ST578: Statistical Learning in Data Science
Chapter 3: Medical Imaging Analysis

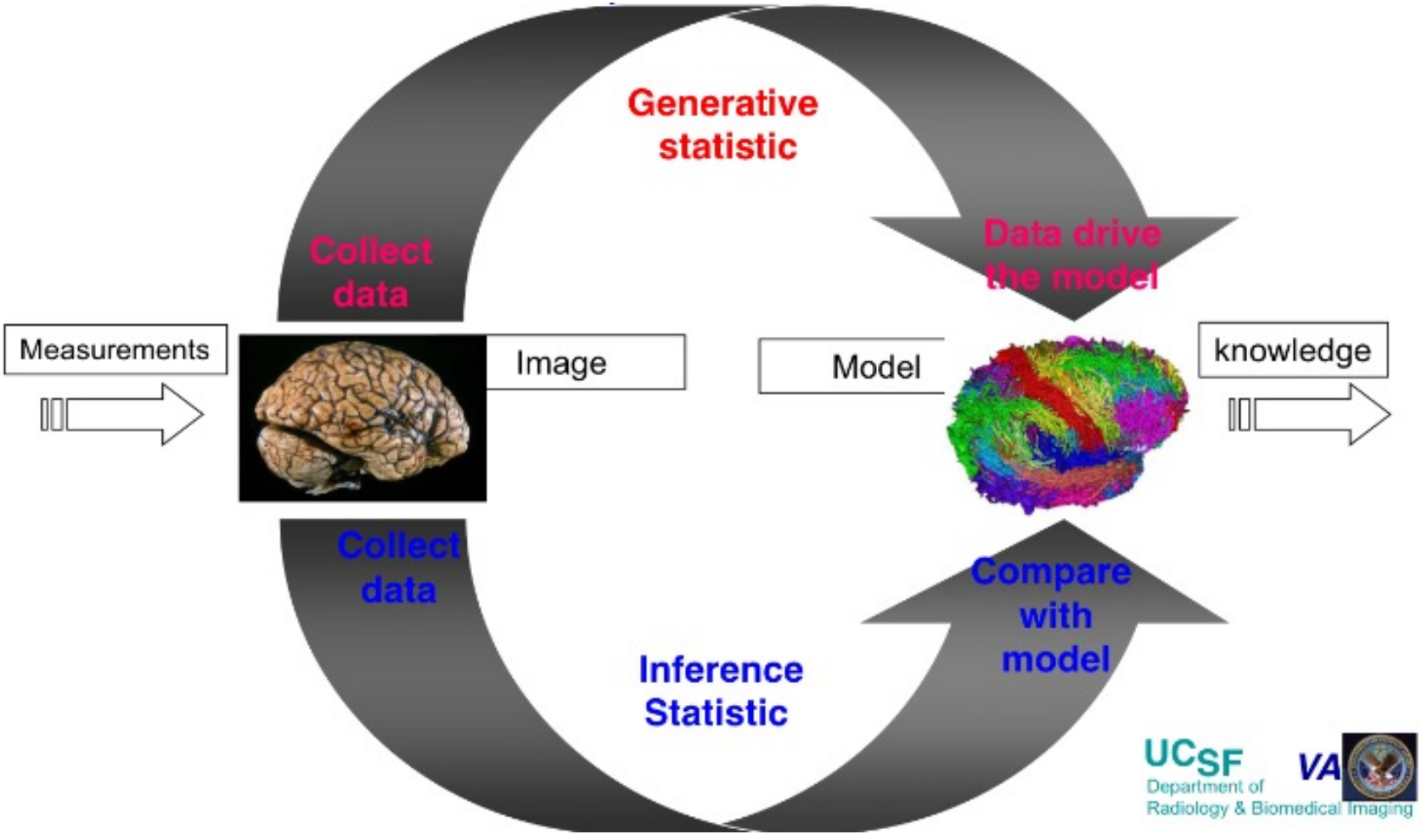
Annie Qu

University of Illinois at Urbana-Champaign
anniequ@illinois.edu

Spring, 2018

Knowledge Discovery Through Imaging Analysis

Figure source: Norbert Schuff, UCSF



Type of Medical Imaging: X-Rays

Figure: X-Rays of emphysema

