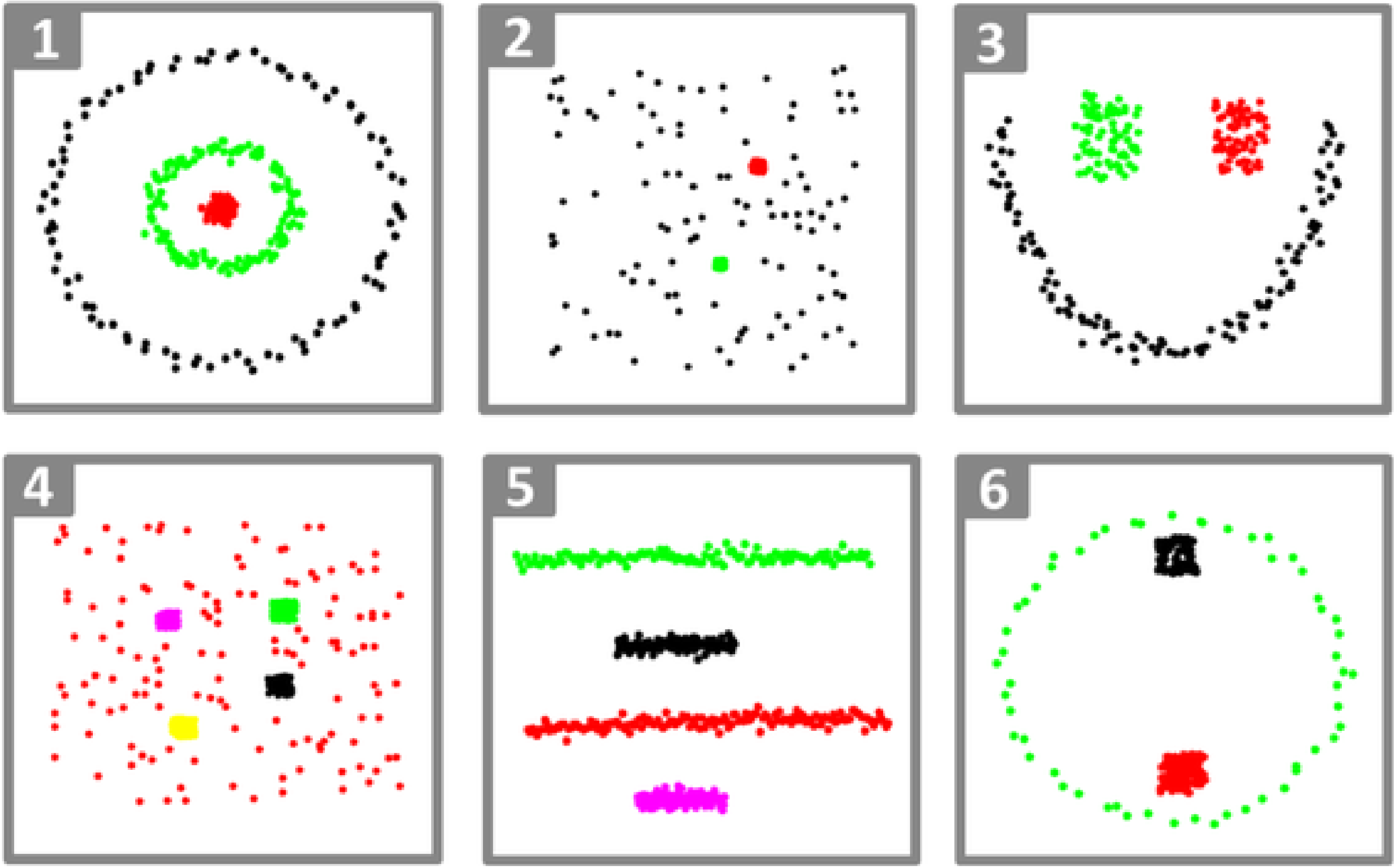
K-means clustering – selection and the High-dimensional-low-sample problem

K-MEANS

CLUSTERING

1. k centerpoints are randomly initialized.
2. Observations are assigned to the closest centerpoint.
3. Centerpoints are moved to the center of their members.
4. Repeat steps 2 and 3 until no observation changes membership in step 2.

Chris Albon


1
2
3
4
5
6

K-means in one of the most widely used unsupervised classification algorithms. It is used to partition observation of data with *N-samples* and *P-features*

k-means clustering, or Lloyd's algorithm [2], is an iterative, data-partitioning algorithm that assigns n observations to exactly one of k clusters defined by centroids, where k is chosen before the algorithm starts.

