DONDERS INSTITUTE

Probing neural representations with task-optimized deep neural networks

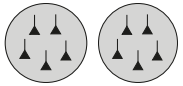
Umut Güçlü

Radboud University Radboudumc

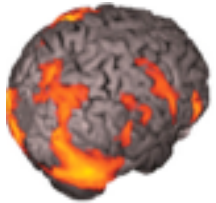
Modeling how the brain responds to the environment

stimulus **x**

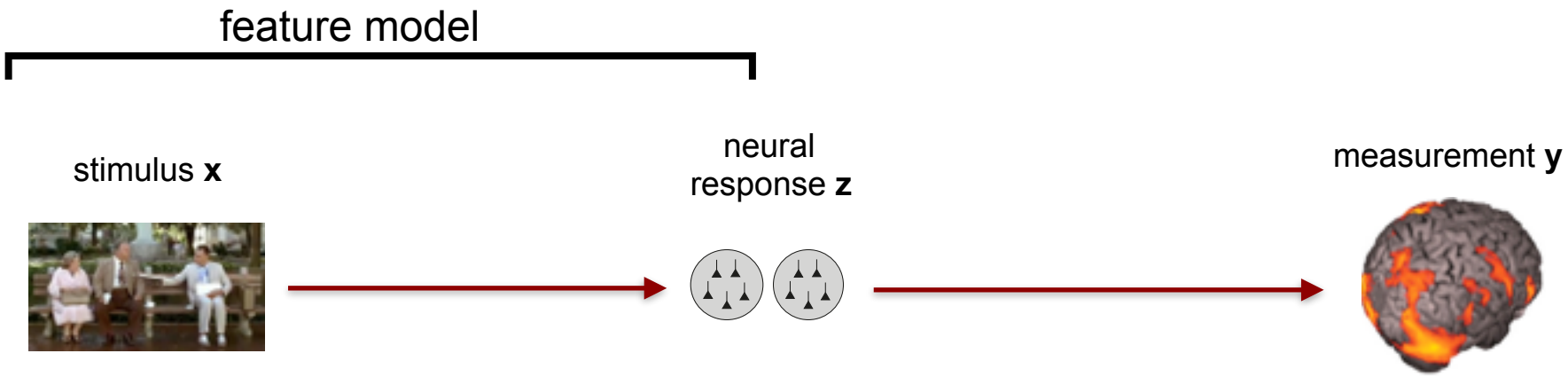
neural response **z**



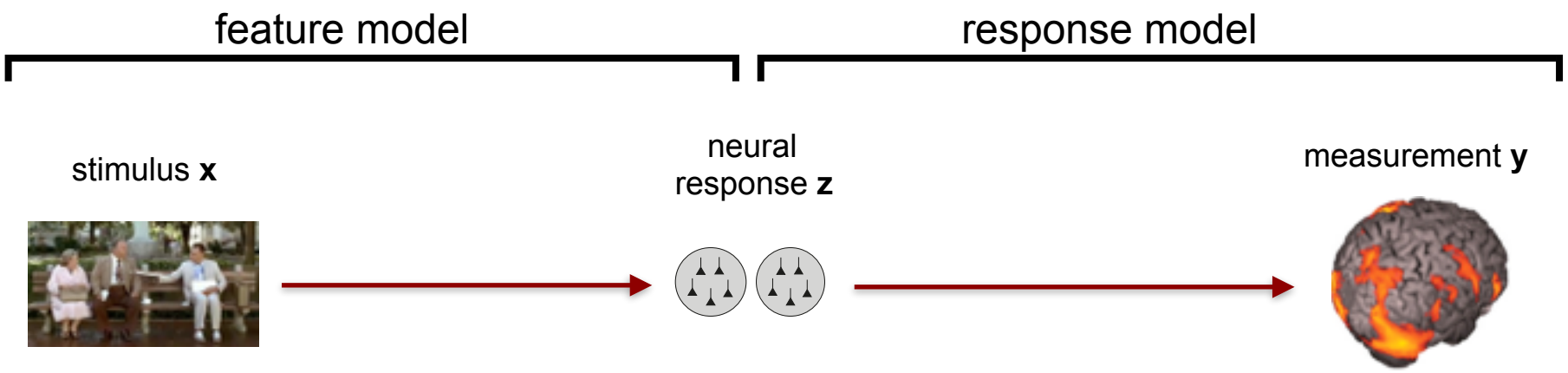
measurement **y**



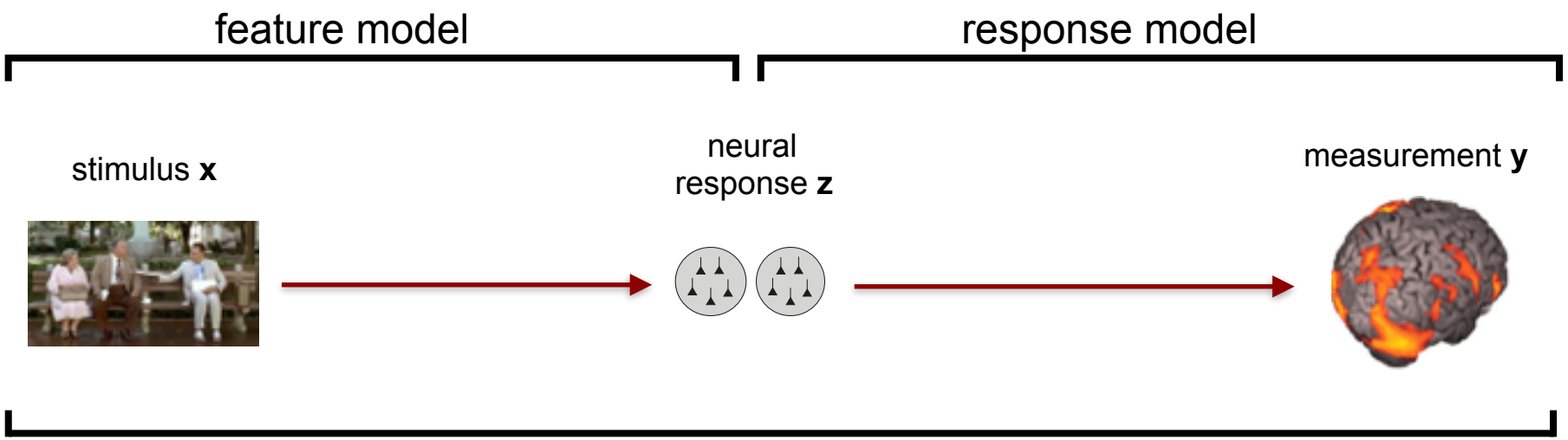
Modeling how the brain responds to the environment



Modeling how the brain responds to the environment



Modeling how the brain responds to the environment



Encoding model

$$p(\mathbf{y} \mid \mathbf{x})$$

Probing neural representations

Probing neural representations

Probing neural representations

Probing neural representations

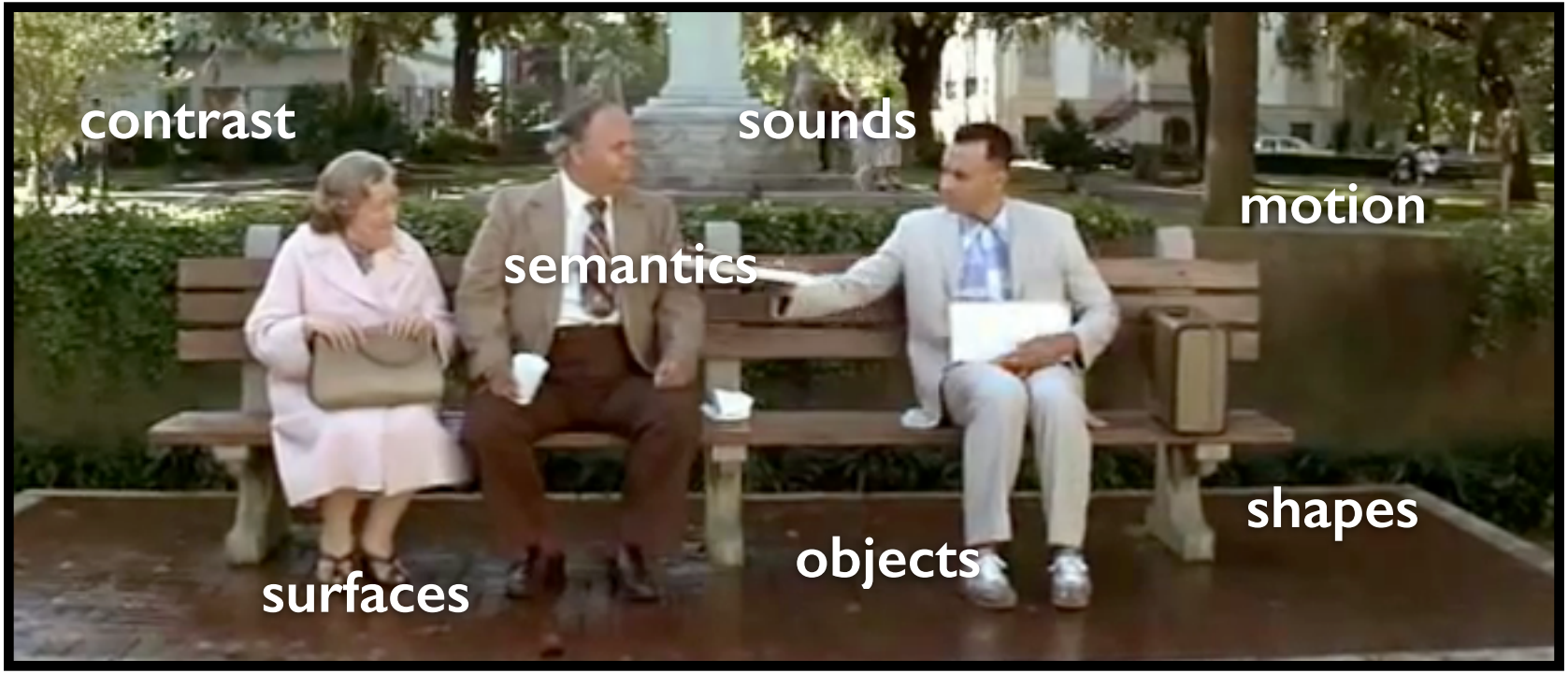
Probing neural representations

Probing neural representations

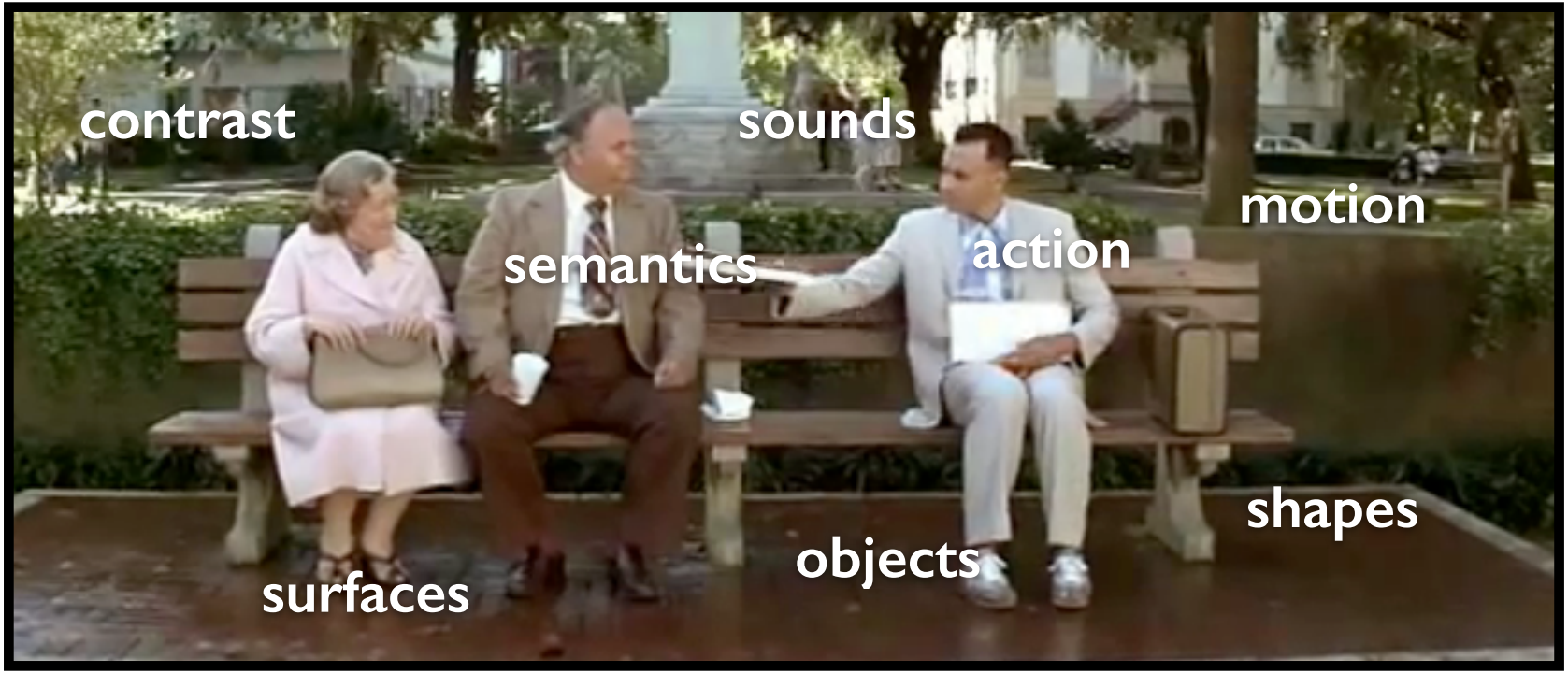
Probing neural representations

Probing neural representations

Probing neural representations



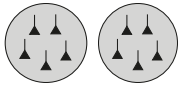
Probing neural representations



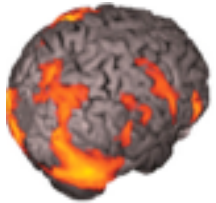
Modeling how the brain responds to the environment

stimulus **x**

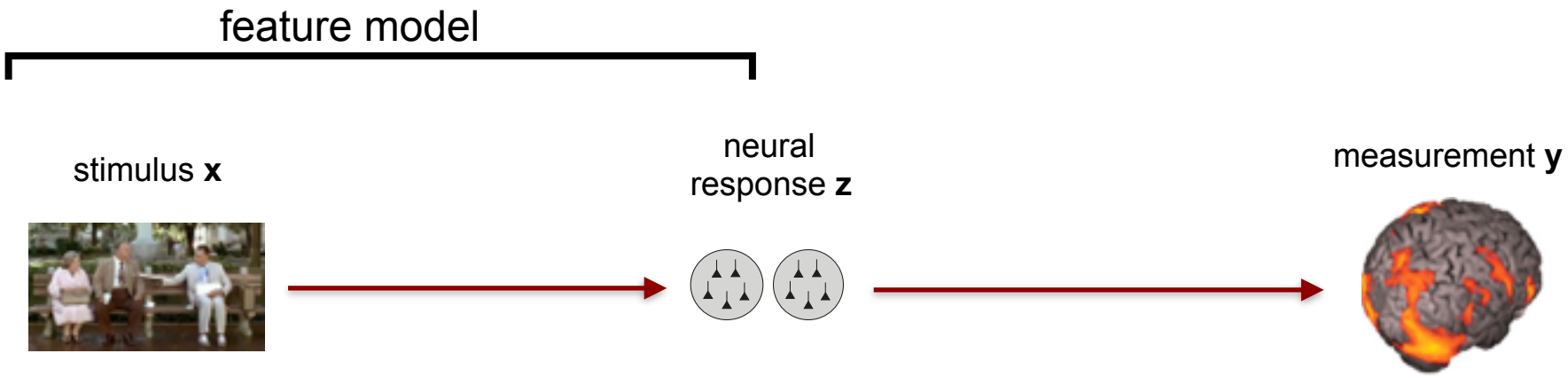
neural response **z**



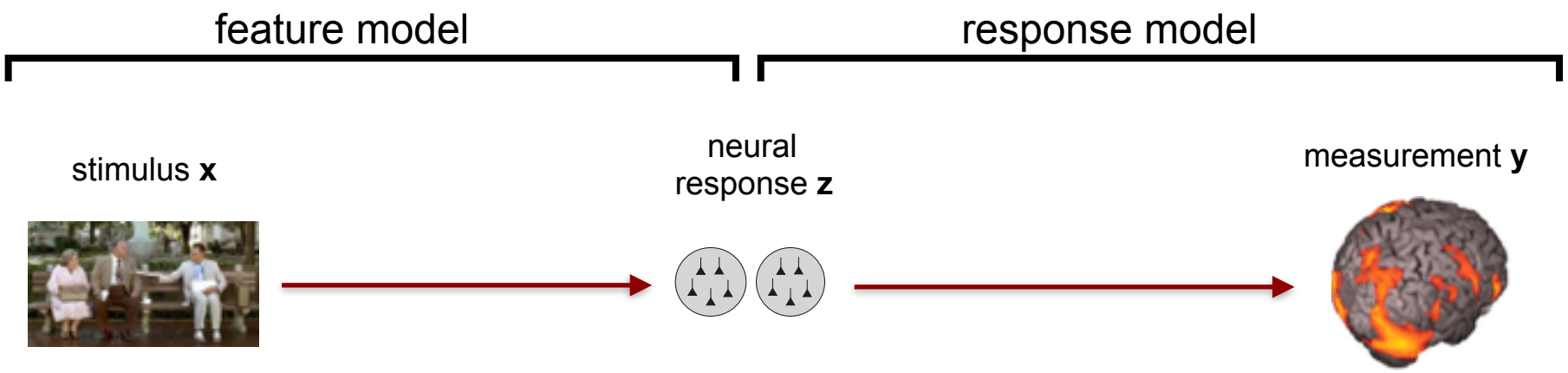
measurement **y**



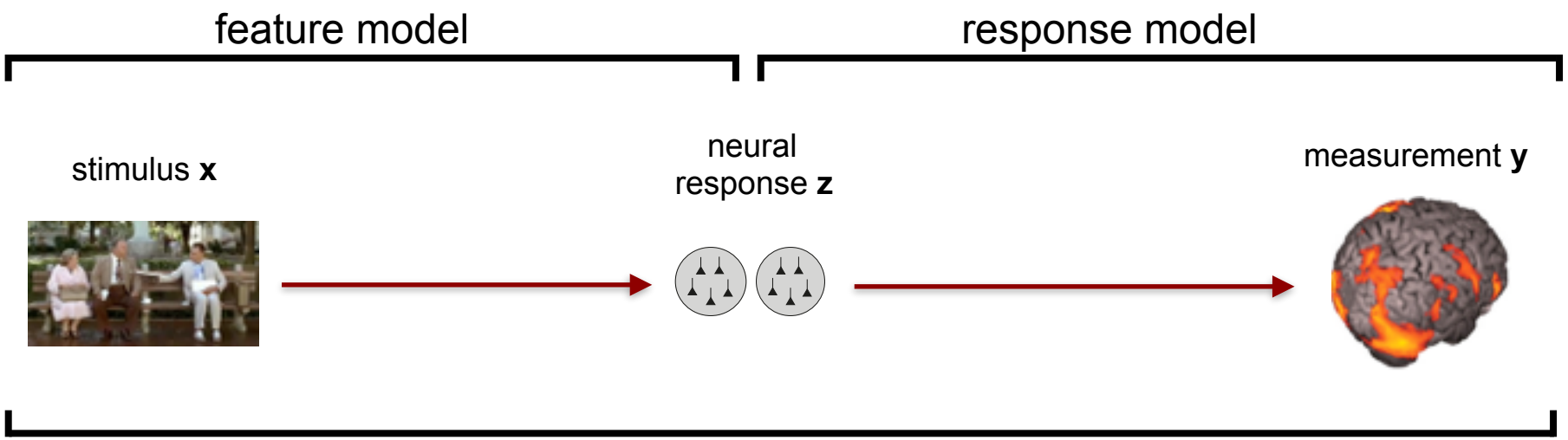
Modeling how the brain responds to the environment



Modeling how the brain responds to the environment



Modeling how the brain responds to the environment

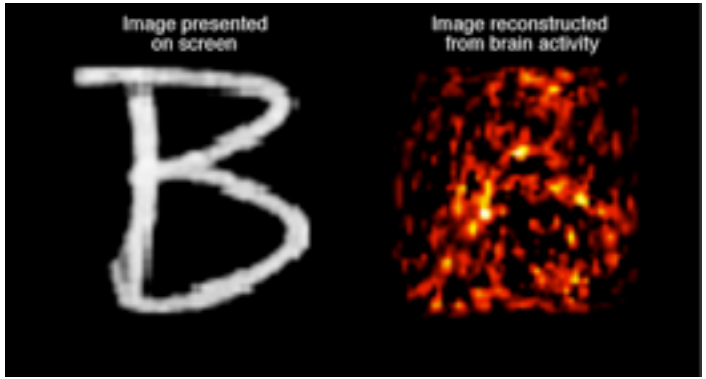


Encoding model

$$p(\mathbf{y} \mid \mathbf{x})$$

Previous work

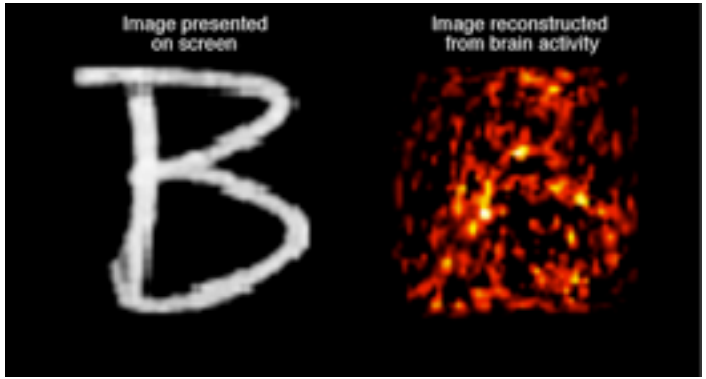
Linear mapping $\mathbf{y} \sim \mathcal{N}(\mathbf{B}^T\mathbf{x}, \mathbf{\Sigma})$



Schoenmakers et al. (2013). NeuroImage; Kay et al. (2008). Nature.

Previous work

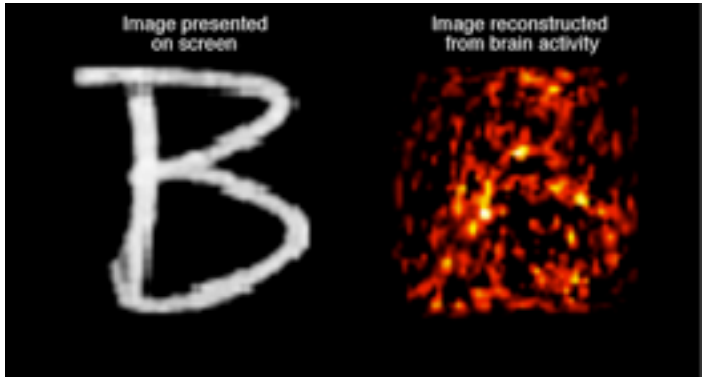
Linear mapping $\mathbf{y} \sim \mathcal{N}(\mathbf{B}^T\mathbf{x}, \mathbf{\Sigma})$



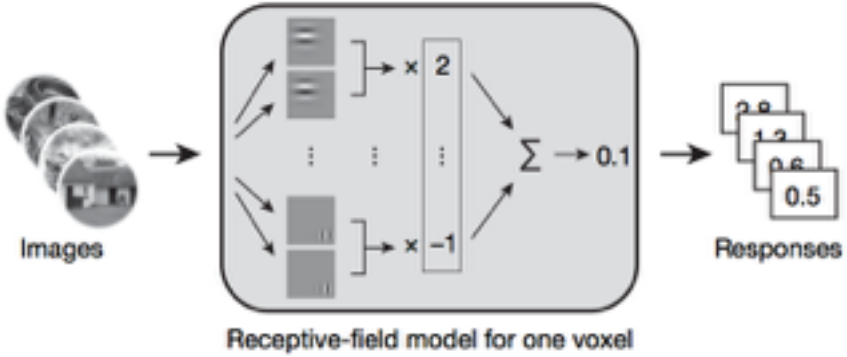
Schoenmakers et al. (2013). NeuroImage; Kay et al. (2008). Nature.

Previous work

Linear mapping $\mathbf{y} \sim \mathcal{N}(\mathbf{B}^T\mathbf{x}, \mathbf{\Sigma})$



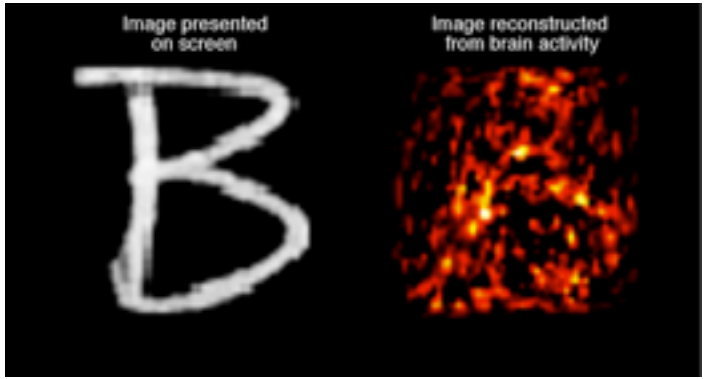
Gabor wavelet model



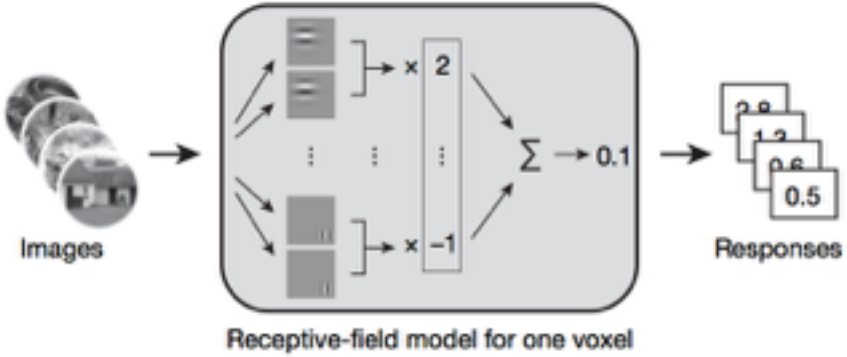
Schoenmakers et al. (2013). NeuroImage; Kay et al. (2008). Nature.

Previous work

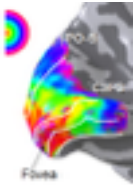
Linear mapping $\mathbf{y} \sim \mathcal{N}(\mathbf{B}^T\mathbf{x}, \mathbf{\Sigma})$



Gabor wavelet model



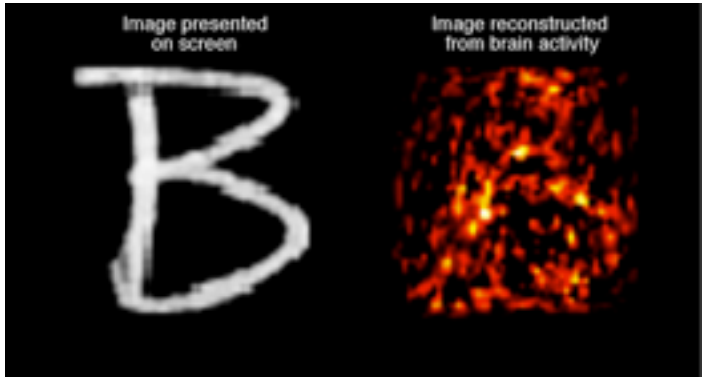
- Models of responses in early visual cortex



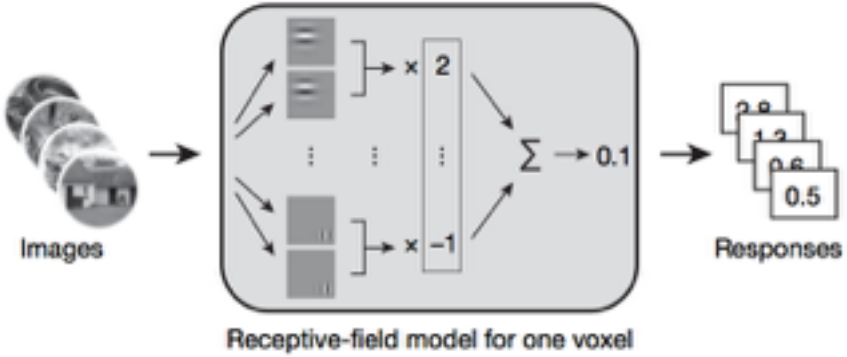
Schoenmakers et al. (2013). NeuroImage; Kay et al. (2008). Nature.

