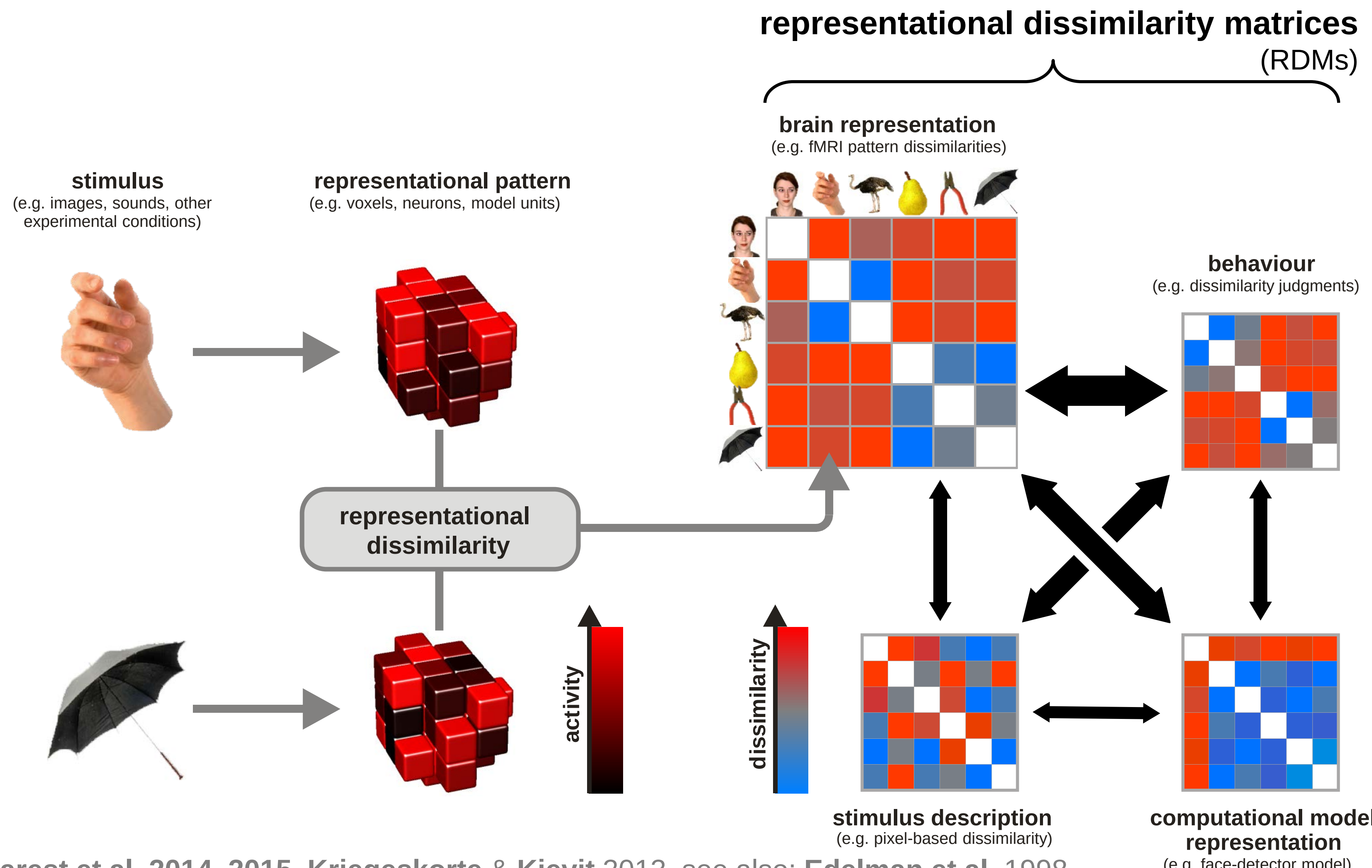
UNIVERSITY OF BIRMINGHAM
COLLEGE OF LIFE AND ENVIRONMENTAL SCIENCES
School of Psychology

Representational similarity analysis

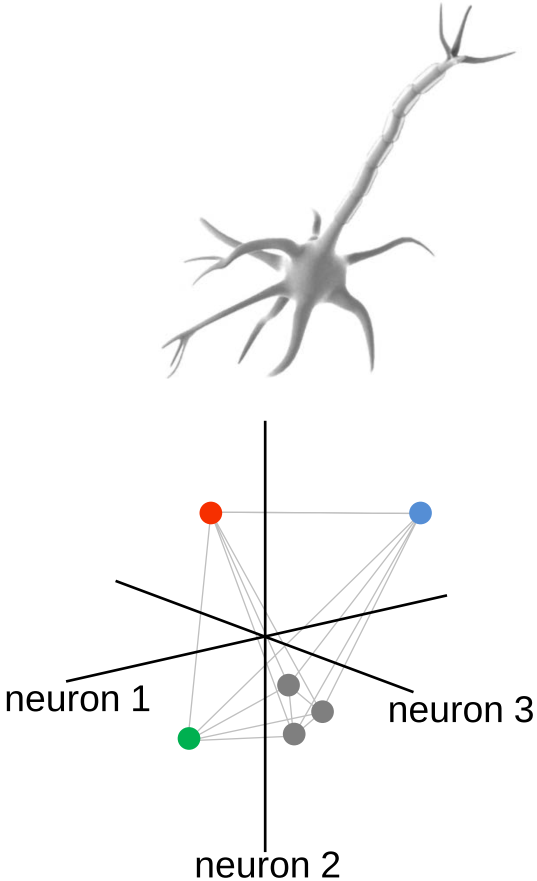
Dr Ian Charest

Representational similarity analysis



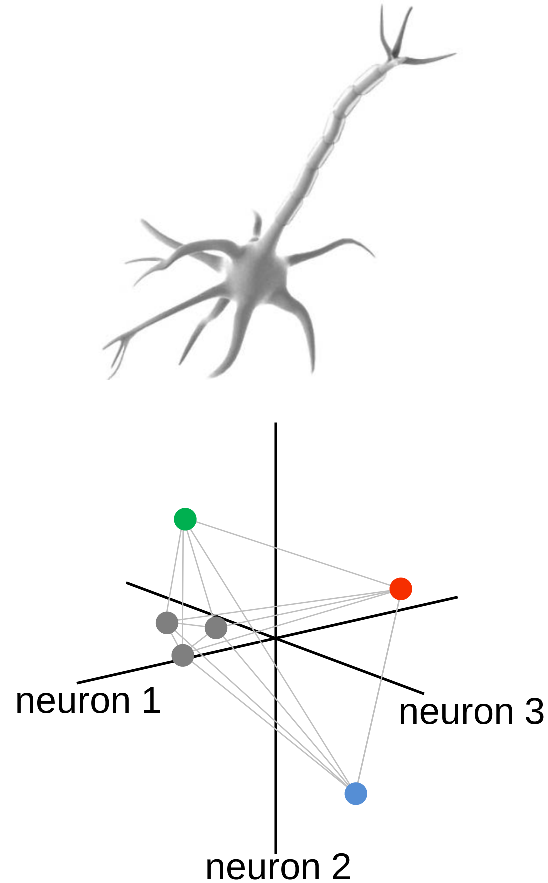
Charest et al. 2014, 2015, Kriegeskorte & **Kievit** 2013, see also: **Edelman et al.** 1998, **Laakso & Cottrell** 2000, **Op de Beeck** et al. 2001, **Haxby** et al. 2001, **Aguirre** 2007, **Kriegeskorte** et al. 2008

Why investigate representational geometries?



downstream neurons
can read out the same information from these codes

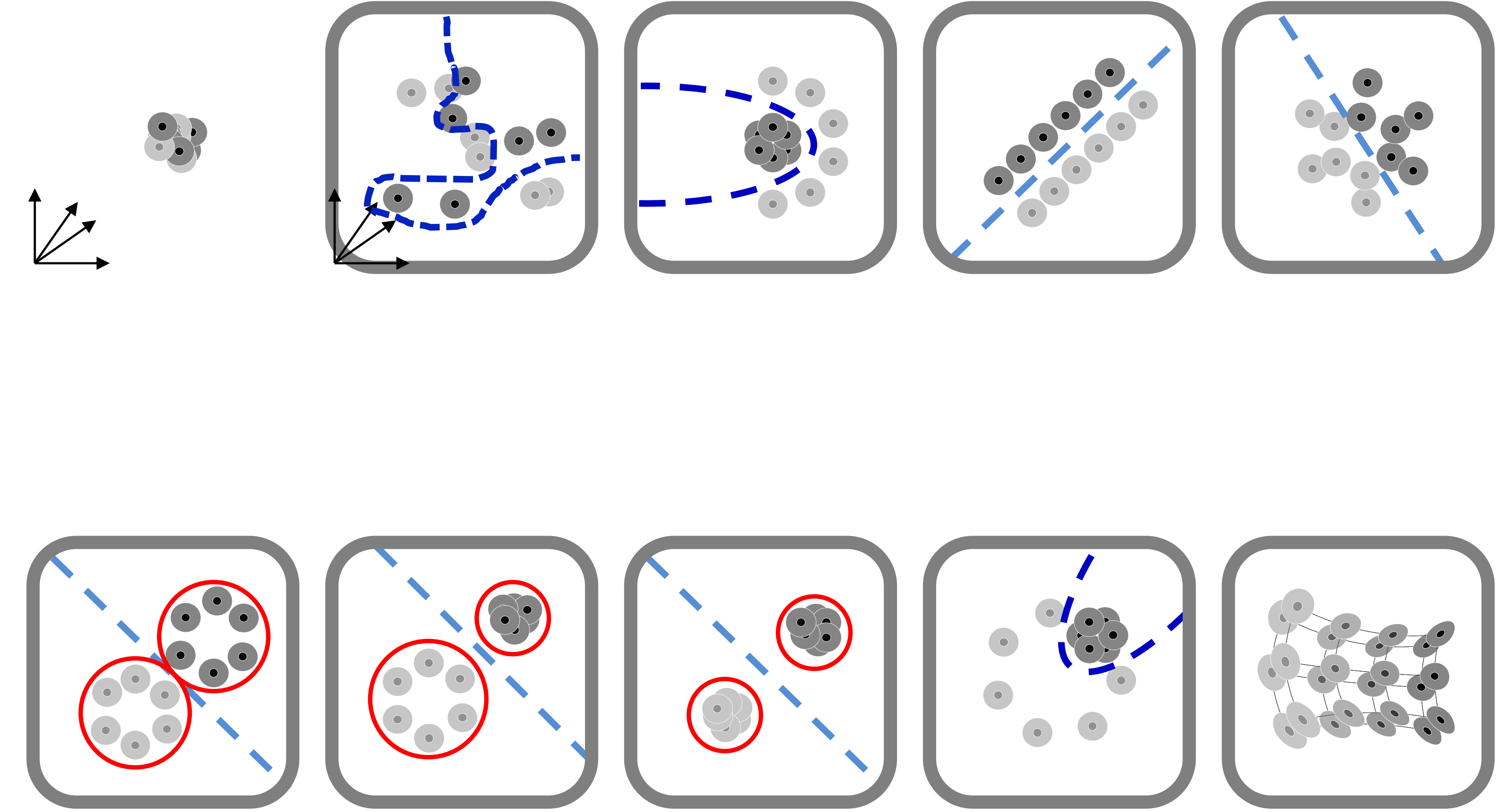
same geometry
→ same information
→ same format



Representational geometry

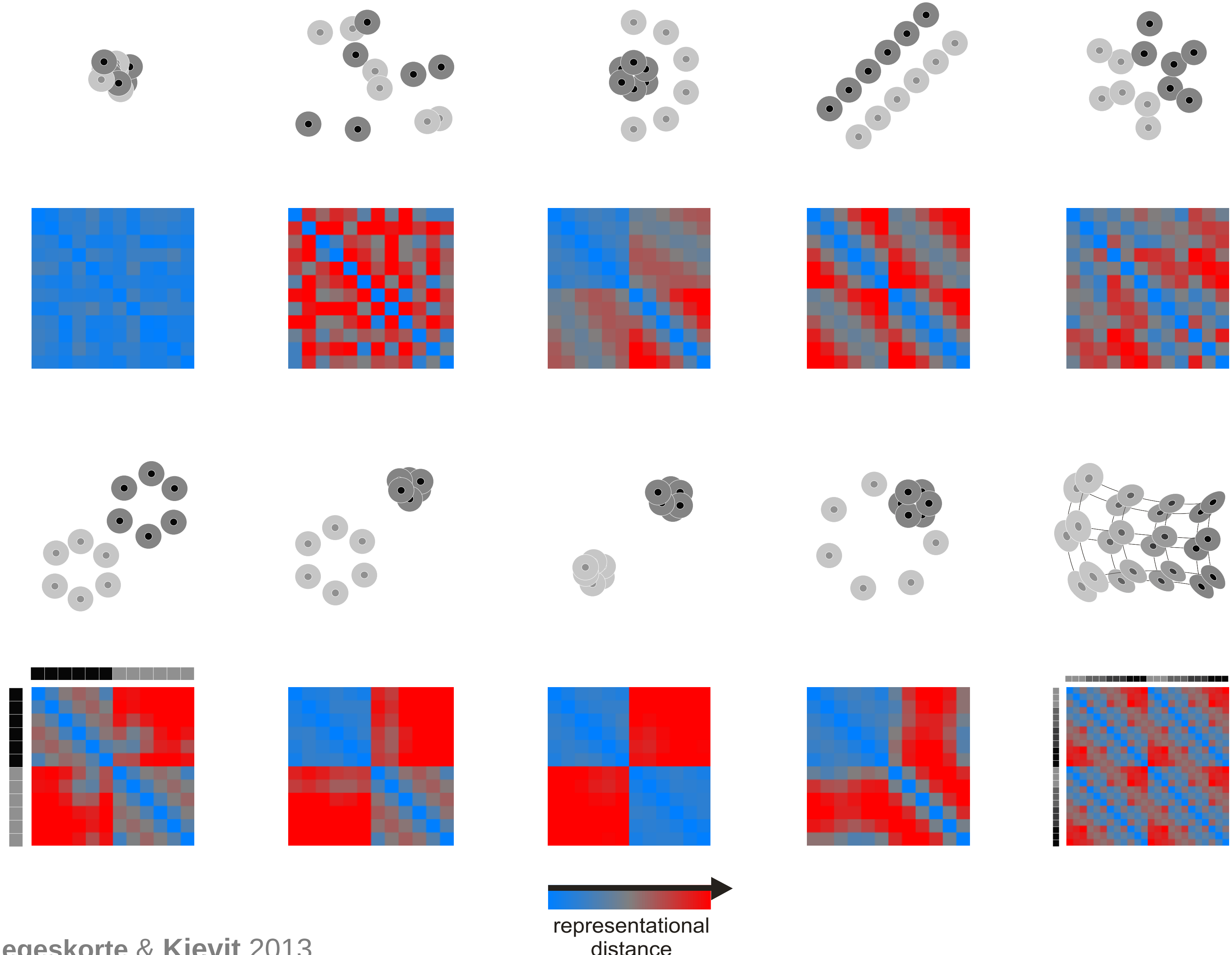
The geometry of the points in a high-dimensional
response pattern space, which are thought to represent particular stimuli.

slide kindly provided by N. Kriegeskorte



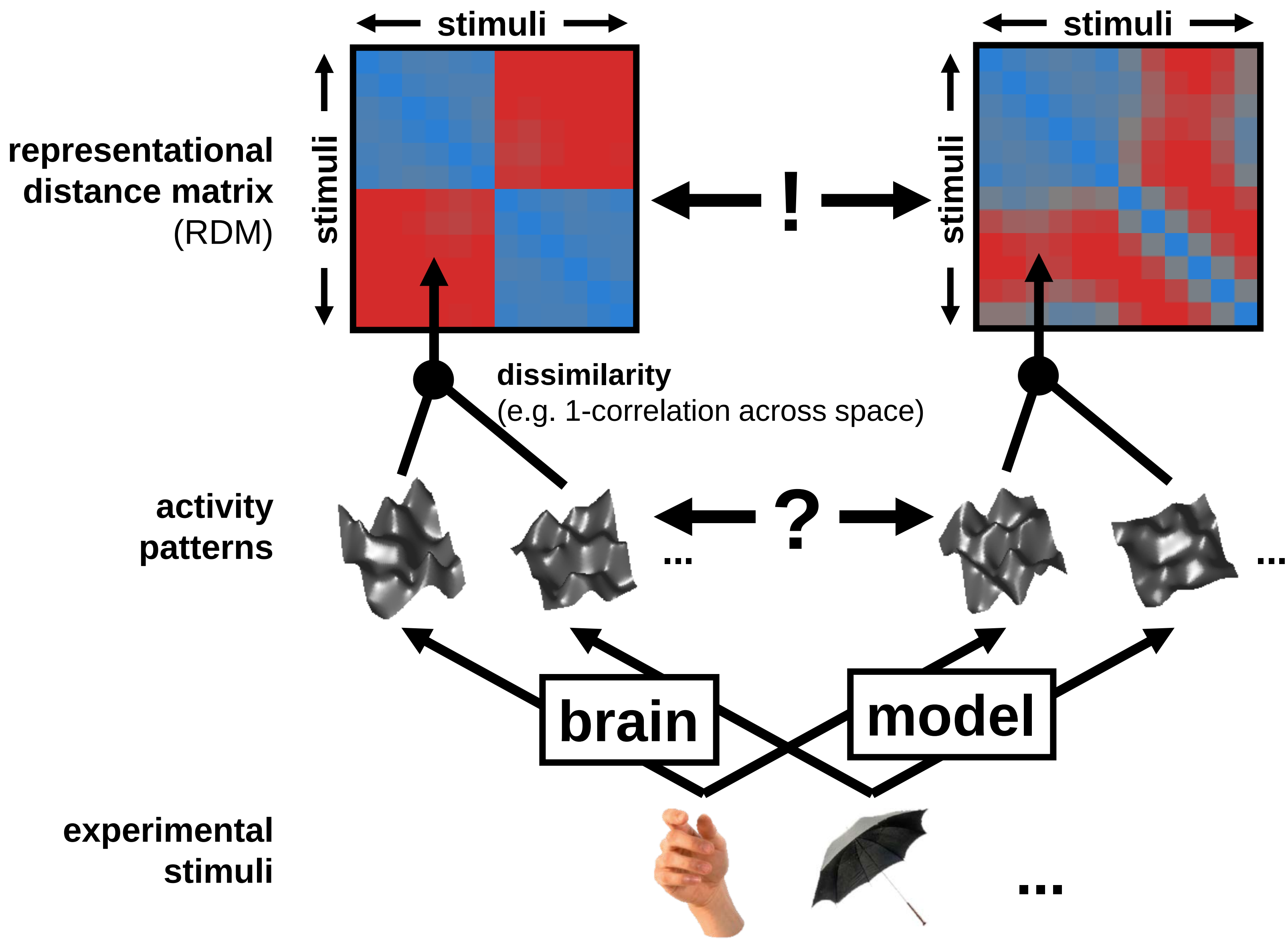
category information
...for linear readout
...for nonlinear readout
...inherently categorical

