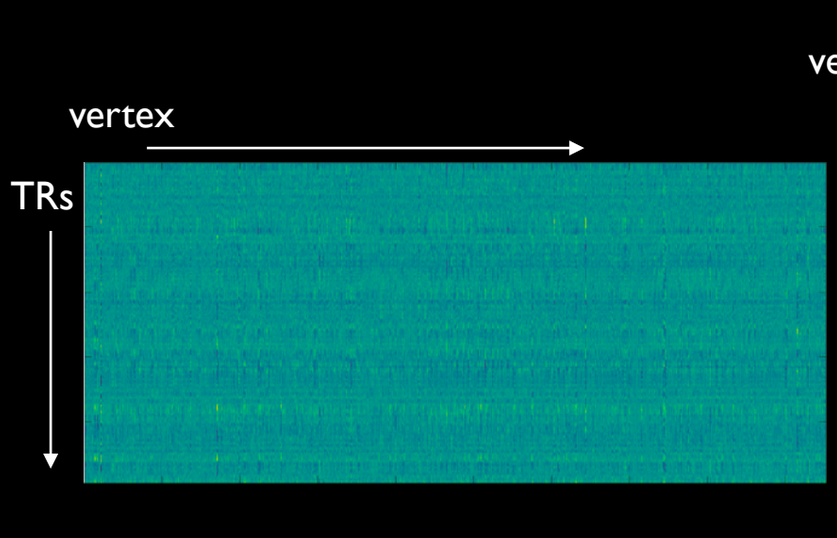
CLUSTERING TECHNIQUES

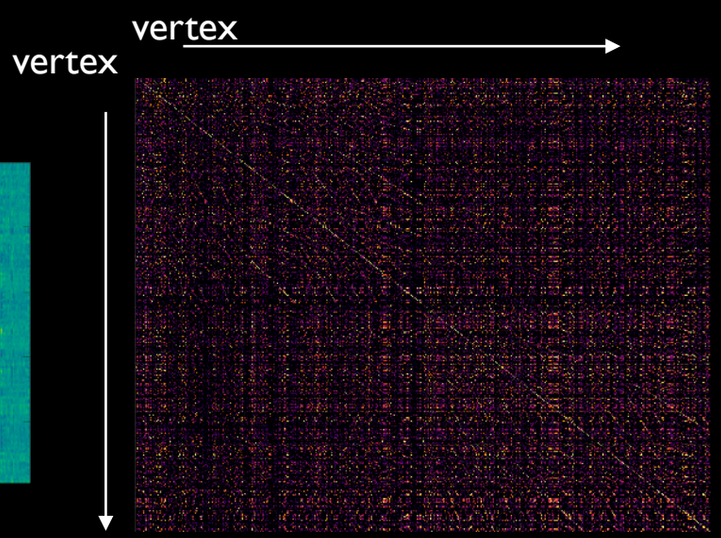
NEUR-608

BORIS BERNHARDT

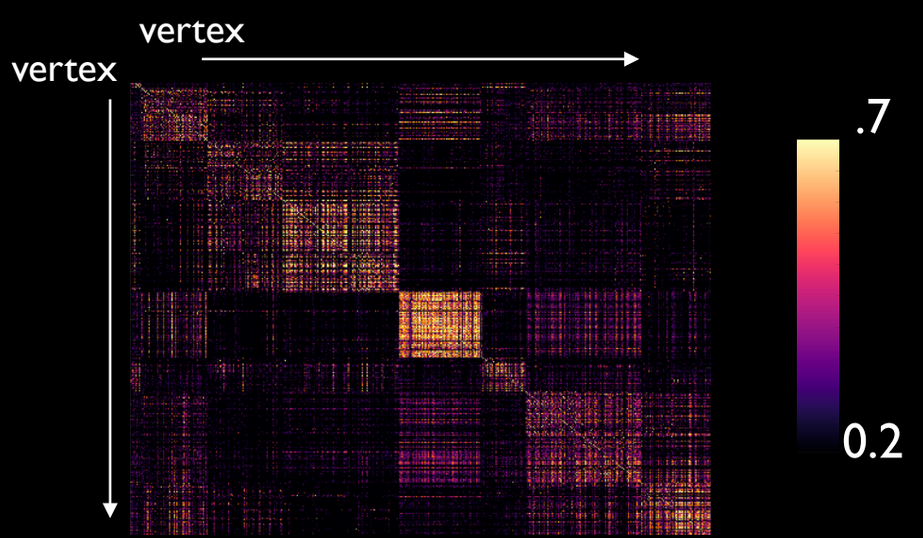
IMAGING DATA:
HIGH DIMENSIONAL AND MULTIVARIATE



Time series of
20k cortical vertices



20k x 20k
cortico-cortical z



20k x 20k
cortico-cortical z
network-sorted

HOW TO IDENTIFY NETWORKS
HOW TO REDUCE DATA DIMENSIONALITY

COMPRESSION

LINEAR (**PCA**, **ICA**, FA, MDS,…)

NON–LINEAR (LE, DME,…..)