Previous work

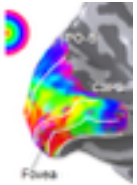
Linear mapping $\mathbf{y} \sim \mathcal{N}(\mathbf{B}^T\mathbf{x}, \mathbf{\Sigma})$



Gabor wavelet model



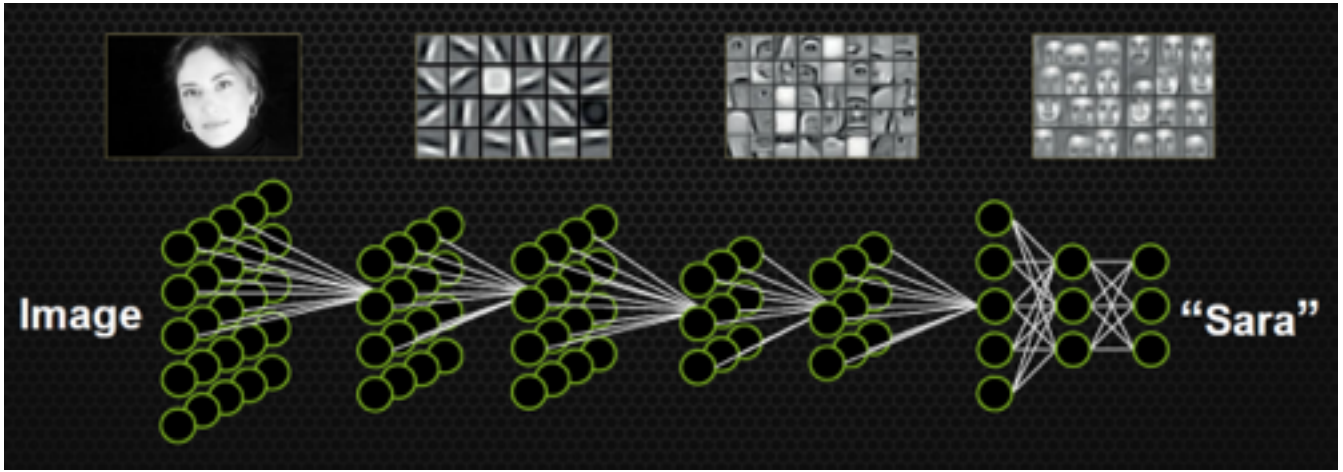
- Models of responses in early visual cortex



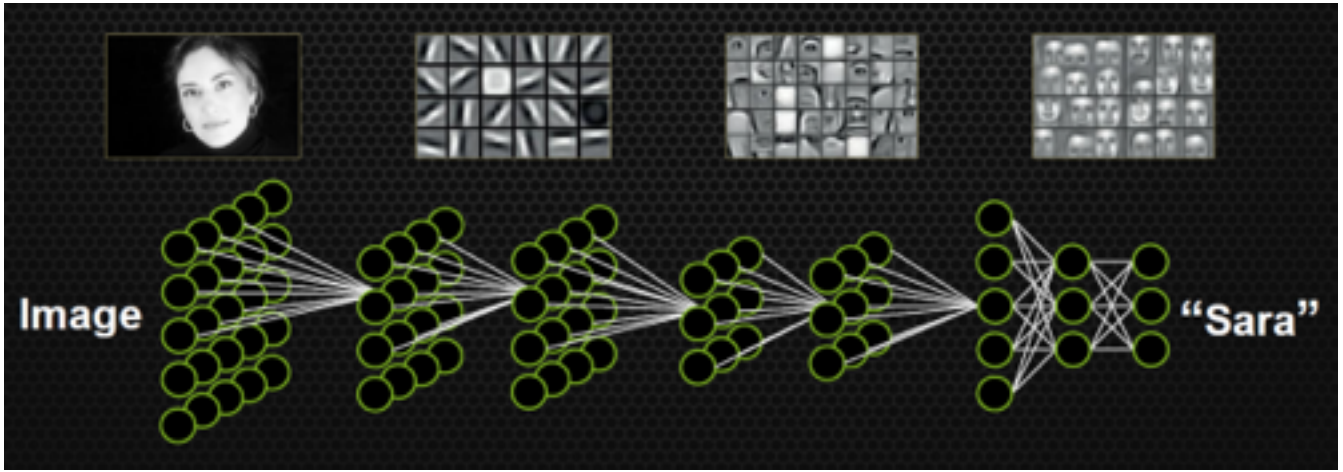
- How to model more complex representations in downstream areas?

Schoenmakers et al. (2013). NeuroImage; Kay et al. (2008). Nature.

Deep neural networks

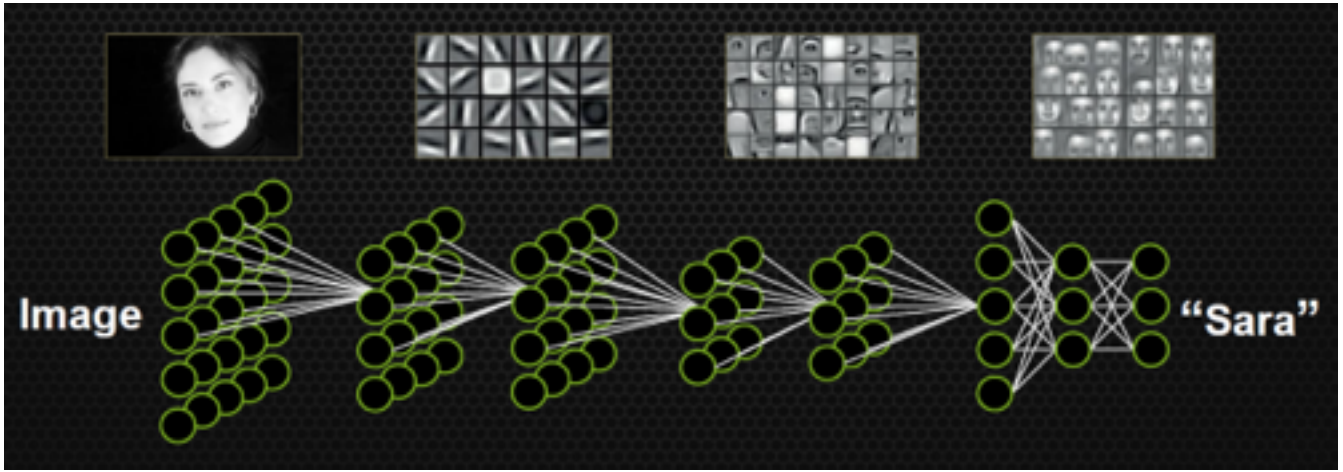


Deep neural networks



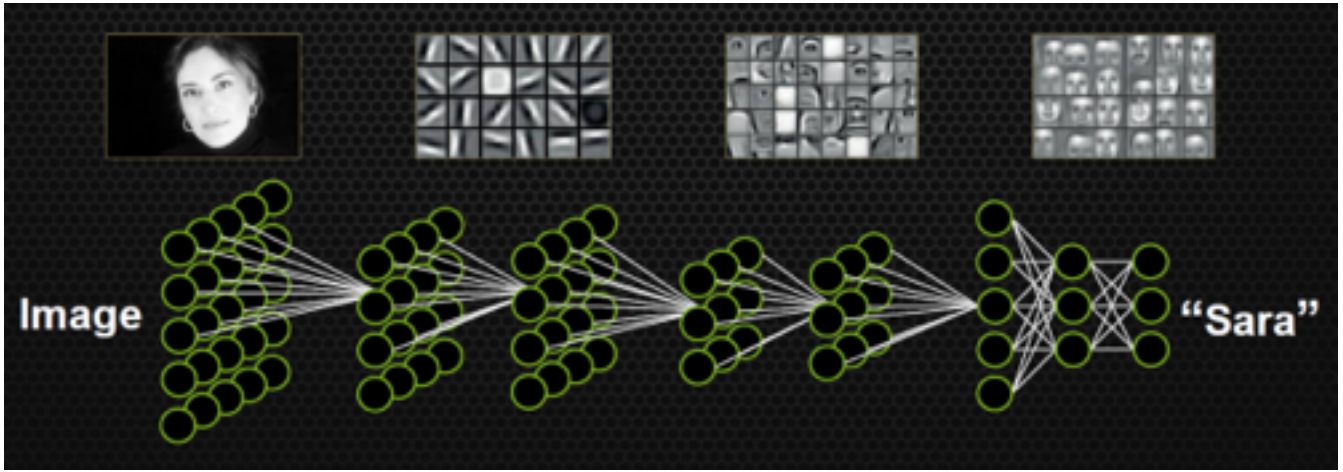
- Computational model consisting of artificial neurons that are organized in terms of layers

Deep neural networks



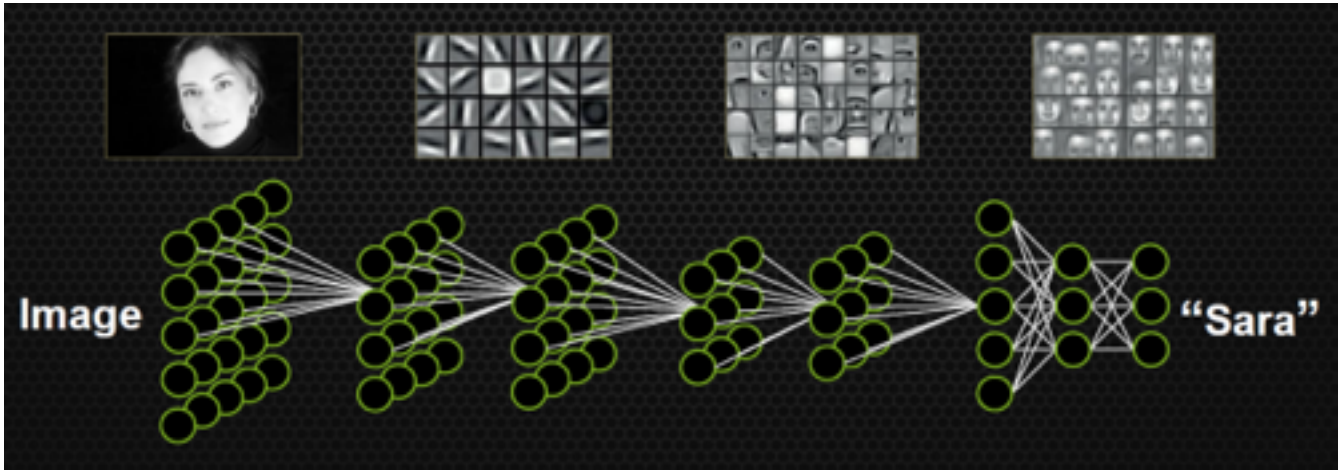
- Computational model consisting of artificial neurons that are organized in terms of layers
- Each layer incorporates particular filtering steps

Deep neural networks



- Computational model consisting of artificial neurons that are organized in terms of layers
- Each layer incorporates particular filtering steps
- Learns a mapping between an input (e.g. image) and an output (e.g. image categories) by adjusting 'synaptic' weights between artificial neurons

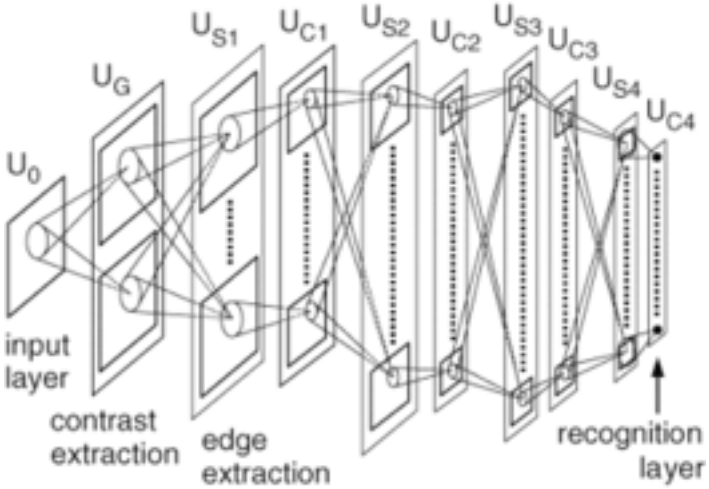
Deep neural networks



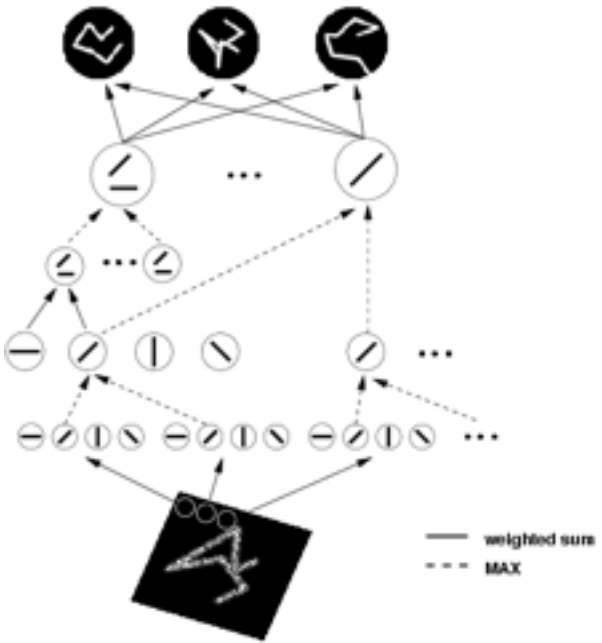
- Computational model consisting of artificial neurons that are organized in terms of layers
- Each layer incorporates particular filtering steps
- Learns a mapping between an input (e.g. image) and an output (e.g. image categories) by adjusting 'synaptic' weights between artificial neurons
- Representations in hidden layers capture increasingly complex characteristics of their input

Neuroscientific inspiration

- Inspired by classical models of stimulus transformations in sensory cortex



Fukushima, 1980, Biol. Cybern.



Riesenhuber & Poggio, 1999, Nat. Neurosci.

Resurgence

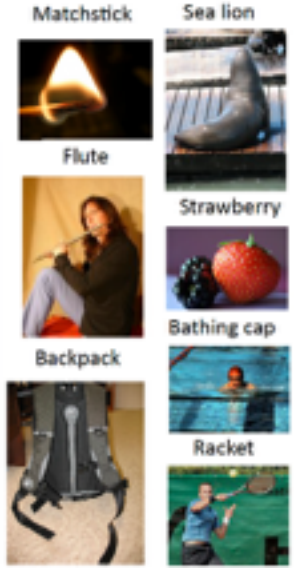
- Deep neural networks (DNNs) consisting of many layers can now be trained efficiently

Resurgence

- Deep neural networks (DNNs) consisting of many layers can now be trained efficiently
- Relies on combination of various tricks (convolution, pooling, dropout, etc.) as well as availability of massive labeled datasets and ample computational resources (GPUs)

The ImageNet dataset [Fei-Fei et al. 2012]
- 1.2 million training samples
- 1000 categories

Fast Graphical Processing Units (GPU)
- Capable of over 1 trillion operations/second

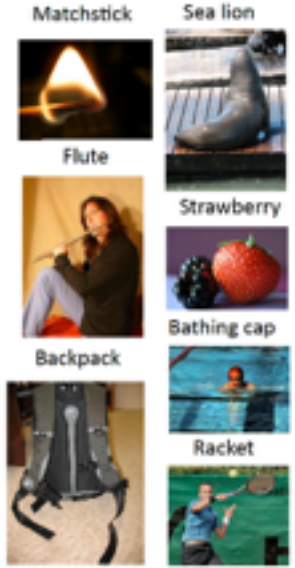


Resurgence

- Deep neural networks (DNNs) consisting of many layers can now be trained efficiently
- Relies on combination of various tricks (convolution, pooling, dropout, etc.) as well as availability of massive labeled datasets and ample computational resources (GPUs)

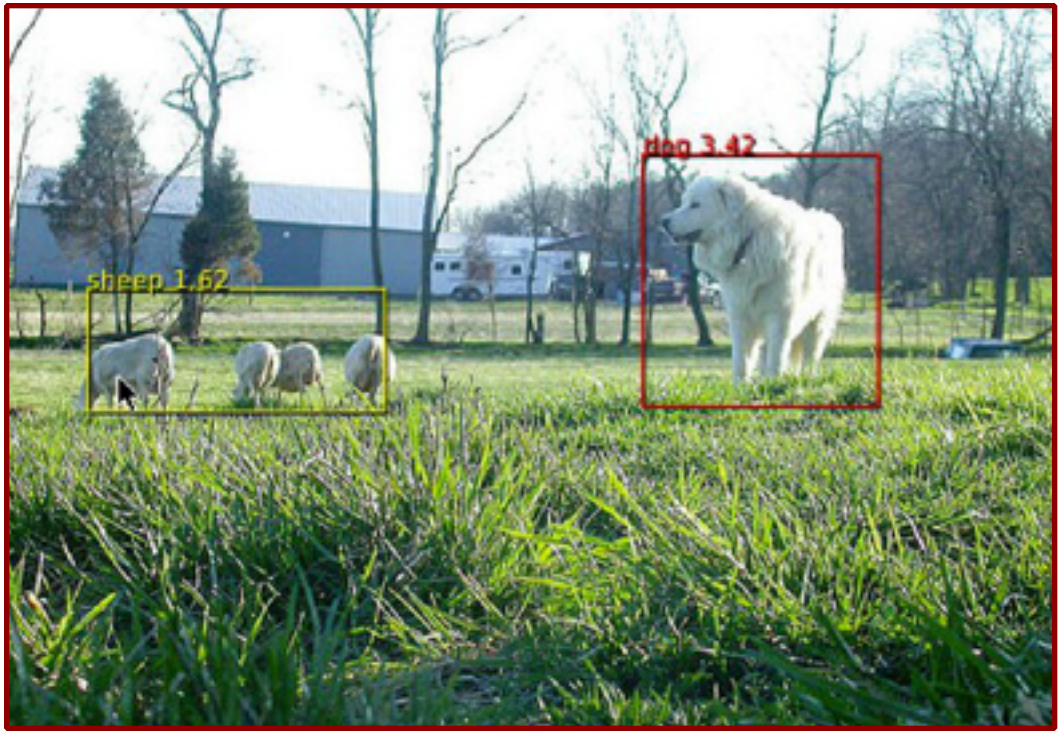
The ImageNet dataset [Fei-Fei et al. 2012]
- 1.2 million training samples
- 1000 categories

Fast Graphical Processing Units (GPU)
- Capable of over 1 trillion operations/second

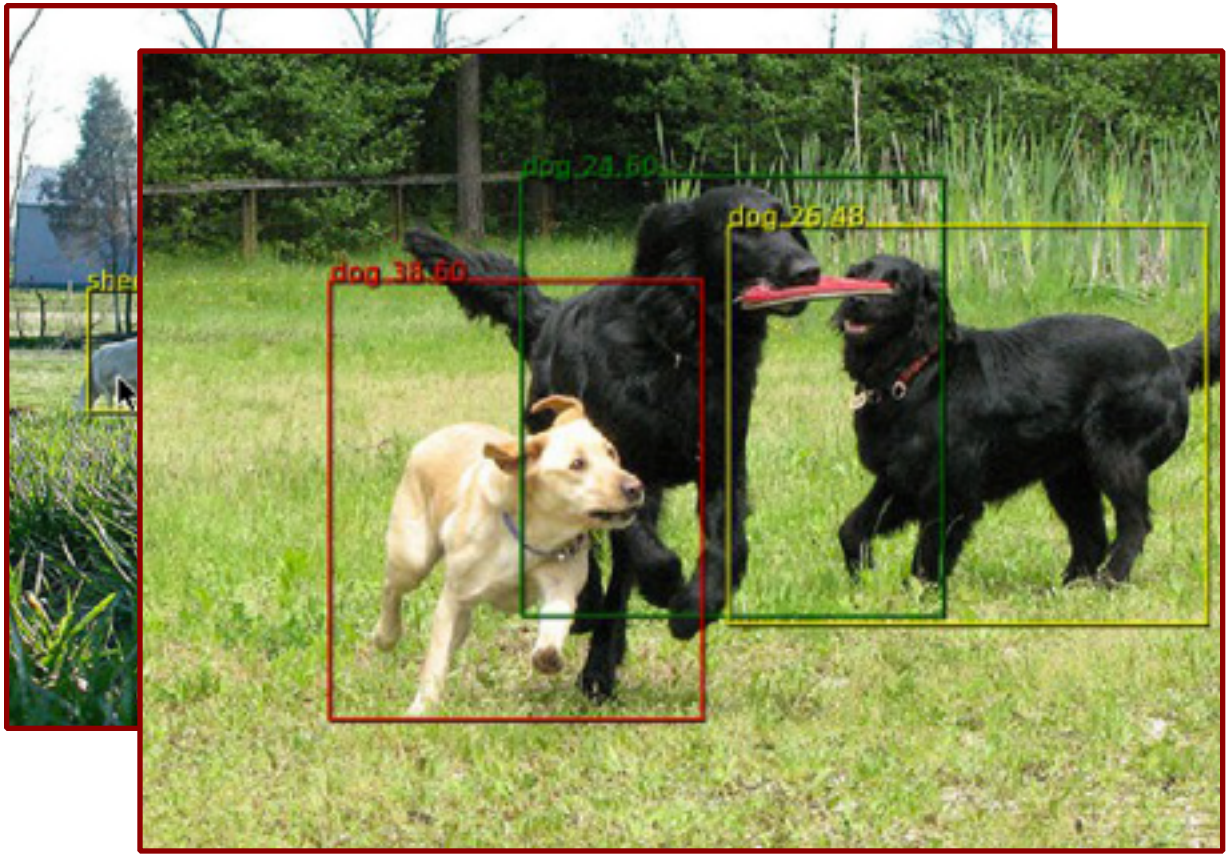


- Revolutionized computer vision, natural language processing, robotics, artificial intelligence

Example: Object detection

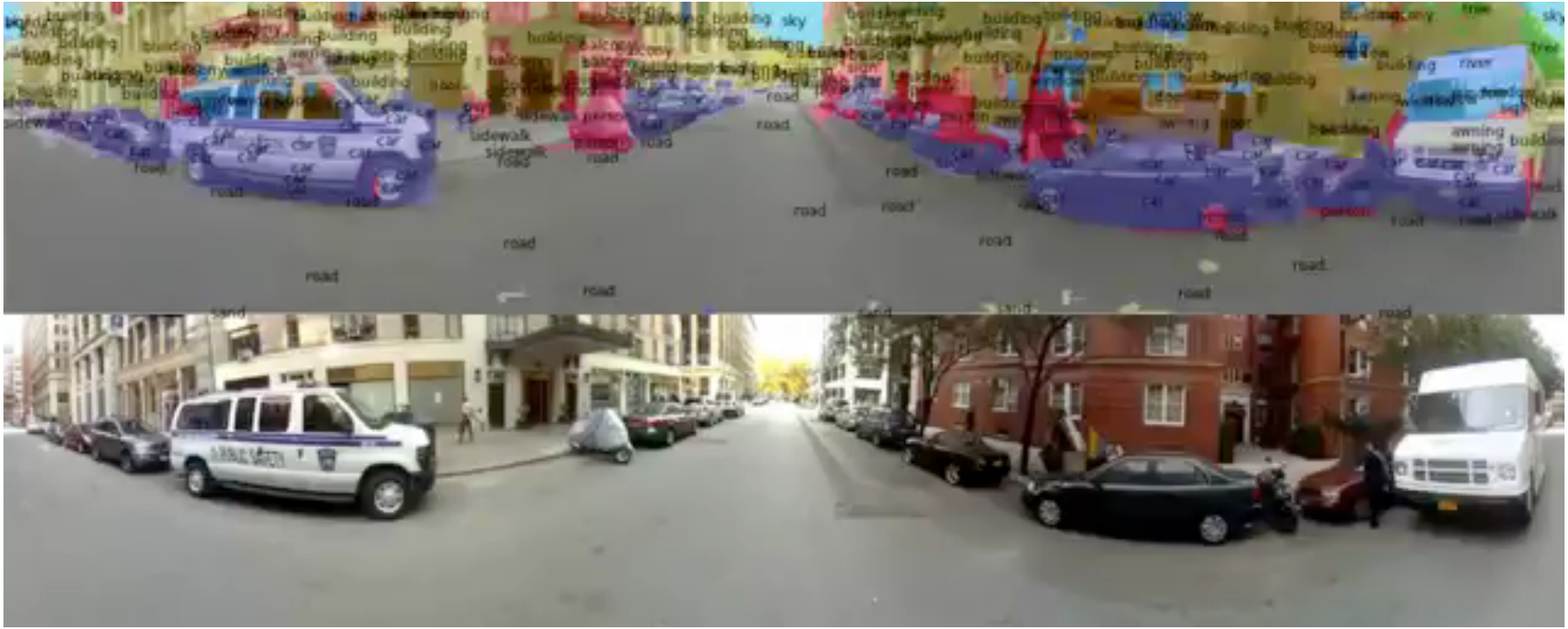


Example: Object detection



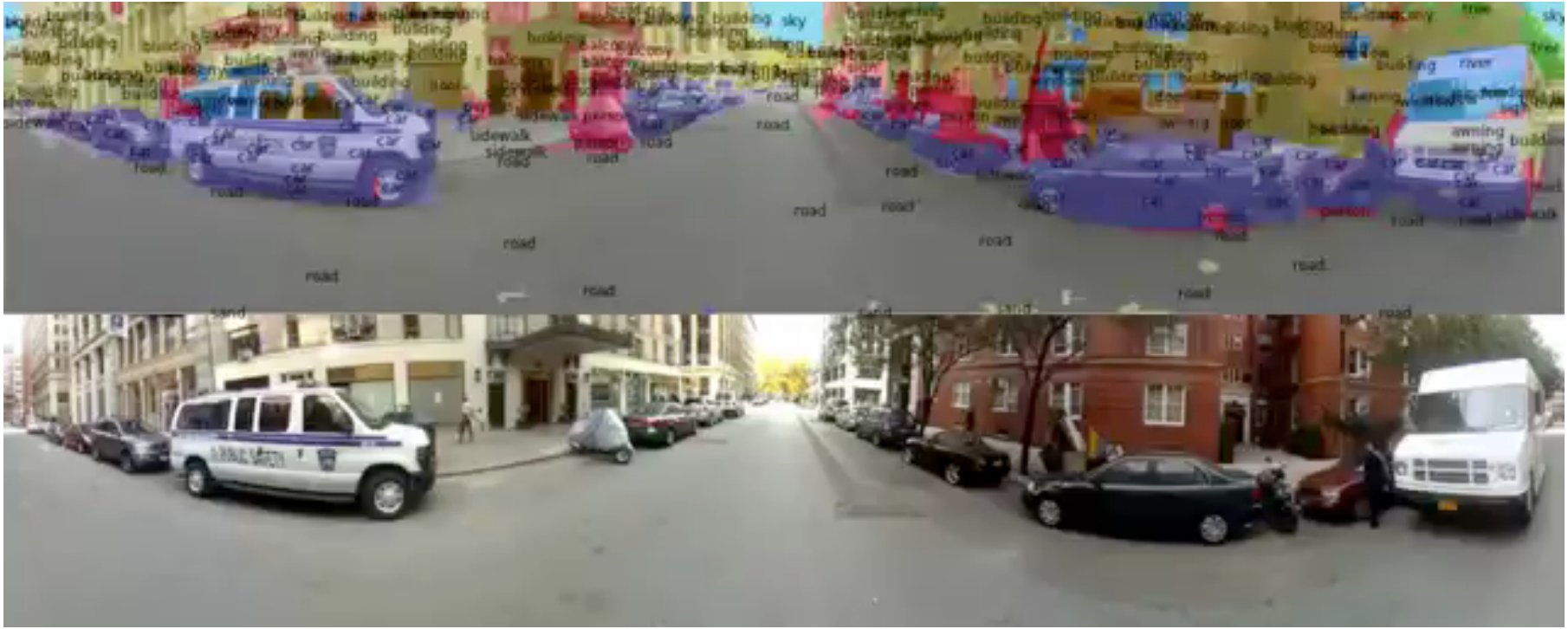
Example: Object detection

Example: Semantic segmentation



Farabet et al. (2012). ICML.

Example: Semantic segmentation



Farabet et al. (2012). ICML.

Example: Neural style transfer

Gatys et al. (2015). arxiv.

Example: Neural style transfer

Gatys et al. (2015). arxiv.

Example: Neural style transfer

Gatys et al. (2015). arxiv.

