University of Wisconsin
SCHOOL OF MEDICINE
AND PUBLIC HEALTH

Clustering

Moo K. Chung
University of Wisconsin-Madison
www.stat.wisc.edu/~mchung →
github.com/laplcebeltrami

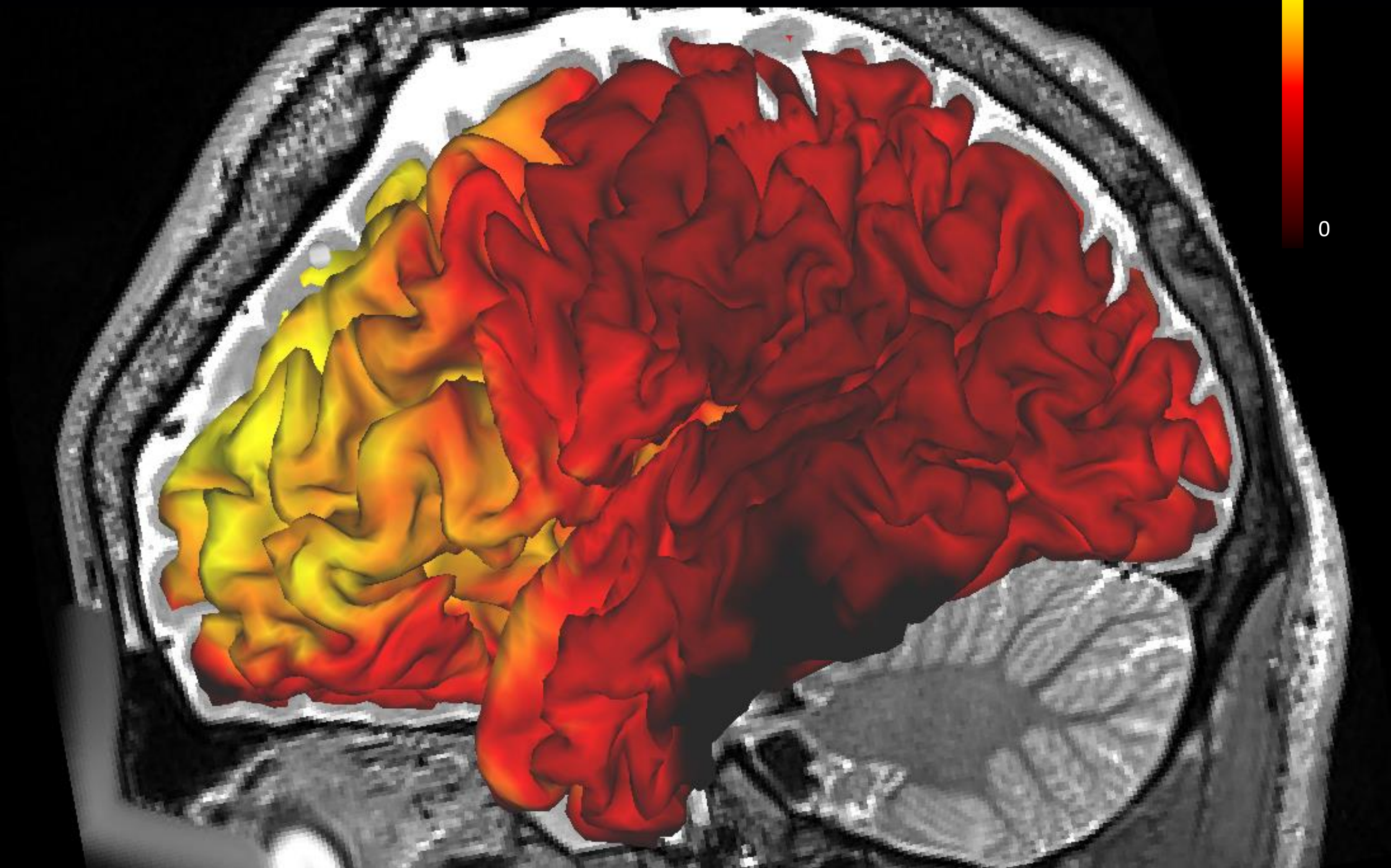
Motivating problem

Magnetic resonance imaging (MRI)

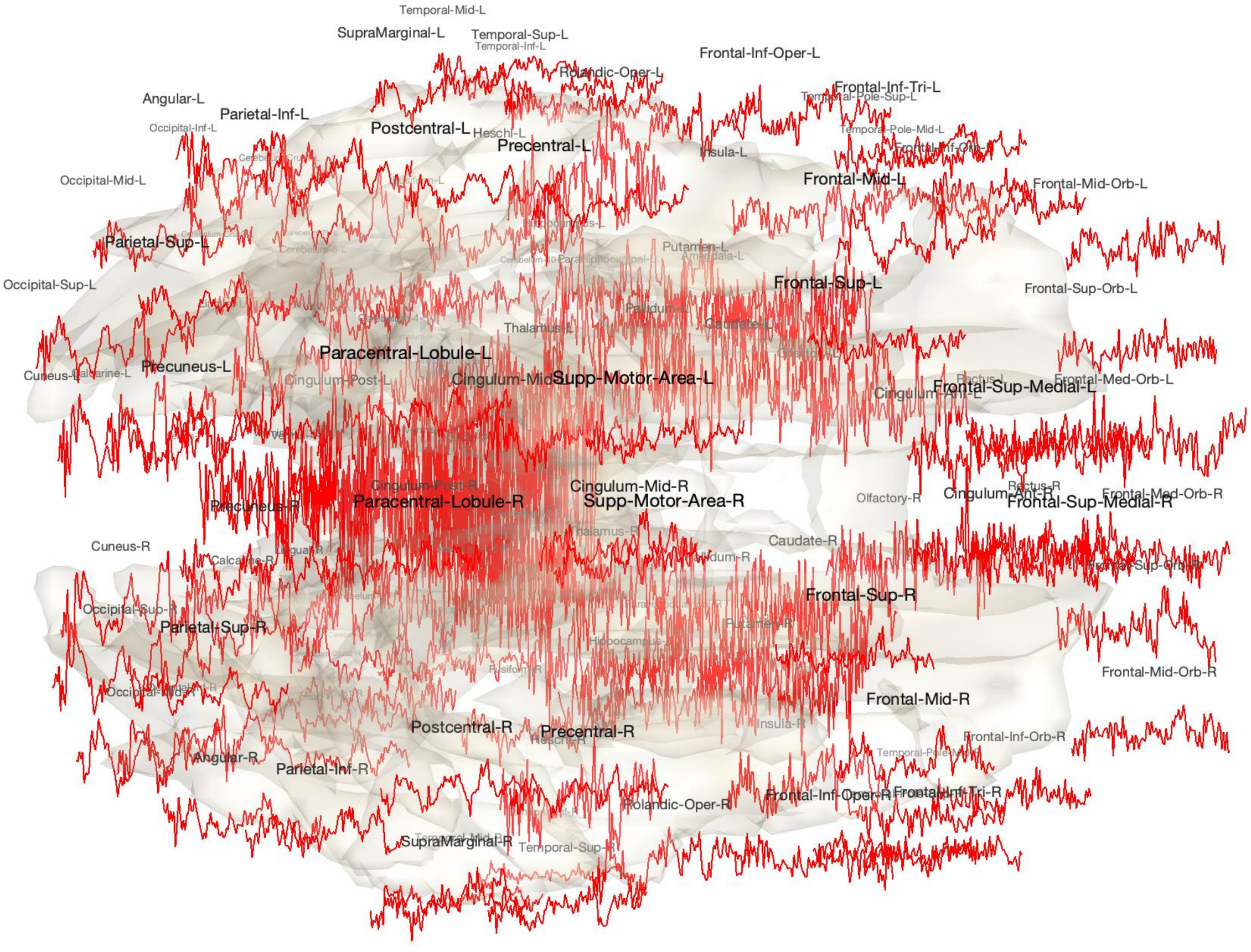
3T GE Discovery X750
Waisman Brain Imaging Laboratory
University of Wisconsin-Madison

