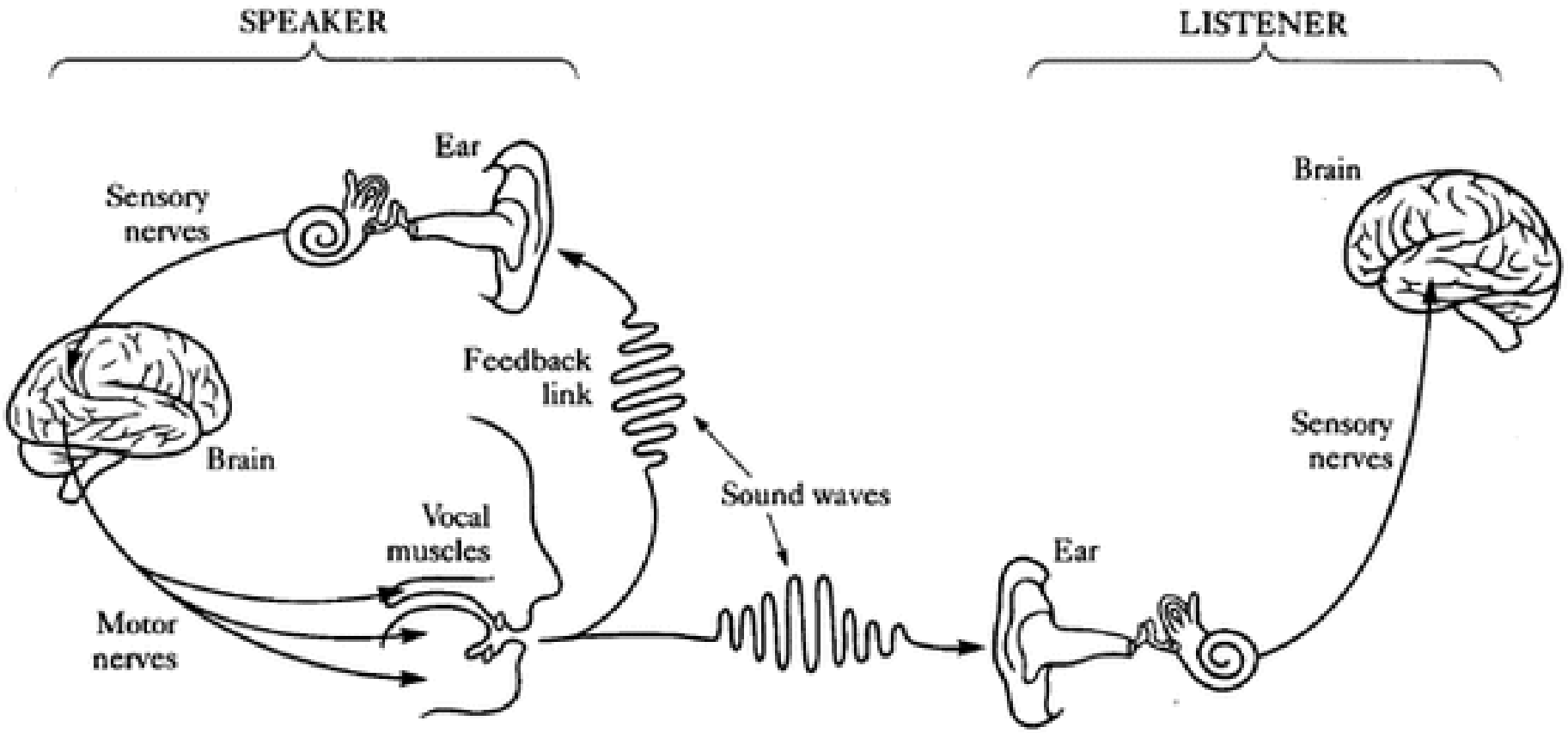
The human perspective
&
Explainability

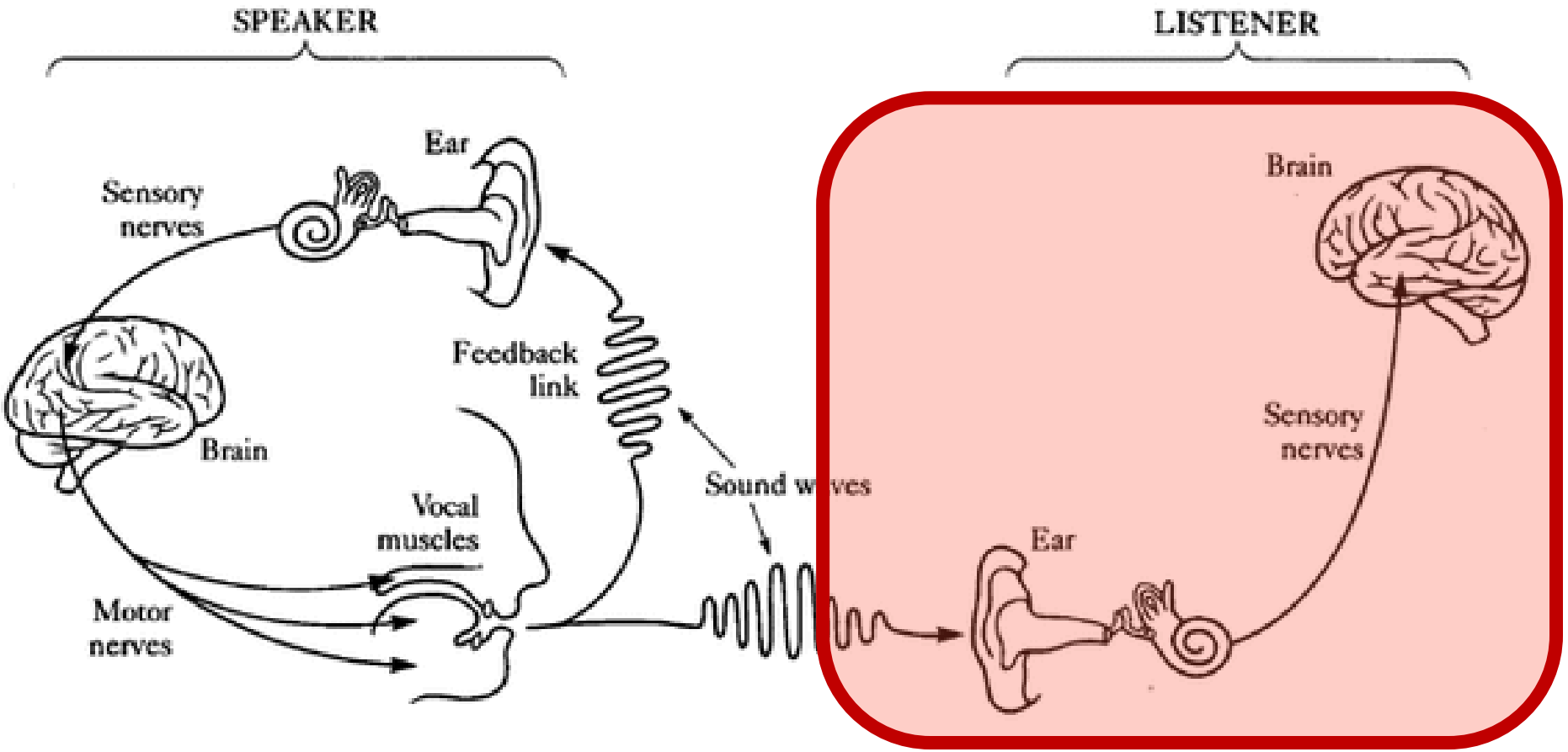
Louis ten Bosch
Nijmegen/Graz 2025

Topic 1.
The human perspective…

Speaker-listener loop



Speaker-listener loop