Example: Deep dreaming

http://deepdreamgenerator.com

Example: Deep dreaming

http://deepdreamgenerator.com

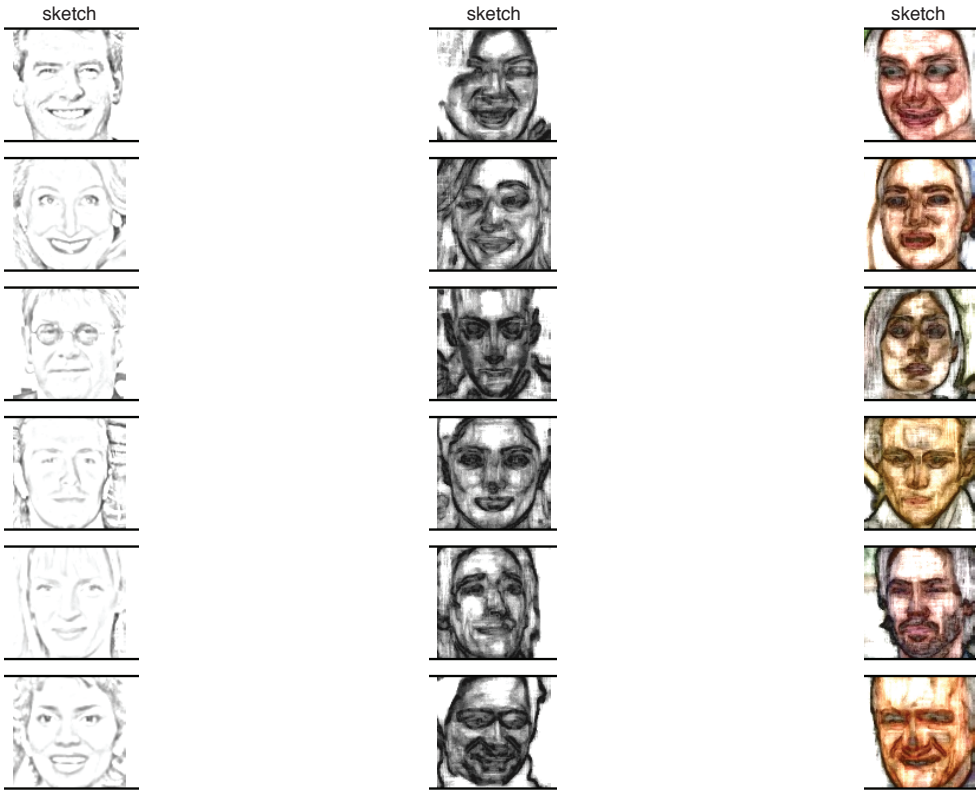
Example: DQN playing Breakout

Mnih et al. (2015). Nature

Example: DQN playing Breakout

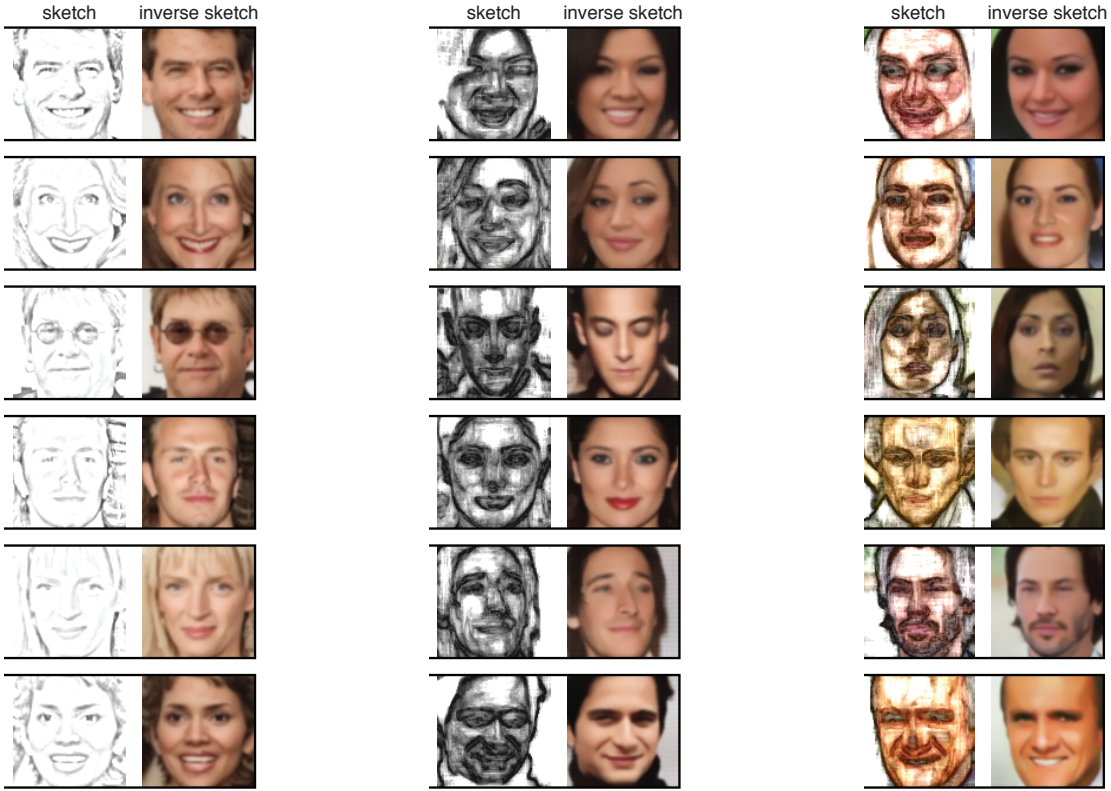
Mnih et al. (2015). Nature

Example: Colorization and filling-in



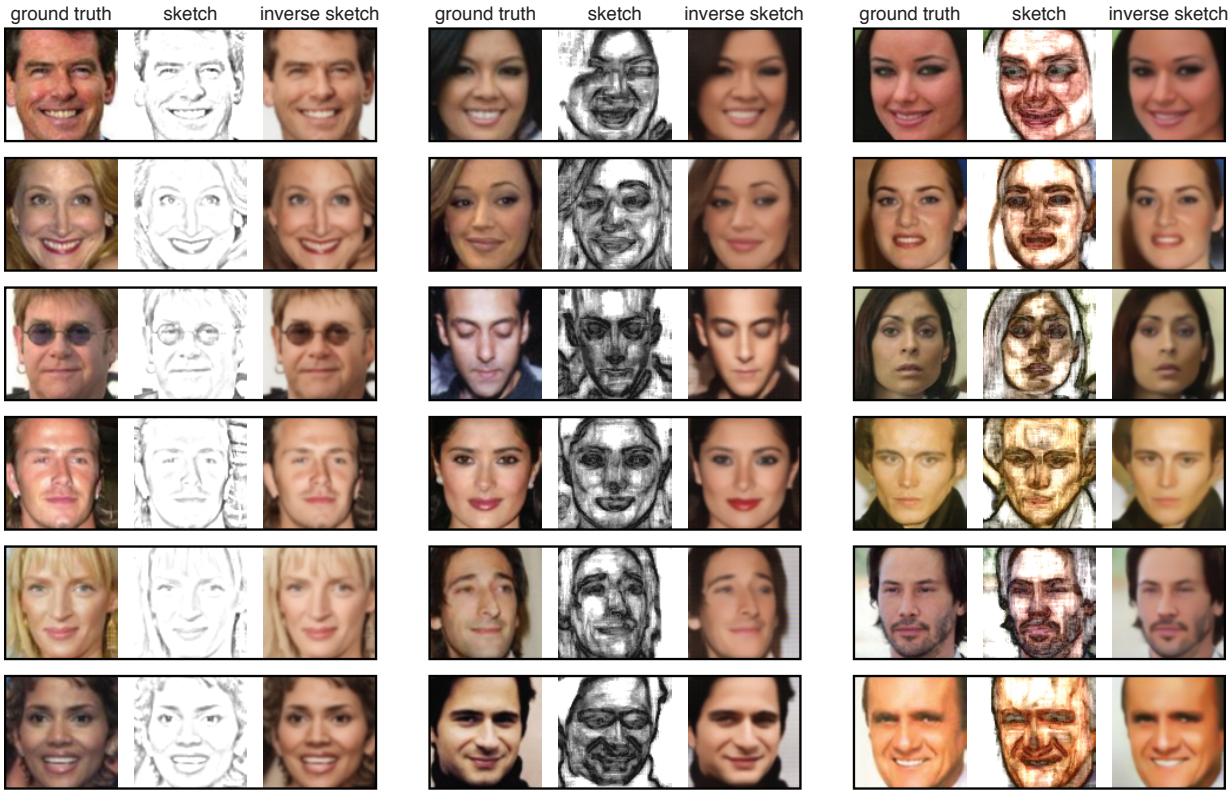
Güçlütürk, Güçlü, van Lier, van Gerven. (2016). arxiv.

Example: Colorization and filling-in



Güçlütürk, Güçlü, van Lier, van Gerven. (2016). arxiv.

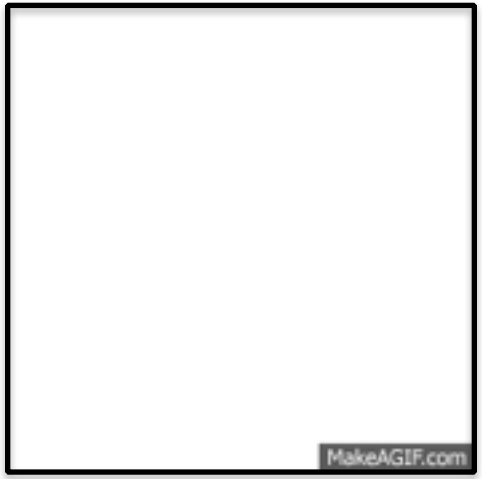
Example: Colorization and filling-in



Güçlütürk, Güçlü, van Lier, van Gerven. (2016). arxiv.

Examples: Colorization and filling-in

sketch

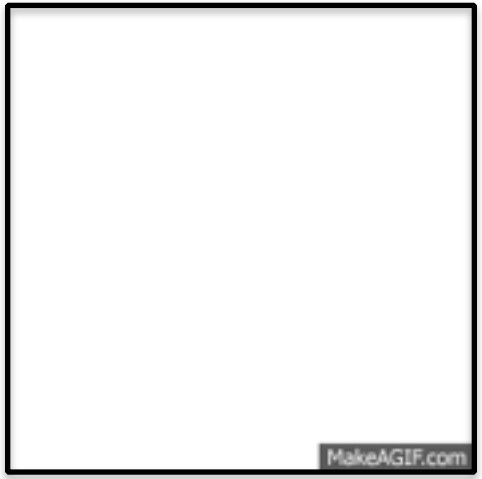


colour reconstruction

grayscale reconstruction

Examples: Colorization and filling-in

sketch



colour reconstruction

grayscale reconstruction

Example: Generating music

Güçlütürk, Güçlü, van Lier, van Gerven. (2016). arxiv.

Example: Generating music

Güçlütürk, Güçlü, van Lier, van Gerven. (2016). arxiv.

Example: Generating music

Güçlütürk, Güçlü, van Lier, van Gerven. (2016). arxiv.

Example: Generating music

Güçlütürk, Güçlü, van Lier, van Gerven. (2016). arxiv.

Probing neural representations with DNNs

Probing neural representations with DNNs

- A task-optimized DNN model learns a hierarchy of nonlinear transformations in a supervised manner with the objective of solving a particular task.

Probing neural representations with DNNs

- A task-optimized DNN model learns a hierarchy of nonlinear transformations in a supervised manner with the objective of solving a particular task.
- fMRI measures local changes in blood-oxygen-level dependent hemodynamic responses to sensory stimulation.

Probing neural representations with DNNs

- A task-optimized DNN model learns a hierarchy of nonlinear transformations in a supervised manner with the objective of solving a particular task.
- fMRI measures local changes in blood-oxygen-level dependent hemodynamic responses to sensory stimulation.
- Any subset of the DNN representations that emerge from this hierarchy of nonlinear transformations can be used to probe neural representations by comparing DNN and fMRI responses to the same sensory stimuli.

Probing neural representations with DNNs

- A task-optimized DNN model learns a hierarchy of nonlinear transformations in a supervised manner with the objective of solving a particular task.
- fMRI measures local changes in blood-oxygen-level dependent hemodynamic responses to sensory stimulation.
- Any subset of the DNN representations that emerge from this hierarchy of nonlinear transformations can be used to probe neural representations by comparing DNN and fMRI responses to the same sensory stimuli.
- Considering that the sensory systems are biological neural networks that routinely perform the same tasks as their artificial counterparts, it is not inconceivable that DNN representations are suitable for probing neural representations.

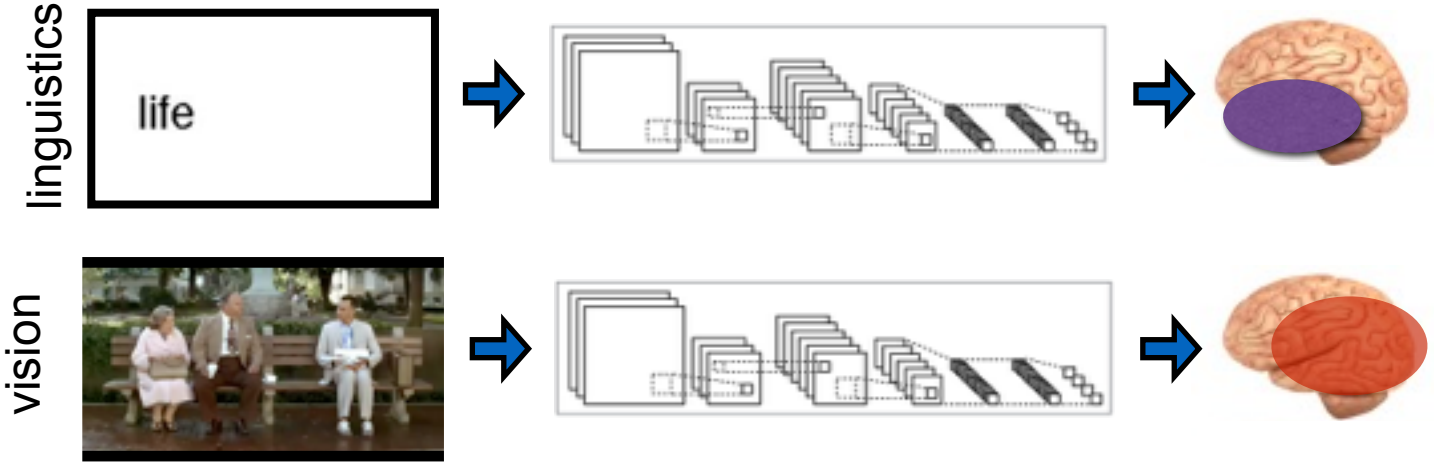
Probing neural representations with DNNs

Güçlü, U., van Gerven, M., 2015. Neuroimage. In Press; Güçlü, U., van Gerven, M.A.J., 2015. Arxiv.

Probing neural representations with DNNs

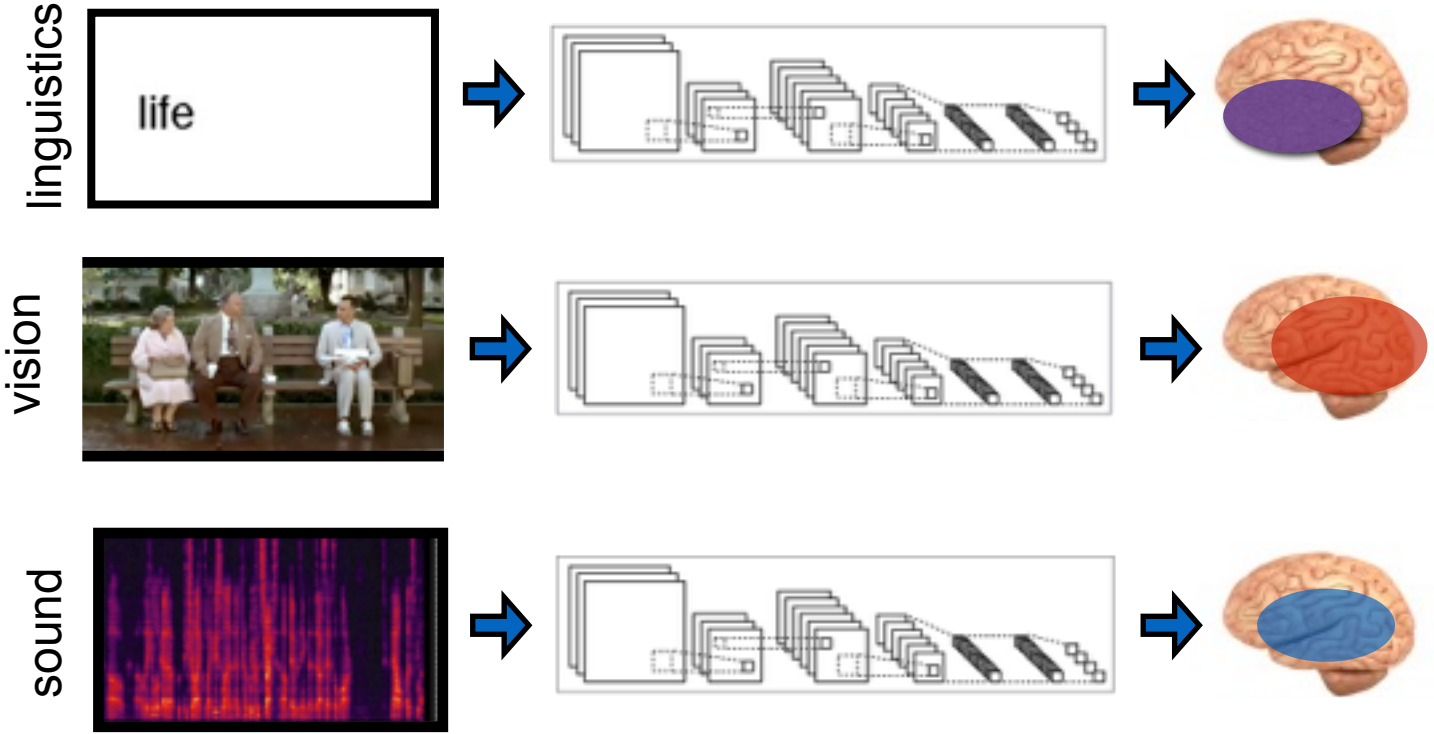
Güçlü, U., van Gerven, M., 2015. Neuroimage. In Press; Güçlü, U., van Gerven, M.A.J., 2015. Arxiv.

Probing neural representations with DNNs



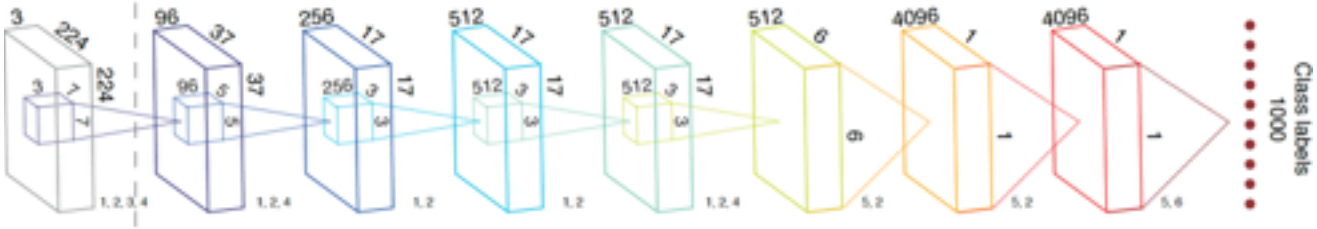
Güçlü, U., van Gerven, M., 2015. Neuroimage. In Press; Güçlü, U., van Gerven, M.A.J., 2015. Arxiv.

Probing neural representations with DNNs



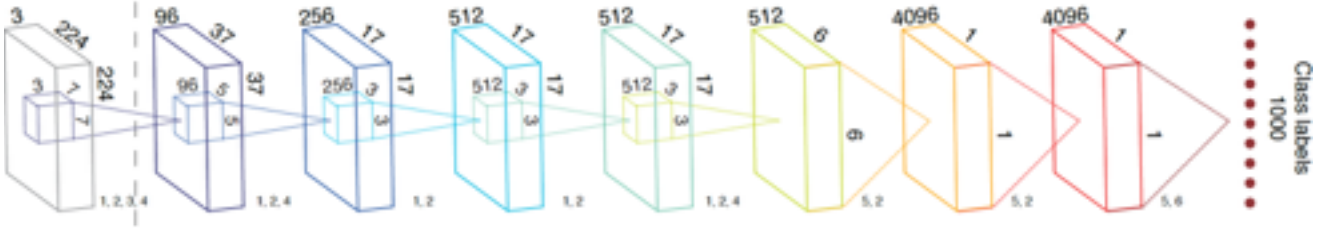
Güçlü, U., van Gerven, M., 2015. Neuroimage. In Press; Güçlü, U., van Gerven, M.A.J., 2015. Arxiv.

DNN-based encoding model



[1] convolution, [2] rectification,
[3] local response normalization,
[4] max pooling, [5] inner product, [6] softmax

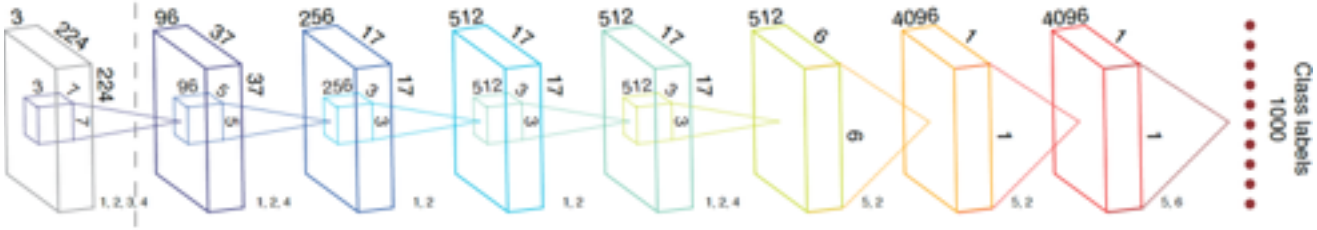
DNN-based encoding model



[1] convolution, [2] rectification,
[3] local response normalization,
[4] max pooling, [5] inner product, [6] softmax

- Model of the ventral stream

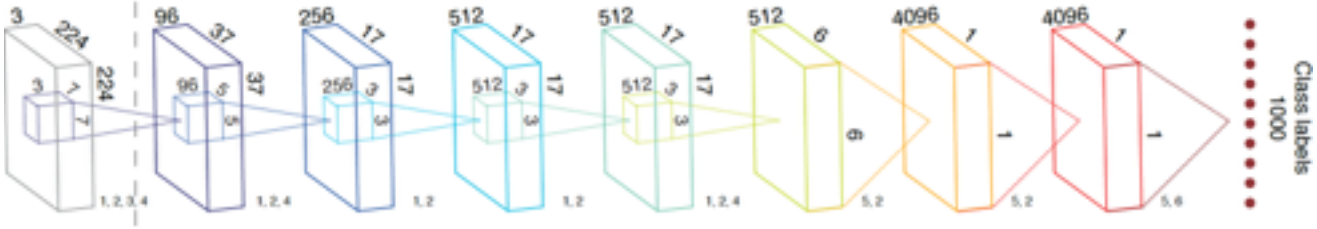
DNN-based encoding model



[1] convolution, [2] rectification,
[3] local response normalization,
[4] max pooling, [5] inner product, [6] softmax

- Model of the ventral stream
- Eight layers of filters tiled in such a way that they respond to overlapping regions in the visual field

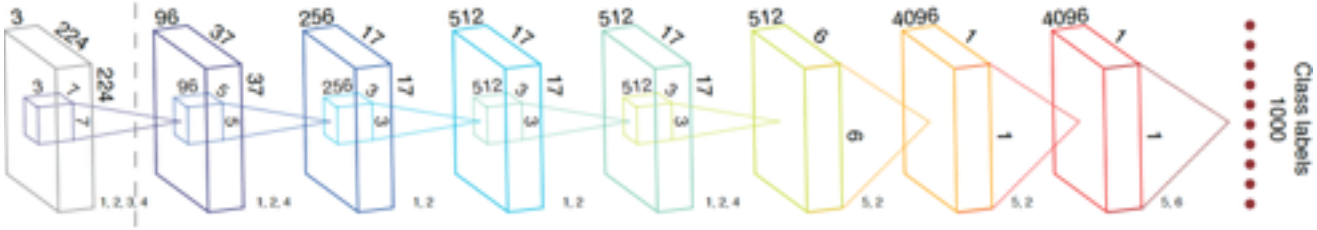
DNN-based encoding model



[1] convolution, [2] rectification,
[3] local response normalization,
[4] max pooling, [5] inner product, [6] softmax

- Model of the ventral stream
- Eight layers of filters tiled in such a way that they respond to overlapping regions in the visual field
- Trained on 1M labelled natural images by back propagation

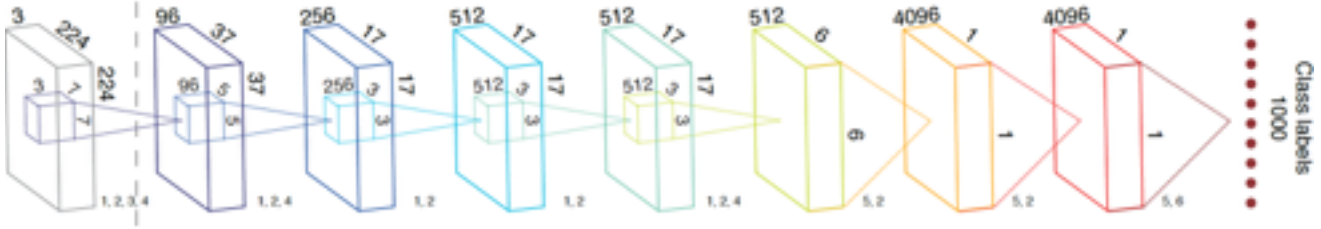
DNN-based encoding model



[1] convolution, [2] rectification,
[3] local response normalization,
[4] max pooling, [5] inner product, [6] softmax

- Model of the ventral stream
- Eight layers of filters tiled in such a way that they respond to overlapping regions in the visual field
- Trained on 1M labelled natural images by back propagation
- The resulting DNN has 60 million parameters and 650K neurons (!)

DNN-based encoding model



[1] convolution, [2] rectification,
[3] local response normalization,
[4] max pooling, [5] inner product, [6] softmax

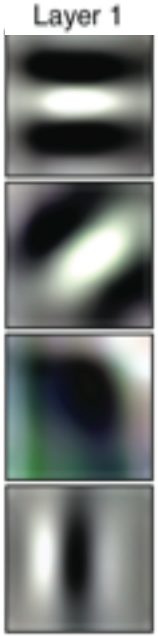
- Model of the ventral stream
- Eight layers of filters tiled in such a way that they respond to overlapping regions in the visual field
- Trained on 1M labelled natural images by back propagation
- The resulting DNN has 60 million parameters and 650K neurons (!)
- DNN features mapped to neural responses using kernel ridge regression

Deep neural network representations

- What stimulus properties are represented by artificial neurons?

Deep neural network representations

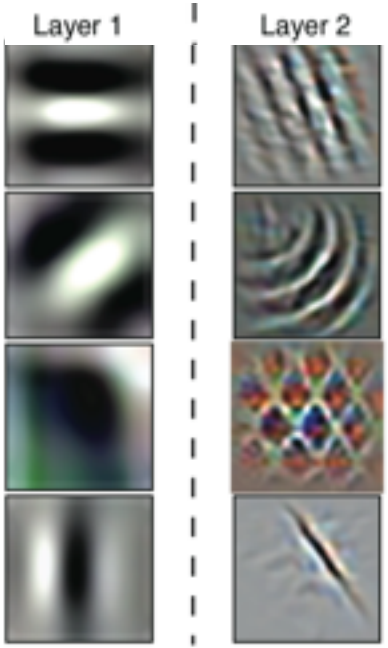
- What stimulus properties are represented by artificial neurons?



Güçlü, U., & van Gerven, M. A. J. (2015). J. Neuroscience.

Deep neural network representations

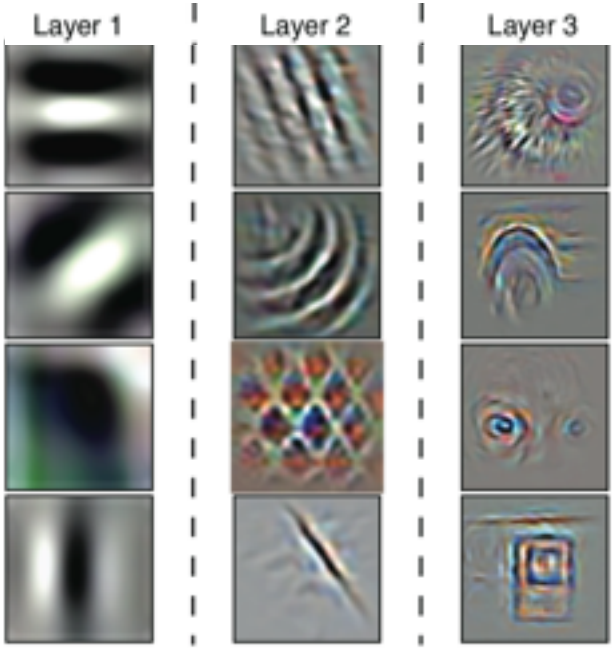
- What stimulus properties are represented by artificial neurons?



Güçlü, U., & van Gerven, M. A. J. (2015). J. Neuroscience.

Deep neural network representations

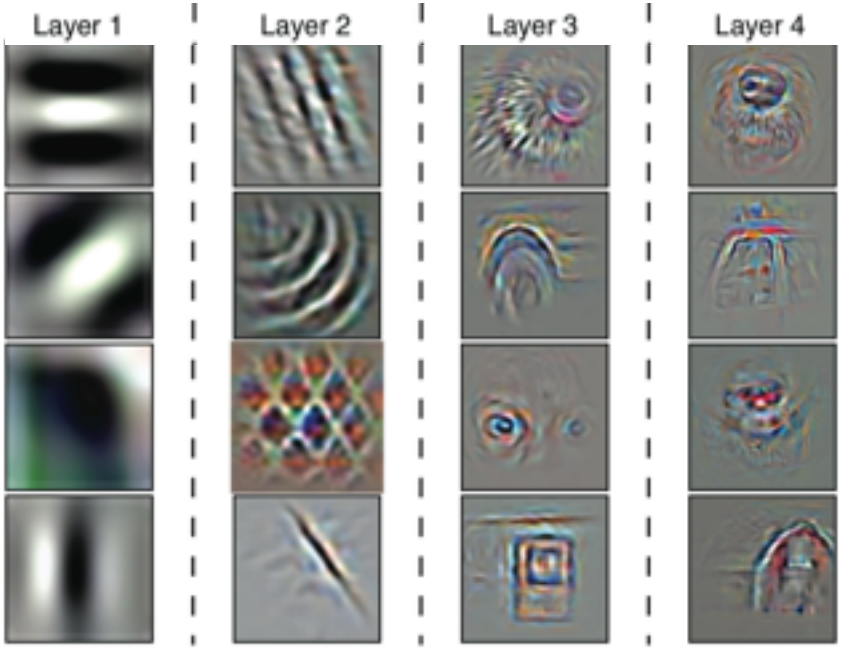
- What stimulus properties are represented by artificial neurons?



Güçlü, U., & van Gerven, M. A. J. (2015). J. Neuroscience.

Deep neural network representations

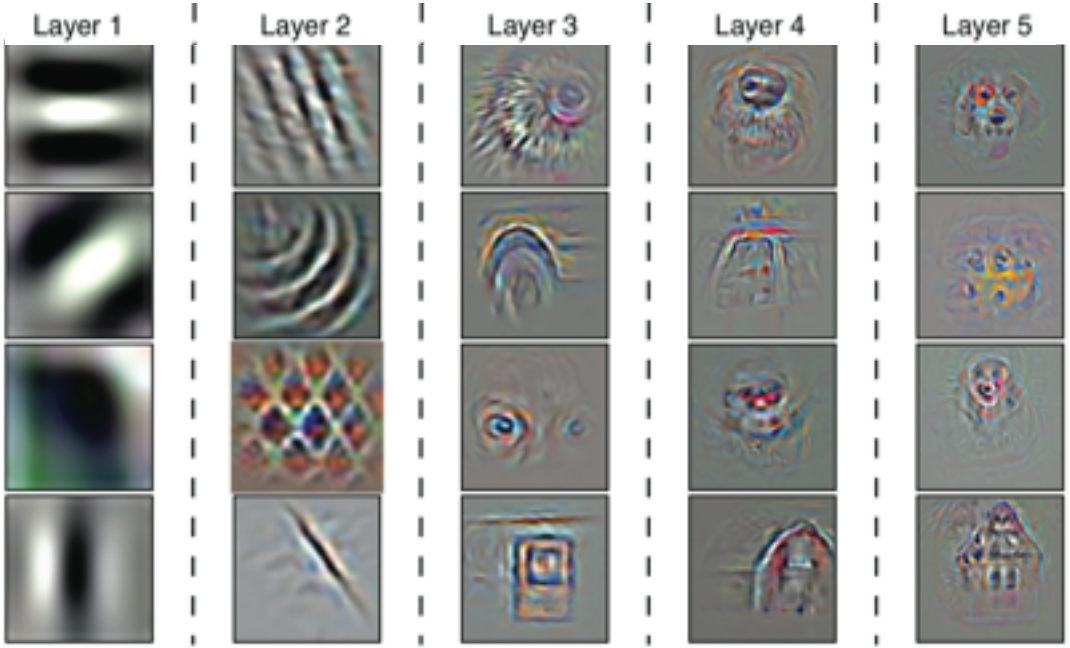
- What stimulus properties are represented by artificial neurons?