3T GE Discovery MR750
Center for Imaging Research
Medical College of Wisconsin, Milwaukee, WI

rs-fMRI (every 30 second)

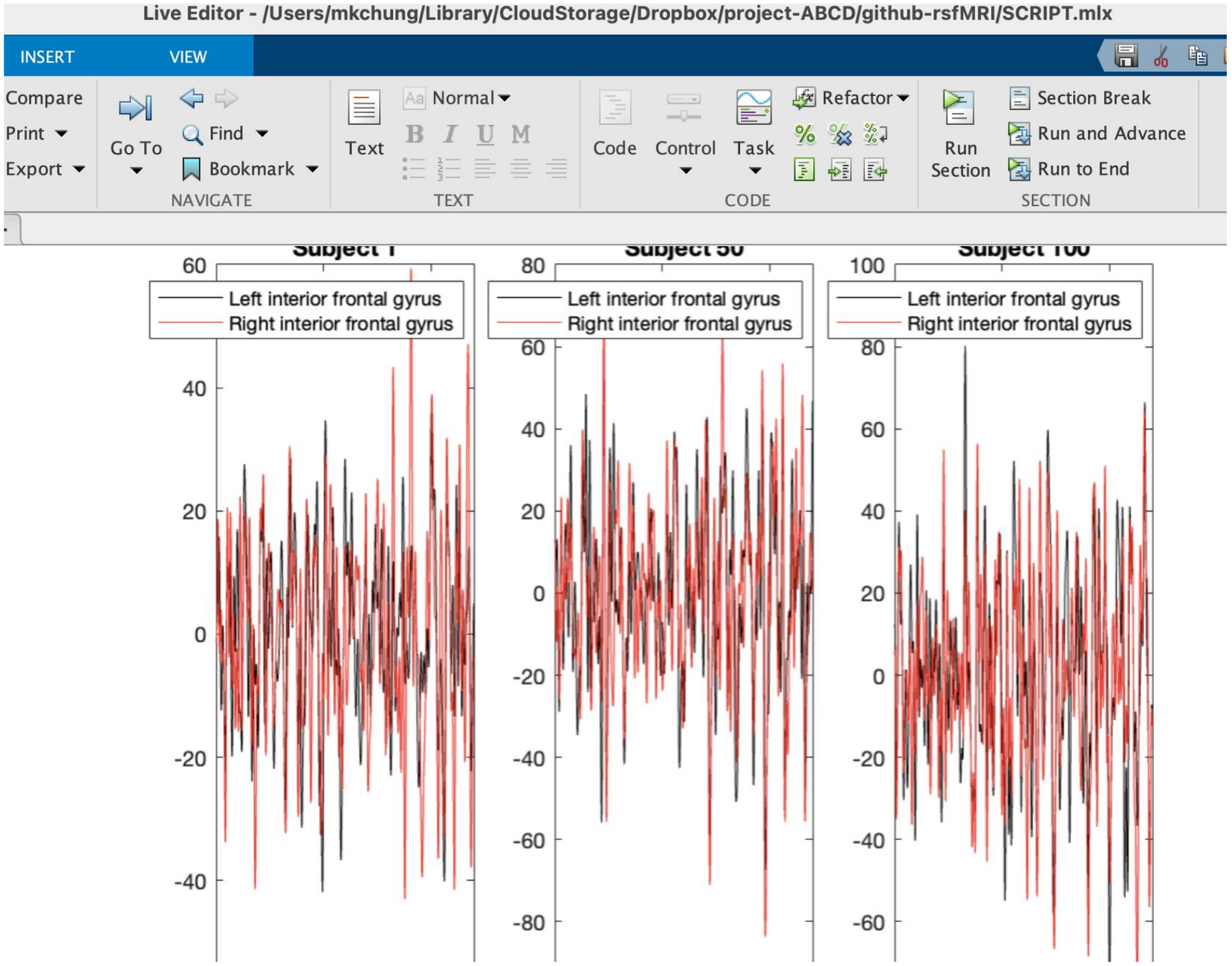


Time series averaged into 116 brain r



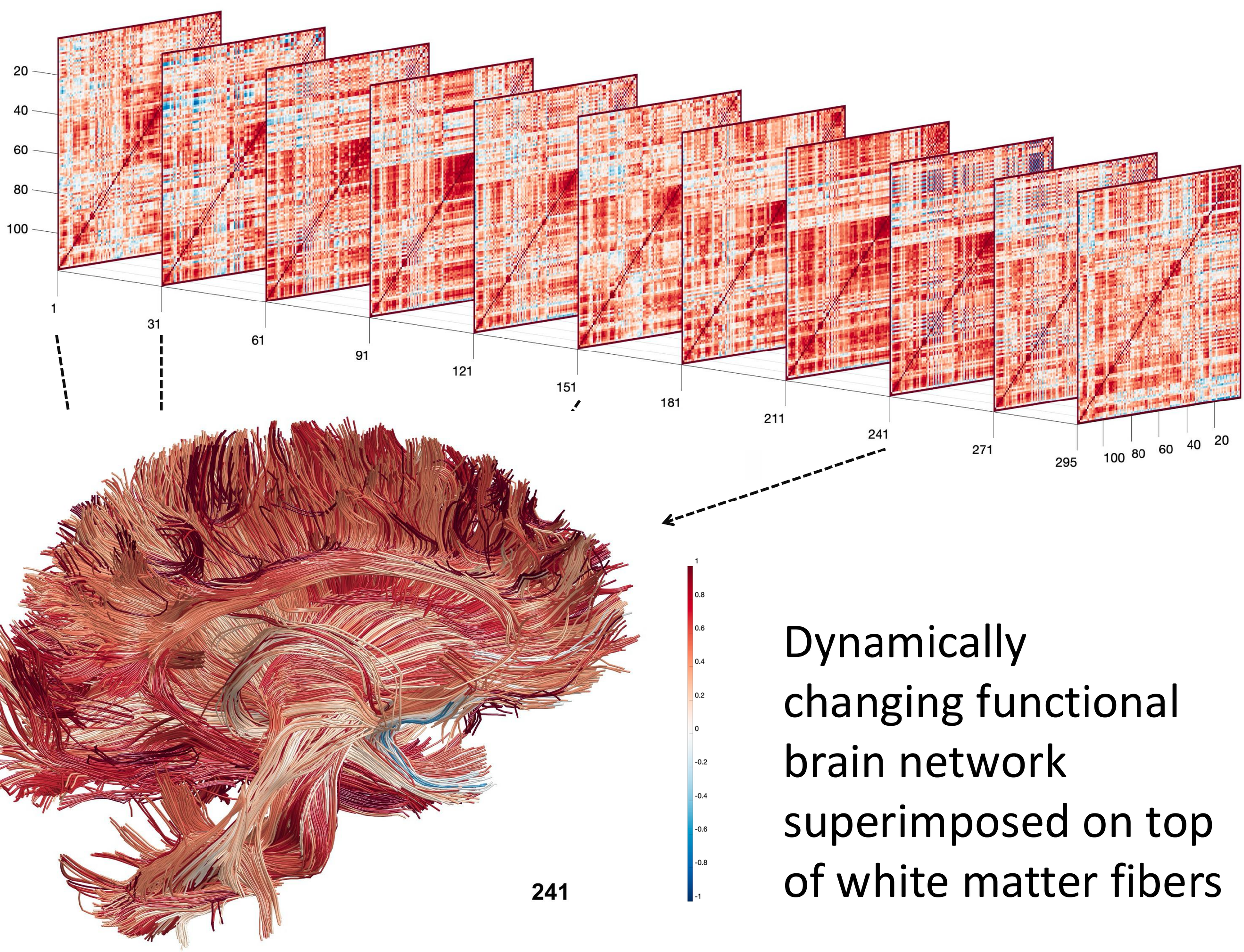
Sample rs-fMRI time series data

rsfMRI



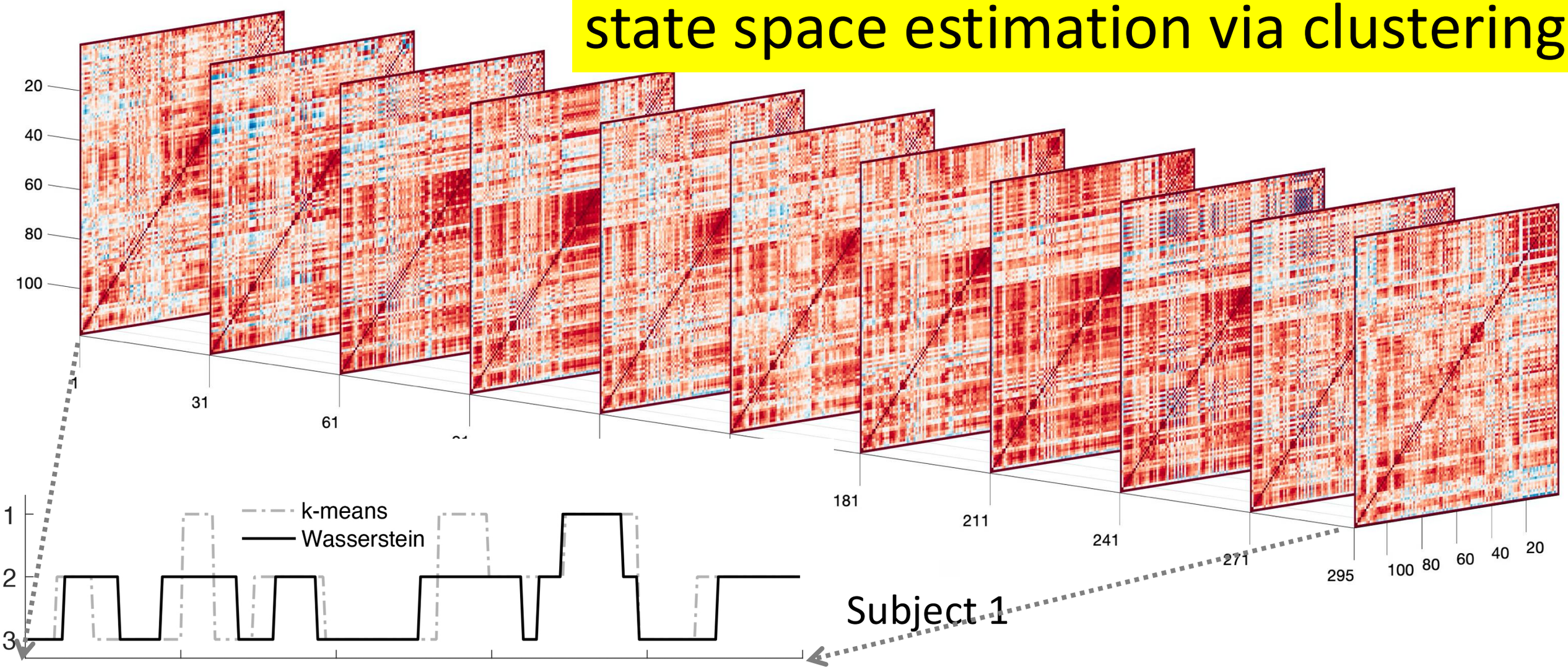
Important biological questions are added

