Perception: Psychophysics and Modeling

10 | Object recognition IV

Felix Wichmann

Neural Information Processing Group
Eberhard Karls Universität Tübingen

Overview

The Problems of Perceiving and Recognising Objects **(VL07-Object Recognition 1)**

Mid-level vision **(VL07-Object Recognition 1)**

- What are ‘‘edges’’ and (illusionary) ‘‘contours’’?
- Gestalt psychology and ‘‘Gestalt laws’’ of perceptual organisation

More on *mid-level vision* **(VL08-Object Recognition 2)**

- Accidental viewpoint and non-accidental features
- Figure-ground, occlusion, wholes and parts
- Texture segmentation, grouping and camouflage

Neuroscience of object recognition **(VL08-Object Recognition 2)**

Object representation **(VL09-Object Recognition 3)**

- Structural description models
- View-based models

Object recognition by algorithms: DNNs **(VL10-Object Recognition 4)**

Overview

The Problems of Perceiving and Recognising Objects **(VL07-Object Recognition 1)**

Mid-level vision **(VL07-Object Recognition 1)**

- What are ‘‘edges’’ and (illusionary) ‘‘contours’’?
- Gestalt psychology and ‘‘Gestalt laws’’ of perceptual organisation

More on *mid-level vision* **(VL08-Object Recognition 2)**

- Accidental viewpoint and non-accidental features
- Figure-ground, occlusion, wholes and parts
- Texture segmentation, grouping and camouflage

Neuroscience of object recognition **(VL08-Object Recognition 2)**

Object representation **(VL09-Object Recognition 3)**

- Structural description models
- View-based models

Object recognition by algorithms: DNNs **(VL10-Object Recognition 4)**

Literature

Geirhos, R., Temme, C. R. M., Rauber, J., Schu¨tt, H. H., Bethge, M., and Wichmann, F. A. (2018). Generalisation in humans and deep neural networks. *Advances in Neural Information Processing Systems (NeurIPS)*, 31:7549–7561.

Yamins, D. L., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., and DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, 111(23):8619–8624.

Supplementary Literature

Cox, D. D. (2014). Do we understand high-level vision? *Current Opinion in Neurobiology*, 25:187–193.

Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., and Brendel, W. (2019). ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. *International Conference on Learning Representations* (ICLR).

Geirhos, R., Meding, K., and Wichmann, F. A. (2020). Beyond accuracy: quantifying trial-by-trial behaviour of CNNs and humans by measuring error consistency. *Advances in Neural Information Processing Systems* (NeurIPS), 33.

Kriegeskorte, N. (2015). Deep neural networks: A new framework for modeling biological vision and brain information processing. *Annual Review of Vision Science*, 1:417–446.

Krizhevsky, A., Sutskever, I. I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Neural Information Processing Systems (NeurIPS)*, 25:1097–1105.

Yamins, D. L. K., & DiCarlo, J. J. (2016). Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience*, 19(3), 356–365.

What changed the world in 2012?

The Mars Science Laboratory or "Curiosity Rover" successfully lands on Mars.

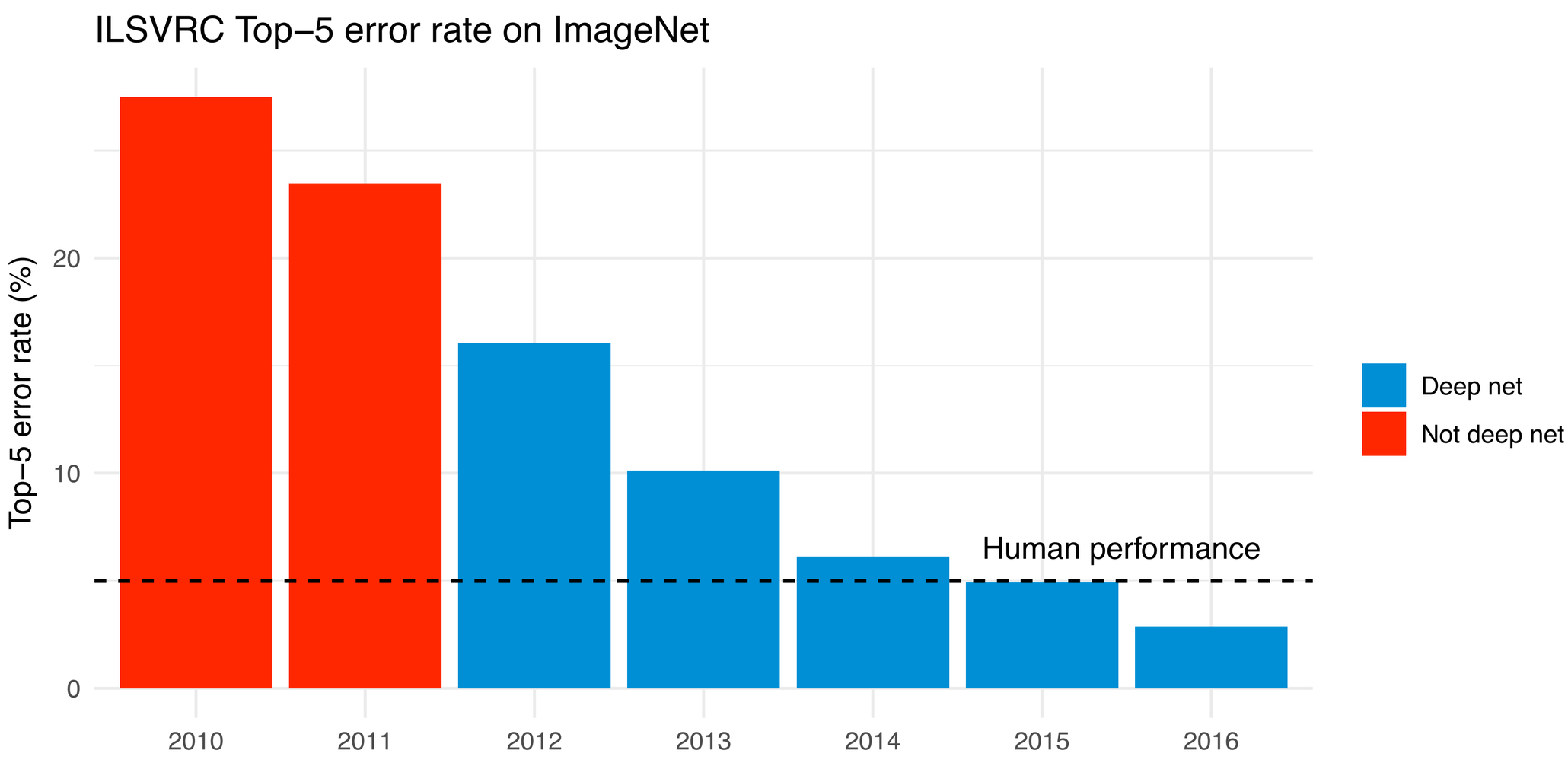
The Summer Olympics are held in London, England from July 27th to August 12th. The top three medal winning countries during the games were the United States, China and Great Britain. US Swimmer Michael Phelps became the most decorated Olympian of all time after bringing his medal count to twenty-two at these games.

The film "Marvel's The Avengers" is released and becomes one of the highest-grossing films.

What changed the world in 2012?

ImageNet challenge: 1000 categories, 1.2 million training images.

AlexNet by Krizhevsky, Sutskever & Hinton (2012) appears on the stage, and basically reduces the prediction error by 50%:



Fundamentals of Neural Networks

Interest in shallow, single-layer artificial neural networks (ANN)—so-called **perceptrons**—began in the late 1950s and early 60s (Frank Rosenblatt), based on Warren McCulloch and Walter Pitts's as well Donald Hebb's ideas of computation by neurons from the 1940s.

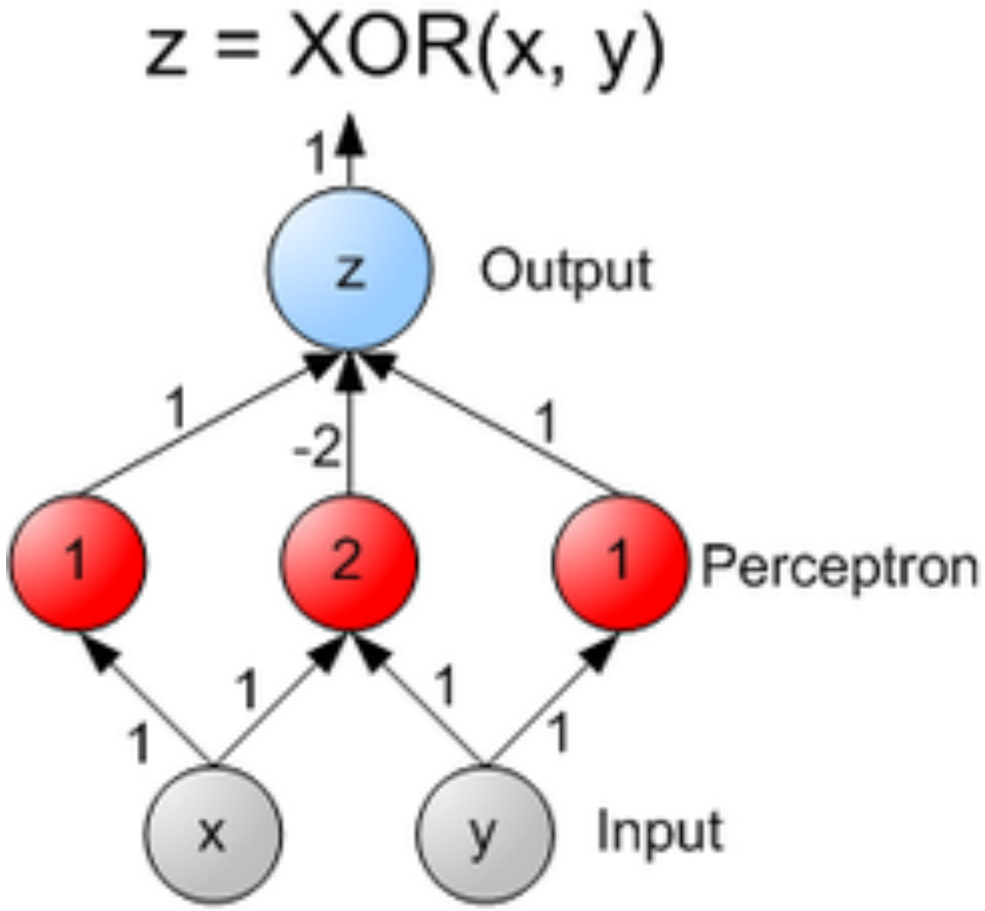
Single-layer perceptrons computed outputs from a simple weighted combination of inputs.

Marvin Minsky & Seymour Papert published *Perceptrons: an Introduction to Computational Geometry* in 1969, basically stopping all perceptron research (''the XOR problem'' or ''the XOR affair'').

Fundamentals of Neural Networks

Second wave of ANN research and interest in psychology—often termed **connectionism**—after the publication of the **parallel distributed processing** (PDP) books by David Rumelhart and James McClelland (1986), introducing the backpropagation algorithm as a learning rule for multi-layer networks to a large audience (mainly three-layer networks).

Three-layer network with (potentially infinitely many) hidden units in the intermediate layer is a universal function approximator: Can compute anything that is computable (Turing equivalent).



Fundamentals of Neural Networks

Lack of theory, non-convex optimization problems during backpropagation training, and lack of computing power limited the usefulness of the ANNs.

Universal function approximator in theory, but could not in general be successfully trained on complex problems in practice.

Thus in the mid-1990s the approach grew unpopular again, and was superseded by kernel methods, most notably the support vector machine (SVM) as the classification algorithm of choice in machine learning (ML).

