Inference on computational models from predictions of representational geometries

Nikolaus Kriegeskorte

MRC Cognition and Brain Sciences Unit

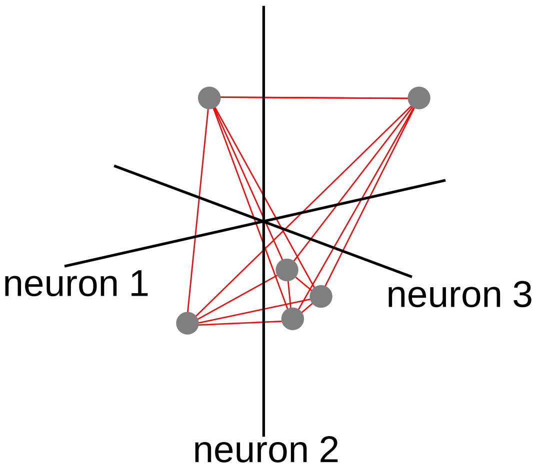
Cambridge, UK

Inference on computational models from predictions of representational geometries

MVPA can reveal what information is present in a region (in a linearly decodable format).

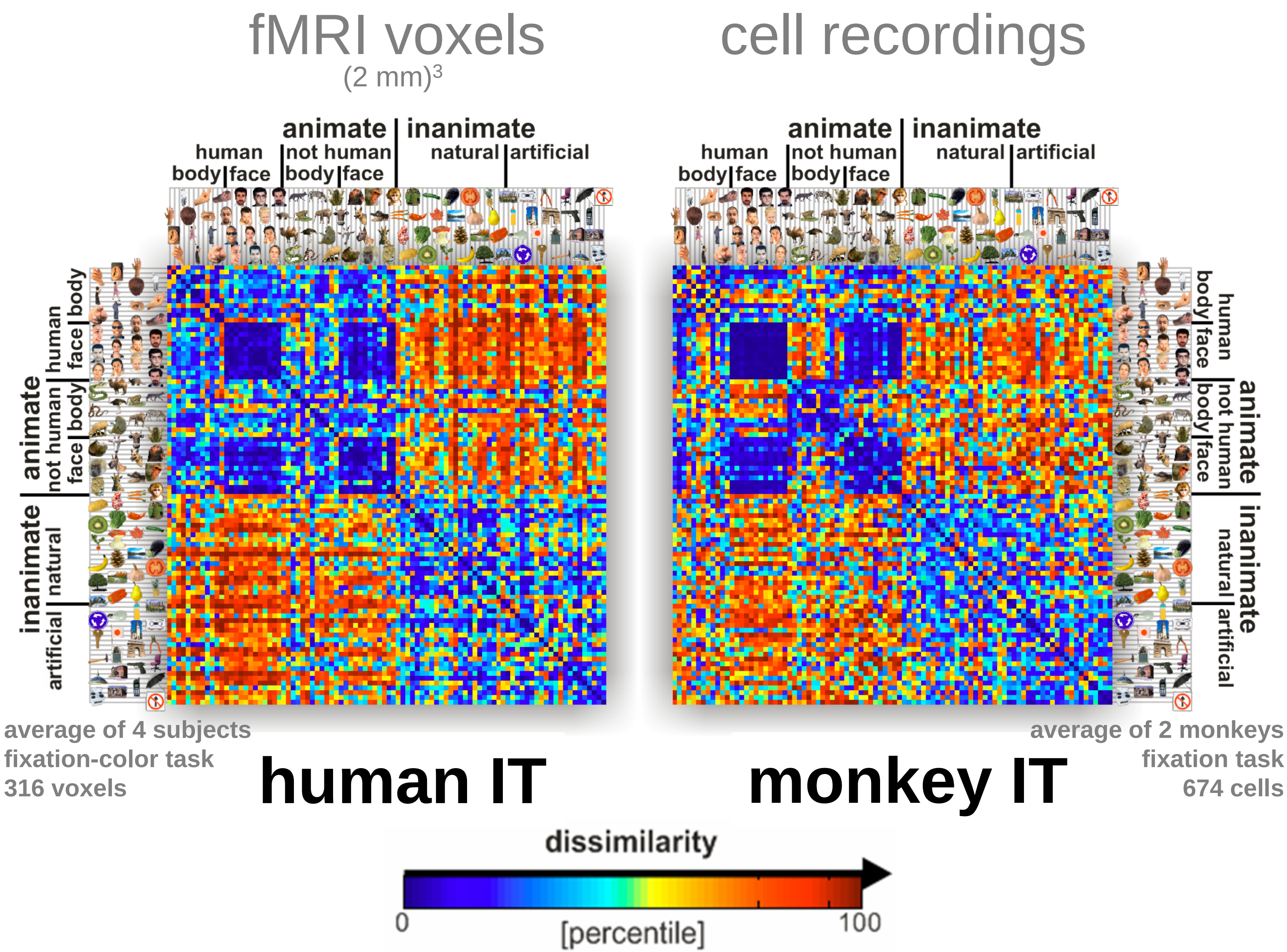
Ultimately, we want to learn about the computational mechanisms of brain information processing.

We can test computational models by comparing their internal representations to brain representations.