Huang et al. 2020 Neuroscience Methods 331:108480

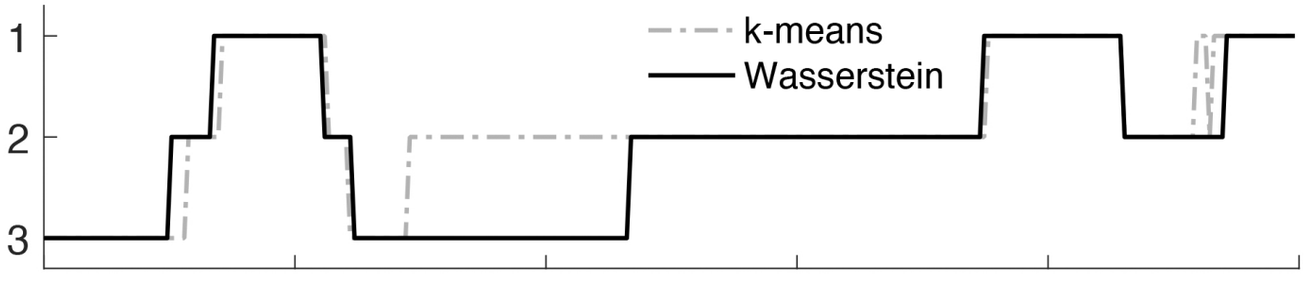


Dynamically changing functional brain network superimposed on top of white matter fibers

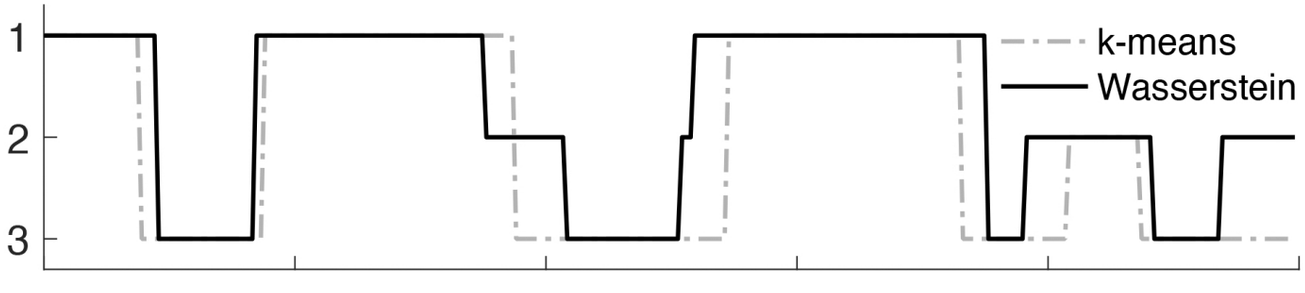
state space estimation via clustering



Subject 1



Subject 2



Subject 3

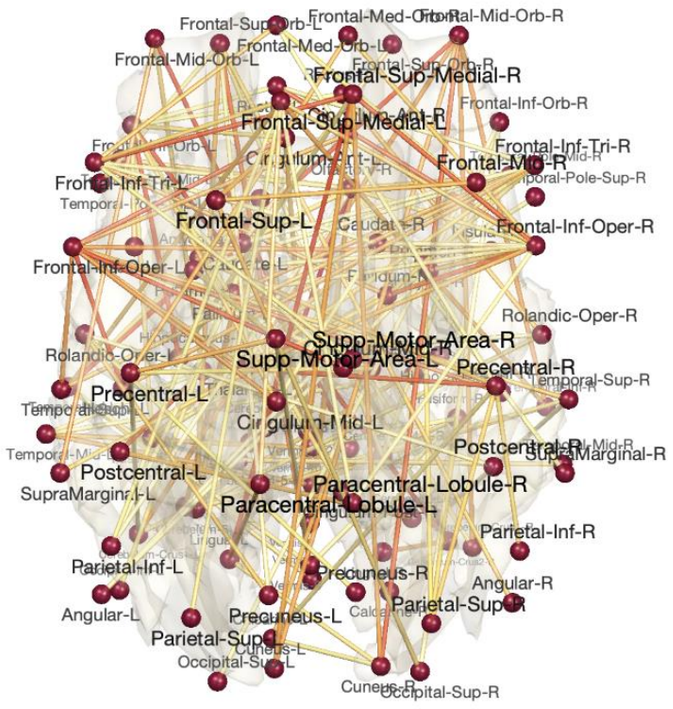
2nd level analysis:

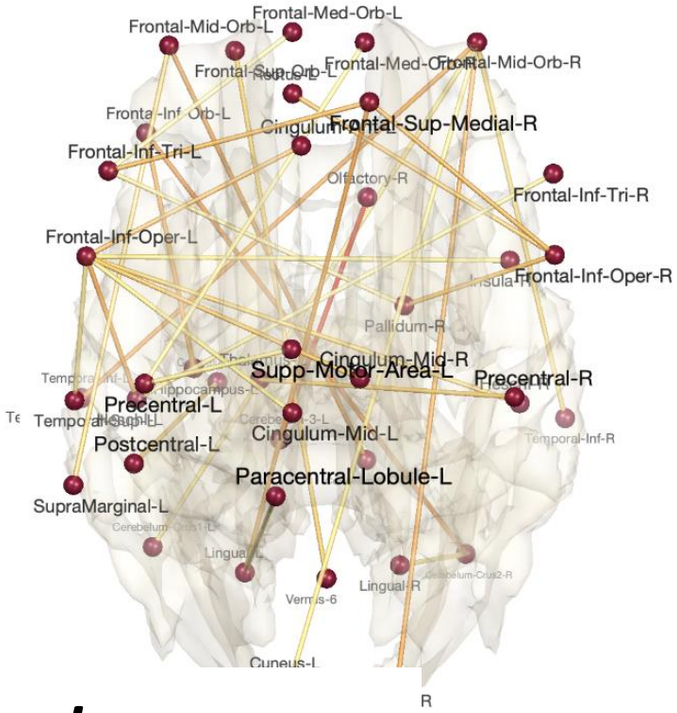
Statistical model

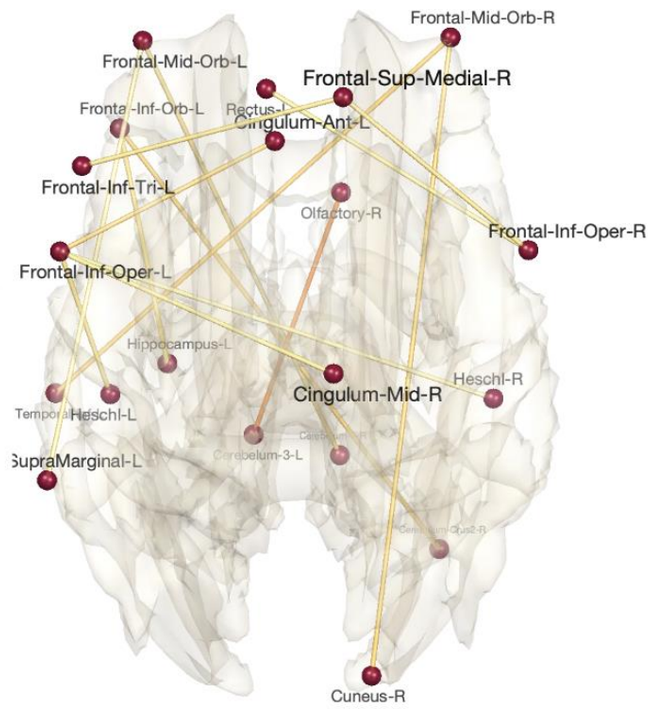
State 1

State 2

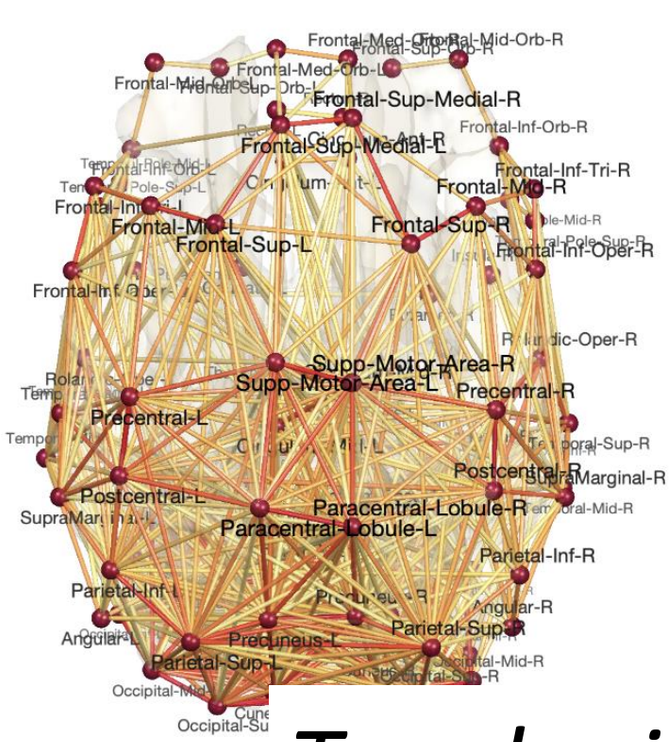
State 3

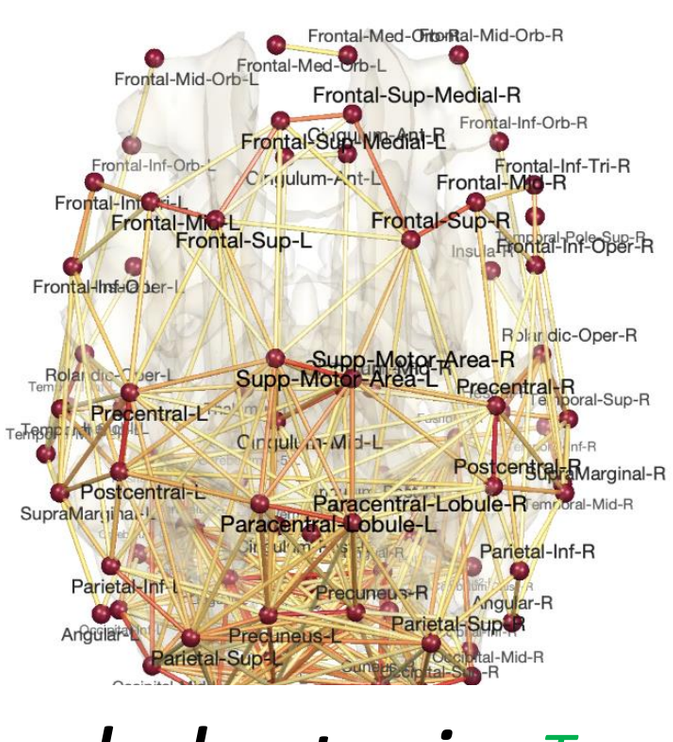


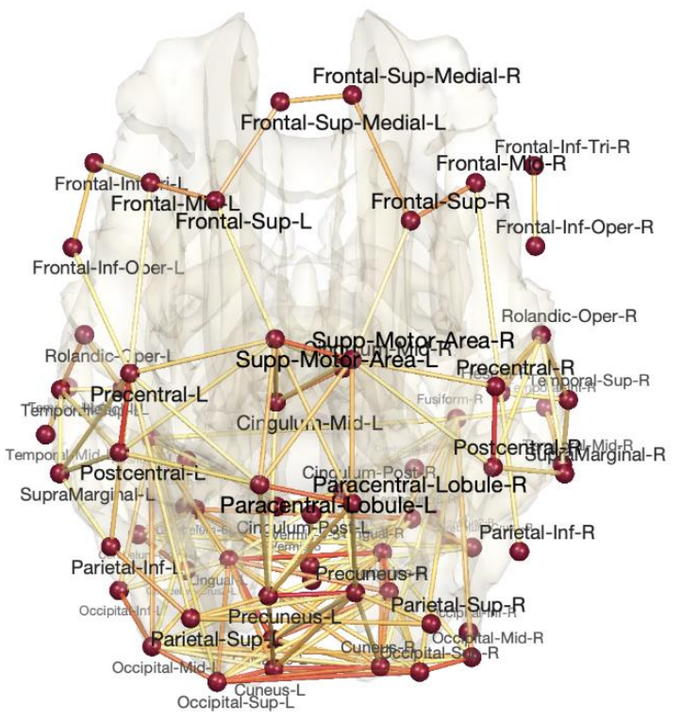




k-means *Sample mean in each state*