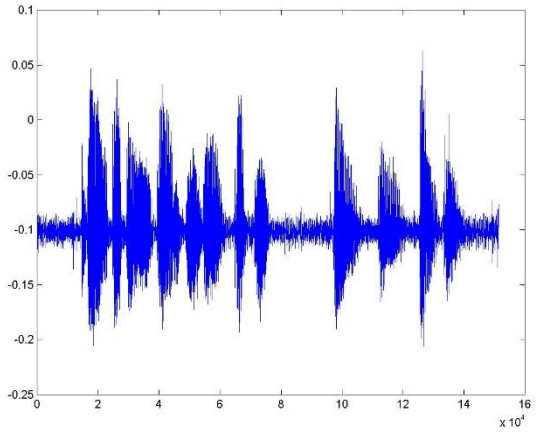


speaker

Concept

Grammar
Mental lexicon

Muscle
activation,
articulation



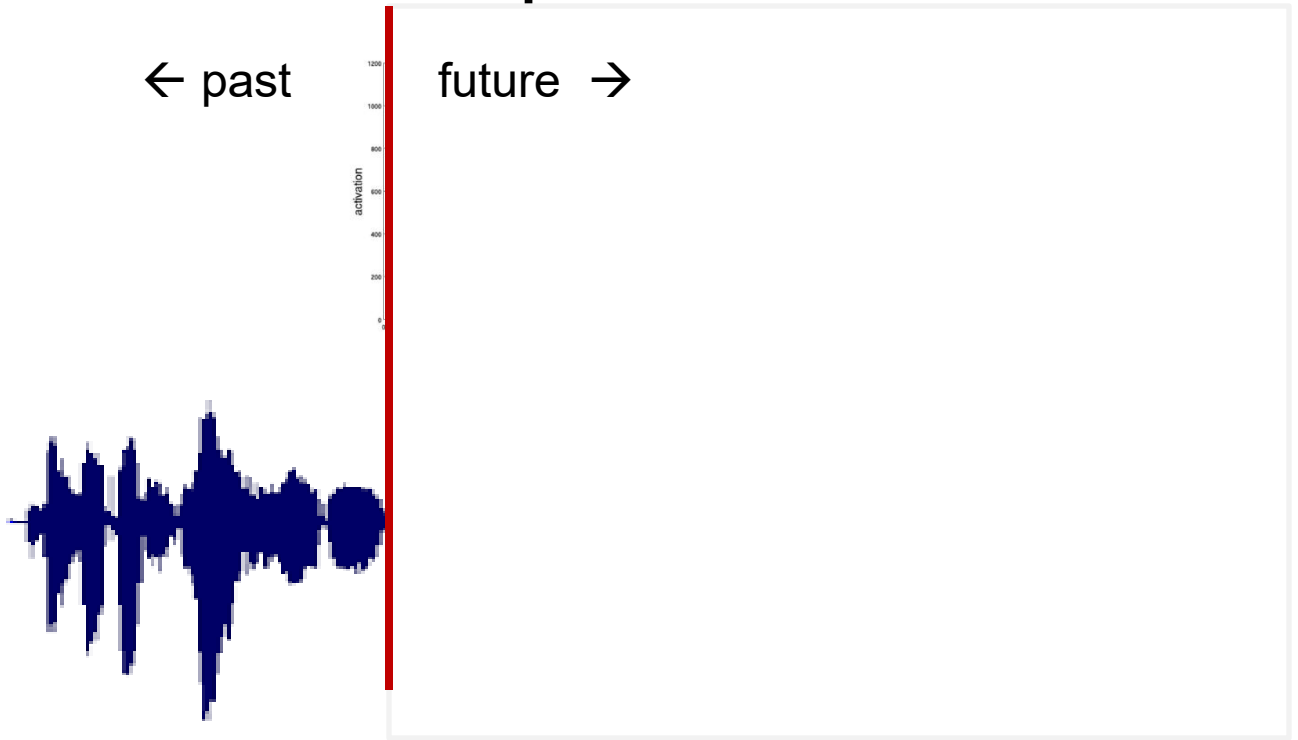
listener

Concept

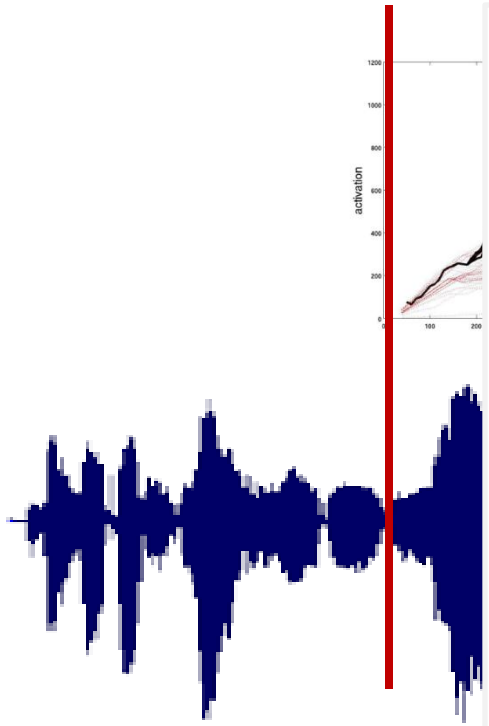
Grammar
Mental
lexicon

Auditory
pathway