Güçlü, U., & van Gerven, M. A. J. (2015). J. Neuroscience.

Deep neural network representations

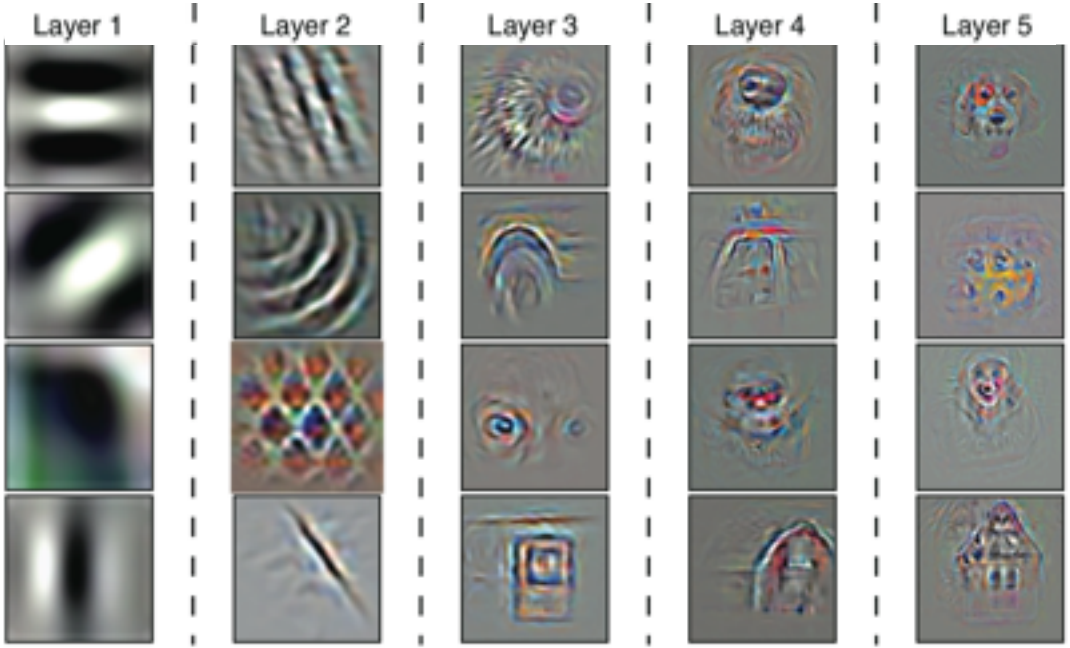
- What stimulus properties are represented by artificial neurons?



Güçlü, U., & van Gerven, M. A. J. (2015). J. Neuroscience.

Deep neural network representations

- What stimulus properties are represented by artificial neurons?

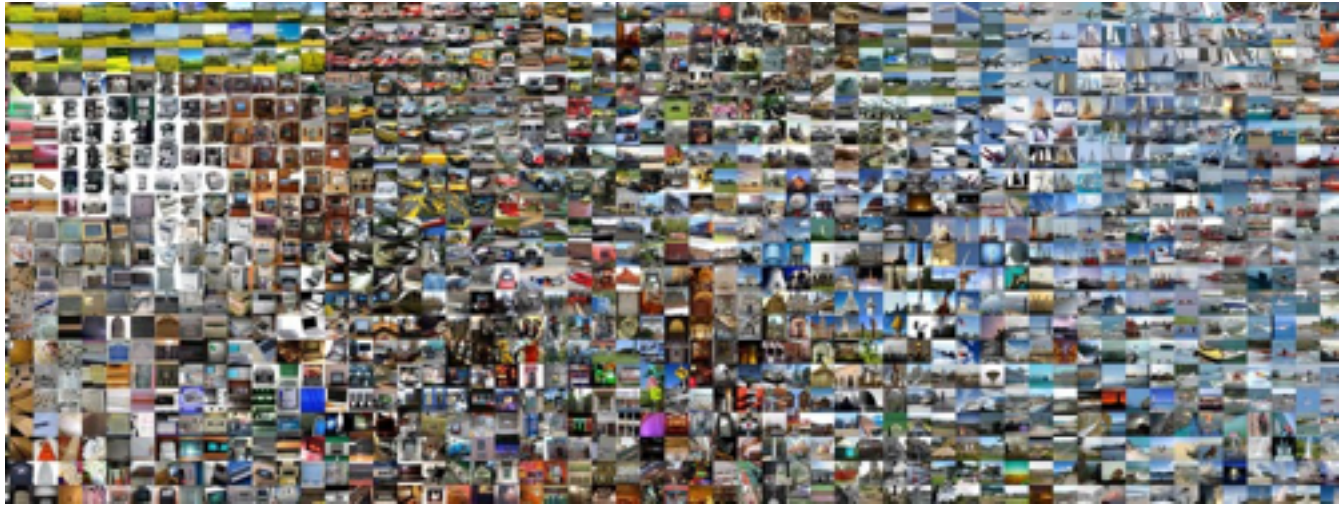


- Deep layers code for increasingly complex stimulus properties

Güçlü, U., & van Gerven, M. A. J. (2015). J. Neuroscience.

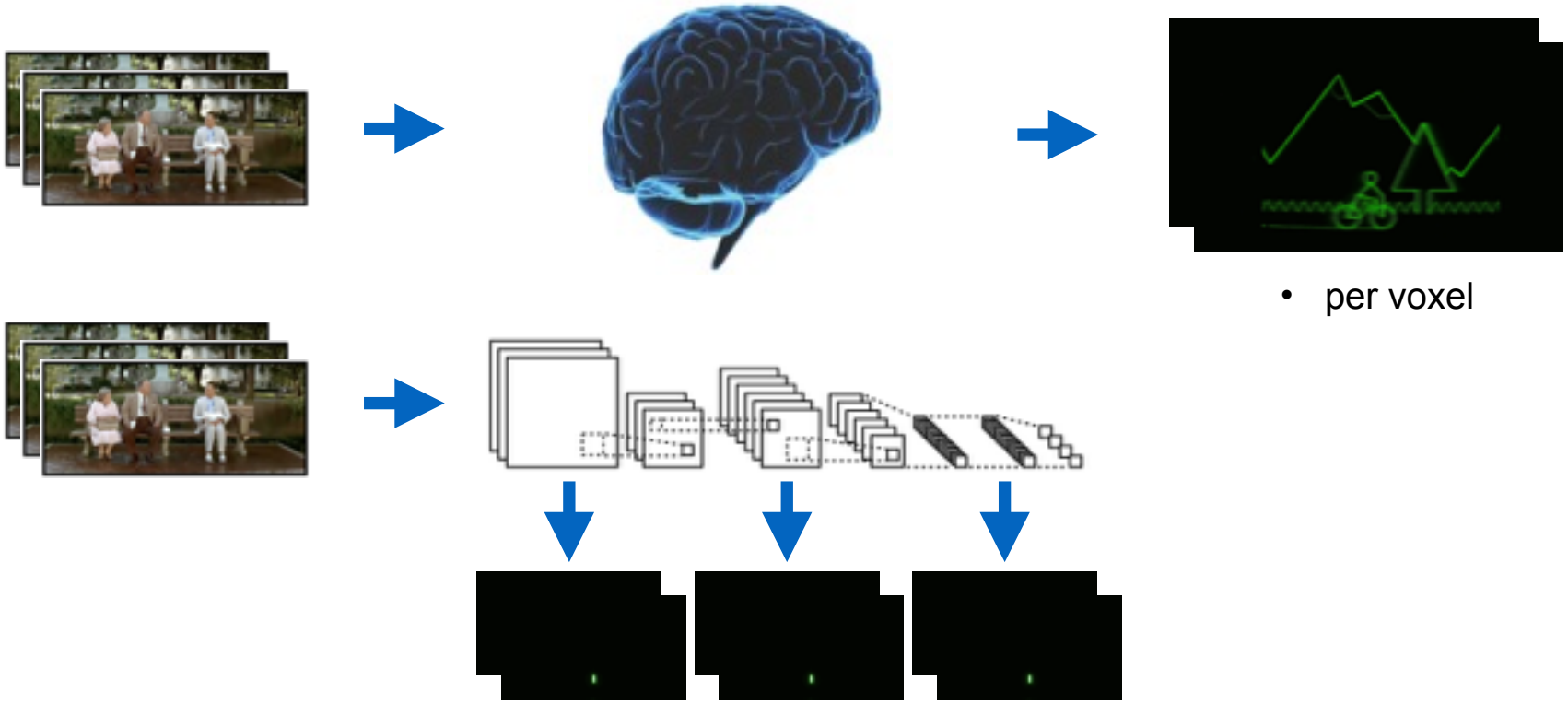
Recipe

- 1870 natural images



Güçlü, U., & van Gerven, M. A. J. (2015). J. Neuroscience.