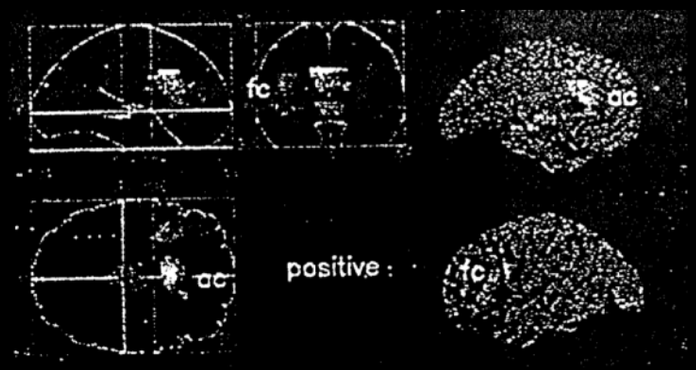
CLUSTERING

K–MEANS

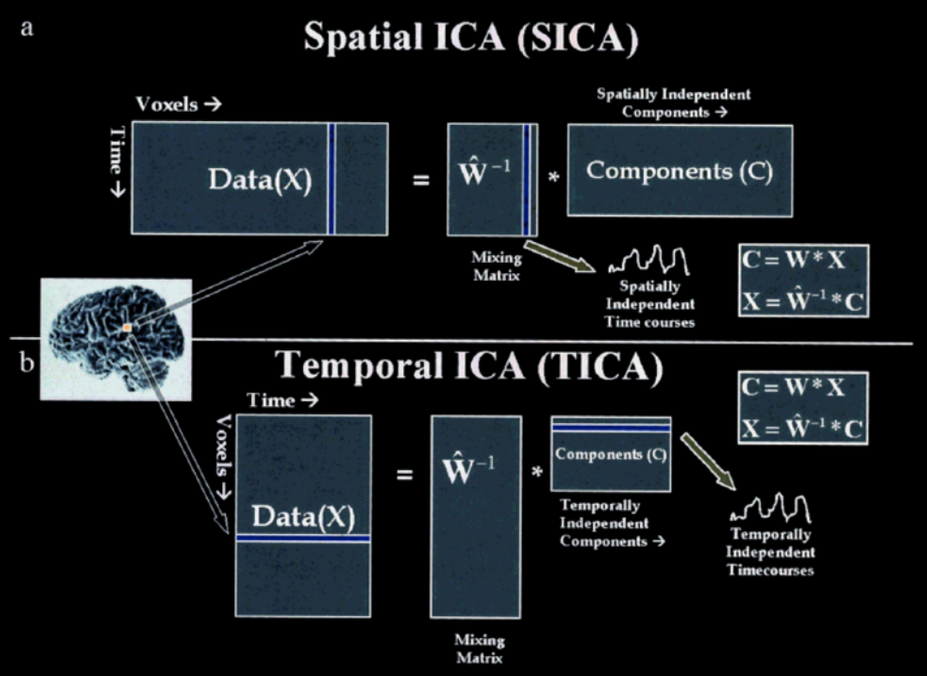
HIERARCHICAL

SPECTRAL

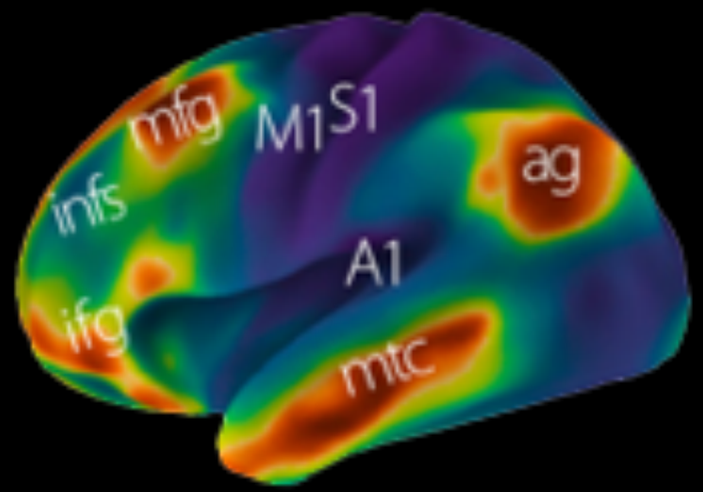
COMPRESSION

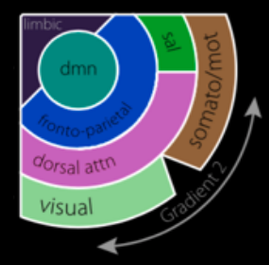


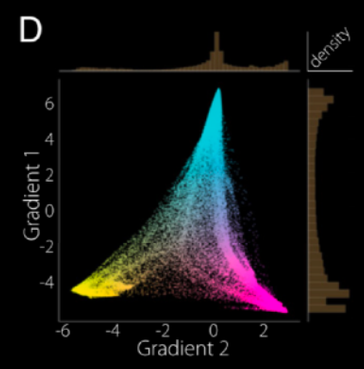
Friston 1993



McKeown et al 1998 HBM
Calhoun 2001
Beckmann 2012 NIMG

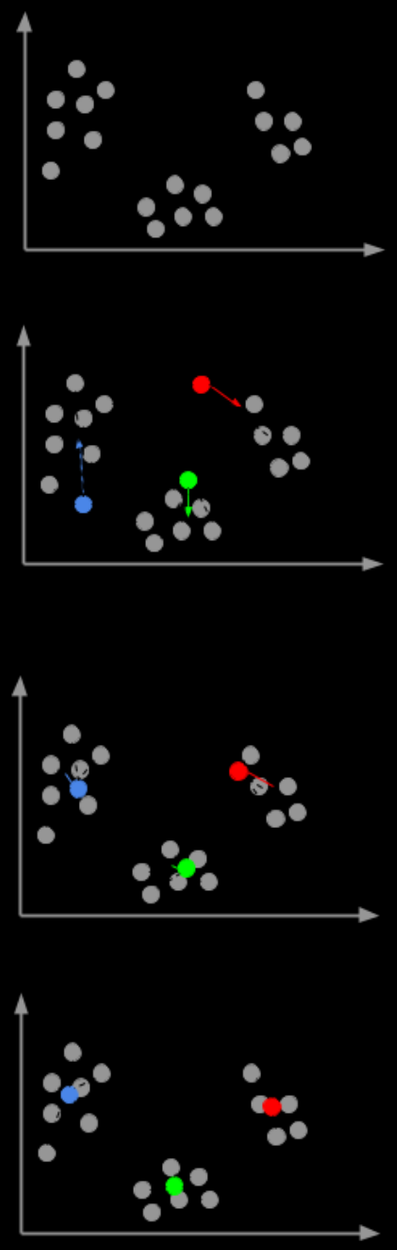




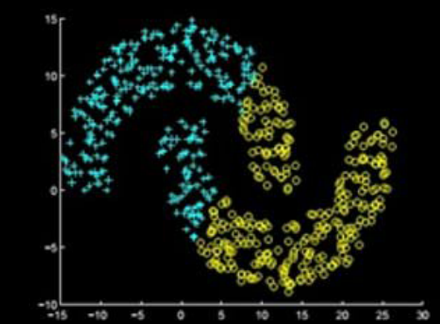


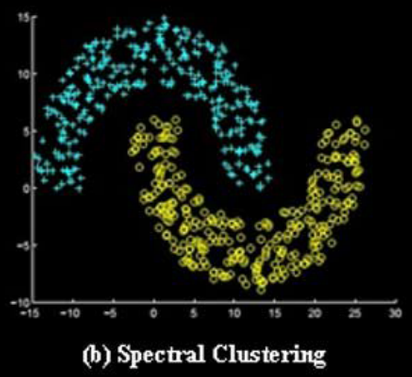
Margulies et al. 2016 PNAS

CLUSTERING ALGORITHMS

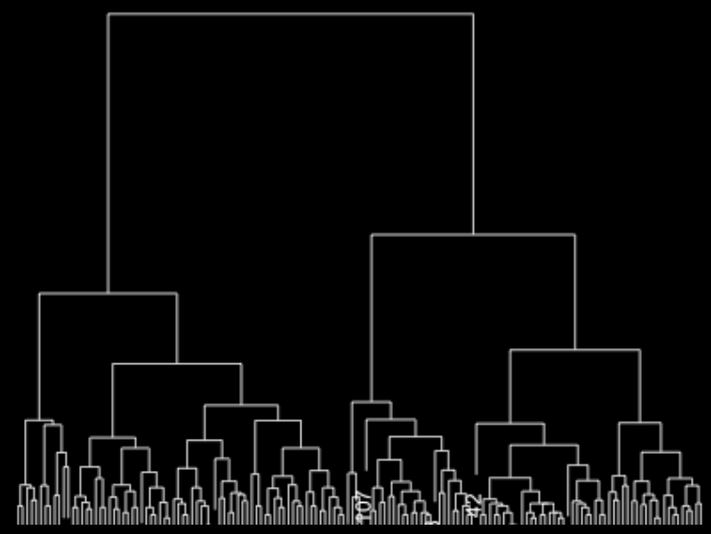


K-means clustering



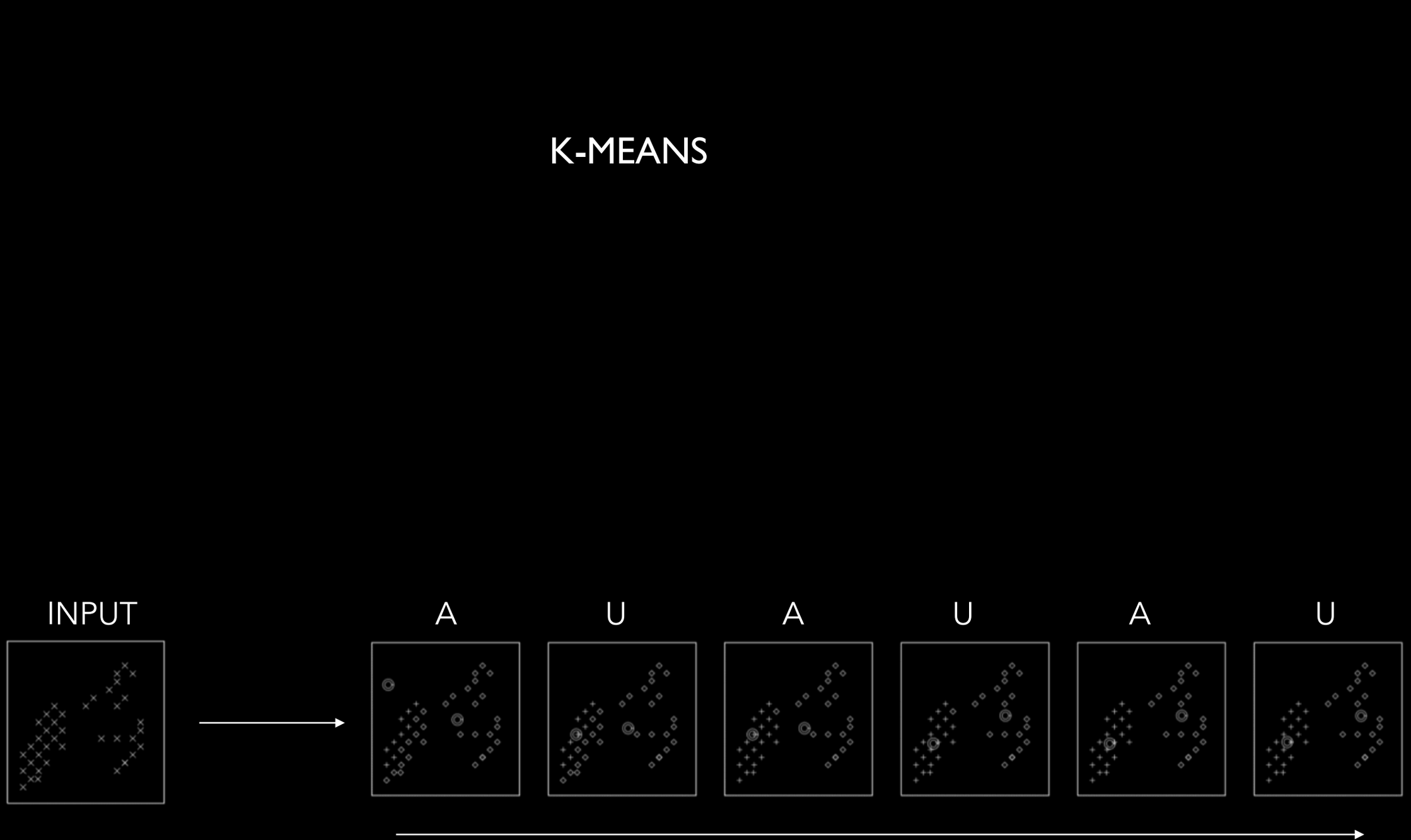


Spectral clustering
Shi Malik 2000
Von Luxburg 2007



Hierarchical clustering

K-MEANS

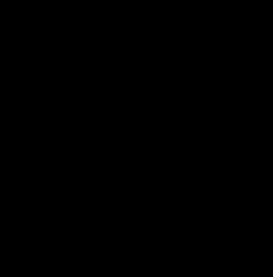


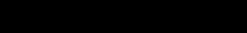
NUMBER OF CLUSTERS K

DISTANCE FUNCTION

CHALLENGE 1: INITIALIZATION

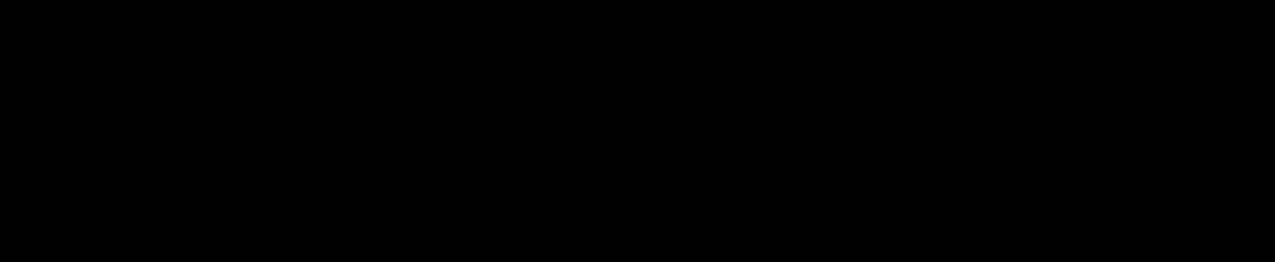
INPUT





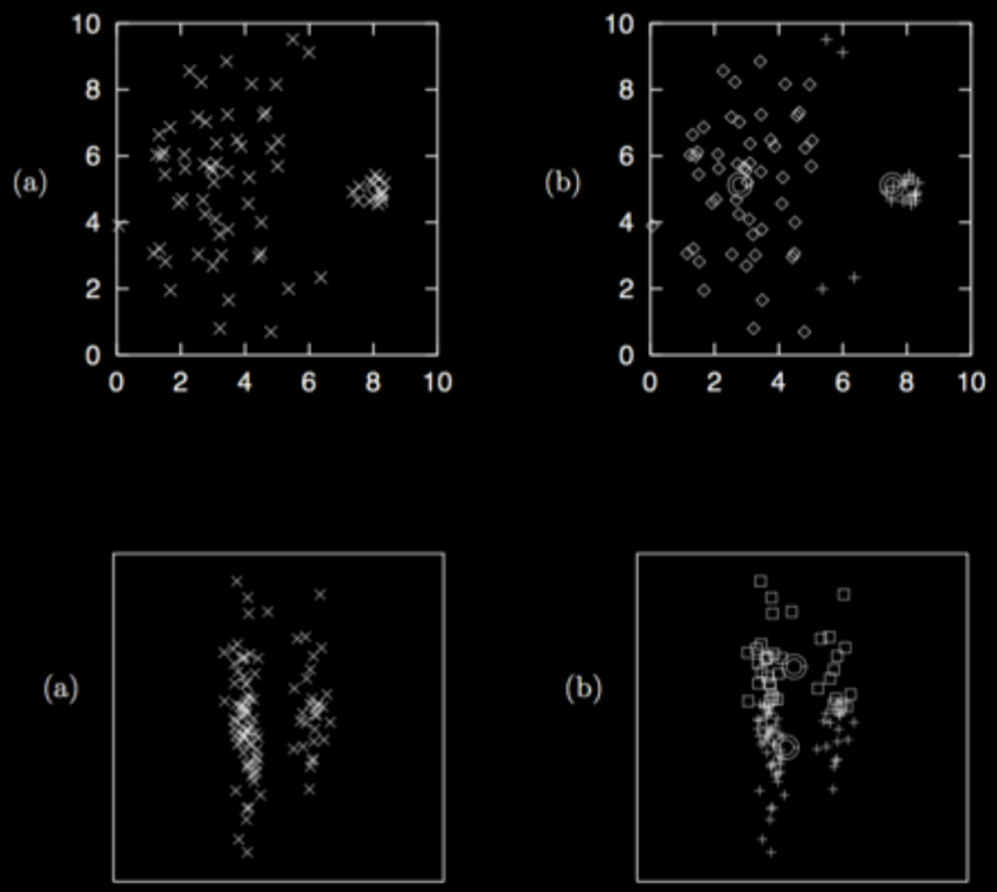
Run 1

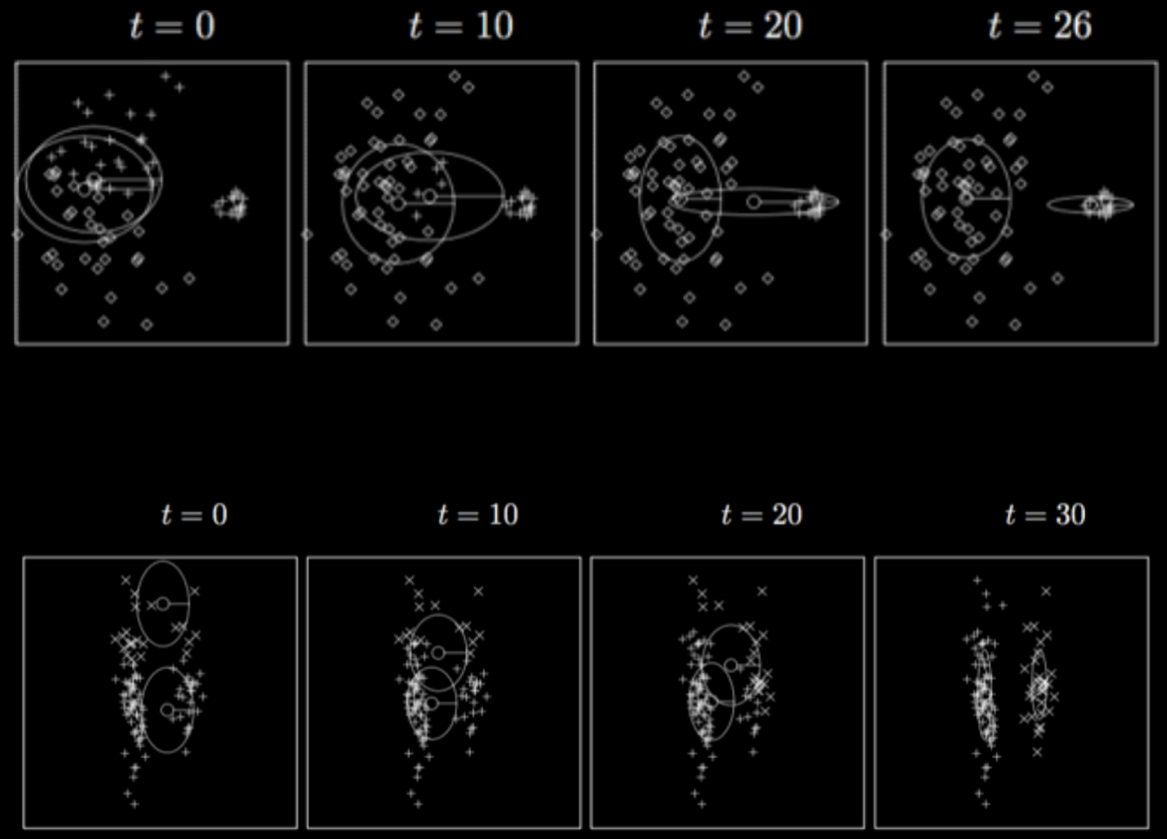
Run 2



SOME SOLUTIONS: RUN MULTIPLE TIME AND IDENTIFY MOST STABLE SOLUTION; BOOTSTRAP

CHALLENGE 2:VERY DISSIMILAR CLUSTERS





SOME SOLUTIONS: SOFTEN THE ASSIGNMENT RULES

ON DISTANCE MEASURES

A ZOO OF DISTANCE MEASURES

EUCLIDEAN, MANHATTAN, CHEBYCHEW
MINKOWSKI, CANBERRA, MAXIMUM, COSINE

Euclidean Distance

$$\sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

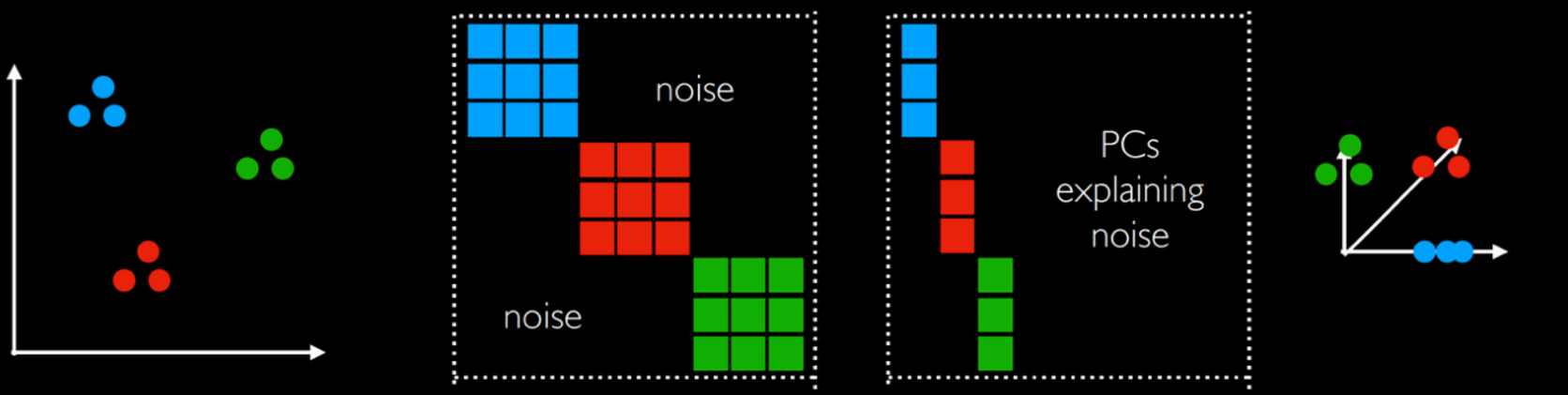
Manhattan Distance

$$|x_1 - x_2| + |y_1 - y_2|$$

Chebyshev Distance

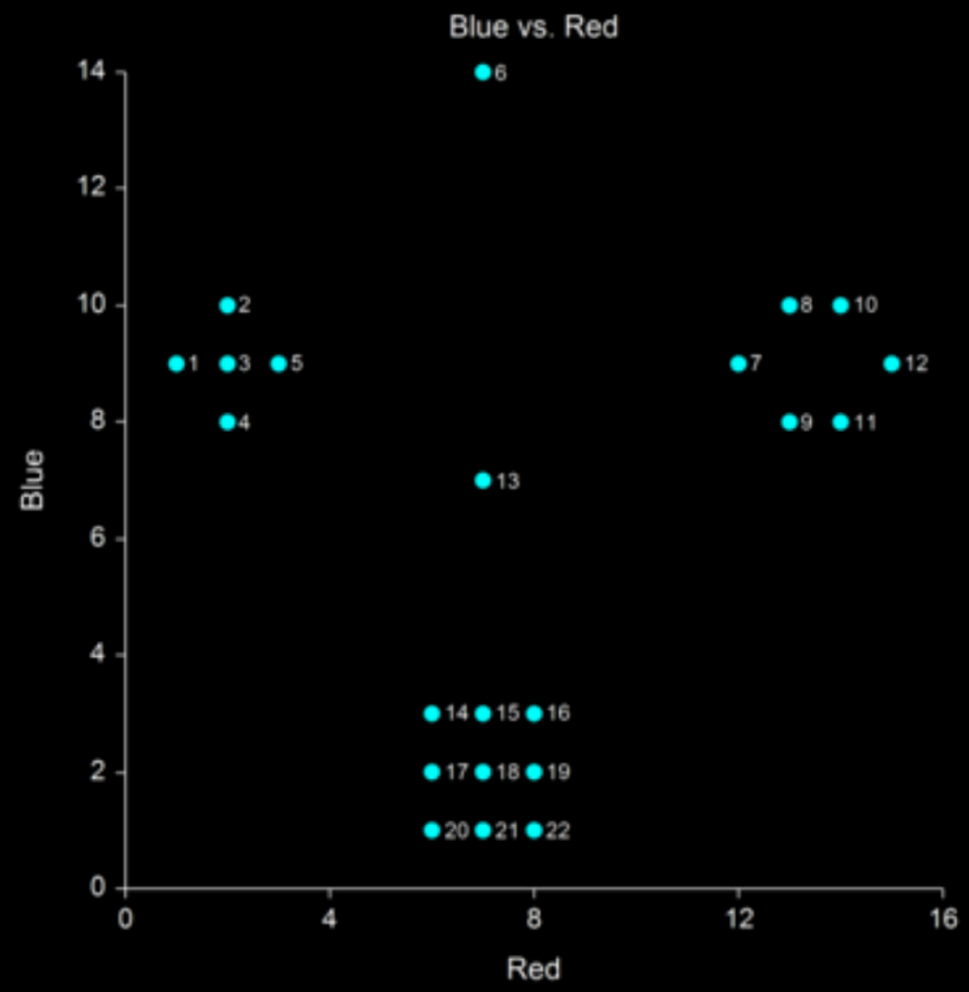
$$\max(|x_1 - x_2|, |y_1 - y_2|)$$

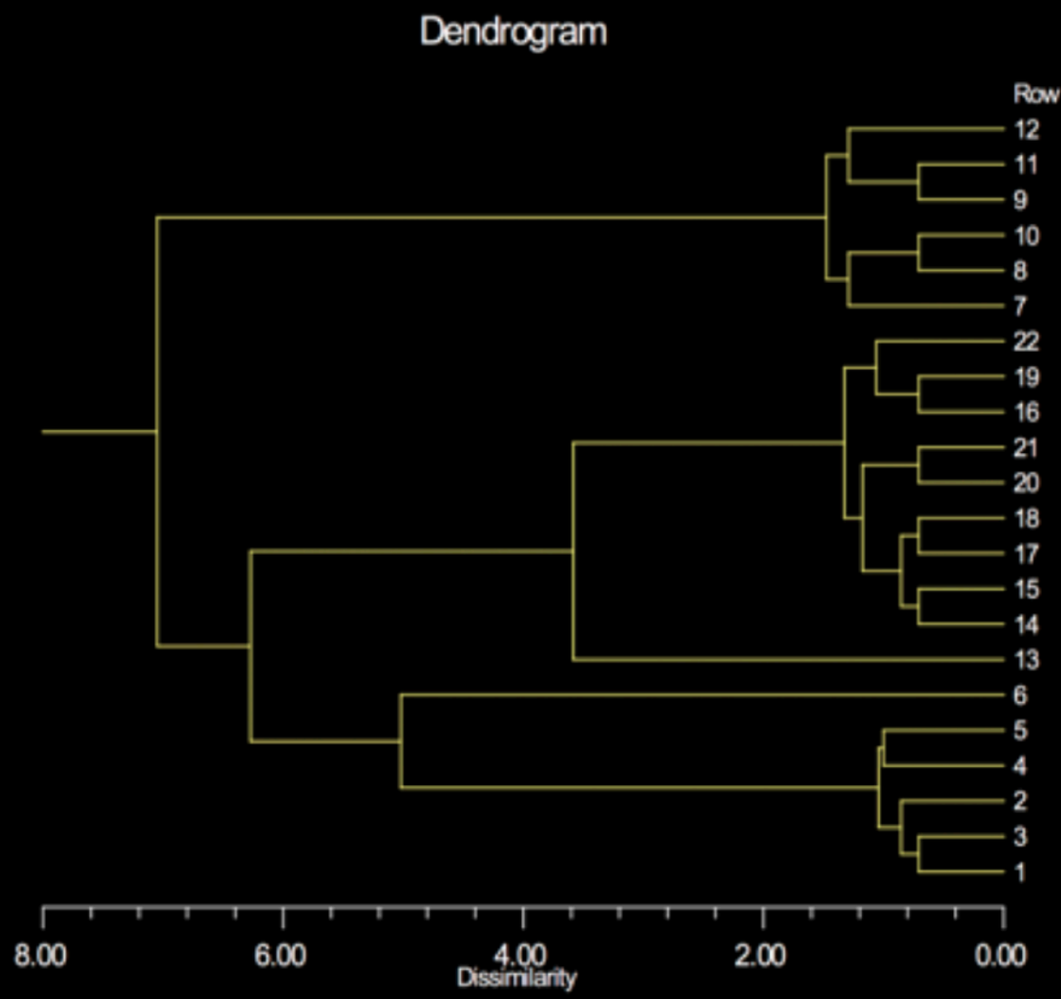
SPECTRAL CLUSTERING



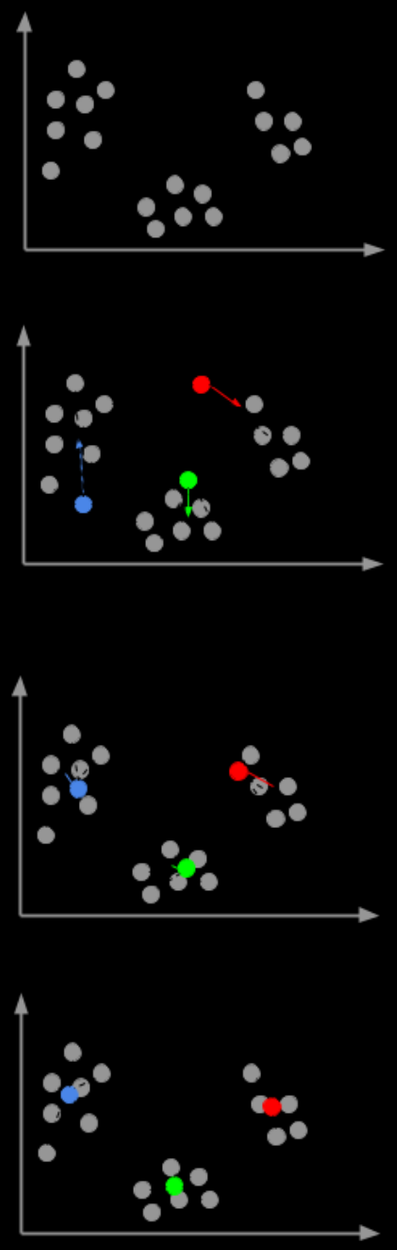
DATA AFFINITY PCA CLUSTERING IN PCA-SUBSPACE

HIERARCHICAL CLUSTERING

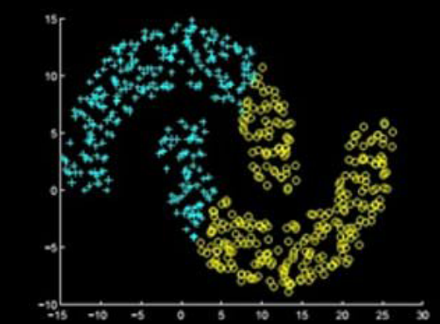


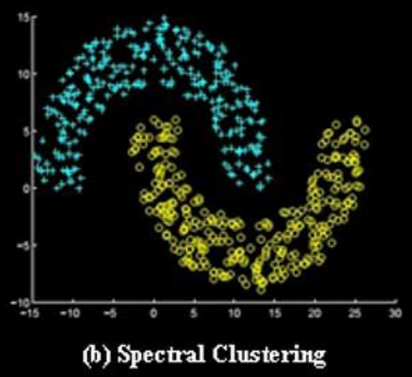


CLUSTERING ALGORITHMS

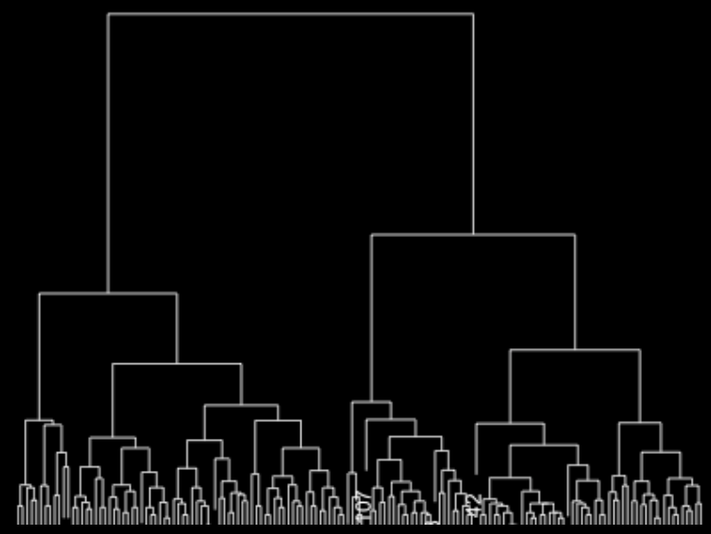


K-means clustering



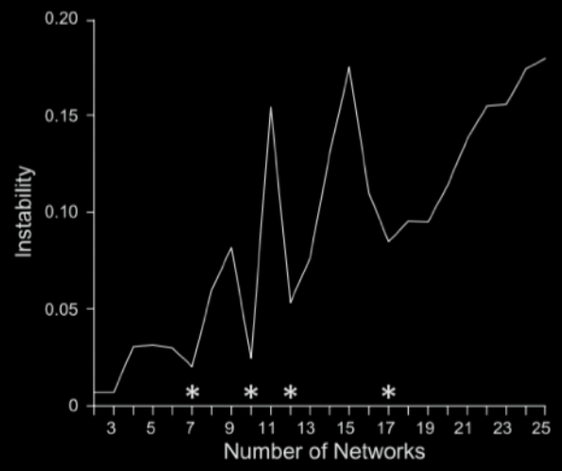


Spectral clustering
Shi Malik 2000