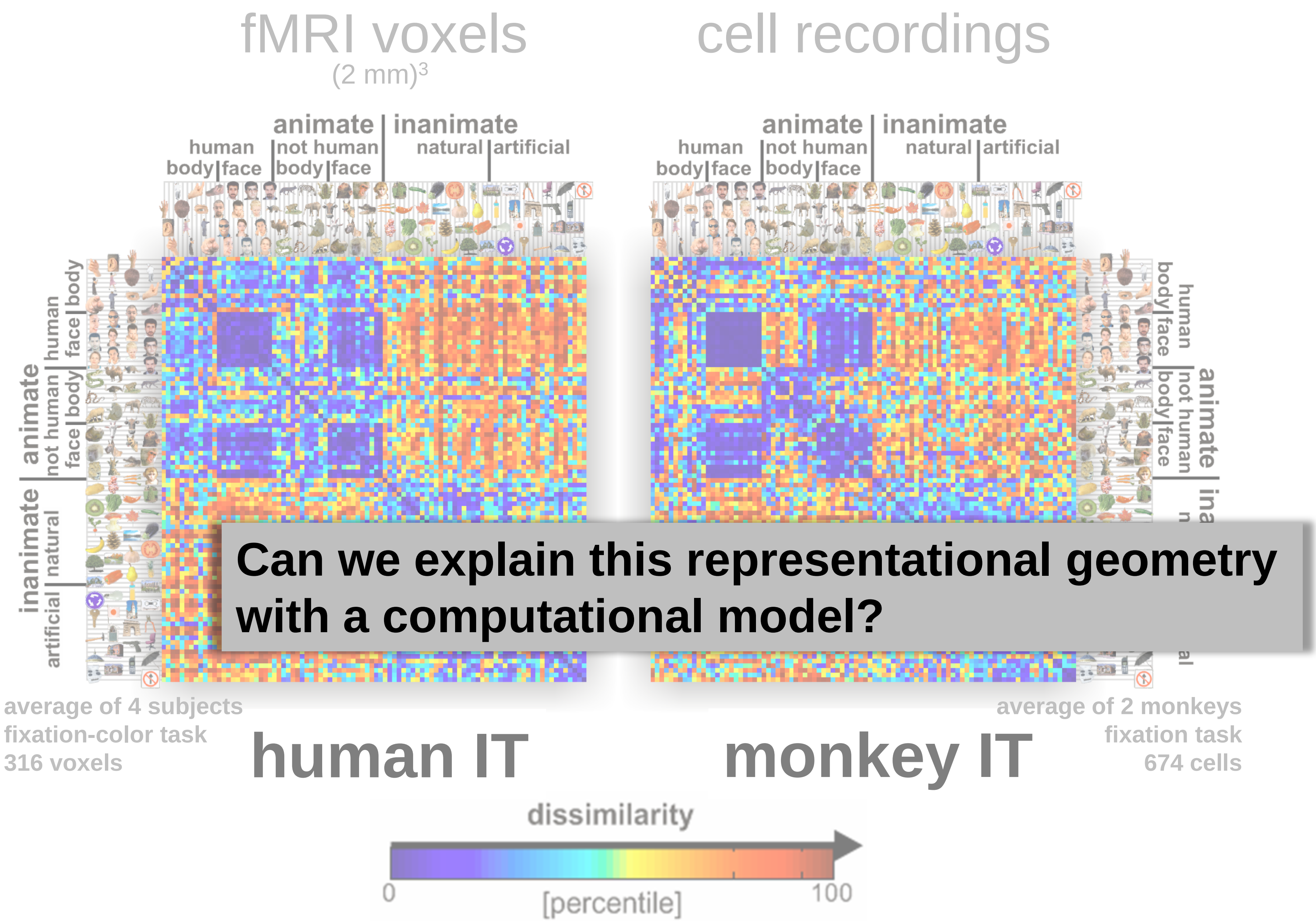


Representational geometry

The geometry of the points in a high-dimensional
response pattern space, which are thought to represent particular stimuli.



Kriegeskorte et al. 2008, cell recordings from Kiani et al. 2007



Can we explain this representational geometry with a computational model?