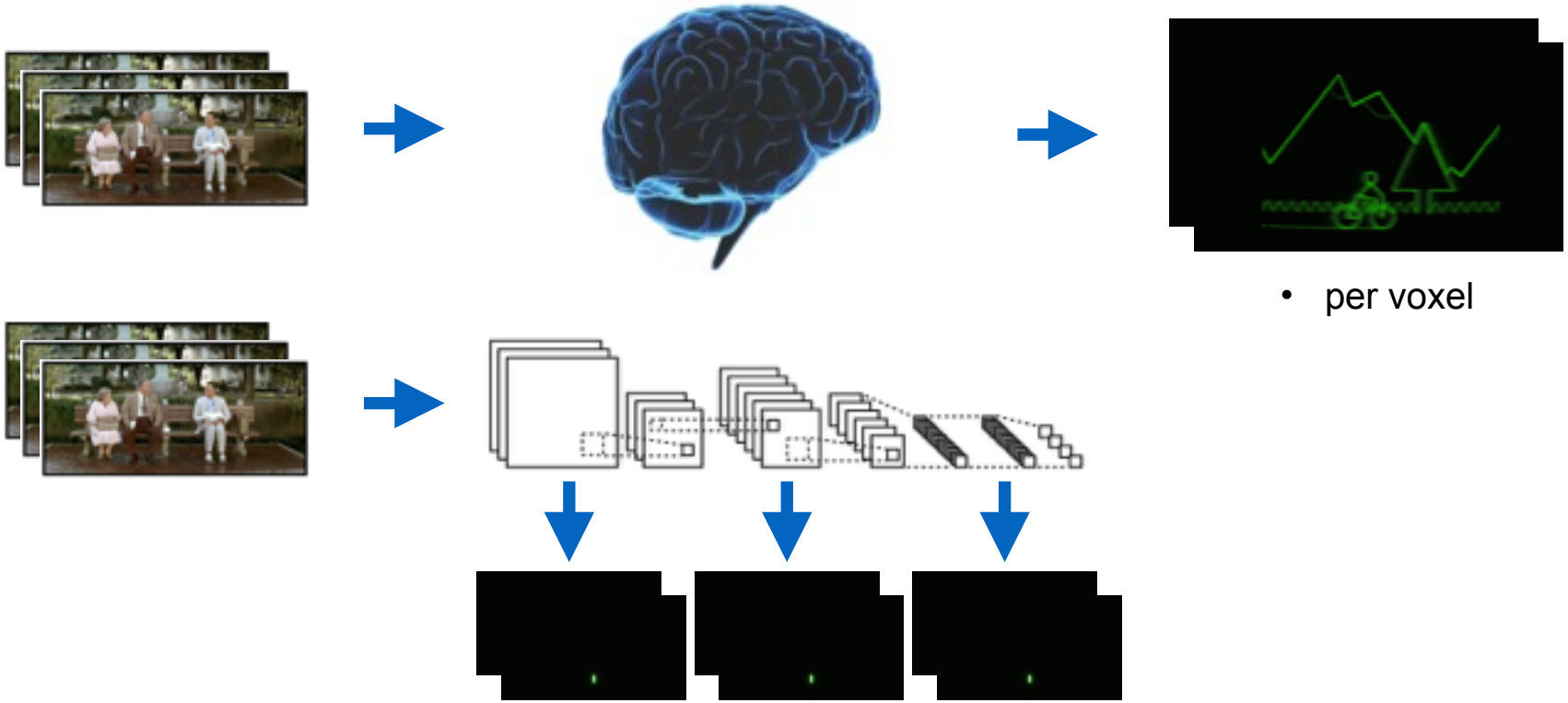
Recipe



- per voxel

- per layer

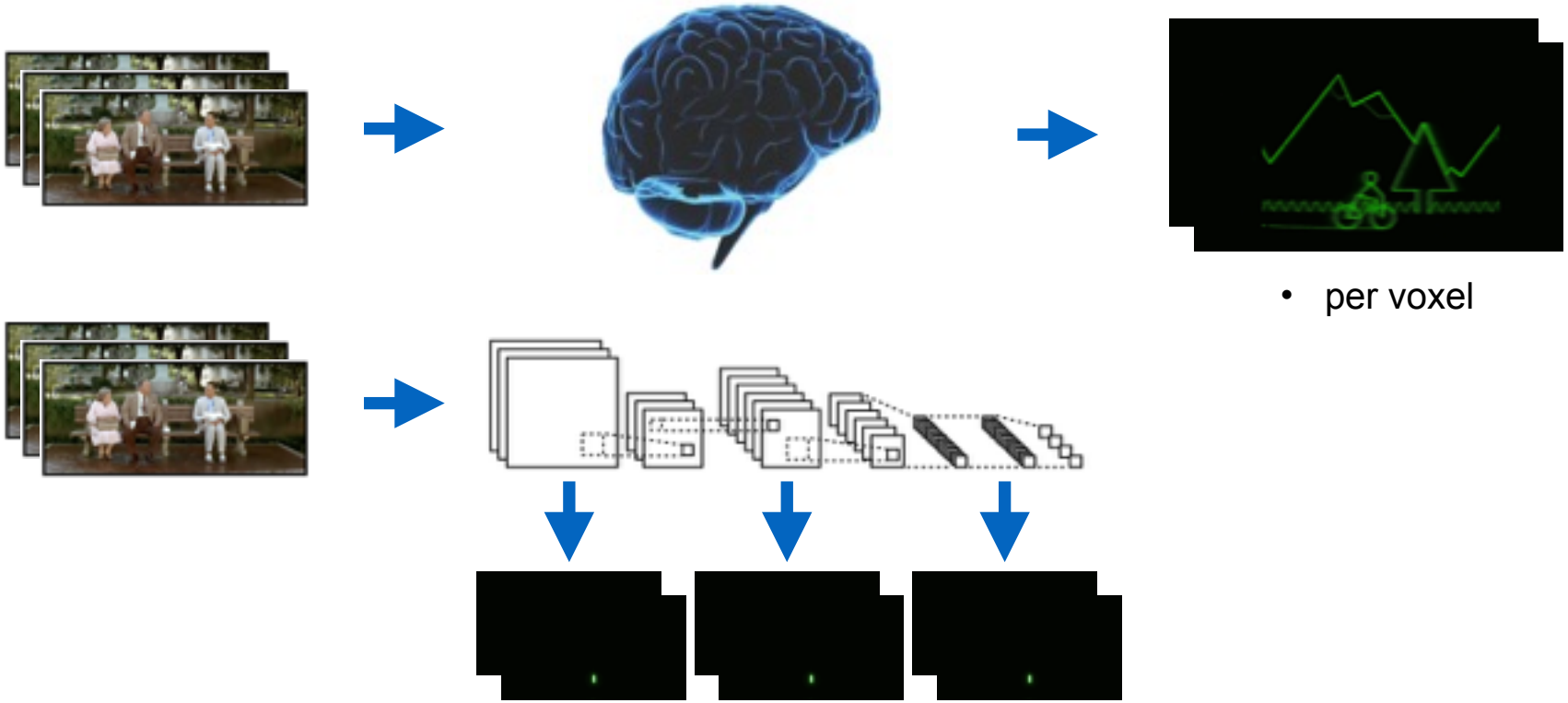
Recipe



- per voxel

- per layer

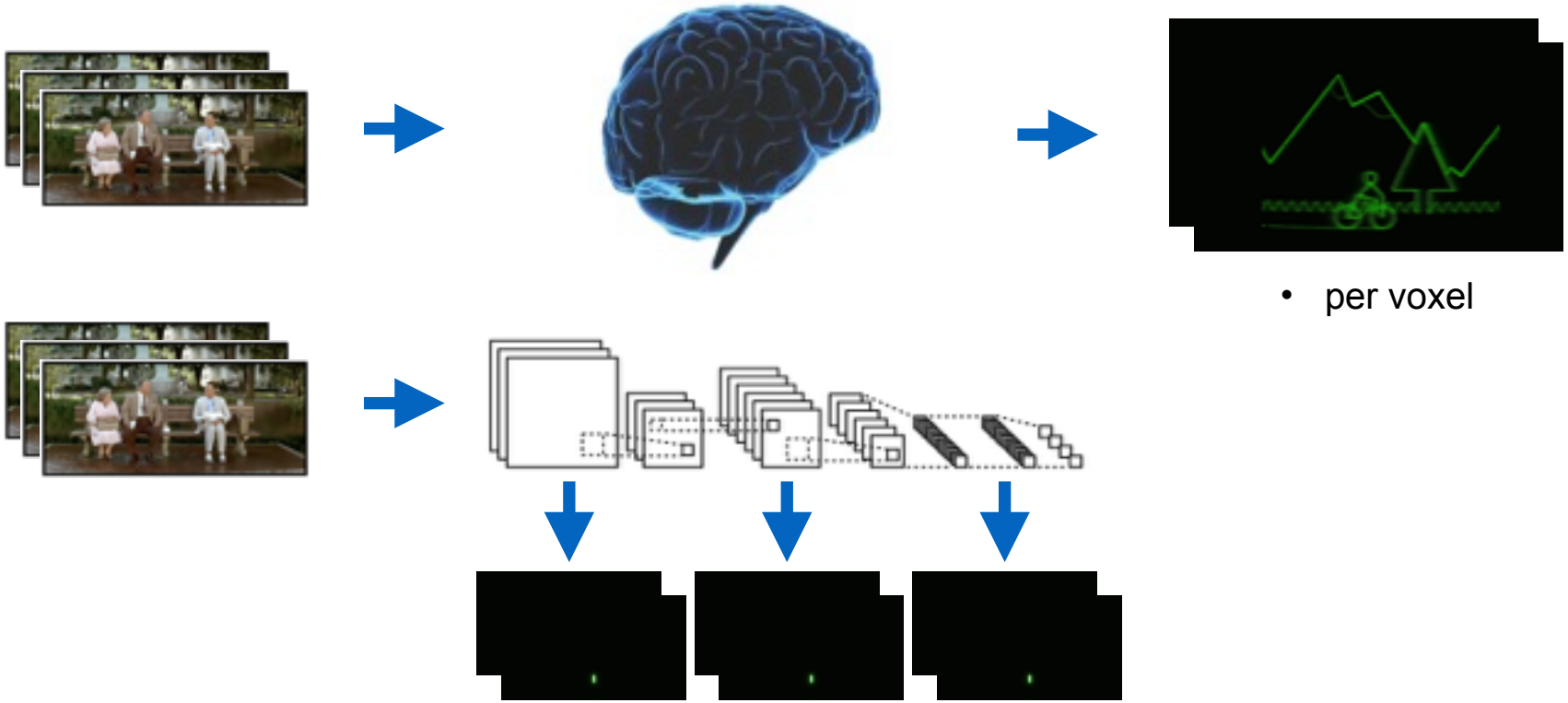
Recipe



- per voxel

- per layer

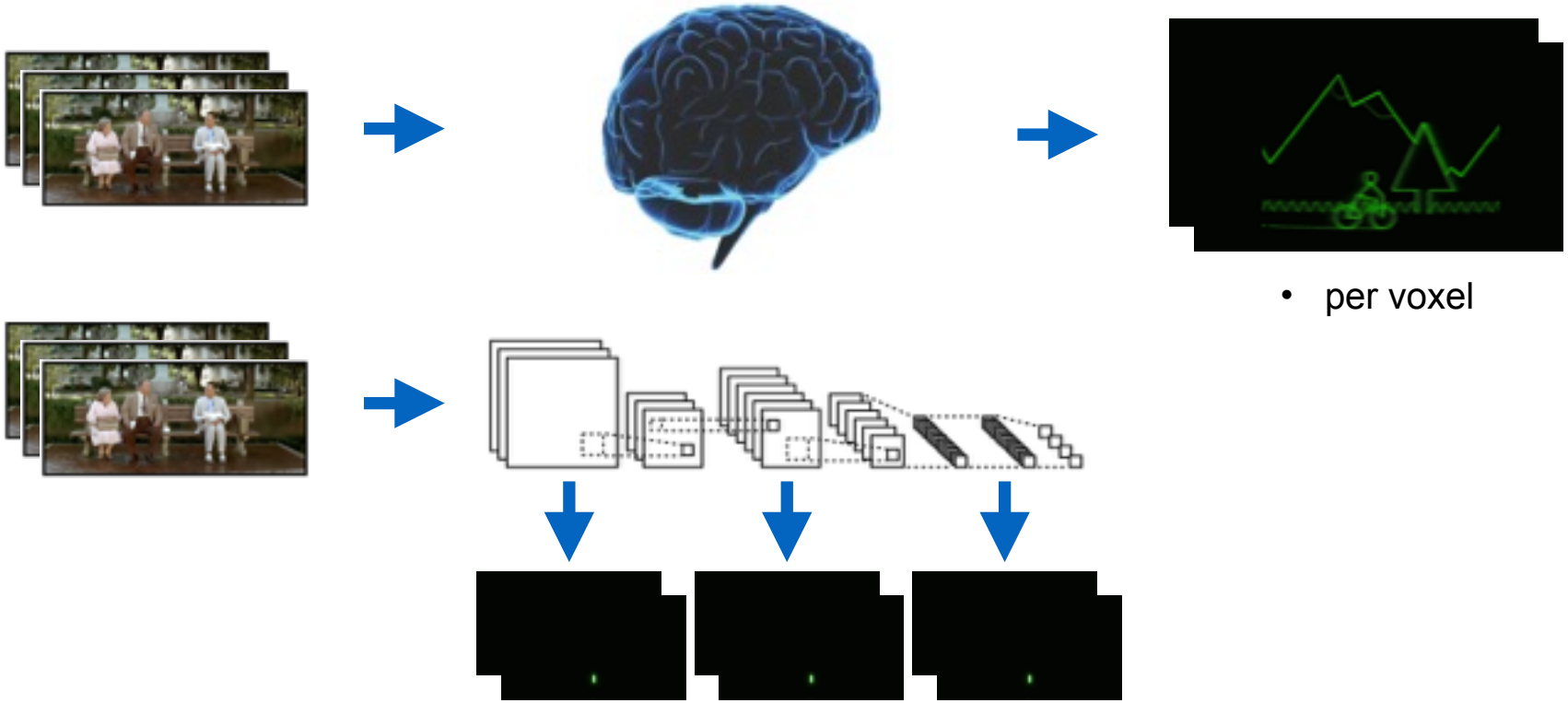
Recipe



- per voxel

- per layer

Recipe

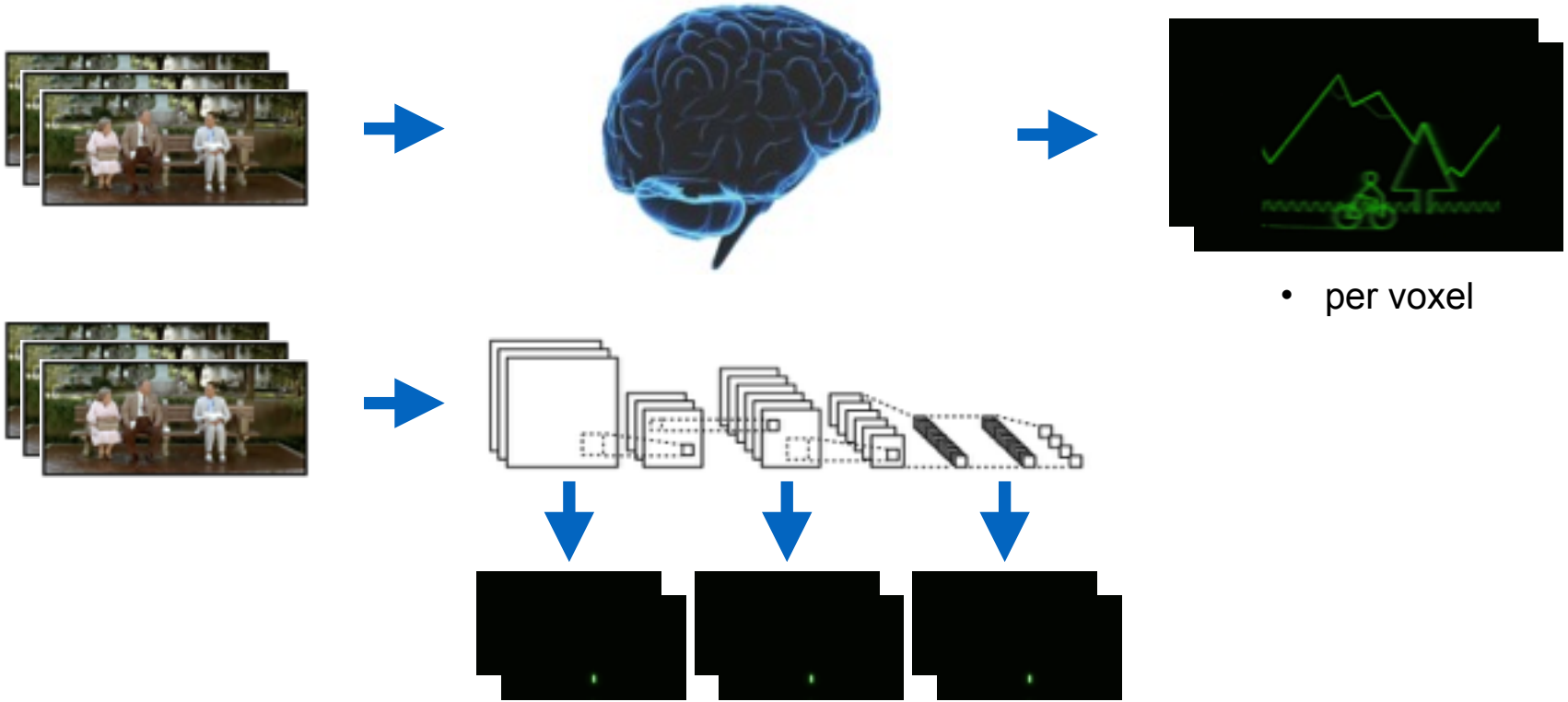


- per voxel

- per layer

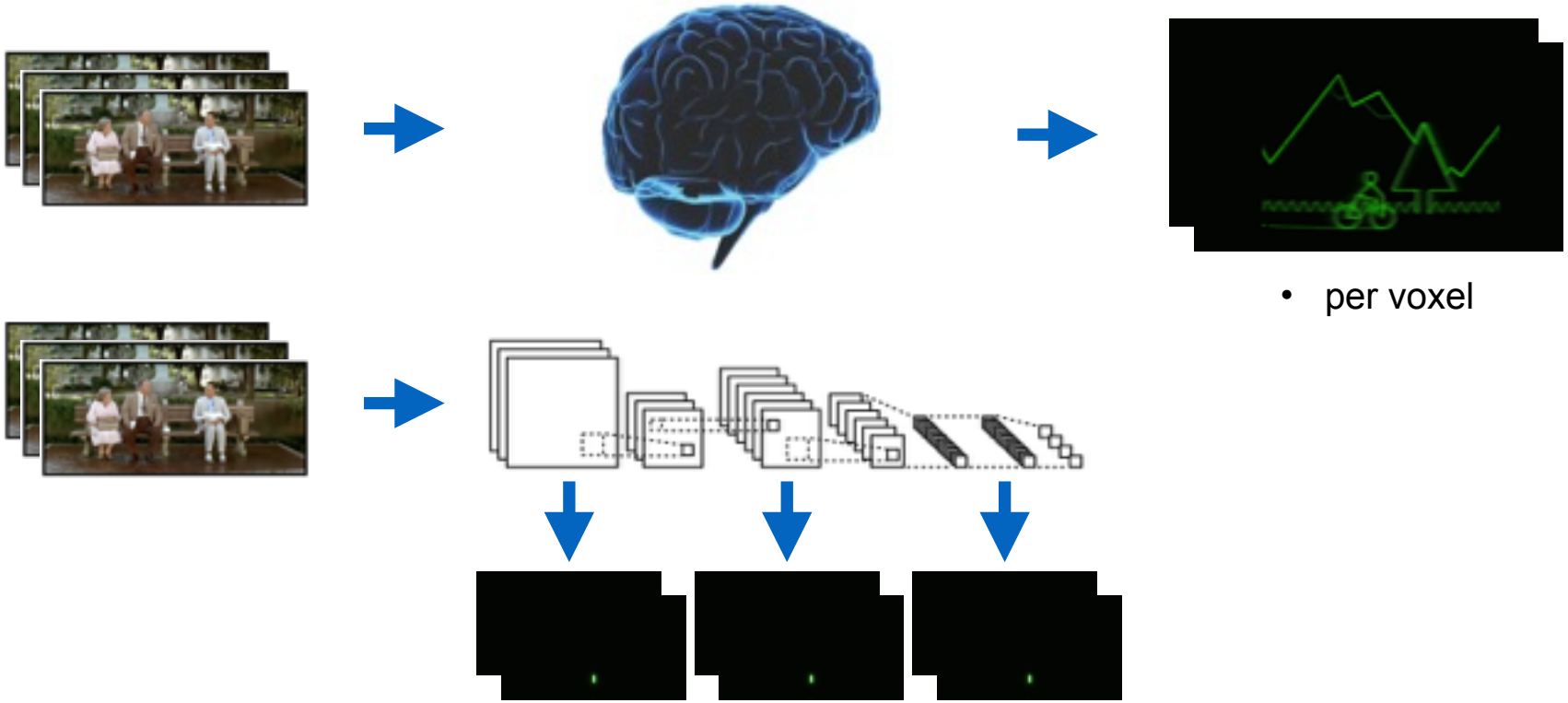
Recipe

- Present each image to the subjects & record activity of voxels



- per voxel
- per layer

Recipe

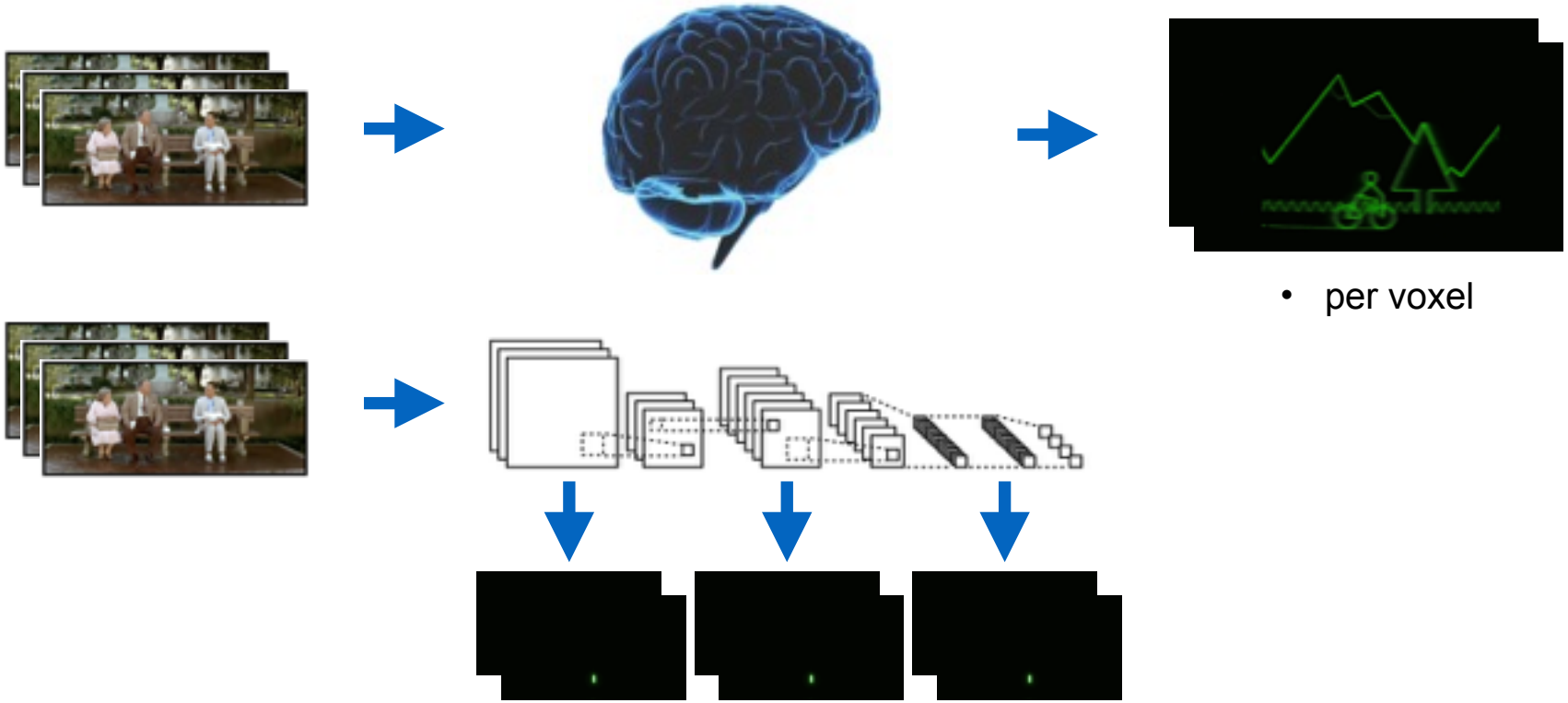


- per voxel

- per layer

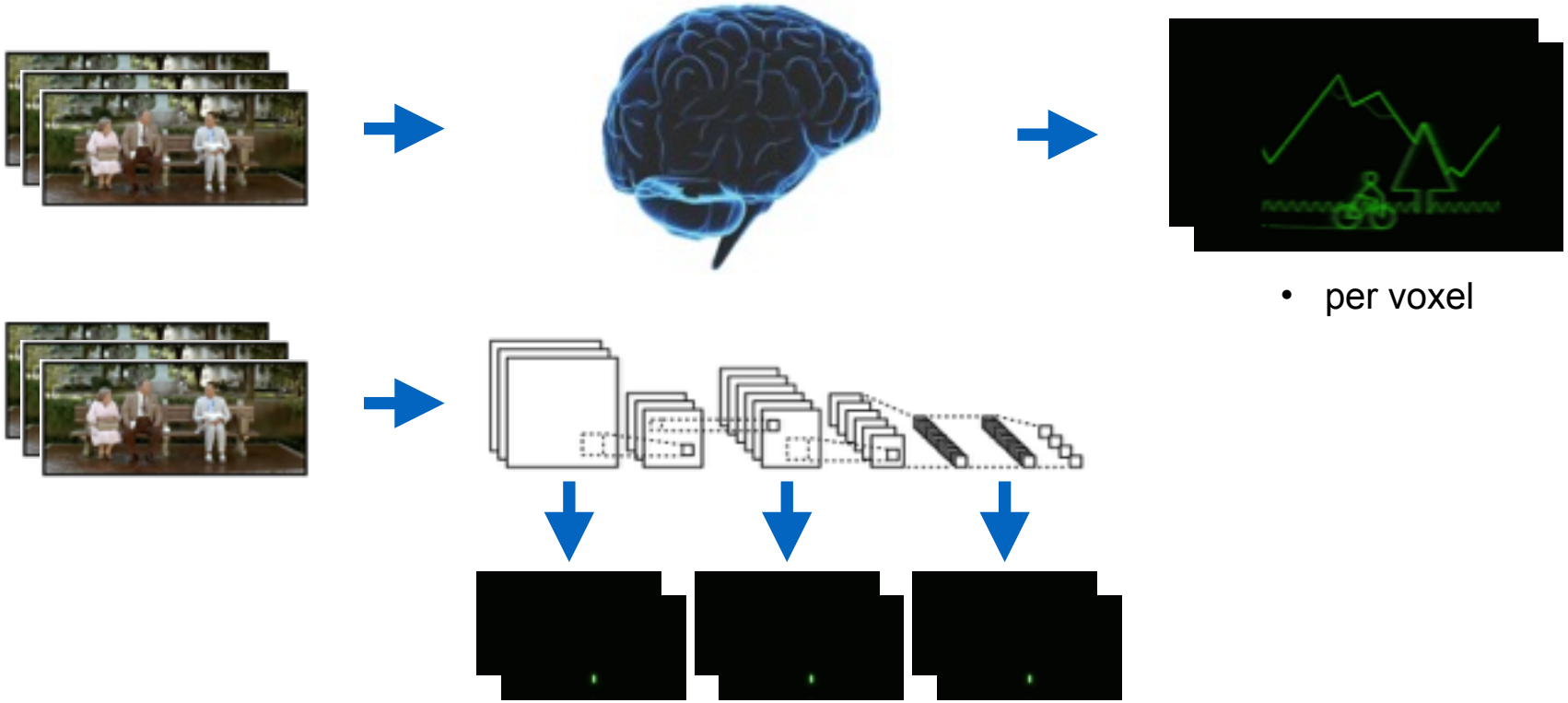
Recipe

- Present each image to the model & record activity of artificial neurons



- per voxel
- per layer

Recipe

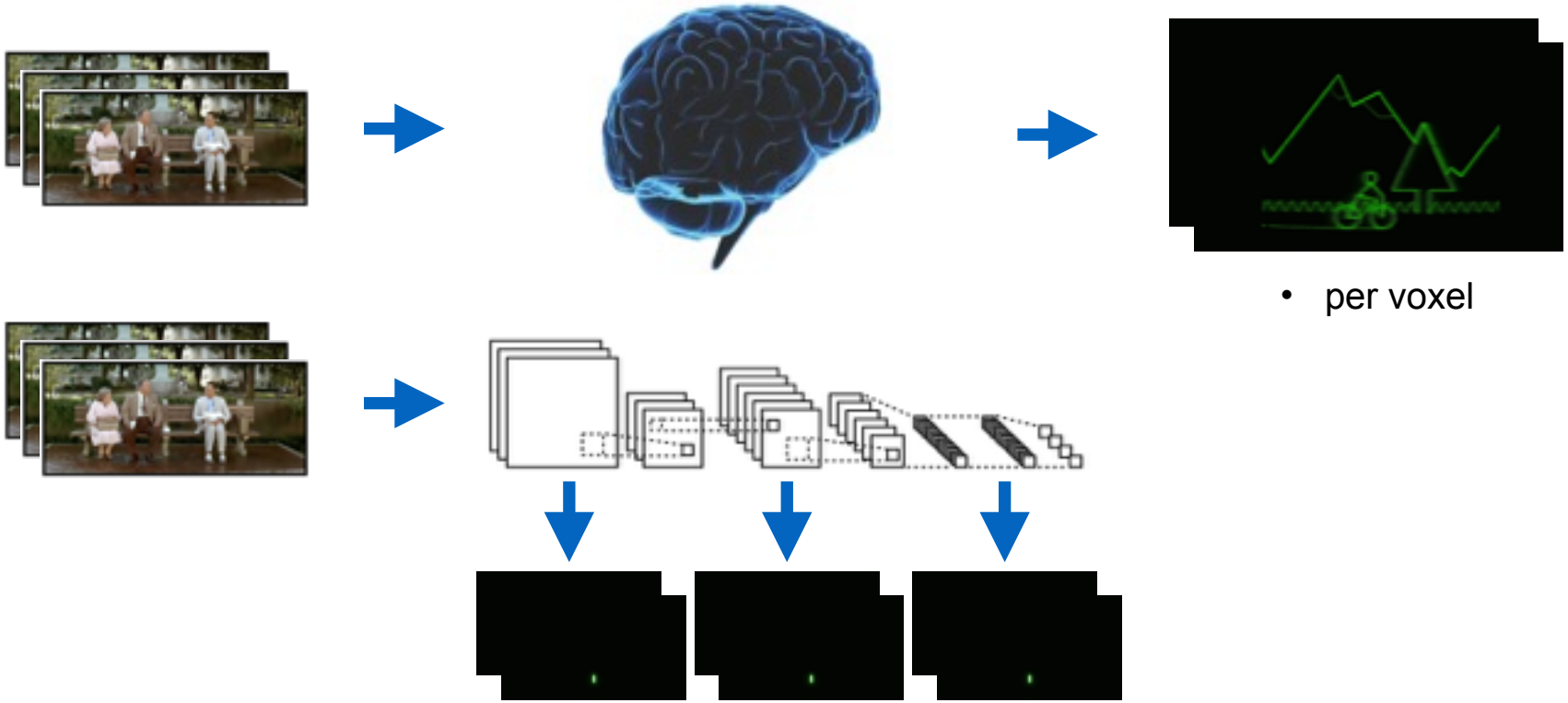


- per voxel

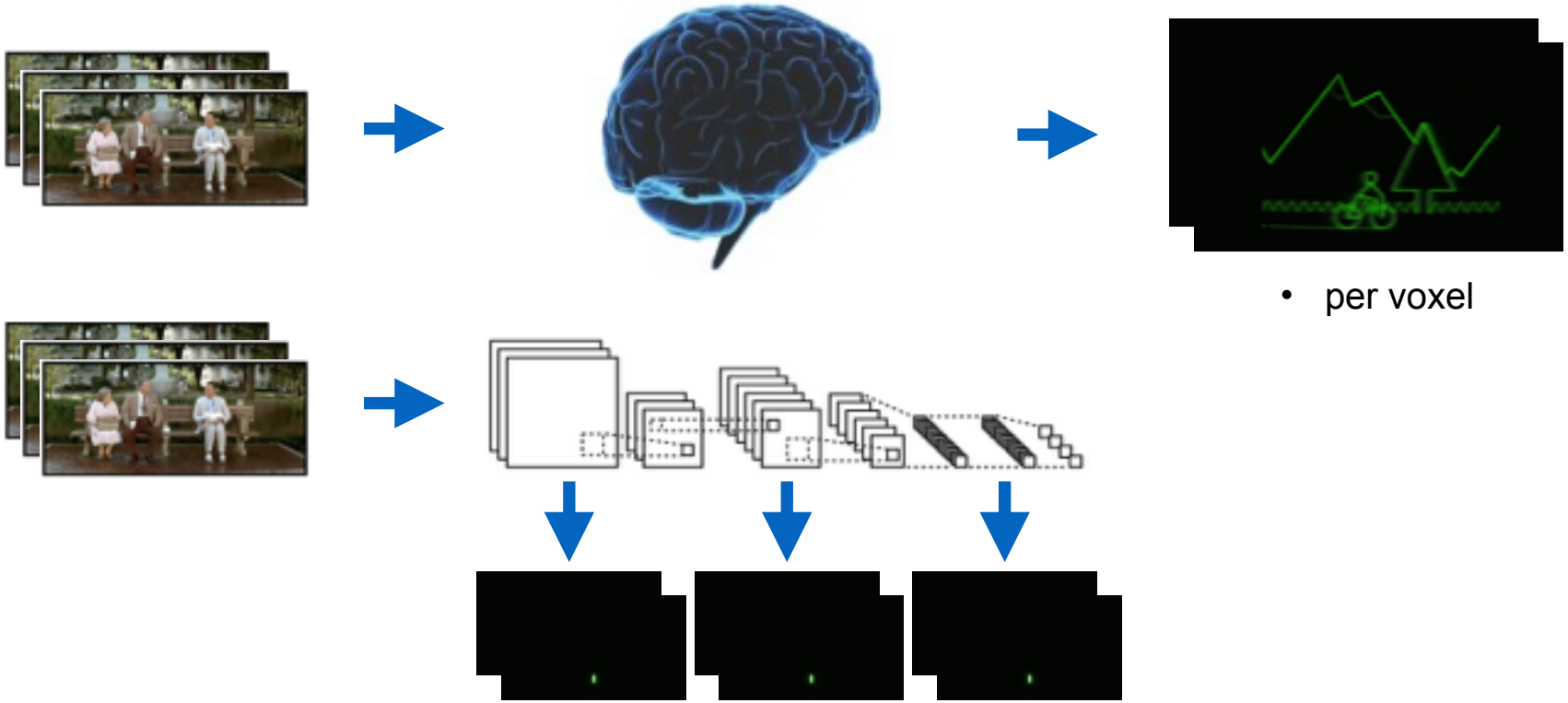
- per layer

Recipe

- Regress activity of artificial neurons in each layer onto activity of voxels



Recipe

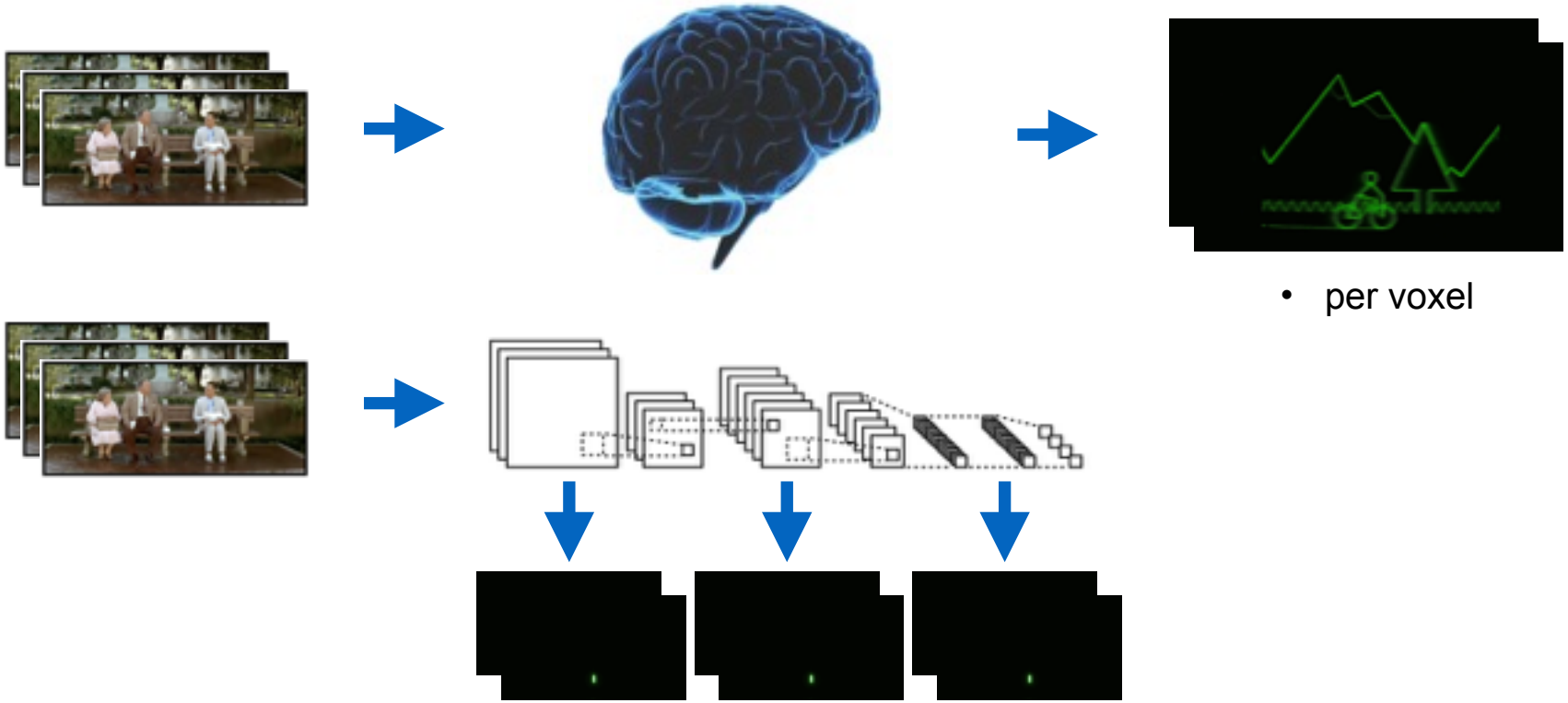


- per voxel

- per layer

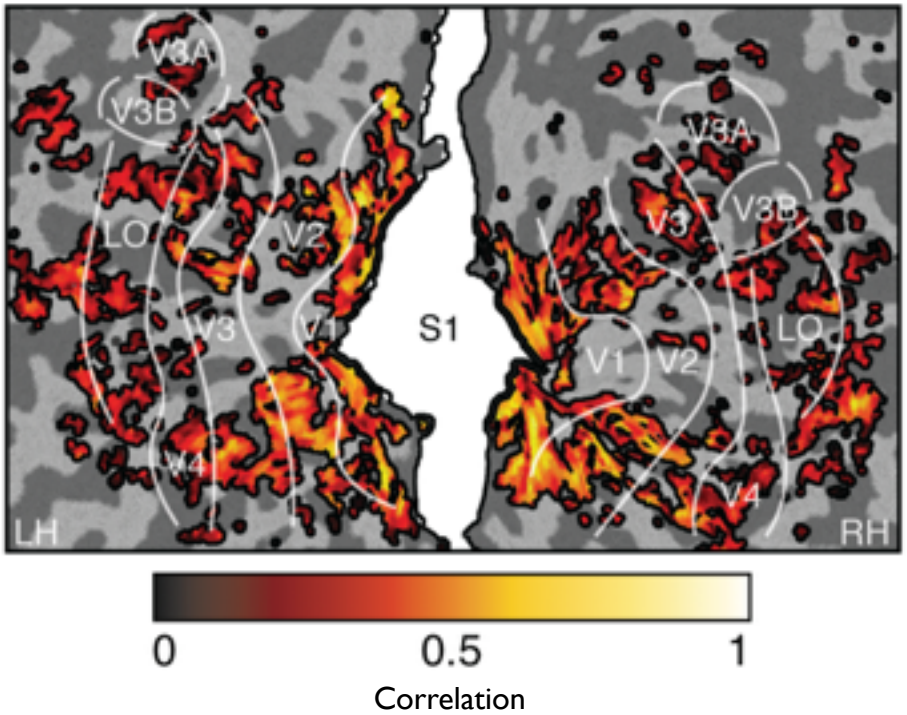
Recipe

- Assign each voxel to the layer that results in the best prediction accuracy



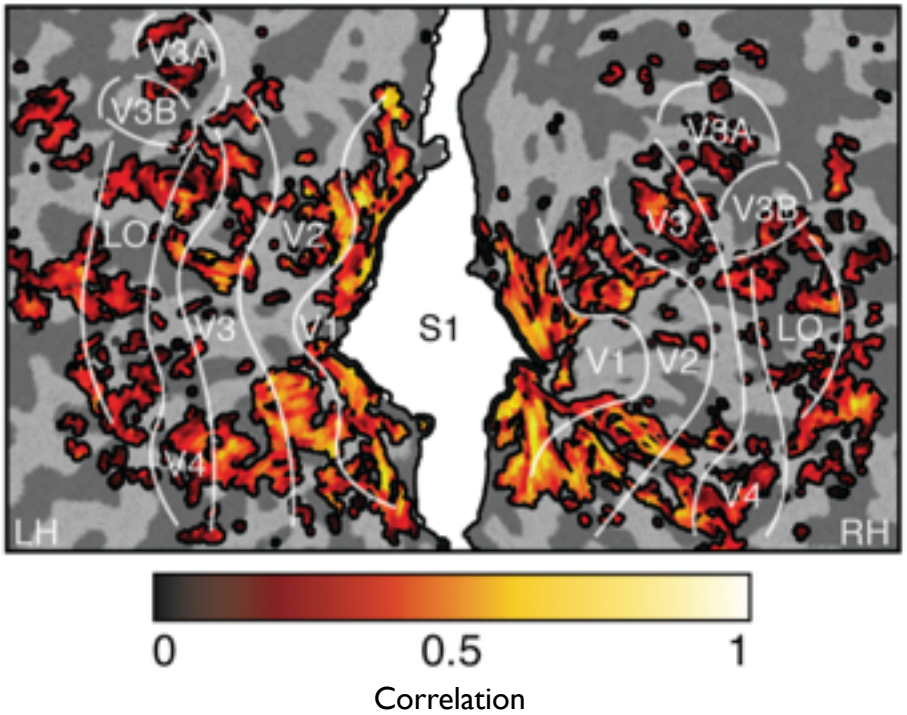
- per voxel
- per layer

Encoding of voxel responses



Güçlü, U., & van Gerven, M. A. J. (2015). J. Neuroscience.

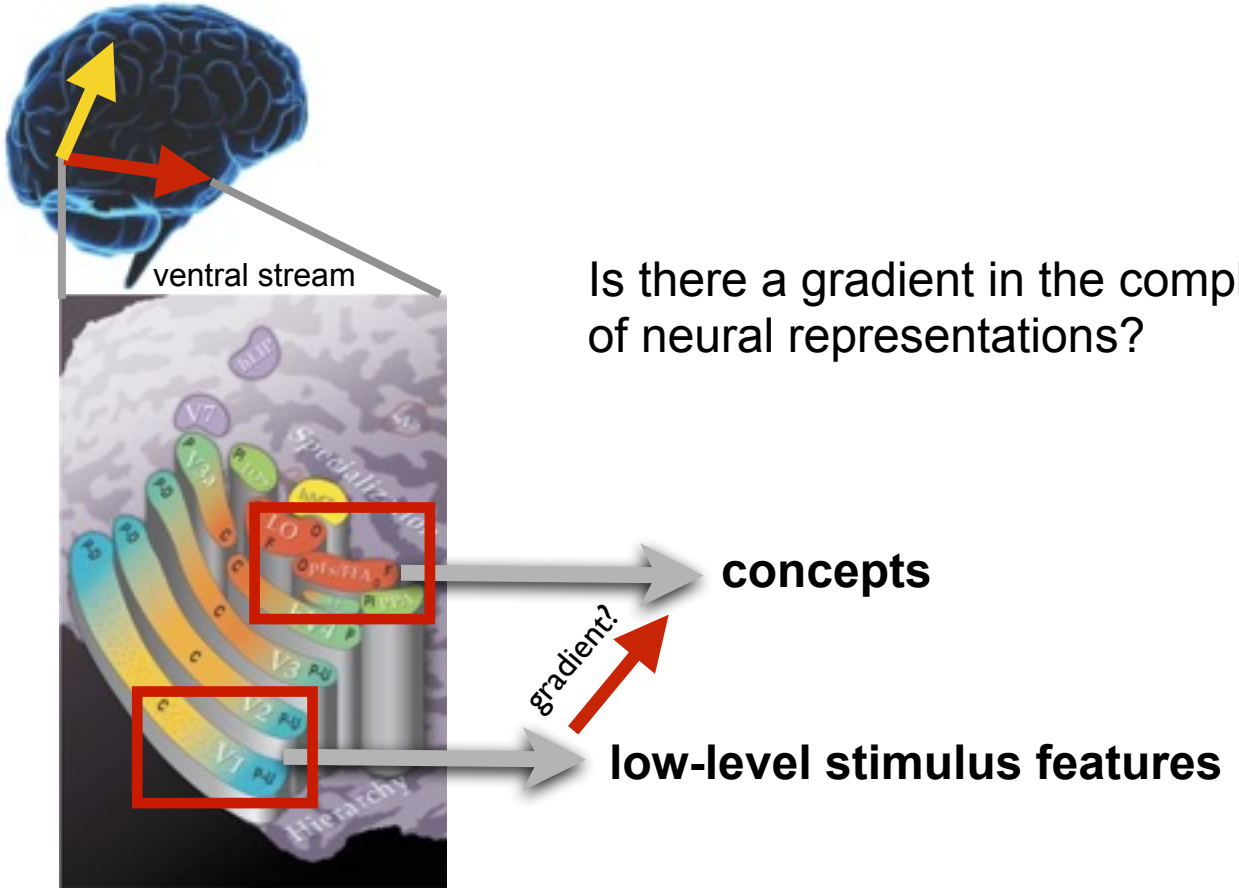
Encoding of voxel responses



- High correlations between measured and predicted responses throughout the ventral stream

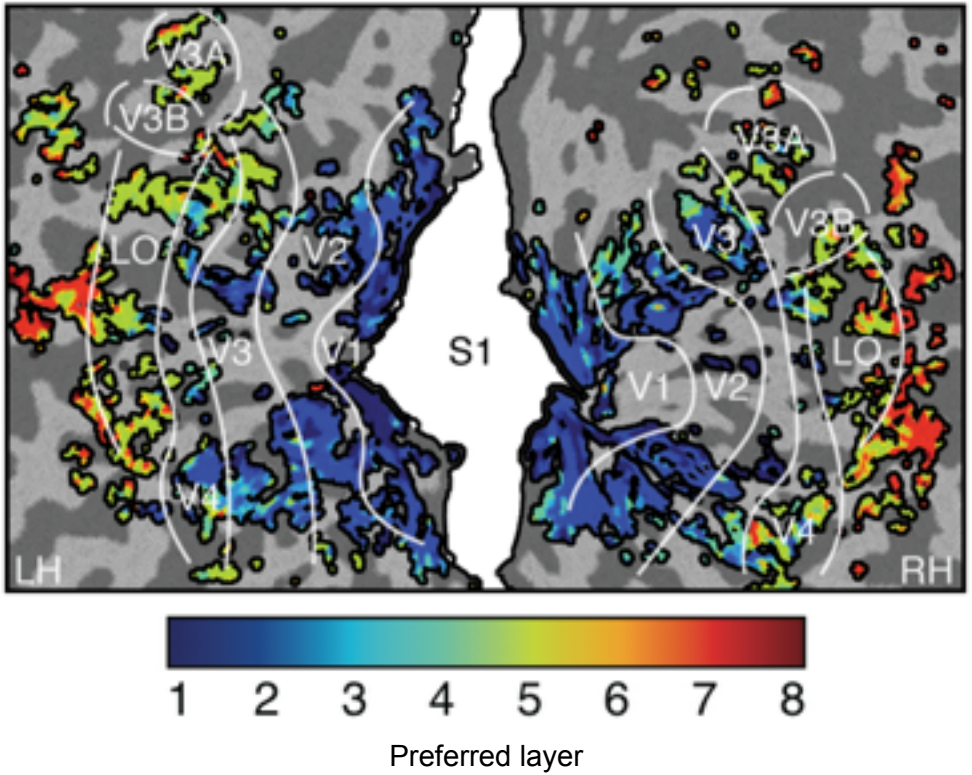
Güçlü, U., & van Gerven, M. A. J. (2015). J. Neuroscience.