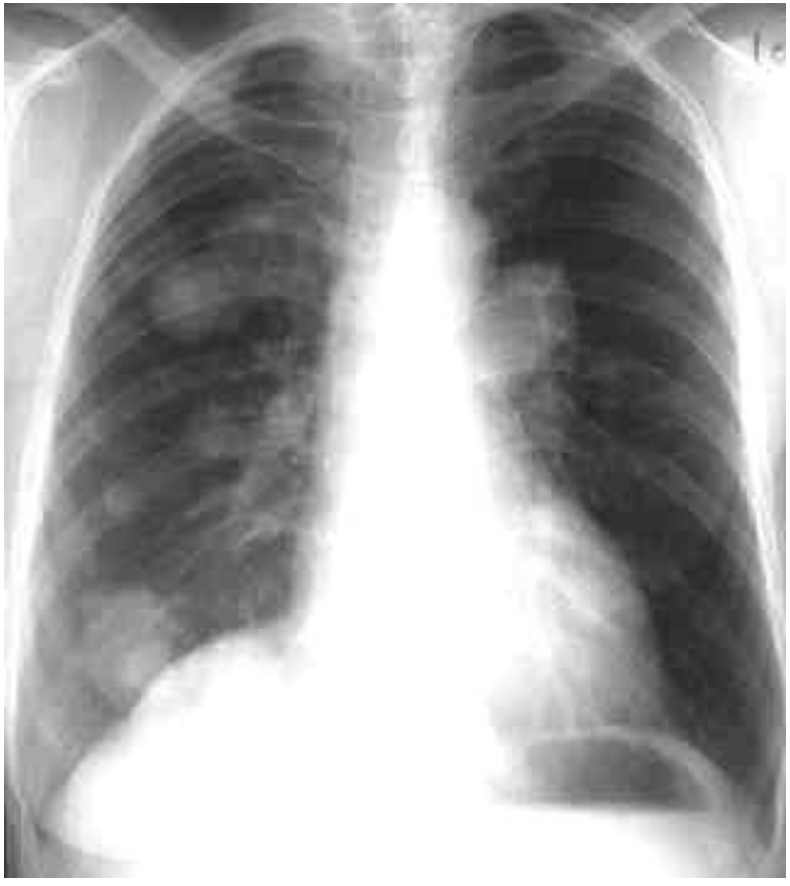


Figure: X-Rays of lung cancer

Emphysema is a long-term, progressive disease of the lungs that primarily causes shortness of breath due to over-inflation of the alveoli.

Computed Tomography (CT)

Imaging procedure uses special X-ray equipment to create detailed pictures, or scans, of areas inside the body.