The algorithm proceeds as follows:

1. Choose k initial cluster centers (centroid). For example, choose k observations at random (by using 'Start','sample') or use the k-means ++ algorithm for cluster center initialization (the default).

2. Compute point-to-cluster-centroid distances of all observations to each centroid.

3. There are two ways to proceed

   Batch update — Assign each observation to the cluster with the closest centroid.

   Online update — Individually assign observations to a different centroid if the reassignment decreases the sum of the within-cluster, sum-of-squares point-to-cluster-centroid distances.

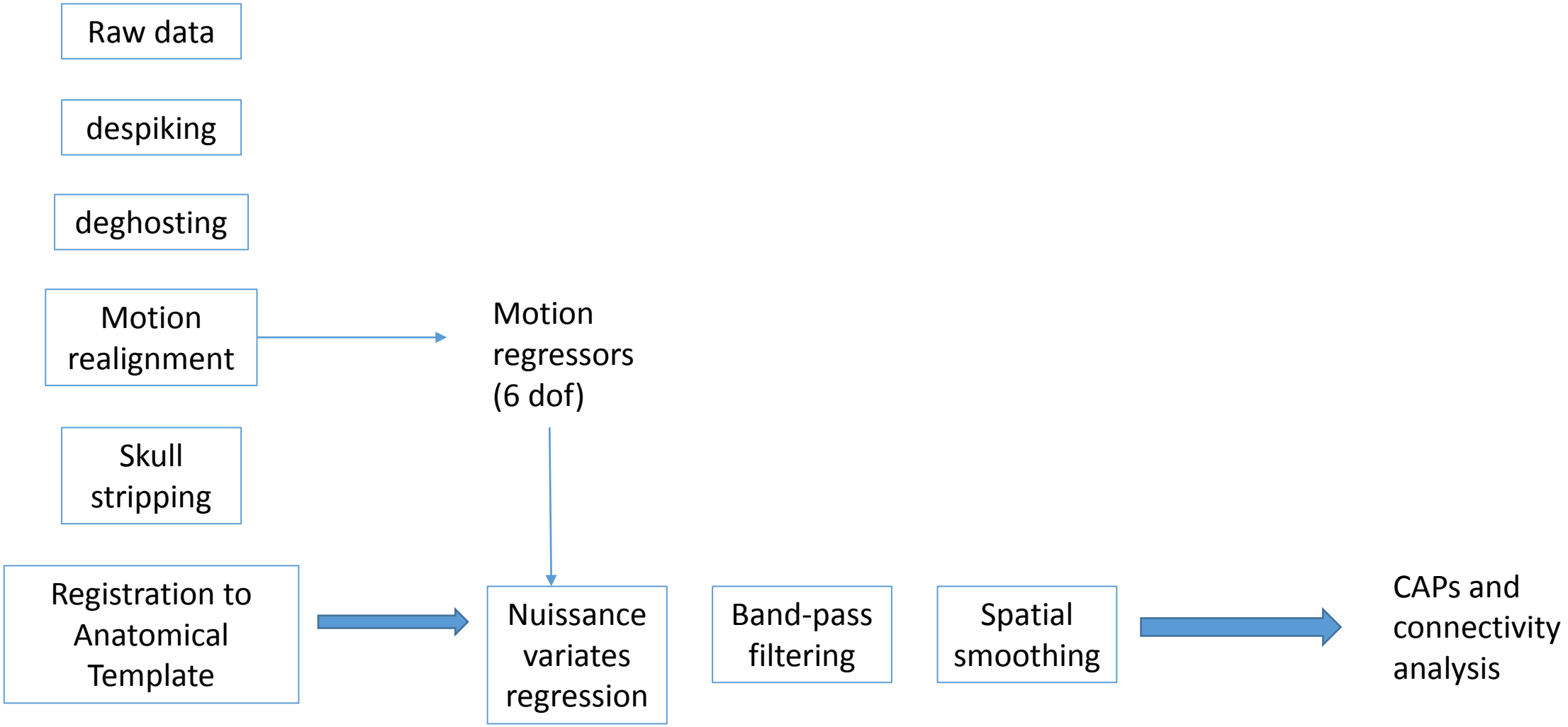
4. Compute the average of the observations in each cluster to obtain k new centroid locations.