From low-level stimulus properties to abstract concepts



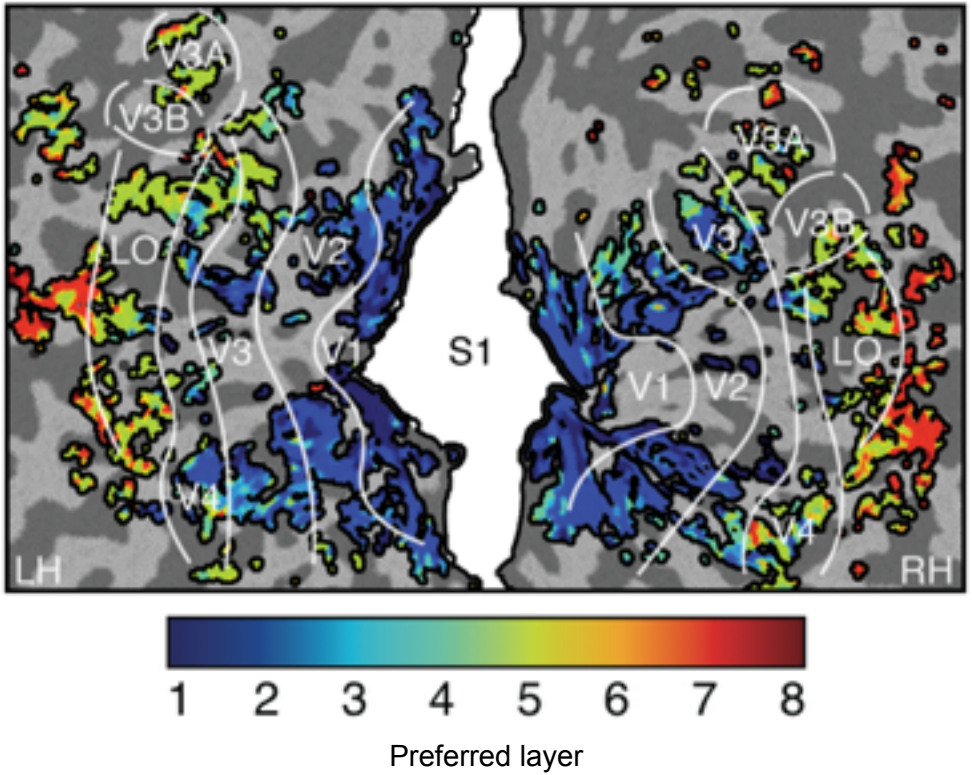
Is there a gradient in the complexity of neural representations?

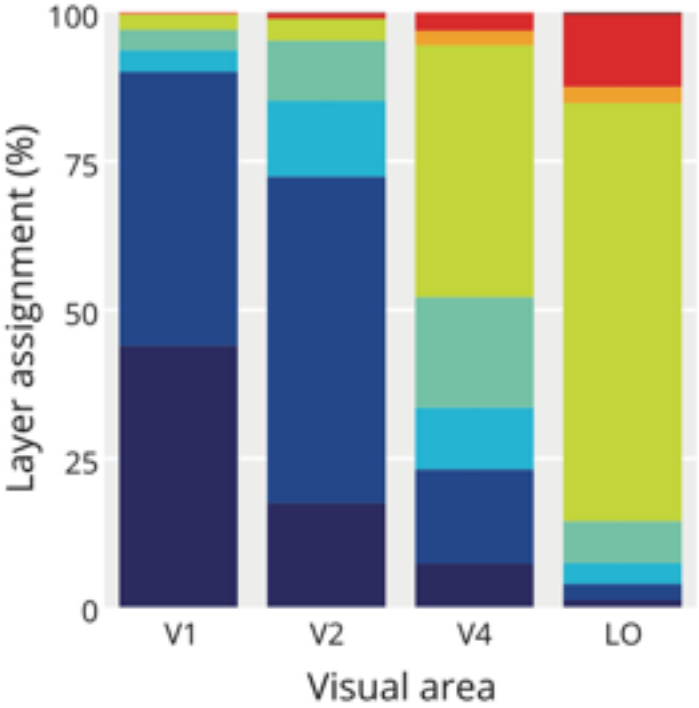
A representational gradient across the ventral stream



Güçlü, U., & van Gerven, M. A. J. (2015). J. Neuroscience.

A representational gradient across the ventral stream





Güçlü, U., & van Gerven, M. A. J. (2015). J. Neuroscience.

Characterizing non-classical receptive fields

- For each feature we can compute properties such as complexity, invariance, and RF size

Characterizing non-classical receptive fields

- For each feature we can compute properties such as complexity, invariance, and RF size
- Complexity K quantified as Kolmogorov complexity (compressed file size)

Characterizing non-classical receptive fields

- For each feature we can compute properties such as complexity, invariance, and RF size
- Complexity K quantified as Kolmogorov complexity (compressed file size)
- Invariance quantified as response changes as function of image displacement

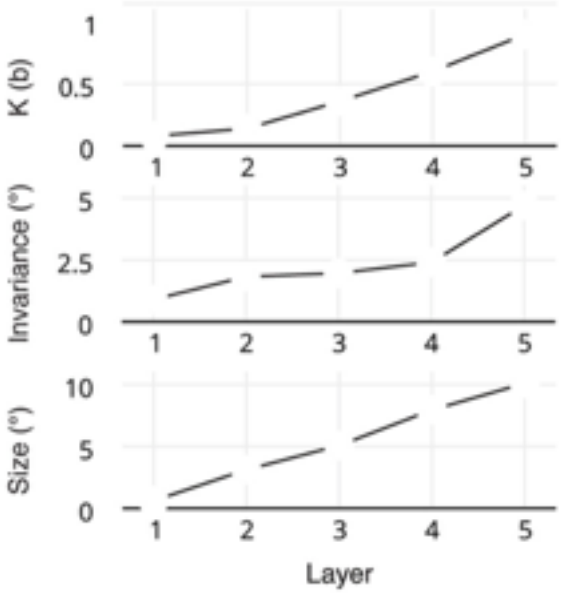
Characterizing non-classical receptive fields

- For each feature we can compute properties such as complexity, invariance, and RF size
- Complexity K quantified as Kolmogorov complexity (compressed file size)
- Invariance quantified as response changes as function of image displacement
- RF size quantified as best fitting Gaussian that predicts feature response

Characterizing non-classical receptive fields

- For each feature we can compute properties such as complexity, invariance, and RF size
- Complexity K quantified as Kolmogorov complexity (compressed file size)
- Invariance quantified as response changes as function of image displacement
- RF size quantified as best fitting Gaussian that predicts feature response

Layer-wise characterization



Characterization in terms of feature classes

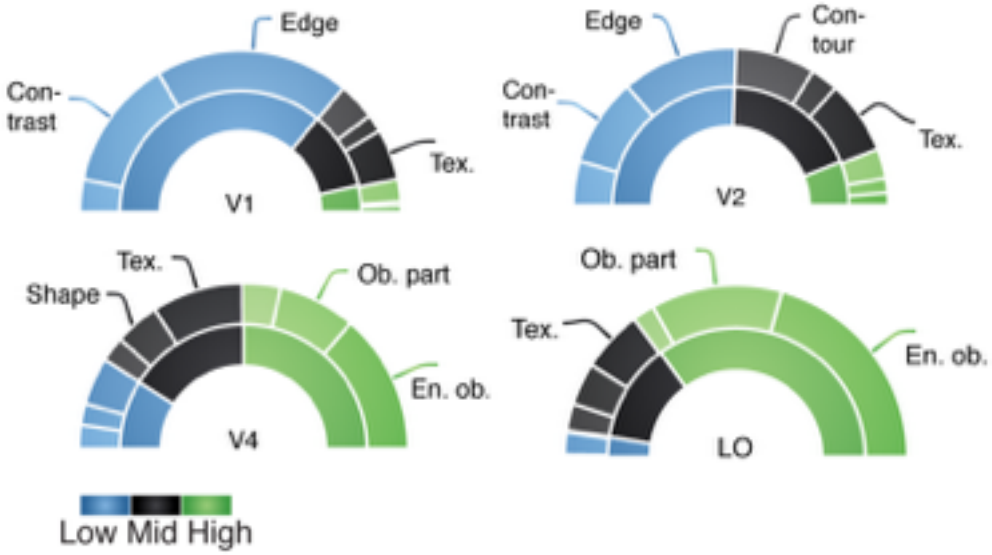
- Features were manually assigned to low-level (blob, contrast, edge), mid-level (contour, shape, texture) and high-level (irregular pattern, object part, entire object) feature classes

Characterization in terms of feature classes

- Features were manually assigned to low-level (blob, contrast, edge), mid-level (contour, shape, texture) and high-level (irregular pattern, object part, entire object) feature classes
- Voxels were assigned to that feature class which best predicted its response

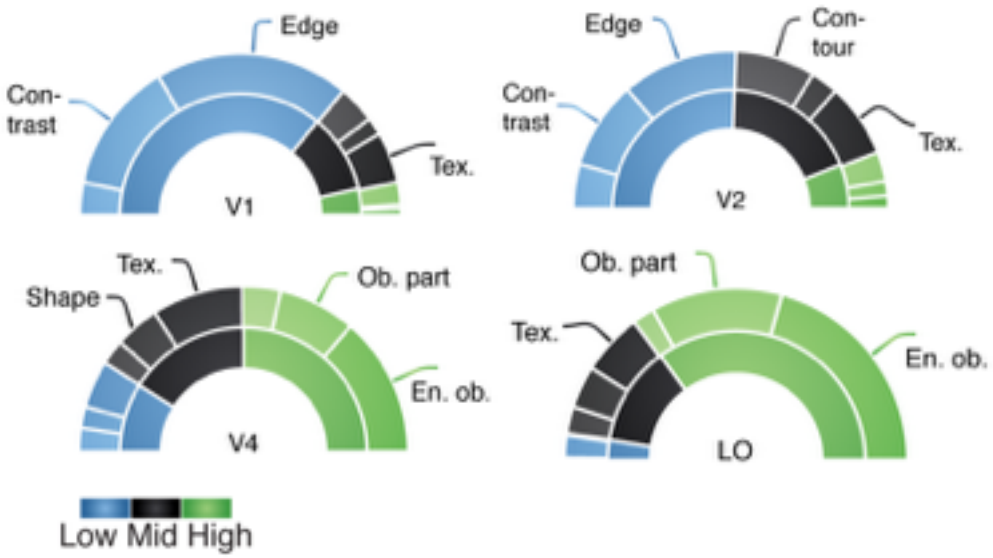
Characterization in terms of feature classes

- Features were manually assigned to low-level (blob, contrast, edge), mid-level (contour, shape, texture) and high-level (irregular pattern, object part, entire object) feature classes
- Voxels were assigned to that feature class which best predicted its response



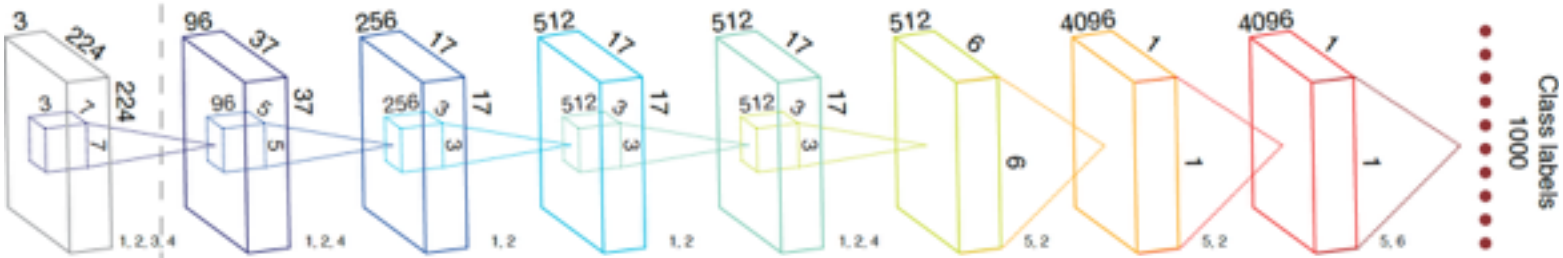
Characterization in terms of feature classes

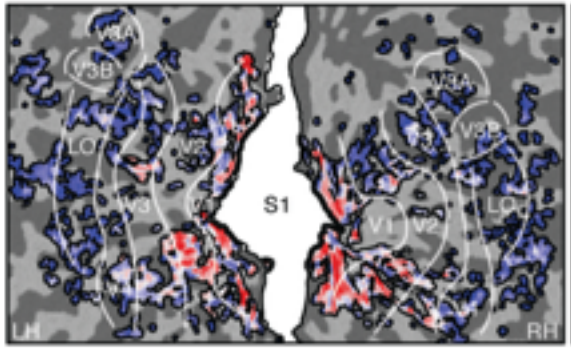
- Features were manually assigned to low-level (blob, contrast, edge), mid-level (contour, shape, texture) and high-level (irregular pattern, object part, entire object) feature classes
- Voxels were assigned to that feature class which best predicted its response



- Confirmation that more complex feature classes better predict downstream responses

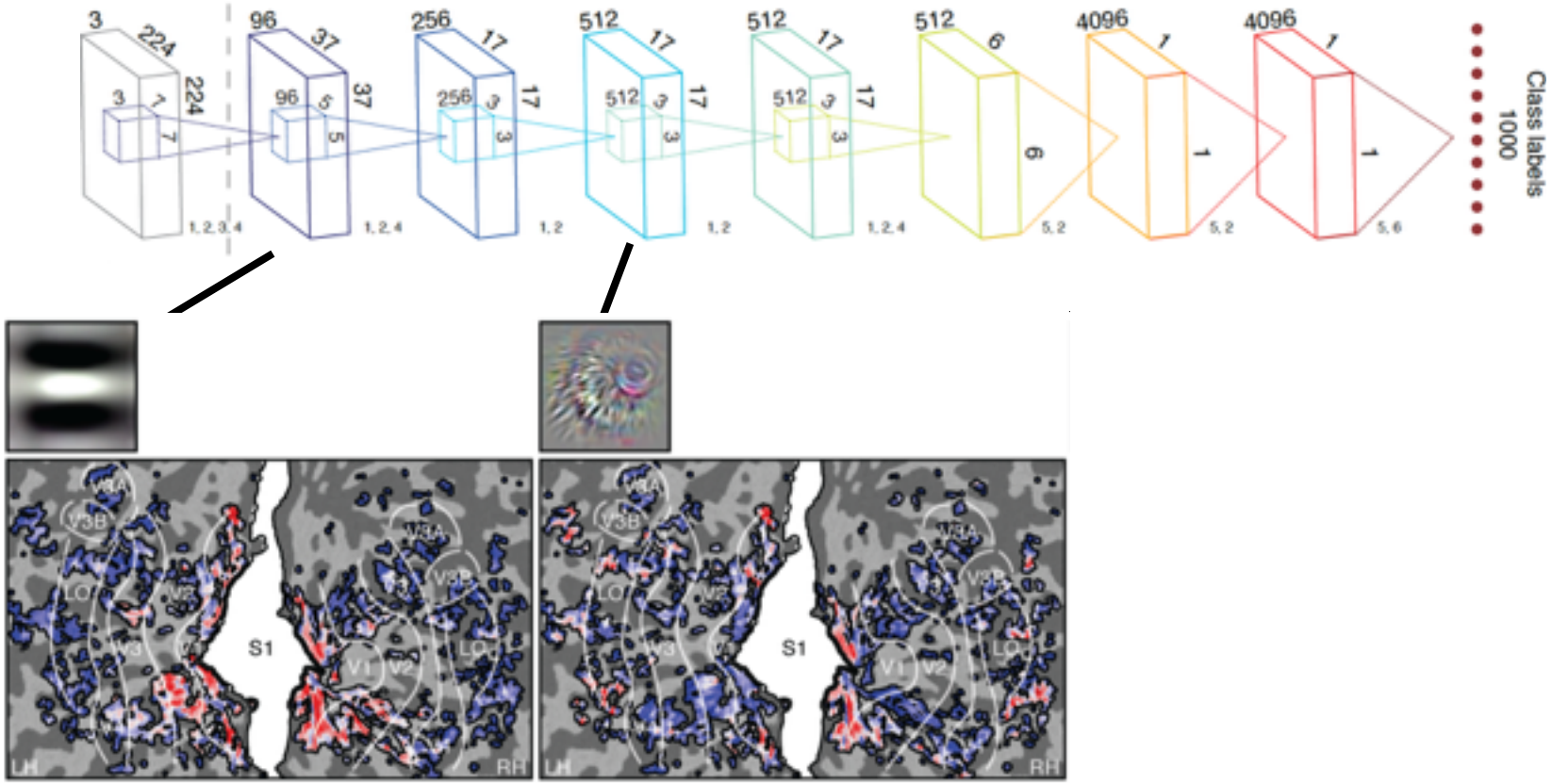
Mapping low-, mid- and high-level features to the cortical sheet





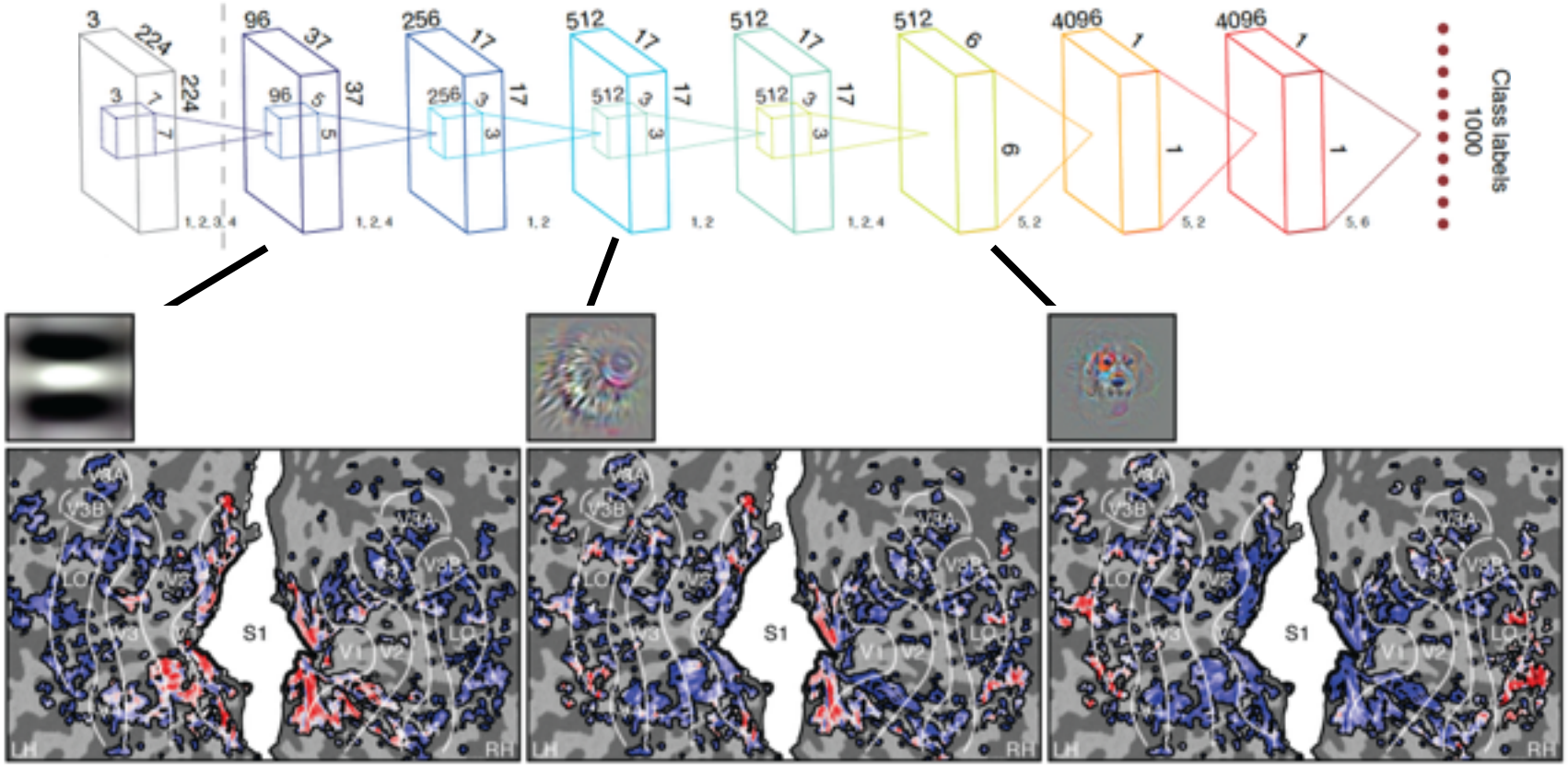
Güçlü, U., & van Gerven, M. A. J. (2015). J. Neuroscience.

Mapping low-, mid- and high-level features to the cortical sheet



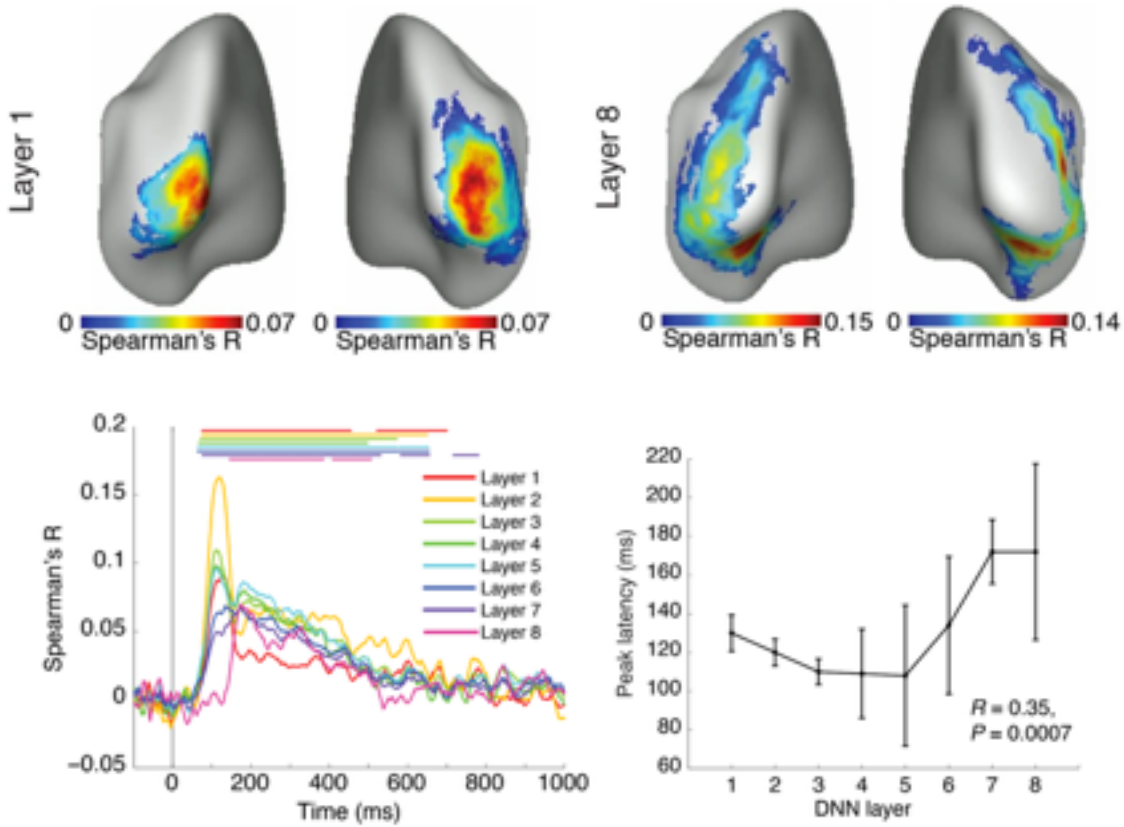
Güçlü, U., & van Gerven, M. A. J. (2015). J. Neuroscience.

Mapping low-, mid- and high-level features to the cortical sheet



Güçlü, U., & van Gerven, M. A. J. (2015). J. Neuroscience.

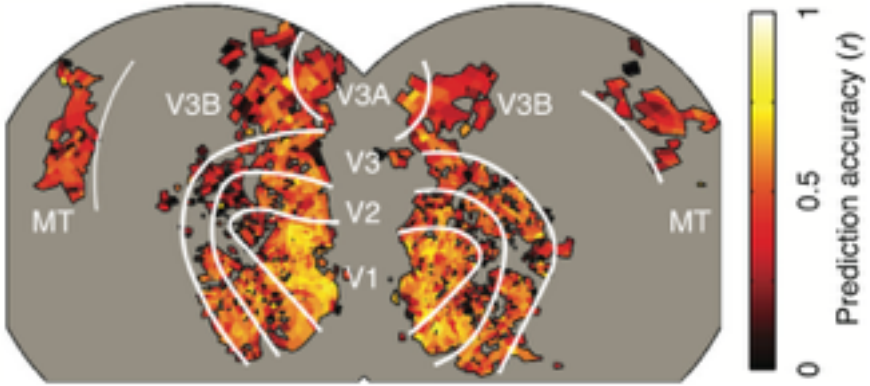
Related work: A representational gradient in space and time

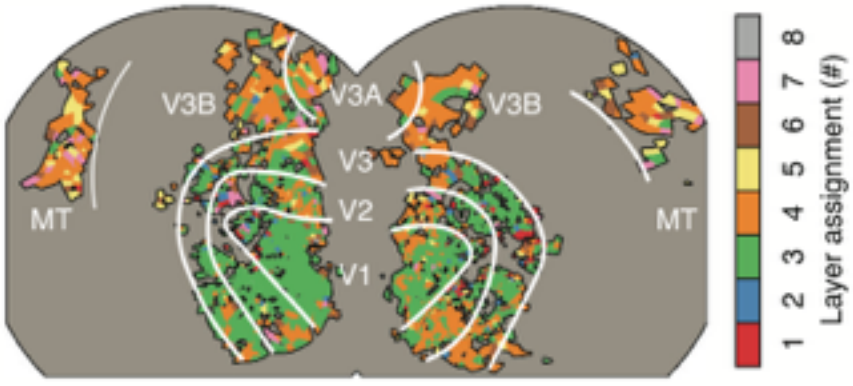


Cichy, R.M. et al. 2016. Deep neural networks predict hierarchical spatio-temporal cortical dynamics of human visual object recognition. arXiv Prepr.

A representational gradient across the dorsal stream

- DNN trained for action recognition
- Learns spatio-temporal receptive fields
- Recovers representational gradient in dorsal stream

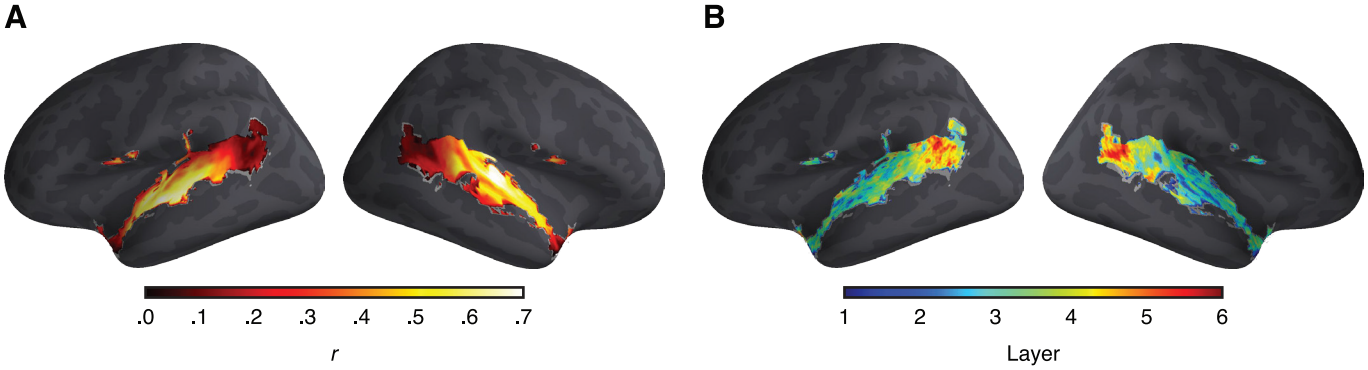




Güçlü, U., & van Gerven, M. A. J. (2015). Neuroimage.

A representational gradient across superior temporal gyrus

- DNN trained on automatic music tagging (genre, instrument, mood, etc.)
- Learns temporal receptive fields
- Recovers representational gradient in superior temporal gyrus

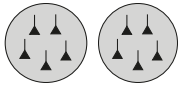


Güçlü, U., & van Gerven, M. A. J. (2016). arxiv.