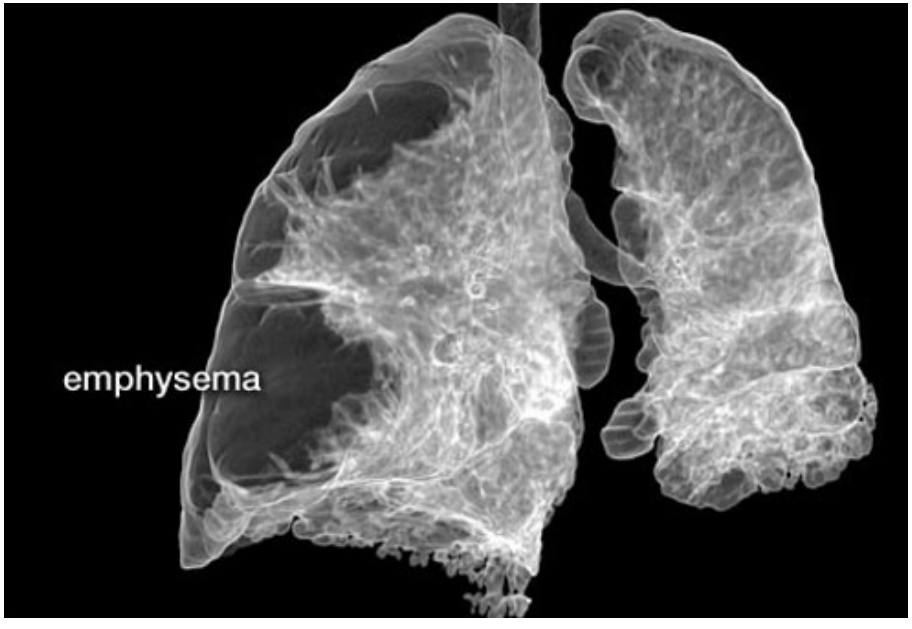


Figure: CT of emphysema

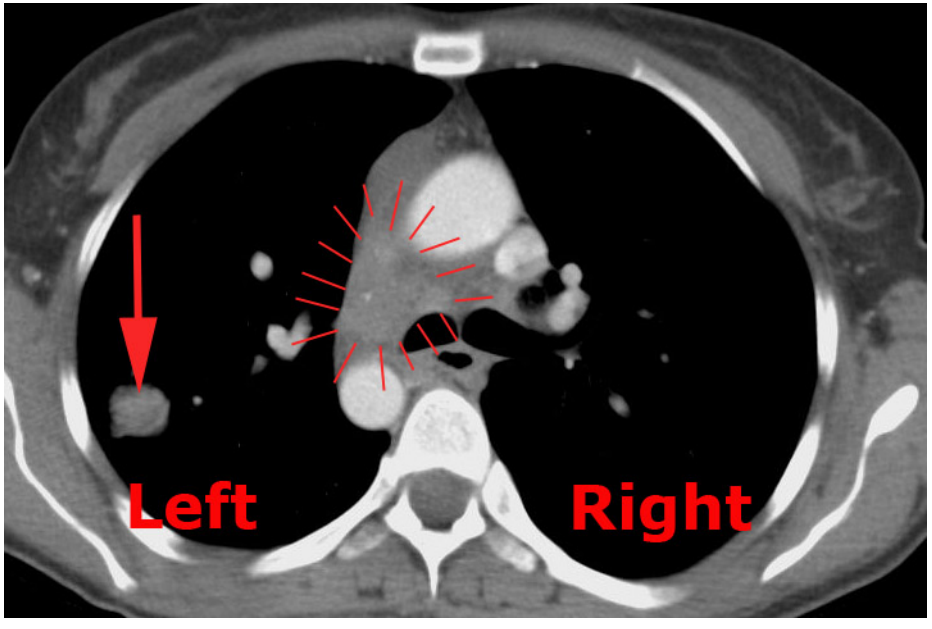


Figure: CT of lung cancer

Brain Anatomy

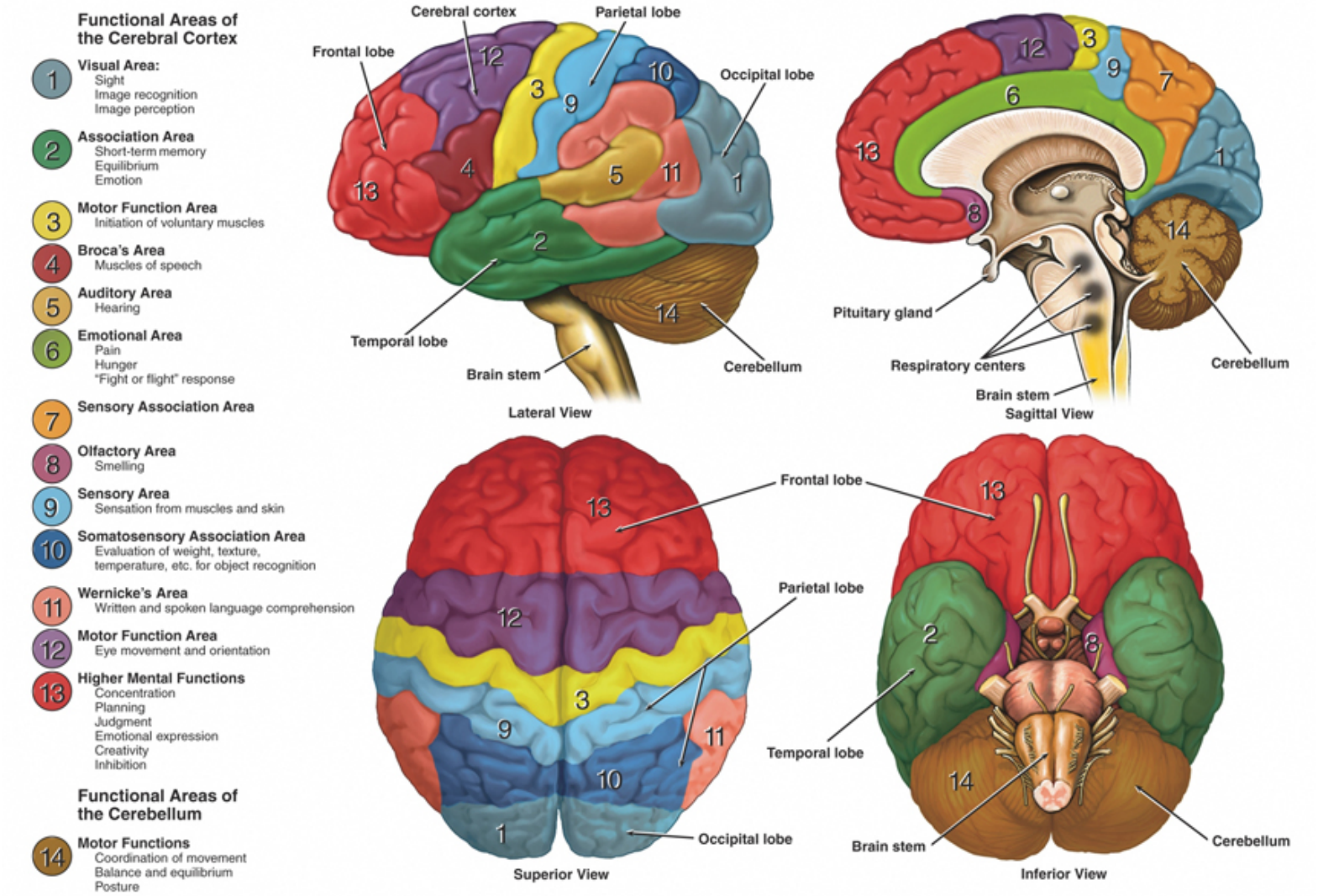


Figure: Brain Anatomy Functions

Single-Photon Emission Computed Tomography (SPECT)

SPECT is a nuclear medicine tomographic imaging technique using gamma rays for measuring the amount of blood flow to the brain.