Von Luxburg 2007

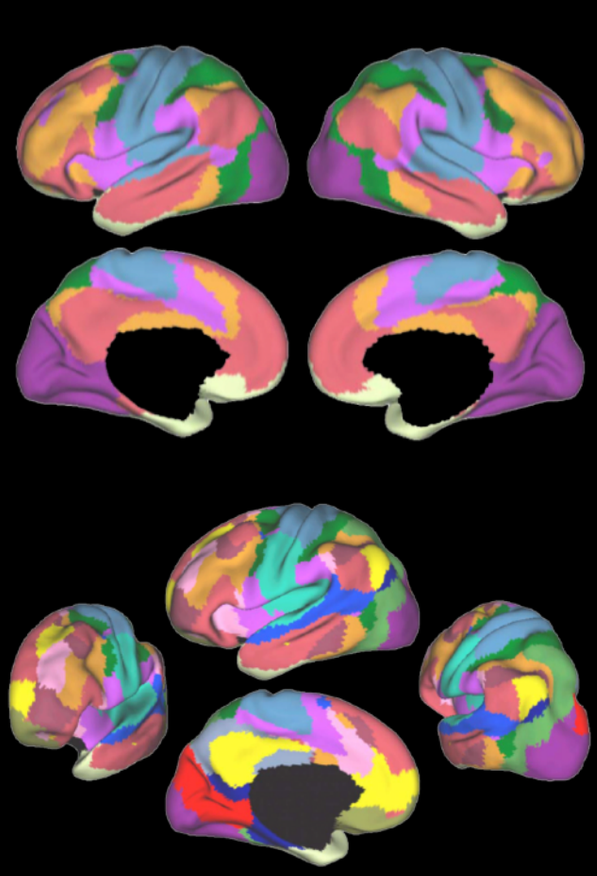


Hierarchical clustering

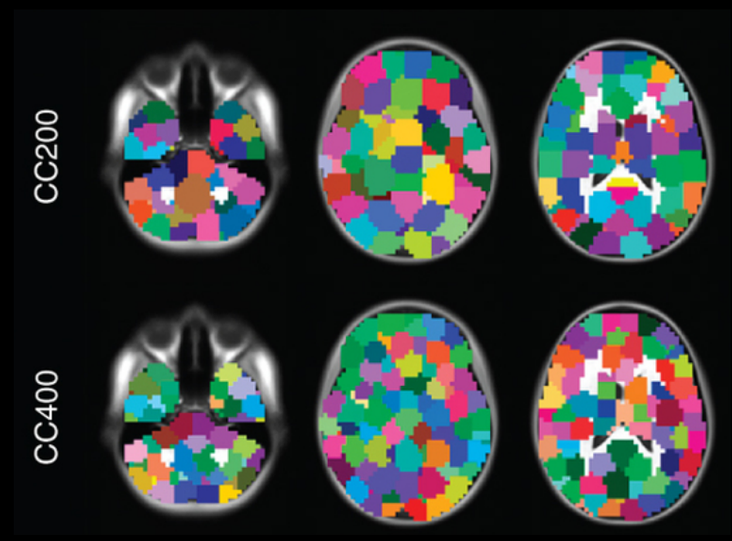
NETWORK CLUSTERING APPLICATIONS



$$w_{ij} = \begin{matrix} s(v_i, v_j) & d_{ij} \leq \varepsilon \\ 0 & d_{ij} > \varepsilon \end{matrix} .$$

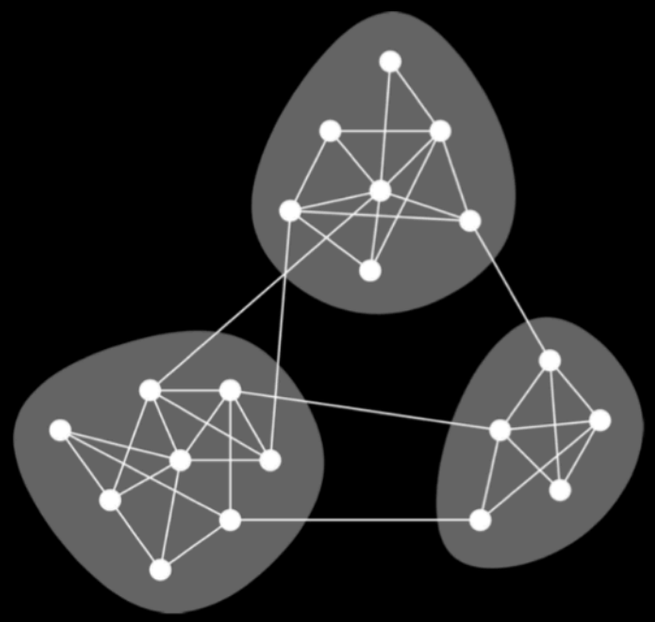


Yeo, Krienen et al 2011 J Neurophysiol



Craddock et al. 2011 NIMG

MODULARITY DETECTION



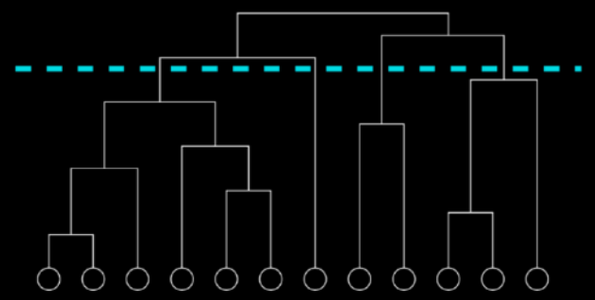


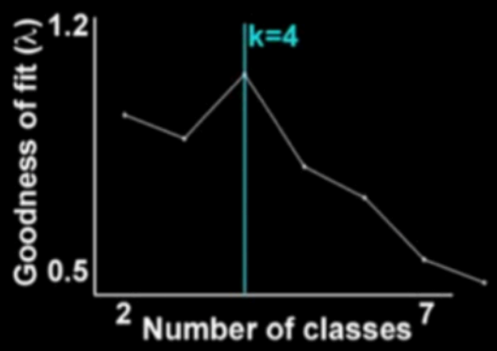
Fig. 1. The vertices in many networks fall naturally into groups or communities, sets of vertices (shaded) within which there are many edges, with only a smaller number of edges between vertices of different groups.

Girvan and Newman 2002, Newman 2006 PNAS, Blondel et al. 2008, Fortunato 2009

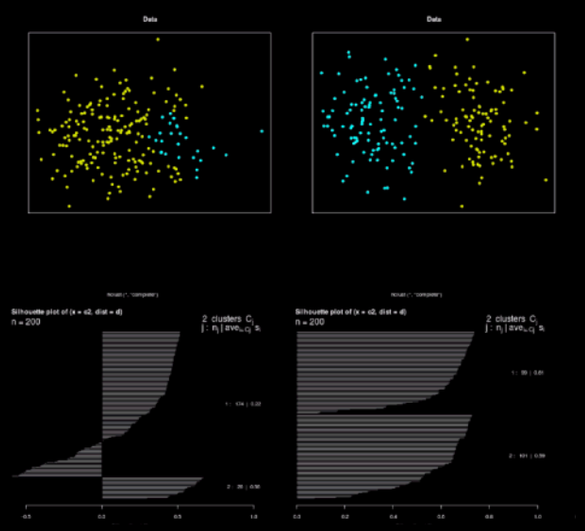
GRAPH THEORY WEEKS

HOW MANY KS?

VARIANCE
WITHIN/BETWEEN CLUSTERS



SILHOUETTE



MULTI-CRITERIA

NbClust {NbClust}

NbClust Package for determining the best number of clust

Description

NbClust package provides 30 indices for determining the number of clusters and proposes to user the scheme from the different results obtained by varying all combinations of number of clusters, distanc clustering methods.

Usage

```
NbClust(data, diss = NULL, distance = "euclidean", min.nc = 2, max.nc =15,
method = "ward.D2", index =  "all", alphaBeale = 0.1)
```

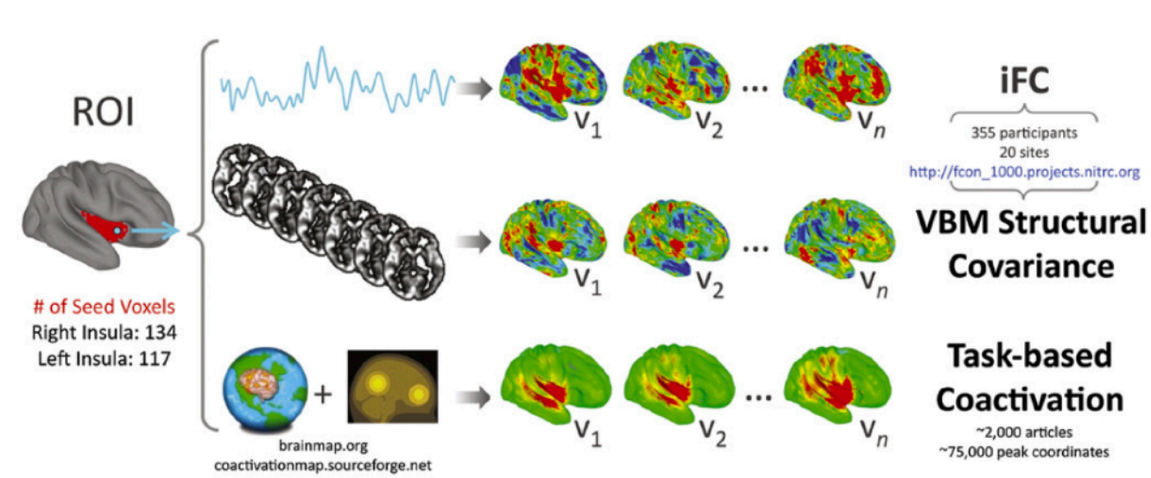
Arguments

data	matrix or dataset (the only mandatory argument)
diss	dissimilarity matrix to be used. By default, diss=NULL, but if it is replaced by a dissimil distance should be "NULL".
distance	the distance measure to be used to compute the dissimilarity matrix. This must be one o "maximum", "manhattan", "canberra", "binary", "minkowski" or "NULL". By default, d If the distance is "NULL", the dissimilarity matrix (diss) should be given by the user. If "NULL", the dissimilarity matrix should be "NULL".
min.nc	minimal number of clusters, between 1 and (number of objects - 1)

