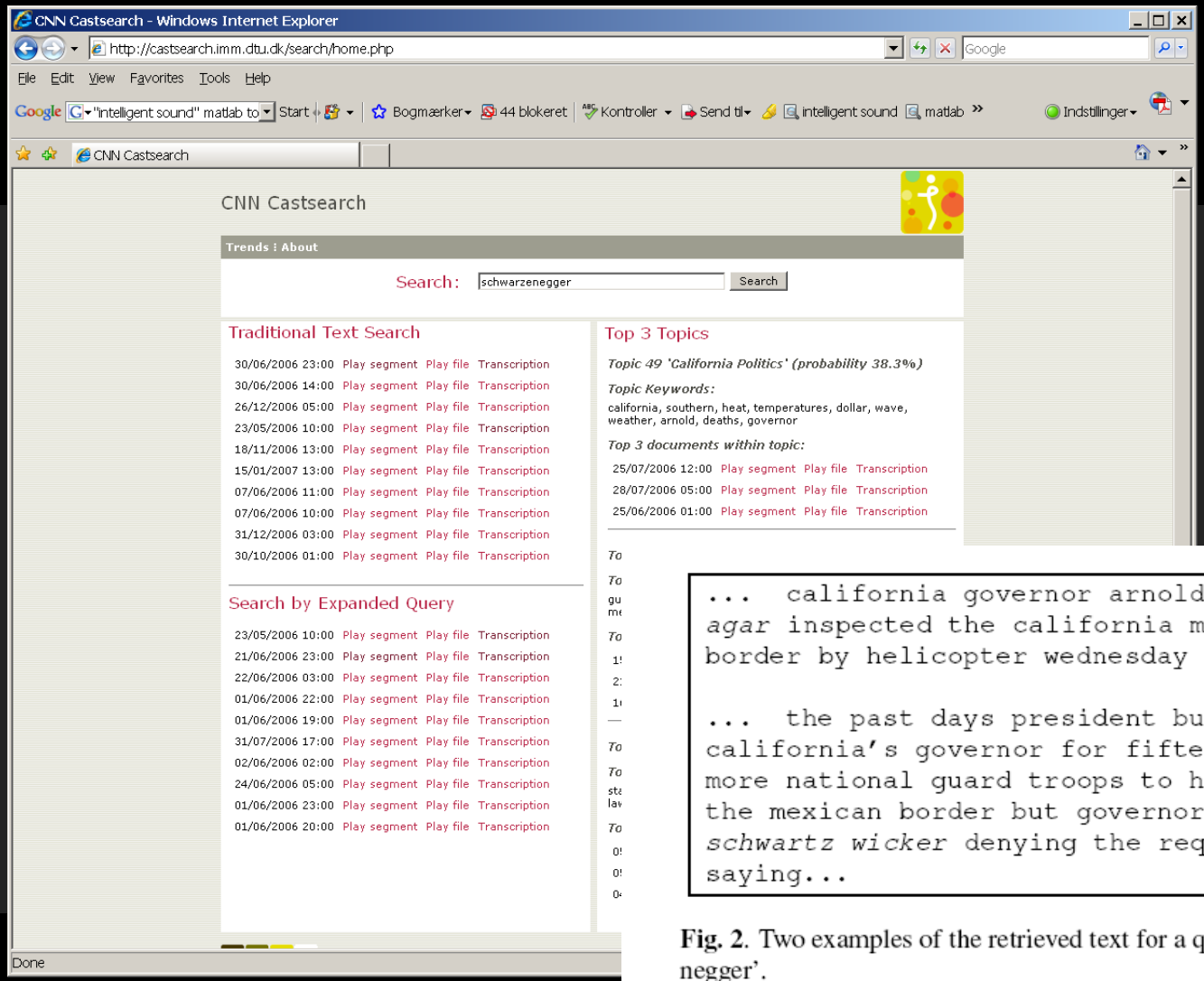
(b) $p(k|d^*)$ for each context. Black means high probability.



(c) The segmentation based on $p(k|d^*)$.

Fig. 3. Figure 3(a) shows the manual segmentation of the news show into 7 classes. Figure 3(b) shows the distribution $p(k|d^*)$ used to do the actual segmentation shown in figure 3(c). The NMF-segmentation is in general consistent with the manual segmentation. Though, the segment that is manually segmented as 'crime' is labeled 'other' by the NMF-segmentation

Mølgaard et al. 2007



... 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'.

castsearch.imm.dtu.dk

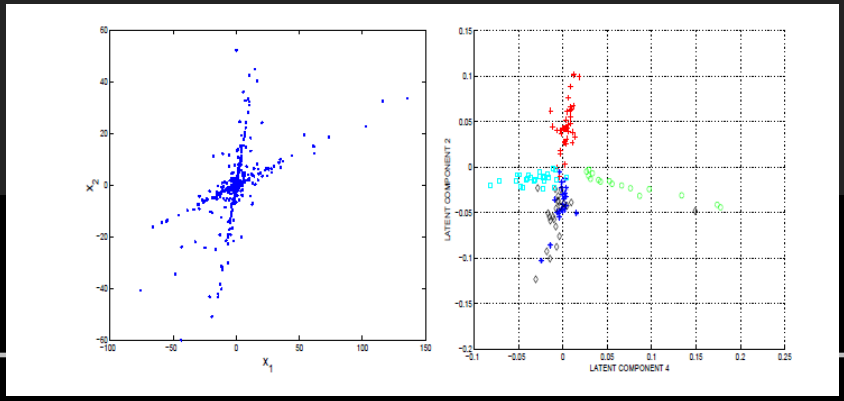
Conclusions & outlook

To take machine to next level of human like behavior we need to understand better what human like behavior is based on...

Evidence that phonemes, gender, identity are independent components 'objects' in the (time stacked) MFCC representation

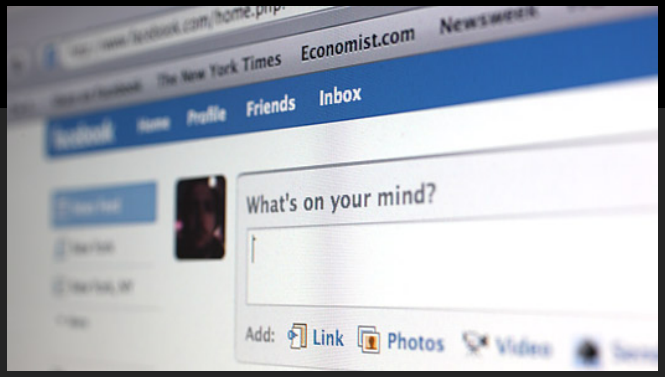
Evidence that human categorization is based on sparse independent components in social networks, text, digital media

Conjecture: Objects in digital media can be identified as "independent components:" The brain uses old tricks from perception to solve complex "modern" problems.



Acknowledgments

- Danish Research Councils
- Innovation fund DK
- EU Commission



For software and demos:
DTU:ICA toolbox (www.imm.dtu.dk/cisp)