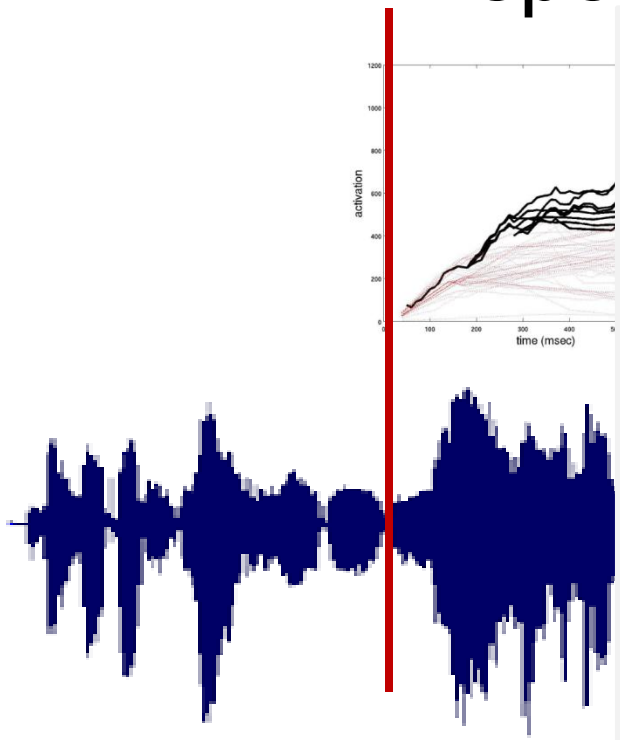
Comprehension of continuous speech



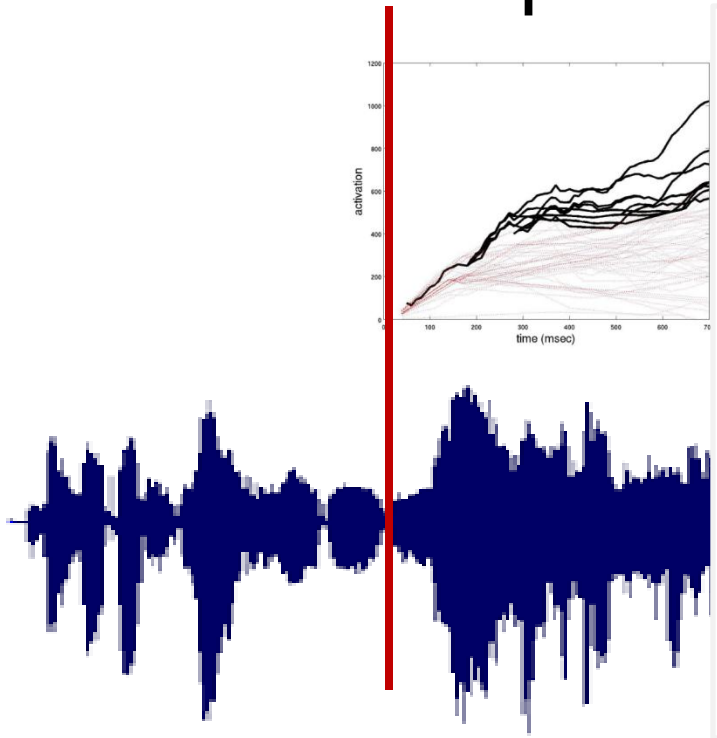
Comprehension of continuous speech



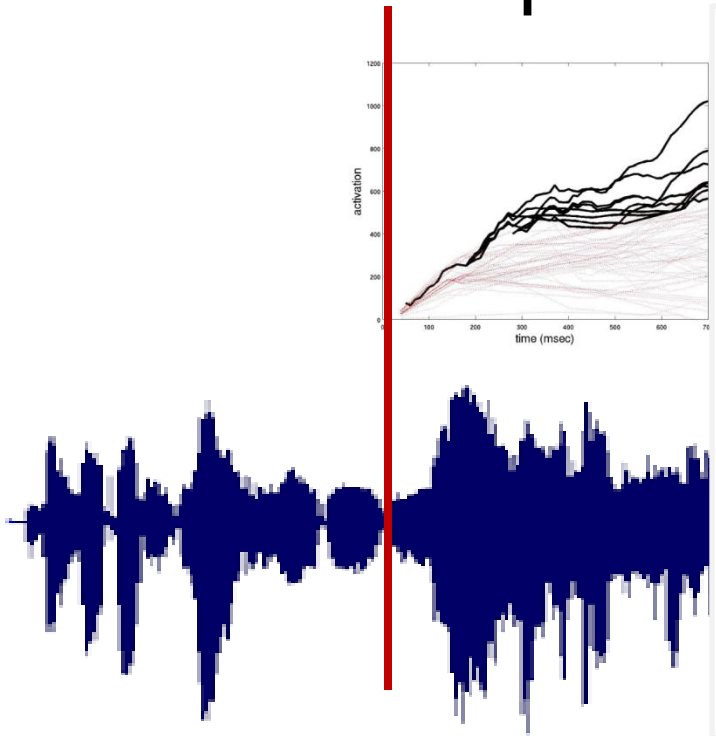
Comprehension of continuous speech



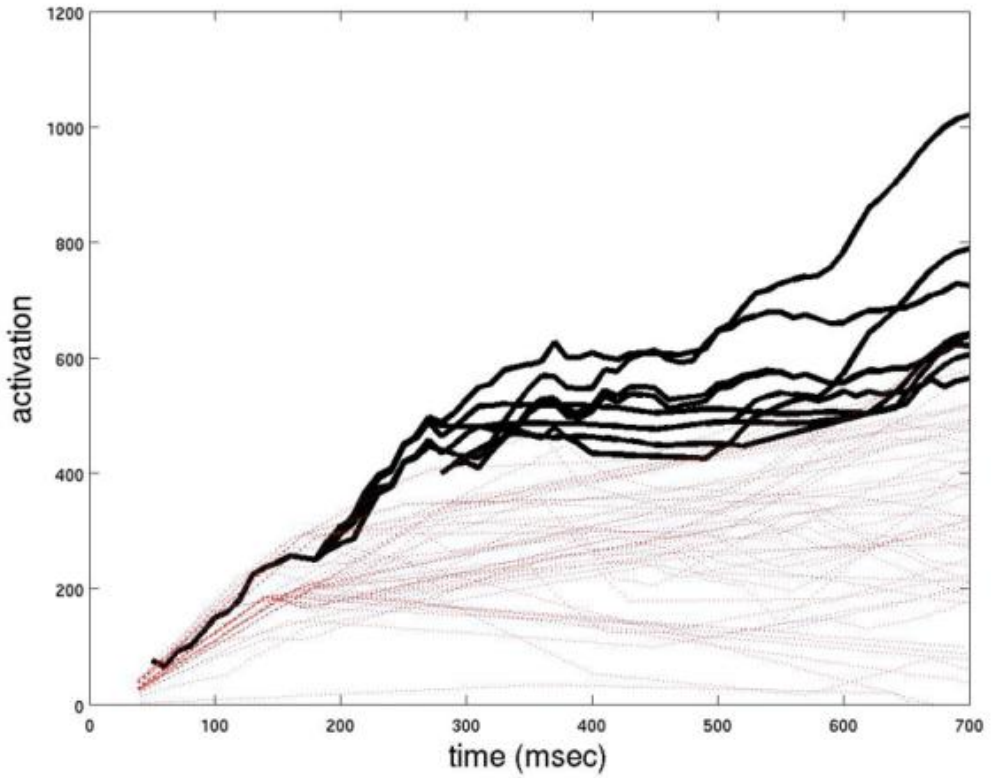