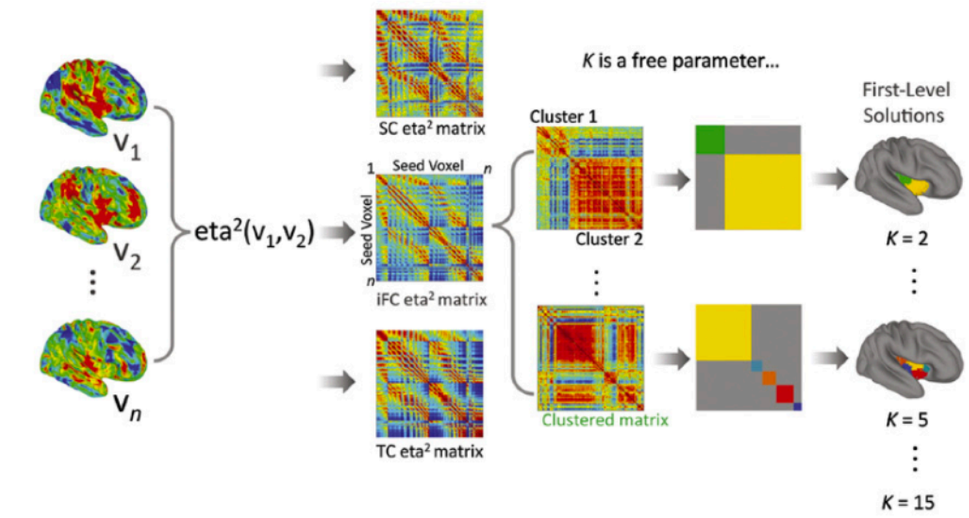
STABILITY
REPLICABILITY

KELLY ET AL

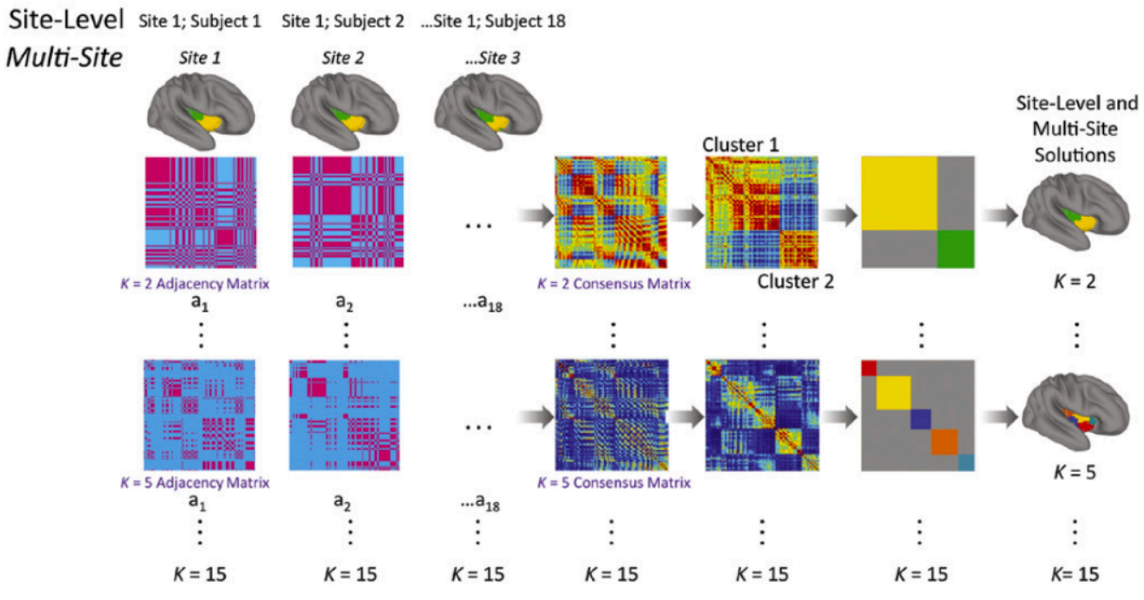
Step 1: Covariance-Based Measures



Step 2: eta^2 and First-Level Clustering



Step 3: Consensus (Site-Level and Multi-Site) Clustering



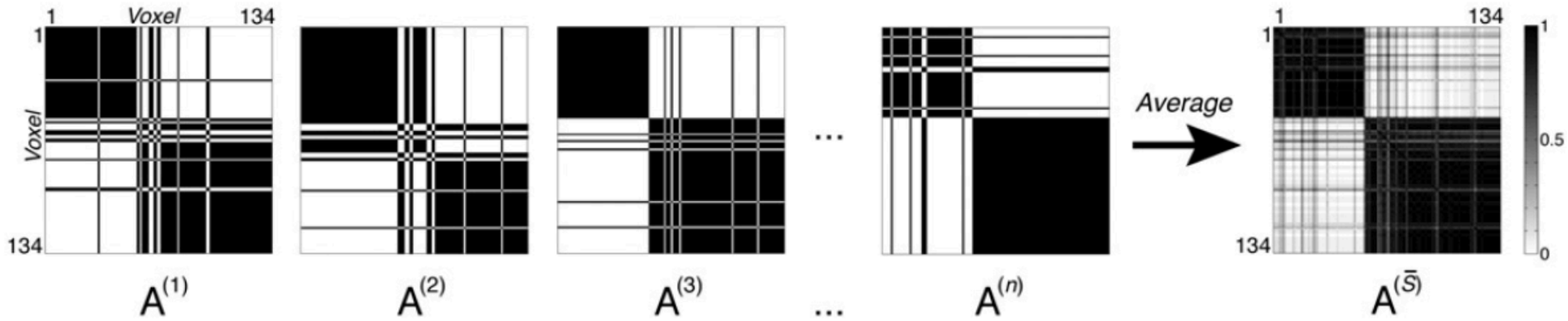
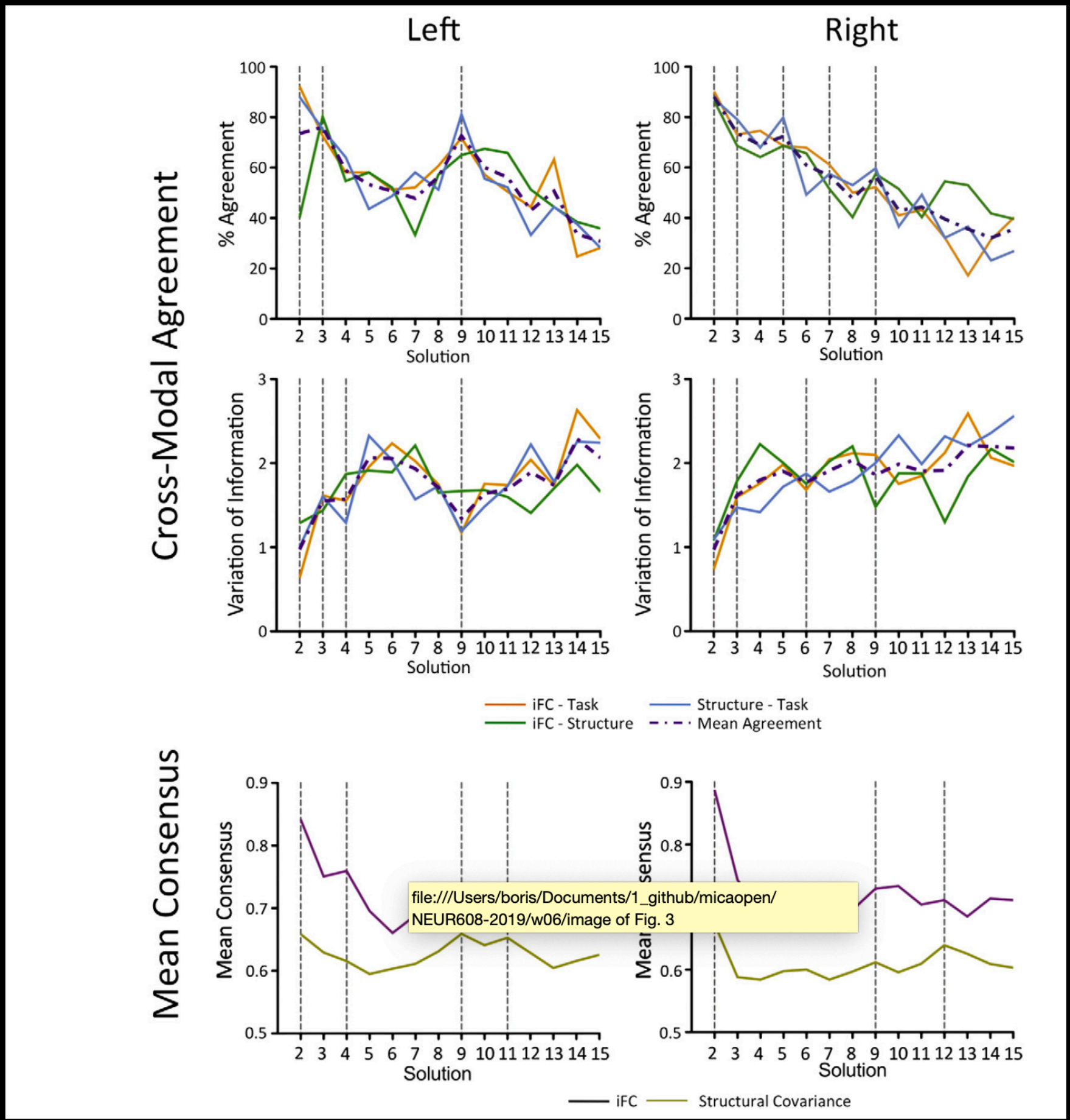
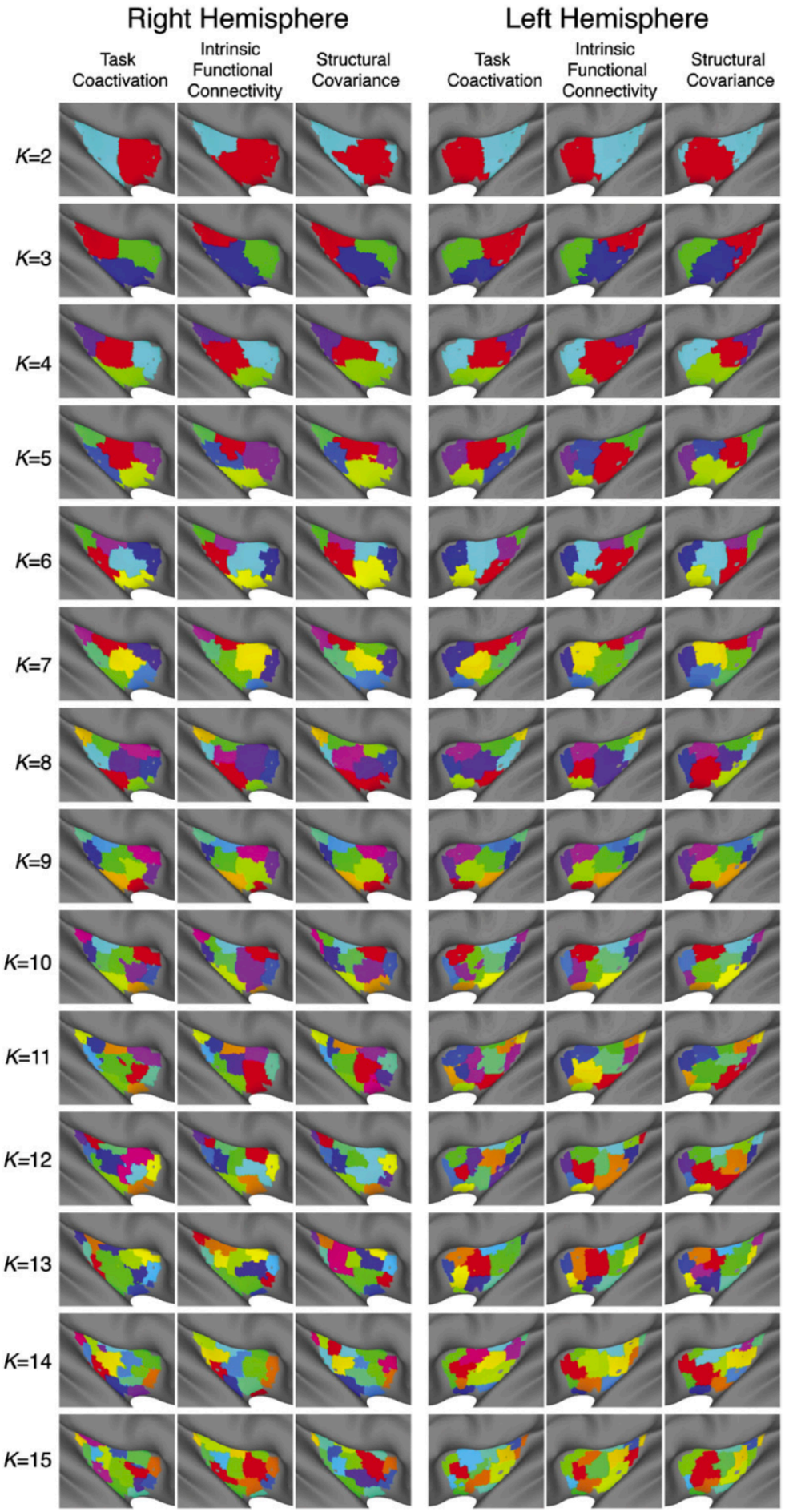


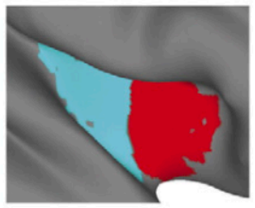
Fig. 2. Consensus clustering schematic for $K=2$. The schematic illustrates the consensus clustering process. For each scale K, each clustering instance contributes an adjacency matrix $A^{(s)}$, each element $a_{ij}^{(s)}$ of which contains a value of 1 if voxels i and j are assigned to the same cluster k, and 0 otherwise. In this example, let each instance be a data collection site, so $A^{(1)}$ is contributed by Bangor; $A^{(2)}$ is contributed by Berlin, etc. A consensus matrix $A^{(\bar{S})}$ is derived by averaging across adjacency matrices. Each element of the consensus matrix thus contains a number between 0 and 1, corresponding to the proportion of times a given pair of voxels appeared in the same cluster, across instances (here, data collection sites). The spectral clustering algorithm can then be applied to identify the most stable pattern of cluster assignments across instances, using the same scale K that was used to generate the consensus matrix (here, $K=2$).


Left
Right
Cross-Modal Agreement
% Agreement
Variation of Information
Solution
iFC - Task
iFC - Structure
Structure - Task
Mean Agreement
Mean Consensus
iFC
Structural Covariance
file:///Users/boris/Documents/1_github/micaopen/NEUR608-2019/w06/image of Fig. 3


Right Hemisphere
Left Hemisphere
Task Coactivation
Intrinsic Functional Connectivity
Structural Covariance
Task Coactivation
Intrinsic Functional Connectivity
Structural Covariance
K=2
K=3
K=4
K=5
K=6
K=7
K=8
K=9
K=10
K=11
K=12
K=13
K=14
K=15

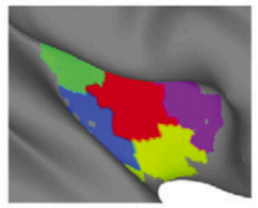
Right Hemisphere

K=2



Cluster	Domain	Subdomain
1	Action Perception *Interoception*	Execution, *Rest* Somesthesis, *Gustation* Sexuality
2	Cognition	Language, Memory, *Social, Space*

K=5



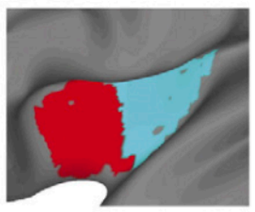
Cluster	Domain	Subdomain
1	Action *Perception*	Execution Somesthesis
2	Perception	Somesthesis, Gustation
3	Action *Perception* *Interoception*	Execution, *Motor, Rest* *Somesthesis* *Sexuality, Hunger*
4	Cognition	Language, Memory
5	Emotion *Cognition*	*Anger, Disgust, Fear, Happiness* *Attention, Social*

K=9

Cluster	Domain	Subdomain
1	*Action* *Perception*	Execution Somesthesis
2	Action Perception	Execution, Motor, *Rest* Somesthesis, *Gustation*
3	Action Perception *Interoception*	Execution Somesthesis *Sexuality*
4	*Perception* *Action*	*Somesthesis, Gustation* *Execution*
5	Cognition	Language, Memory, *Space*
6	Cognition *Emotion*	Language, Memory, *Reasoning* --
7	Cognition	*Language, Memory, Social*
8	Emotion	*Disgust, Sadness, Anger, Anxiety, Happiness*
9	Emotion	*Anxiety, Disgust, Fear, Sadness*

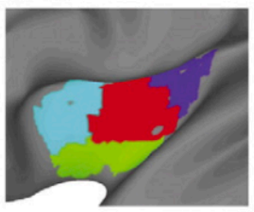
Left Hemisphere

K=2



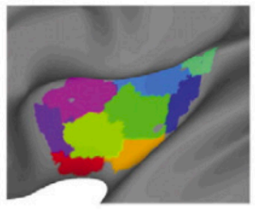
Cluster	Domain	Subdomain
1	Action Perception *Interoception*	Execution, Rest Somesthesis, *Audition* *Sexuality, Hunger*
2	Cognition *Emotion*	Memory, *Attention* --

K=4



Cluster	Domain	Subdomain
1	Perception *Action* *Interoception*	Audition, *Somesthesis* Execution *Air Hunger, Hunger*
2	Action Perception *Interoception*	Execution, *Imagination* Somesthesis *Sexuality, Air Hunger*
3	Cognition	Memory
4	Emotion	*Fear, Happiness, Anger,* *Disgust, Sadness*

K=9



Cluster	Domain	Subdomain
1	*Action* *Perception* *Pharmacology*	Execution, *Rest* Somesthesis, *Audition* *Caffeine, Capsaicin, Psychiatric*
2	*Perception* *Interoception*	Audition *Hunger, Thirst*
3	Action *Interoception*	Execution, *Imagination, Rest* *Hunger, Air Hunger*
4	Perception *Action*	Somesthesis, *Audition* Execution, *Imagination*
5	Cognition	Memory
6	Cognition	Memory, *Attention, Reasoning*
7	Emotion *Cognition*	*Fear* *Attention*
8	Emotion *Cognition*	Sadness, *Fear, Happiness, Anger, Disgust* *Memory*
9	*Interoception* *Emotion*	*Air Hunger* *Happiness, Anger*