Kriegeskorte & **Kievit** 2013

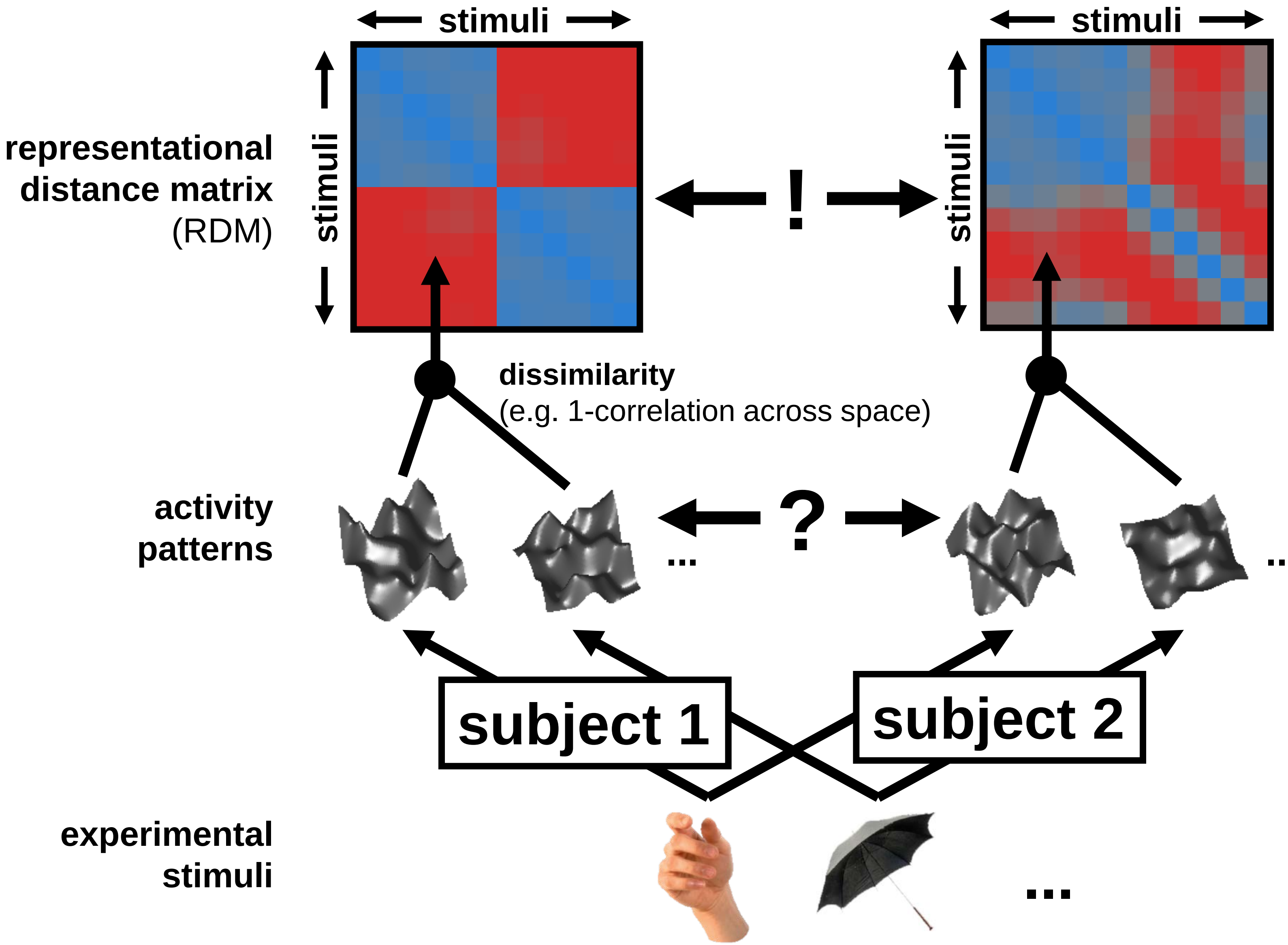


Kriegeskorte & **Kievit** 2013

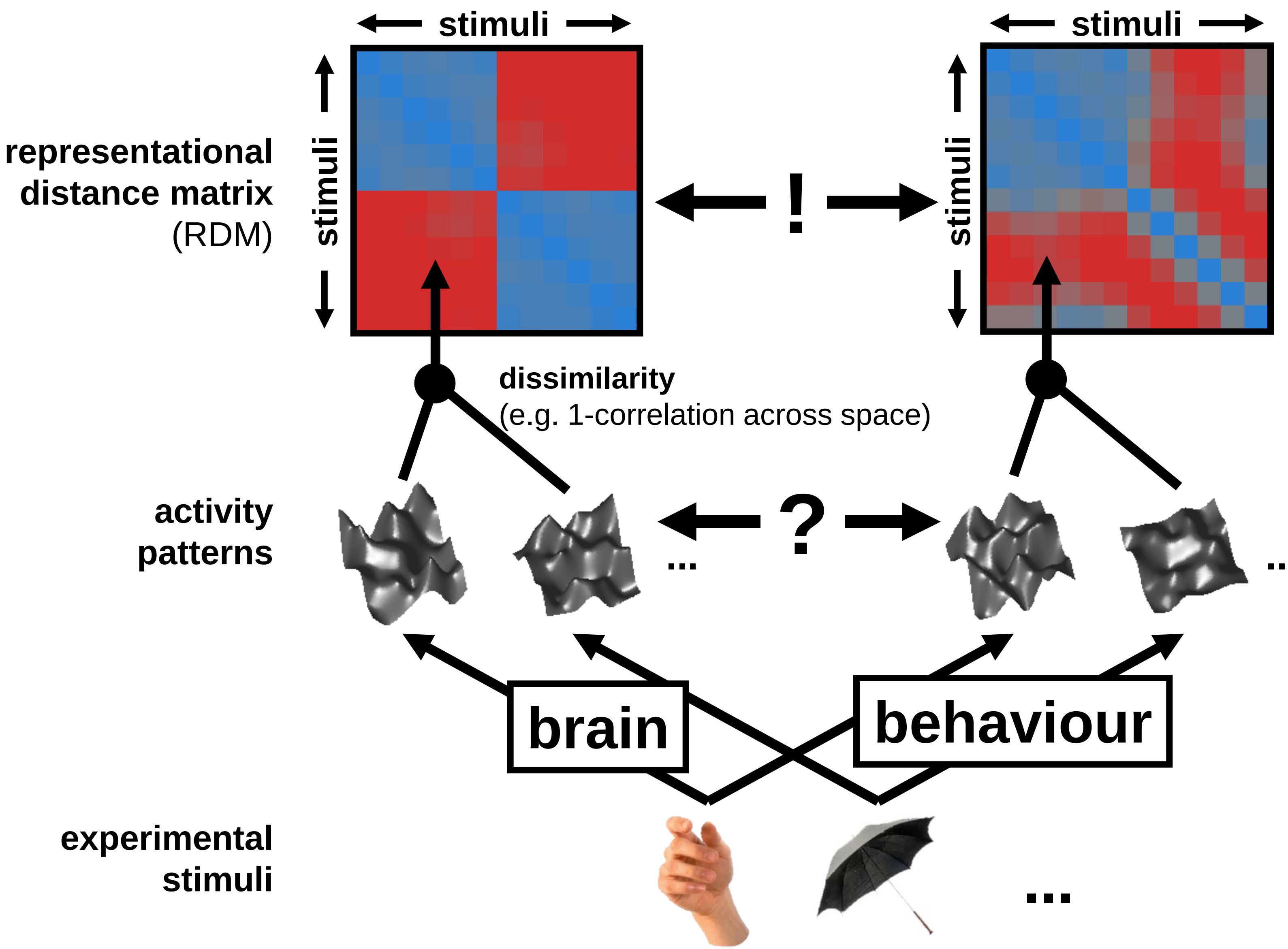
The representational similarity trick



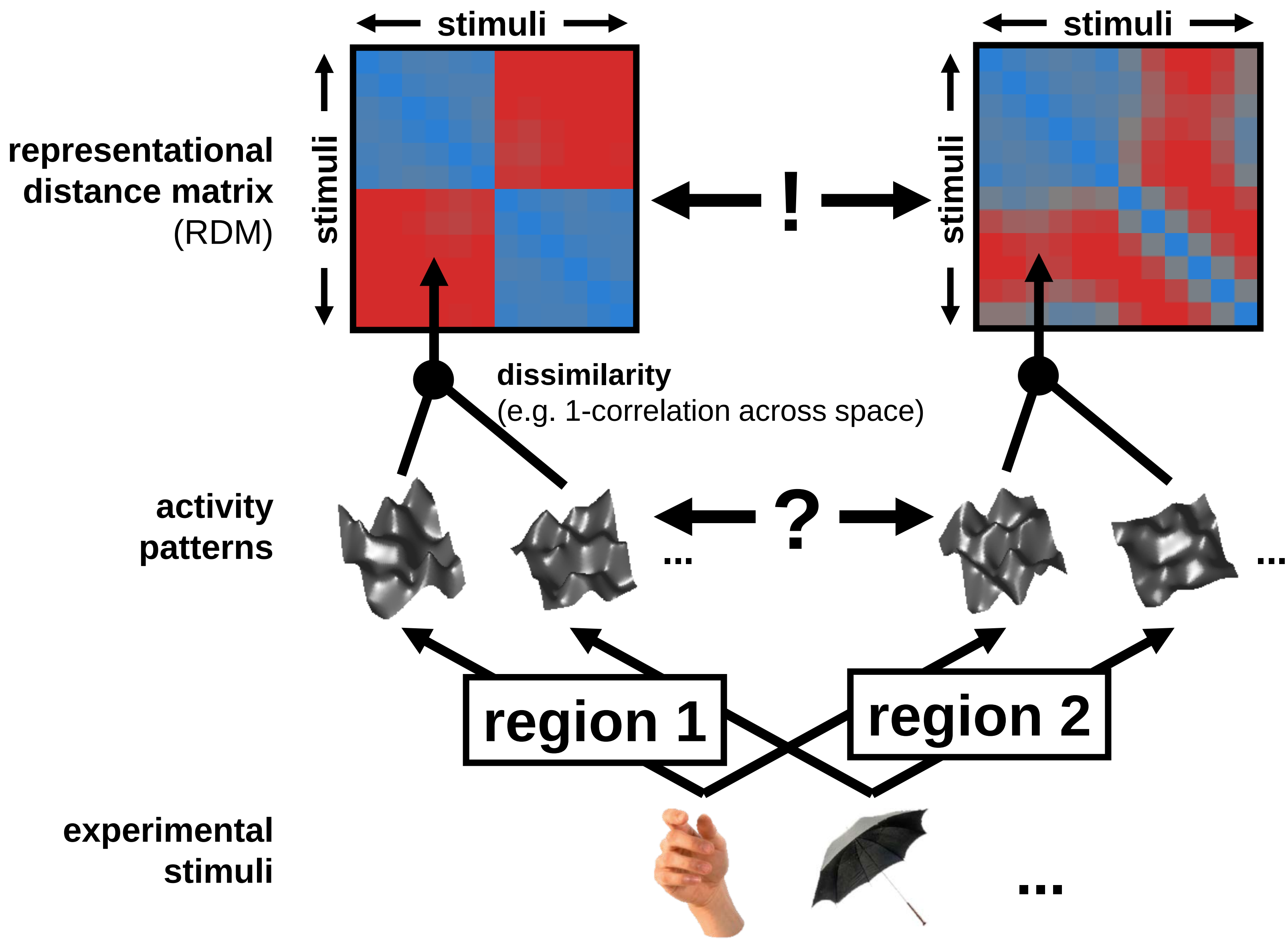
The representational similarity trick



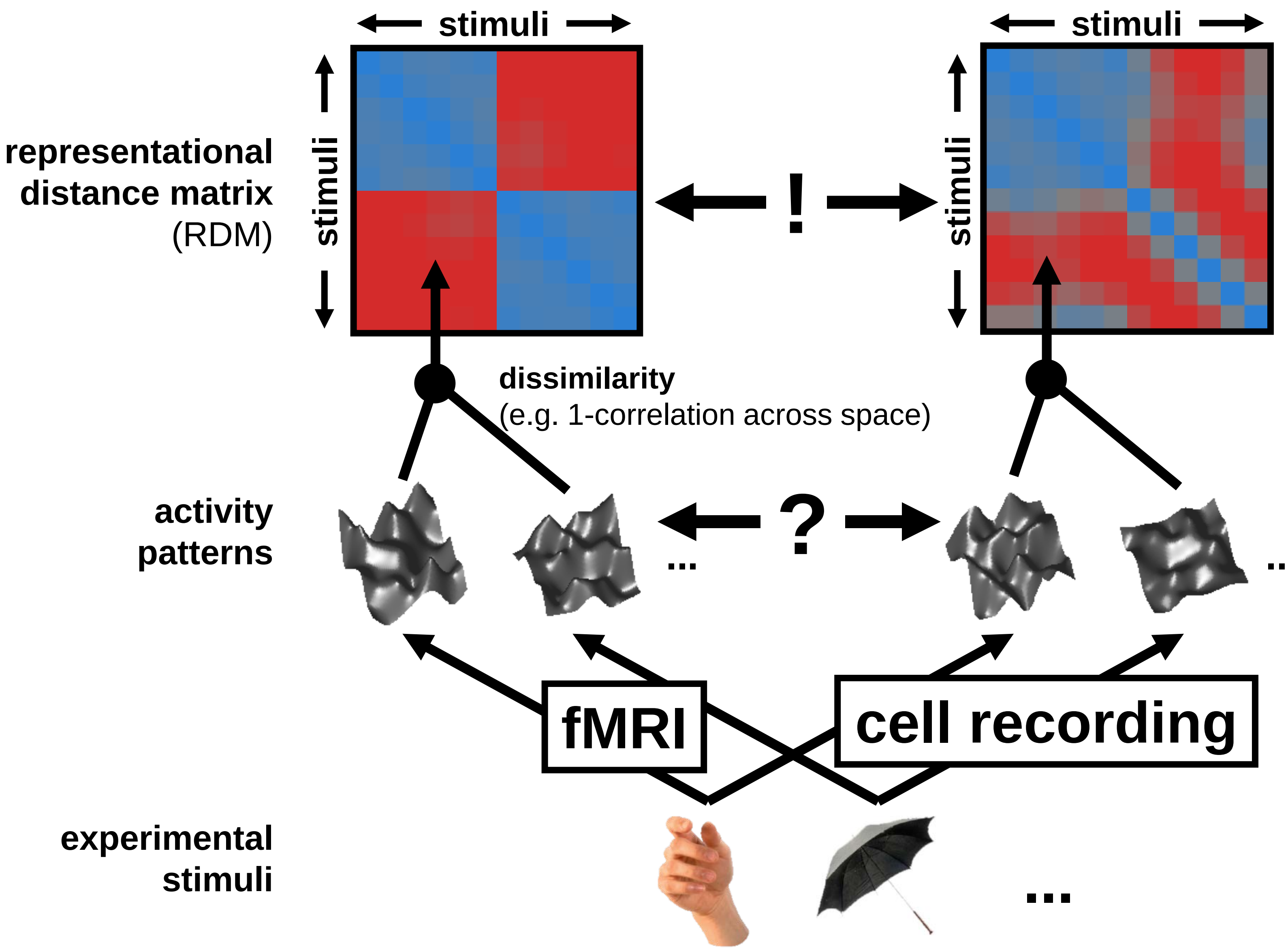
The representational similarity trick



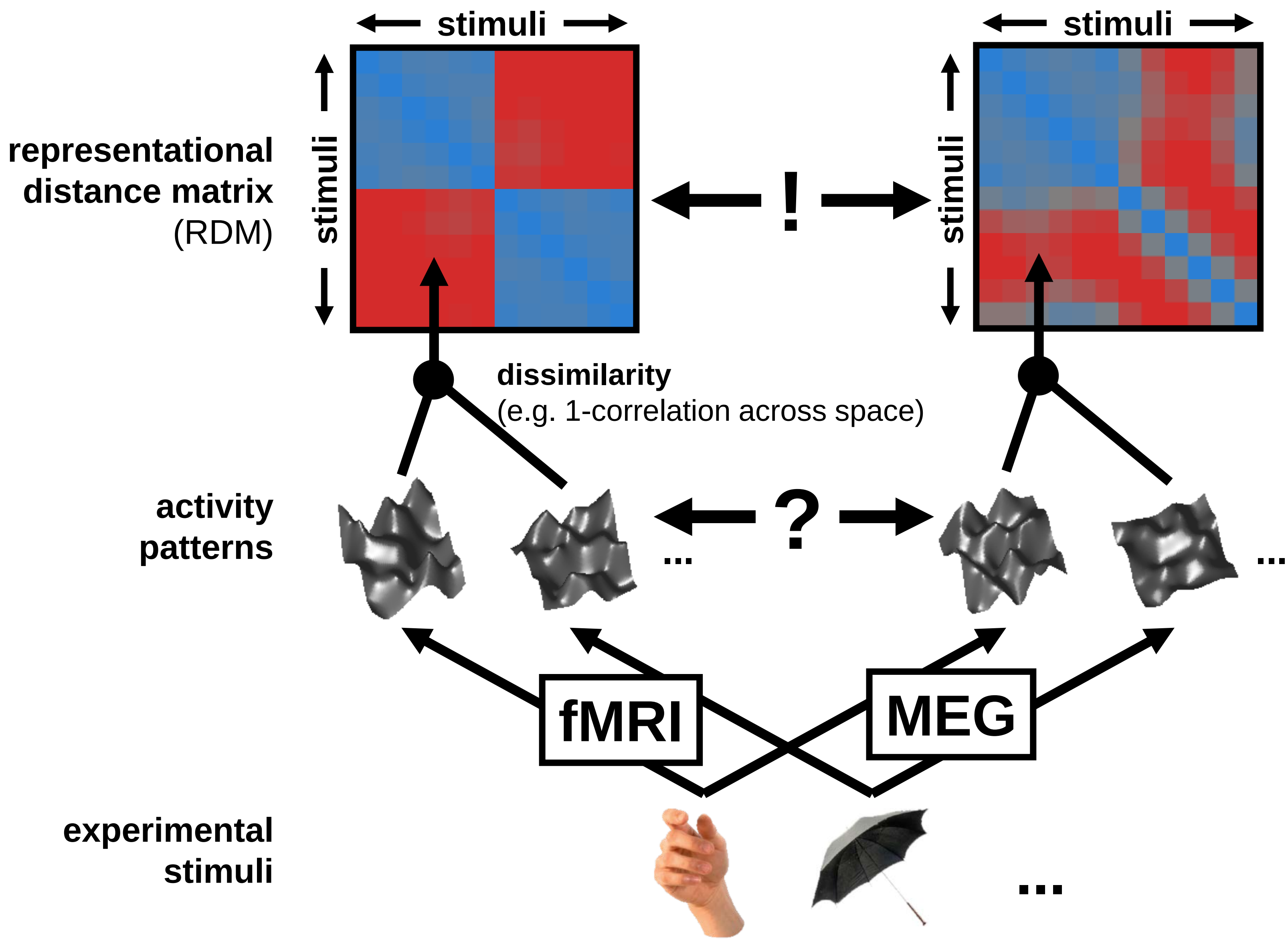
The representational similarity trick



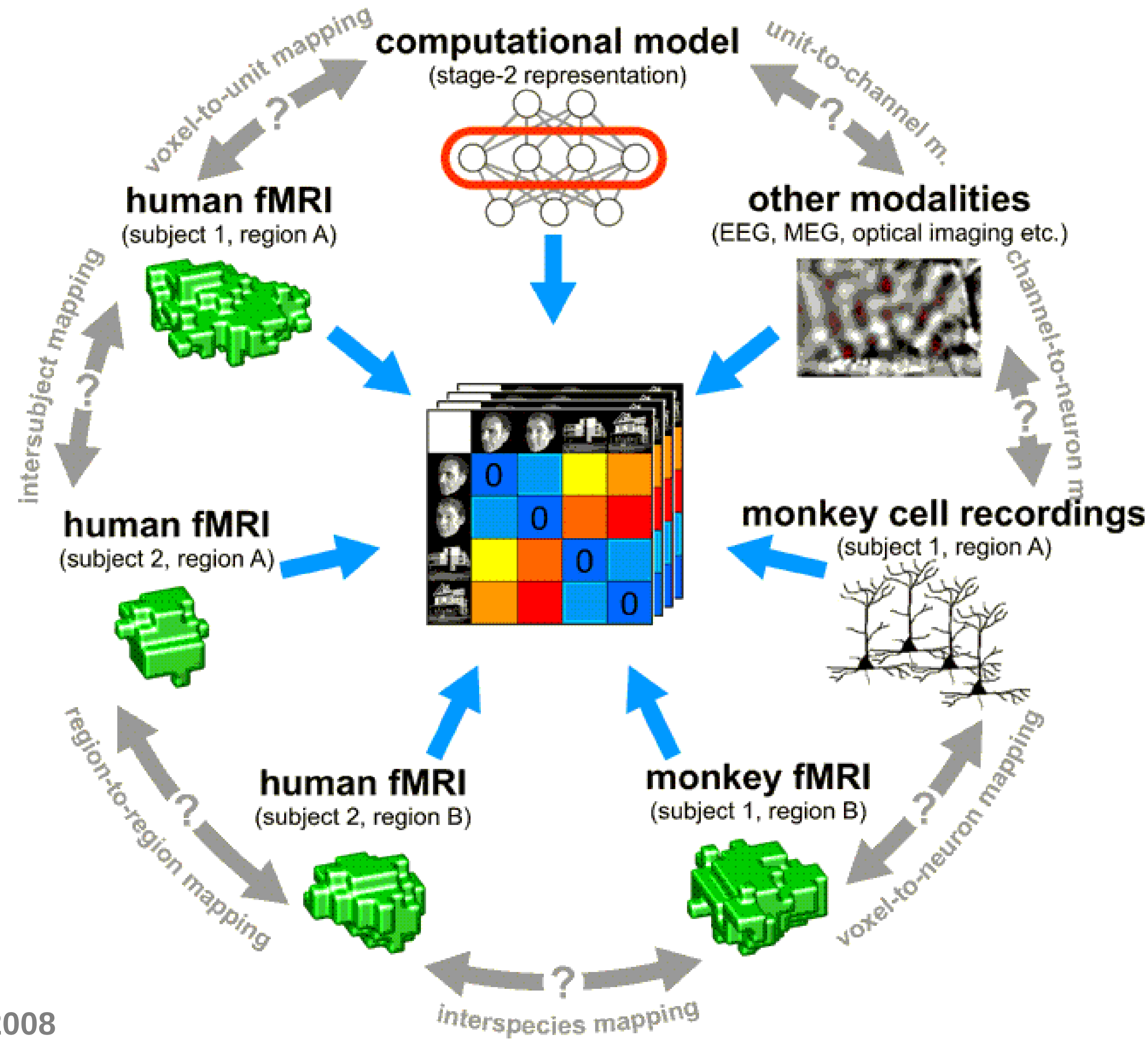
The representational similarity trick



The representational similarity trick

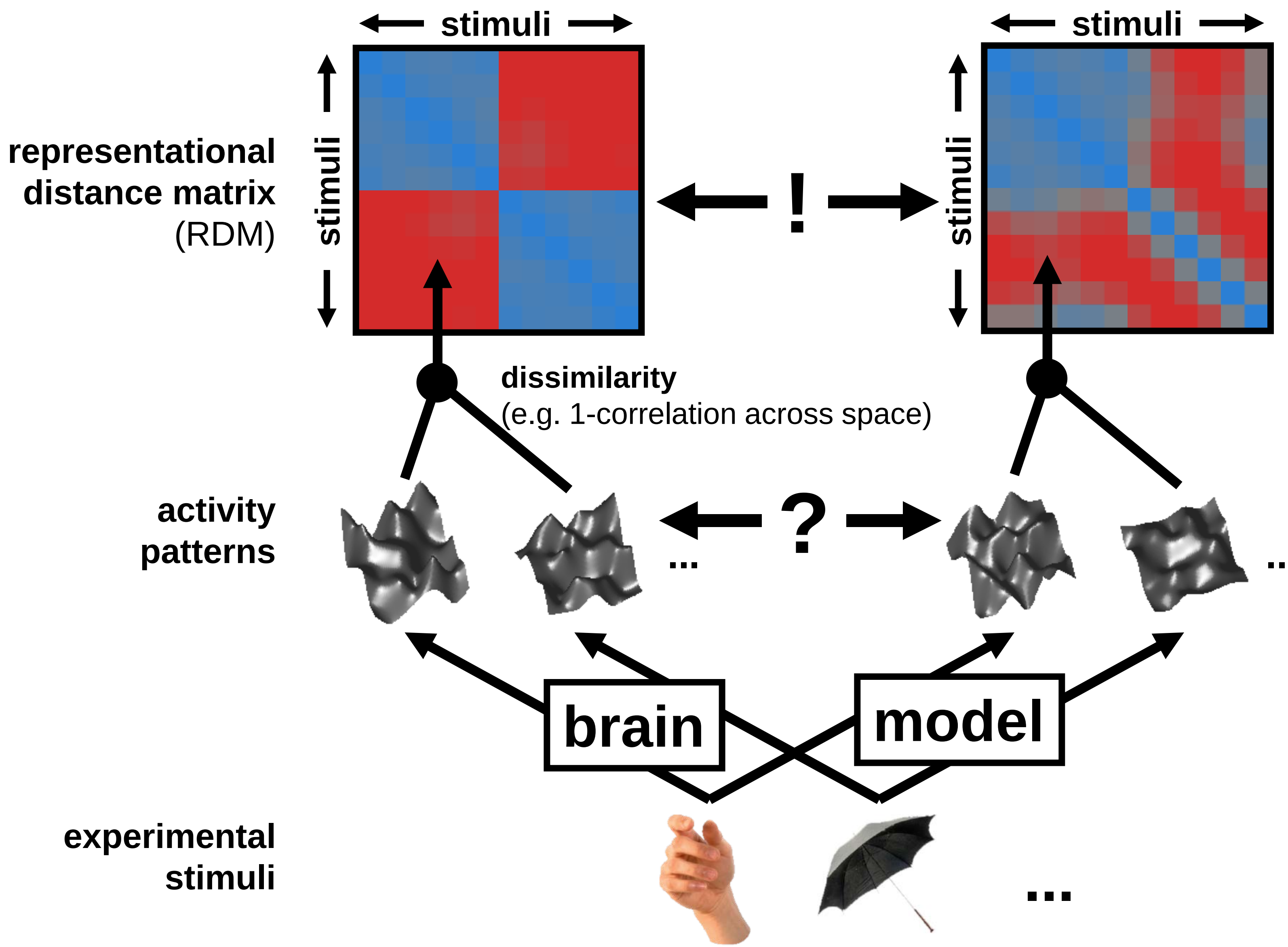


The RSA trick



Kriegeskorte, 2008

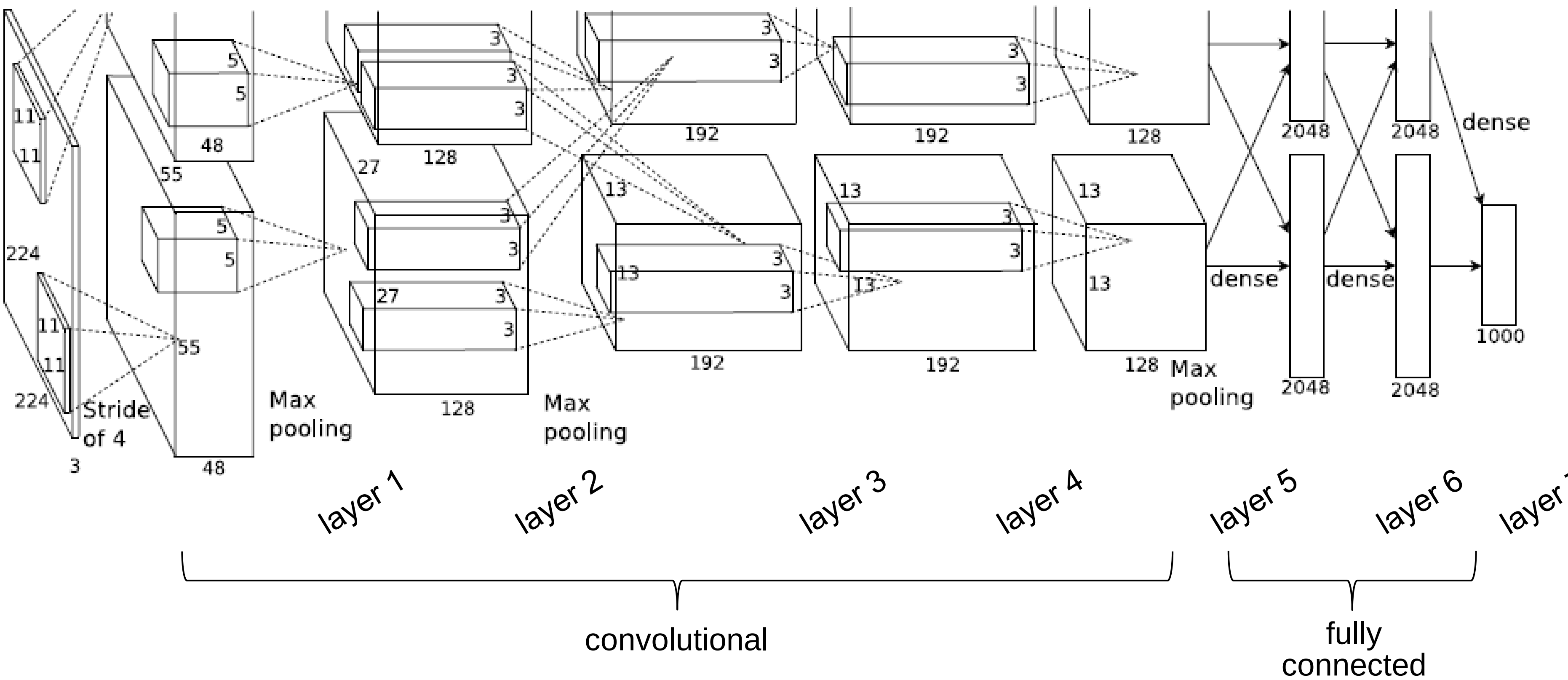
Relating brain and model RDMs



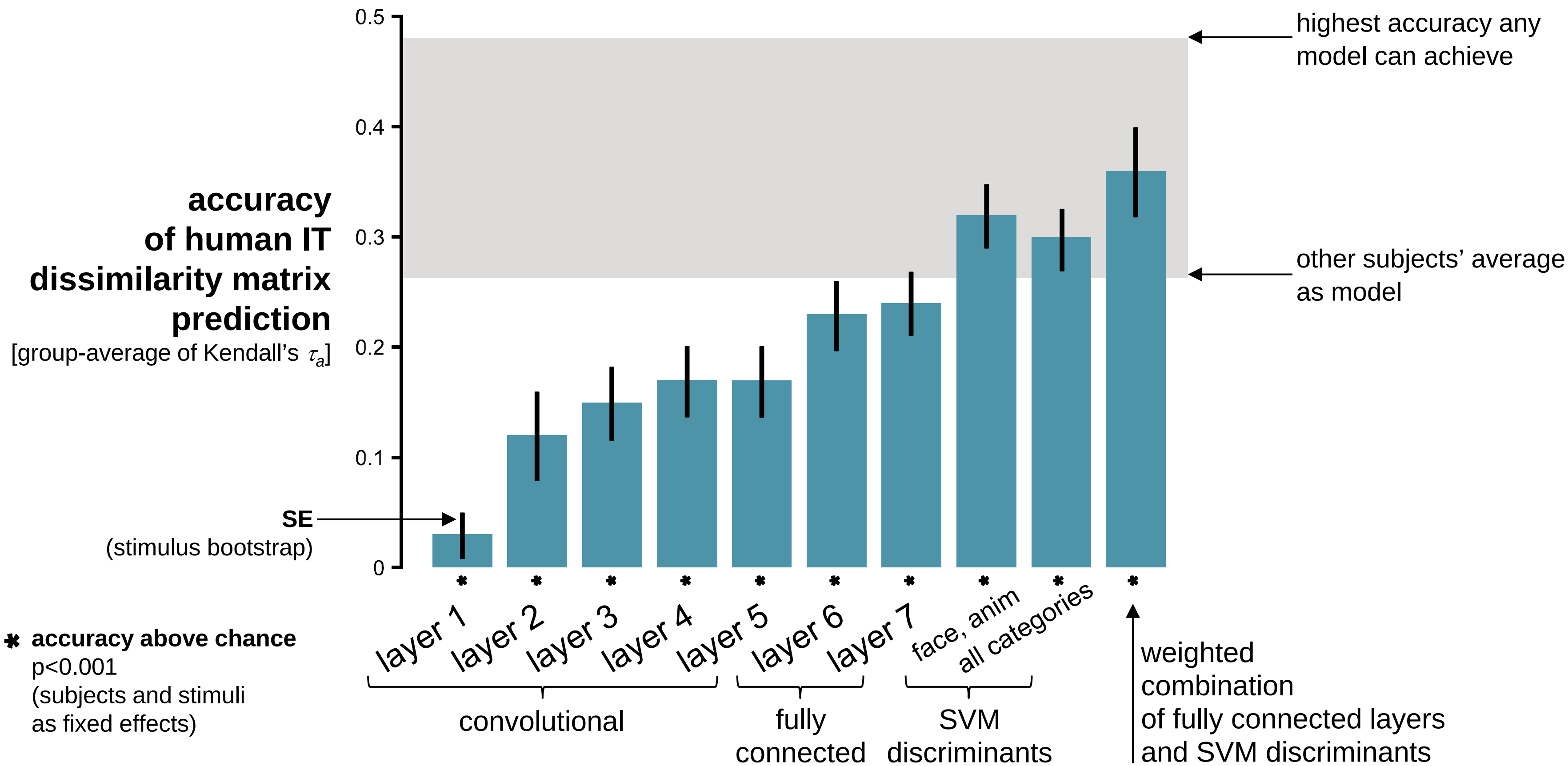
Deep convolutional neural network

- 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