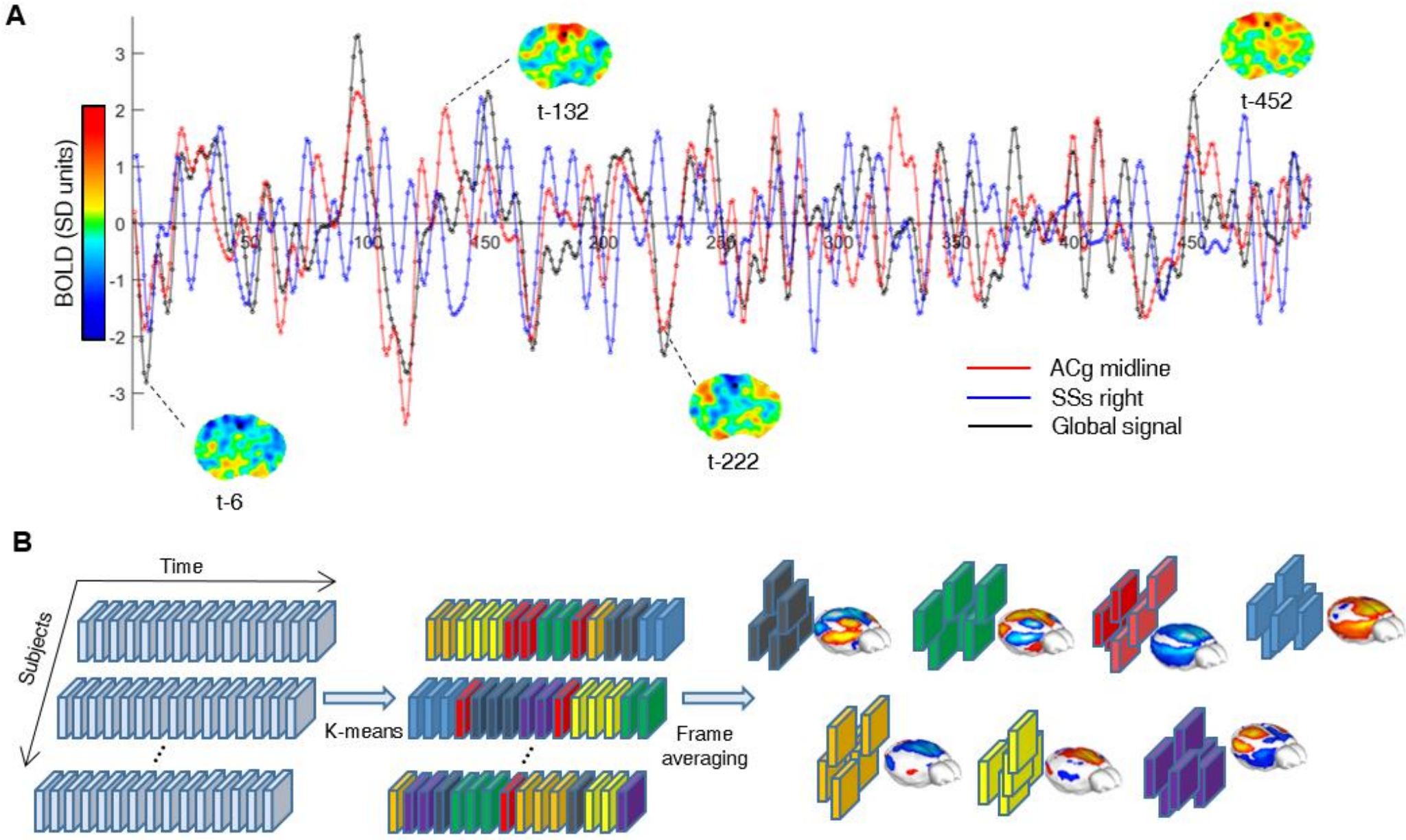
5. Repeat steps 2 through 4 until cluster assignments do not change, or the maximum number of iterations is reached.