YEO, KRIENEN ET AL

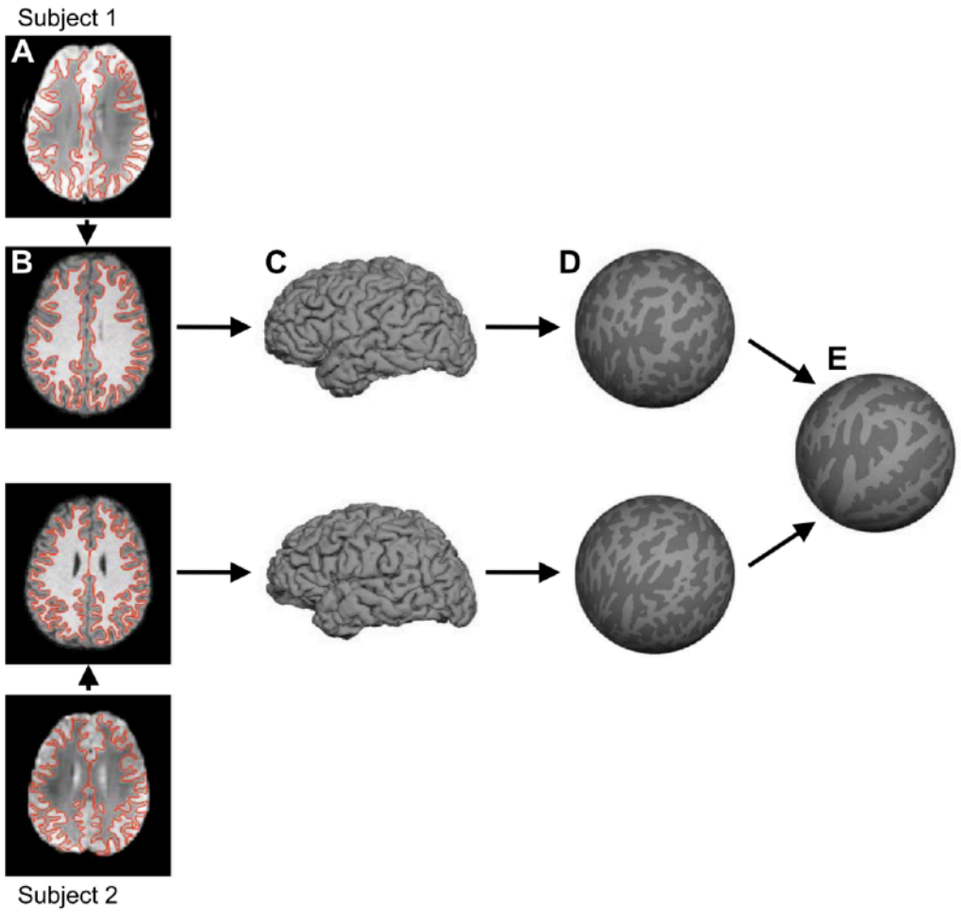


Fig. 1. Surface coordinate system for functional magnetic resonance imaging (fMRI) analysis. For each subject, the T2* images yielding blood oxygenation level-dependent (BOLD) contrast fMRI data (*A*) were registered to the T1-weighted structural data (*B*). The cortical gray-white and pial surfaces were estimated from the structural data. The red lines show the estimated gray-white surface (*A* and *B*). Pial surface is shown in *C*. The gray-white surface was inflated into a sphere (*D*). The inflated spheres were then aligned across subjects using surface-based registration of the cortical folding pattern, resulting in a common spherical coordinate system (*E*). BOLD data of individual subjects (*A*) can then be projected onto the spherical coordinate system (*E*) in a single transformation step to reduce artifacts due to multiple interpolations.

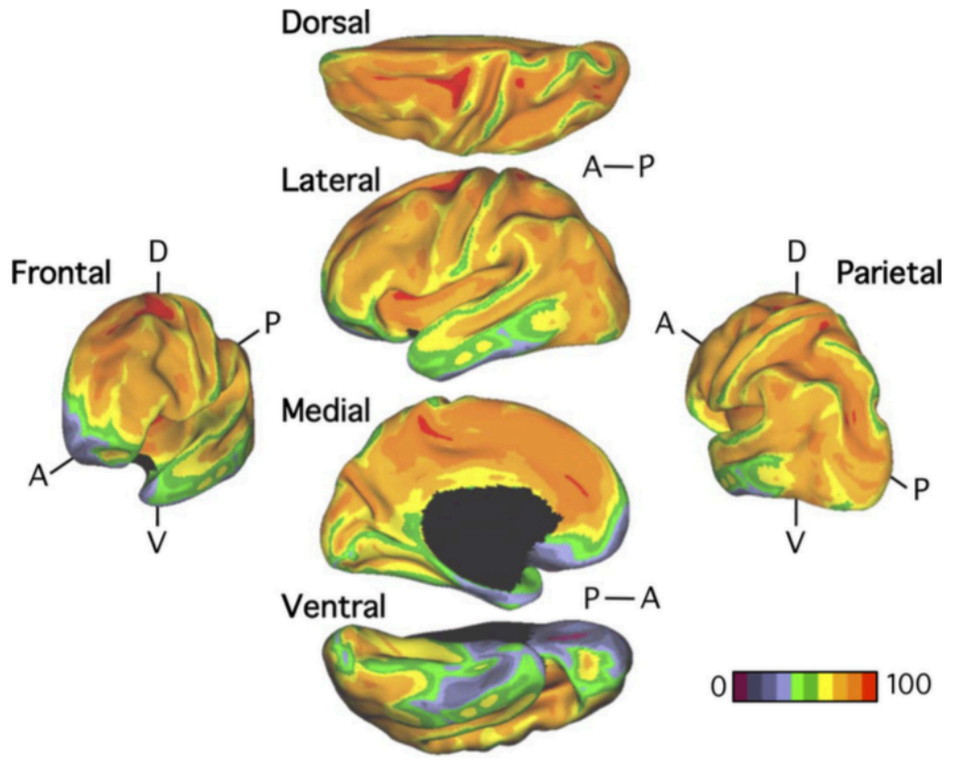
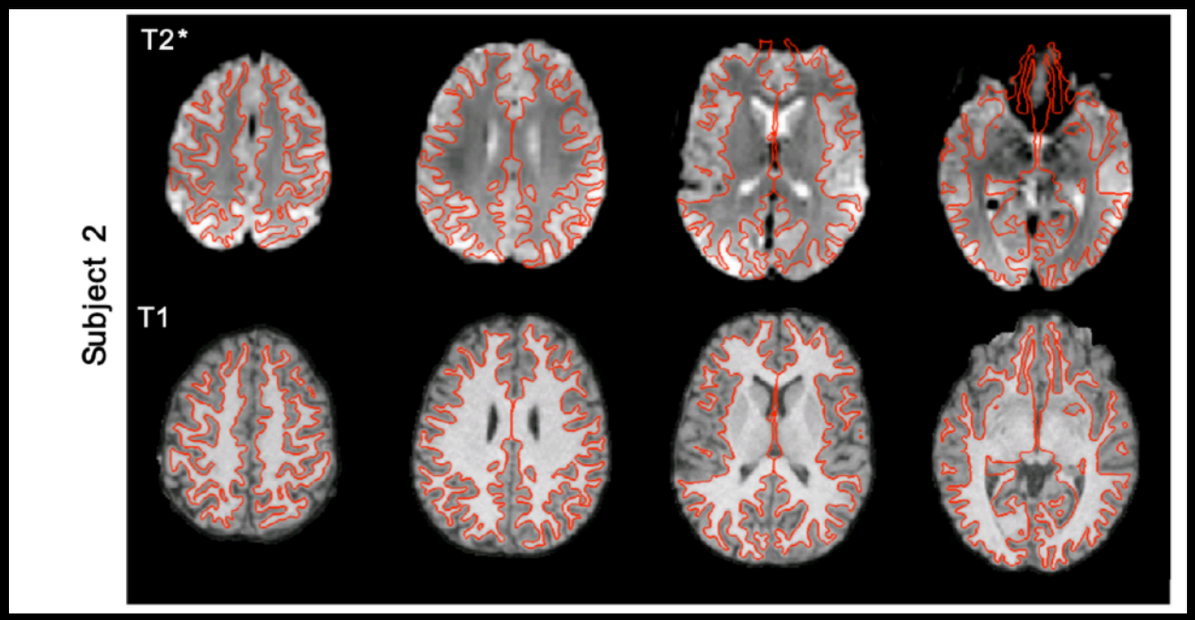
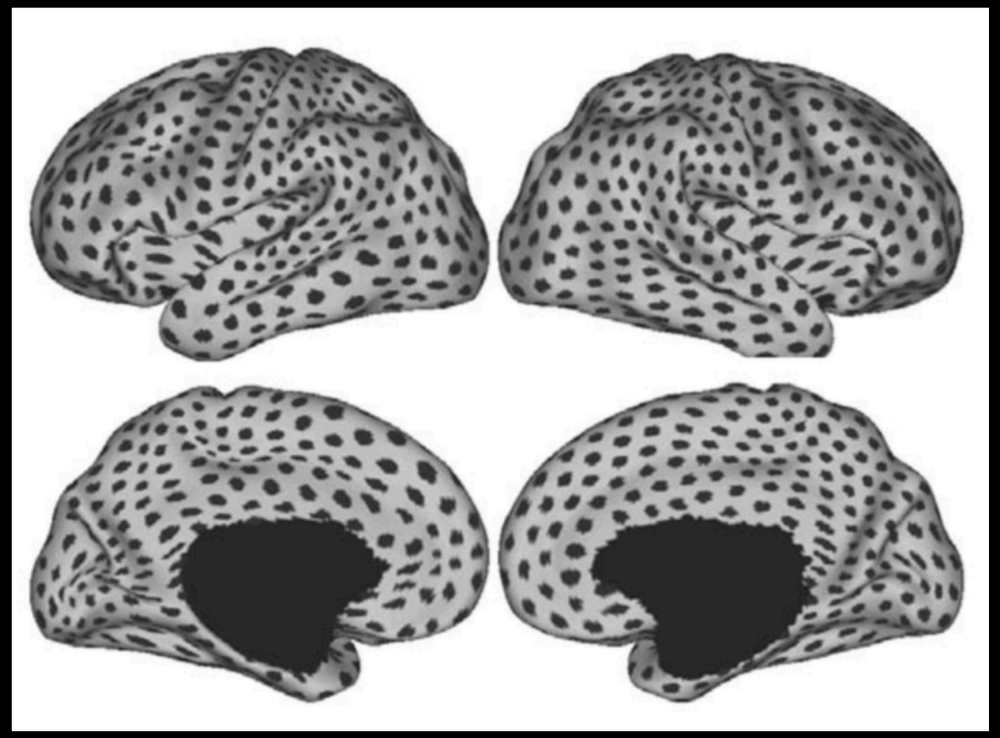
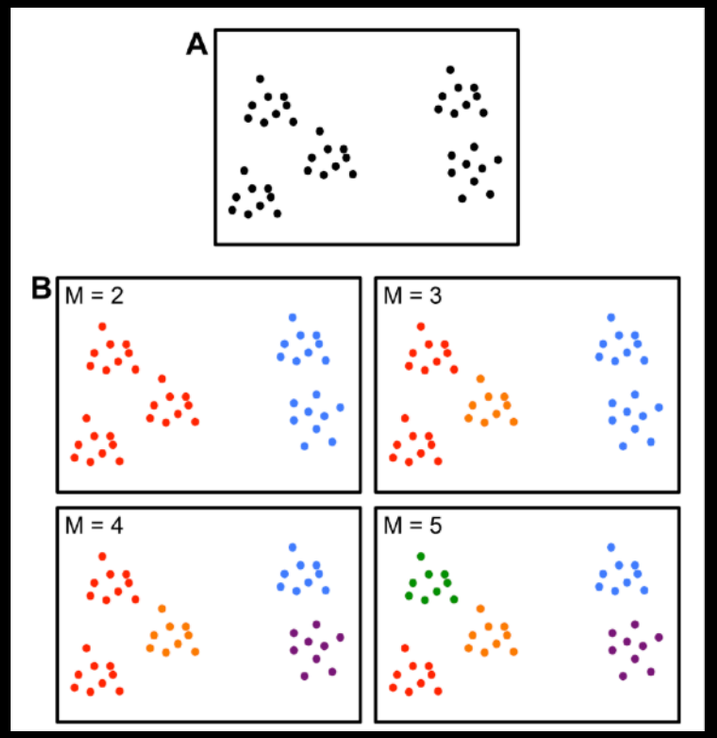
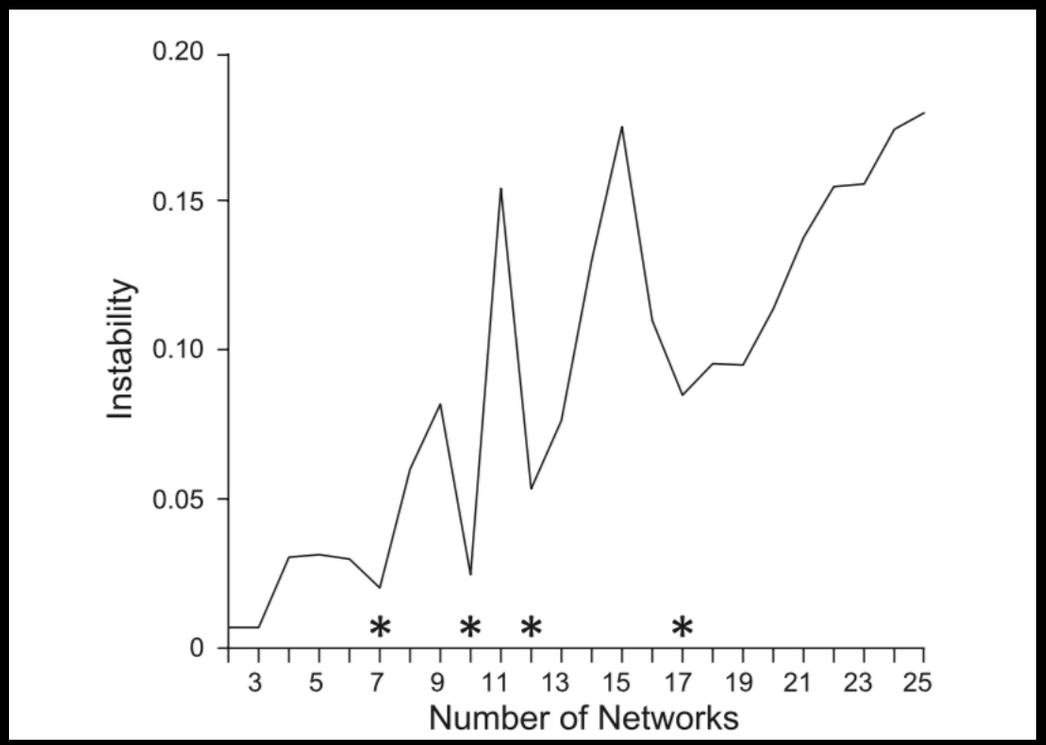


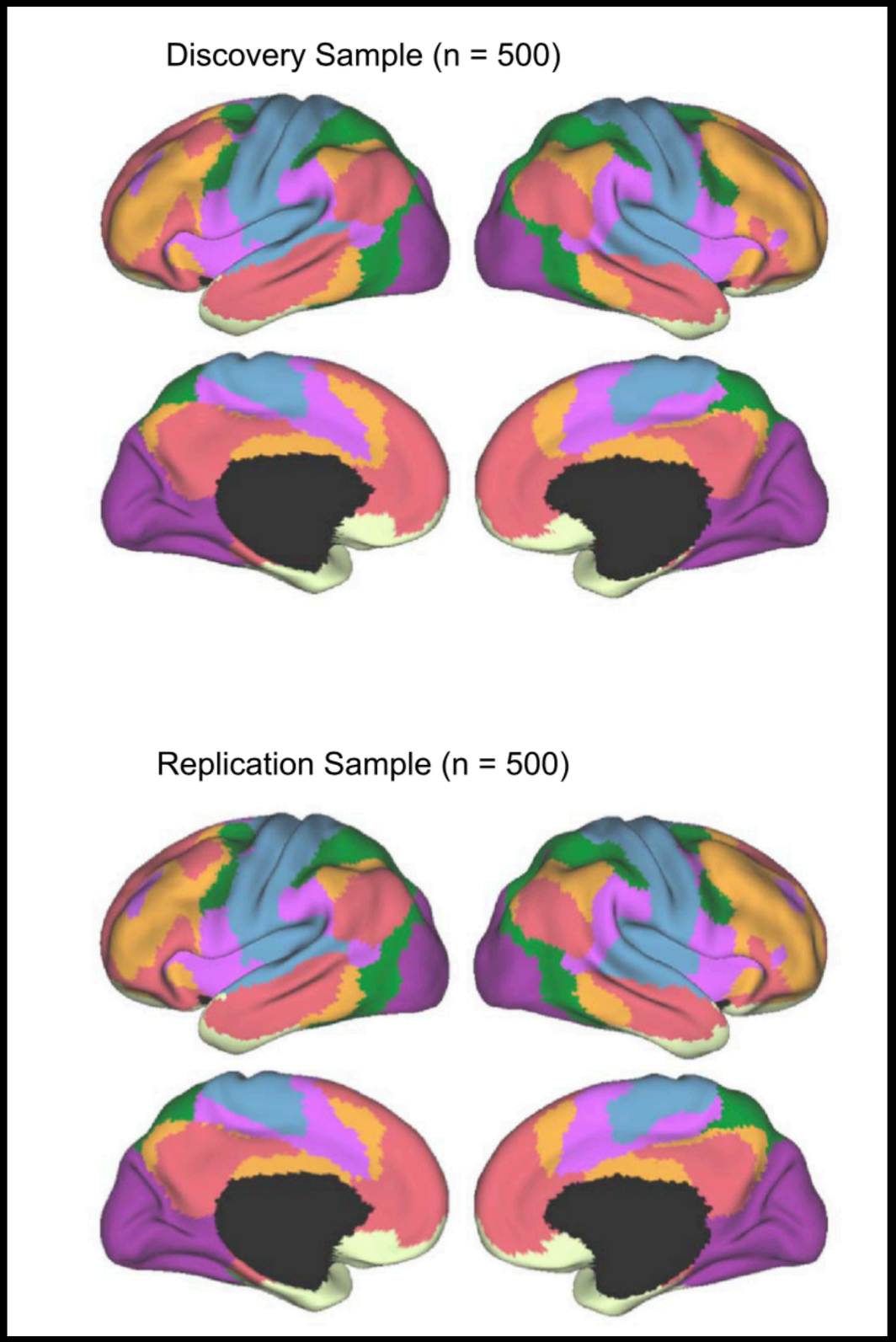
Fig. 3. Signal-to-noise ratio (SNR) maps of the functional data from the full sample (N = 1,000). The mean estimate of the BOLD fMRI data SNR is illustrated for multiple views of the left hemisphere in Caret PALS space. A, anterior; P, posterior; D, dorsal; V, ventral.