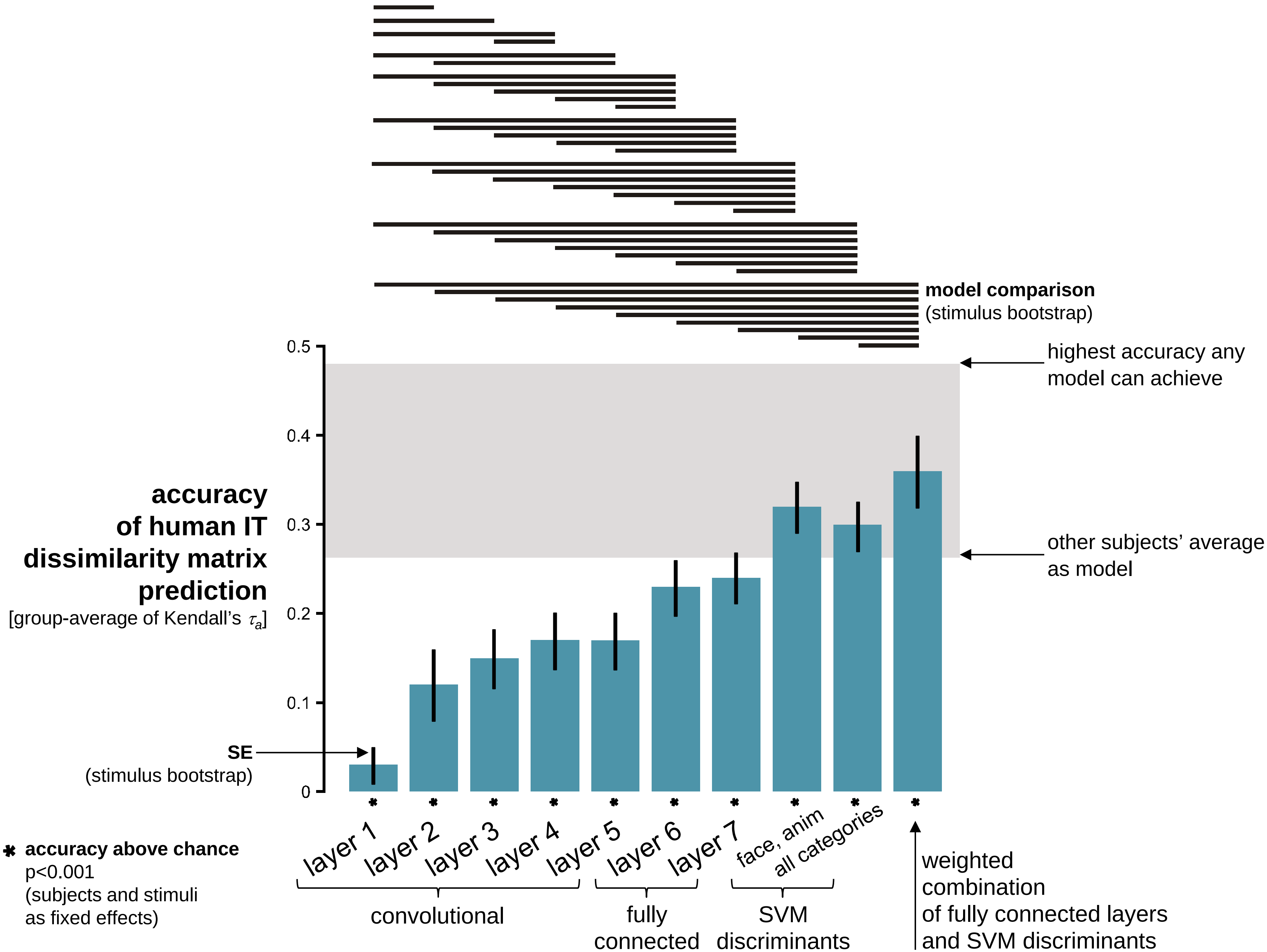
Is this network functionally similar to the brain?



Krizhevsky et al. 2012

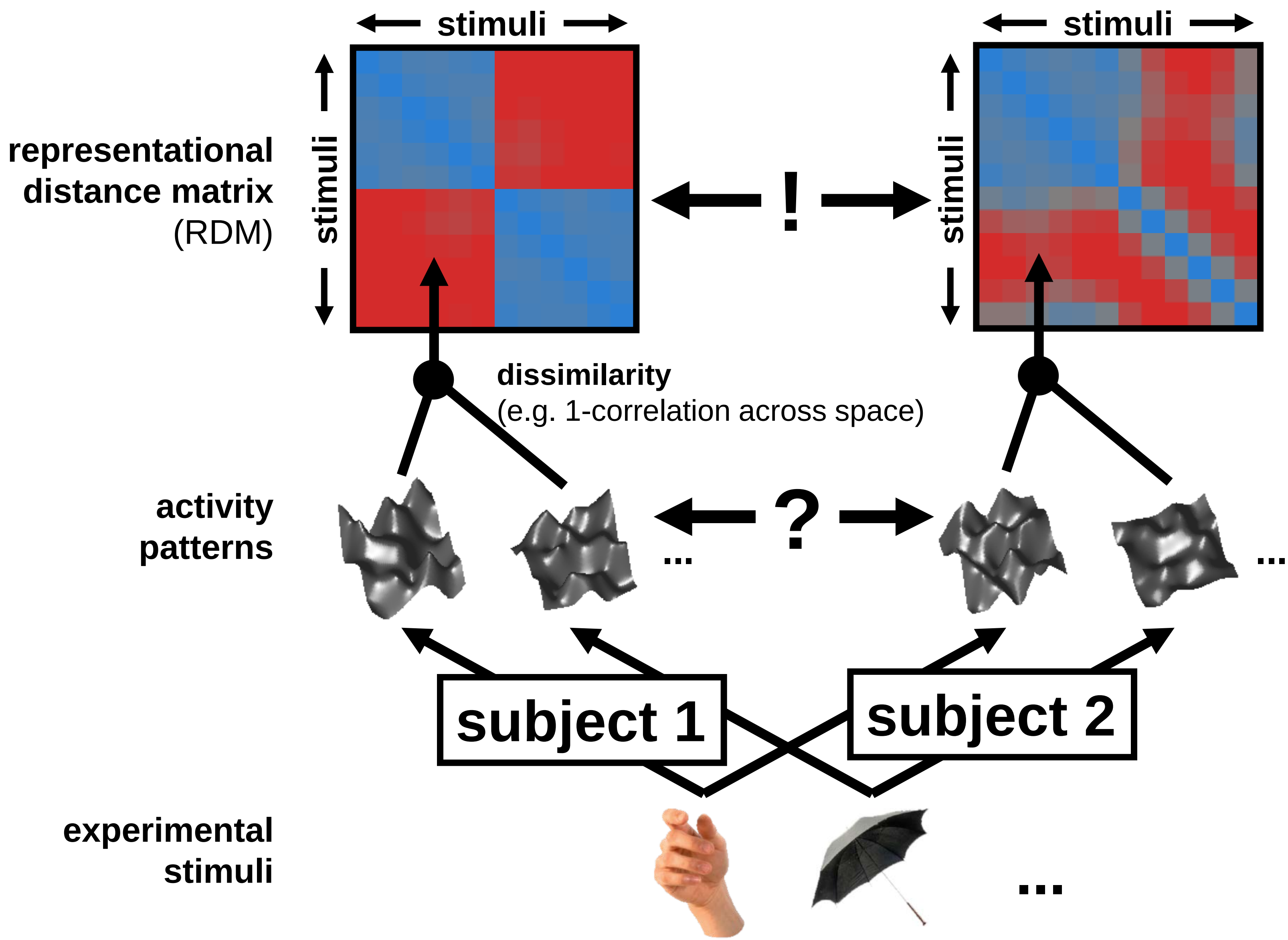


Khaligh-Razavi & Kriegeskorte (2014), **Nili** et al. 2014 (RSA Toolbox)



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

Comparing brain RDMs between people



Charest et al. 2014 *PNAS*

Stimuli

Objects from the subject's own photo-album

animate

bodies

faces

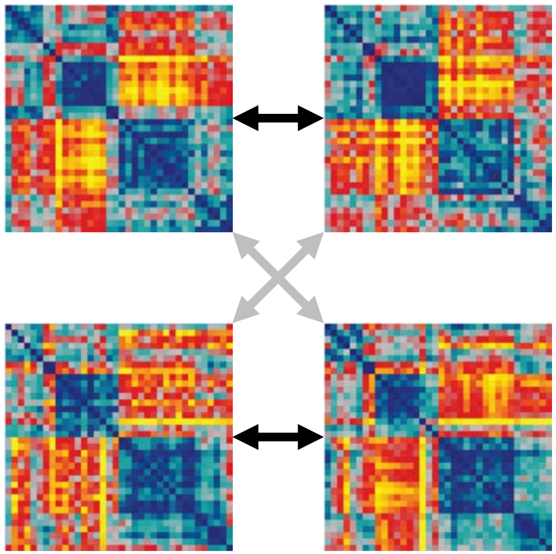
inanimate

places

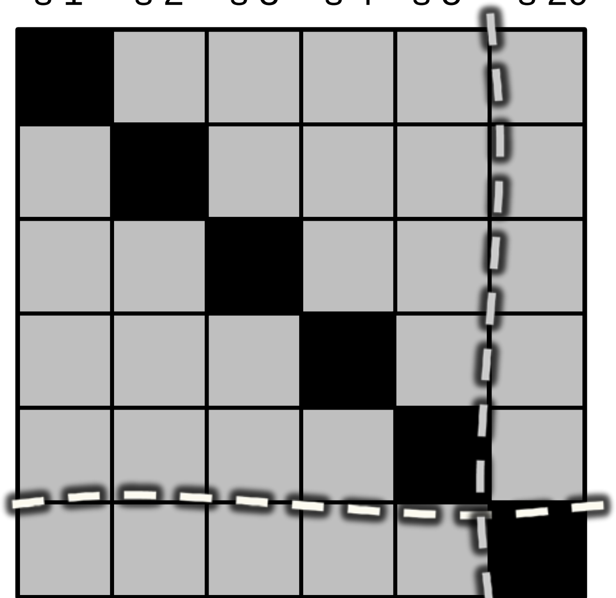
objects

Charest et al. 2014 *PNAS*

Comparing brain RDMs between people



subject similarity matrix



correlation

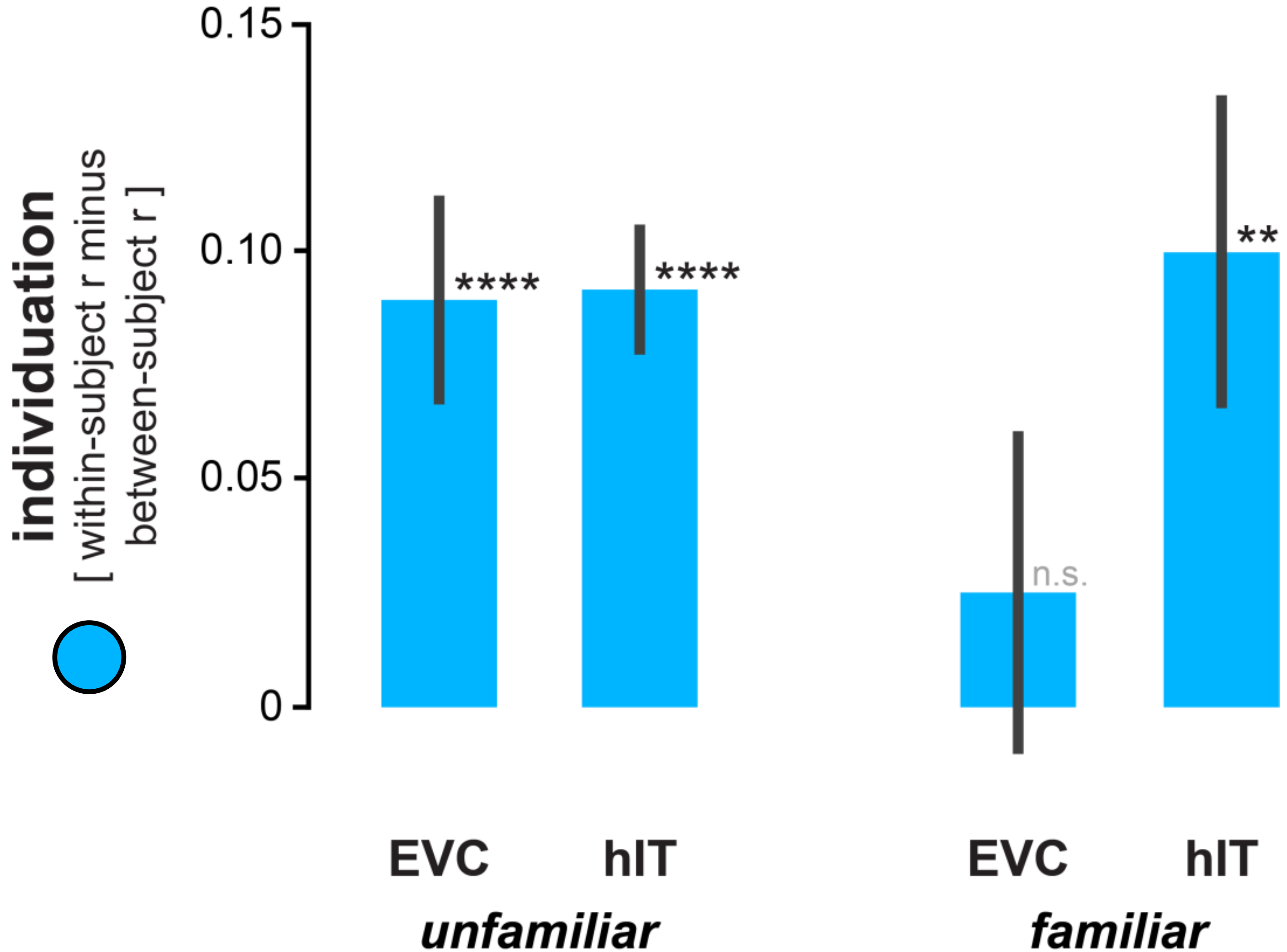
within-subject (ws) ✔

between-subject (bs) ✔

individuation index (ws - bs) ?

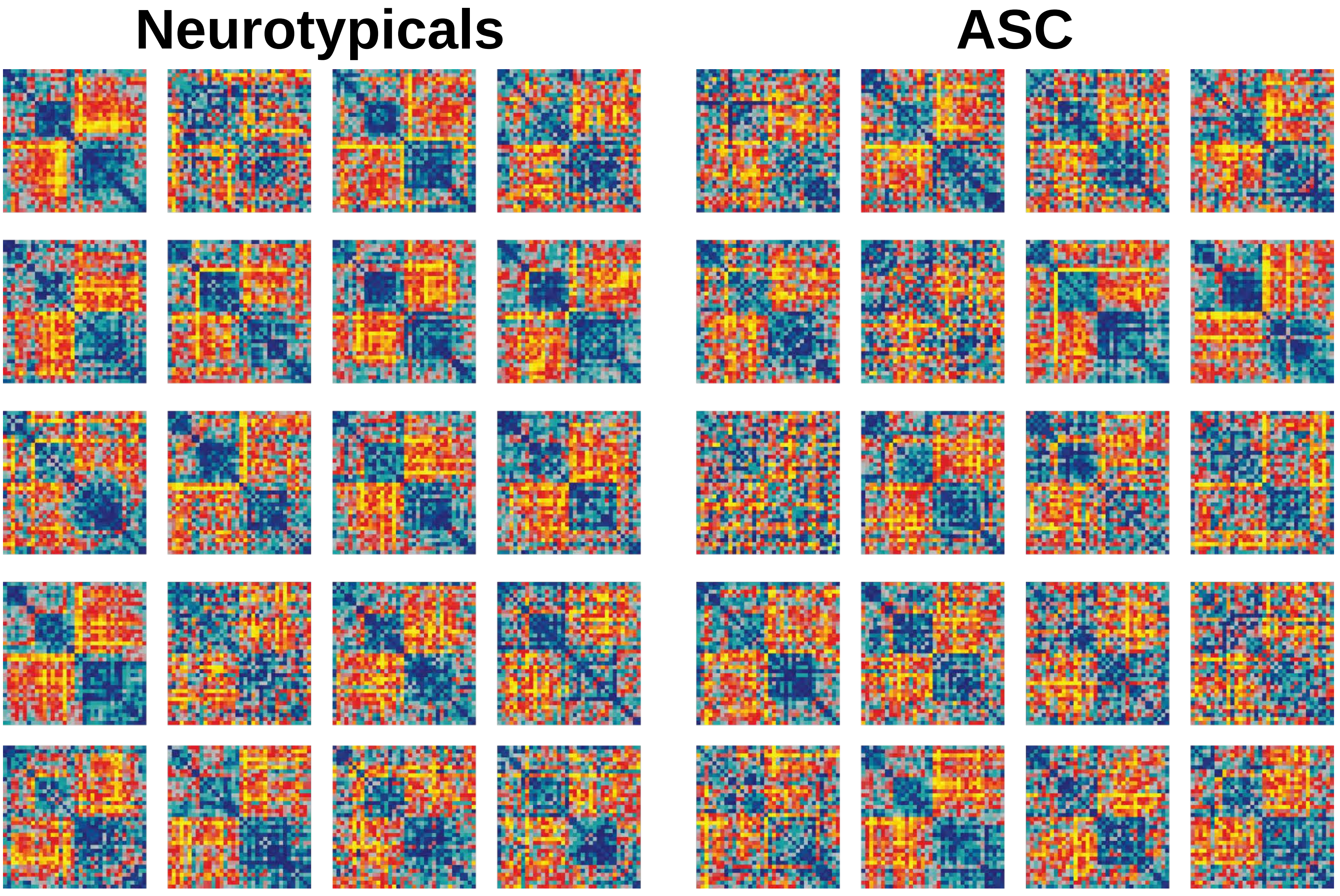
Charest et al. 2014 *PNAS*

Brain representations unique?