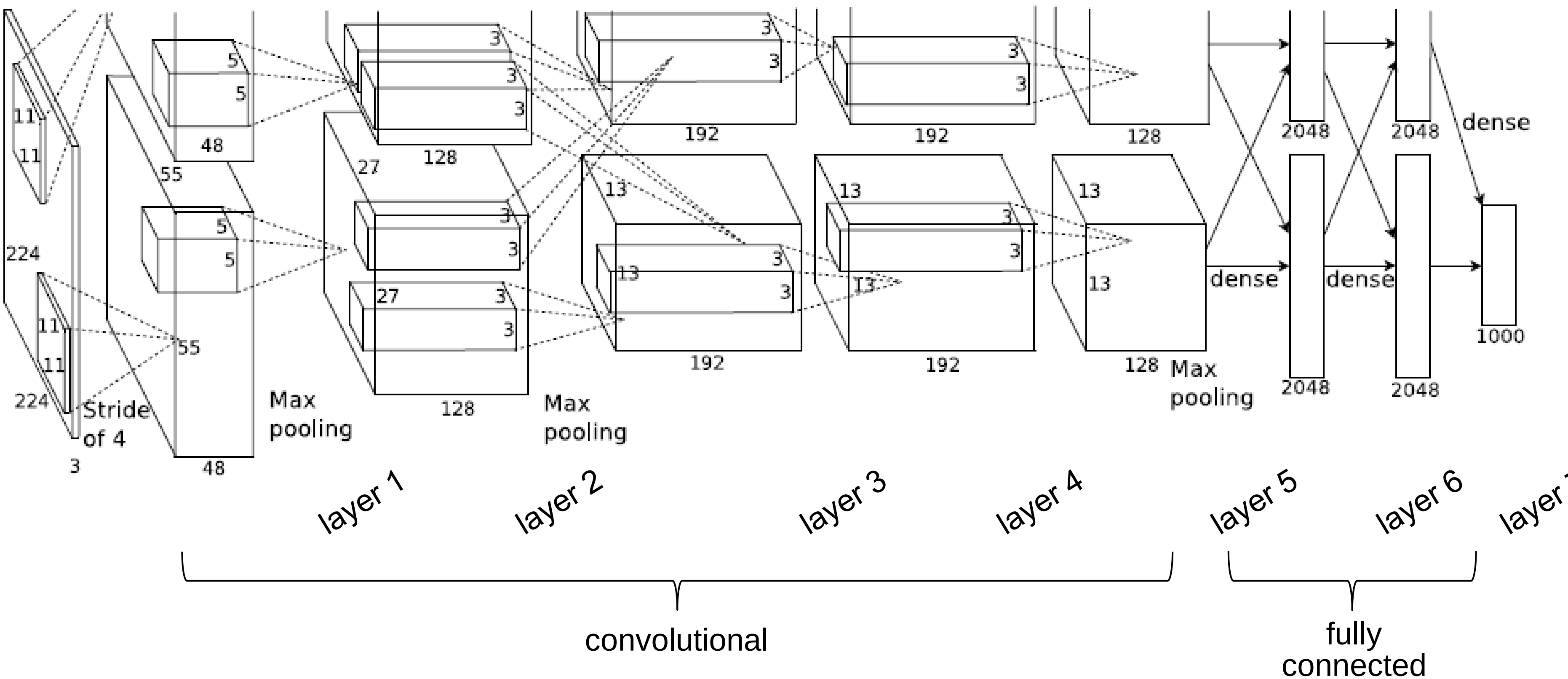
Kriegeskorte et al. 2008, cell recordings from **Kiani** et al. 2007

Seyed-Mahdi Khaligh-Razavi

Deep convolutional neural network

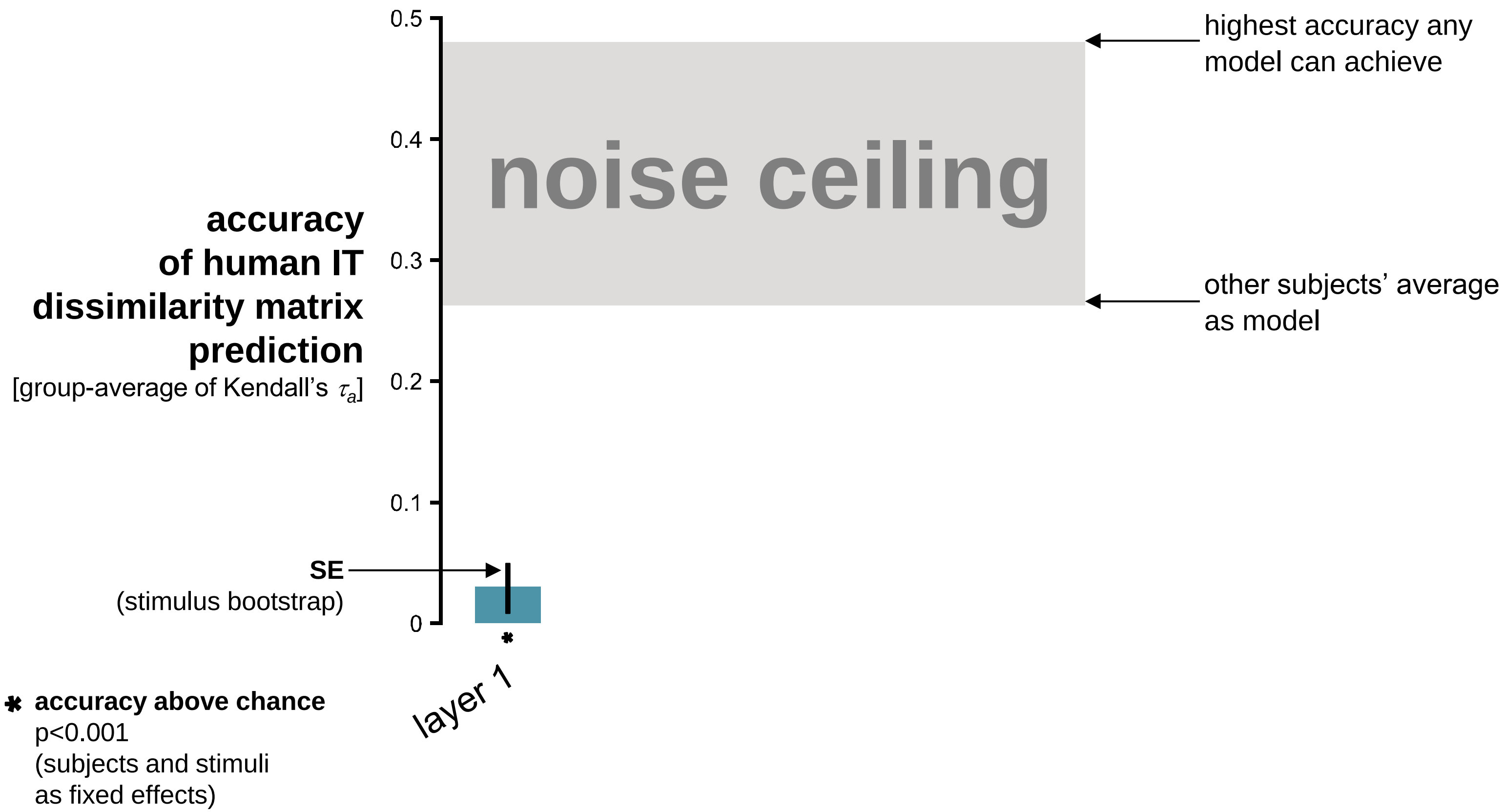
- state of the art in computer vision
- trained with stochastic gradient descent
- supervised with 1.2 million category-labeled images
- 60 million parameters and 650,000 neurons