Comprehension of continuous speech



Comprehension of continuous speech



Activation

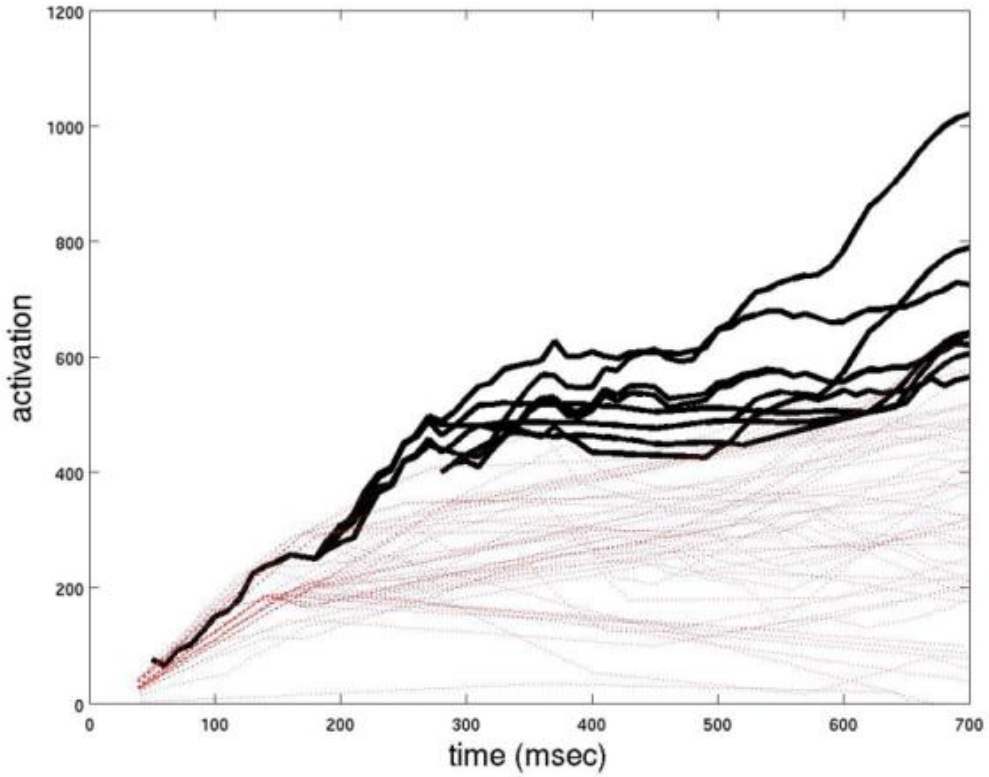


Words compete in the listener's head

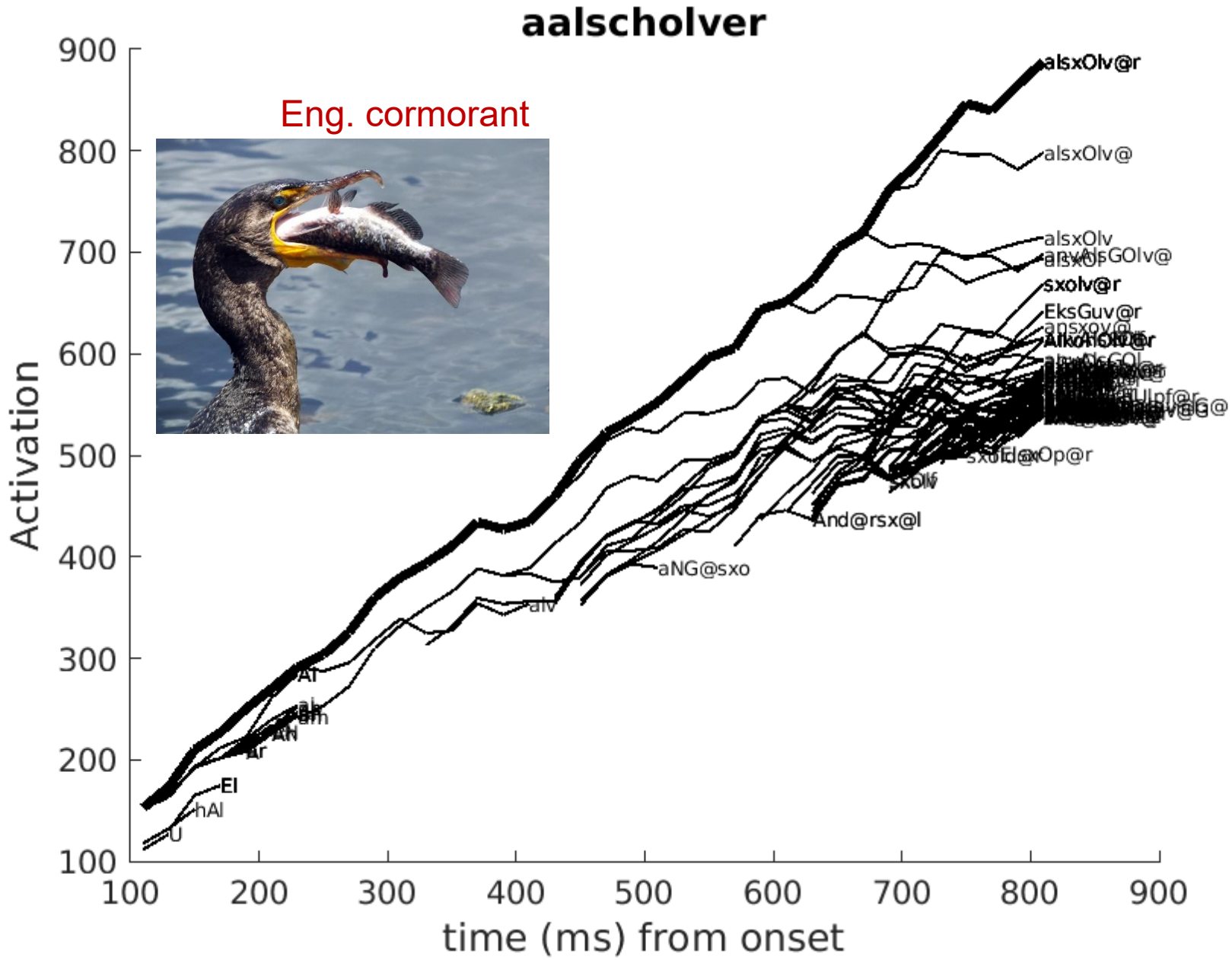
You have to select a winner 3 times a second