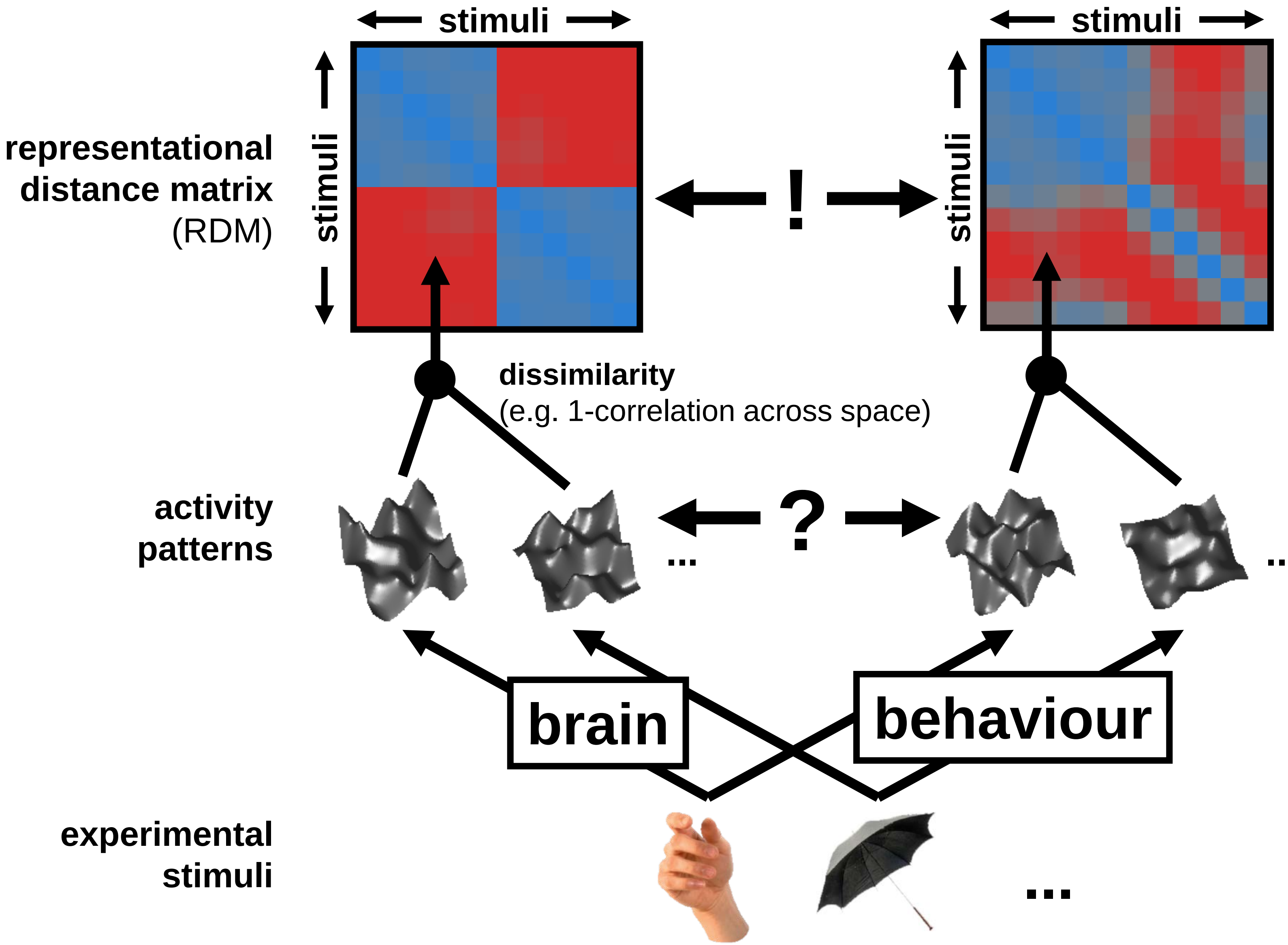


Charest et al. 2014 *PNAS*

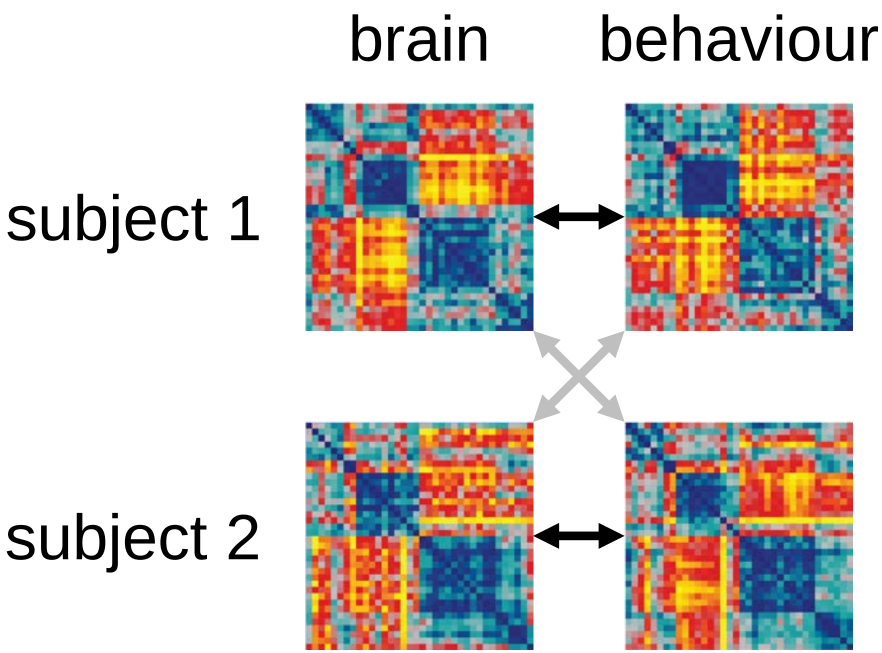
Representational geometries in human inferior temporal cortex



Comparing brain RDMs and behavioural RDMs



Comparing brain RDMs and behavioural RDMs



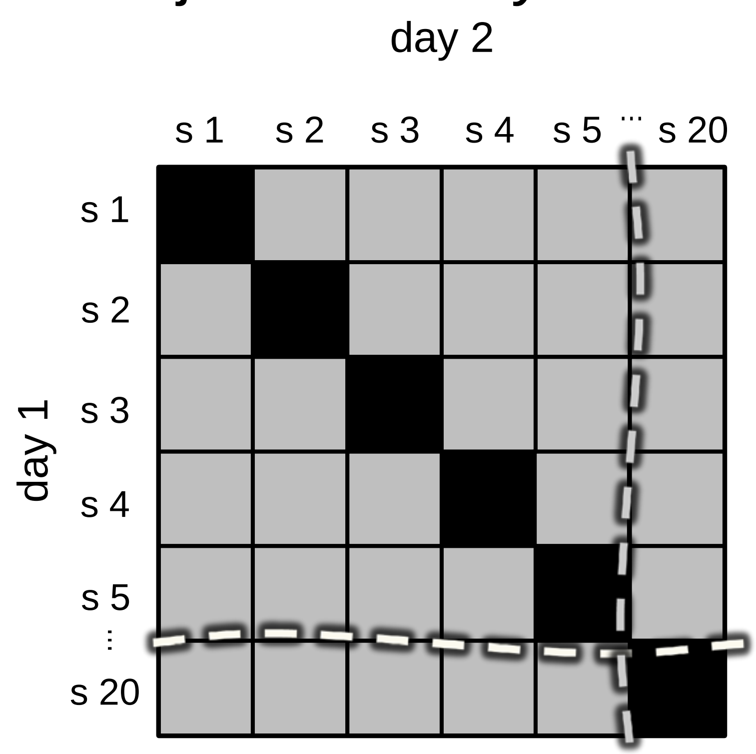
correlation

within-subject (ws) ✔

between-subject (bs) ✔

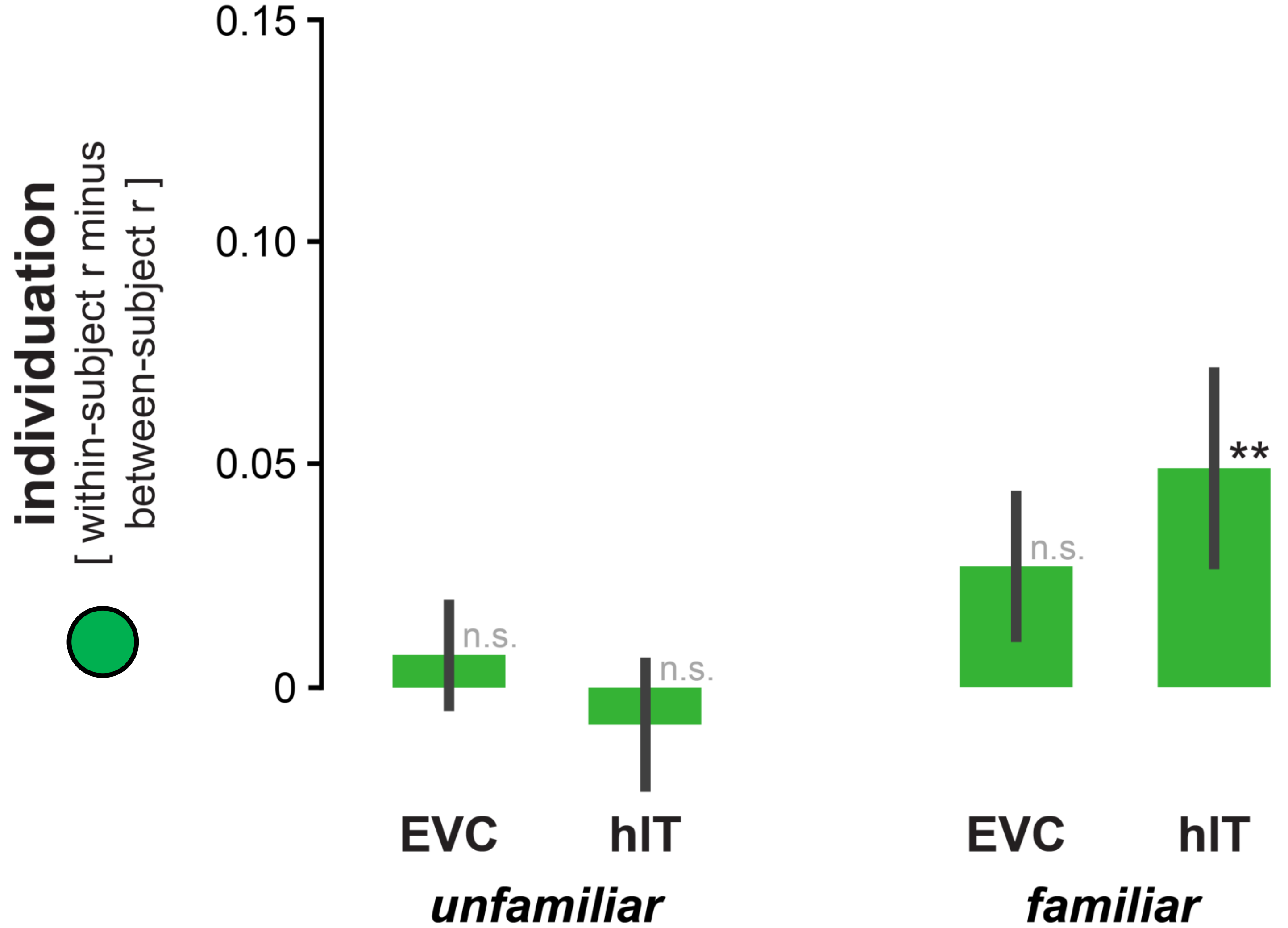
individuation index (ws - bs) ?

subject similarity matrix



Charest et al. 2014 *PNAS*

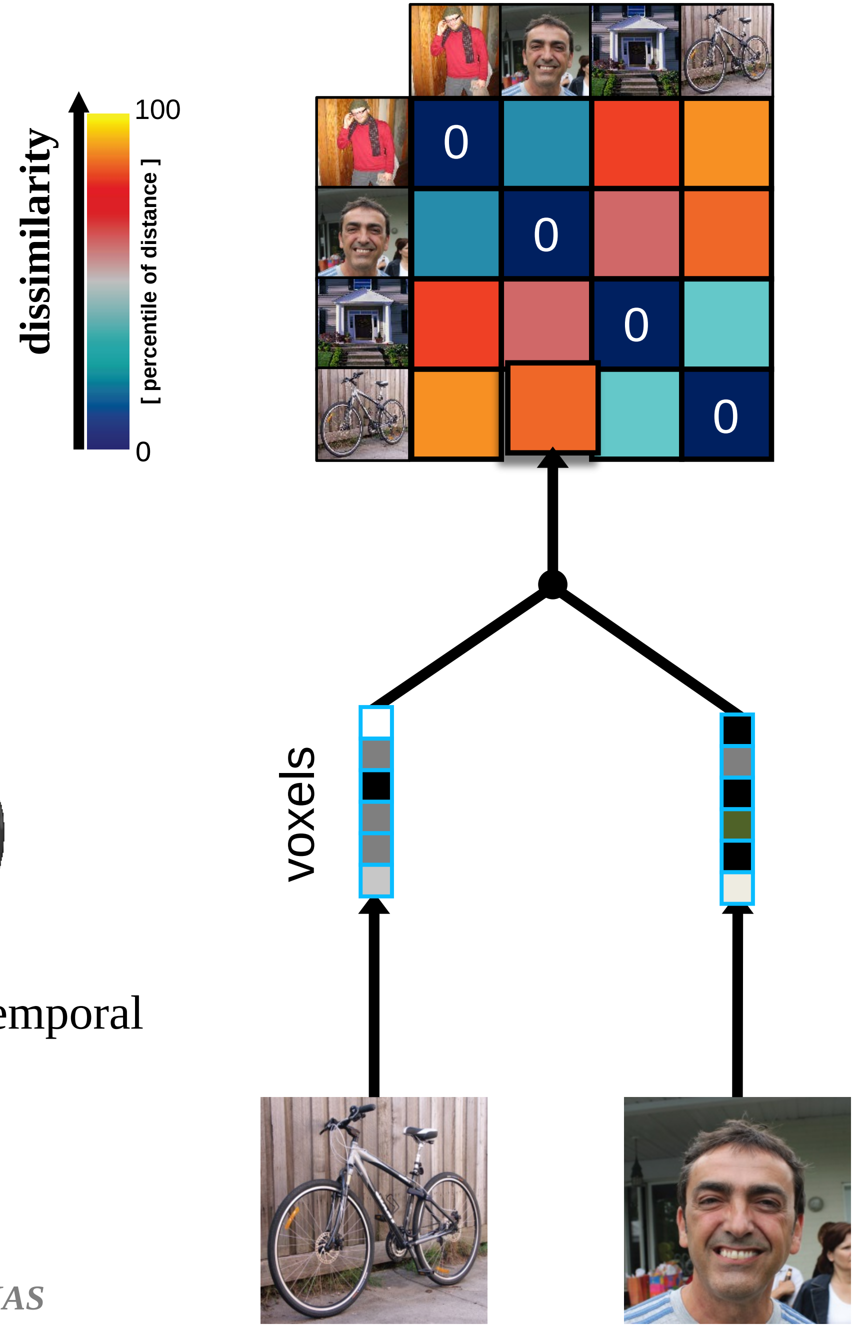
Brain-behavior relationship unique?



Charest et al. 2014 *PNAS*

RSA

Representational Dissimilarity Matrix (RDM)



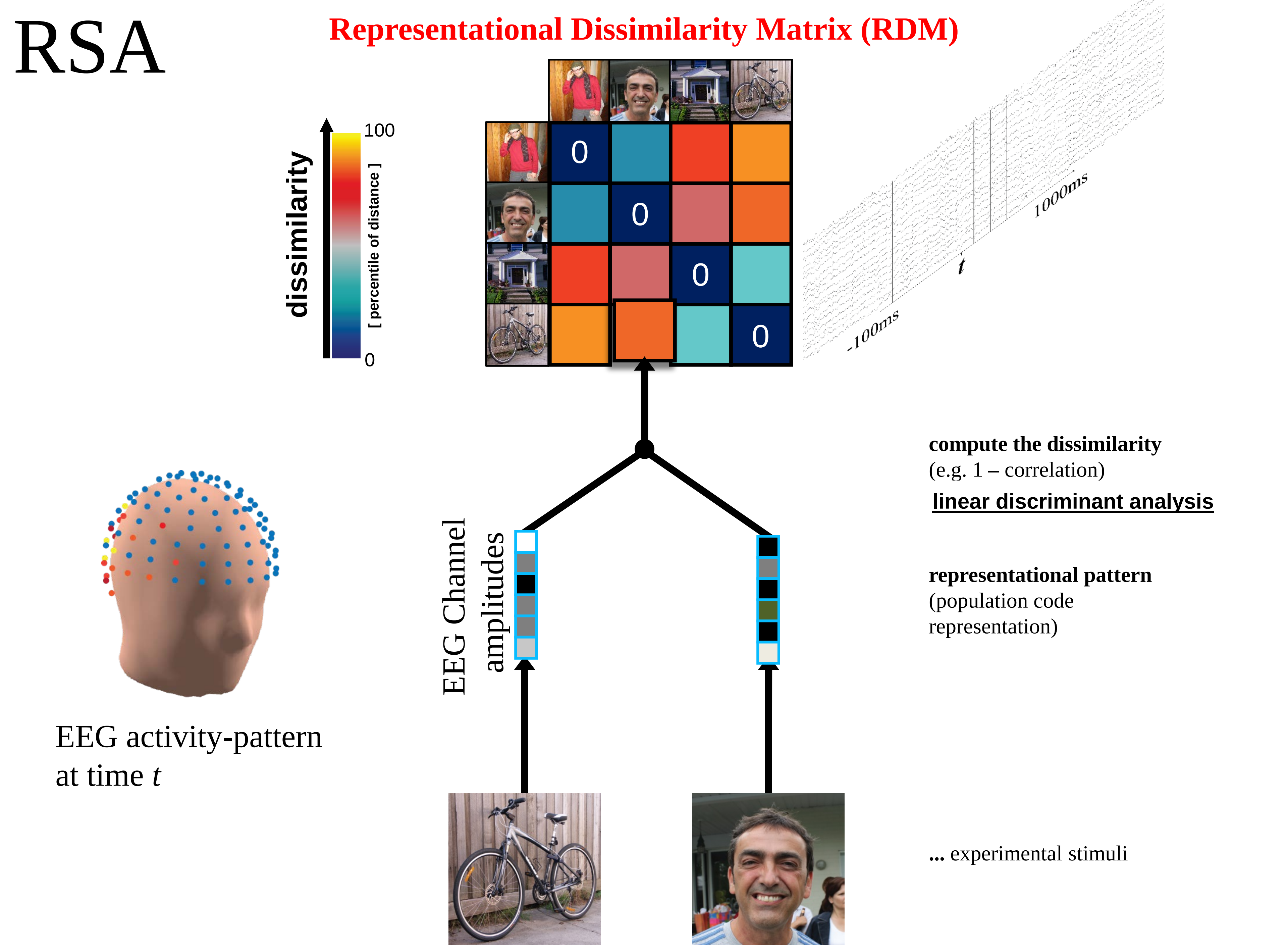
compute the dissimilarity
(e.g. 1 – correlation)

representational pattern
(population code representation)

... experimental stimuli

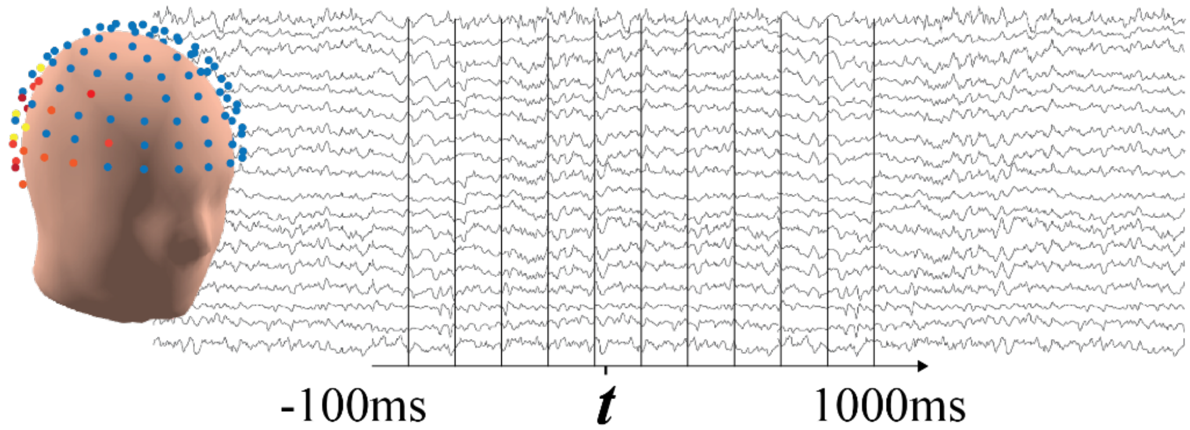
human inferior temporal (hIT)

Charest et al. 2014 *PNAS*


RSA
Representational Dissimilarity Matrix (RDM)
dissimilarity
100
[percentile of distance]
0
0
0
0
0
-100ms
t
1000ms
compute the dissimilarity
(e.g. 1 – correlation)
linear discriminant analysis
EEG Channel
amplitudes
representational pattern
(population code
representation)
EEG activity-pattern
at time t
... experimental stimuli