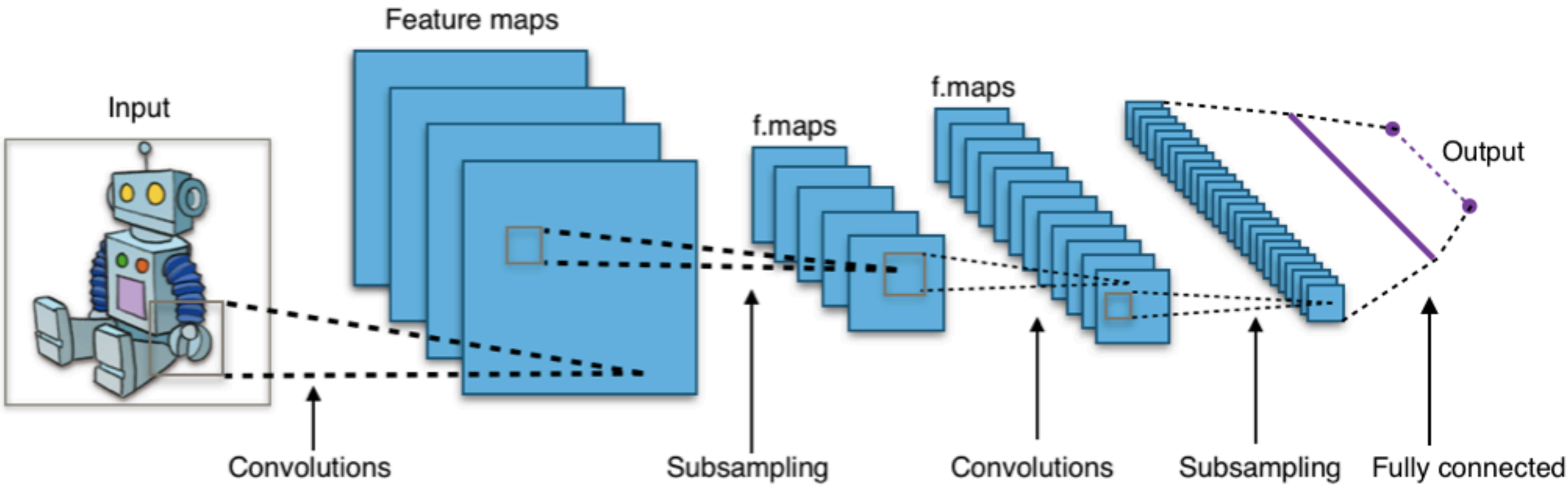
Breakthrough again in the mid-2000s with so-called **deep neural networks** or **DNN**s, widely known since the 2012 NIPS-paper by Alex Krizhevsky.

Fundamentals of Neural Networks

DNN: loose terminology to refer to networks with at least two hidden or intermediate layers, typically at least five to ten or twenty.

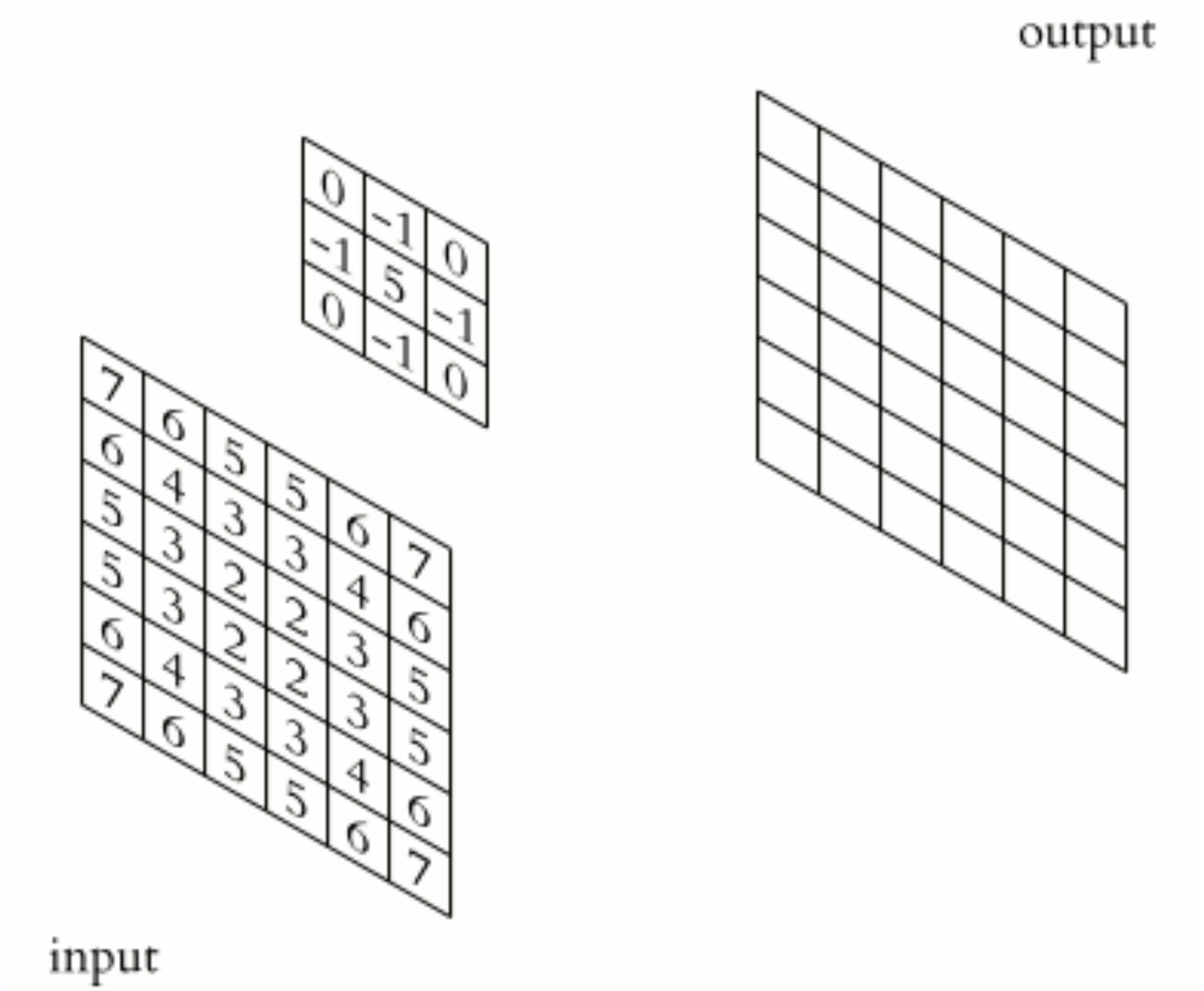
Massive increase in labelled training data (“the internet”), computing power (GPUs), and tricks with simple non-linearities (ReLU) and convolutional rather than fully connected layers make them the current method of choice in ML.

Fundamentals of Neural Networks



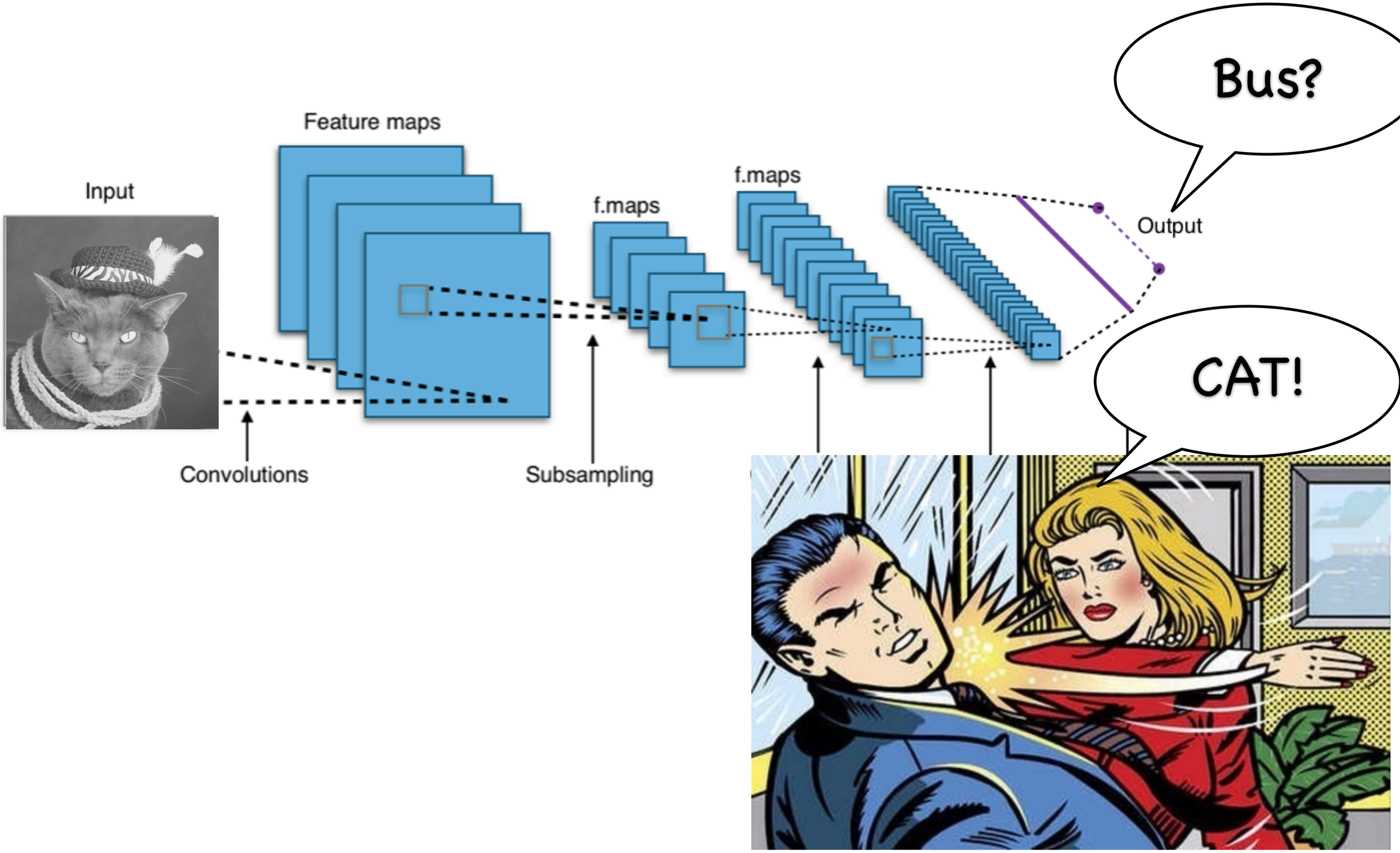
Source: Wikipedia (de)

Fundamentals of Neural Networks



Source: Wikipedia (de)