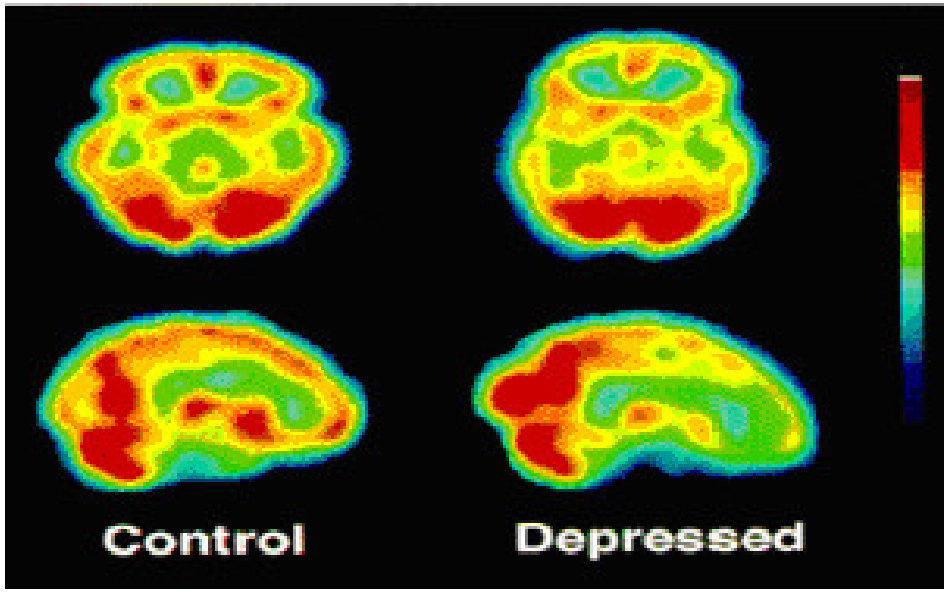


Figure: SPECT images from a depressed patient showing characteristic hypofrontality (decreased cerebral blood flow) relative to a healthy control subject.

Positron Emission Tomography (PET)

PET uses positron-emitting radioactive tracers that are attached to molecules entering biological pathways of interest.