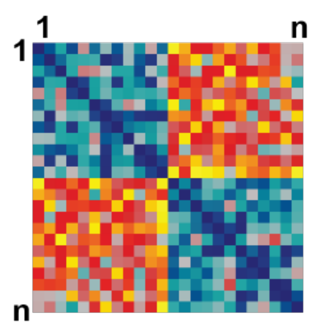
RSA

EEG Sensor
Activation
Patterns

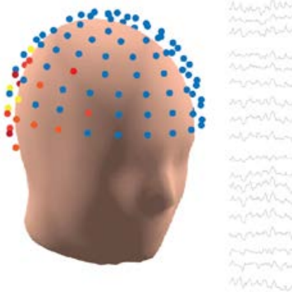


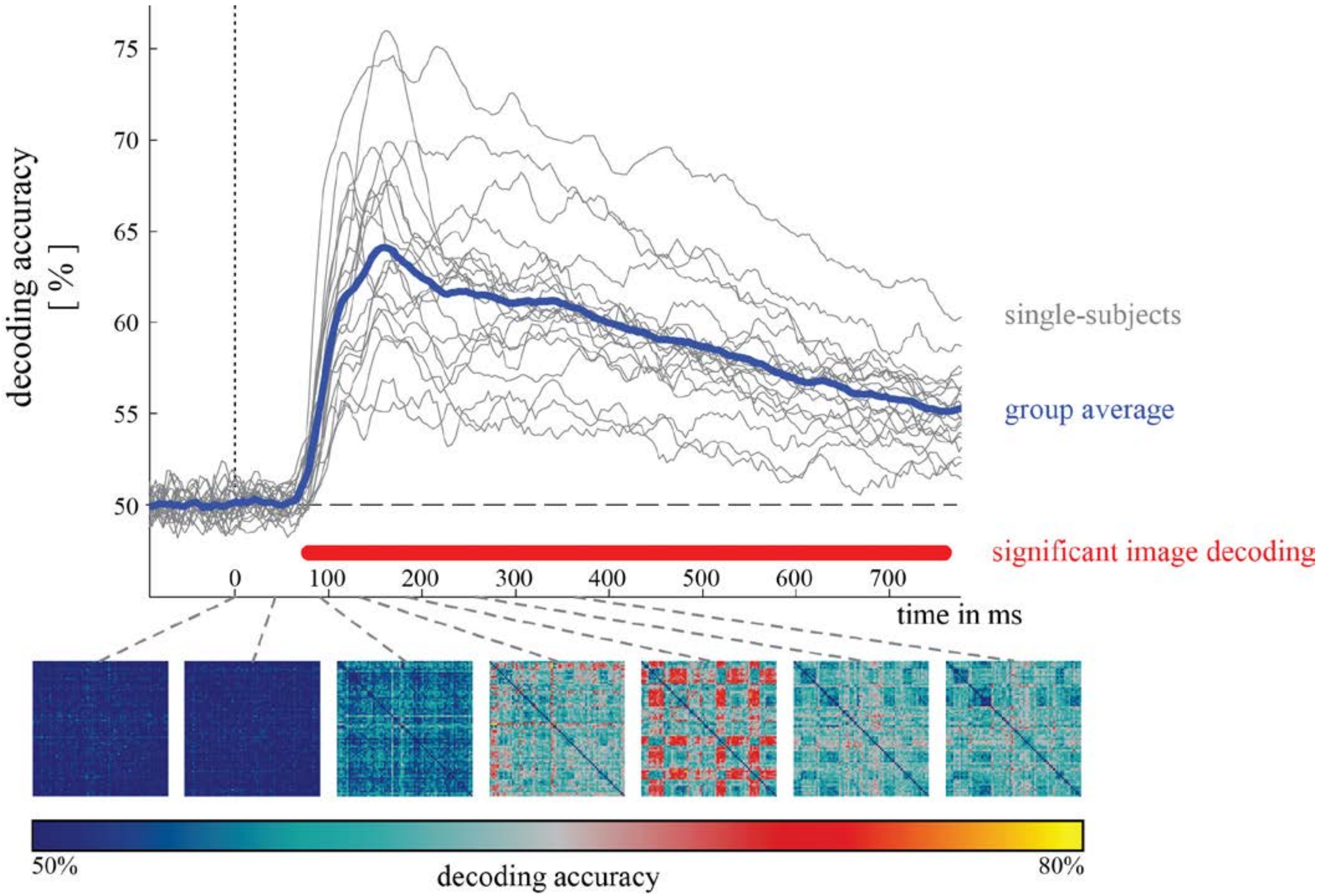
decode pair-wise activation
patterns for objects 1:n

representational
dissimilarity matrix
at time *t*

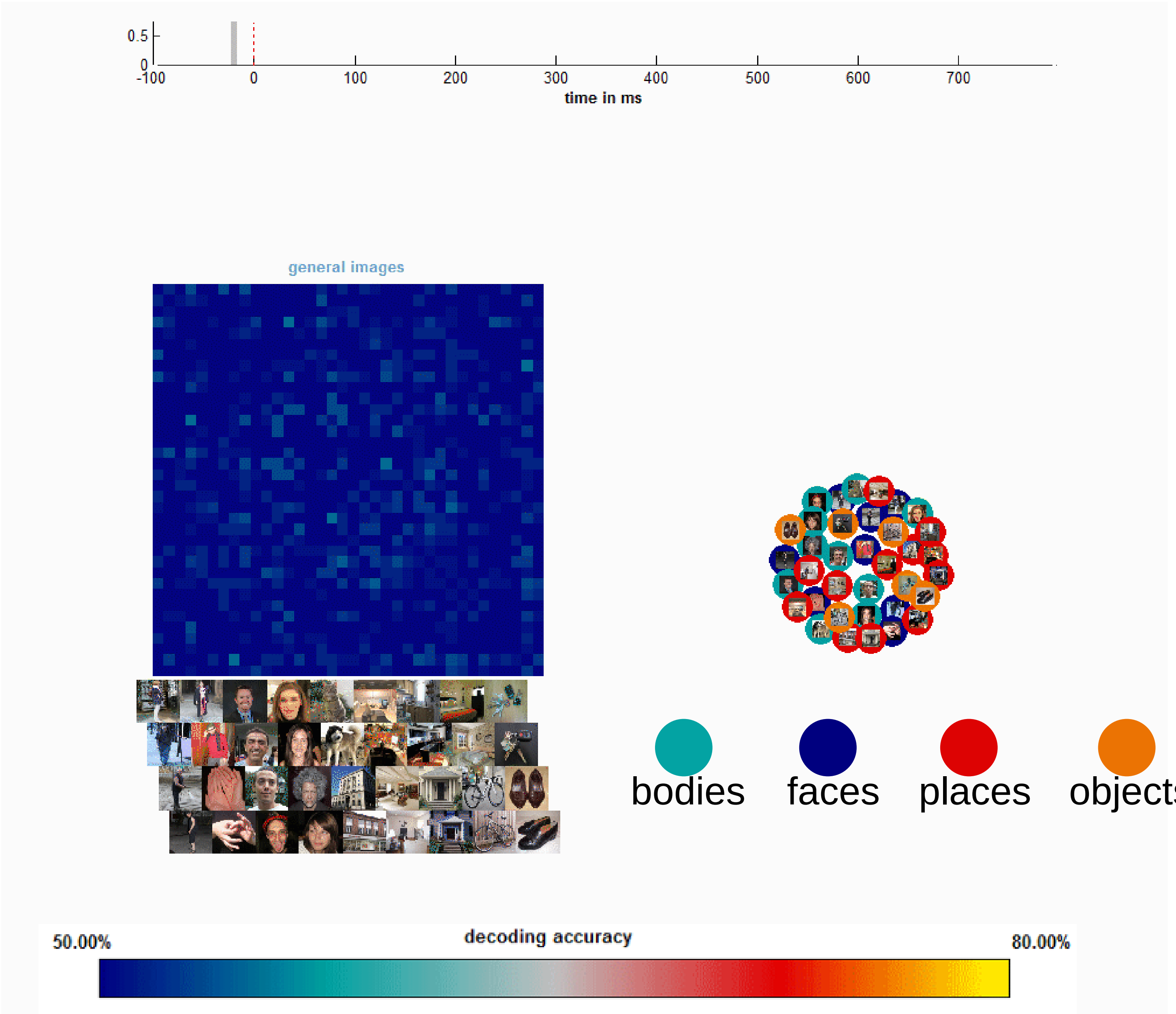


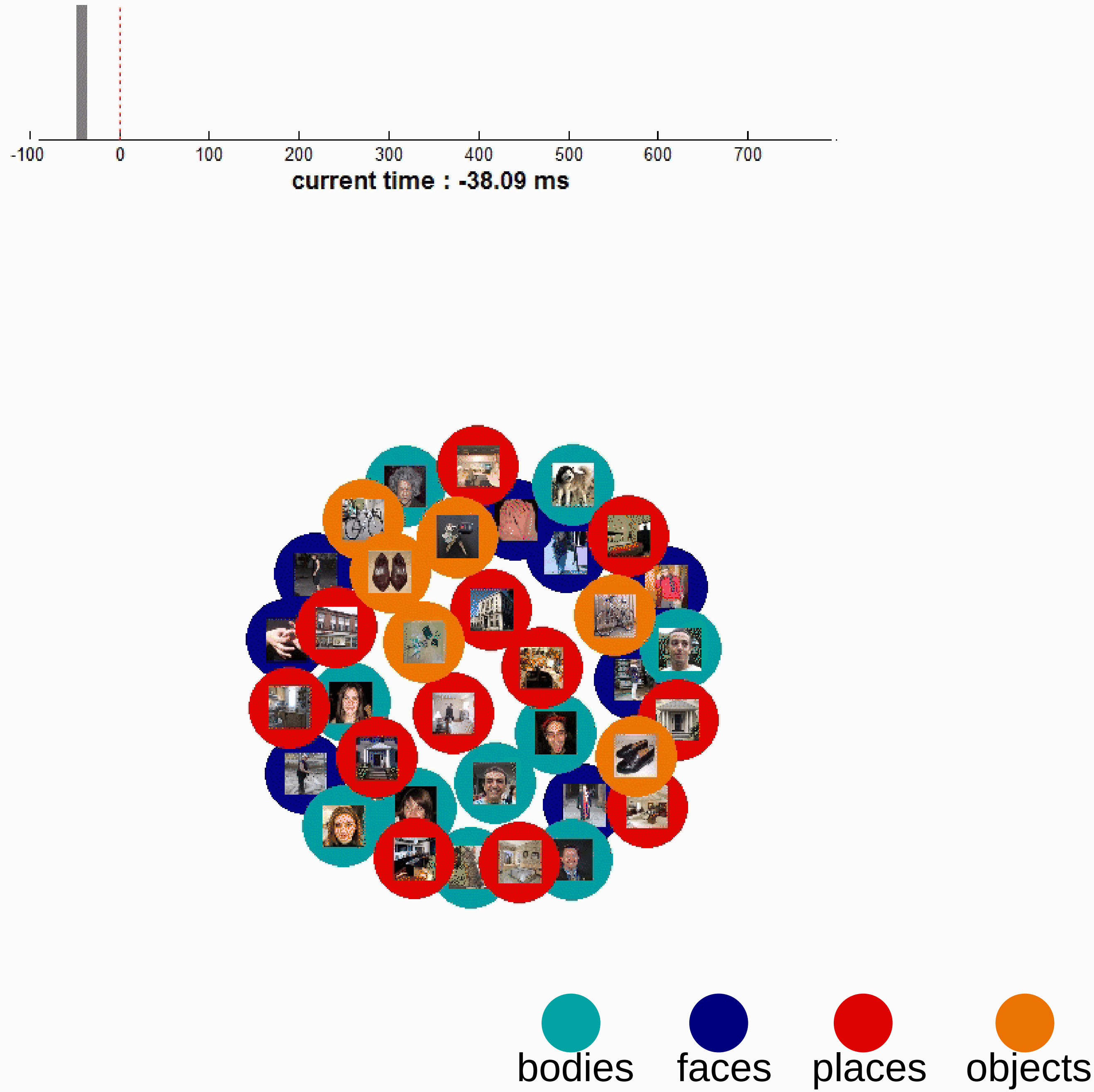
EEG contains rich topographic information from which you can distinguish mental states





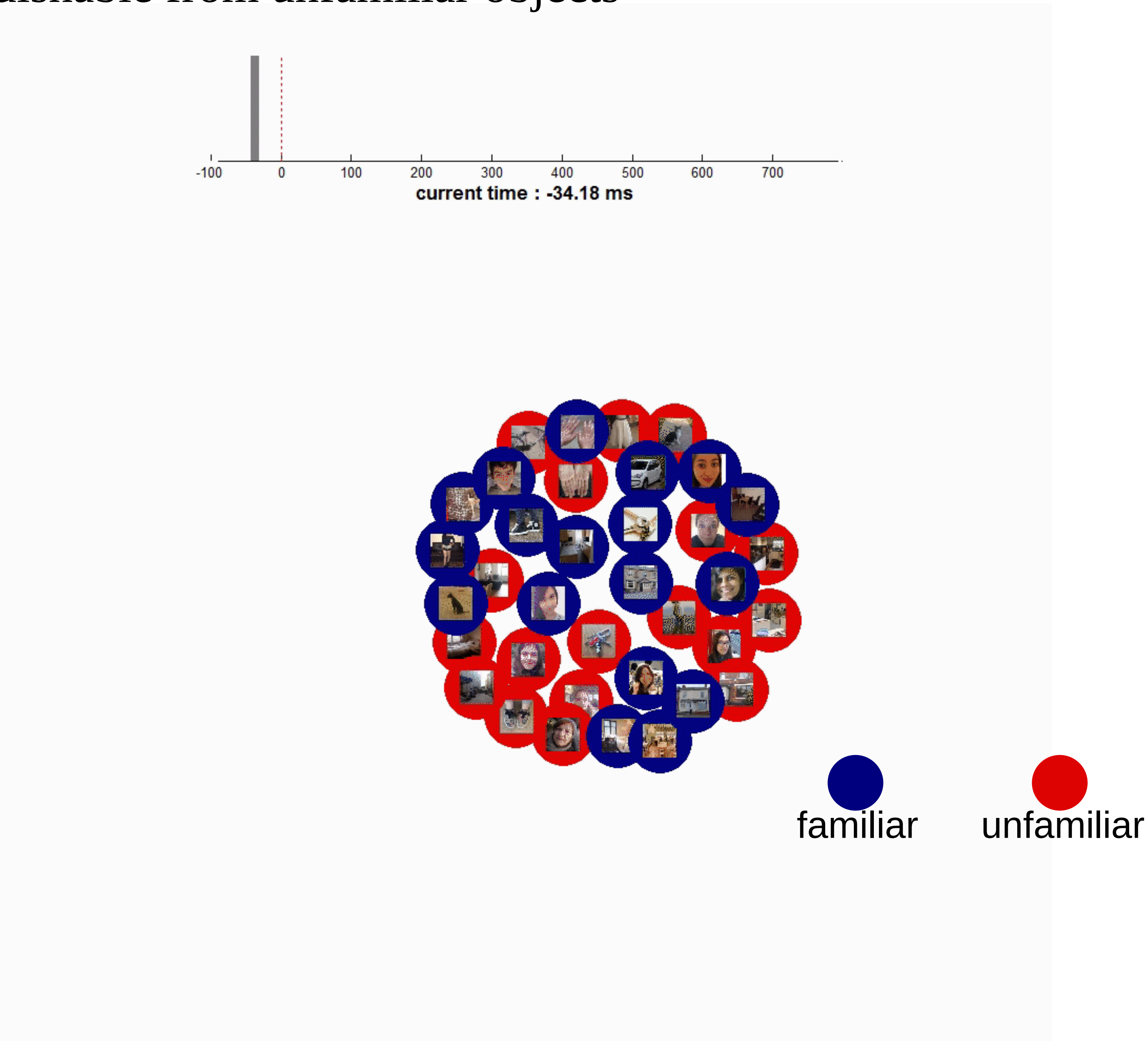
EEG contains rich topographic information from which you can distinguish mental states



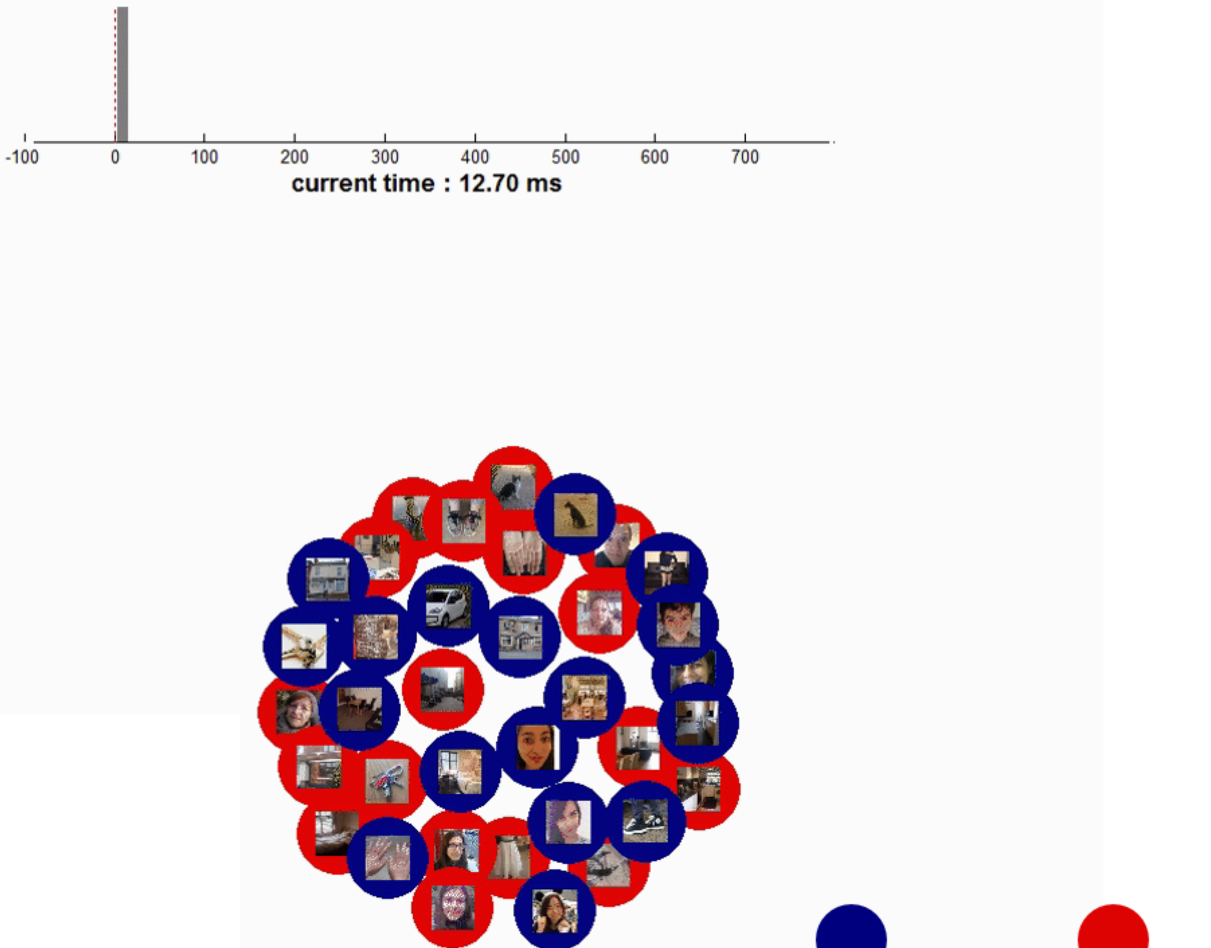


-100
0
100
200
300
400
500
600
700
current time : -38.09 ms
bodies
faces
places
objects

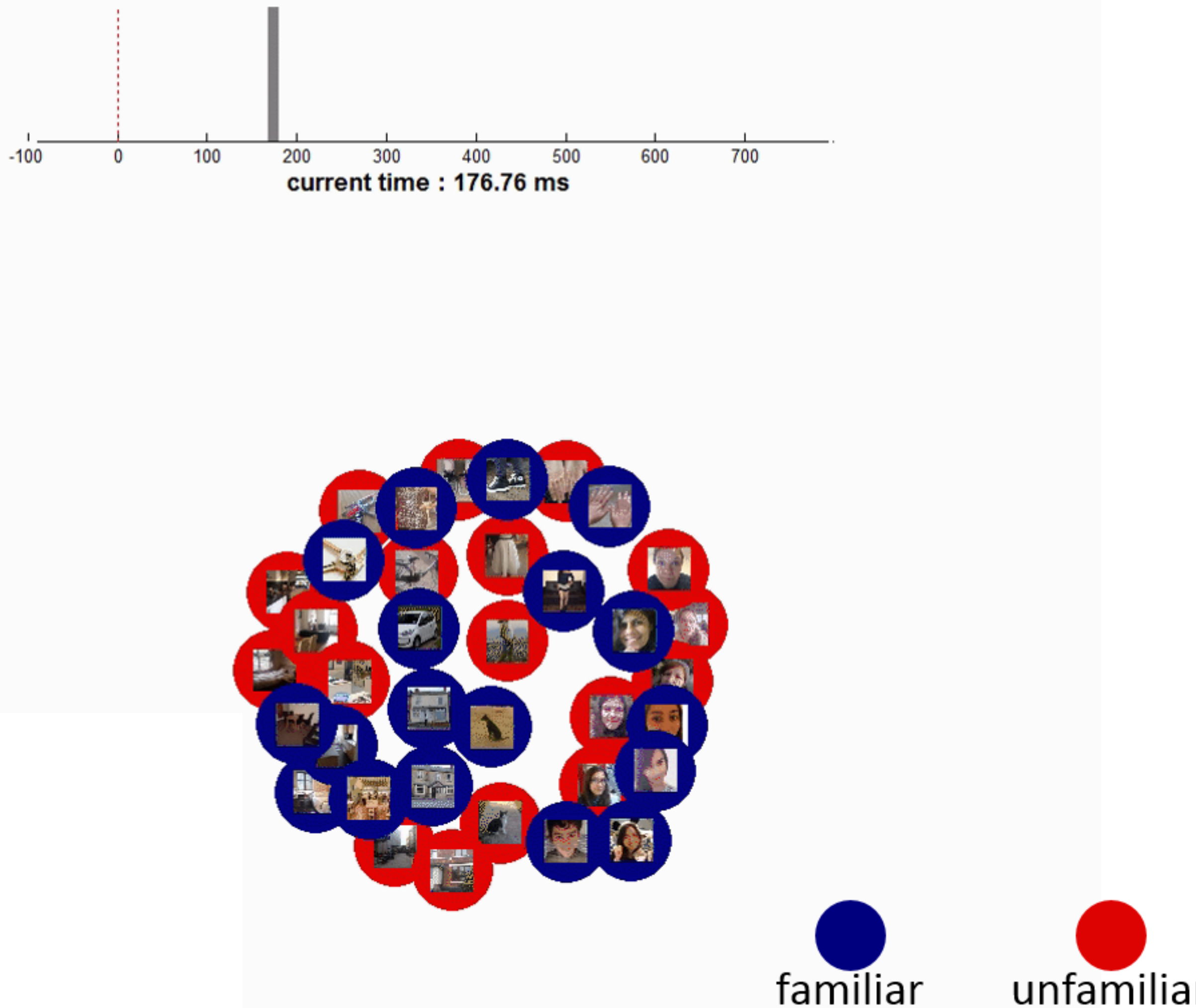
Personally meaningful objects elicit activity patterns that are distinguishable from unfamiliar objects



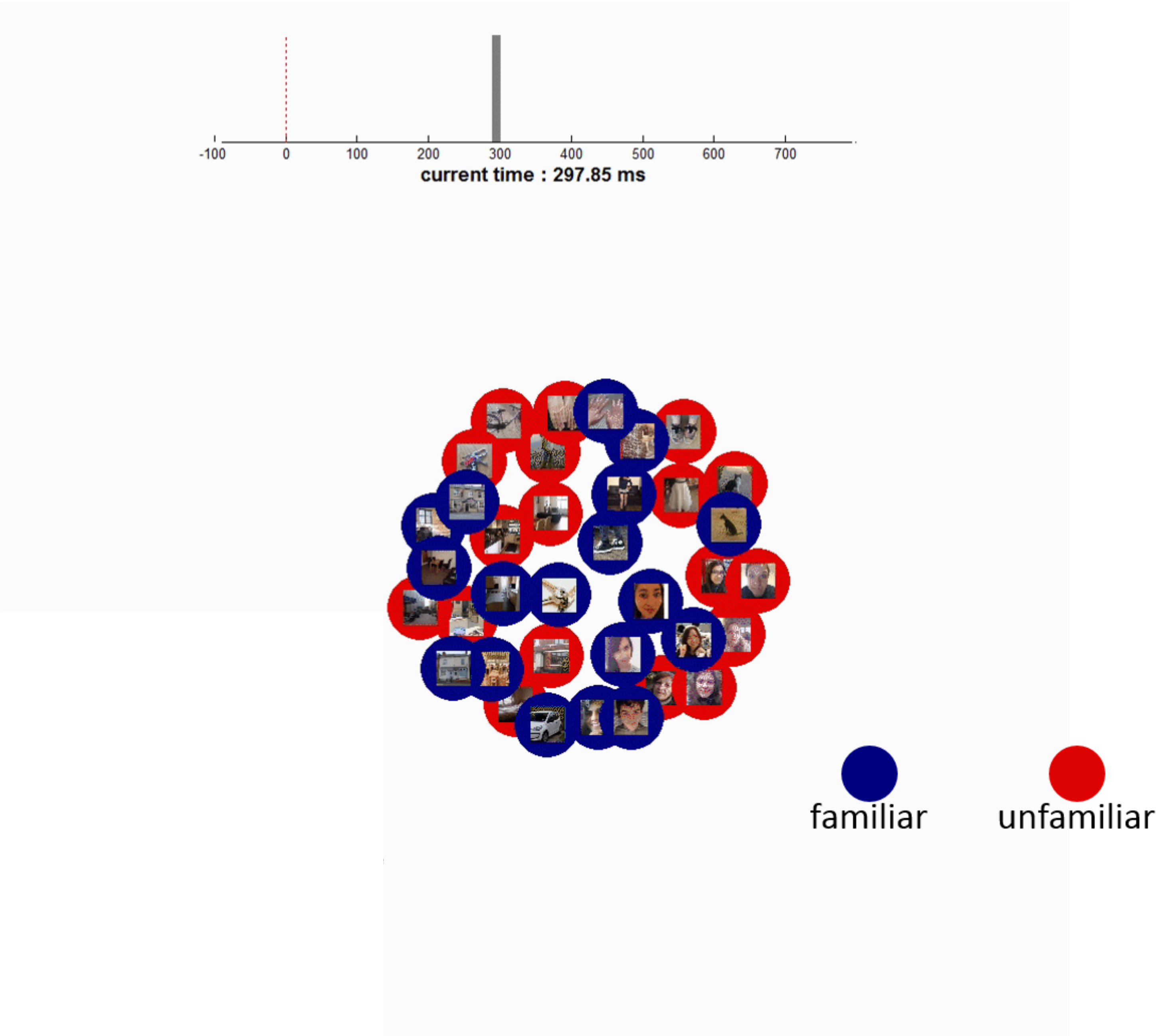
Personally meaningful objects elicit activity patterns that are distinguishable from unfamiliar objects