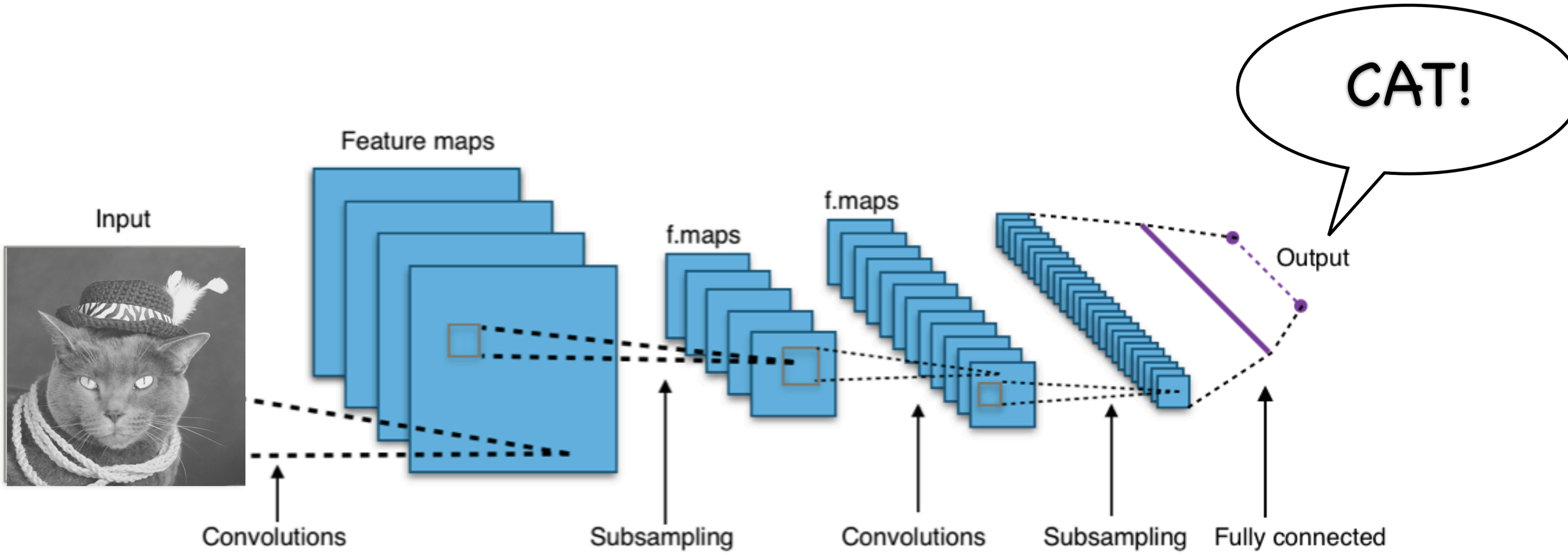
Fundamentals of Neural Networks



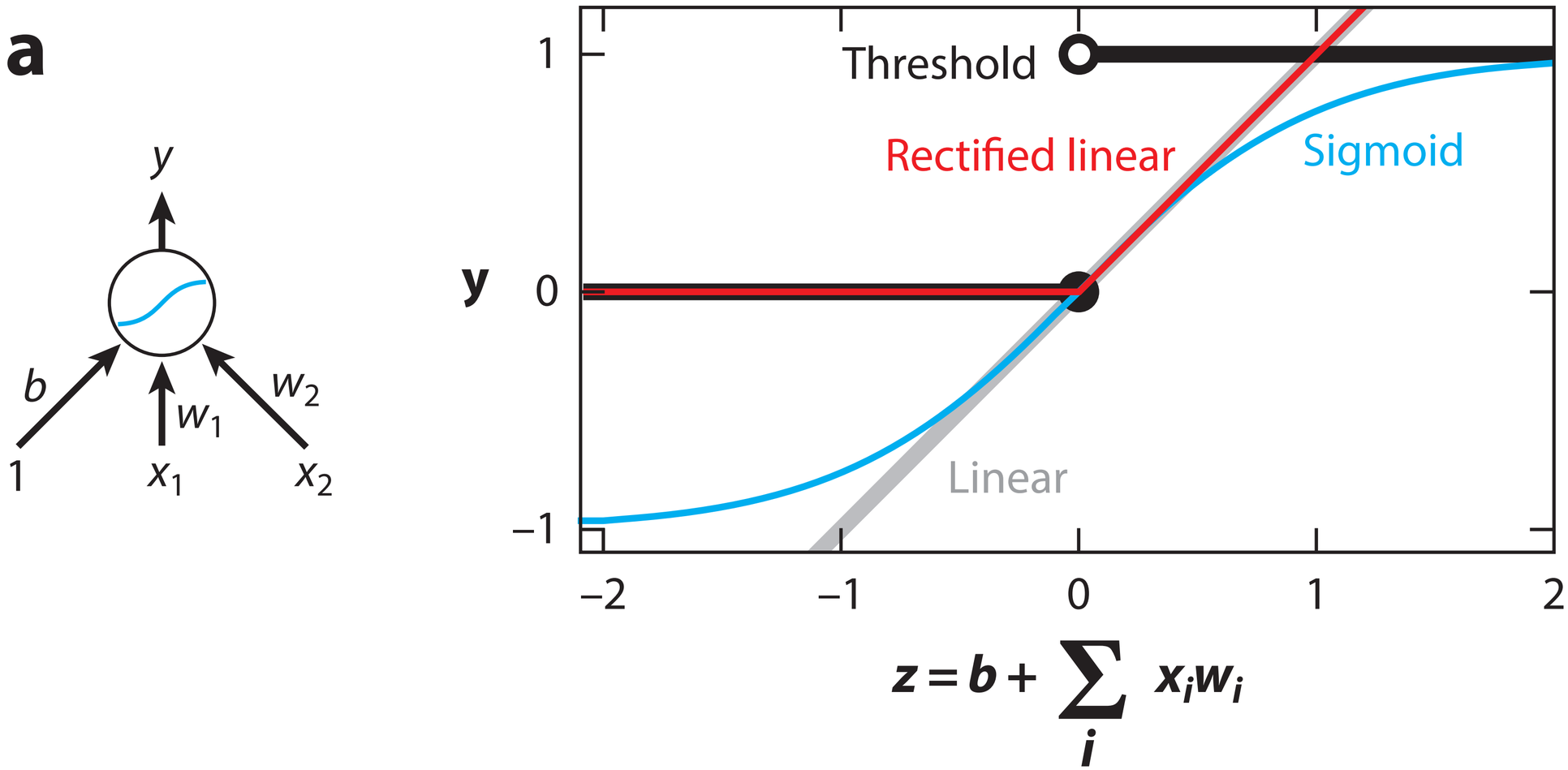
Fundamentals of Neural Networks

Millions of training images later...

Fundamentals of Neural Networks



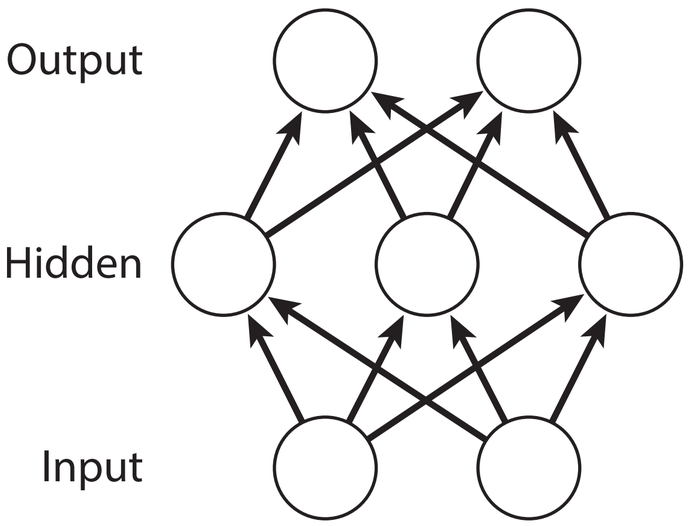
Fundamentals of Neural Networks



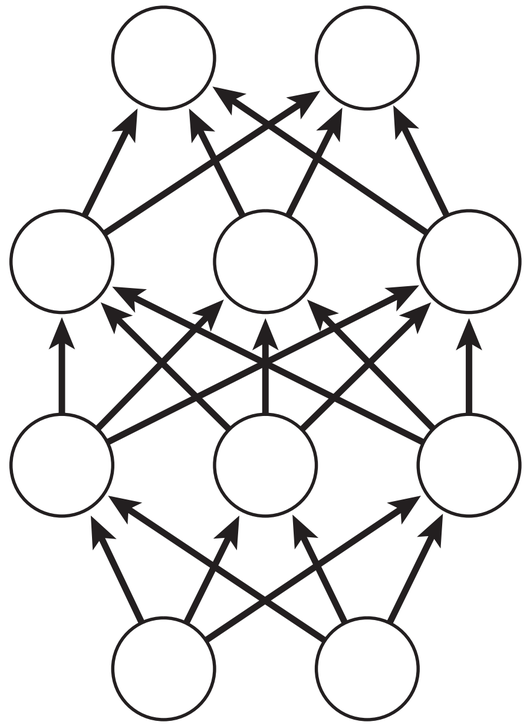
Kriegeskorte (2015)

Fundamentals of Neural Networks

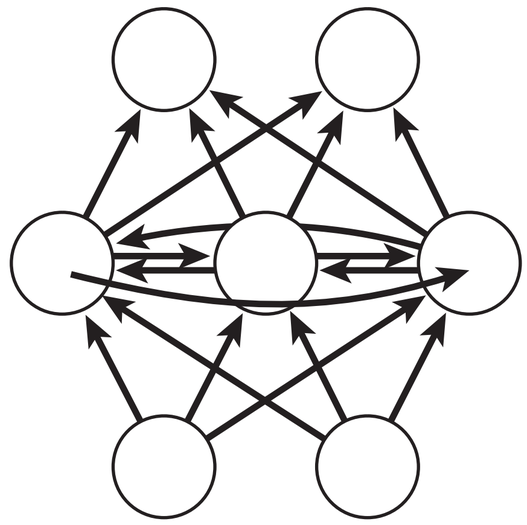
b **Shallow feedforward**
(1 hidden layer)



c **Deep feedforward**
(>1 hidden layer)

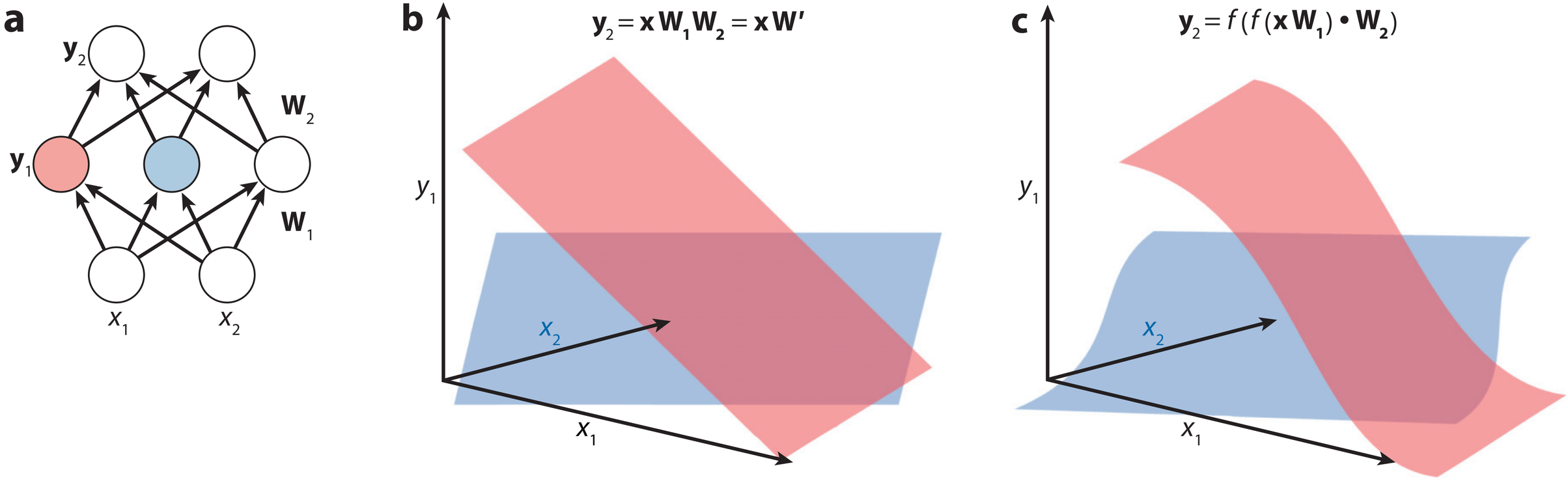


d **Recurrent**



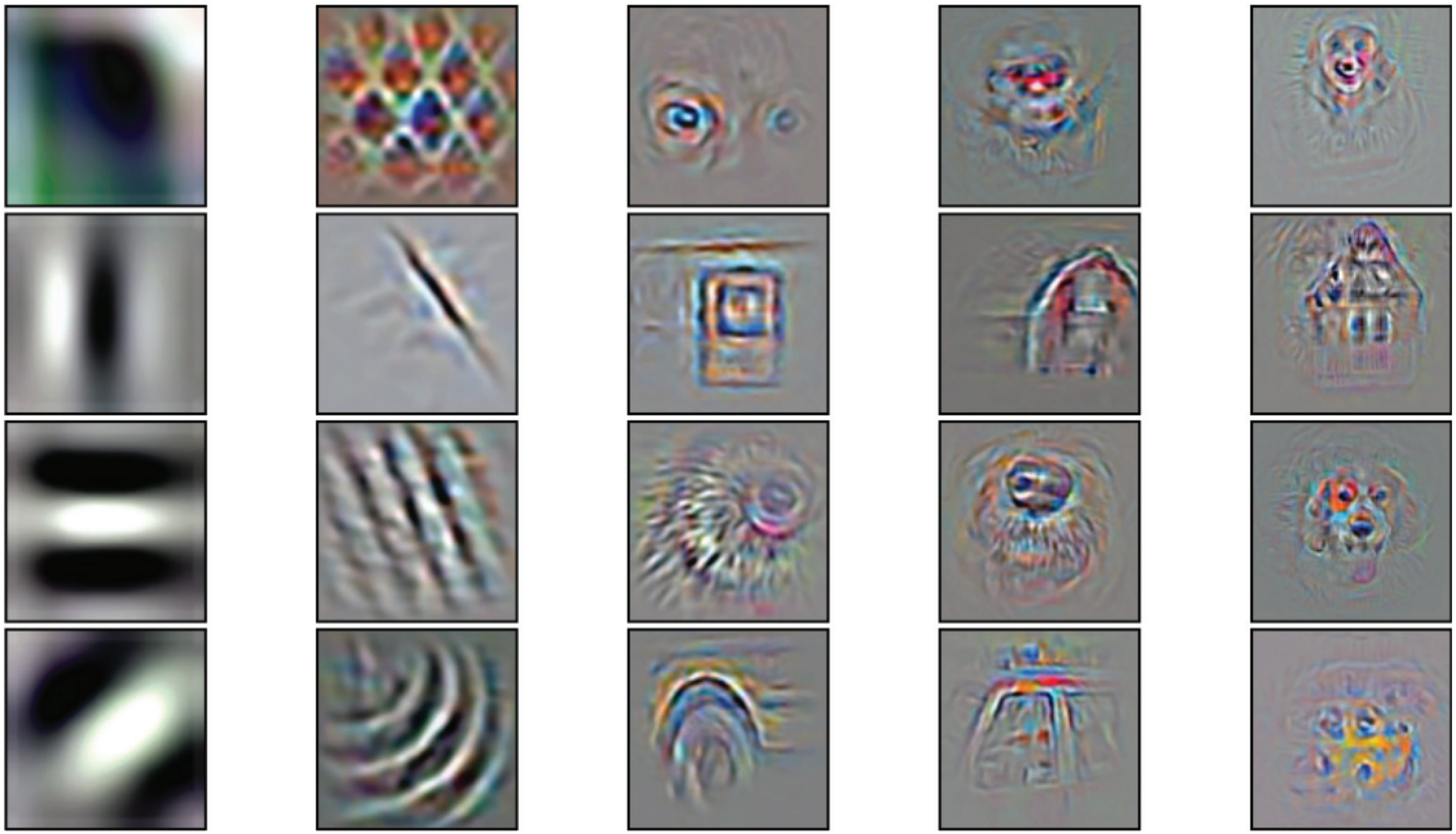
Kriegeskorte (2015)