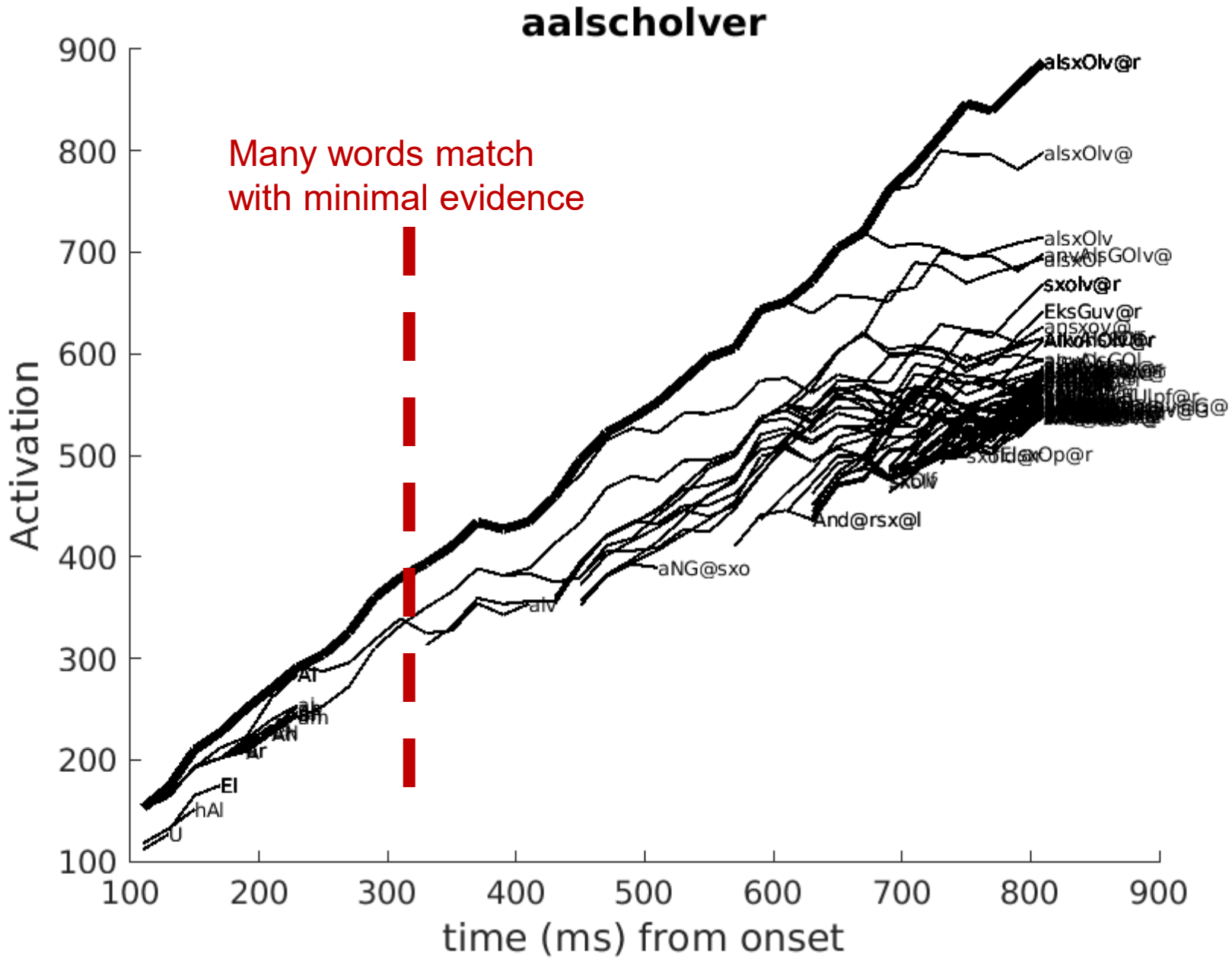
±60000 (±100) options per word for L1

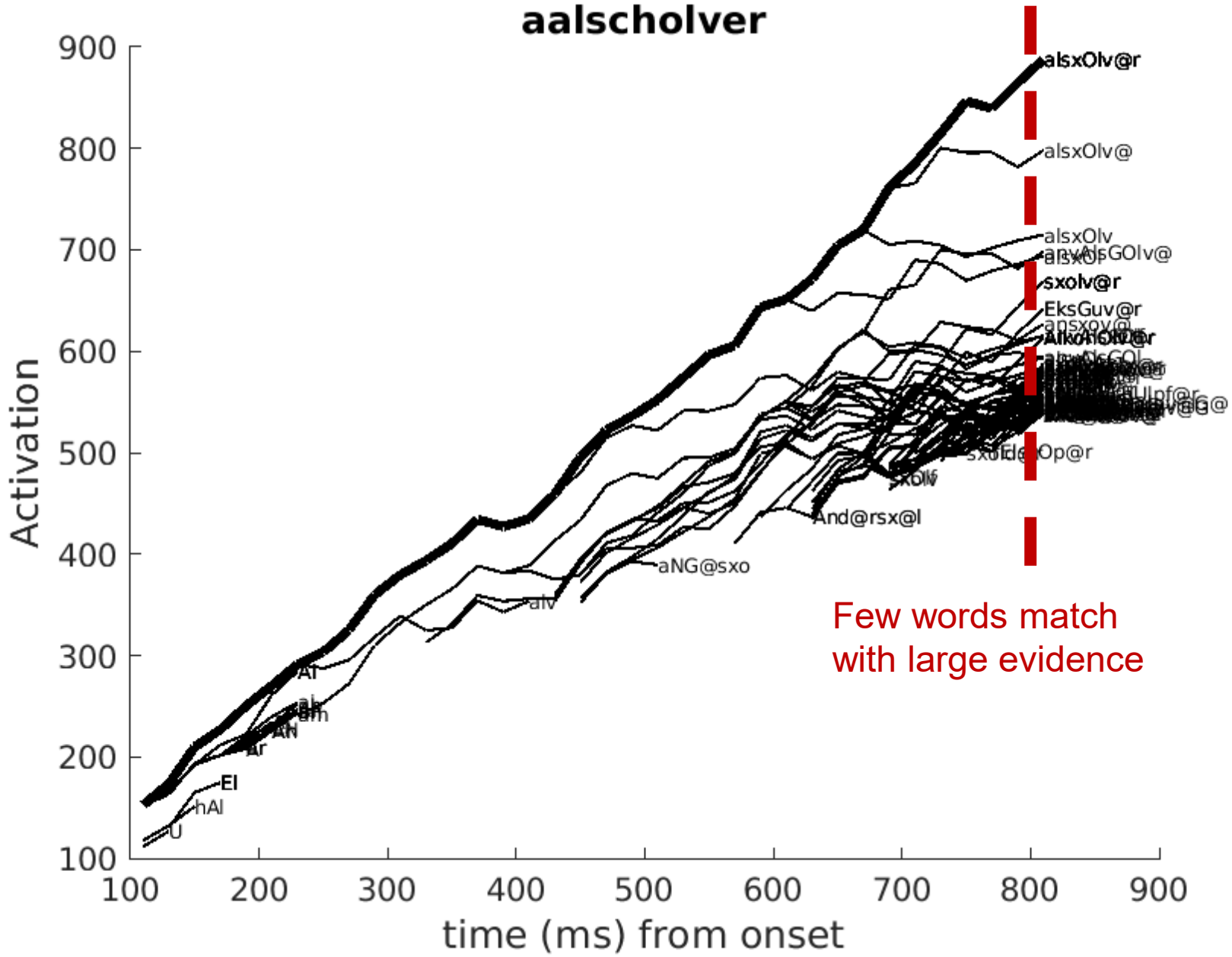
Activation

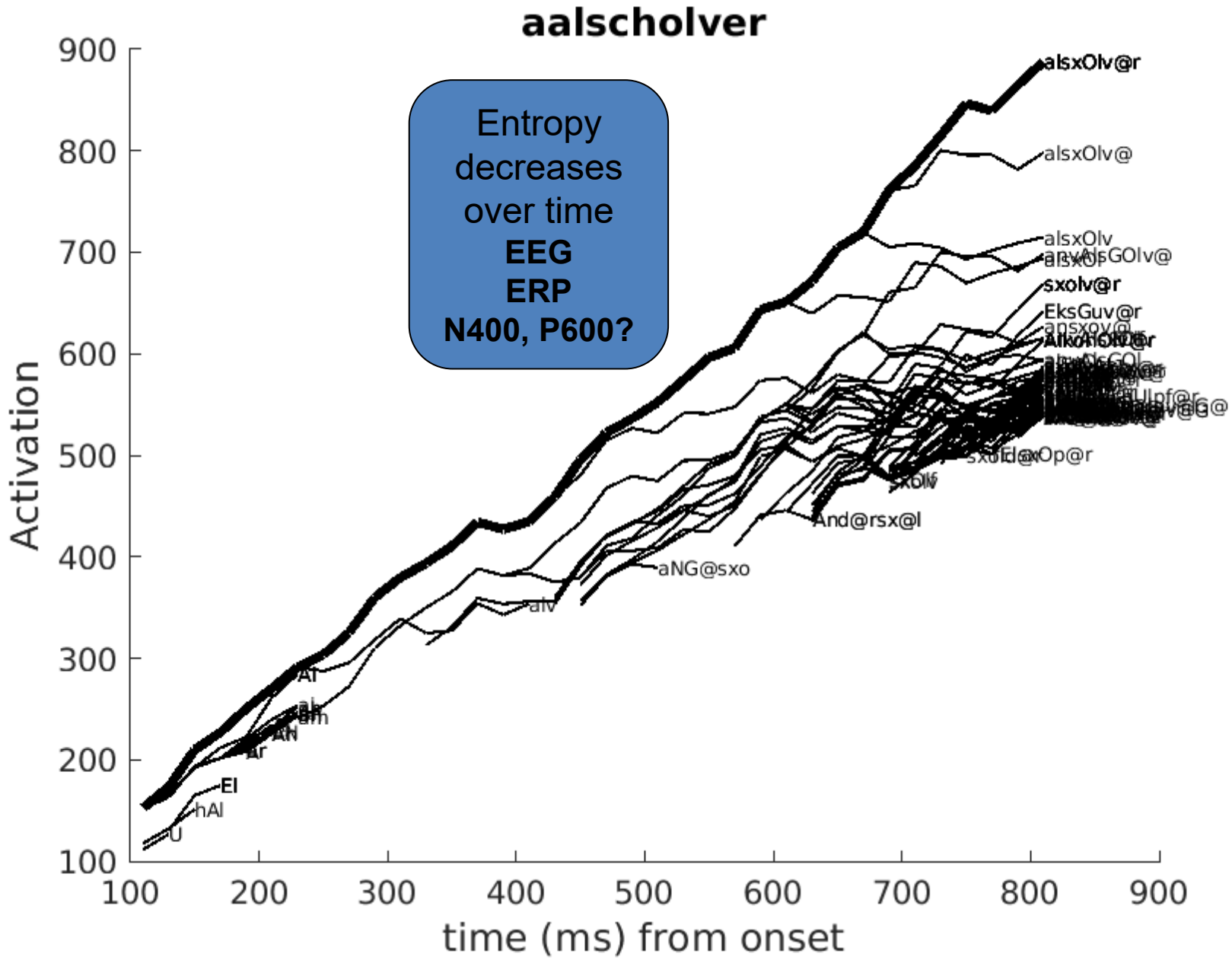


Activation = log(P(signal|word)) + λ log P(word)
Example here: duration 0.68 sec.
Shown: top 200 candidates, recomputed each 10 ms


aalscholver
Eng. cormorant
Activation
time (ms) from onset
100
200
300
400
500
600
700
800
900
alsxOlv@r
alsxOlv@
alsxOlv
sxolv@r
EksGuv@r
aNG@sxo
And@rsx@l
hAl


aalscholver
Many words match
with minimal evidence
Activation
time (ms) from onset
100
200
300
400
500
600
700
800
900
alsxOlv@r
alsxOlv@
alsxOlv
sxolv@r
EksGuv@r
aNG@sxo
And@rsx@l
hAl


aalscholver
Activation
time (ms) from onset
100
200
300
400
500
600
700
800
900
alsxOlv@r
alsxOlv@
alsxOlv
sxolv@r
EksGuv@r
aNG@sxo
And@rsx@l
Few words match
with large evidence


aalscholver
Entropy decreases over time
EEG
ERP
N400, P600?
Activation
time (ms) from onset
100
200
300
400
500
600
700
800
900
alsxOlv@r
alsxOlv@
alsxOlv
sxolv@r
EksGuv@r
aNG@sxo
And@rsx@l
hAl