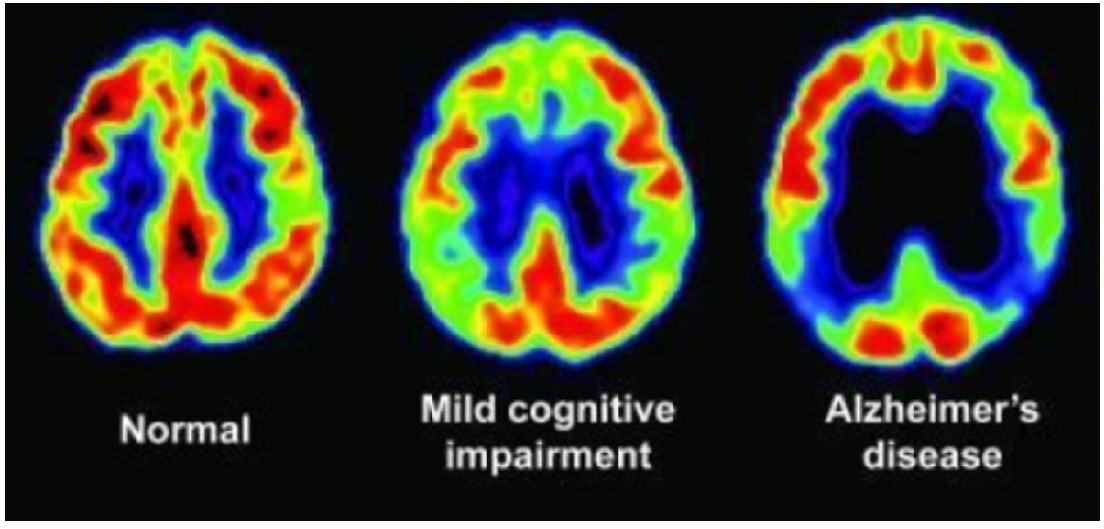


Figure: PET scans can detect the decline in glucose metabolism associated with decreased cognitive function, particularly in the temporal and parietal lobes, the regions associated with memory formation and language

Ultrasound imaging

Involves exposing part of the body to high frequency sound waves to produce pictures of the inside of the body.