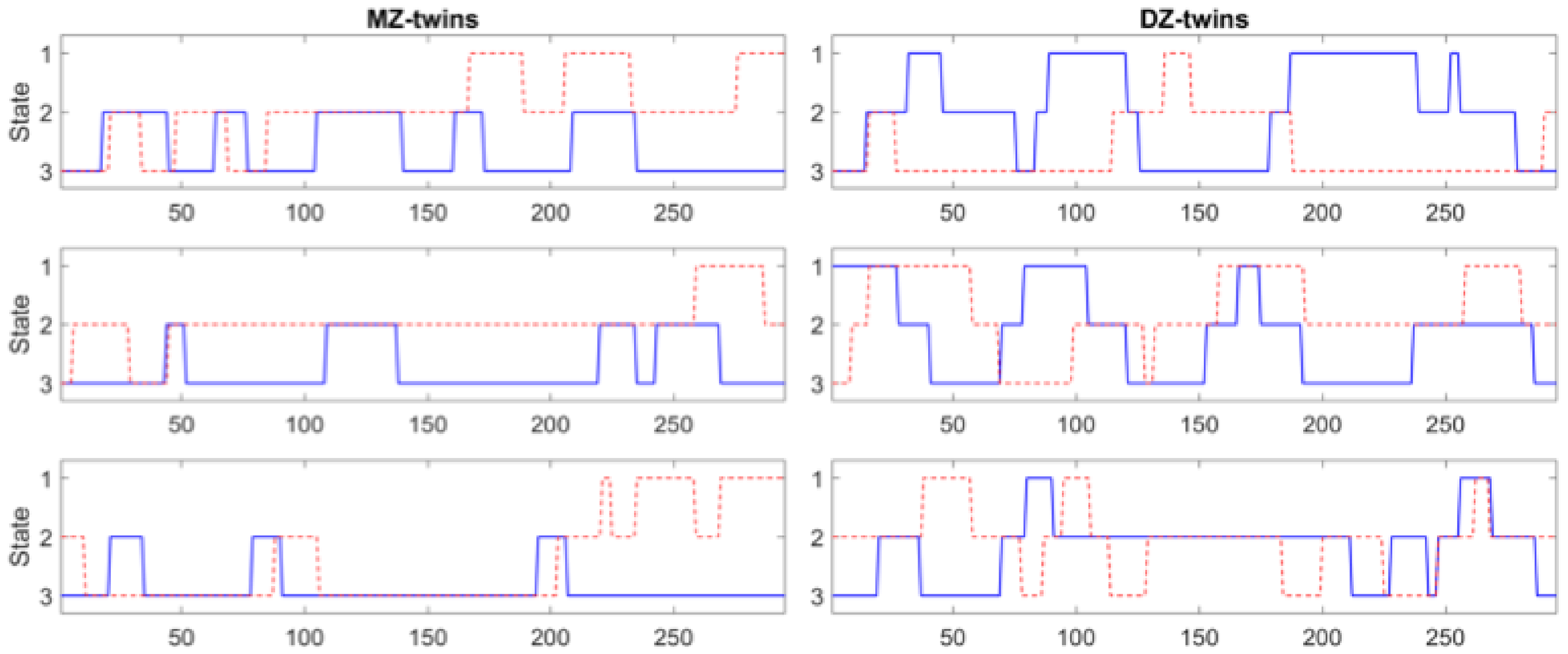
1
0.98
0.96
0.94
0.92
0.9
0.88
0.86
0.84
0.82
0.8

Topological clustering *Topological mean in each state*

State space estimation on 479 subjects

Chung at al. 2023 arXiv:2201.00087

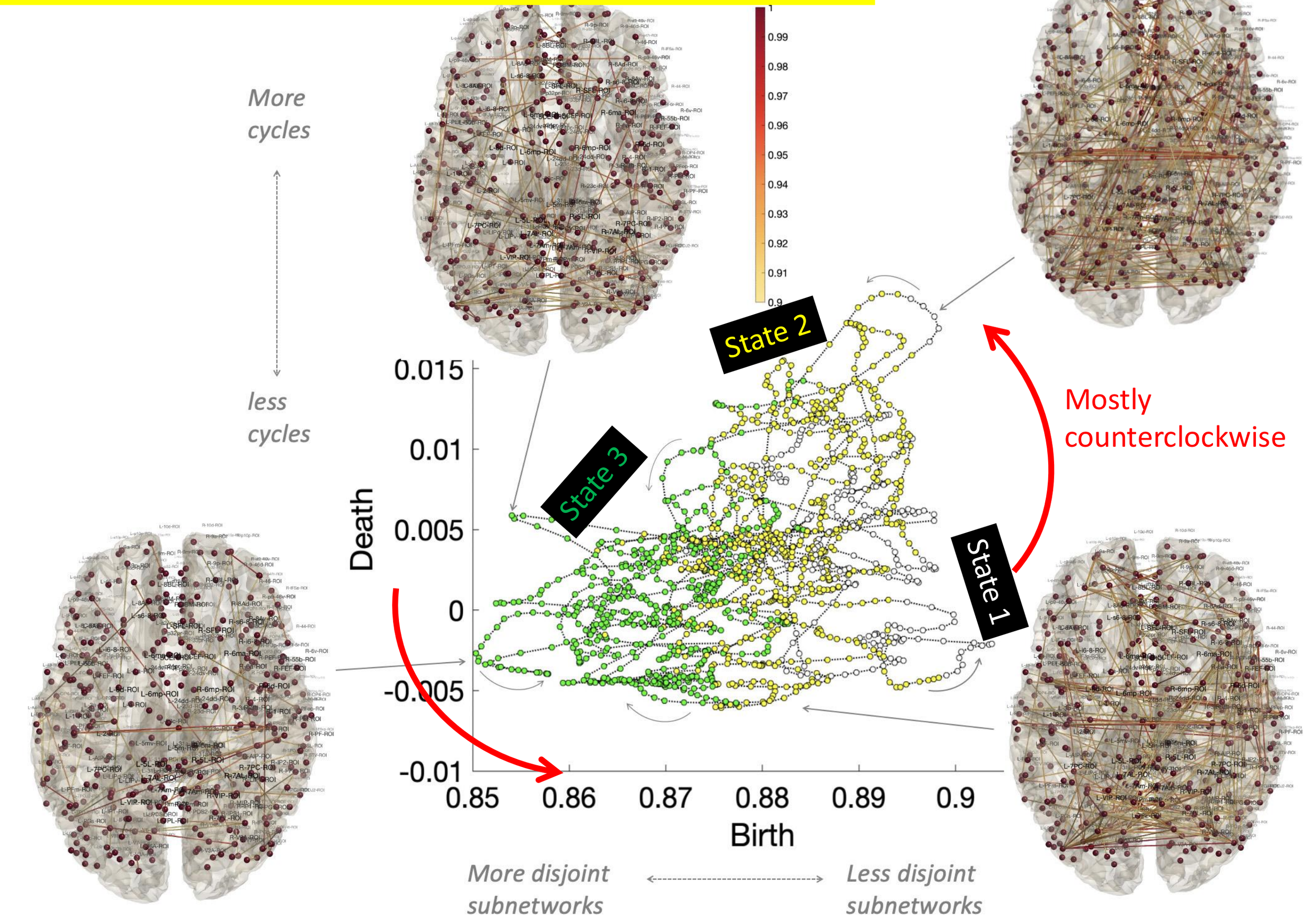
Is the state-change heritable?