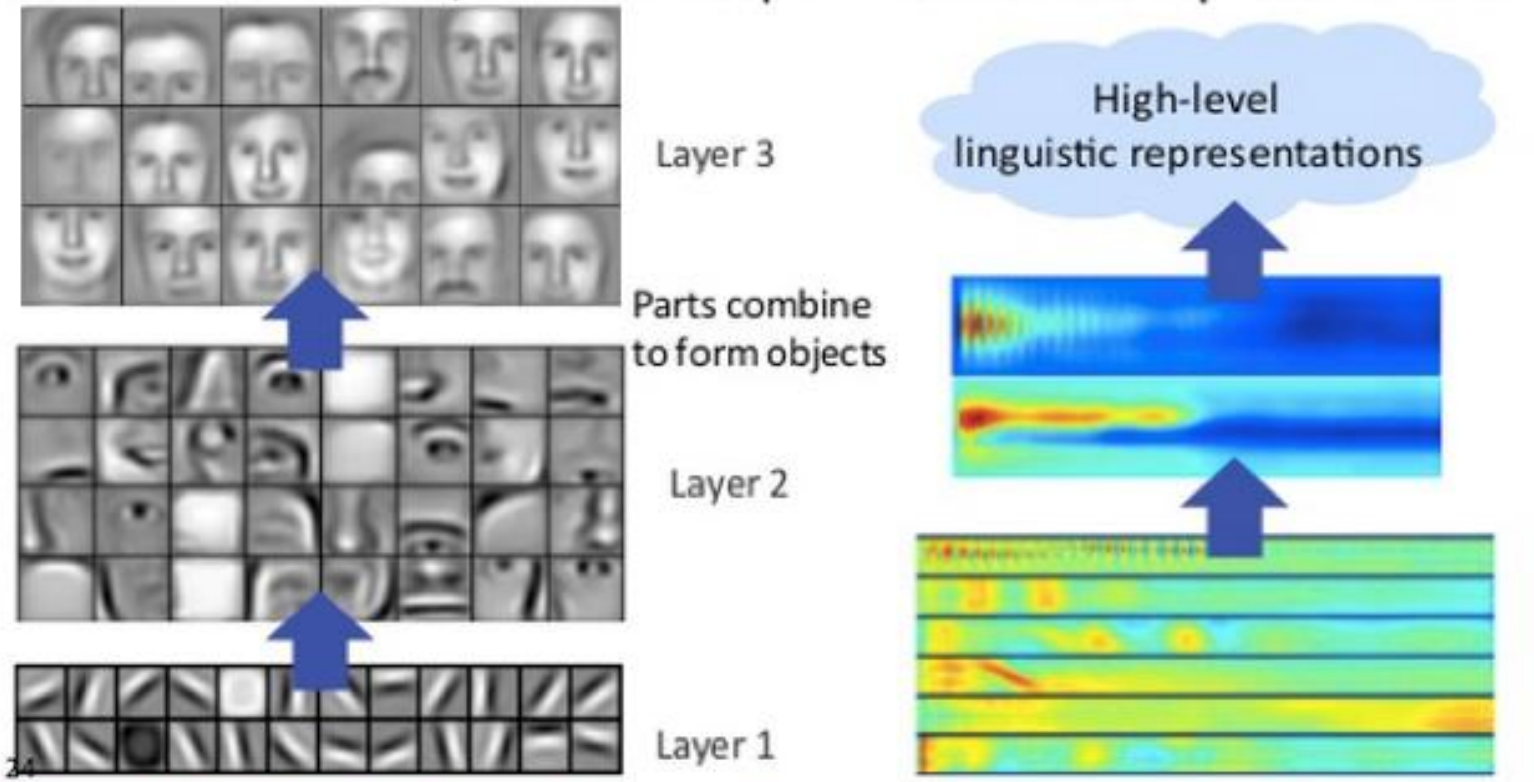
Topic 2.
Four views of Deep Networks

Why DNN are useful: view 1

- DNNs may discover structure in data sets because subsequent layers ignore more and more details that are irrelevant for correctly predicting output labels
- Increasingly abstract representations emerge by cascading multiple (nonlinear) transformations

  - so far, most convincing in image classification

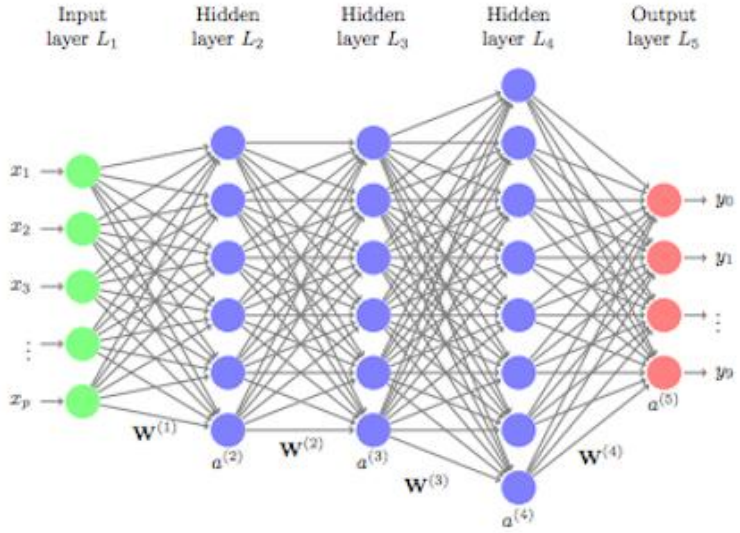
Successive model layers learn deeper intermediate representations



Prior: underlying factors & concepts compactly expressed w/ multiple levels of abstraction

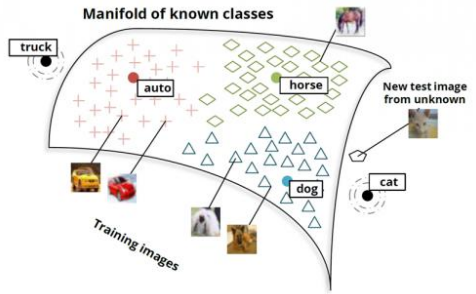
View 2

- focus on the role of DNNs to find optimal representations, in particular in the sense of features.
- learning of optimal representations can be achieved if the network is able to disentangle the underlying explanatory factors hidden in the observed data (Bengio et al., 2013)



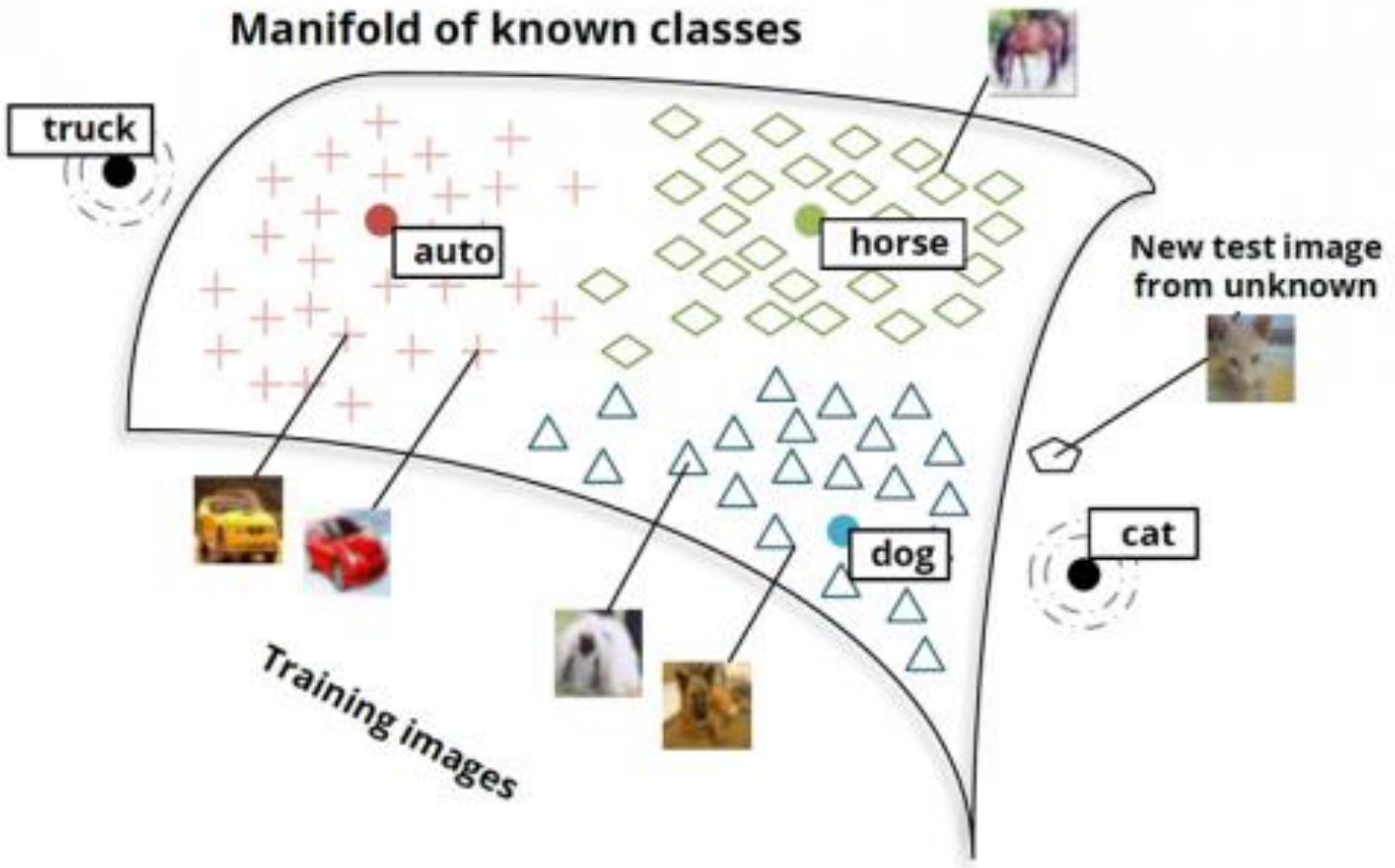
View 3

- more geometrically inspired interpretation of a DNN is based on the manifold assumption
  - the application of manifold learning methods on speech signals is (also) based on the relatively slow ballistic movements of articulators



- Deep networks may provide a very effective method for dimensionality reduction (remember the bottleneck!)
- directions tangent to the manifold are well preserved while directions orthogonal to the manifolds aren't
- DNNs are related to **manifold learning** (Tenenbaum, Bengio, Singh Tomar, ….).