A
B
M = 2
M = 3
M = 4
M = 5


0.20
0.15
0.10
0.05
0
Instability
3
5
7
9
11
13
15
17
19
21
23
25
Number of Networks



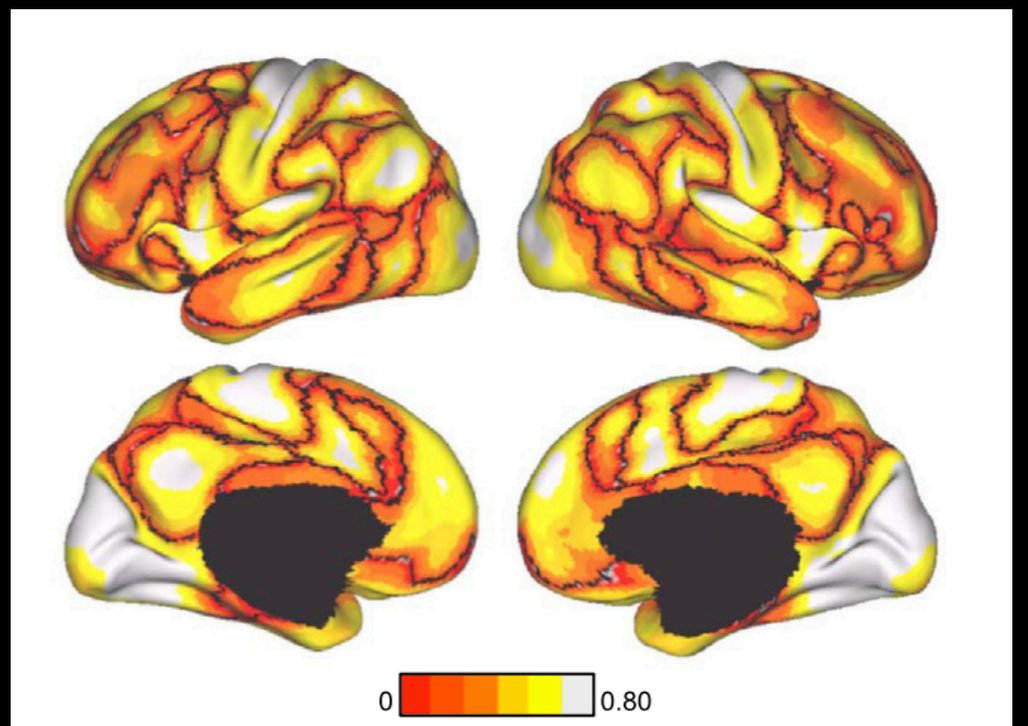
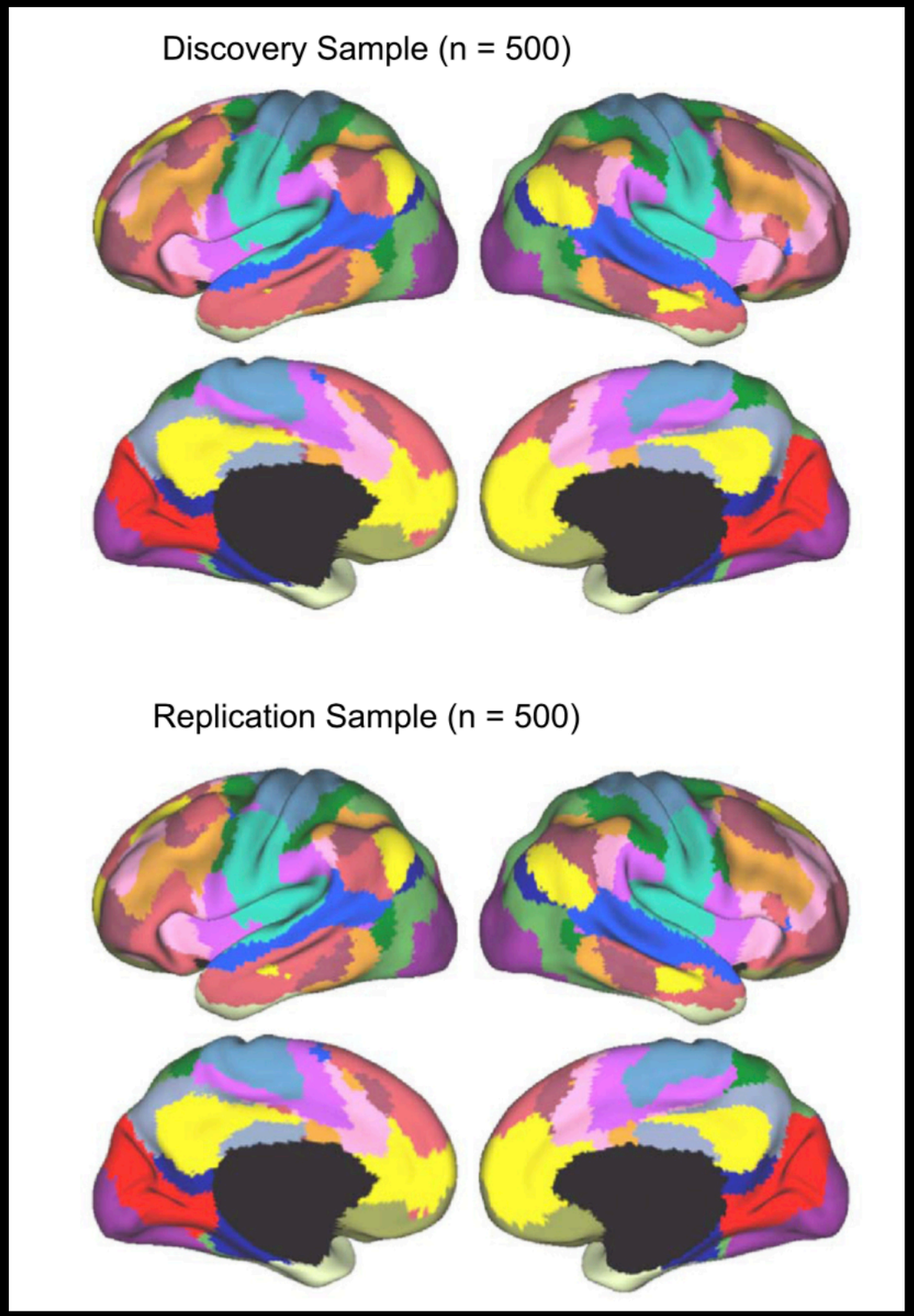


Fig. 8. Confidence of the 7-network estimate in the discovery data set. Confidence (silhouette) value for each vertex with respect to its assigned network is shown for the discovery data set. Regions close to the boundaries between networks were less confident of their assignment, although we also observed structured spatial variation within individual components of the estimated networks, such as lateral prefrontal cortex, which foreshadows its division in the 17-network estimate (see Fig. 9).



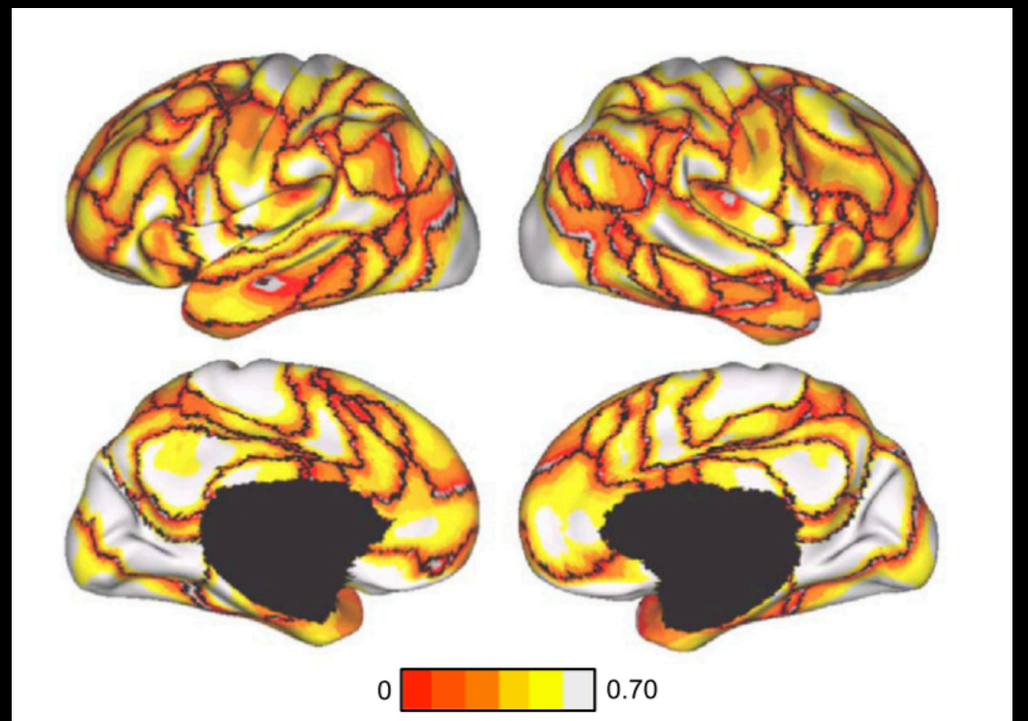
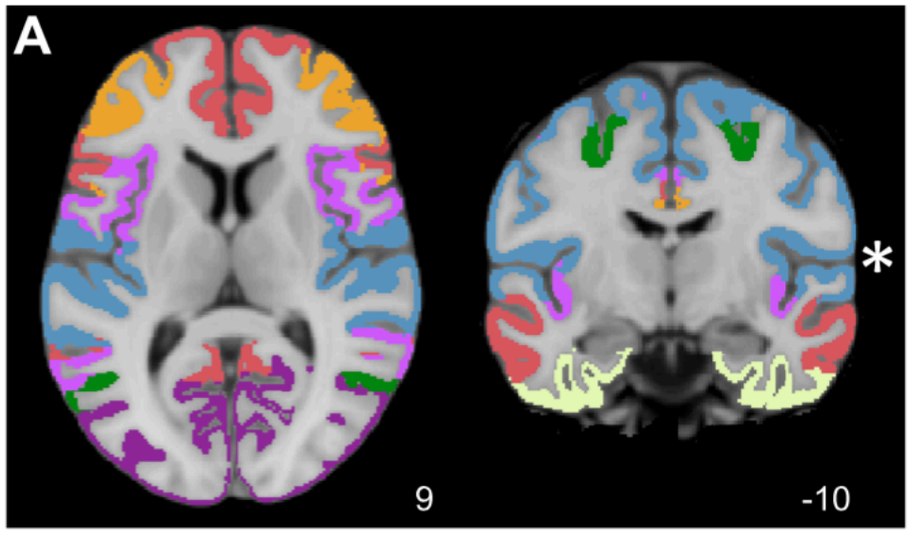
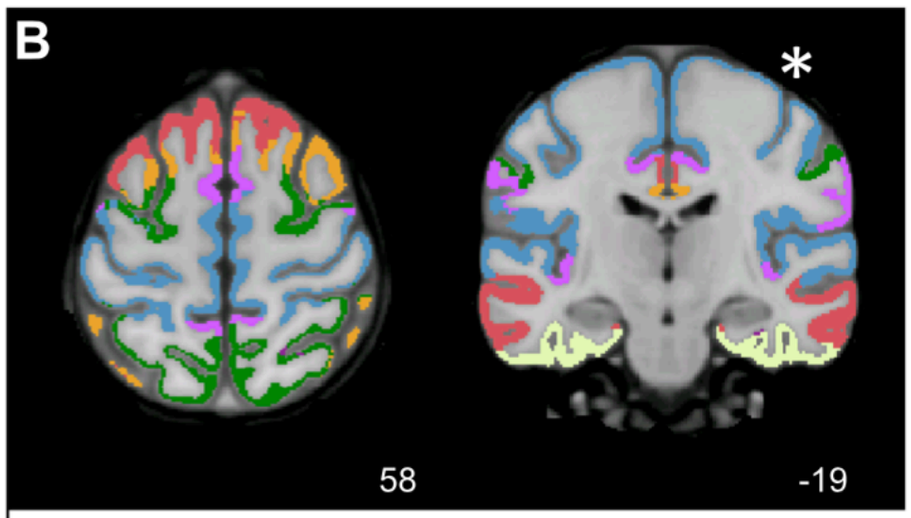
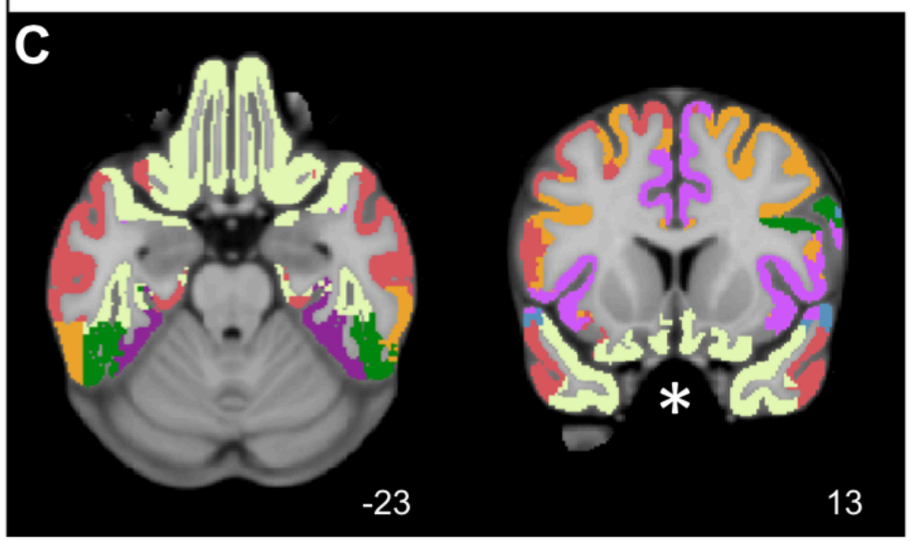


Fig. 10. Confidence of 17-network estimate in the discovery data set. Confidence (silhouette) value for each vertex with respect to its assigned network is shown for the discovery data set. Again, regions close to the boundaries between networks were less confident of their assignment.