Fundamentals of Neural Networks



Kriegeskorte (2015)

Fundamentals of Neural Networks



Kriegeskorte (2015)

Performance-optimized hierarchical models predict neural responses in higher visual cortex

Daniel L. K. Yamins[a,1], Ha Hong[a,b,1], Charles F. Cadieu[a], Ethan A. Solomon[a], Darren Seibert[a], and James J. DiCarlo[a,2]

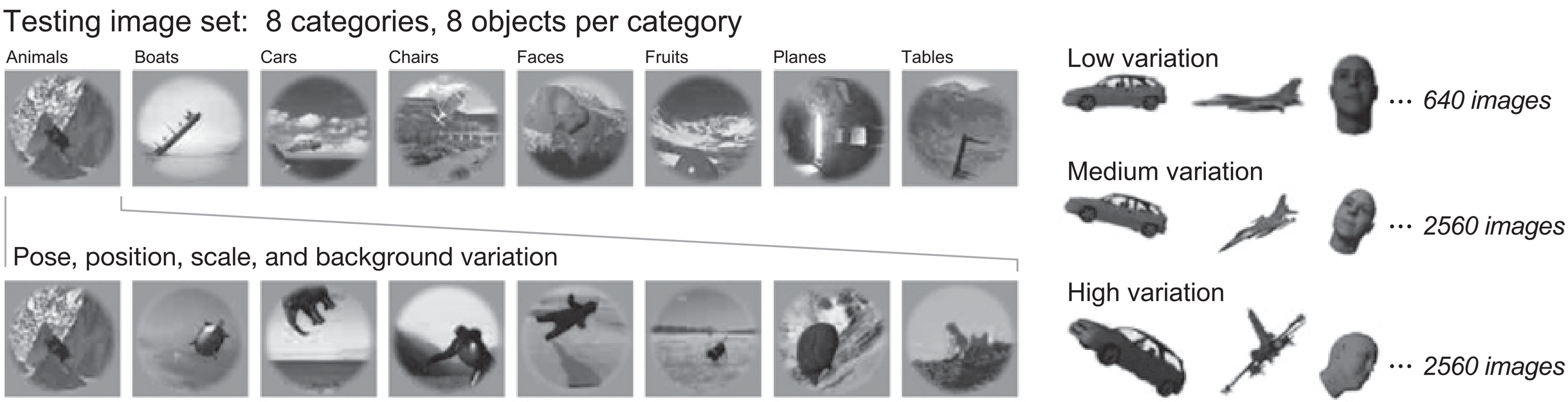
[a]Department of Brain and Cognitive Sciences and McGovern Institute for Brain Research, Massachusetts Institute of Technology, Cambridge, MA 02139; and [b]Harvard-MIT Division of Health Sciences and Technology, Institute for Medical Engineering and Science, Massachusetts Institute of Technology, Cambridge, MA 02139

Edited by Terrence J. Sejnowski, Salk Institute for Biological Studies, La Jolla, CA, and approved April 8, 2014 (received for review March 3, 2014)

The ventral visual stream underlies key human visual object recognition abilities. However, neural encoding in the higher areas of the ventral stream remains poorly understood. Here, we describe a modeling approach that yields a quantitatively accurate model of inferior temporal (IT) cortex, the highest ventral cortical area. Using high-throughput computational techniques, we discovered that, within a class of biologically plausible hierarchical neural network models, there is a strong correlation between a model's categorization performance and its ability to predict individual IT neural unit response data. To pursue this idea, we then identified a high-performing neural network that matches human performance on a range of recognition tasks. Critically, even though we did not

Explaining the neural encoding in these higher ventral areas thus remains a fundamental open question in systems neuroscience.

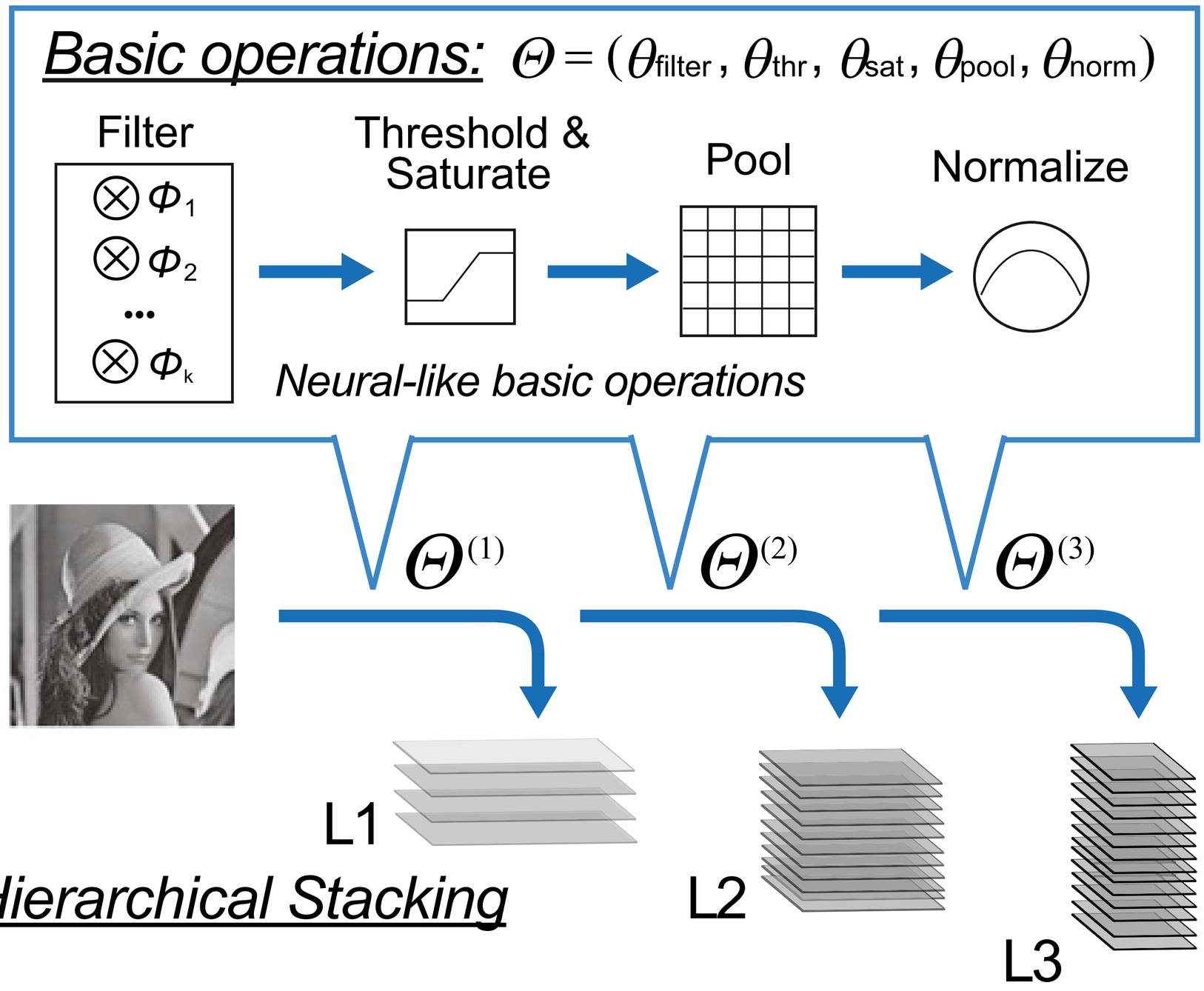
As with V1, models of higher ventral areas should be neurally predictive. However, because the higher ventral stream is also believed to underlie sophisticated behavioral object recognition capacities, models must also match IT on performance metrics, equalling (or exceeding) the decoding capacity of IT neurons on object recognition tasks. A model with perfect neural predictivity in IT will necessarily exhibit high performance, because IT itself does. Here we demonstrate that the converse is also true, within a biologically appropriate model class. Combining high-throughput computational and electrophysiology techniques, we explore a wide range of biologically plausible hierarchical neural network models



Note: Nonsense scenes with floating objects …
what if object recognition were generative in nature?

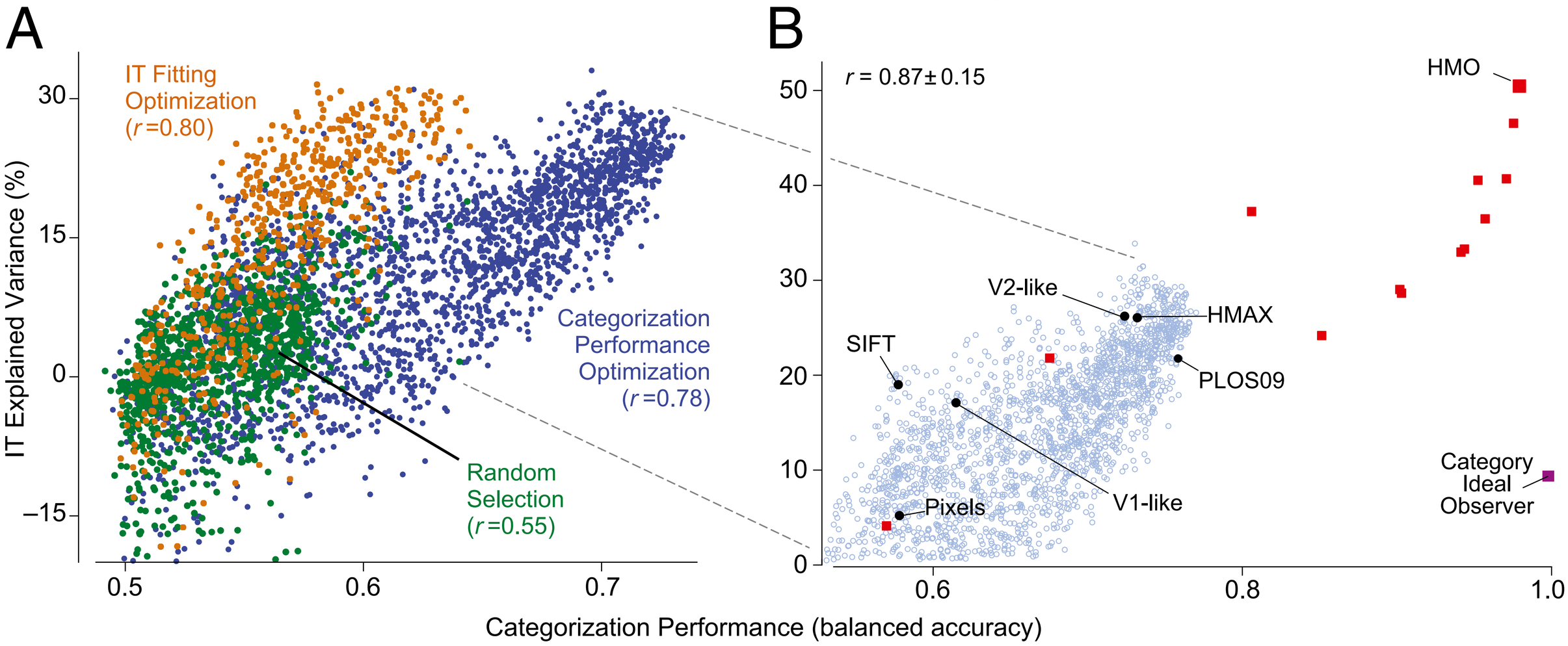
Yamins et al. (2014)

a



Yamins et al. (2014)