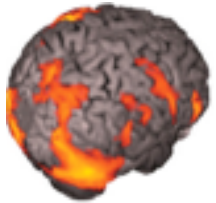
Modeling how the brain responds to the environment

stimulus **x**

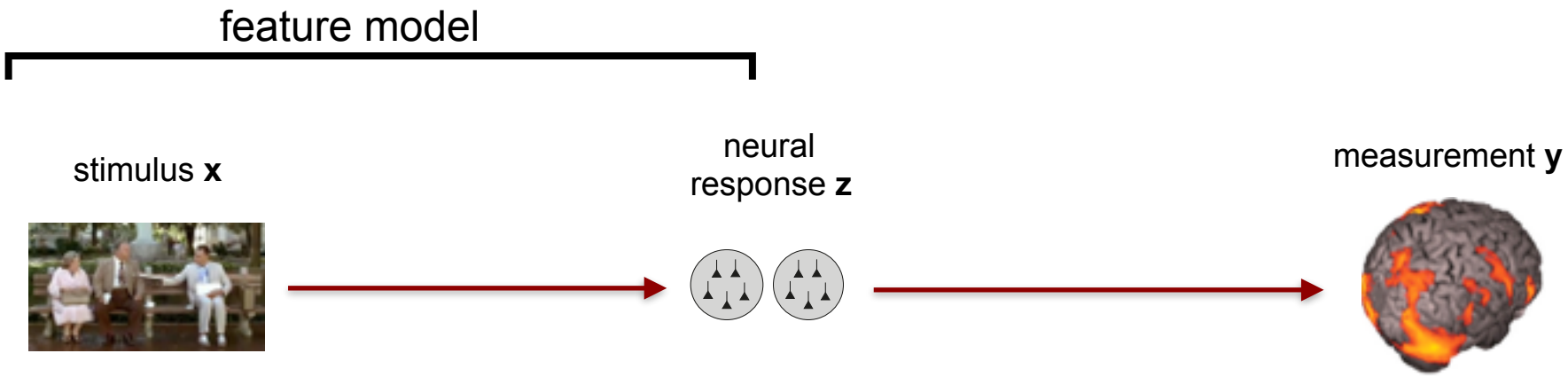
neural response **z**



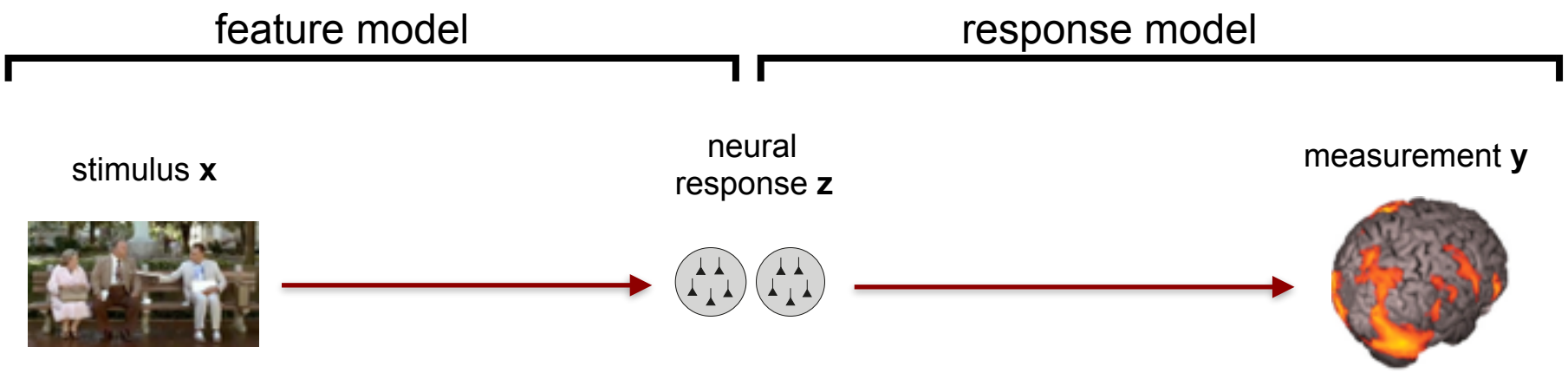
measurement **y**



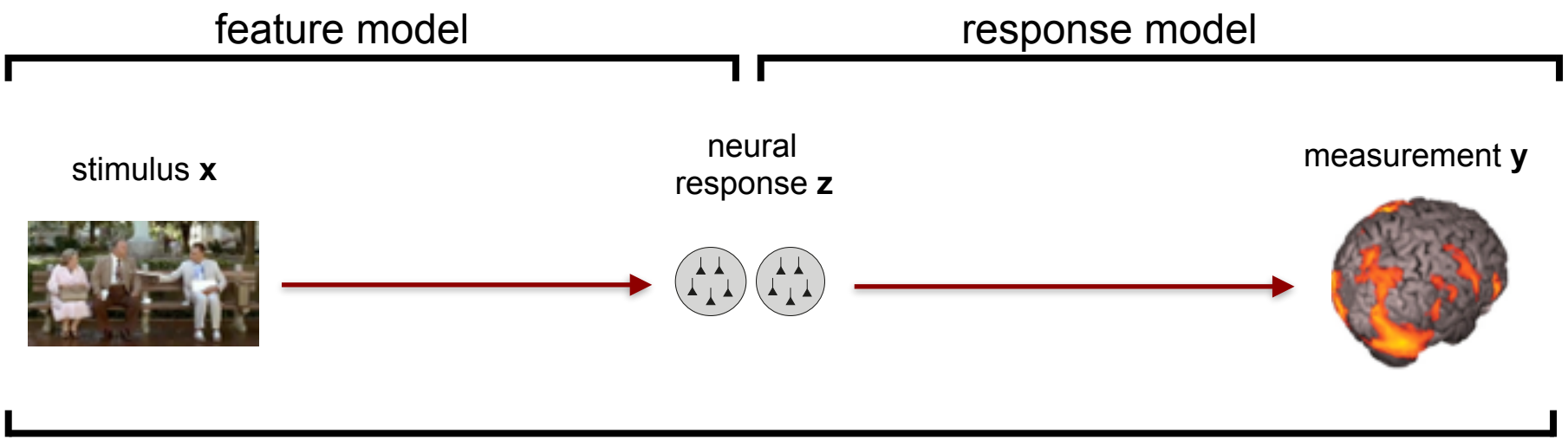
Modeling how the brain responds to the environment



Modeling how the brain responds to the environment



Modeling how the brain responds to the environment



Encoding model

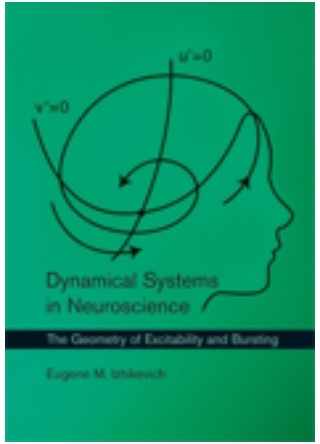
$$p(\mathbf{y} \mid \mathbf{x})$$

The brain as a dynamical system

- The human brain is not a passive device which idles until a stimulus is presented

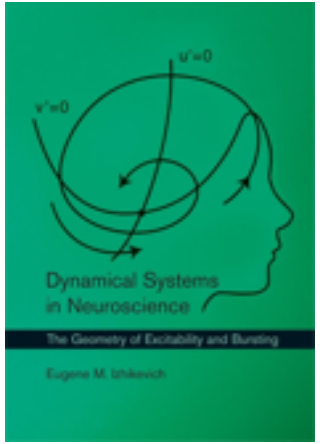
The brain as a dynamical system

- The human brain is not a passive device which idles until a stimulus is presented
- It can be seen as a complex dynamical system whose dynamics are continuously perturbed by bottom-up and top-down processes



The brain as a dynamical system

- The human brain is not a passive device which idles until a stimulus is presented
- It can be seen as a complex dynamical system whose dynamics are continuously perturbed by bottom-up and top-down processes



- How to model these dynamics?

Modeling dynamics: Recurrent neural networks

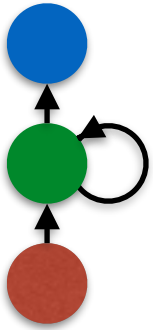
- Recurrent networks capture temporal dependencies

Modeling dynamics: Recurrent neural networks

- Recurrent networks capture temporal dependencies
- Models of dynamical systems rather than static functions (cf. feedforward ANNs)

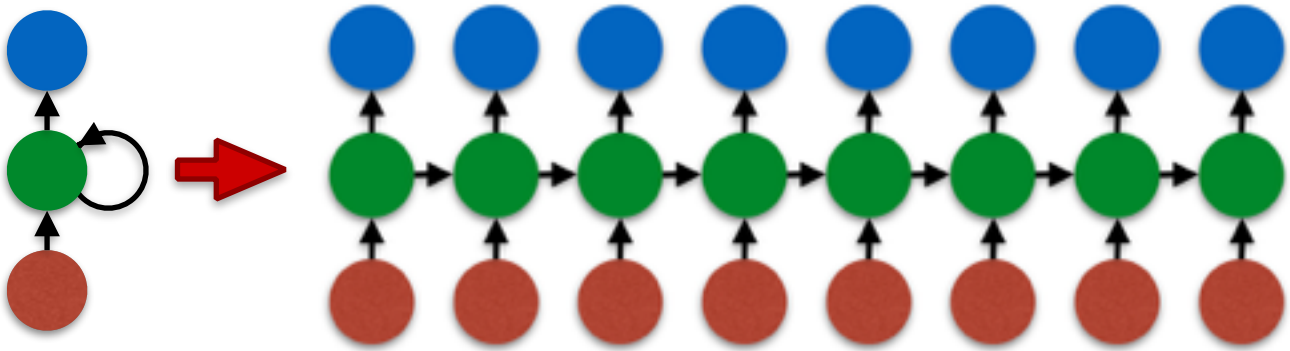
Modeling dynamics: Recurrent neural networks

- Recurrent networks capture temporal dependencies
- Models of dynamical systems rather than static functions (cf. feedforward ANNs)



Modeling dynamics: Recurrent neural networks

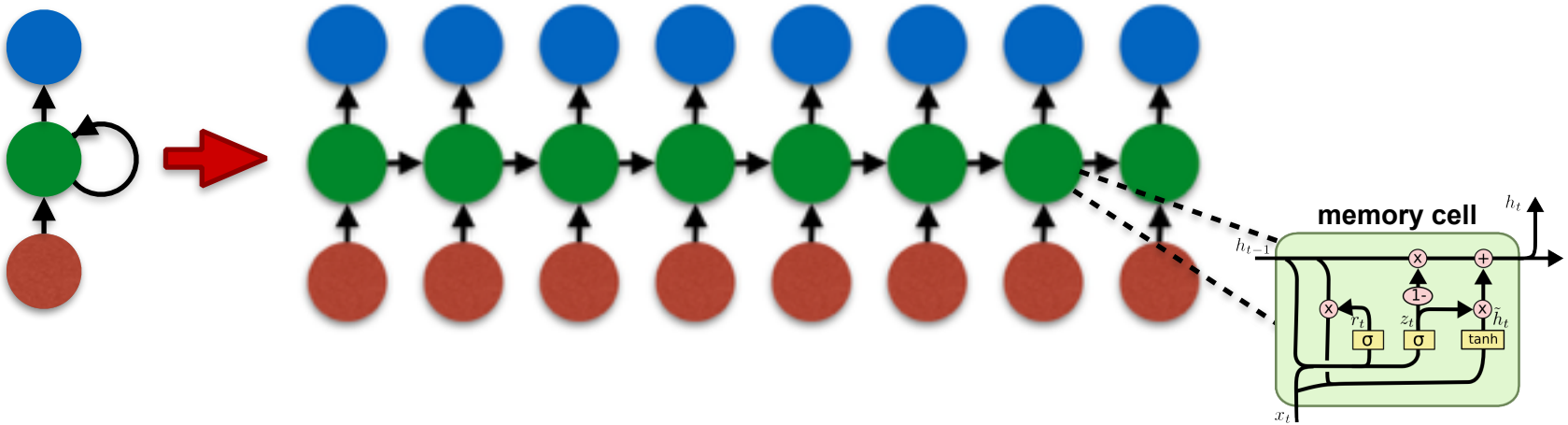
- Recurrent networks capture temporal dependencies
- Models of dynamical systems rather than static functions (cf. feedforward ANNs)



- Inference realized by unrolling the network and applying backpropagation through time

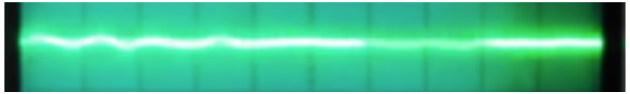
Modeling dynamics: Recurrent neural networks

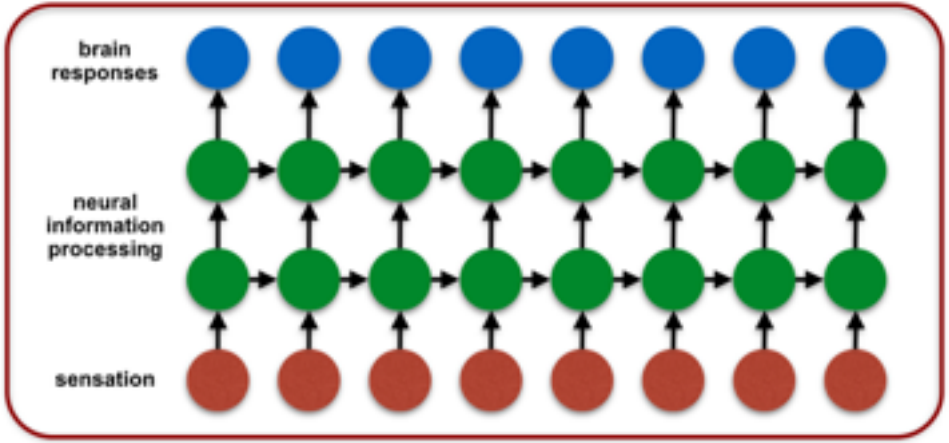
- Recurrent networks capture temporal dependencies
- Models of dynamical systems rather than static functions (cf. feedforward ANNs)



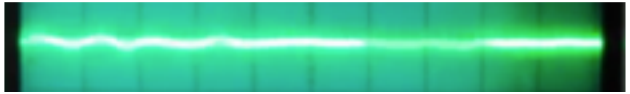
- Inference realized by unrolling the network and applying backpropagation through time
- Make use of memory cells to keep explicit track of long-term dependencies

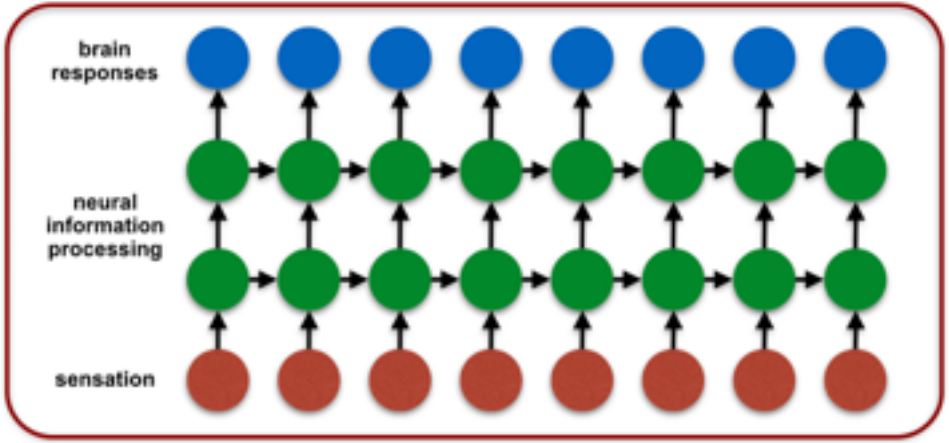
Using RNNs to model stimulus-response relationships



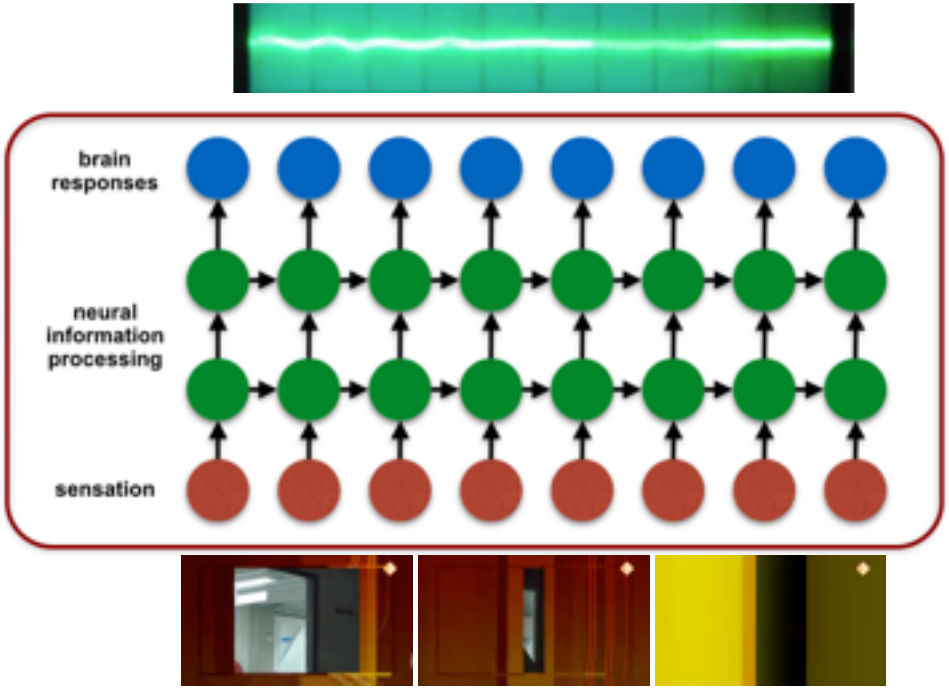


Using RNNs to model stimulus-response relationships



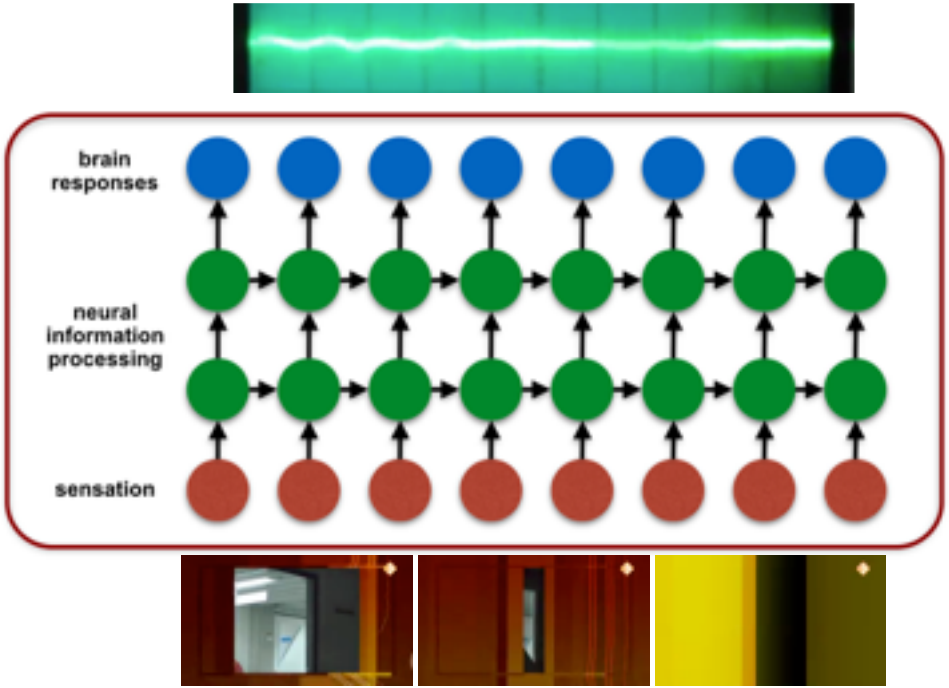


Using RNNs to model stimulus-response relationships



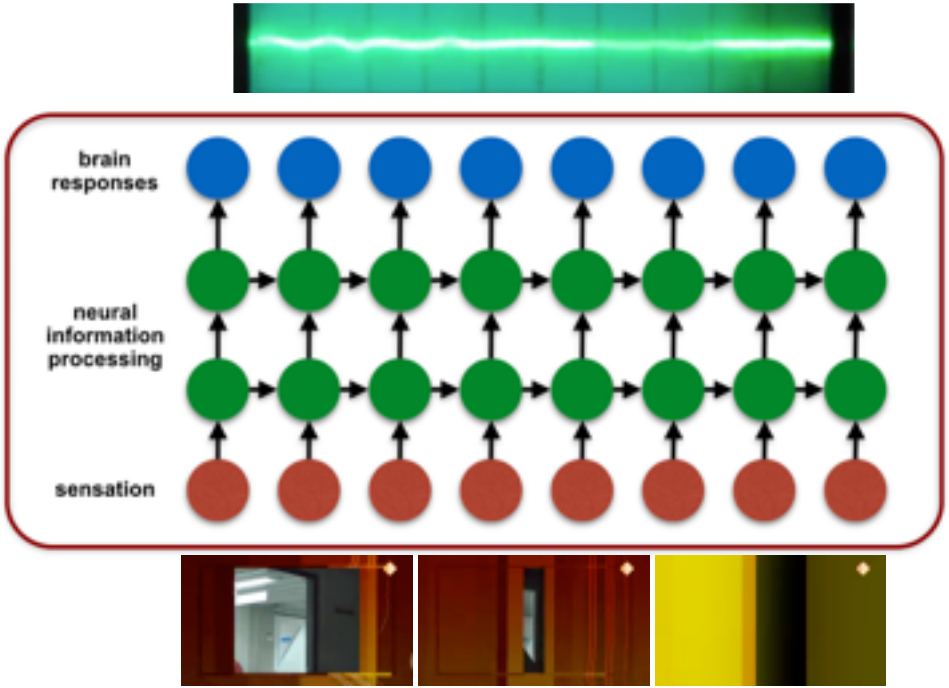
- Subjects presented with movies

Using RNNs to model stimulus-response relationships



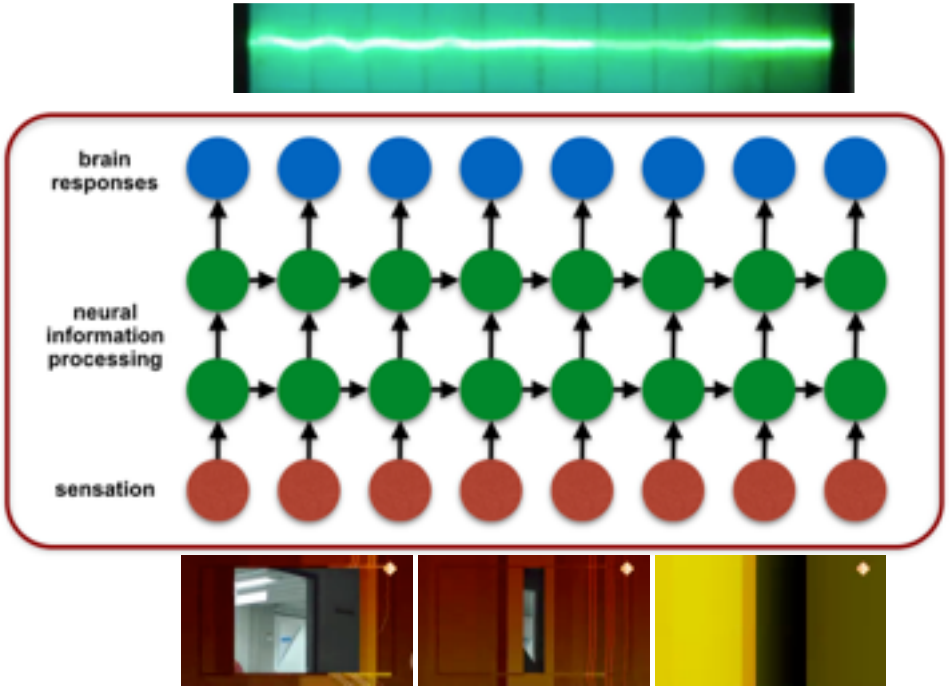
- Subjects presented with movies
- BOLD response measured along visual pathway

Using RNNs to model stimulus-response relationships



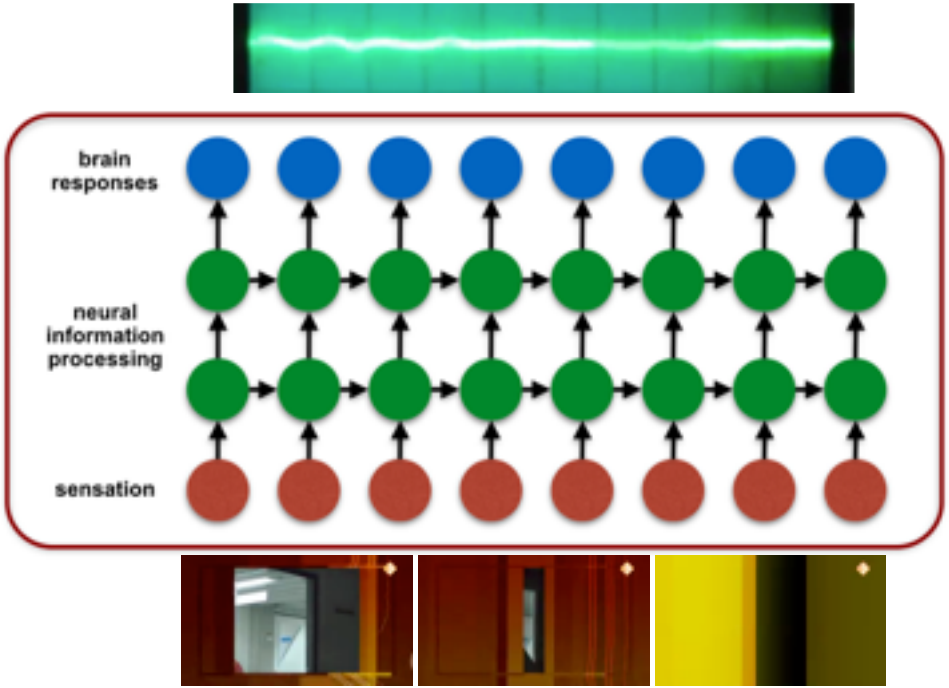
- Subjects presented with movies
- BOLD response measured along visual pathway
- S-R mapping modelled by RNN with two hidden layers

Using RNNs to model stimulus-response relationships



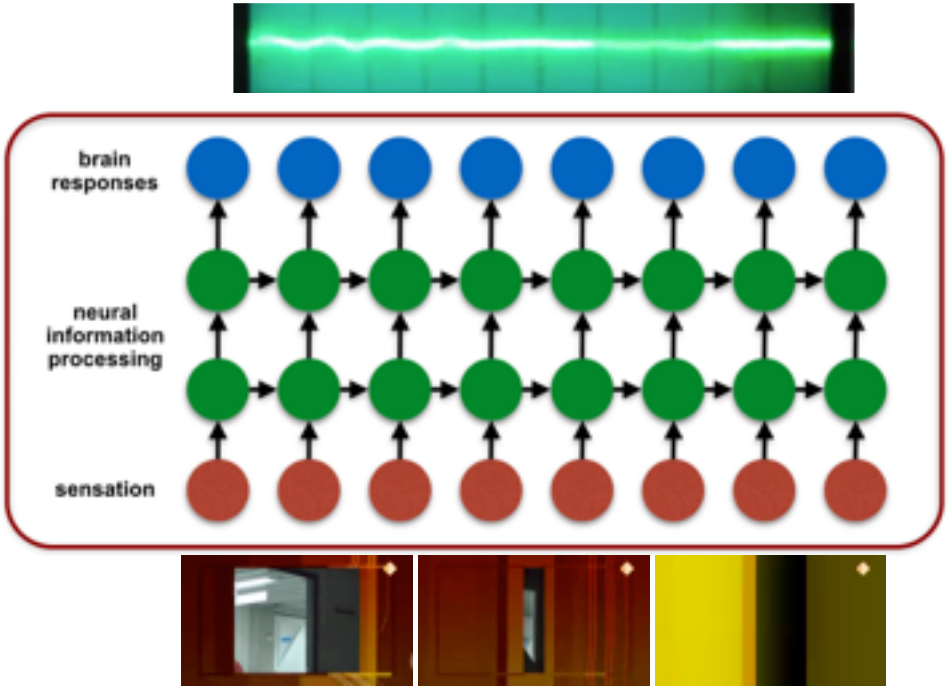
- Subjects presented with movies
- BOLD response measured along visual pathway
- S-R mapping modelled by RNN with two hidden layers
- Input was either:

Using RNNs to model stimulus-response relationships



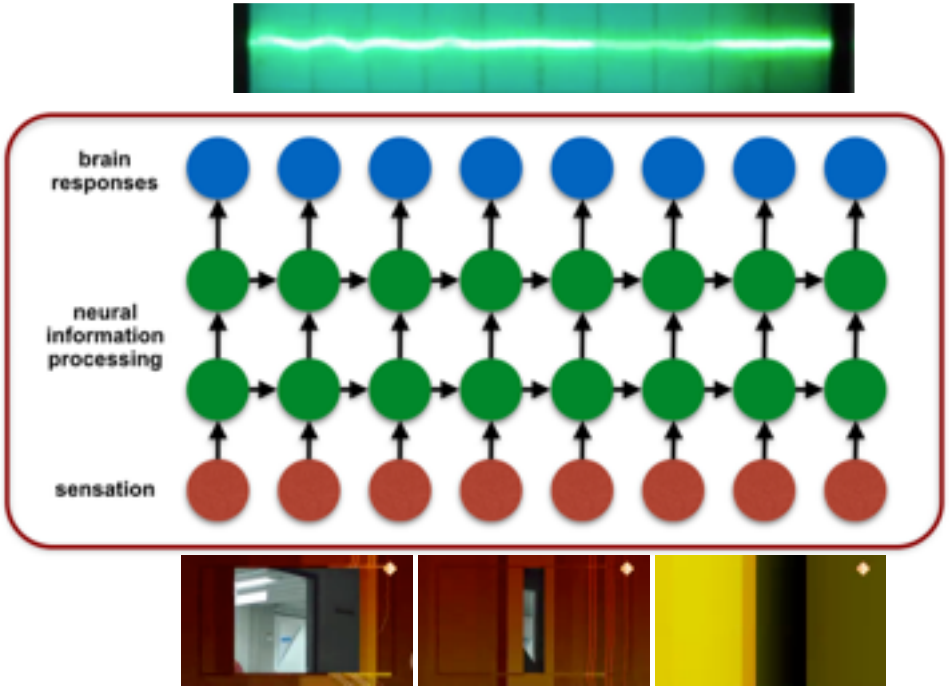
- Subjects presented with movies
- BOLD response measured along visual pathway
- S-R mapping modelled by RNN with two hidden layers
- Input was either:
    - low-level feature model (GIST)

Using RNNs to model stimulus-response relationships



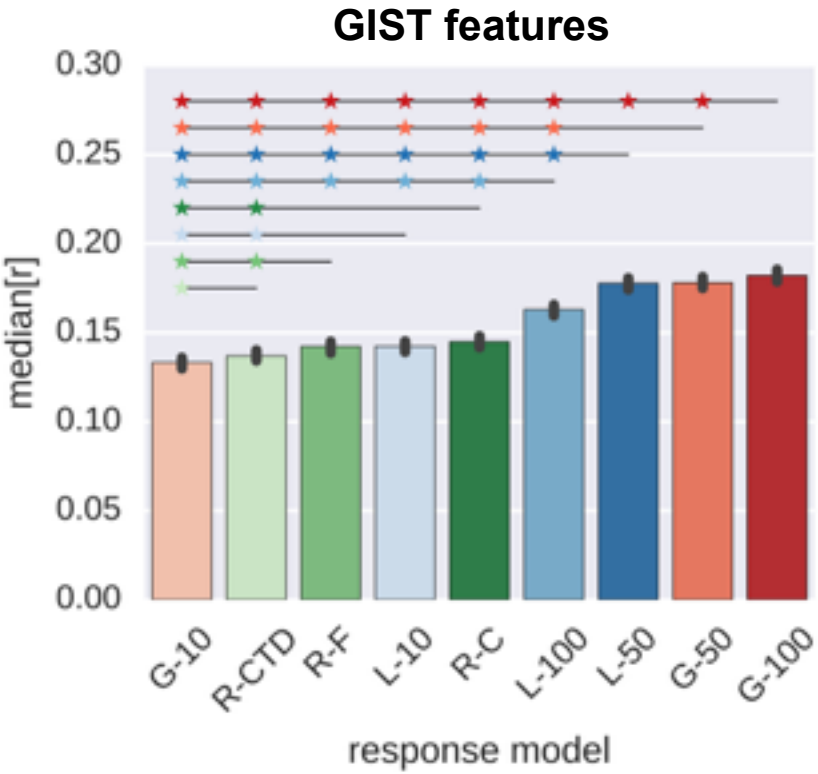
- Subjects presented with movies
- BOLD response measured along visual pathway
- S-R mapping modelled by RNN with two hidden layers
- Input was either:
  - low-level feature model (GIST)
  - high-level feature model (Word2Vec)