Is this network functionally similar to the brain?



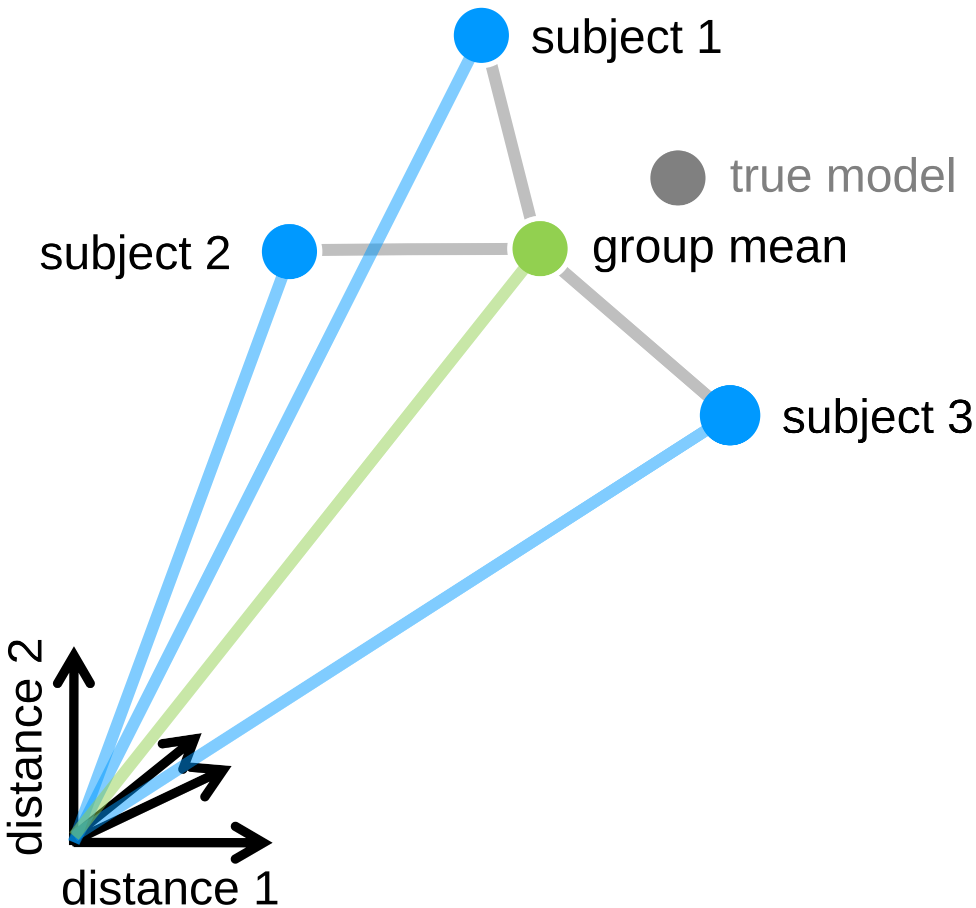
Krizhevsky et al. 2012

Deep supervised convolutional network



Khaligh-Razavi & **Kriegeskorte** 2014, **Nili** et al. 2014 (RSA Toolbox)