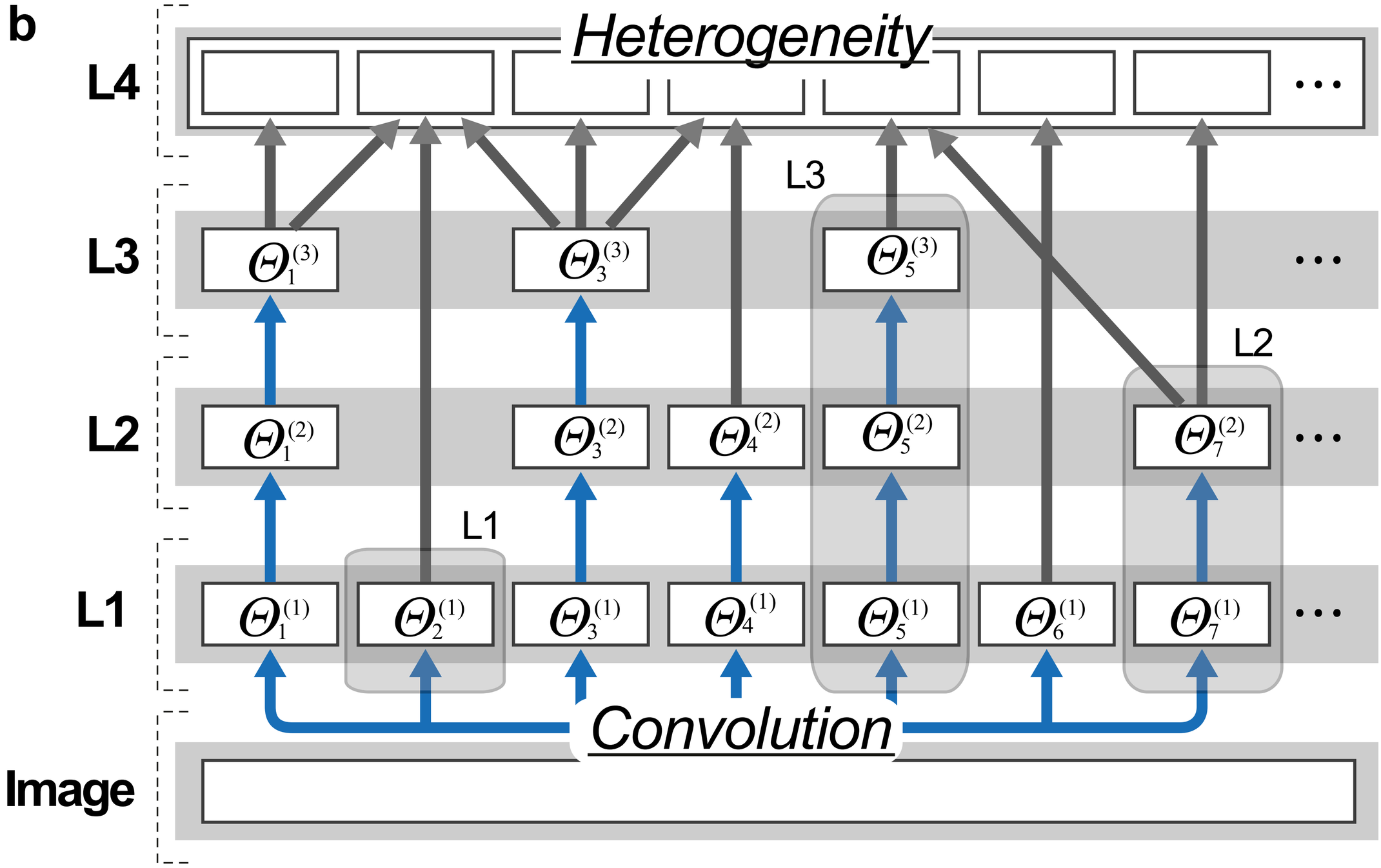
hierarchical modular optimisation (HMO)



Every point in the figures corresponds to a new network architecture, parametrised by 57 parameters. (Total of nearly 6.000 architectures!)

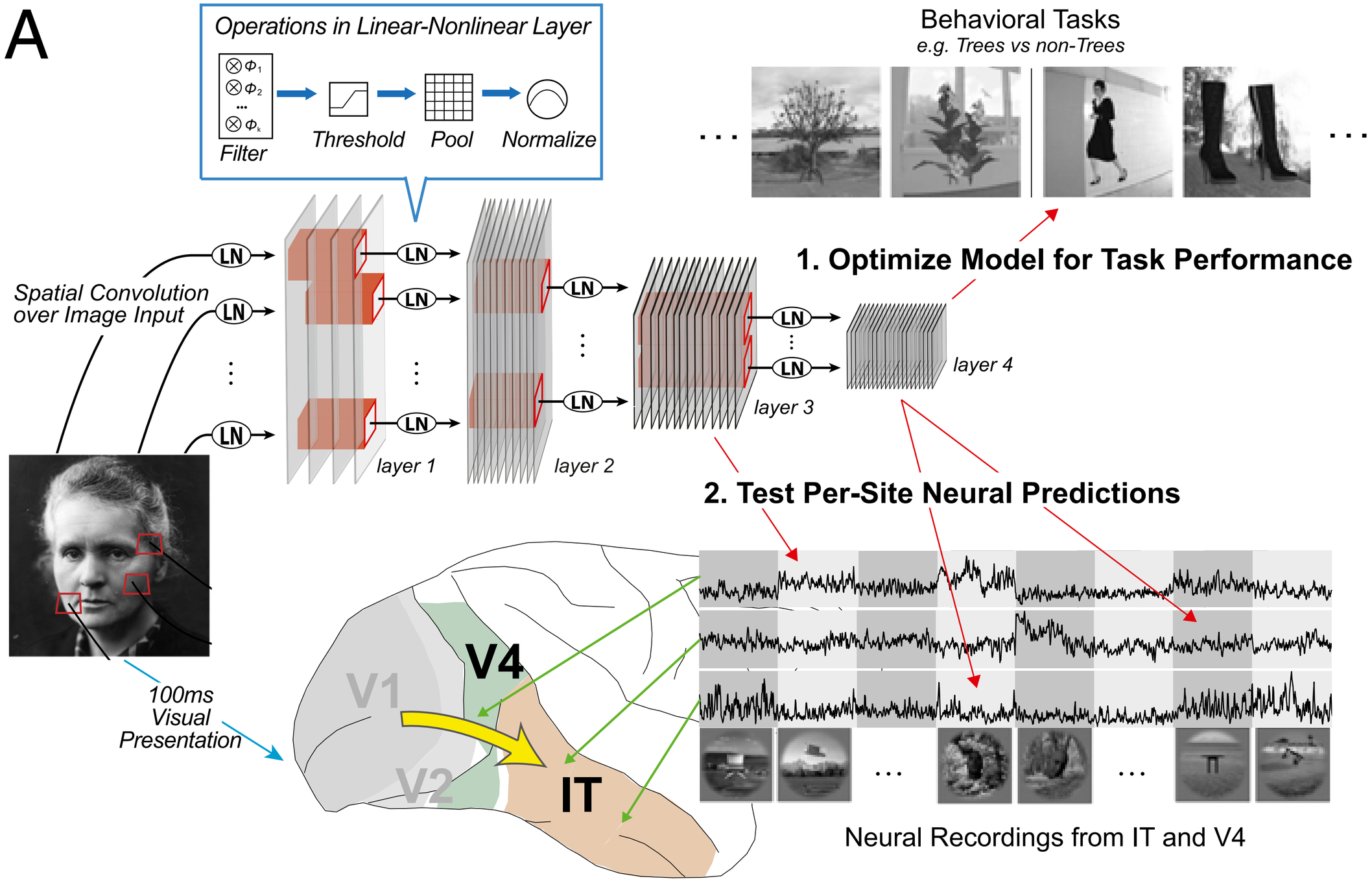
Yamins et al. (2014)

b



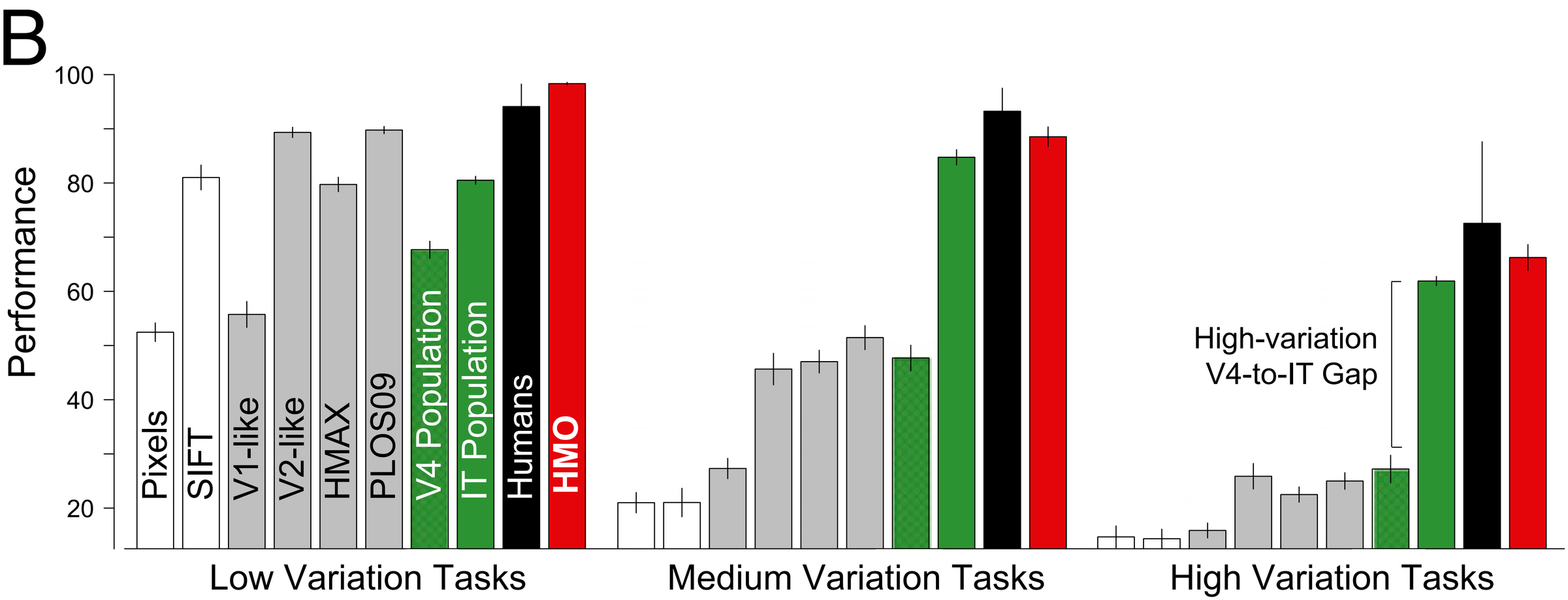
Yamins et al. (2014)

A



Yamins et al. (2014)