UW-Madison twin study (200 twin pairs)

Chung et al. 2024 arXiv:2201.00087 (PLOS Computational

Interpreting clustering results is often needed



Distances

Wasserstein distance between networks

$$C_1 \cup C_2 \cup \cdots \cup C_k = \{\mathcal{X}_1, \cdots, \mathcal{X}_n\}, \quad C_i \cap C_j = \emptyset$$

Between-group distance

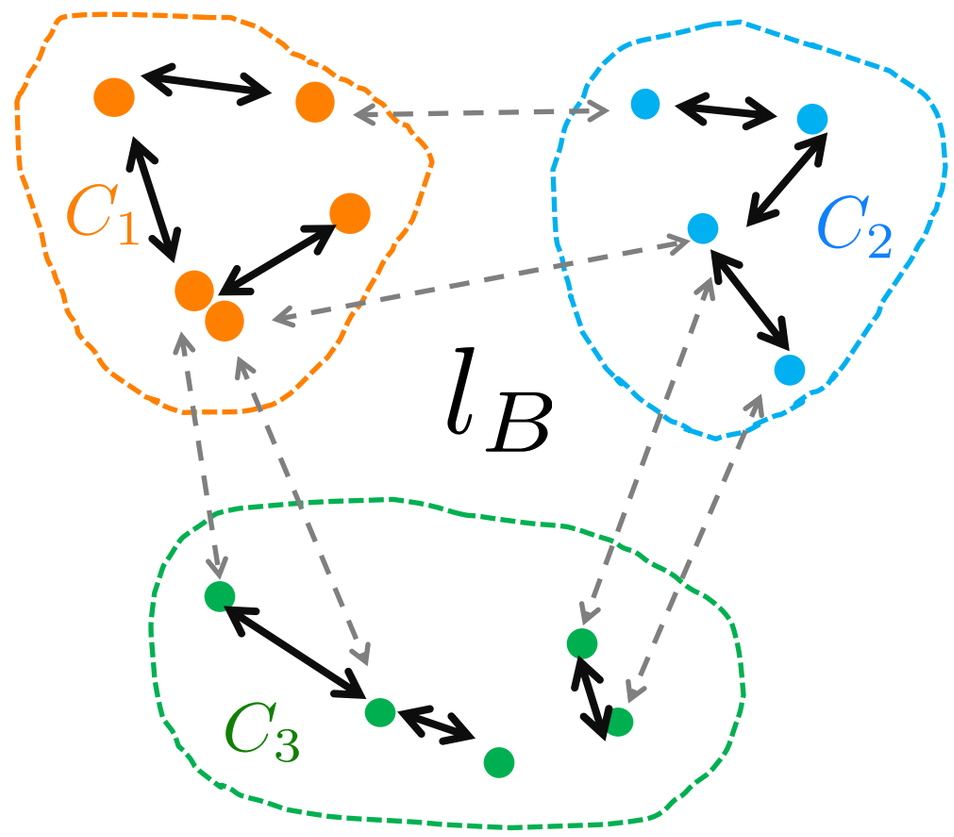
$$l_B \propto \sum_{i \in C_1, j \in C_2} \mathcal{L}(\mathcal{X}_i, \mathcal{X}_j)$$

l_W

Within-group distance

$$l_W \propto \sum_k \sum_{i,j \in C_k} \mathcal{L}(\mathcal{X}_i, \mathcal{X}_j)$$

$$l_B + l_W = \sum_{i,j} \mathcal{L}(\mathcal{X}_i, \mathcal{X}_j)$$



Clustering

Minimize the within cluster distance over all possible clustering configurations

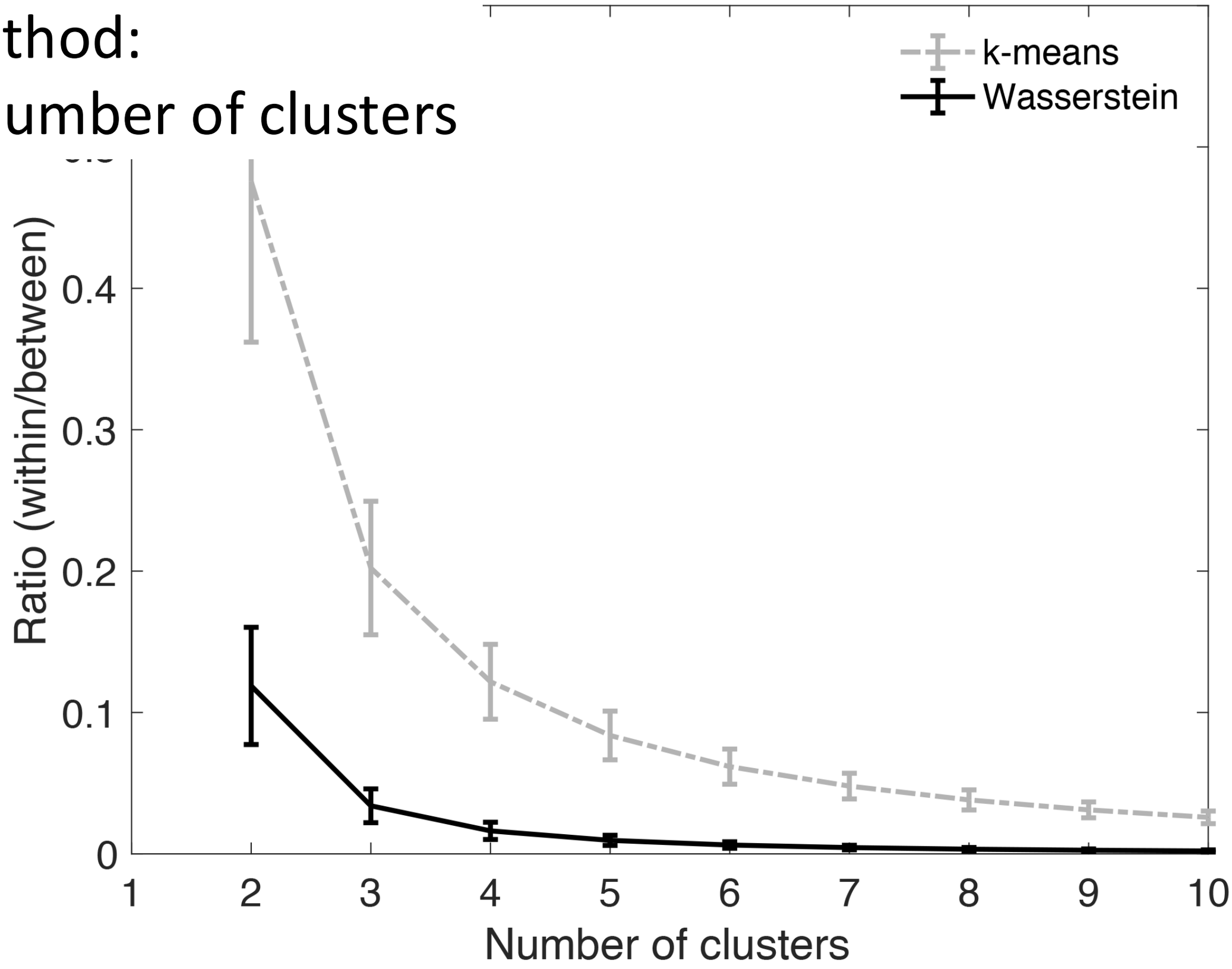
$$l_W \propto \sum_k \sum_{i,j \in C_k} \mathcal{L}(\mathcal{X}_i, \mathcal{X}_j)$$