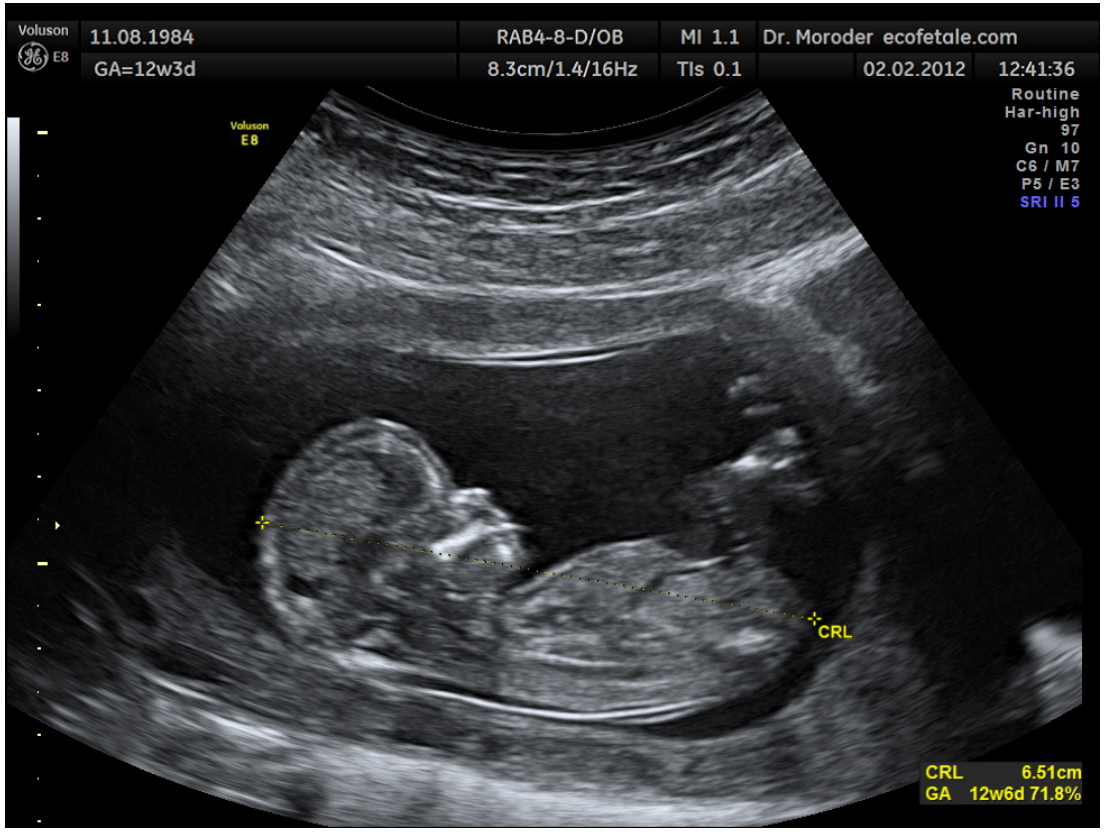


Figure: Ultrasound image of a fetus in the womb, viewed at 12 weeks of pregnancy

Magnetic Resonance Imaging (MRI)

Use a powerful magnetic field to align the magnetization of some atoms in the body, then uses radio frequency fields to systematically alter the alignment of this magnetization. This causes the nuclei to produce a rotating magnetic field detectable by the scanner. There is no ionizing radiation.

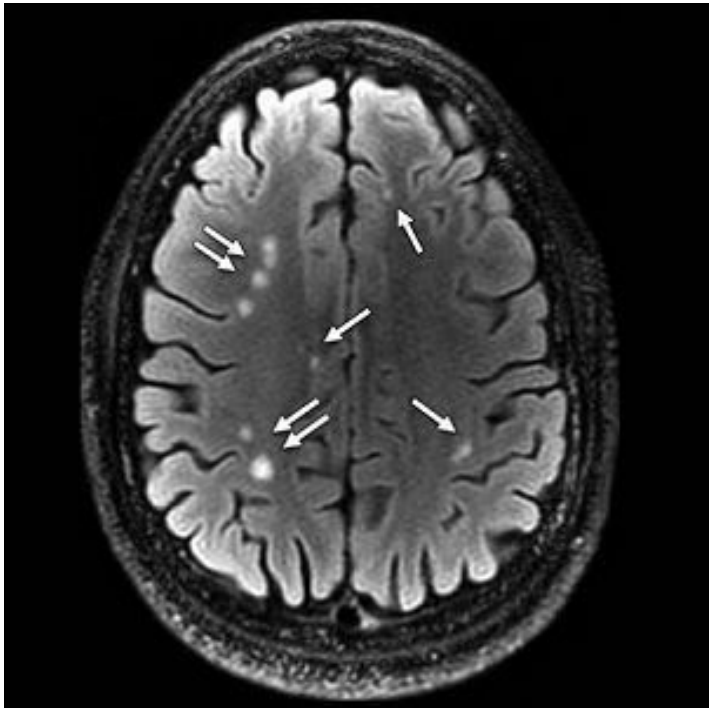


Figure: MRI shows brain damage (white matter scars) among U.S. military personnel who suffered blast-related mild traumatic brain injury or concussion.

Functional MRI (fMRI)

Measures the hemodynamic response (change in blood flow) related to neural activity in the brain or spinal cord of humans or other animals.