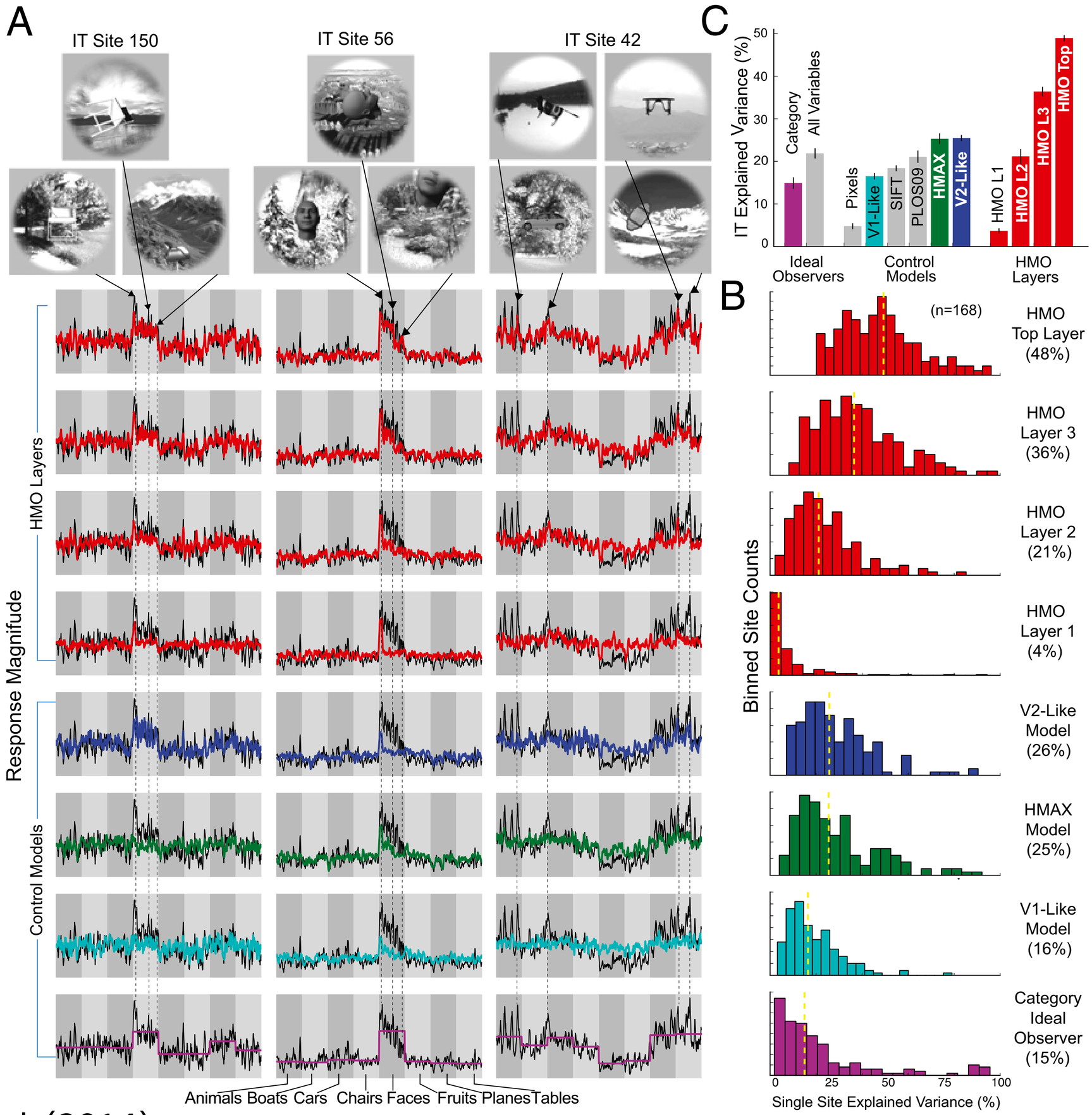
B



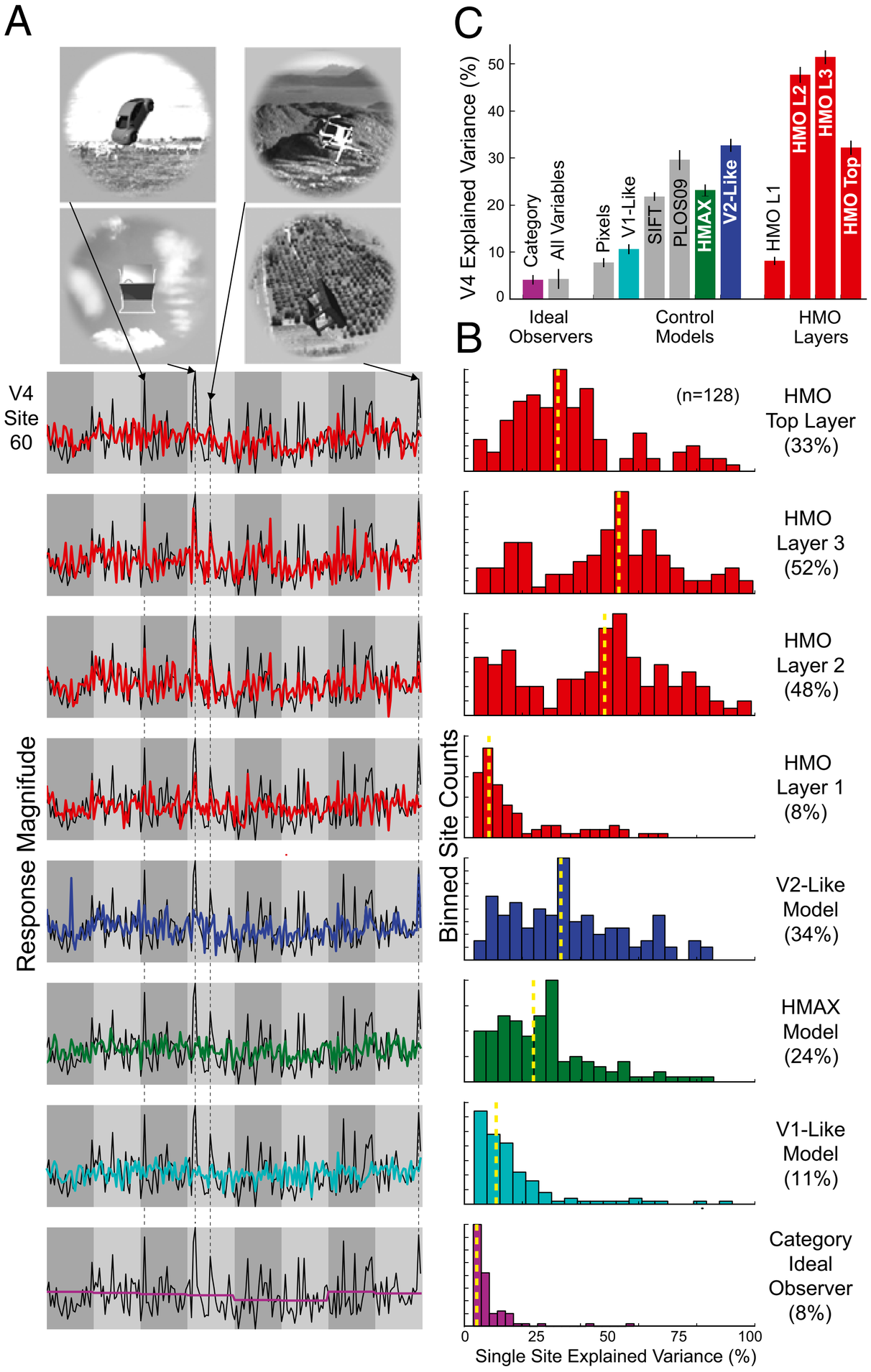
Notes:

1. Classification performance via linear SVM trained on 1.259 top-layer units of HMO; new training (set of weights) for every object category and level of difficulty (low, medium, high).

Yamins et al. (2014)



Yamins et al. (2014)



Yamins et al. (2014)

Summary of DNN's (1/2)

Yamins et al. (2014) present a DNN model that can recognise objects under strong background, pose & illumination changes rather well (70% correct in high variance condition).

The model's computations are locally build on computations believed to be those of visual cortex (filtering/convolution, pooling, normalisation, thresholding).

A key ingredient for the success of the HMO model is its heterogeneity: Bypass connections, sub-parts of the network with very different filtering and normalisation parameters, etc.

The successful performance necessitated a retraining of the more than 1.000 output units for every object category and difficulty; predicting IT neuronal responses required different weights (and procedures) from categorisation performance.

Object perception is one of the central goals of human vision

2

Object recognition is considered (almost) “solved” in (parts of) computer vision thanks to the “deep learning revolution”

DNNs as models of human vision?

DNNs as models of human vision?

DNNs are clearly inspired by the architecture of human visual system:
Large number of individual, simple processing units, strong connectivity, hierarchical organisation, convolution followed by nonlinearities, receptive fields getting progressively more complex and larger at later processing stages.

Equally clearly, in DNNs the network units and connections are an enormous simplification given the sophisticated nature and diversity of neurones, axons and dendrites in the brain.

On the other hand, often the strength of a model lies not in replicating the original system but rather in its ability to capture the important aspects while abstracting from details of the implementation (e.g. Box, 1976).

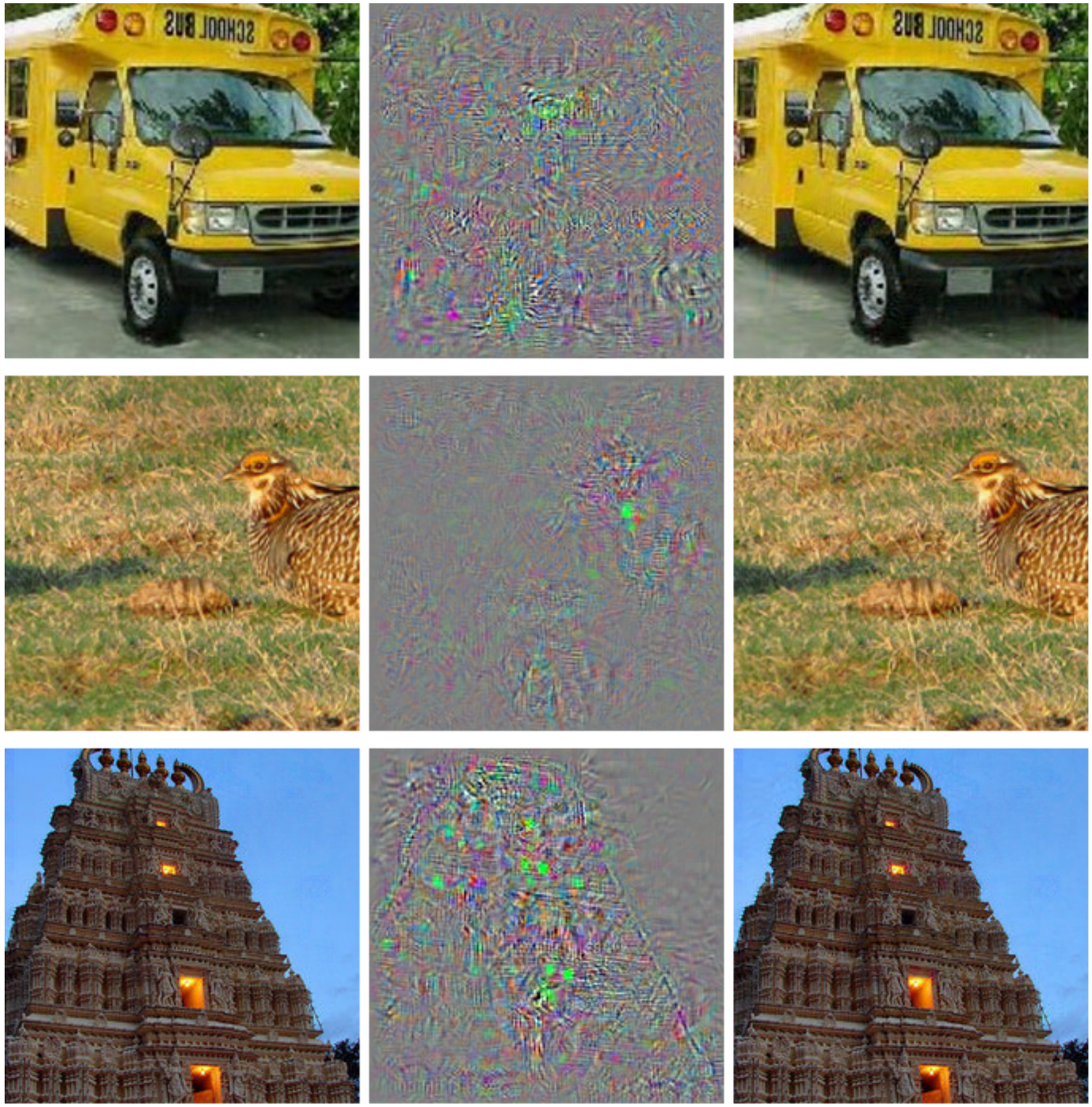
Do current DNNs contain the essential ingredients and the correct abstractions?

It is a capital mistake to theorize before one has data.

(Sherlock Holmes)

Arthur Conan Doyle (1891). A Scandal in Bohemia. *The Strand Magazine*, July issue.

Adversarial attacks?



Szegedy et al. (2014), Intriguing properties of neural networks. *arXiv*:1312.6199v4, 1–10.