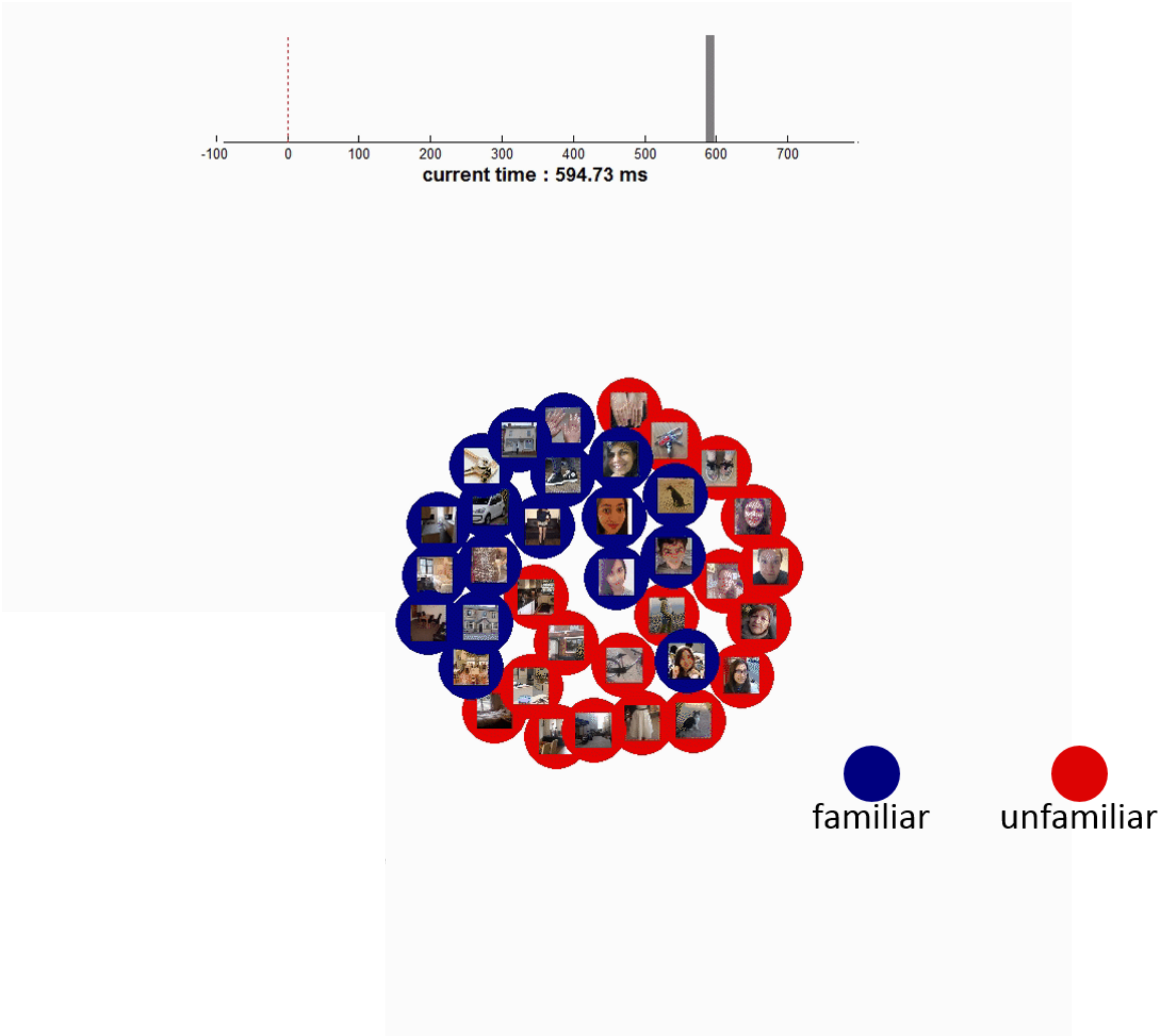
Personally meaningful objects elicit activity patterns that are distinguishable from unfamiliar objects



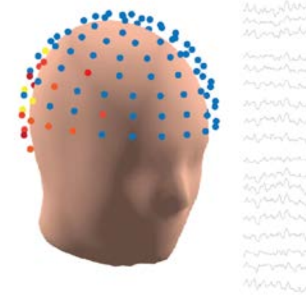
Personally meaningful objects elicit activity patterns that are distinguishable from unfamiliar objects

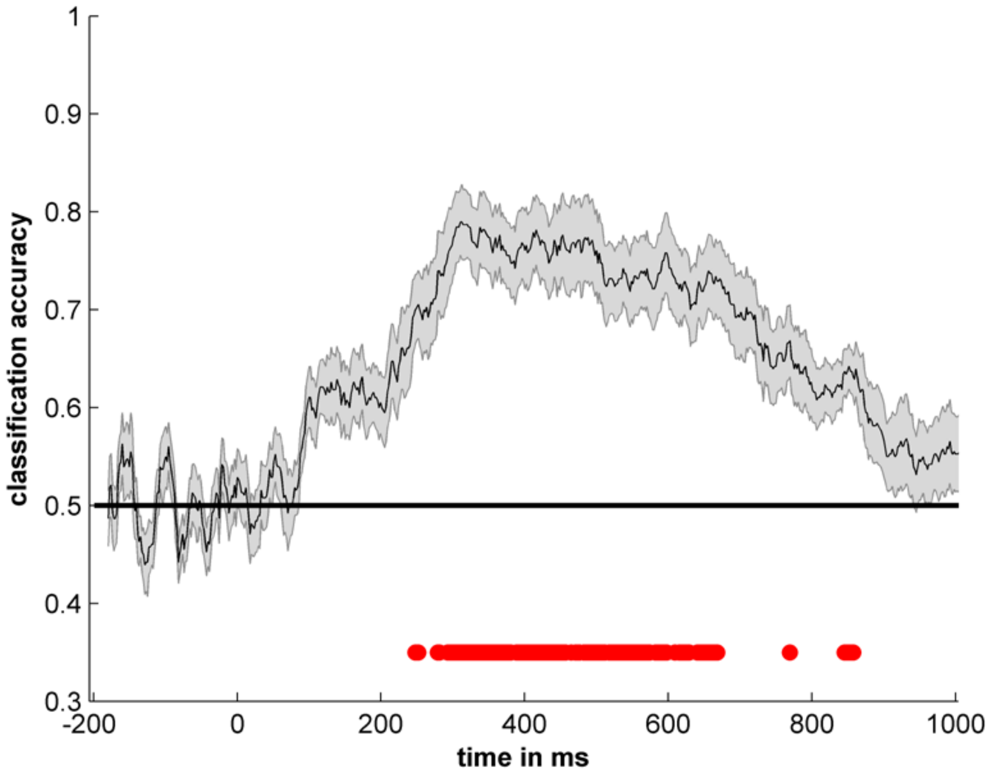


Personally meaningful objects elicit activity patterns that are distinguishable from unfamiliar objects



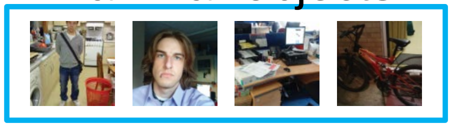
Object familiarity decoding from EEG activity patterns





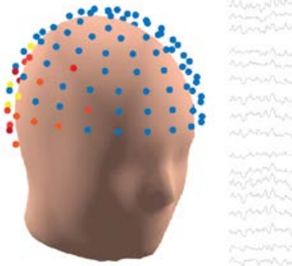
unfamiliar objects

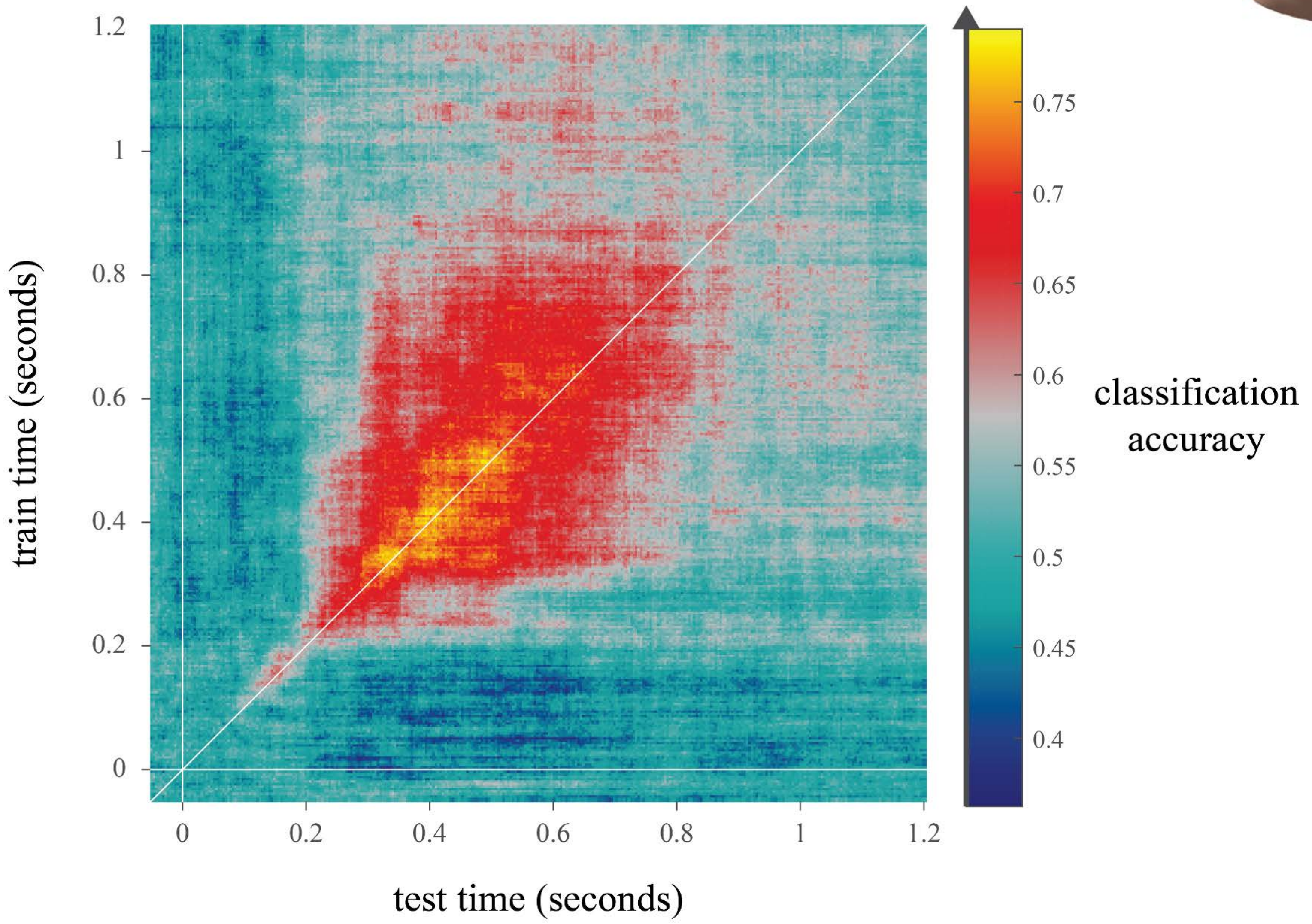
familiar objects



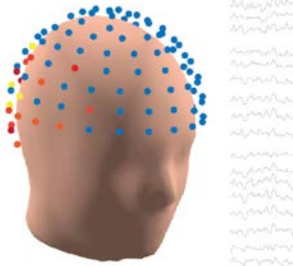
signifiant above-chance decoding

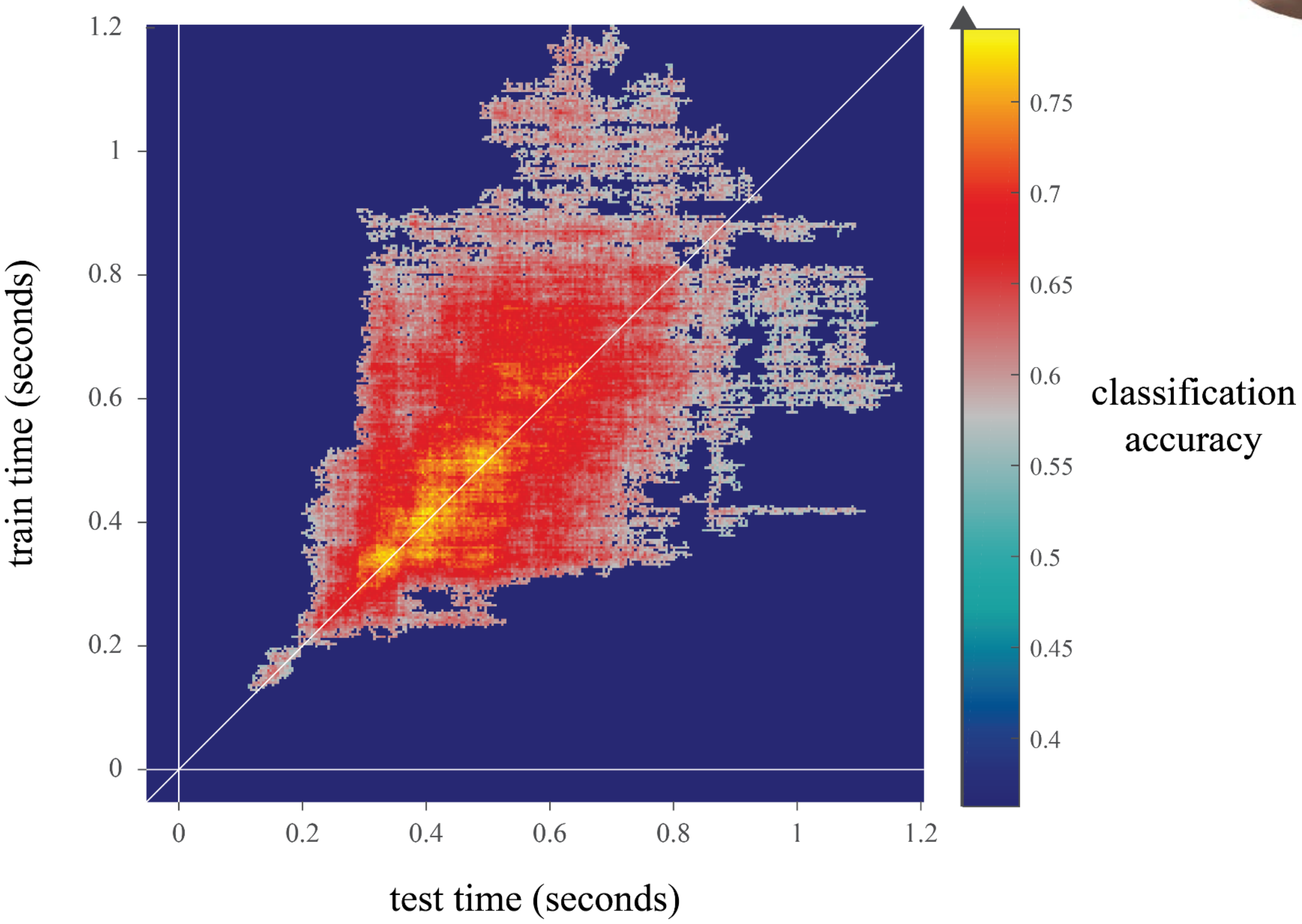
Object familiarity decoding from EEG activity patterns



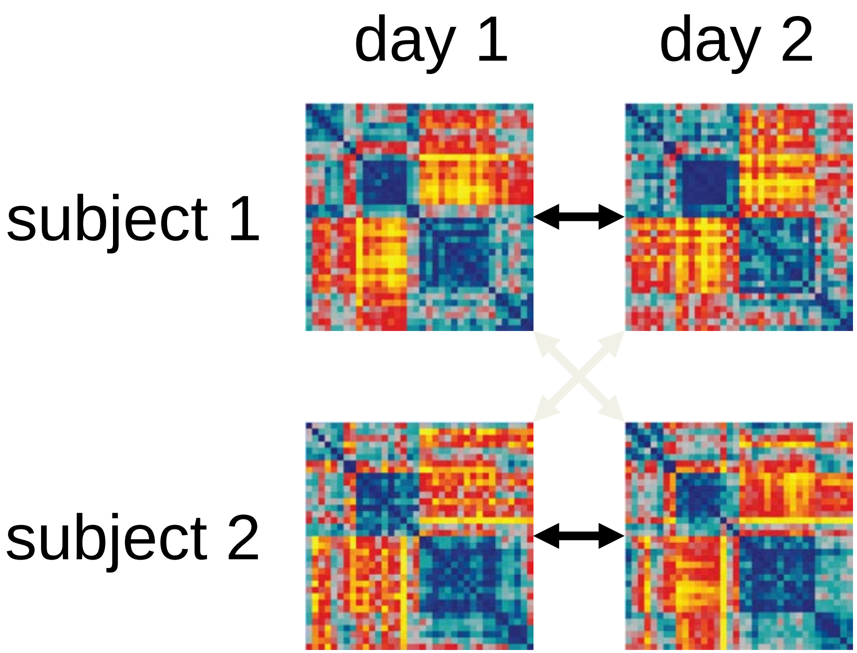


Object familiarity decoding from EEG activity patterns





Comparing individuals' representations

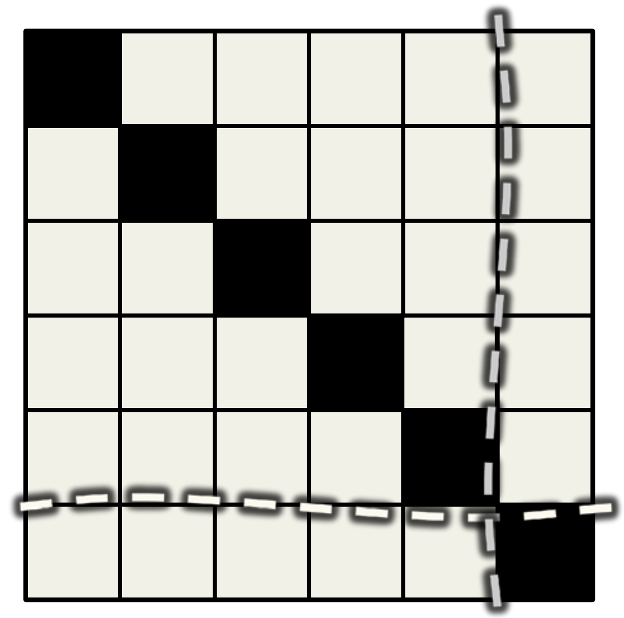


correlation

within-subject (ws) ✔

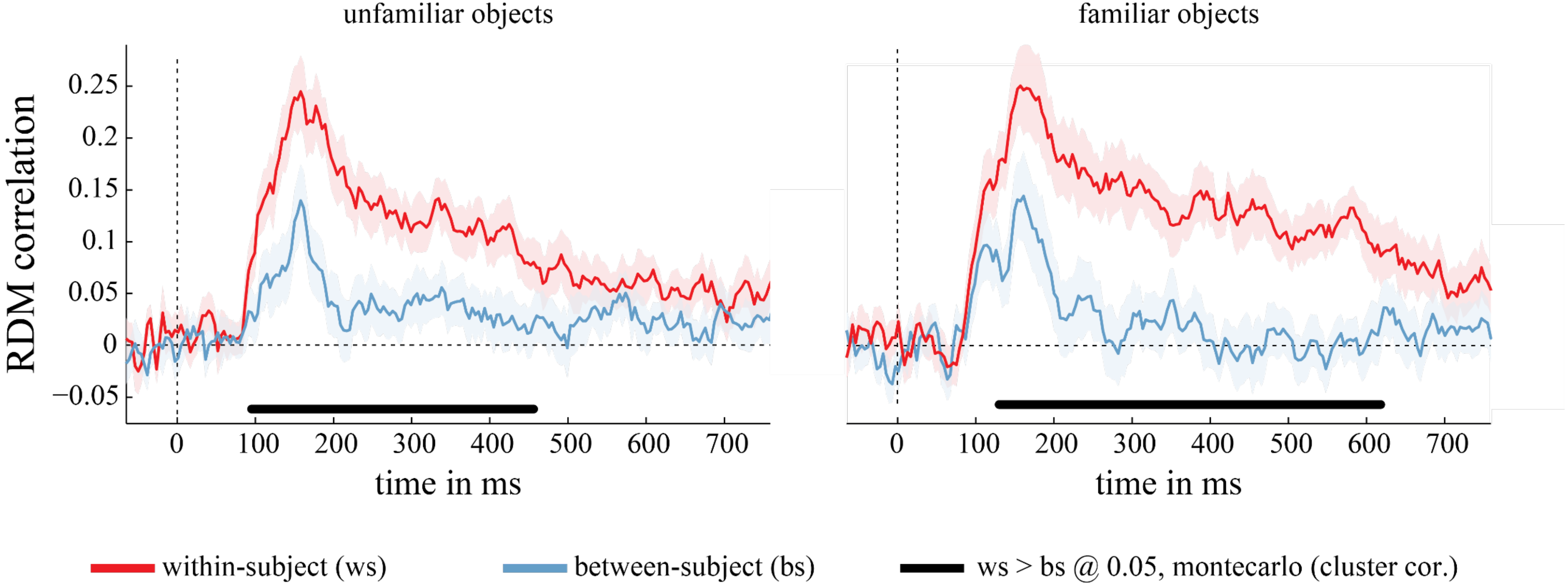
between-subject (bs) ✔

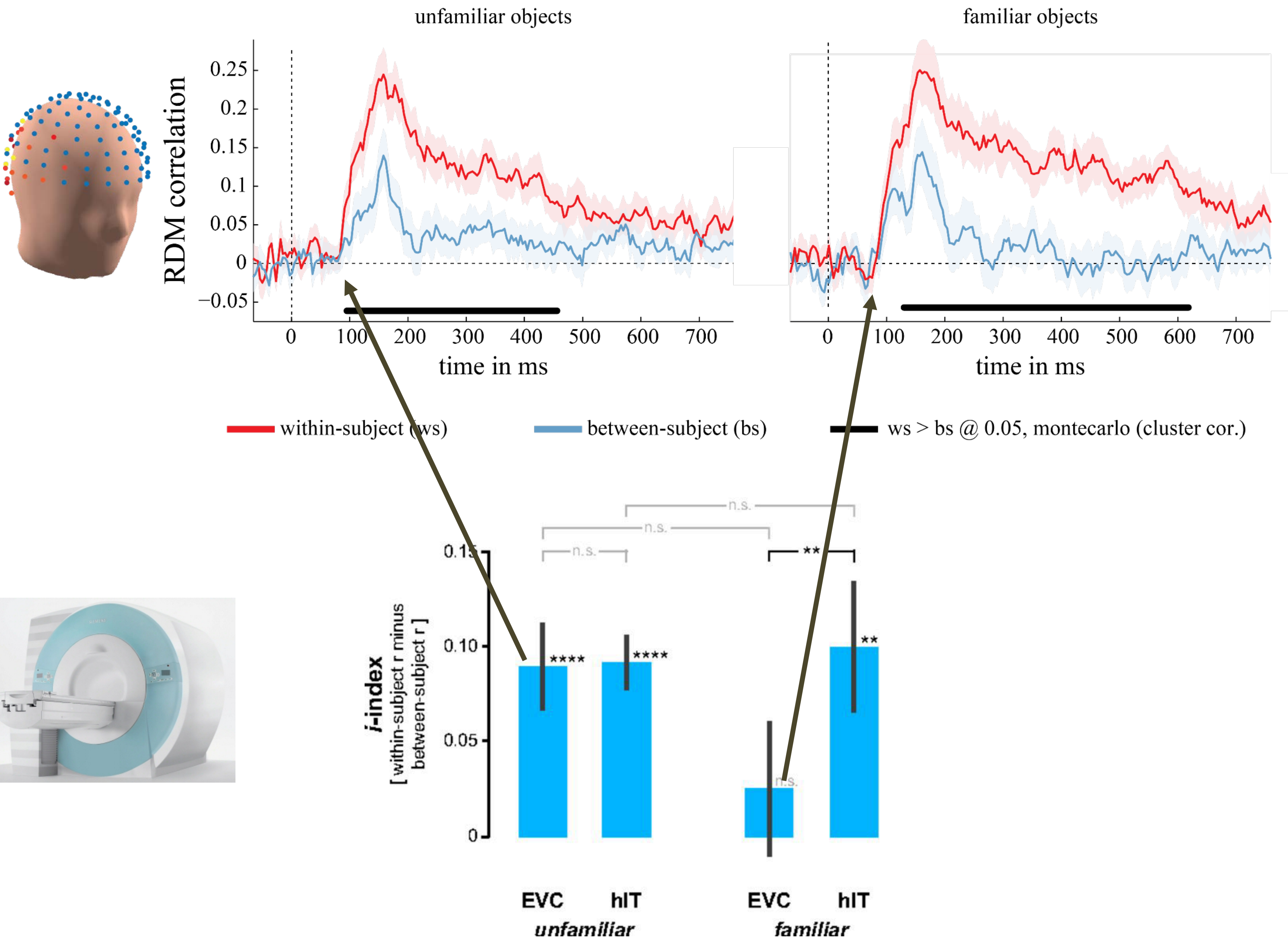
individuation index (ws - bs) ?