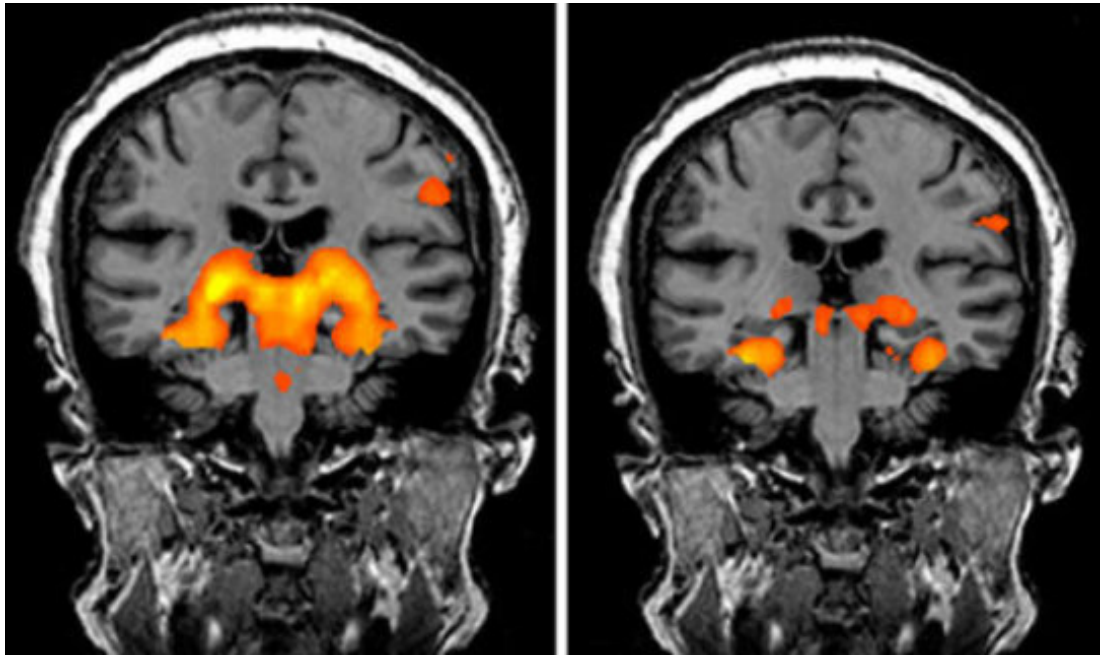


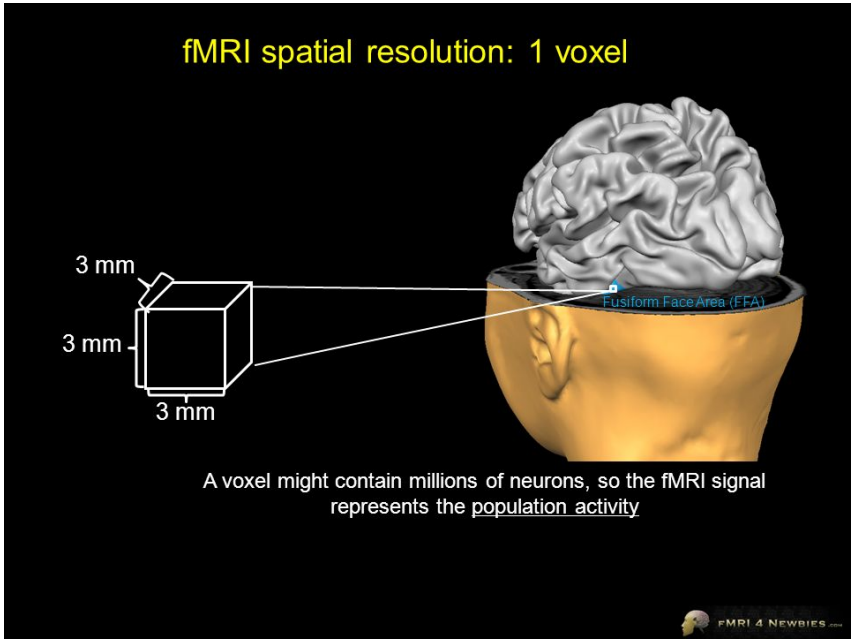
Figure: fMRI scans of study participants trying to remember object locations. Activity in the hippocampus is stronger in healthy elderly controls (left) compared with people with mild cognitive impairment (right).