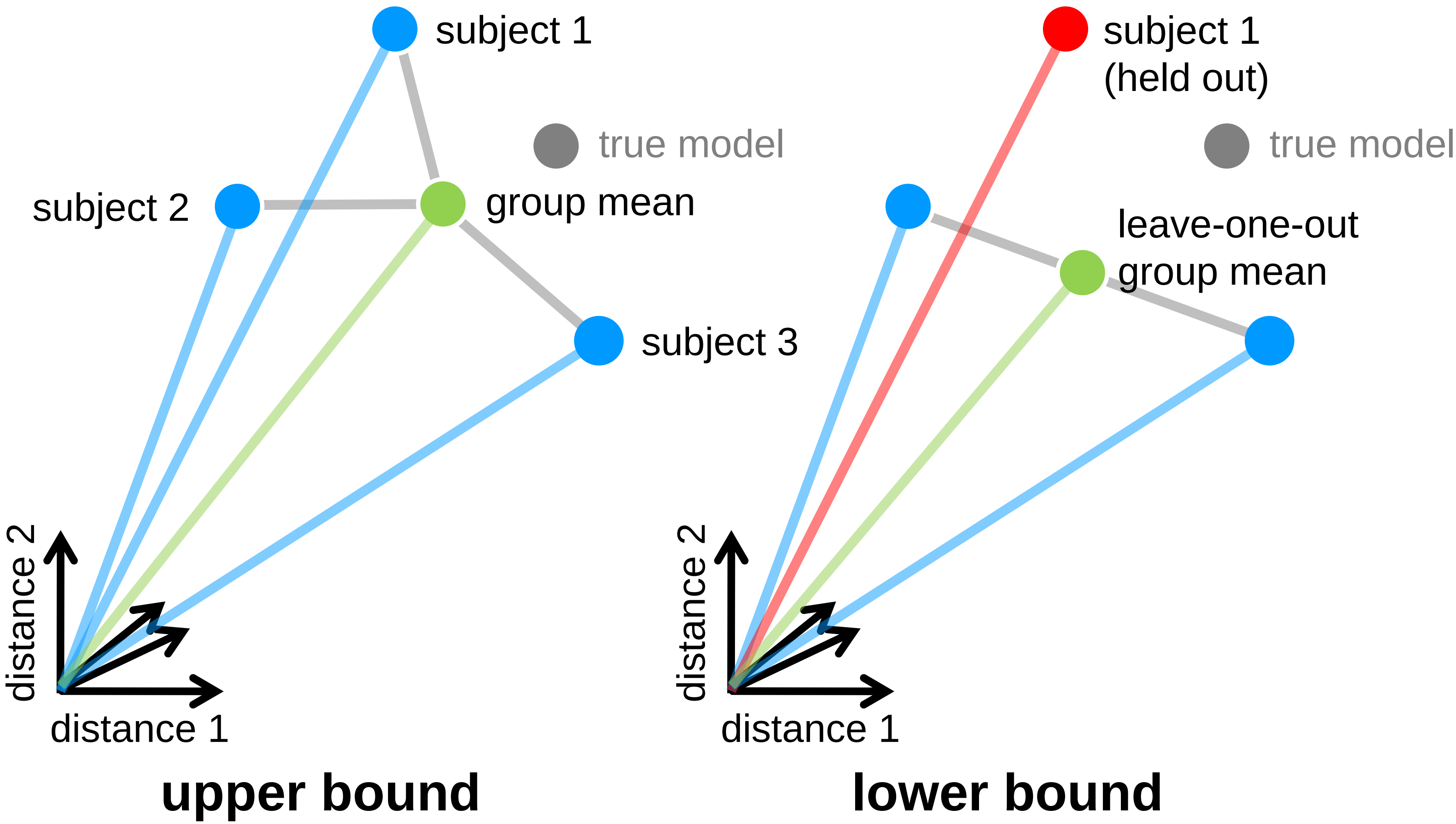
Estimating the noise ceiling



upper bound

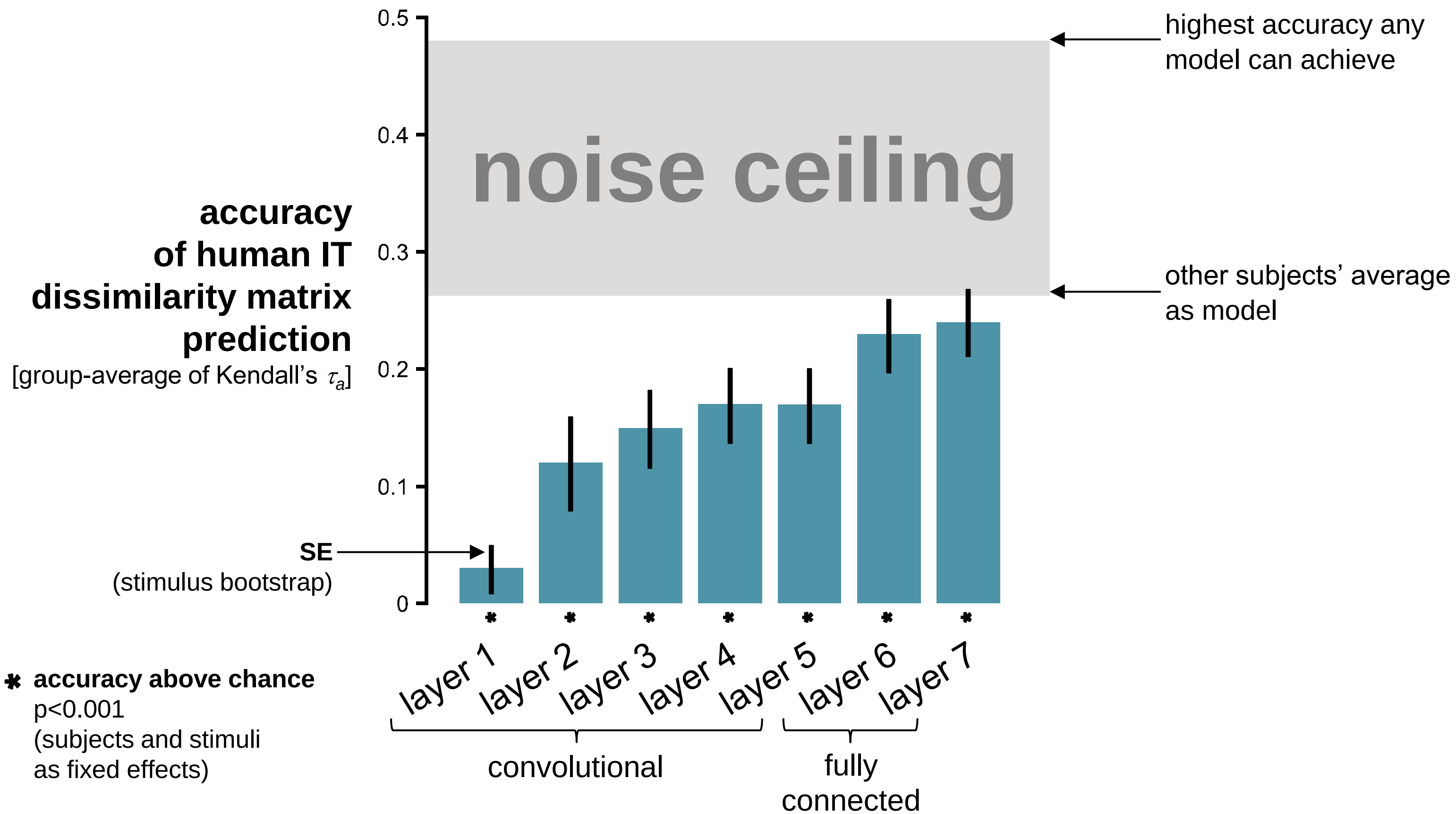
Nili et al. 2014

Estimating the noise ceiling



Nili et al. 2014

Deep supervised convolutional network