A manifold


Manifold of known classes
truck
auto
horse
New test image
from unknown
cat
dog
Training images

View 4

- A fourth approach is more theoretical and analyzes DNNs on an ’information plane’ using ’information bottleneck’
- Any DNN can be characterized by the mutual information between a hidden layer and the input and output variables, as a function of hidden layer depth
- Tishby et al. argue that the optimal architecture (number of layers and features/connections at each layer) is related to the bifurcation points of the information bottleneck plane.

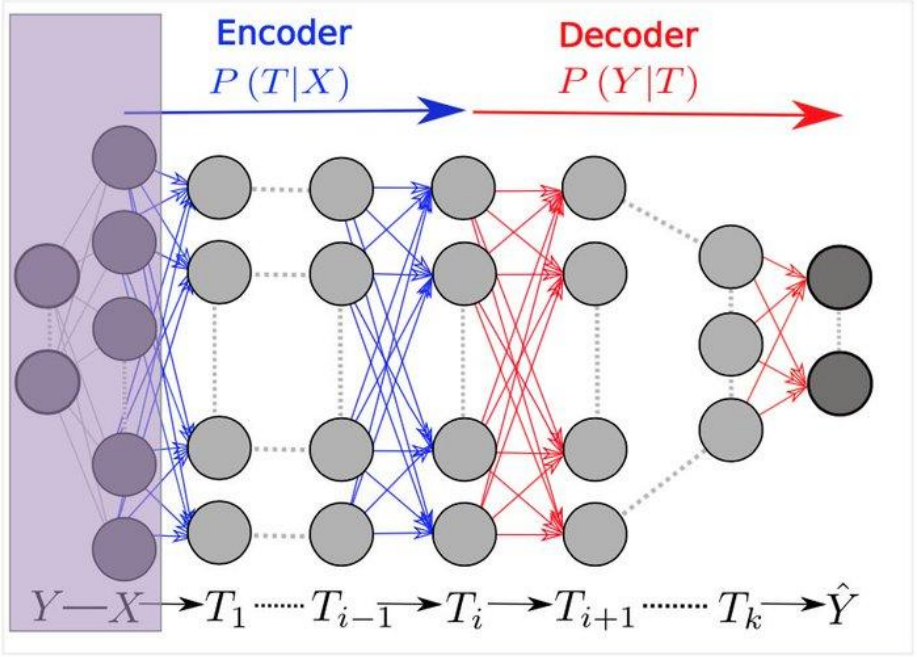


Figure 1: The DNN layers form a Markov chain of successive internal representations of the input layer X. Any representation of the input, T, is defined through an encoder, $P(T|X)$, and a decoder $P(\hat{Y}|T)$, and can be quantified by its *information plane* coordinates: $I_X = I(X;T)$ and $I_Y = I(T;Y)$. The Information Bottleneck bound characterizes the optimal representations, which maximally compress the input X, for a given mutual information on the desired output Y. After training, the network receives an input X, and successively processes it through the layers, which form a Markov chain, to the predicted output $\hat{Y}$. $I(Y;\hat{Y})/I(X;Y)$ quantifies how much of the relevant information is captured by the network.

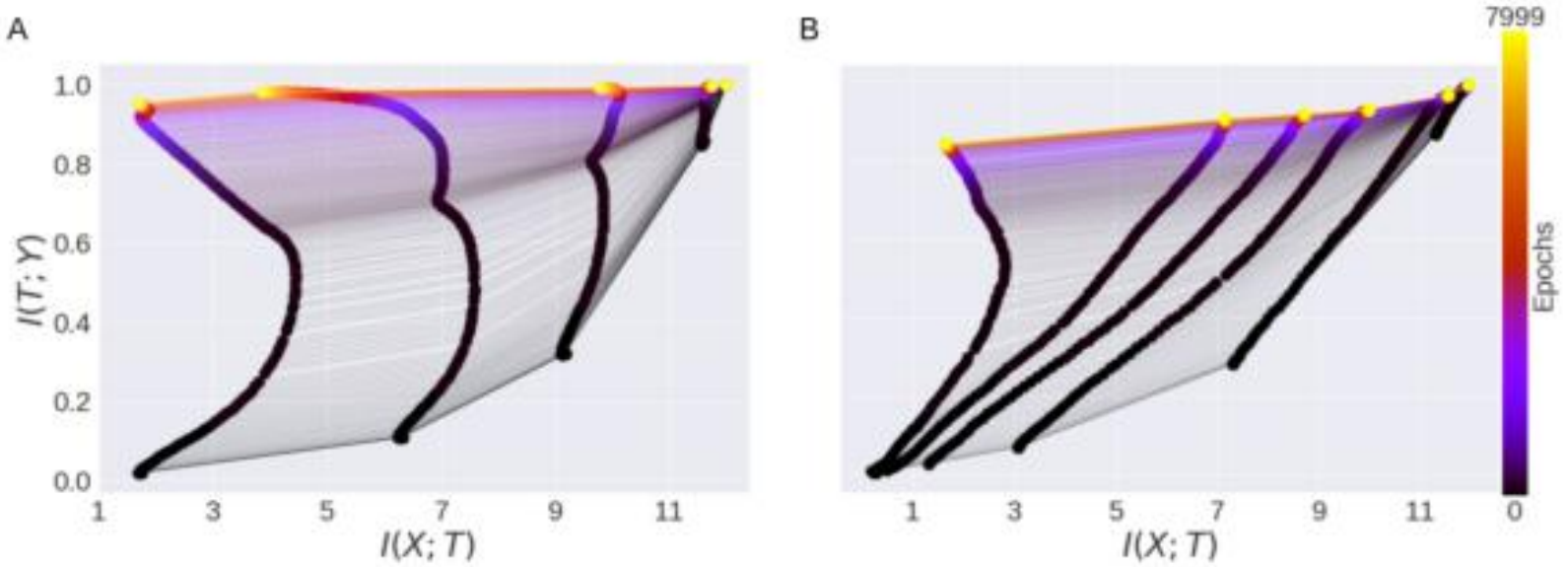


Figure 1: Information plane dynamics and neural nonlinearities. (A) Replication of Shwartz-Ziv & Tishby (2017) for a network with tanh nonlinearities (except for the final layer which contains sigmoidal neurons). The x-axis plots information between each layer and the input, while the y-axis plots information between each layer and the output. The color scale indicates training time in epochs. Each of the six layers produces a curve in the information plane with the input layer at far right, output layer at the far left. Different layers at the same epoch are connected by fine lines. (B) Information plane dynamics with ReLU nonlinearities (except for the final layer of 2 sigmoidal neurons). Here no compression phase is visible in the ReLU layers. For learning curves of both networks, see Appendix A