Equivalently, maximize the between cluster distance

$$l_B = \sum_{i \in C_1, j \in C_2} \mathcal{L}(\mathcal{X}_i, \mathcal{X}_j)$$

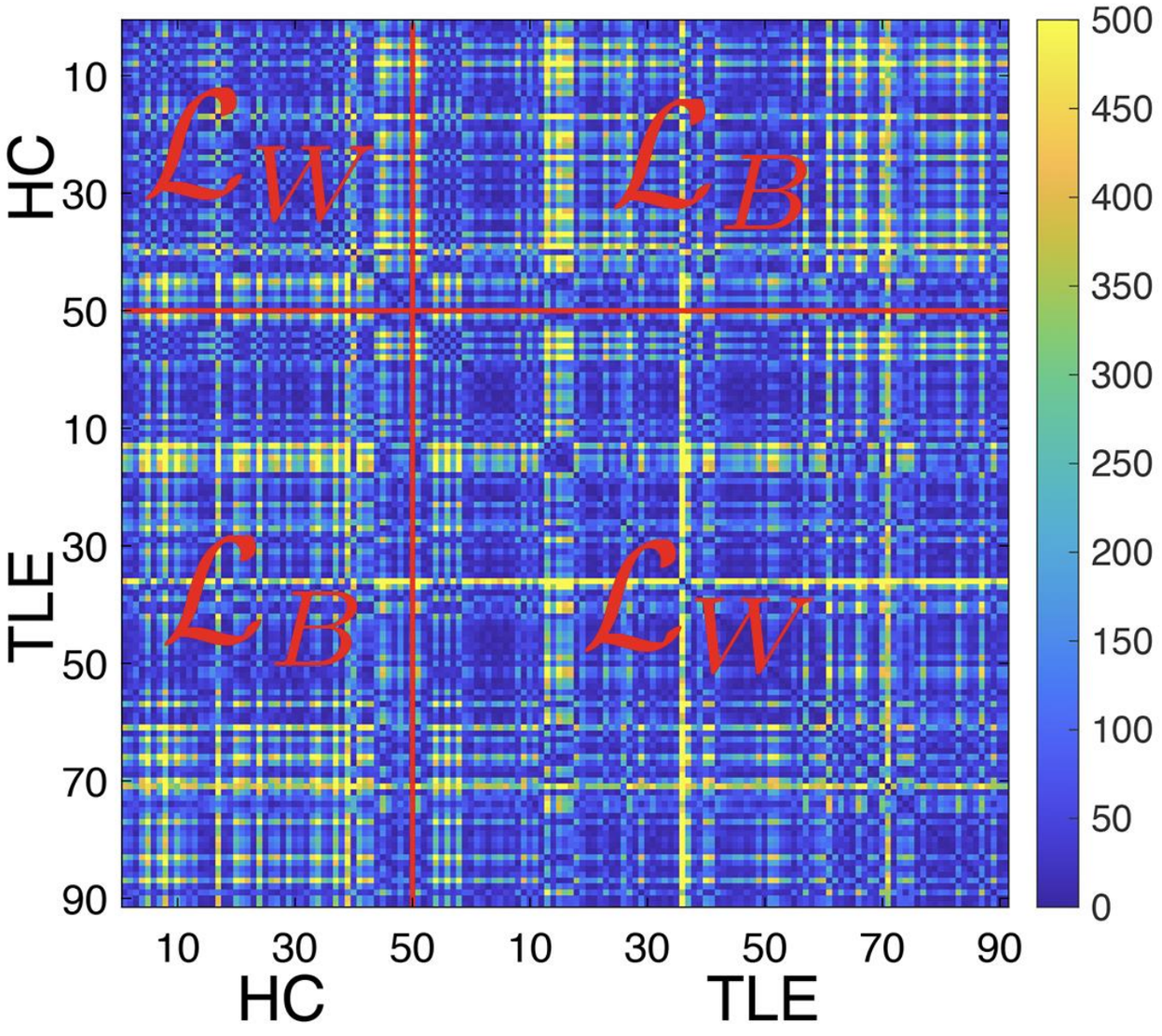
Elbow method:
Optimal number of clusters

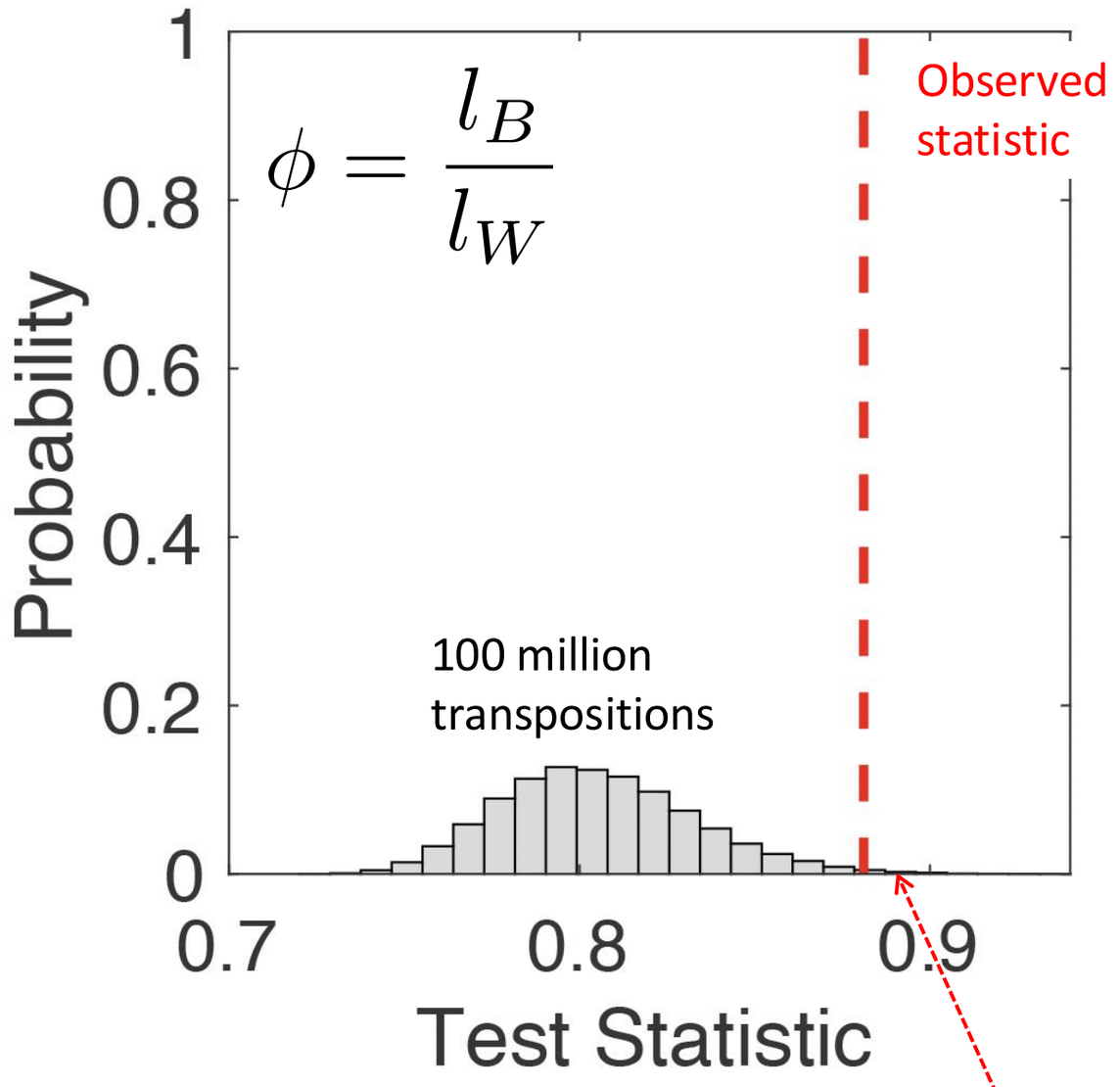
$$\frac{1}{\phi} = \frac{l_W}{l_B}$$



The within cluster variance *6 times* smaller

Statistical inference on the equivalence of clusters





P-value 0.0086

$$l_W \rightarrow l_W + \Delta(tranposition)$$

$$l_B \rightarrow l_B + \Delta(tranposition)$$

Songdechakraiwut and Chung 2023

Mathematical equivalence of topological clustering and topological inference

There exists a monotonically decreasing function f satisfying

p-value = f (clustering accuracy)