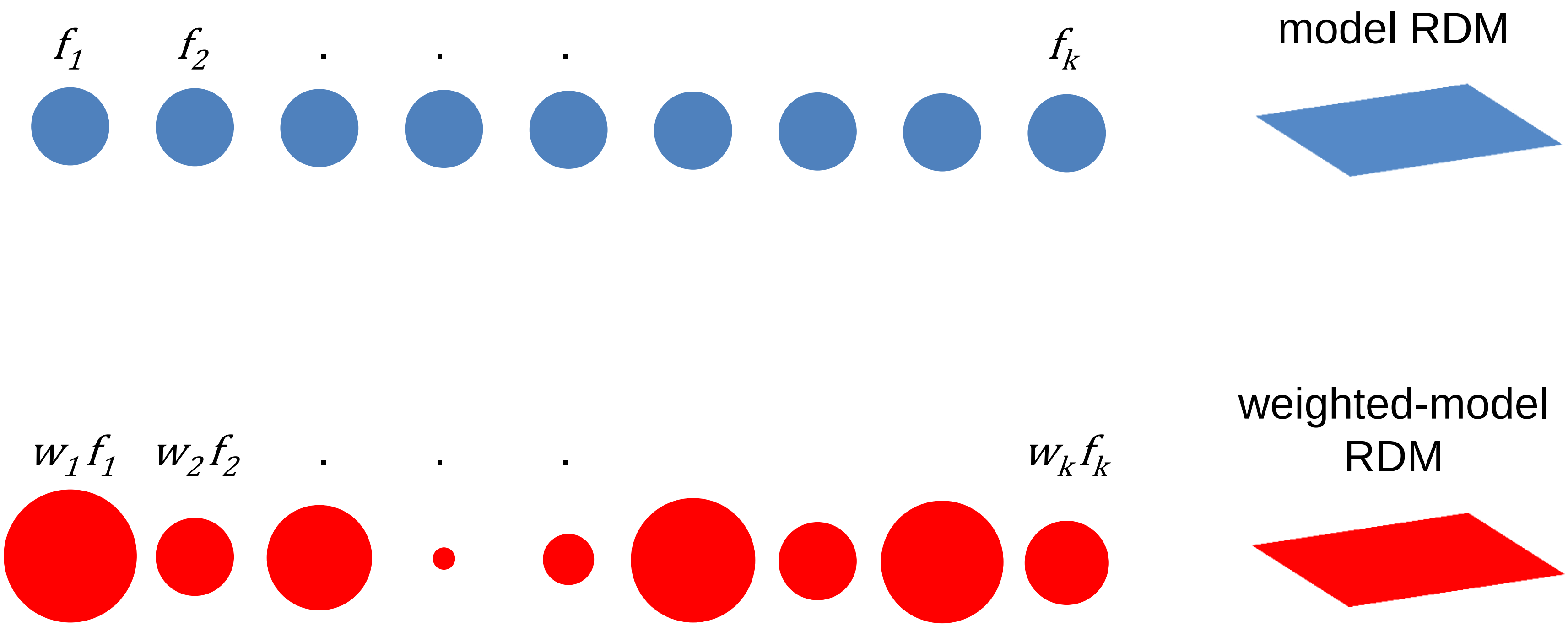
Khaligh-Razavi & **Kriegeskorte** 2014, **Nili** et al. 2014 (RSA Toolbox)

The deep net almost reached the noise ceiling!

Can weighted combinations of its units fully explain the IT representation?