Using RNNs to model stimulus-response relationships

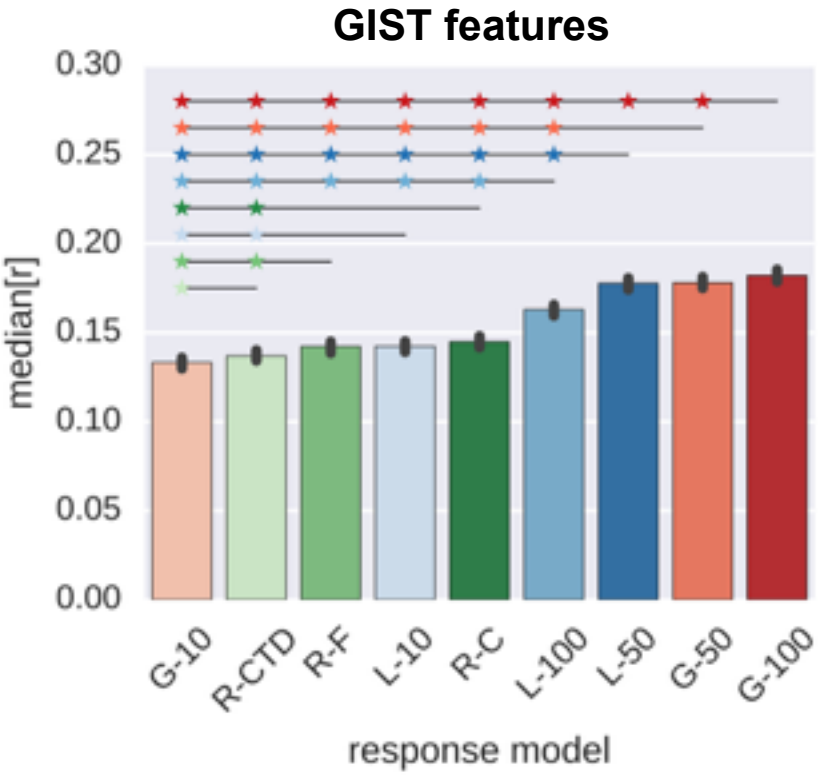


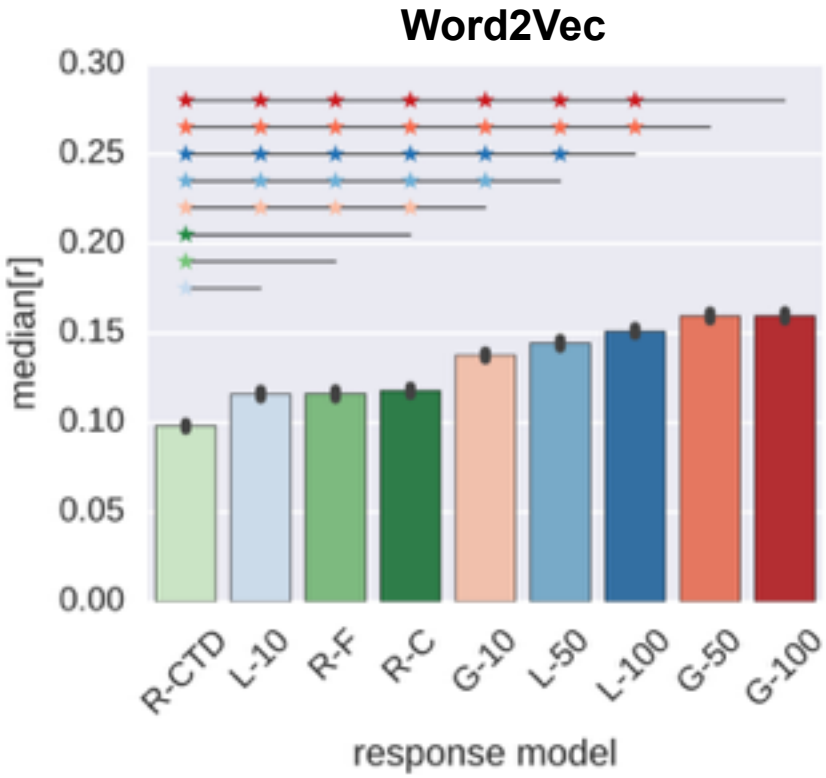
- Subjects presented with movies
- BOLD response measured along visual pathway
- S-R mapping modelled by RNN with two hidden layers
- Input was either:
  - low-level feature model (GIST)
  - high-level feature model (Word2Vec)
- Indirect coupling modelled via memory cells (LSTM vs. GRU)

Comparing RNNs with standard approaches

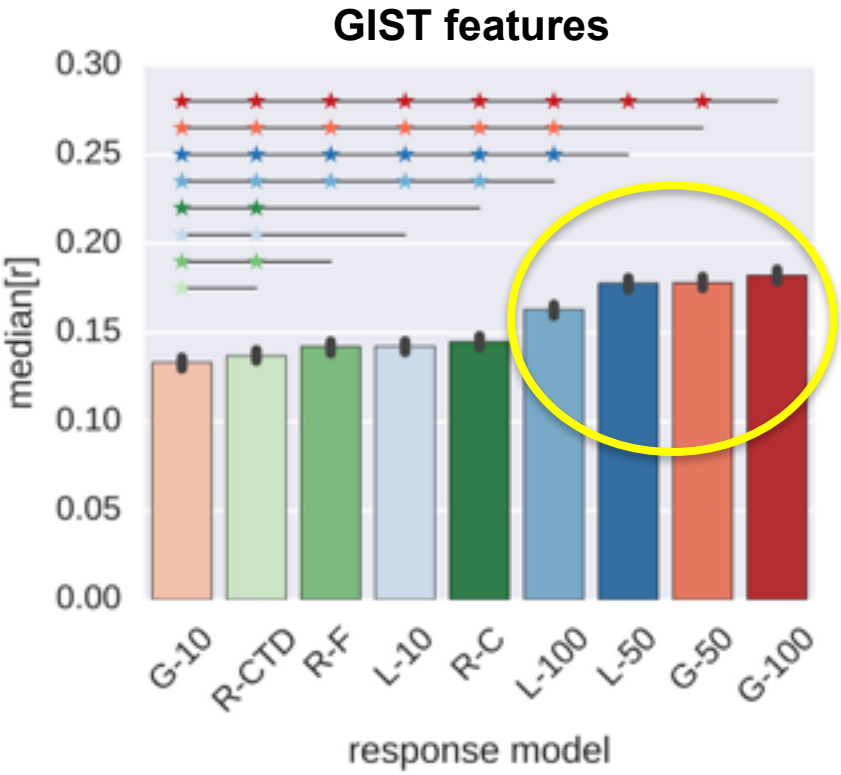


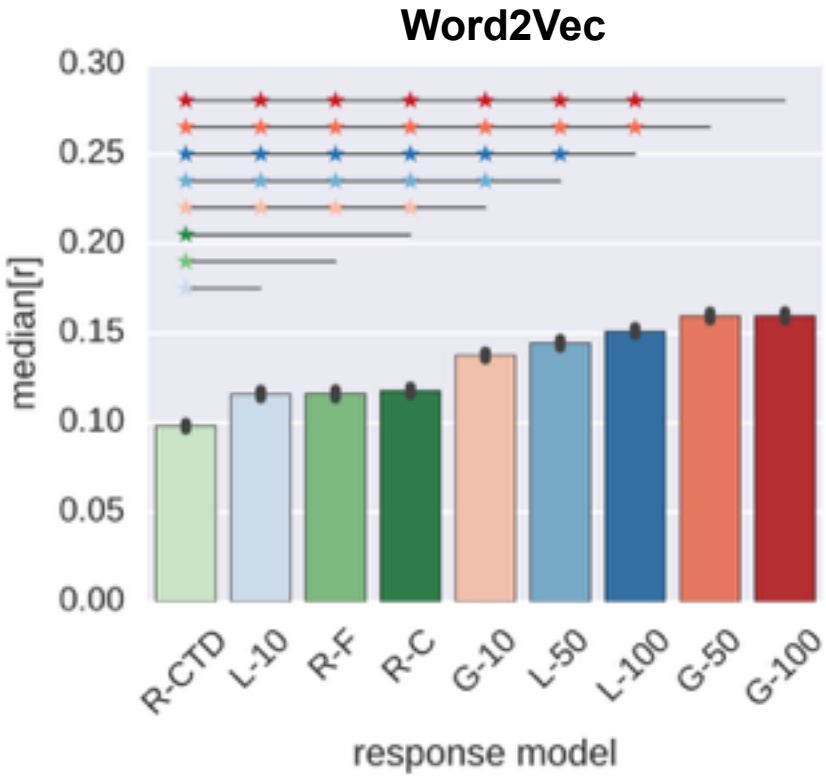
Comparing RNNs with standard approaches



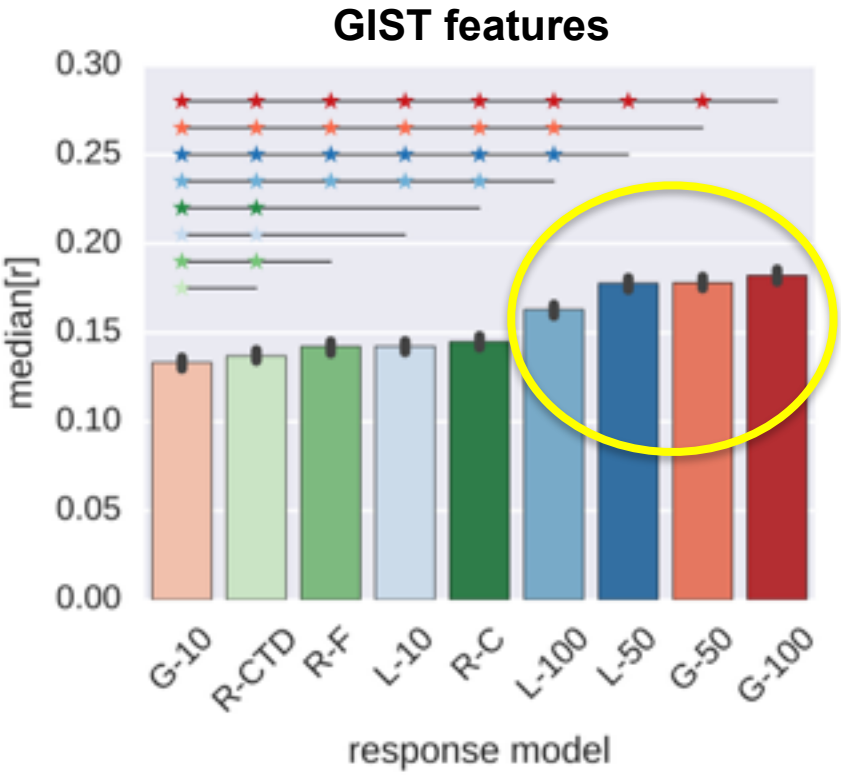


Comparing RNNs with standard approaches



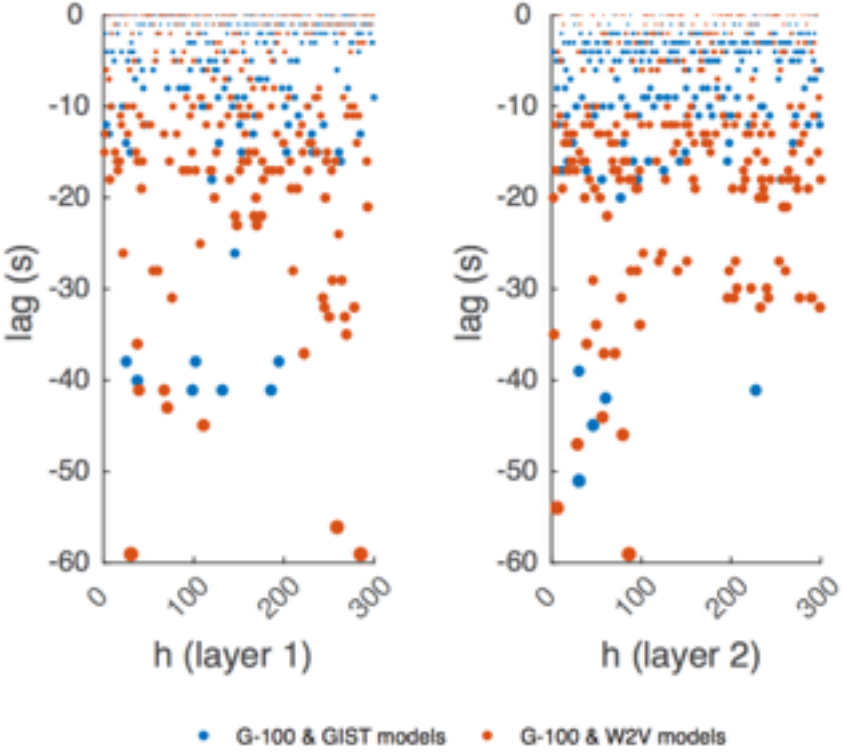


Comparing RNNs with standard approaches



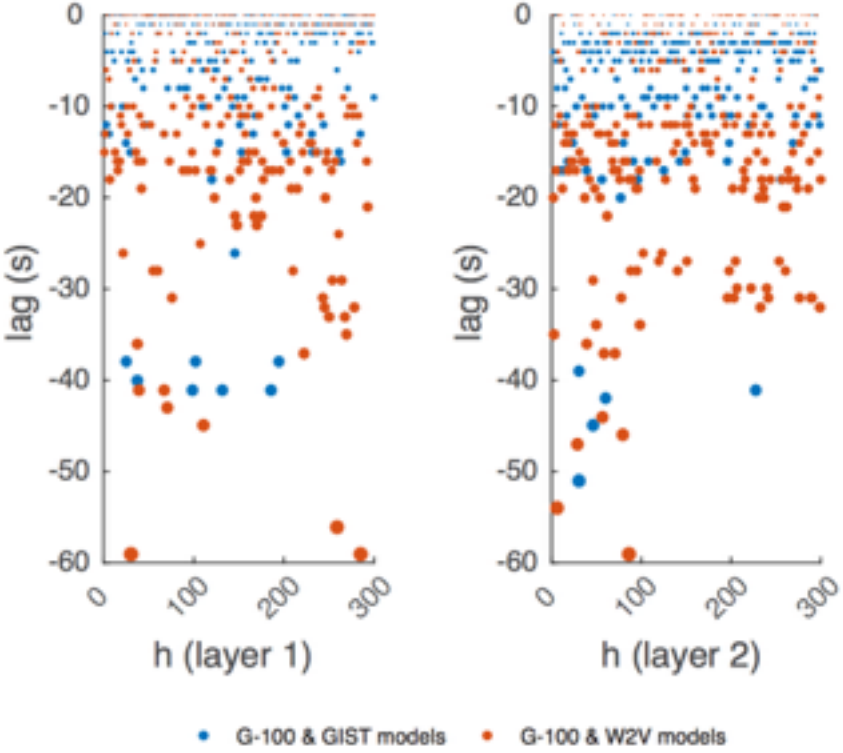
Interrogating RNNs

- Compute cross-correlations between input features and RNN hidden states



Interrogating RNNs

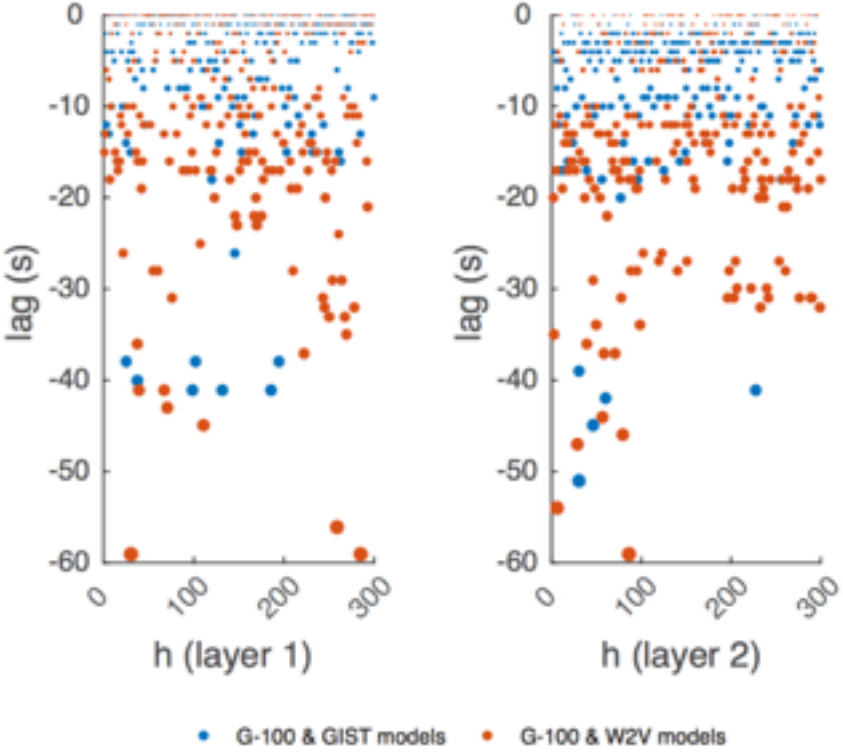
- Compute cross-correlations between input features and RNN hidden states



- Majority of the optimal lags were between 0 s and -10 s

Interrogating RNNs

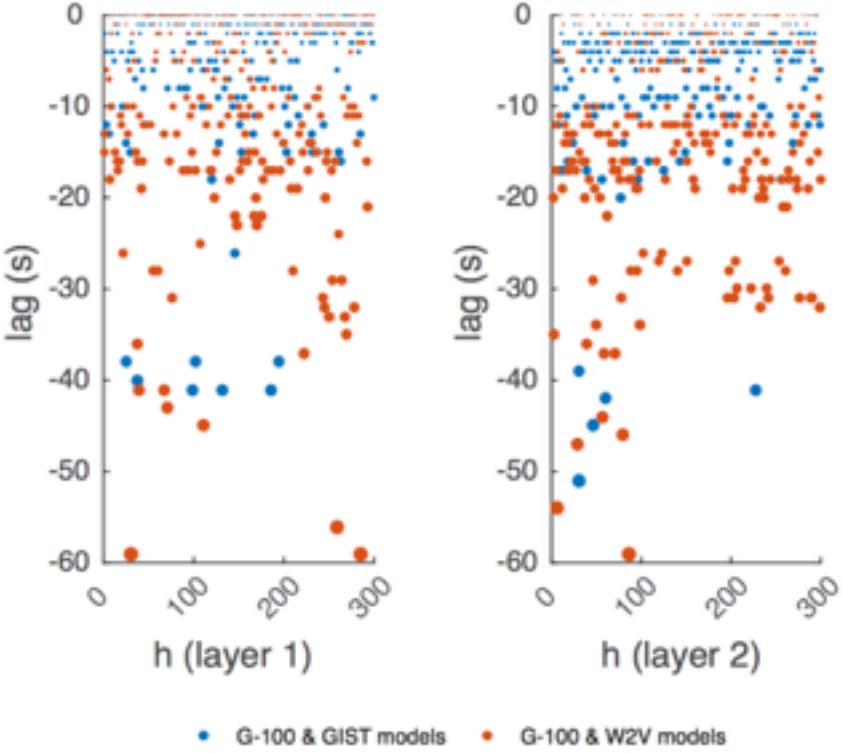
- Compute cross-correlations between input features and RNN hidden states



- Majority of the optimal lags were between 0 s and -10 s
- Substantial number of hidden units (in particular in layer 2) with optimal lags beyond this period

Interrogating RNNs

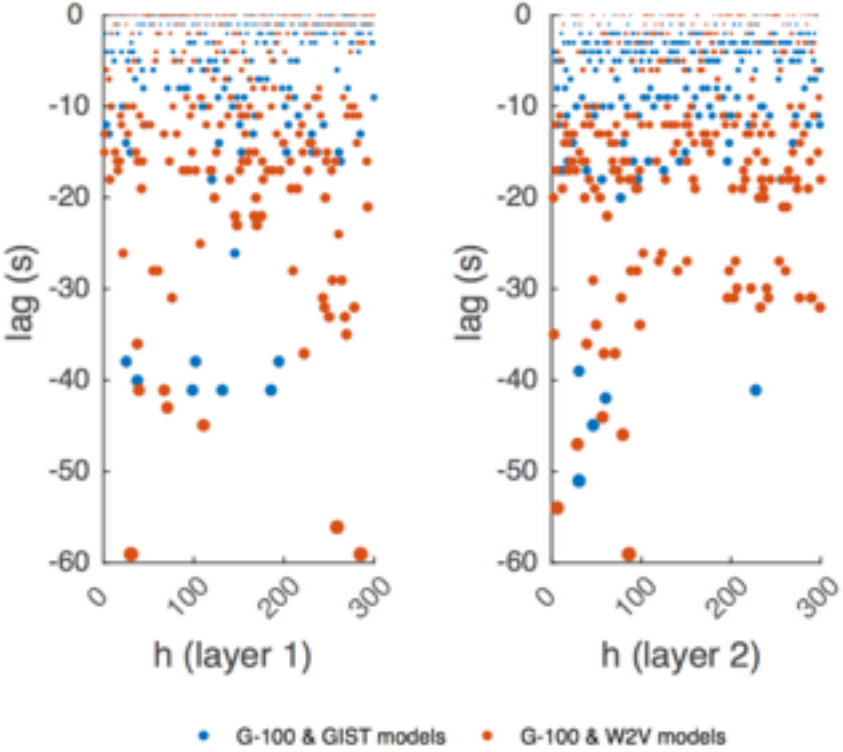
- Compute cross-correlations between input features and RNN hidden states



- Majority of the optimal lags were between 0 s and -10 s
- Substantial number of hidden units (in particular in layer 2) with optimal lags beyond this period
- Optimal lags of the hidden units in the W2V model significantly higher than those in the GIST model

Interrogating RNNs

- Compute cross-correlations between input features and RNN hidden states



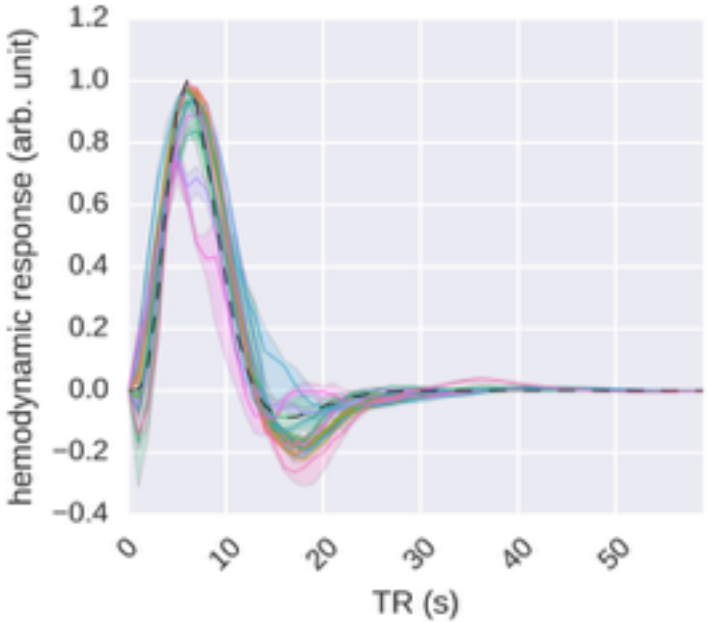
- Majority of the optimal lags were between 0 s and -10 s
- Substantial number of hidden units (in particular in layer 2) with optimal lags beyond this period
- Optimal lags of the hidden units in the W2V model significantly higher than those in the GIST model
- May reflect the fact that semantic states are more persistent than low-level visual features across the stimulus sequence

Interrogating RNNs

- Provide impulse as input to the RNN

Interrogating RNNs

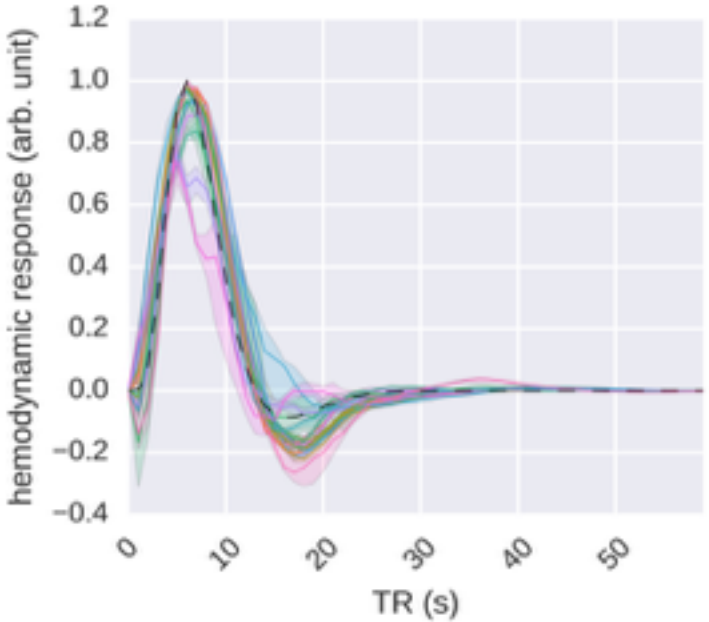
- Provide impulse as input to the RNN



- Results in a hemodynamic response at the output

Interrogating RNNs

- Provide impulse as input to the RNN



- Results in a hemodynamic response at the output
- Only model we know of which estimates HRFs from continuous designs

What ANNs tell us about neural representations

- Hierarchical organization as an essential ingredient
  - No-flattening theorems
  - Approximates the compositional structure of physical processes

What ANNs tell us about neural representations

- Hierarchical organization as an essential ingredient
  - No-flattening theorems
  - Approximates the compositional structure of physical processes
- Results suggest functional organization in terms of gradients rather than patches

What ANNs tell us about neural representations

- Hierarchical organization as an essential ingredient
  - No-flattening theorems
  - Approximates the compositional structure of physical processes
- Results suggest functional organization in terms of gradients rather than patches
- Recurrency essential to deal with perceptual aliasing

What ANNs tell us about neural representations

- Hierarchical organization as an essential ingredient
  - No-flattening theorems
  - Approximates the compositional structure of physical processes
- Results suggest functional organization in terms of gradients rather than patches
- Recurrency essential to deal with perceptual aliasing
- ANNs can learn to represent state uncertainty as in Bayesian modeling approaches

What ANNs tell us about neural representations

- Hierarchical organization as an essential ingredient
  - No-flattening theorems
  - Approximates the compositional structure of physical processes
- Results suggest functional organization in terms of gradients rather than patches
- Recurrency essential to deal with perceptual aliasing
- ANNs can learn to represent state uncertainty as in Bayesian modeling approaches
- ANNs can be validated by (directly?) fitting them to behavioural and/or neural data

What ANNs tell us about neural representations

- Hierarchical organization as an essential ingredient
  - No-flattening theorems
  - Approximates the compositional structure of physical processes
- Results suggest functional organization in terms of gradients rather than patches
- Recurrency essential to deal with perceptual aliasing
- ANNs can learn to represent state uncertainty as in Bayesian modeling approaches
- ANNs can be validated by (directly?) fitting them to behavioural and/or neural data
- General approach to model-based cognitive neuroscience

Summary

- ANNs provide an elegant and powerful framework for modeling cognitive processes:

Summary

- ANNs provide an elegant and powerful framework for modeling cognitive processes:
  - ANNs trained for object and action recognition model ventral/dorsal stream responses

Summary

- ANNs provide an elegant and powerful framework for modeling cognitive processes:
  - ANNs trained for object and action recognition model ventral/dorsal stream responses
  - ANNs trained to recognize music reveal functional differentiation across STG

Summary

- ANNs provide an elegant and powerful framework for modeling cognitive processes:
  - ANNs trained for object and action recognition model ventral/dorsal stream responses
  - ANNs trained to recognize music reveal functional differentiation across STG
  - ANNS that are trained to play games may reveal neural correlates of reward estimation and decision making

Further reading

Güçlü, U., van Gerven, M.A.J., 2015. Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. J. Neurosci. 35, 10005–10014.

LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature 521, 436–444.

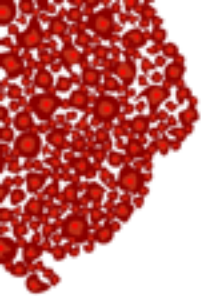
Kriegeskorte, N., 2015. Deep neural networks: A new framework for modeling biological vision and brain information processing. Annu. Rev. Vis. Sci. 1, 417–446.

Yamins, D.L.K., Dicarlo, J.J., 2016. Using goal-driven deep learning models to understand sensory cortex. doi:10.1038/nn.4244

Horikawa, T., Kamitani, Y., 2016. Generic decoding of seen and imagined objects using hierarchical visual features. ArXiv Prepr.

Cichy, R.M., Khosla, A., Pantazis, D., Torralba, A., Oliva, A., 2016. Deep neural networks predict hierarchical spatio-temporal cortical dynamics of human visual object recognition. arXiv Prepr.

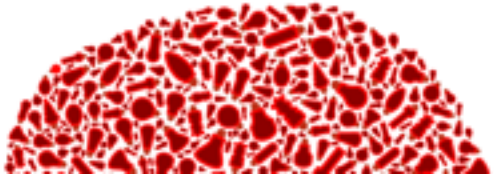
Yuste, R., 2015. From the neuron doctrine to neural networks. Nat. Rev. Neurosci. 16, 487–497. doi:10.1038/nrn3962



DONDERS INSTITUTE

www.ru.nl/donders

www.ccnlab.net



Radboud University Radboudumc