Adversarial attacks, random perturbations and generalisation in DNNs

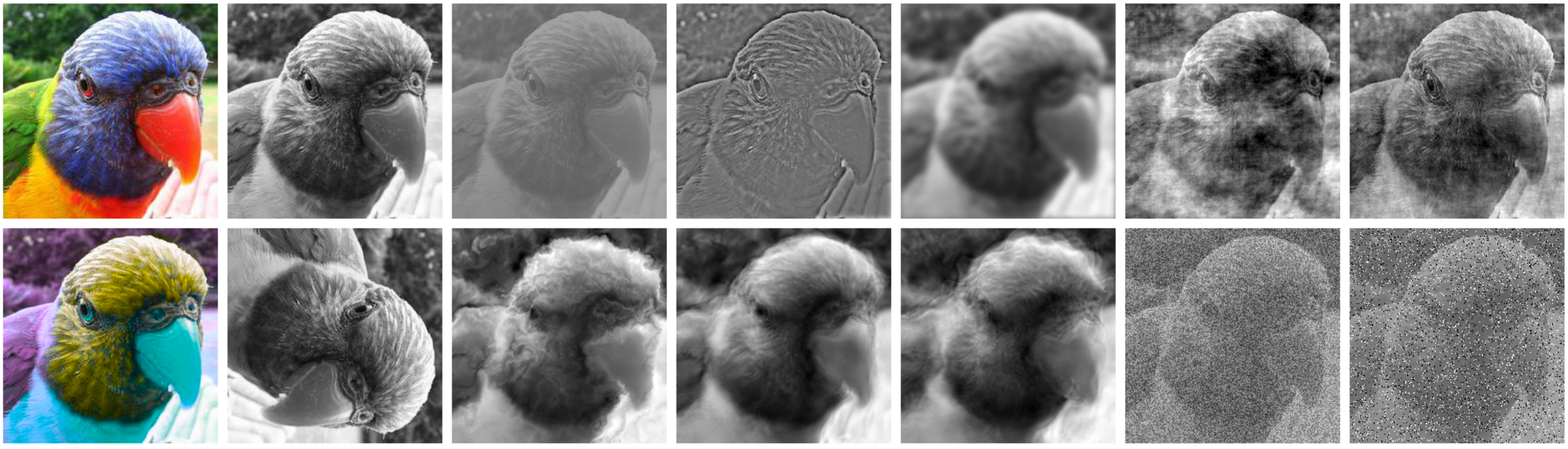
Adversarial attacks show generalisation errors of DNNs

To be fair, however, only to carefully designed stimuli, exploiting the knowledge of all the weights and gradients in the DNN.

Data augmentation (re-training) often leads to robustness against a specific adversarial attack, but it does not guarantee robustness against adversarial perturbations in general.

Is this a strong argument against DNNs using similar computations as human vision?

What about generalisation abilities—robustness— of DNNs and humans to weak signals and (randomly) degraded stimuli rather than carefully engineered “freak” stimuli?



Geirhos, Medina Temme, Rauber, Schütt, Bethge and Wichmann. (2018). Generalisation in humans and deep neural networks. *Advances in Neural Information Processing Systems (NeurIPS) 31, 7549–7561.*

Images and categories

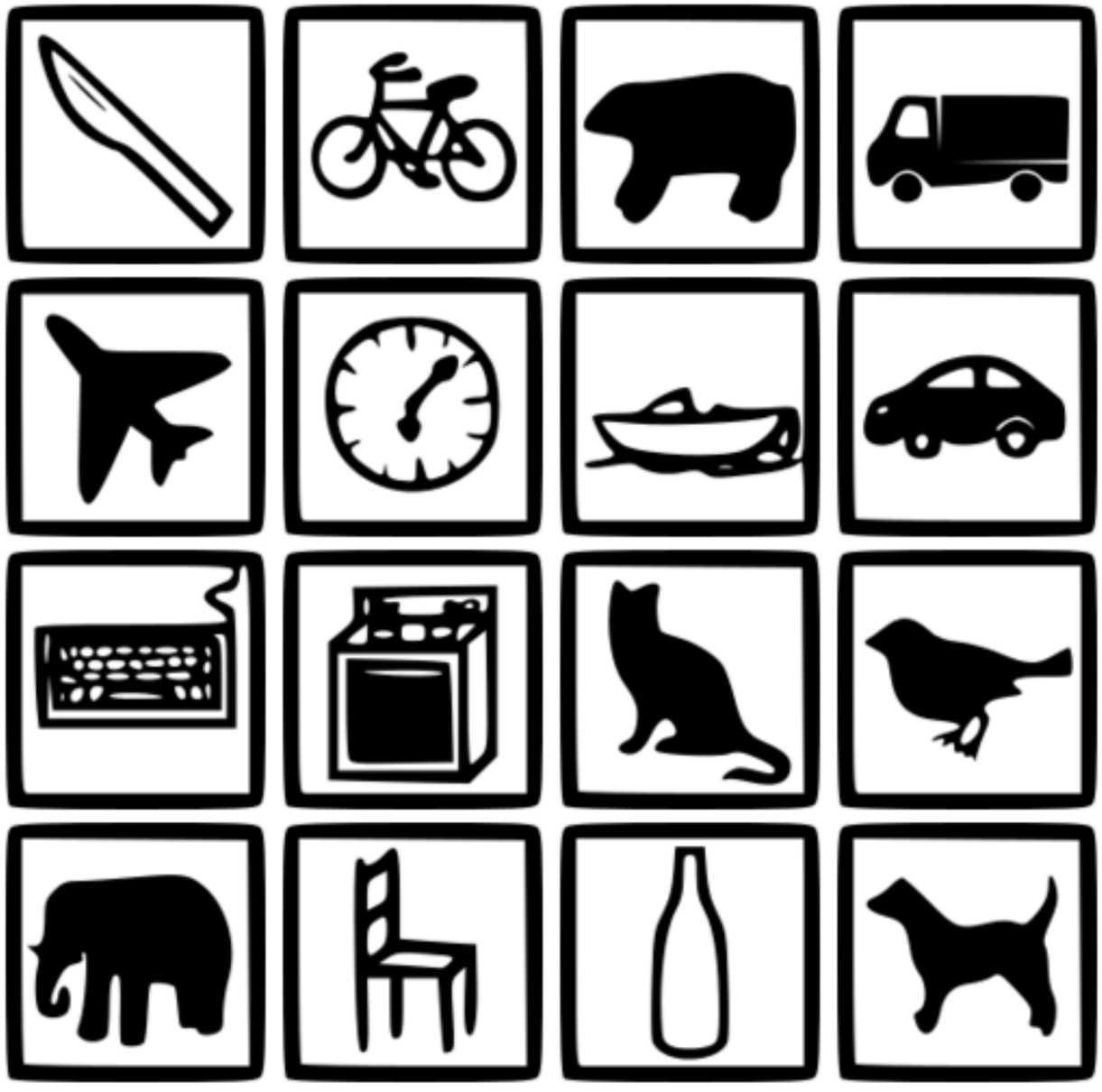
All images from the training set of ImageNet 2012 database.

To compare human observers to DNNs, a categorisation in 1000+ classes at different psychological levels is not optimal.

MS COCO database is structured according to 91 basic or entry-level categories, making it an excellent source for an object recognition task using human observers.

We used MS COCO categories with images from ImageNet, mapping, if possible, the ImageNet label to a MS COCO entry-level category.

We retained 16 non-ambiguous categories with 213,555 images.

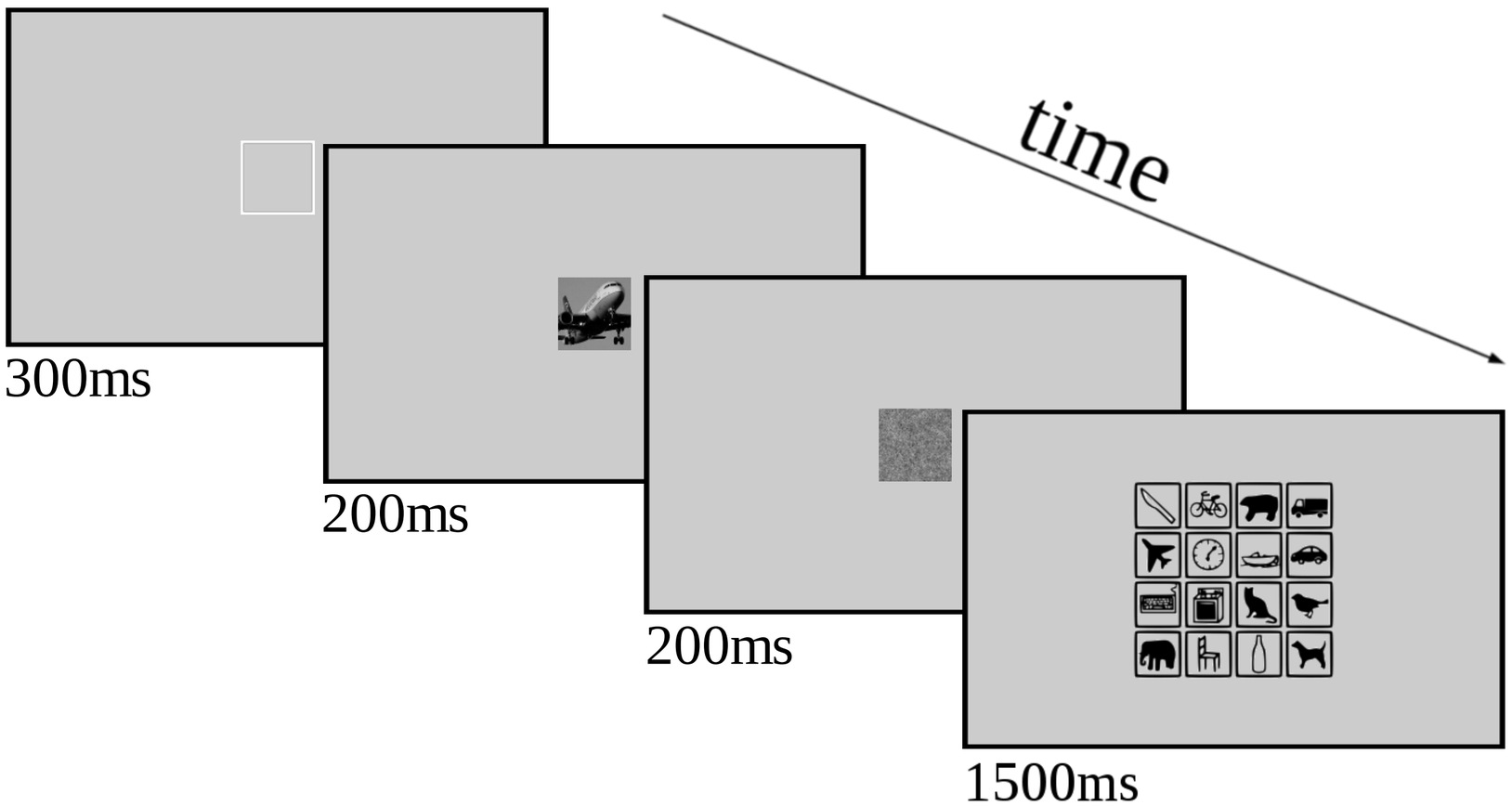


DNNs and methods

Number of well-known, successful and architecturally different DNNs:
AlexNet, VGG-16, GoogleLeNet, ResNet-50 … .

Experimental protocol chosen to allow fair comparison between humans and DNNs as models of the human visual system for core object recognition:

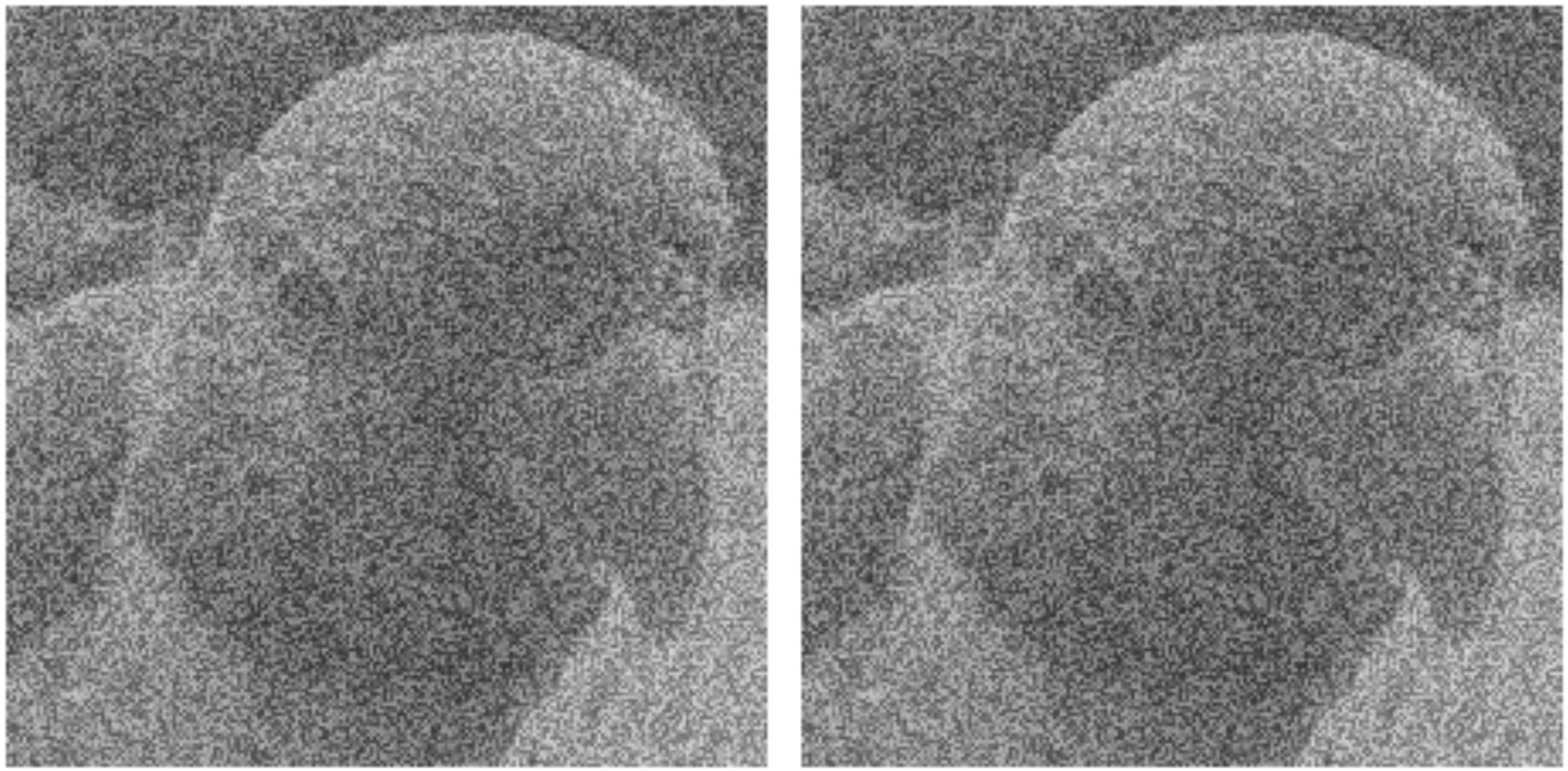
- short presentation time (200 ms)
- followed by a high contrast 1/f noise mask (200 ms)
- fast-paced responding (1500 ms, mouse to select one of 16 icons)



Train	Test
	

Super-human performance of ResNet-50 if trained and tested on the (original) full colour ImageNet images.

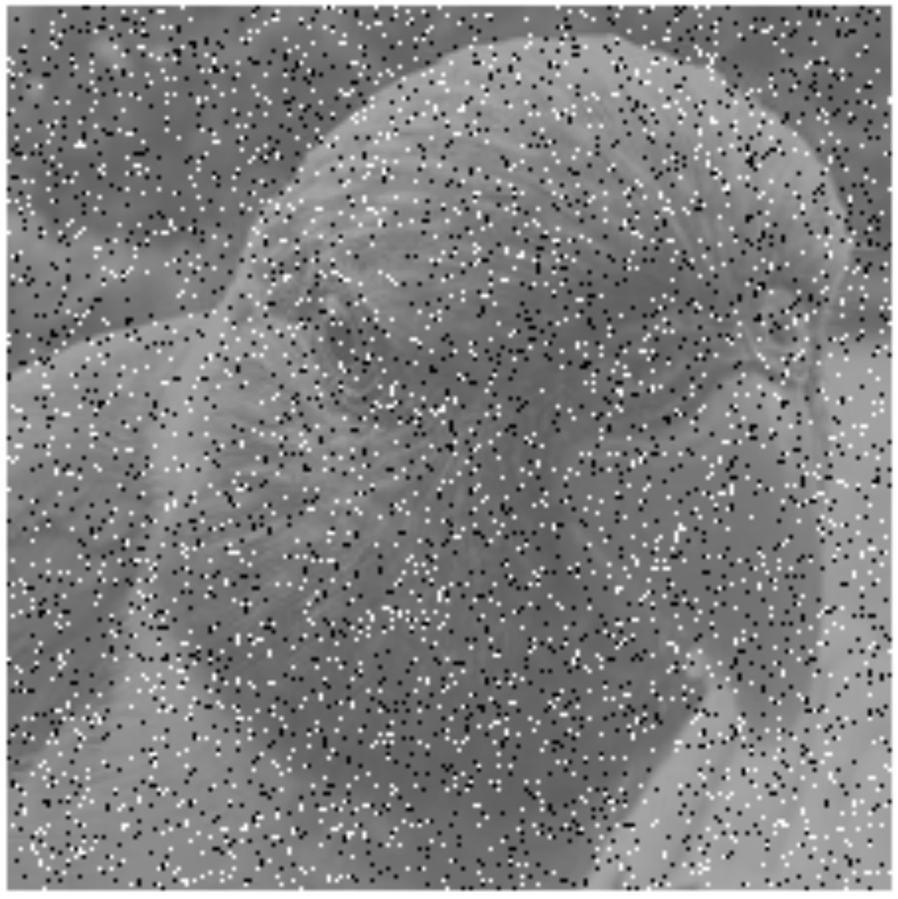
GEIRHOS et al. (2018)



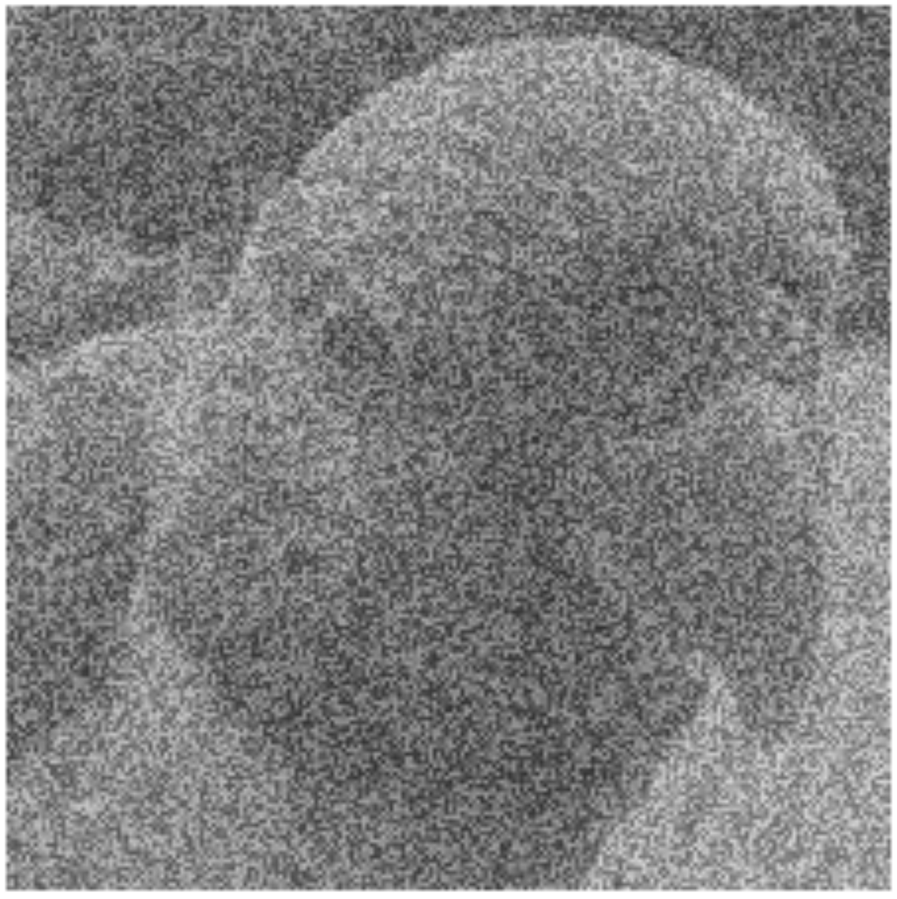
Super-human performance of ResNet-50 if trained and tested on grey-scale converted ImageNet images with additive uniform noise.

Geirhos et al. (2018)

Train

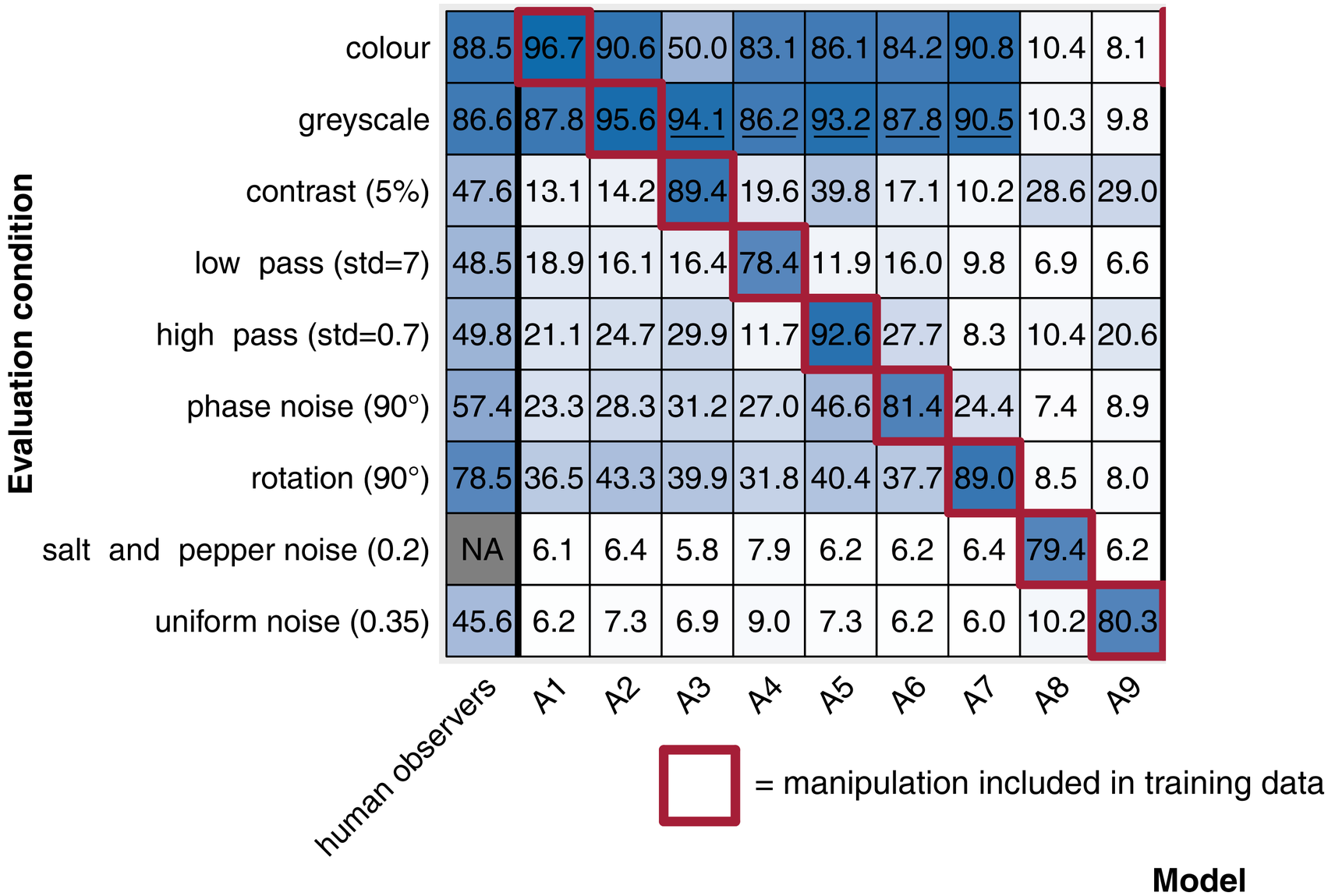


Test



Chance performance of ResNet-50 if trained on grey-scale converted ImageNet images with additive salt-and-pepper noise but tested on the same images but with additive uniform noise.
Striking generalisation failure

Geirhos et al. (2018)



GEIRHOS et al. (2018)

Summary of DNN's (2/2)

Have we replaced our wet and messy grey-box in our skull by an equally inscrutable black-box (DNNs) in our computer?

Prediction is a necessary but not sufficient condition for an idea or theory to be a scientific: We need to understand what is going on, too!

Furthermore, care is needed when comparing humans to algorithms, or different animal species, different algorithms—similar performance (''behaviour'') in one condition does not necessarily imply similar performance (''behaviour'') in a different condition.

DNNs show remarkably little generalisation to image distortions: super-human performance if part of the training set, but virtually no generalisation.

Claims about strong behavioural—and implied algorithmic—similarities between current DNNs trained on standard image databases and human observers appear overstated.

The End

Felix Wichmann

Neural Information Processing Group
Eberhard Karls Universität Tübingen