Khaligh-Razavi & **Kriegeskorte** 2014

Representational feature weighting with non-negative least-squares



Khaligh-Razavi & Kriegeskorte (2014)

Representational feature weighting with non-negative least-squares

predicted distance

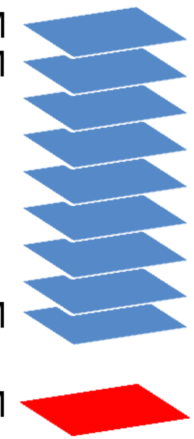
weight k model feature k

$$\hat{d}_{i,j}{}^{2} = \sum_{k=1}^{n} [w_k f_k(i) - w_k f_k(j)]^2$$

stimuli i, j

$$= \sum_{k=1}^{n} w_k{}^{2} [f_k(i) - f_k(j)]^2$$

The squared distance RDM of weighted model features equals a weighted sum of single-feature RDMs.



measured distance predicted distance

$$\mathbf{w} = \arg \min_{\mathbf{w} \in \mathbf{R}^{+n}} \sum_{i \neq j} \left[d_{i,j}{}^{2} - \hat{d}_{i,j}{}^{2} \right]^2 = \arg \min_{\mathbf{w} \in \mathbf{R}^{+n}} \sum_{i \neq j} \left[d^2 - \sum_{k=1}^{n} w_k{}^{2} \cdot \mathrm{RDM}_k \right]_{i,j}^{2}$$

w_k weight given to model feature k
$f_k(i)$ model feature k for stimulus i
$d_{i,j}$ distance between stimuli i,j
$\mathbf{w}$ is the weight vector $[w_1\ w_2\ \ldots\ w_k]$ minimising the squared errors

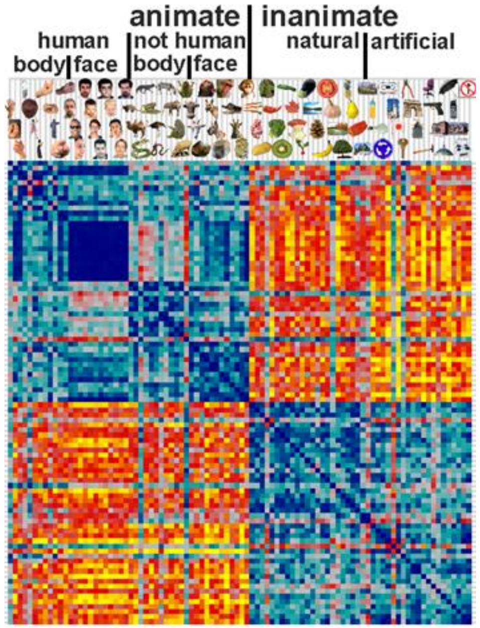
Khaligh-Razavi & Kriegeskorte (2014)

Deep supervised convolutional network remixed & reweighted

3 SVM discriminants

- animate/inanimate
- face/nonface
- body/nonbody

non-neg. least squares
one weight for each layer and each SVM discriminant



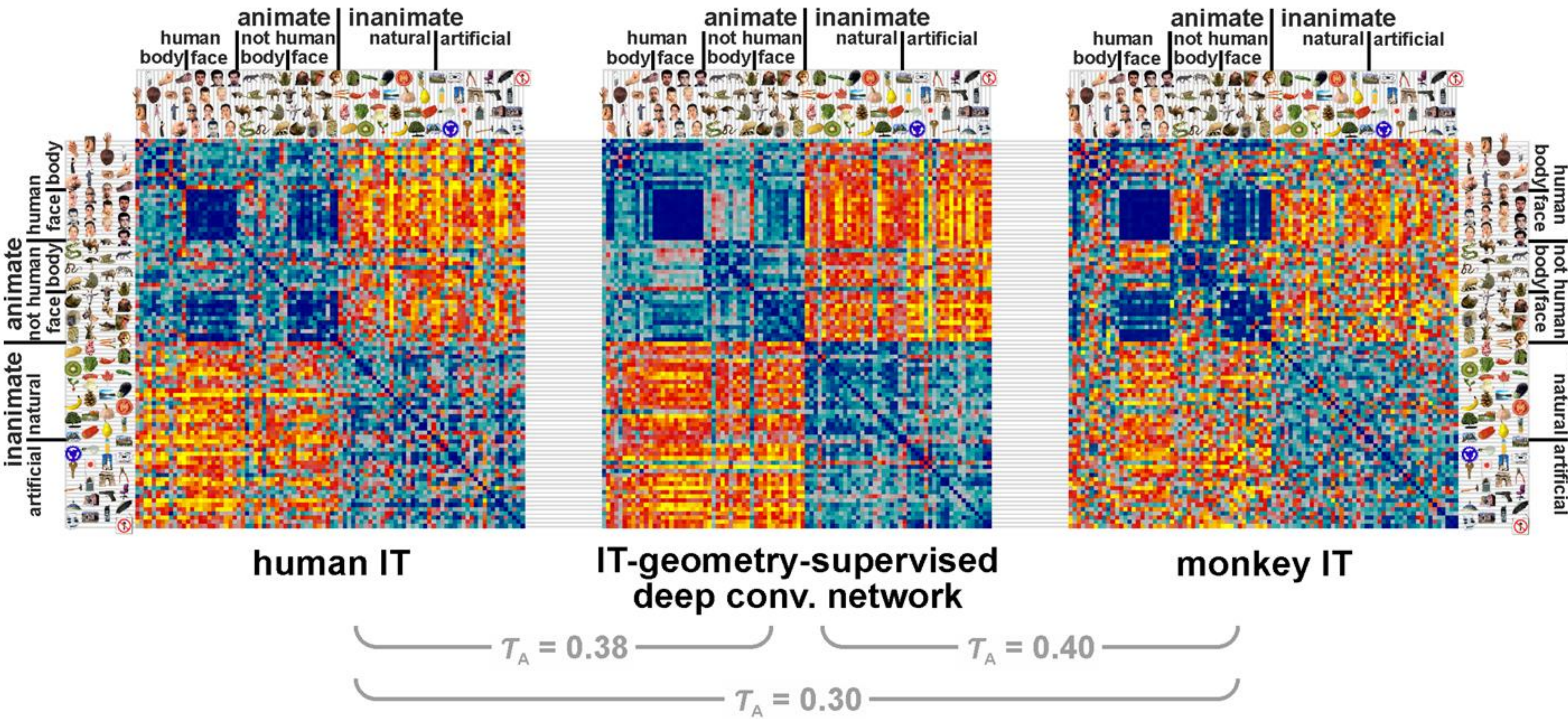
IT-geometry-supervised deep conv. network