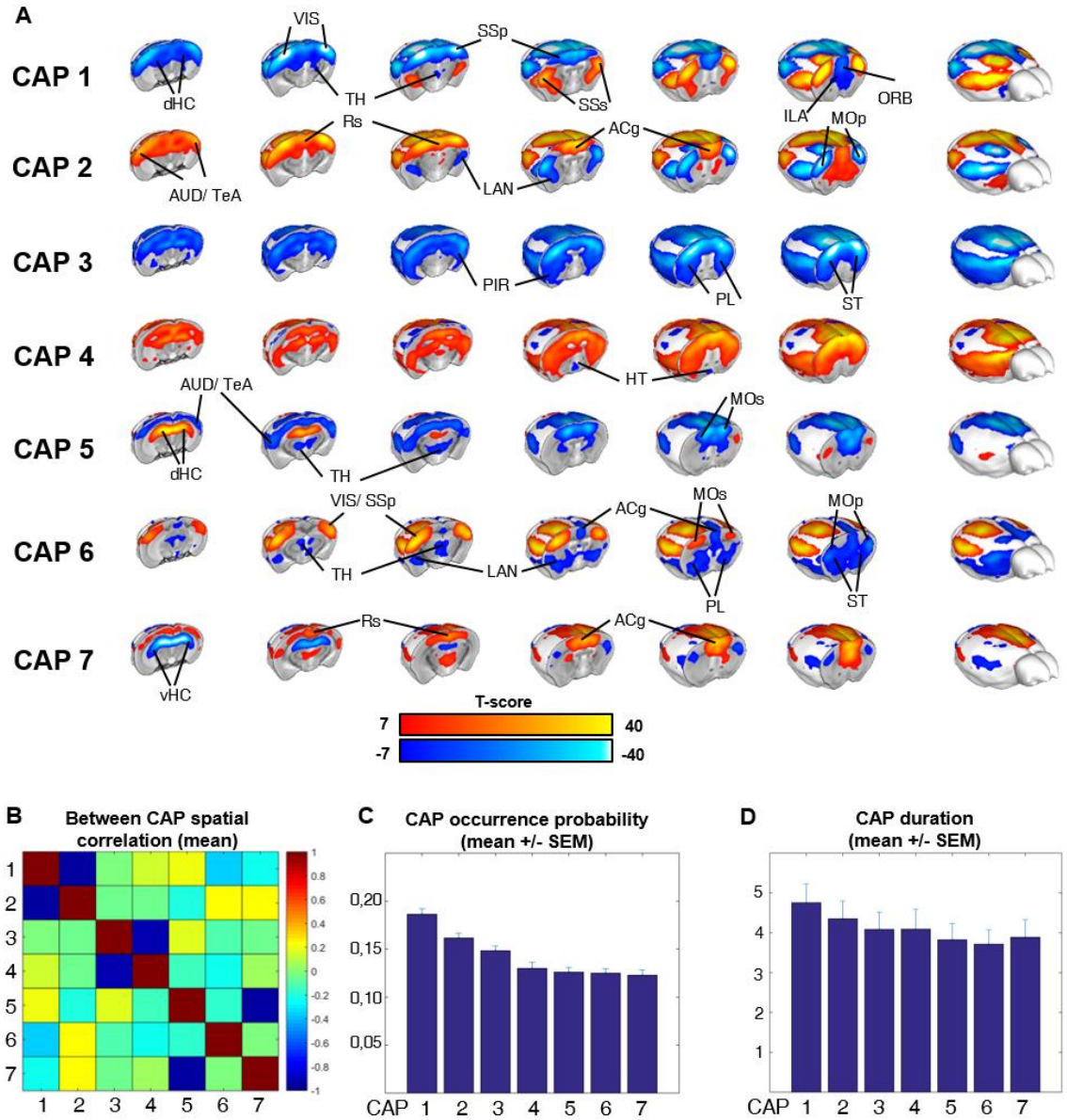
fMRI Data

- fMRI data is heavily preprocessed before any type of analysis as raw data is contaminated with motion of the subject; magnetic susceptibility artifacts; low frequency machine drifts; high frequency cardiac and respiratory nuisance; and between brain size variability.