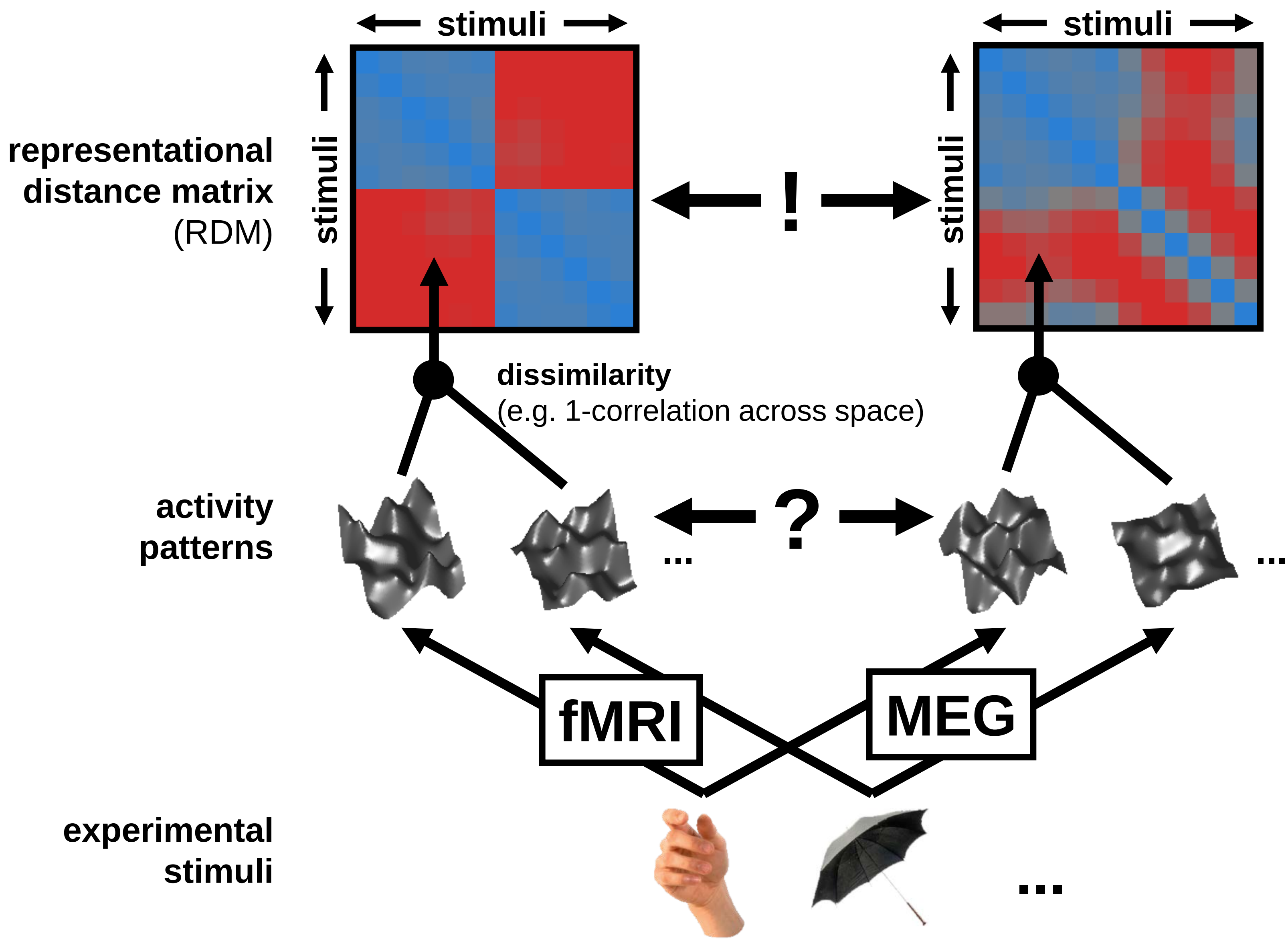
Charest et al. 2014 *PNAS*

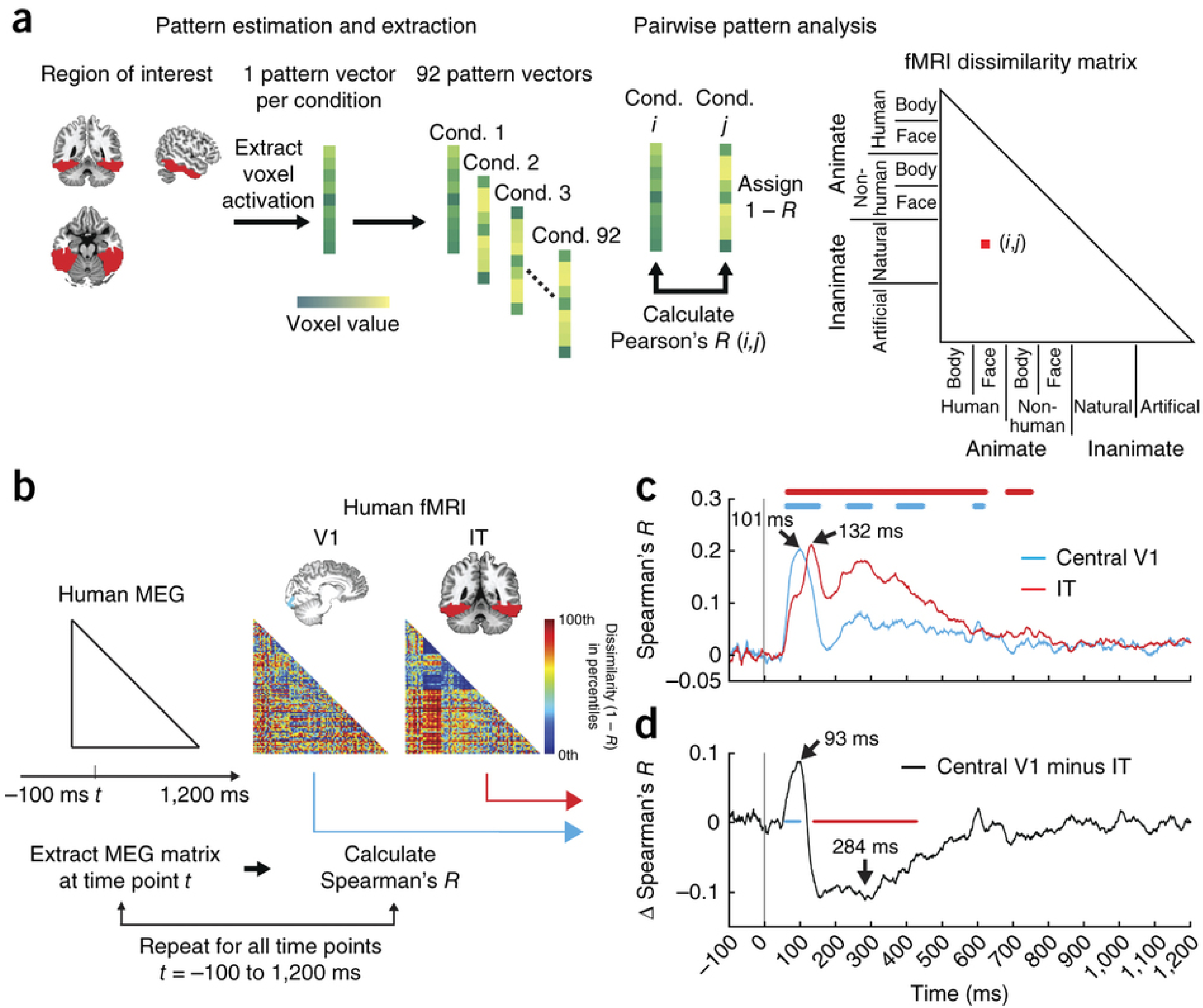
Comparing individuals' representations




unfamiliar objects
familiar objects
RDM correlation
0.25
0.2
0.15
0.1
0.05
0
−0.05
0
100
200
300
400
500
600
700
time in ms
within-subject (ws)
between-subject (bs)
ws > bs @ 0.05, montecarlo (cluster cor.)
n.s.
n.s.
n.s.
**
****
****
**
n.s.
0.15
0.10
0.05
0
i-index
[within-subject r minus between-subject r]
EVC
hIT
EVC
hIT
unfamiliar
familiar

Comparing RDMs between measurement modalities



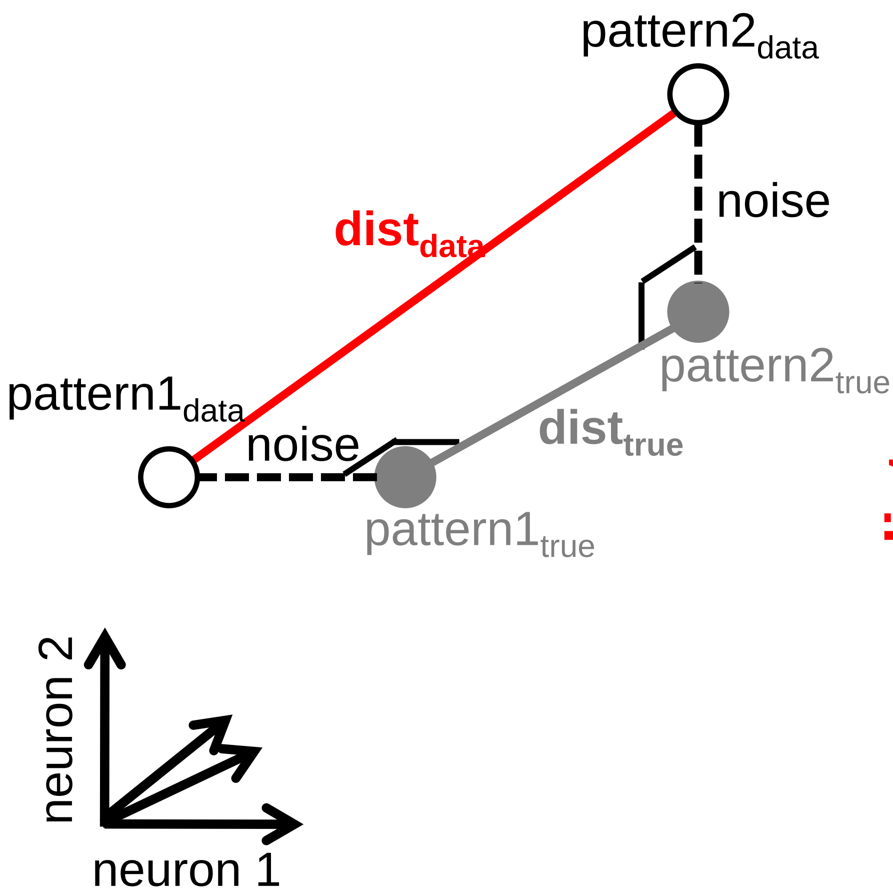


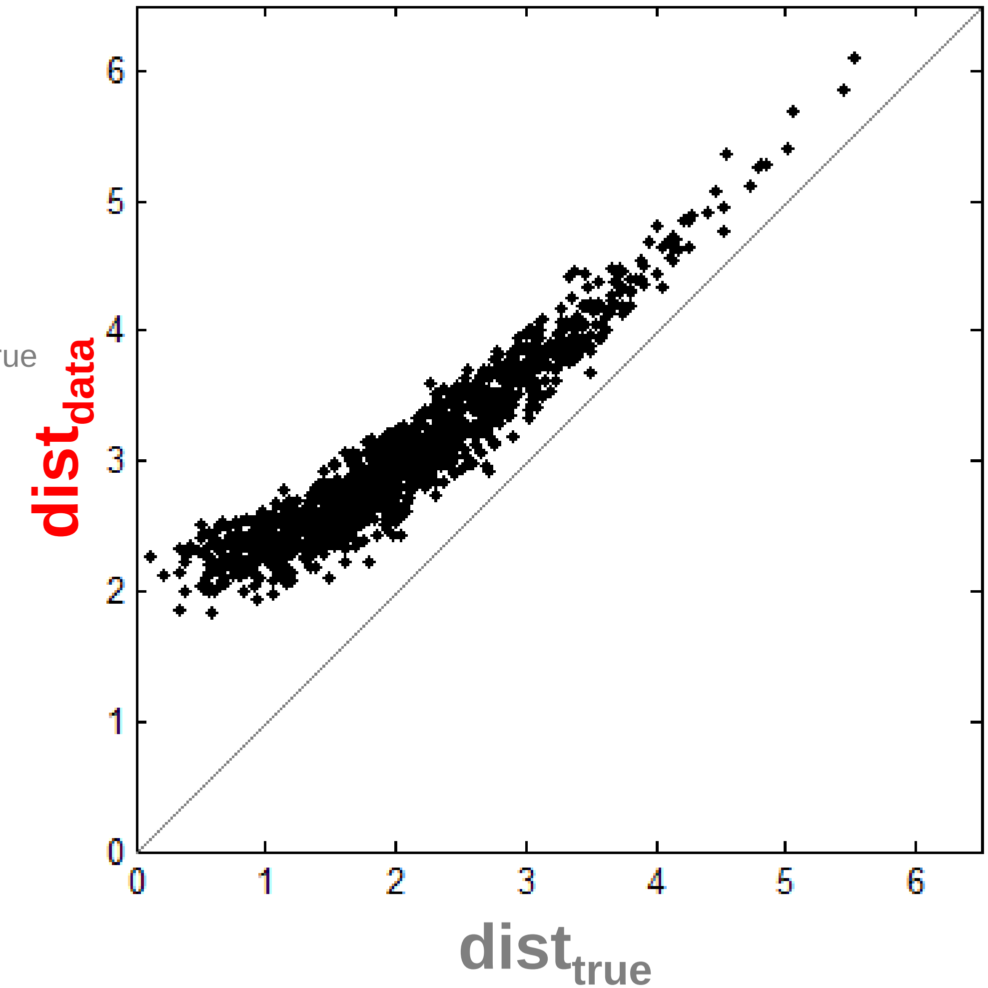
Cichy et al. 2014 *Nature Neuroscience*

https://www.youtube.com/watch?v=YBv_Bju4_aM

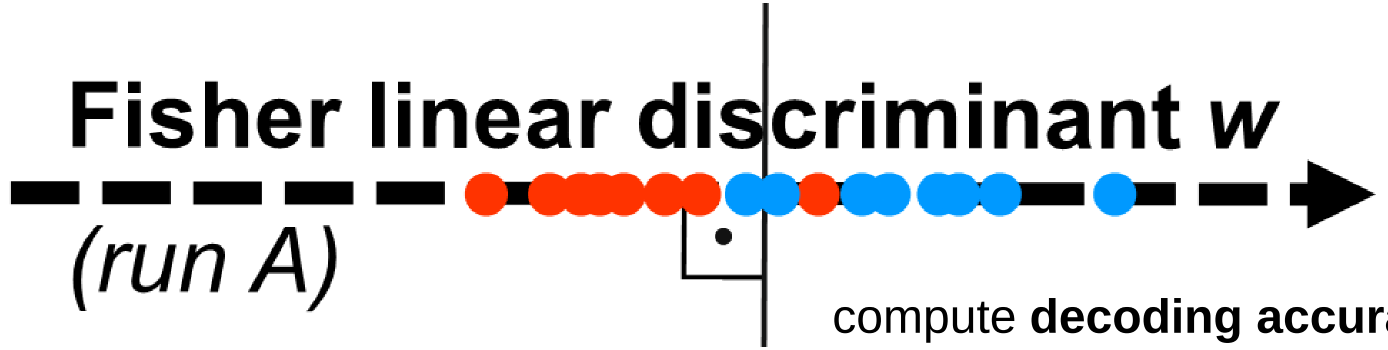
How can we best measure representational distances?

Distance estimates are positively biased



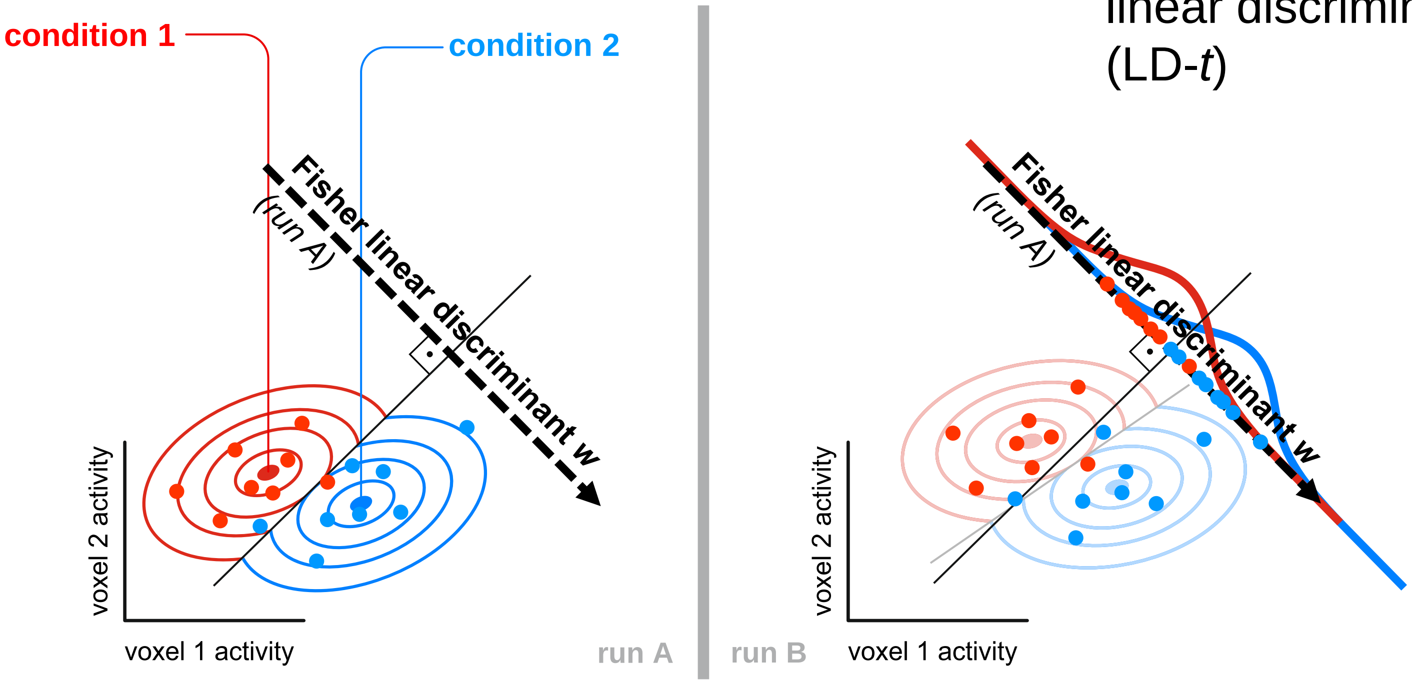


Distances are positively biased
– just like training-set decoding accuracies!



compute **decoding accuracy**?
or just do a ***t* test**?

linear discriminant *t* value
(LD-*t*)



Kriegeskorte et al. 2007, **Nili** et al. 2014

Unbiased distance estimates through crossvalidation

true distance = 0

average angle = 90°

E(inner product) = 0

true distance = 1

average angle < 90°

E(inner product) > 0

data set A

data set B

Unbiased distance estimates through crossvalidation

Estimated activity patterns

True activity patterns

Estimated error

"Training" $\hat{\mathbf{u}}_1 = \mathbf{u}_1 + \epsilon_1$

"Testing" $\hat{\mathbf{u}}_2 = \mathbf{u}_2 + \epsilon_2$

Run 1 *"Training"* *Run 2* *"Testing"*

$$\hat{d}(\hat{\mathbf{u}}_A,\hat{\mathbf{u}}_B) = \left(\hat{\mathbf{u}}_A^{(1)} - \hat{\mathbf{u}}_B^{(1)}\right)\left(\hat{\mathbf{u}}_A^{(2)} - \hat{\mathbf{u}}_B^{(2)}\right)^T$$

$$= \hat{\mathbf{u}}_A^{(1)}\hat{\mathbf{u}}_A^{(2)^T} - \hat{\mathbf{u}}_A^{(1)}\hat{\mathbf{u}}_B^{(2)^T} - \hat{\mathbf{u}}_B^{(1)}\hat{\mathbf{u}}_A^{(2)^T} + \hat{\mathbf{u}}_B^{(1)}\hat{\mathbf{u}}_B^{(2)^T}$$

$$= \left(\mathbf{u}_A^{(1)} + \epsilon^{(1)}\right)\left(\mathbf{u}_A^{(2)} + \epsilon^{(2)}\right)^T - \left(\mathbf{u}_A^{(1)} + \epsilon^{(1)}\right)\left(\mathbf{u}_B^{(2)} + \epsilon^{(2)}\right)^T - \left(\mathbf{u}_B^{(1)} + \epsilon^{(1)}\right)\left(\mathbf{u}_A^{(2)} + \epsilon^{(2)}\right)^T \ldots$$

$$+ \left(\mathbf{u}_B^{(1)} + \epsilon^{(1)}\right)\left(\mathbf{u}_B^{(2)} + \epsilon^{(2)}\right)^T$$

$$= \mathbf{u}_A^{(1)}\mathbf{u}_A^{(2)^T} + \mathbf{u}_A^{(1)}\epsilon^{(2)^T} + \epsilon^{(1)}\mathbf{u}_A^{(2)^T} + \epsilon^{(1)}\epsilon^{(2)^T} - \mathbf{u}_A^{(1)}\mathbf{u}_A^{(2)^T} - \mathbf{u}_A^{(1)}\epsilon^{(2)^T} - \epsilon^{(1)}\mathbf{u}_B^{(2)^T} - \epsilon^{(1)}\epsilon^{(2)^T} \ldots$$

$$- \mathbf{u}_B^{(1)}\mathbf{u}_A^{(2)^T} - \mathbf{u}_B^{(1)}\epsilon^{(2)^T} \ldots - \epsilon^{(1)}\mathbf{u}_A^{(2)^T} - \epsilon^{(1)}\epsilon^{(2)^T} + \mathbf{u}_B^{(1)}\mathbf{u}_B^{(2)^T} + \epsilon^{(2)}\mathbf{u}_B^{(1)^T} + \epsilon^{(1)}\mathbf{u}_B^{(2)^T} + \epsilon^{(1)}\epsilon^{(2)^T}$$

$$E\left(\hat{d}(\hat{\mathbf{u}}_A,\hat{\mathbf{u}}_B)\right) = \mathbf{u}_A^{(1)}\mathbf{u}_A^{(2)^T} - \mathbf{u}_A^{(1)}\mathbf{u}_A^{(2)^T} - \mathbf{u}_B^{(1)}\mathbf{u}_A^{(2)^T} + \mathbf{u}_B^{(1)}\mathbf{u}_B^{(2)^T}$$

$$= \left(\mathbf{u}_A^{(1)} - \mathbf{u}_B^{(1)}\right)\left(\mathbf{u}_A^{(2)} - \mathbf{u}_B^{(2)}\right)^T = d(\mathbf{u}_A,\mathbf{u}_B)$$