Khaligh-Razavi & Kriegeskorte 2014

Deep supervised convolutional network remixed & reweighted

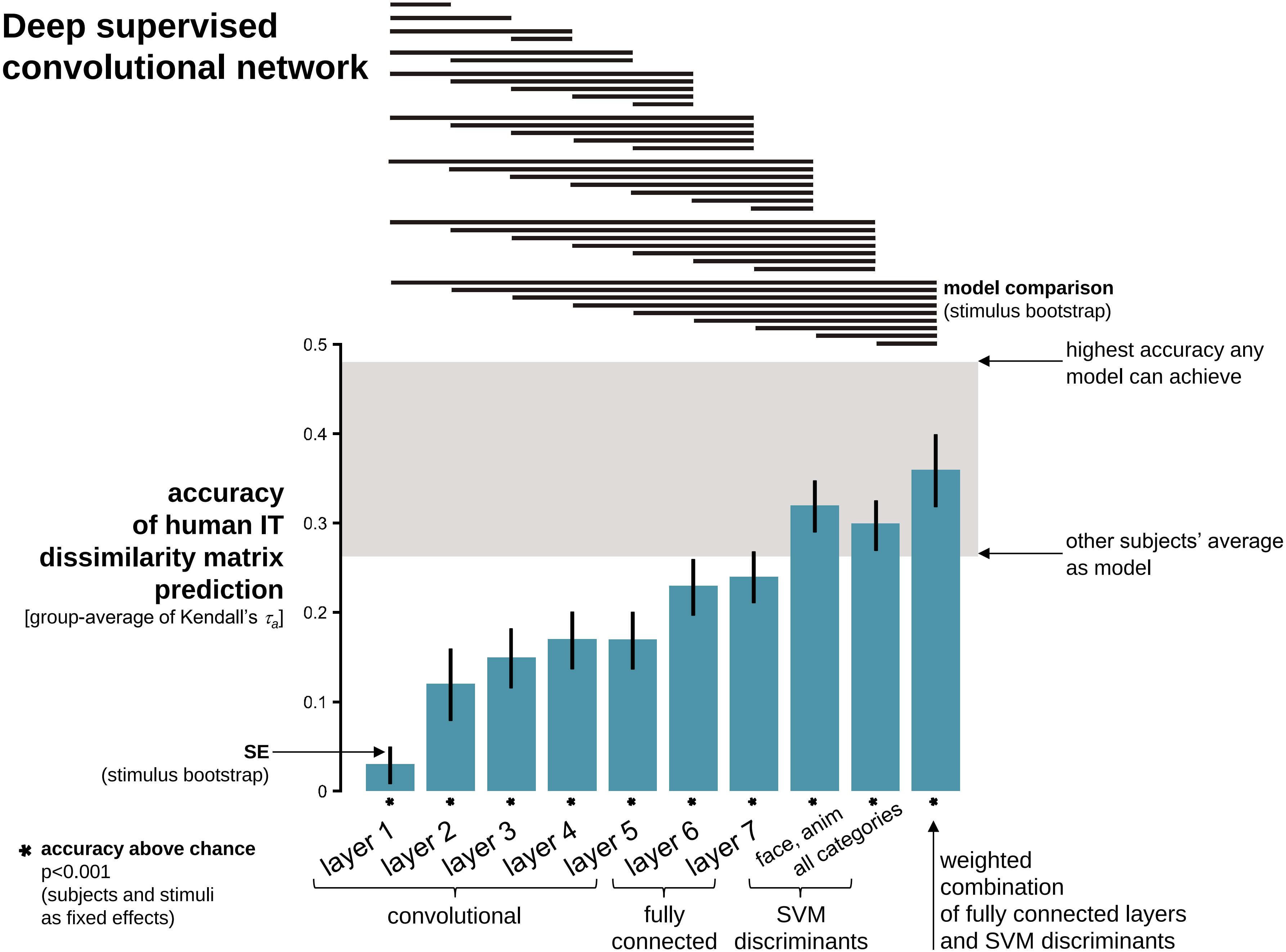
3 SVM discriminants
- animate/inanimate
- face/nonface
- body/nonbody

non-neg. least squares
one weight for each layer
and each SVM discriminant



Khaligh-Razavi & **Kriegeskorte** 2014

Deep supervised convolutional network



Khaligh-Razavi & **Kriegeskorte** 2014, **Nili** et al. 2014 (RSA Toolbox)

Conclusions

We can go beyond decoding stimulus information and test explicit computational models of brain information processing.

Representations in brains and models can be characterised by representational dissimilarity matrices (RDMs).

Multiple models can be statistically compared by nonparametric statistical inference.

The noise ceiling tells us whether a model fully explains our data, guiding us to seek either a better model or more data.

Combinations of representational features correspond to additive squared distance components, which can be fitted with nonnegative least squares.

Nili et al. (2014) **Khaligh-Razavi** et al. (2014) **Diedrichsen** et al. (2011)