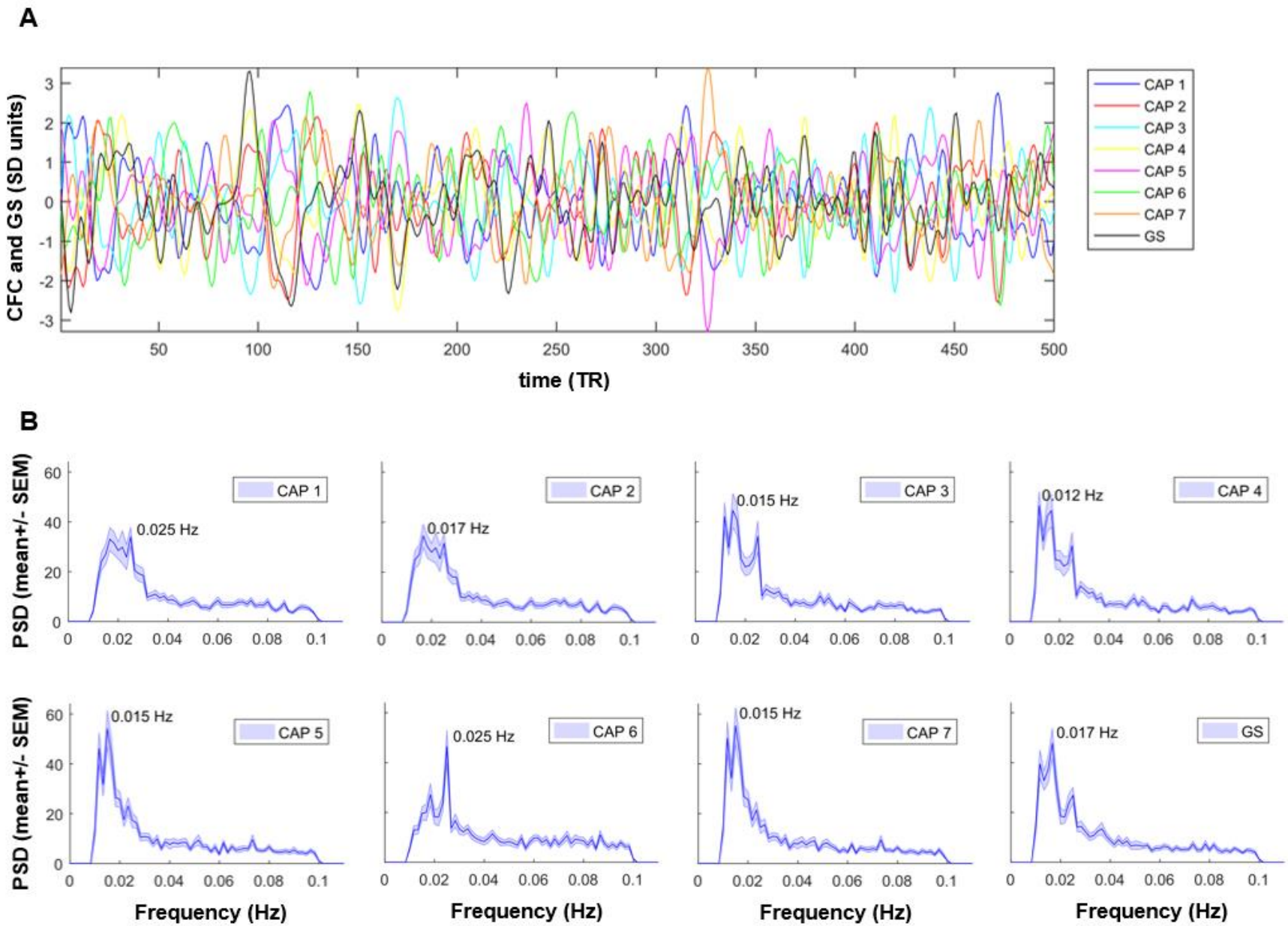
Preprocessing




A
BOLD (SD units)
3
2
1
0
-1
-2
-3
50
100
150
200
250
300
350
400
450
t-132
t-452
t-6
t-222
ACg midline
SSs right
Global signal
B
Time
Subjects
K-means
Frame
averaging



Abbreviations: ACg – Anterior Cingulate; ILA – Infralimbic Area; SSp – Primary somatosensory cortex; SSs – Secondary somatosensory cortex; VIS – Visual Cortex; Rs – Retrosplenial cortex; AUD – Auditory cortex; TeA – Temporal Association cortex; TH – Thalamus; dHC – Hippocampus dorsal; vHC – Hippocampus ventral; MOp – Primary motor cortex; LAN – Lateral Amygdalar nucleus; PL – Pallidum; ST – Striatum; HT – Hypothalamus.


A
CFC and GS (SD units)
time (TR)
CAP 1
CAP 2
CAP 3
CAP 4
CAP 5
CAP 6
CAP 7
GS
B
PSD (mean+/- SEM)
0.025 Hz
0.017 Hz
0.015 Hz
0.012 Hz
0.015 Hz
0.025 Hz
0.015 Hz
0.017 Hz
Frequency (Hz)