Inference

Clustering

Proof in Chung et al. 2023 NeuroImage

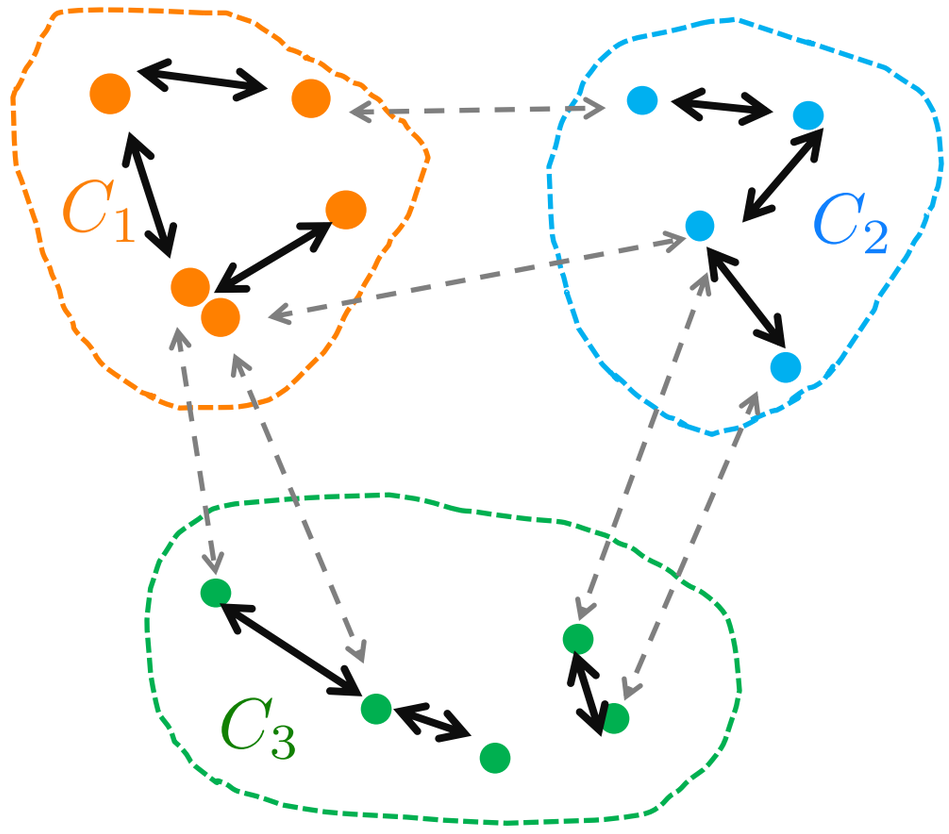
Topological clustering = topological inference

Within-group distance $d_W \propto \sum_k \sum_{i,j \in C_k} d_0(\mathcal{X}_i, \mathcal{X}_j) + d_1(\mathcal{X}_i, \mathcal{X}_j)$

Clustering accuracy

$$\frac{dA}{dp} = \frac{dA}{d(d_W)} \frac{d(d_W)}{dp} \leq 0$$

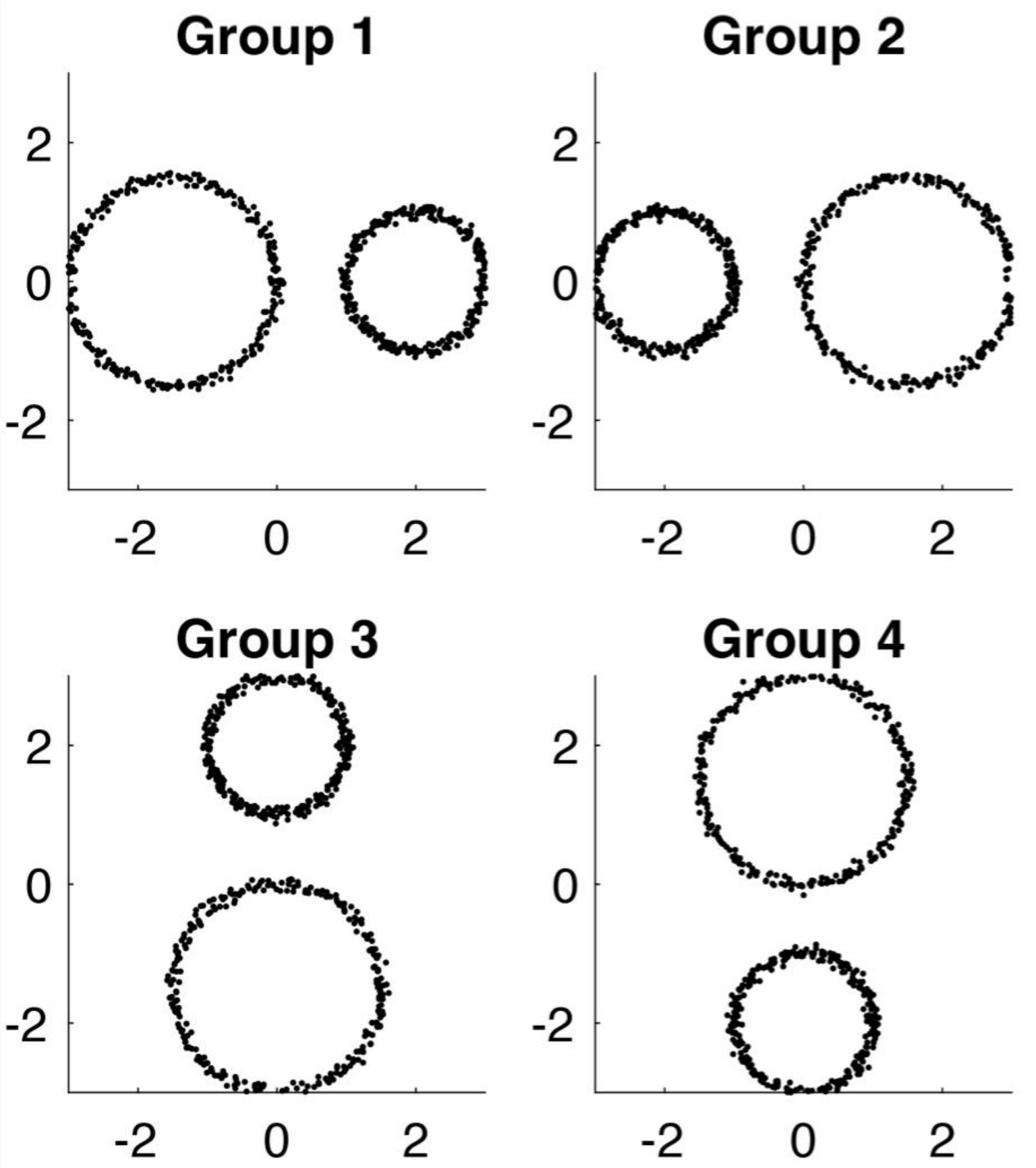
p-value



There exists a monotone decreasing function f satisfying p = f (A)

Sketch of Proof in Chung et al. 2023 NeuroImage

Geometric clustering methods fail topological task



All false positives

K-means clustering
0.98 +- 0.01

Hierarchical clustering
1.00 +- 0.00

Topological clustering
0.63 +- 0.04