A
9
-10


B
58
-19


C
-23
13

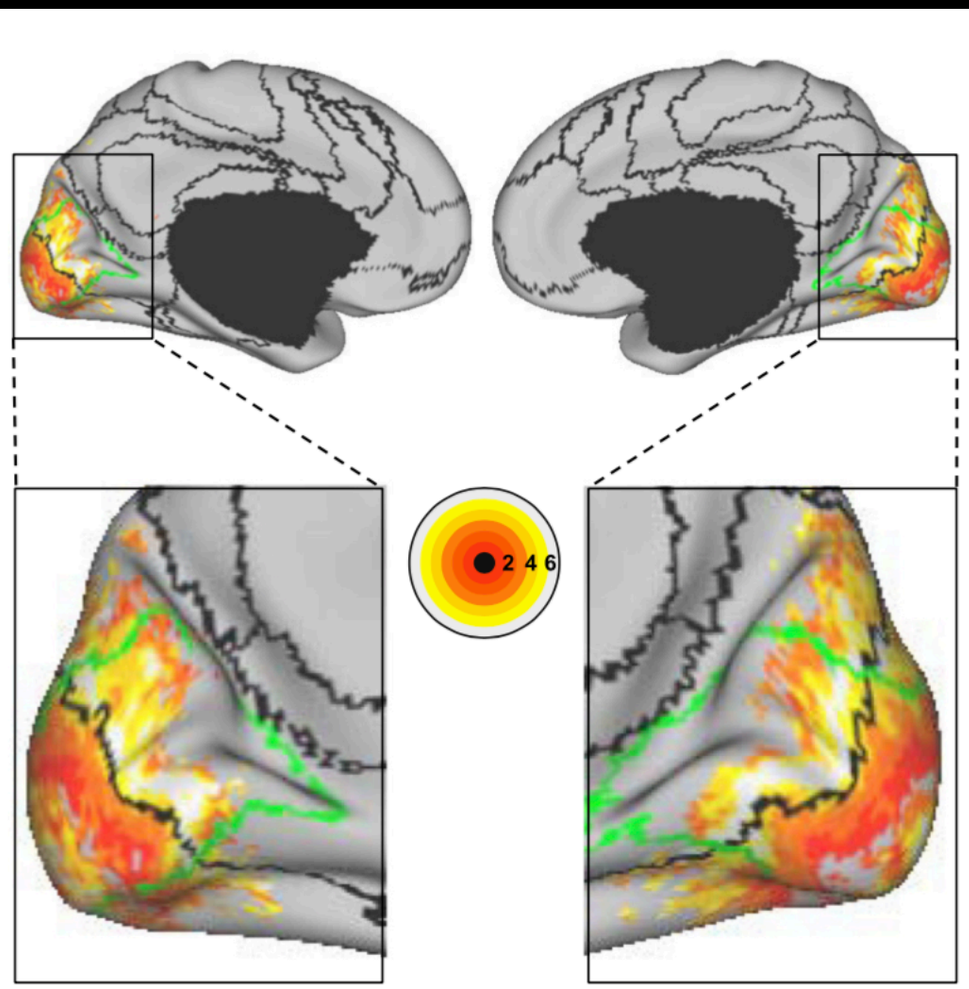


Fig. 15. Eccentricity estimates quantify the division of the early visual cortex into central and peripheral systems. Eccentricity estimates in the early visual areas of 4 subjects were averaged and overlaid on the boundaries (in black) of the 17-network estimate. The boundary between areas 18 and 19 estimated from the histological data set is overlaid in green. The 17-network estimate divides the early visual areas along an isoeccentricity line of ~4°. Note that the eccentricity estimates are not reliable outside the V1-V3 complex.

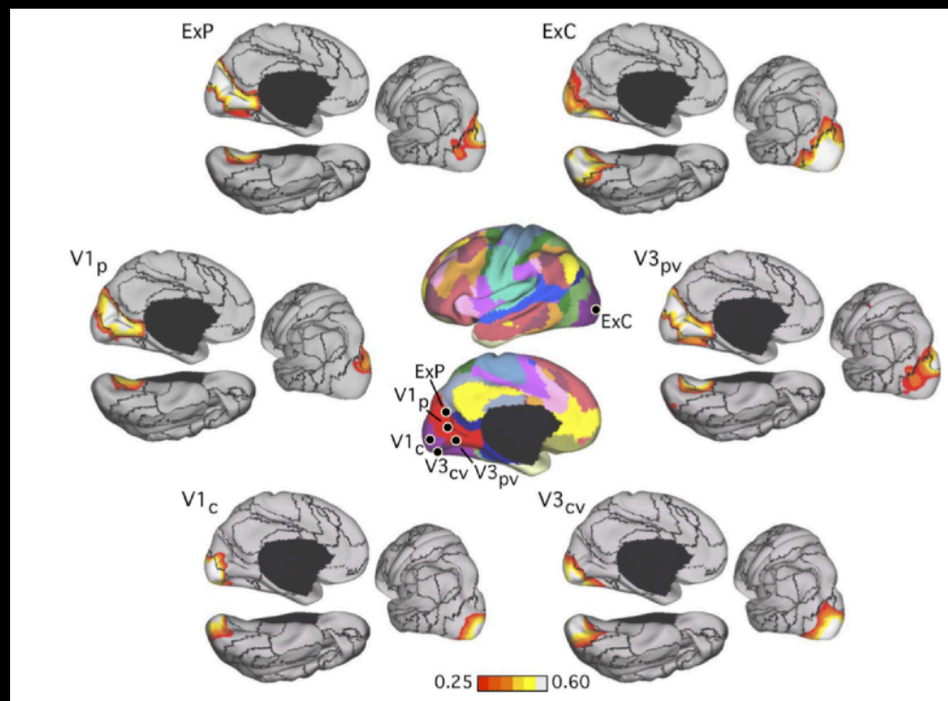
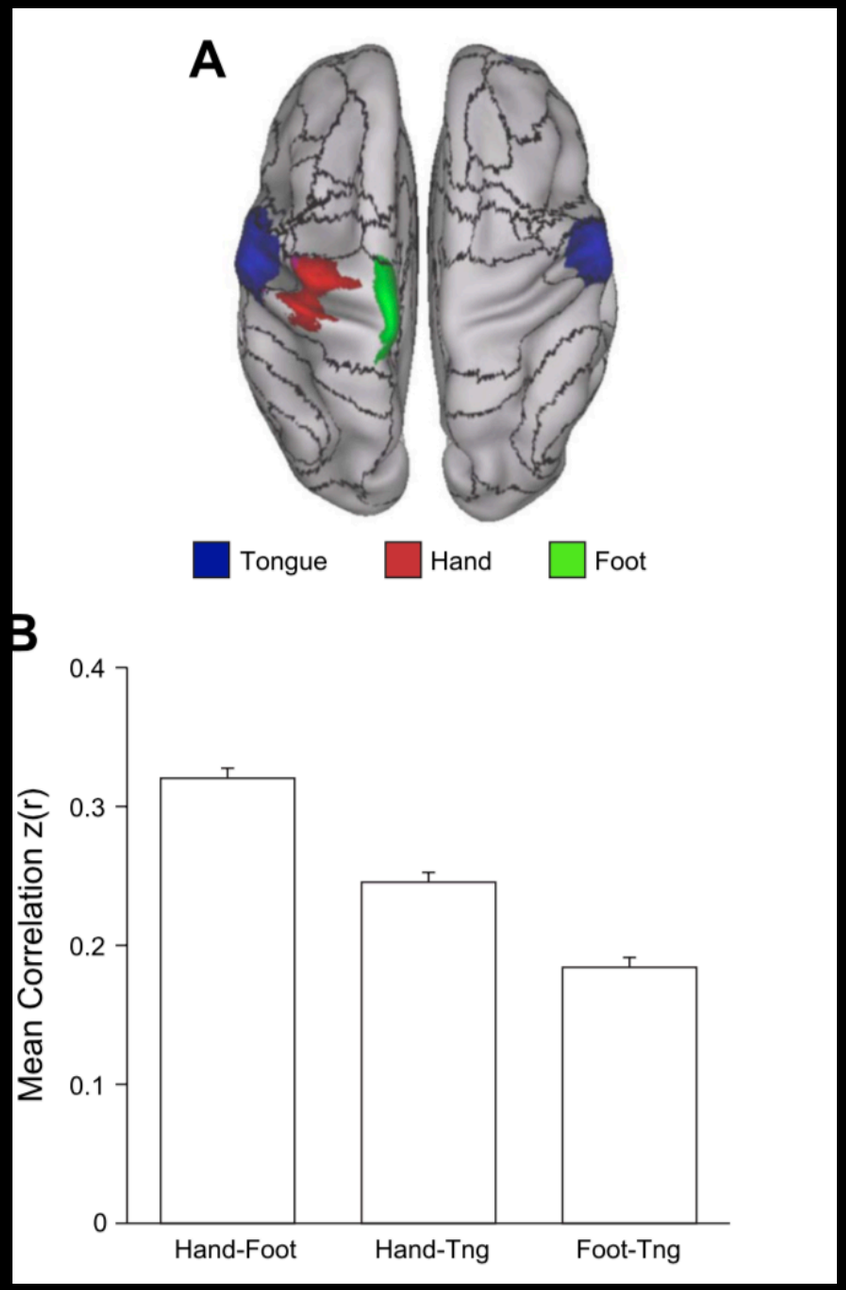


Fig. 16. Evidence that the fractionation of the visual system reflects functional connectivity MRI (fcMRI) topography within the visual cortex. Six left hemisphere seed regions were picked from the discovery dataset: $V1_c$ and $V1_p$ correspond to central and peripheral visual field representation within V1, respectively; $V3_{cv}$ and $V3_{pv}$ correspond to central and peripheral V3v, respectively; ExC and ExP correspond to 2 seed regions within the extrastriate visual cortex in the estimated locations of the central and peripheral visual fields (purple and bright red at *center*). The 6 seed regions are illustrated at *center*, and their coordinate locations are reported in Table 1. Their left hemisphere fcMRI maps were computed using the replication data set and arranged around the *center* images. Note that the central visual seed regions are selectively correlated with the central visual representation, whereas the peripheral visual seed regions are selectively correlated with the peripheral visual representation.



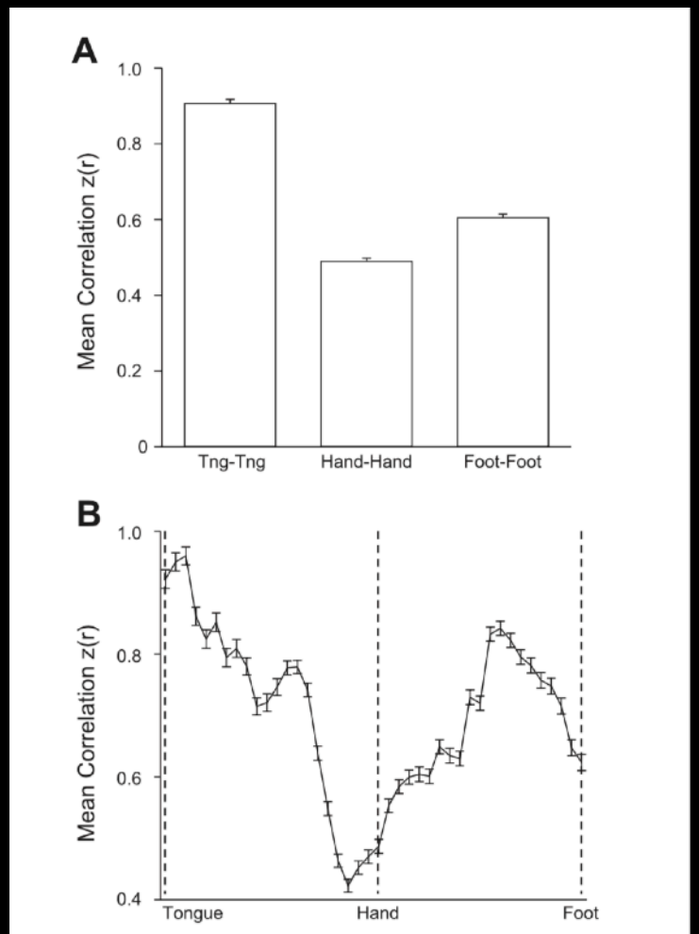


Fig. 21. Evidence that the interhemispheric fcMRI of homotopic regions within the primary motor cortex is topographically organized. *A*: correlation strength of left hemisphere tongue, hand, and foot seed regions with corresponding contralateral seed regions averaged over all 1,000 subjects. Right hemisphere vertices were obtained by reflection across the midline. Hand coordinates = ±41, −20, 62; foot coordinates = ±6, −26, 76; and tongue coordinates = ±55, −4, 26. The tongue representation has the strongest interhemispheric correlation, followed by the foot and then the hand. *B*: plot of interhemispheric correlation along the ventral (tongue) to dorsal (foot) extent of motor cortex. Maximal interhemispheric correlation is highest near the tongue representation and also peaks between the hand and foot representations, possibly corresponding to the trunk representation.

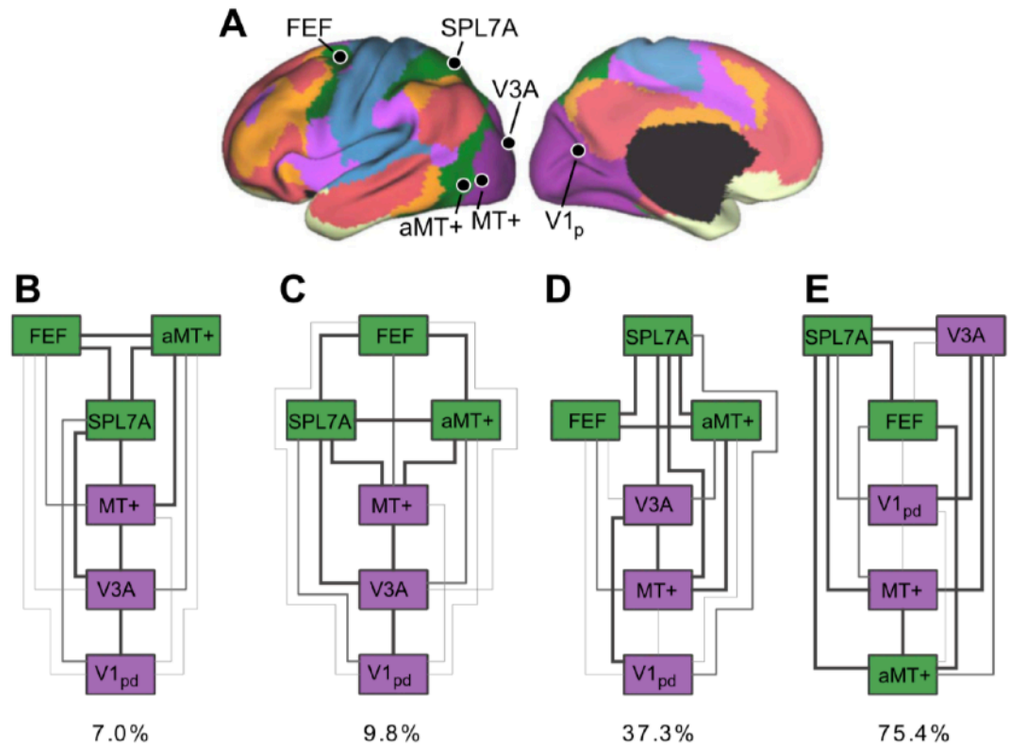


Fig. 29. Functional connectivity estimates of the hierarchical organization of a canonical sensory-motor pathway. *A*: 6 seed regions arranged into a 5-level functional hierarchy using the replication data set. *B* and *C*: 2 best hierarchical arrangements of the seed regions as measured by the proportion of violated hierarchical and lateral constraints. A violation occurred when the ordering placed more strongly correlated regions farther apart in the hierarchy than more weakly correlated regions (see Fig. 28). *D* and *E*: 2 poor hierarchical arrangements of the seed regions as measured by the proportion of violated hierarchical and lateral constraints. Relative ordering of the seed regions (*A* and *B*) within the functional hierarchy agrees well with the proposed macaque visual hierarchy (see text).