Walther et al. (2016)

Unbiased distance estimates through crossvalidation

Estimated activity patterns

True activity patterns

Estimated error

"Training" $\hat{\mathbf{u}}_1 = \mathbf{u}_1 + \epsilon_1$

"Testing" $\hat{\mathbf{u}}_2 = \mathbf{u}_2 + \epsilon_2$

Run 1 *"Training"* *Run 2* *"Testing"*

$$\hat{d}(\hat{\mathbf{u}}_A,\hat{\mathbf{u}}_B) = \left(\hat{\mathbf{u}}_A^{(1)} - \hat{\mathbf{u}}_B^{(1)}\right)\left(\hat{\mathbf{u}}_A^{(2)} - \hat{\mathbf{u}}_B^{(2)}\right)^T$$

$$= \hat{\mathbf{u}}_A^{(1)}\hat{\mathbf{u}}_A^{(2)^T} - \hat{\mathbf{u}}_A^{(1)}\hat{\mathbf{u}}_B^{(2)^T} - \hat{\mathbf{u}}_B^{(1)}\hat{\mathbf{u}}_A^{(2)^T} + \hat{\mathbf{u}}_B^{(1)}\hat{\mathbf{u}}_B^{(2)^T}$$

$$= \left(\mathbf{u}_A^{(1)} + \epsilon^{(1)}\right)\left(\mathbf{u}_A^{(2)} + \epsilon^{(2)}\right)^T - \left(\mathbf{u}_A^{(1)} + \epsilon^{(1)}\right)\left(\mathbf{u}_B^{(2)} + \epsilon^{(2)}\right)^T - \left(\mathbf{u}_B^{(1)} + \epsilon^{(1)}\right)\left(\mathbf{u}_A^{(2)} + \epsilon^{(2)}\right)^T \ldots$$

$$+ \left(\mathbf{u}_B^{(1)} + \epsilon^{(1)}\right)\left(\mathbf{u}_B^{(2)} + \epsilon^{(2)}\right)^T$$

$$= \mathbf{u}_A^{(1)}\mathbf{u}_A^{(2)^T} + \mathbf{u}_A^{(1)}\epsilon^{(2)^T} + \epsilon^{(1)}\mathbf{u}_A^{(2)^T} + \epsilon^{(1)}\epsilon^{(2)^T} - \mathbf{u}_A^{(1)}\mathbf{u}_A^{(2)^T} - \mathbf{u}_A^{(1)}\epsilon^{(2)^T} - \epsilon^{(1)}\mathbf{u}_B^{(2)^T} - \epsilon^{(1)}\epsilon^{(2)^T} \ldots$$

$$-\mathbf{u}_B^{(1)}\mathbf{u}_A^{(2)^T} - \mathbf{u}_B^{(1)}\epsilon^{(2)^T} \ldots - \epsilon^{(1)}\mathbf{u}_A^{(2)^T} - \epsilon^{(1)}\epsilon^{(2)^T} + \mathbf{u}_B^{(1)}\mathbf{u}_B^{(2)^T} + \epsilon^{(2)}\mathbf{u}_B^{(1)^T} + \epsilon^{(1)}\mathbf{u}_B^{(2)^T} + \epsilon^{(1)}\epsilon^{(2)^T}$$

$$E\left(\hat{d}(\hat{\mathbf{u}}_A,\hat{\mathbf{u}}_B)\right) = \mathbf{u}_A^{(1)}\mathbf{u}_A^{(2)^T} - \mathbf{u}_A^{(1)}\mathbf{u}_A^{(2)^T} - \mathbf{u}_B^{(1)}\mathbf{u}_A^{(2)^T} + \mathbf{u}_B^{(1)}\mathbf{u}_B^{(2)^T}$$

$$= \left(\mathbf{u}_A^{(1)} - \mathbf{u}_B^{(1)}\right)\left(\mathbf{u}_A^{(2)} - \mathbf{u}_B^{(2)}\right)^T = d(\mathbf{u}_A,\mathbf{u}_B)$$

Walther et al. (2016)

The linear discriminant contrast (LDC) is a crossvalidated variant of the Mahalanobis distance

Mahalanobis distance (single data set)

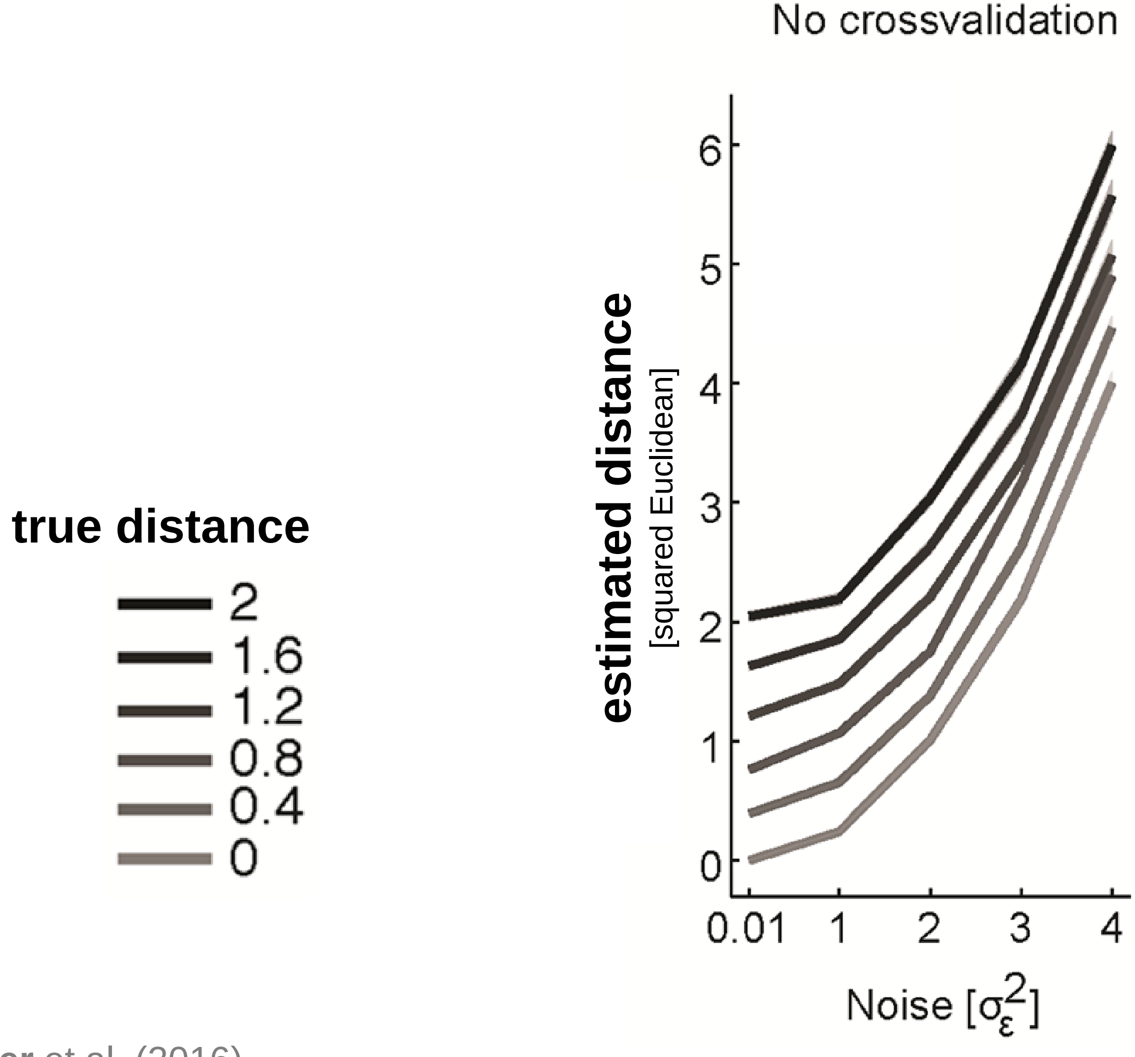
training set $(\mathbf{p2}-\mathbf{p1})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{p2}-\mathbf{p1})$

Fisher linear discriminant contrast (crossvalidated)

training set $(\mathbf{p2}-\mathbf{p1})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{p2}'-\mathbf{p1}')$ test set

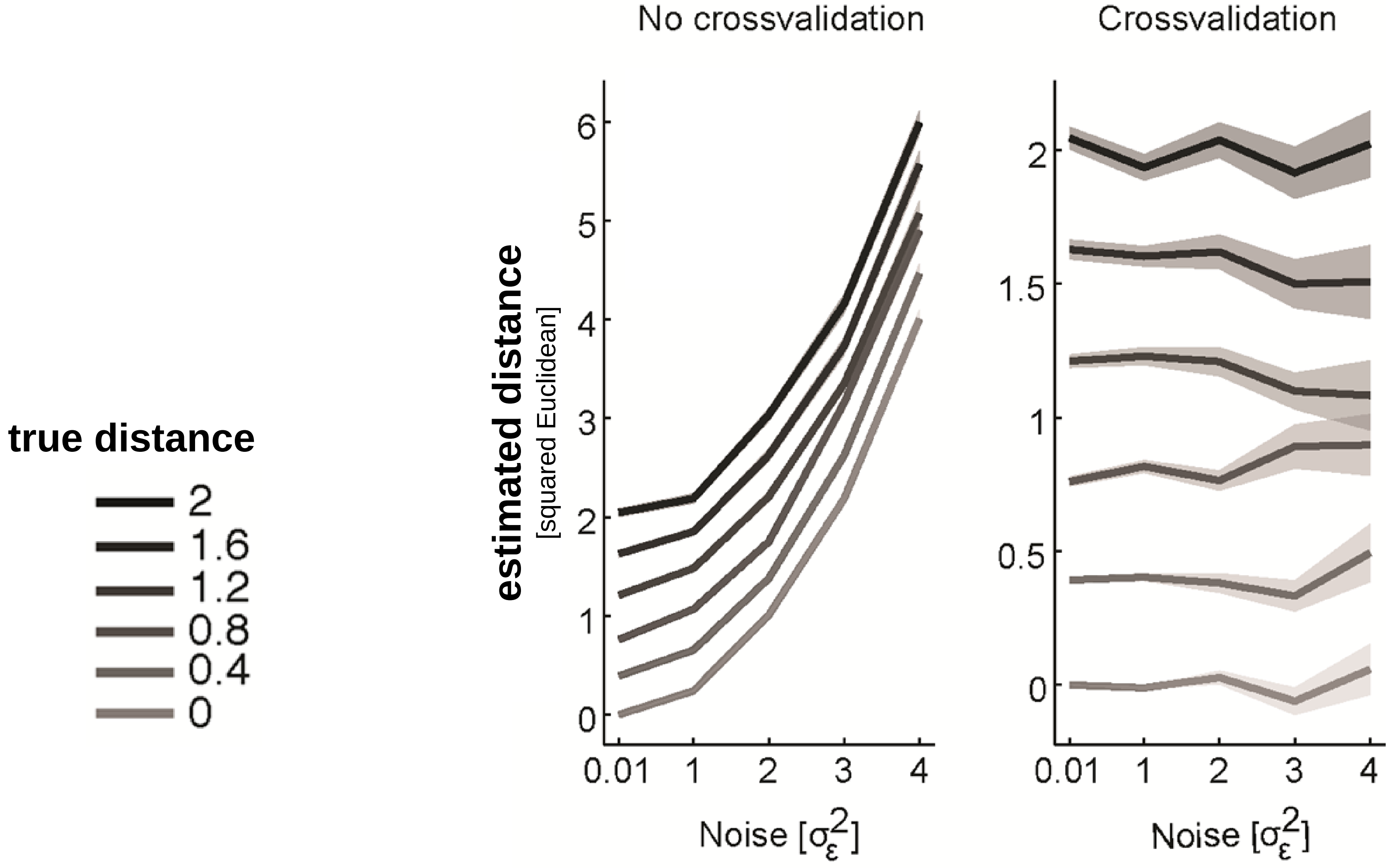
Nili et al. (2014)

Crossvalidation removes the bias of distance estimates



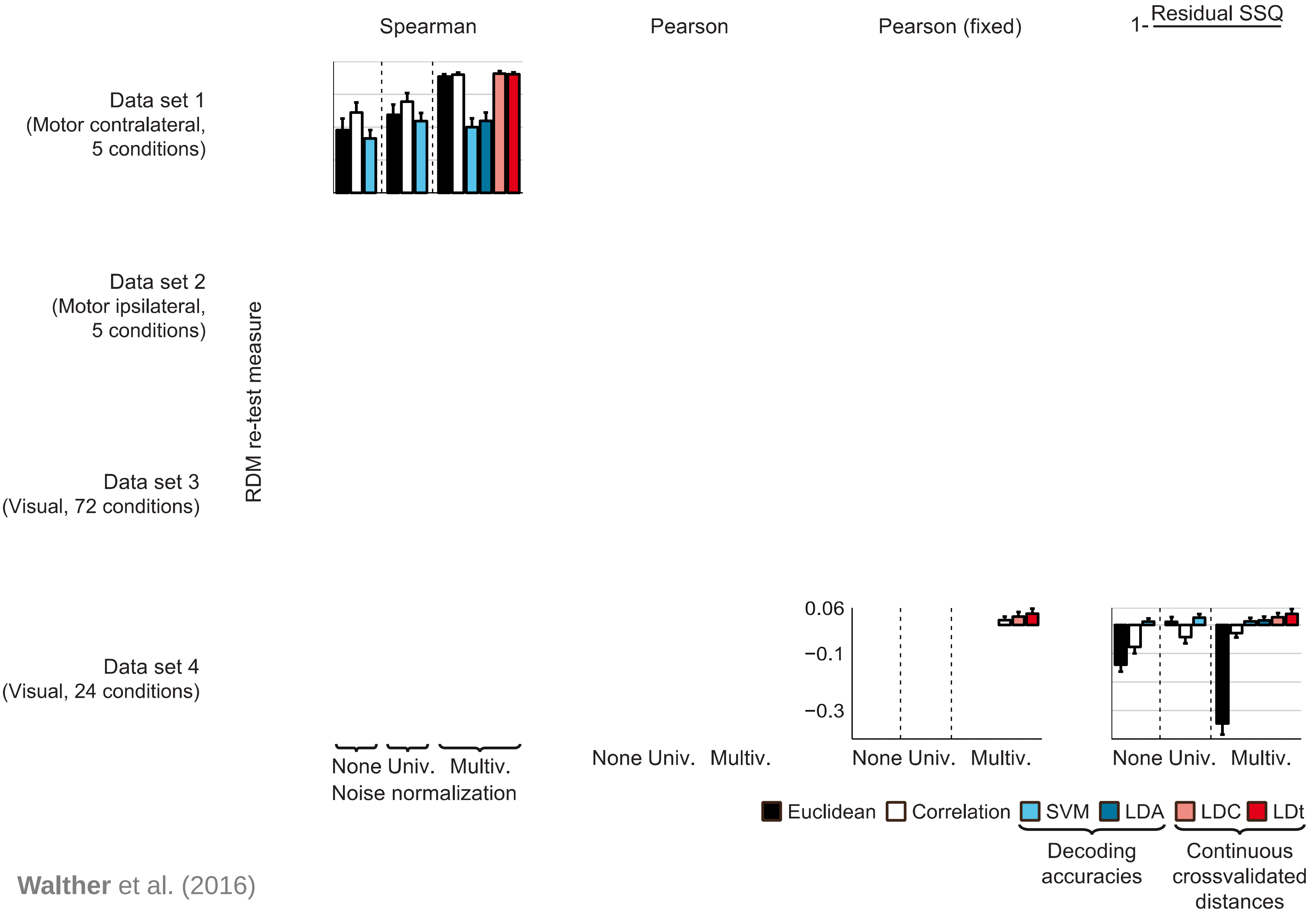
Walther et al. (2016)

Crossvalidation removes the bias of distance estimates

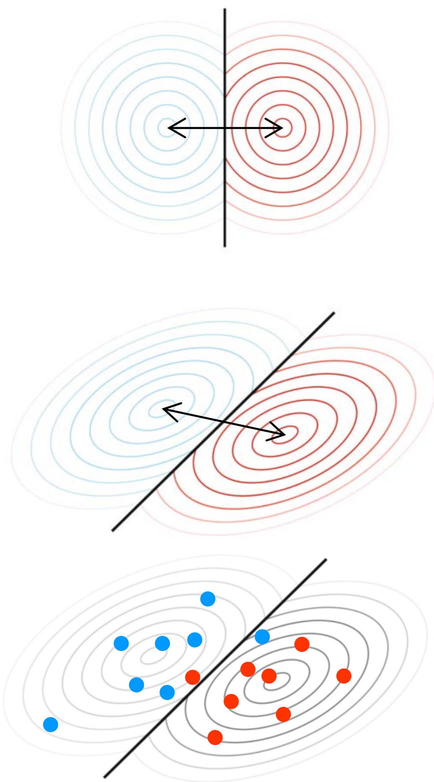


Walther et al. (2016)

Test-retest reliability for different dissimilarity measures



Walther et al. (2016)

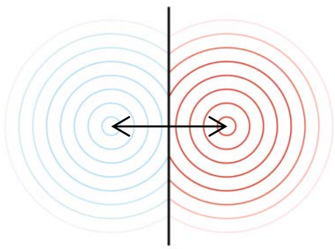


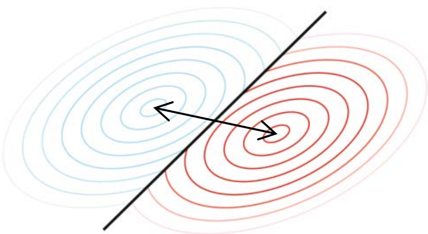
no crossvalidation	crossvalidation
Euclidean distance	Centroid connection discriminant contrast
Mahalanobis distance	Fisher linear discriminant contrast
Training-set decoding accuracy	Test-set decoding accuracy

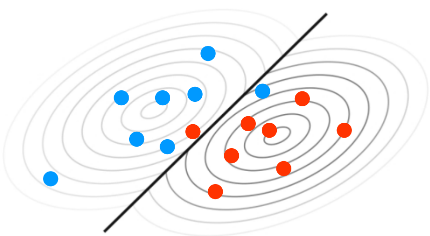
covariance-blind

ceiling-limited and quantized

positively biased







no crossvalidation	crossvalidation
Euclidean distance	Centroid connection discriminant contrast
Mahalanobis distance	**Fisher linear discriminant contrast**
Training-set decoding accuracy	Test-set decoding accuracy

covariance-blind

ceiling-limited and quantized

positively biased

The best of both worlds...

Multivariate statistics	**Machine learning**
multinormal distribution	pattern classifiers
continuous measures of multivariate separation	crossvalidation
inference relying on multinormality	nonparametric inference procedures

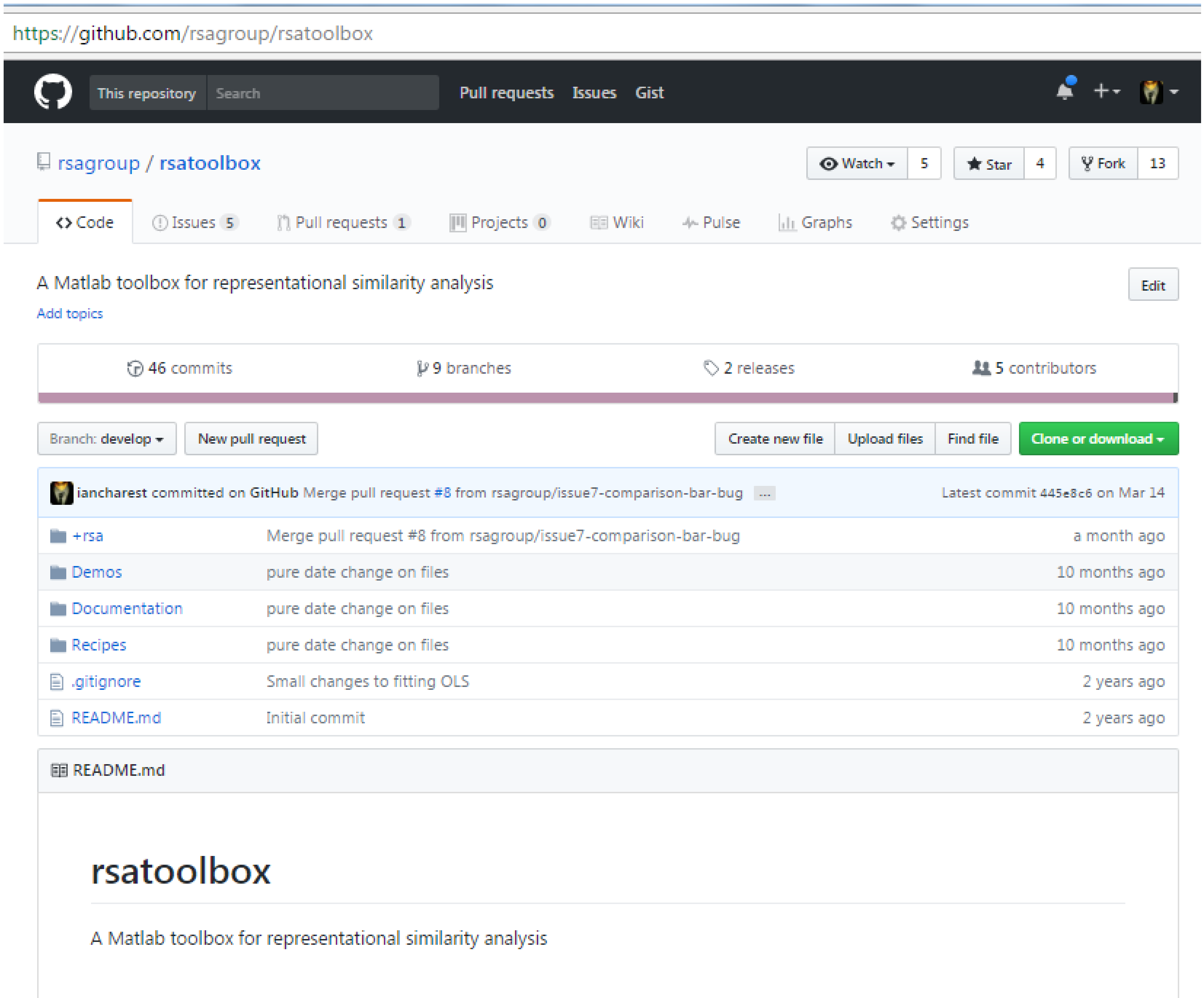
Key insights

Representational geometries encapsulate the *content* and *format* of brain representations.

Representational geometries can be characterised by representational dissimilarity matrices (RDMs).

RDMs can easily be compared between brains and models, individuals and species, different brain regions, different measurement modalities, and brain and behaviour.

We can statistically compare multiple computational models and assess whether they fully explain the measured brain response patterns.


https://github.com/rsagroup/rsatoolbox
This repository
Search
Pull requests
Issues
Gist
rsagroup / rsatoolbox
Watch 5
Star 4
Fork 13
Code
Issues 5
Pull requests 1
Projects 0
Wiki
Pulse
Graphs
Settings
A Matlab toolbox for representational similarity analysis
Edit
Add topics
46 commits
9 branches
2 releases
5 contributors
Branch: develop
New pull request
Create new file
Upload files
Find file
Clone or download
iancharest committed on GitHub Merge pull request #8 from rsagroup/issue7-comparison-bar-bug
Latest commit 445e8c6 on Mar 14
+rsa
Merge pull request #8 from rsagroup/issue7-comparison-bar-bug
a month ago
Demos
pure date change on files
10 months ago
Documentation
pure date change on files
10 months ago
Recipes
pure date change on files
10 months ago
.gitignore
Small changes to fitting OLS
2 years ago
README.md
Initial commit
2 years ago
README.md
rsatoolbox
A Matlab toolbox for representational similarity analysis