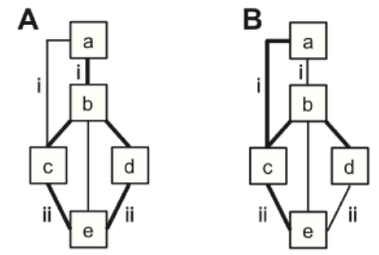
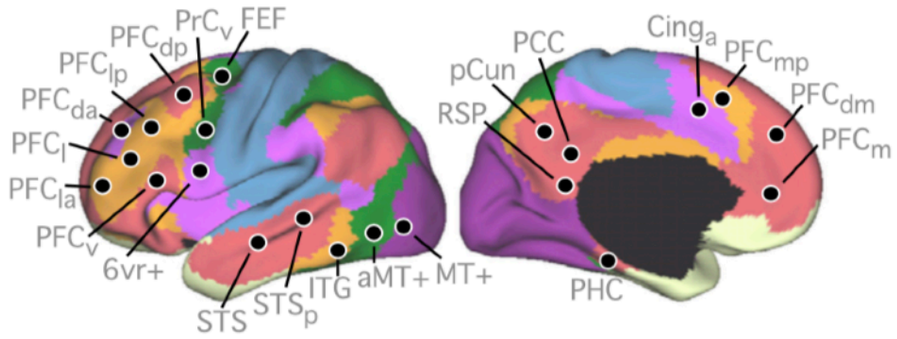
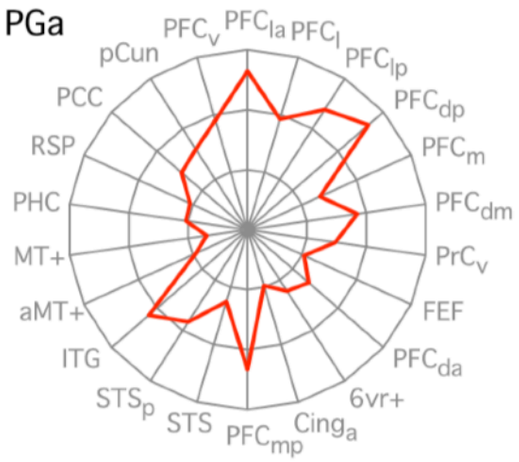


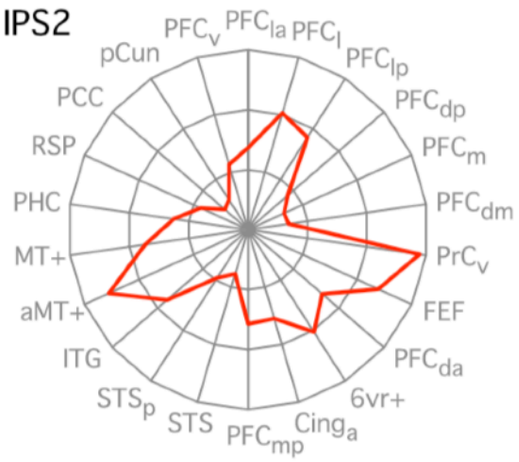
Fig. 28. Examples of satisfied and violated constraints in estimating the functional hierarchy of cortical regions based on fcMRI. A functional hierarchy is estimated based on the assumption that regions closer in a hierarchy have stronger correlation. *A*: 5 cortical regions are arranged in a 4-level hierarchy whose functional connectivity strengths satisfy both hierarchical and lateral constraints. *B*: identical arrangement of 5 cortical regions in a 4-level hierarchy with different functional connectivity strengths that violate both hierarchical and lateral constraints. Thick lines correspond to strong correlations. Thin lines correspond to weak correlations. *i*: *regions a* and *c* are farther apart than *regions a* and *b*. In the example in *A*, correlation of *regions a* and *c* is weaker than correlation of *regions a* and *b*, so a hierarchical constraint is satisfied. In the example in *B*, correlation of *regions a* and *c* is stronger than correlation of *regions a* and *b*, so a hierarchical constraint is violated. *ii*: *regions c* and *d* are on the same hierarchical level. In the example in *A*, correlation of *regions c* and *e* is approximately the same as the correlation of *regions d* and *e*, so a lateral constraint is satisfied. In the example in *B*, correlation of *regions c* and *e* is stronger than the correlation of *regions d* and *e*, so a lateral constraint is violated. In the context of hierarchy estimation in this article, we consider 2 correlations within 0.2 of each other to be approximately the same when assessing lateral constraints. Given the pairwise correlations of a set of seed regions and a known number of levels in the hierarchy, we can seek the best hierarchical arrangement of the seed regions with the following local optimization procedure: *1*) randomly arrange the seed regions into a hierarchy, *2*) consider swapping a pair of seed regions or shifting a single seed region to a different hierarchical level without changing the number of levels in the hierarchy such that the proportion of violated constraints is maximally decreased, *3*) terminate when no further improvement in the proportion of violated constraints can be achieved, and *4*) repeat the preceding steps 20 times, picking the solution with the least proportion of violated constraints. The best solution obtained using this optimization procedure is (in practice) the same as a brute-force search over all possible hierarchical arrangements of the seed regions. We note that we are generally unable to infer the number of levels in the functional hierarchy, since the number of possible constraints can be drastically different for hierarchies with a different number of levels, and so the proportion of violated or satisfied constraints is not comparable across hierarchies with different levels. In practice, however, the solution space is robust; for example, the best solution for a 5-level hierarchy typically differs from the best solution for a 4-level hierarchy by the collapsing of regions from 2 adjacent levels into 1 level. Uncovering the true hierarchical structure in the macaque visual hierarchy based on anatomical connectivity has also proved to be problematic (Hilgetag et al. 1996).



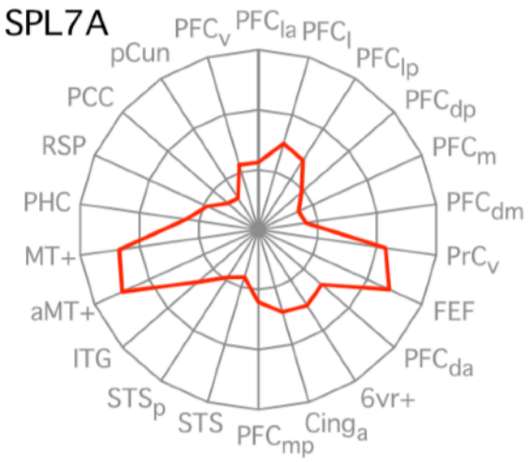
PGa



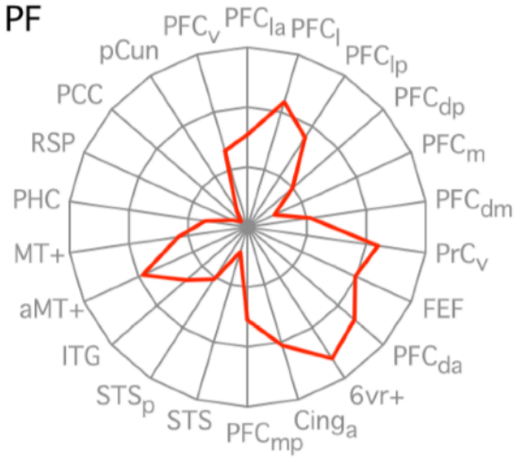
IPS2

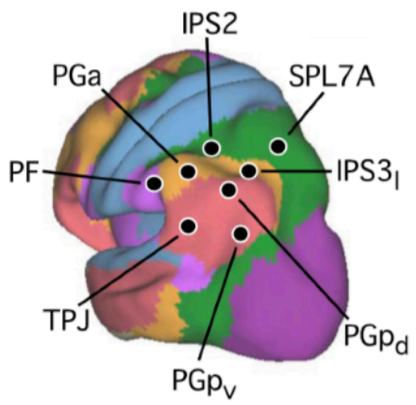


SPL7A

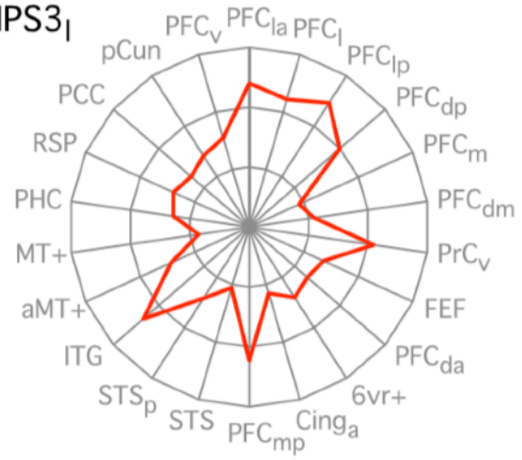


PF

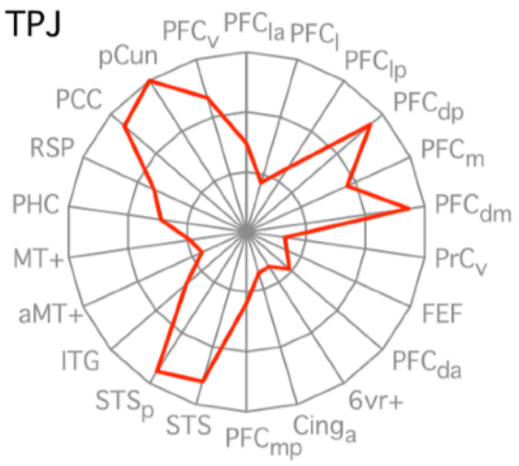




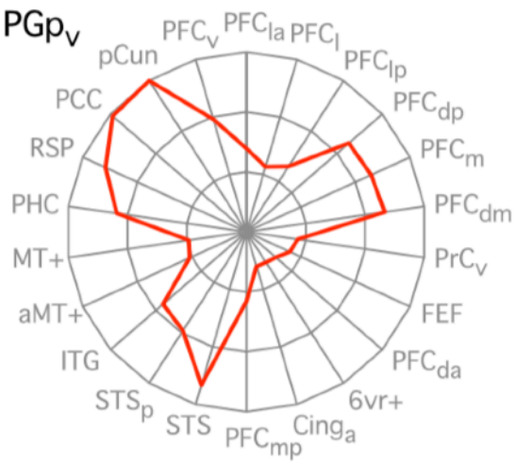
IPS3$_l$



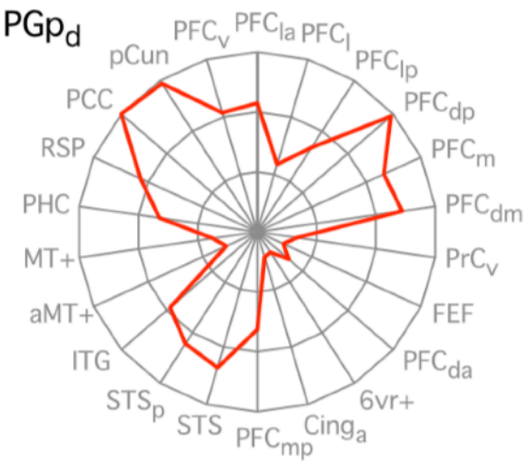
TPJ

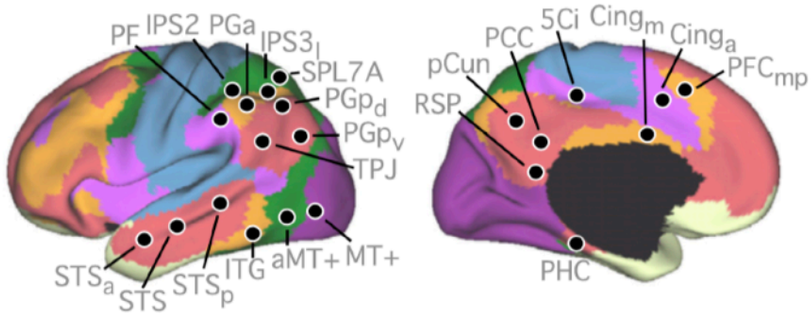


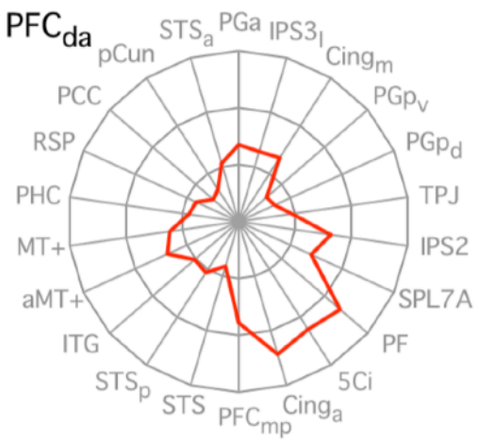
PGp$_v$

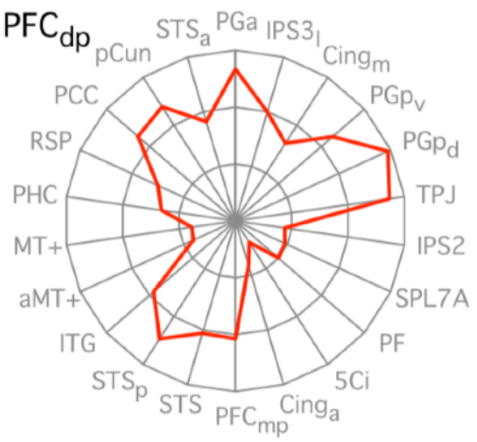


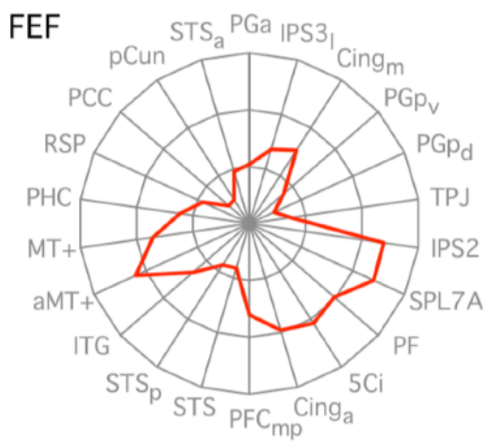
PGp$_d$

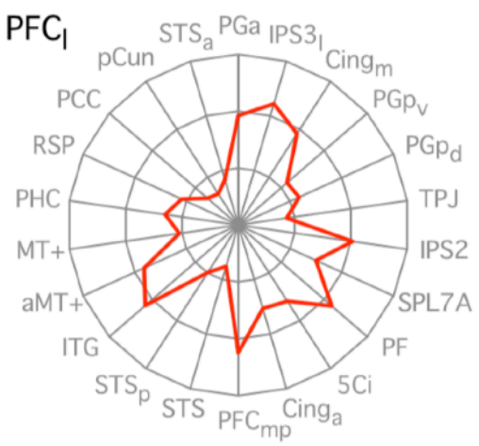


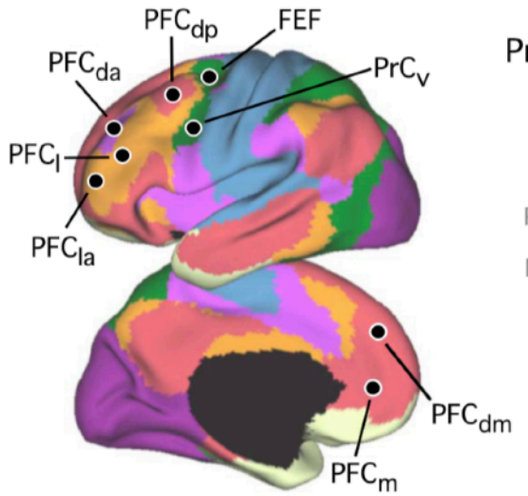


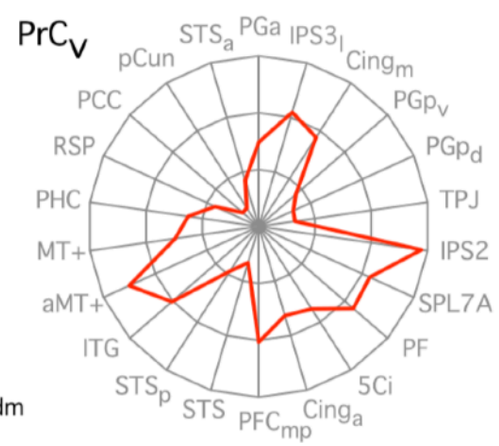


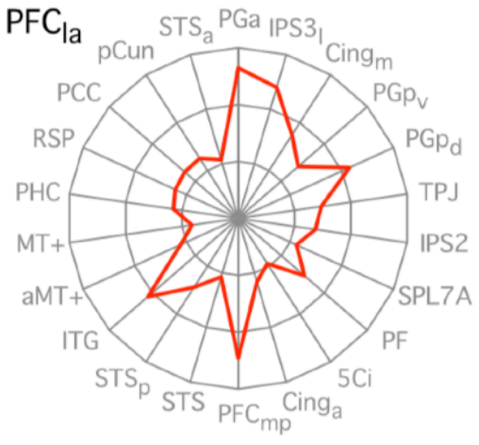


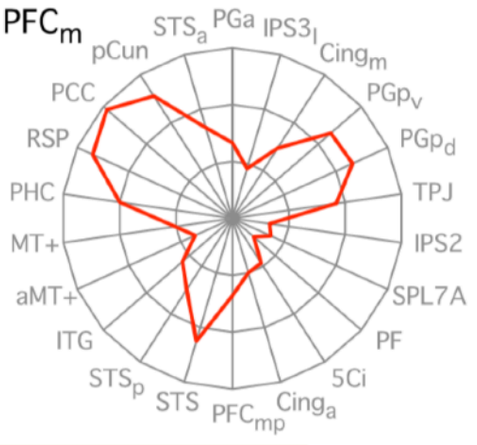












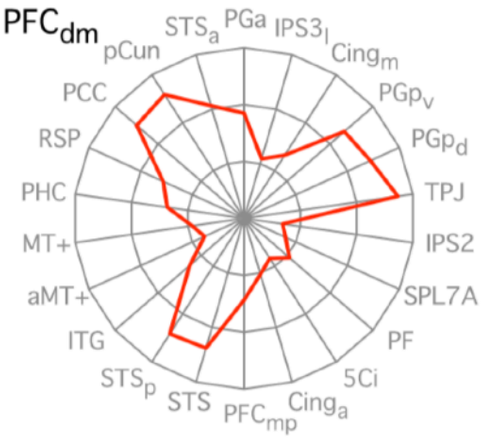


Fig. 31. Adjacent frontal regions exhibit distinct functional connectivity fingerprints. The format and plotting are the same as for Fig. 30 with regions tailored for exploration of frontal cortex. The coordinate locations are reported in Table 4. The polar scales range from $r = -0.4$ (center) to $r = 0.5$ (outer boundary) in 0.3-step increments.