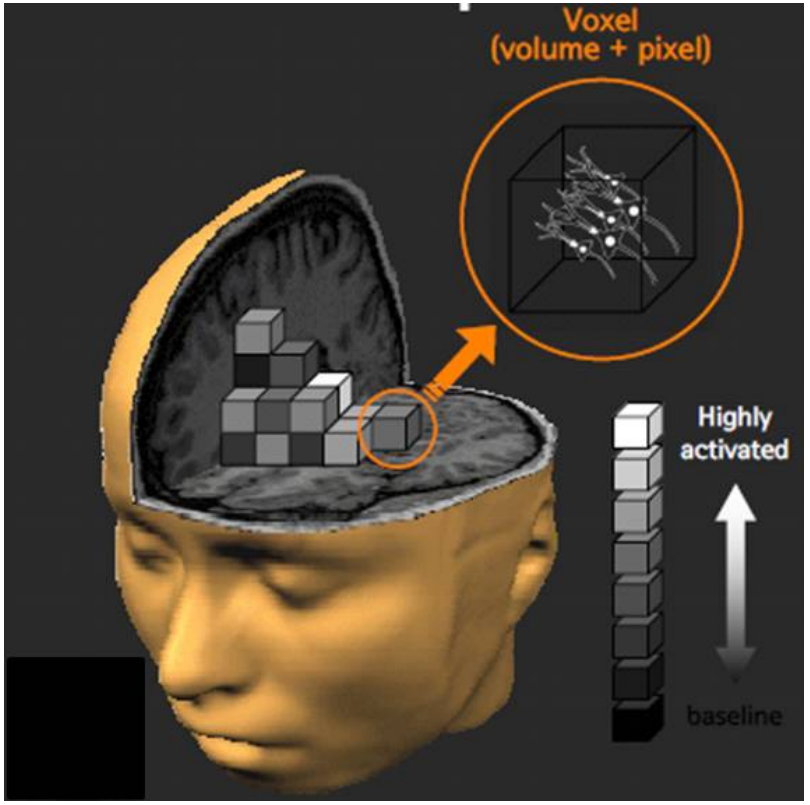
Bold fMRI

- The most common approach in fMRI uses the Blood Oxygenation Level Dependent (BOLD) contrast.
- BOLD fMRI allows us to measure the ratio of oxygenated to deoxygenated hemoglobin in the blood.
- It is important to note that BOLD fMRI does not measure neuronal activity directly, instead it measures the metabolic demands (oxygen consumption) of active neurons.

fMRI: Voxel

- Each fMRI image consists of about 100,000 brain voxel (cubic volumes that spans the 3D space of the brain
- During the course of an experiment, several hundreds of images are obtained (about 1 every 2 seconds)





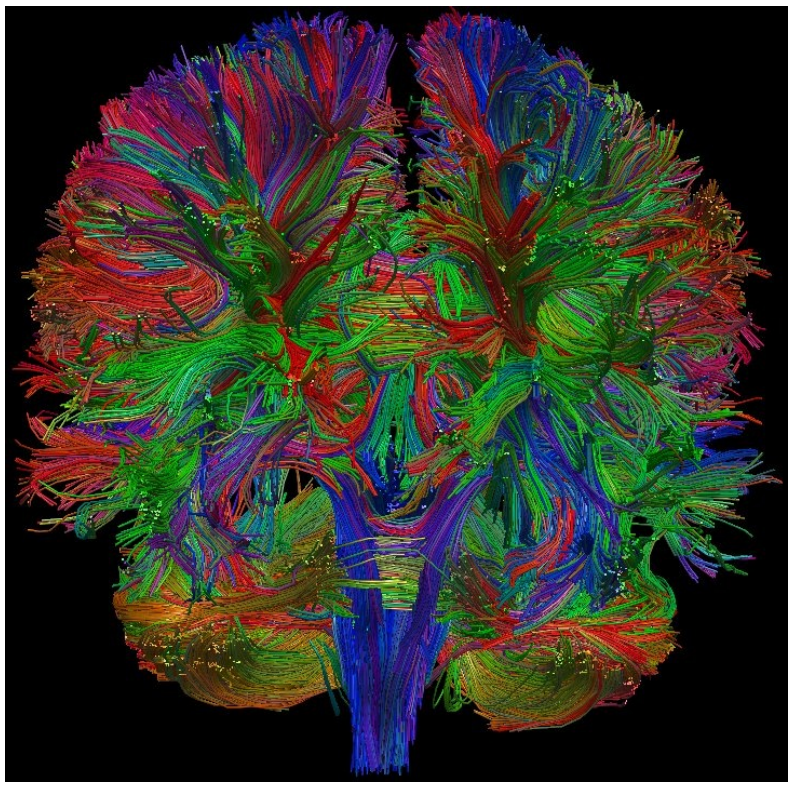
Experiment Design

- Block design: Each condition (active or sleeping) is presented for an extended period of time.

- Event-related design: Each event (given a certain task) is presented for a short duration

Diffusion Tensor MRI

- Use the diffusion of water molecules to generate contrast in MR images.
- Water molecule diffusion patterns reveal microscopic details about tissue architecture, either normal or in a diseased state. It is used extensively to map white matter tractography in the brain.
- Provide models of brain connectivity



Histology

Study the microscopic anatomy (microanatomy) of cells and tissues of plants and animals through examining cells and tissues under a light microscope or electron microscope. The specimen is cut into a thin cross section with a microtome, stained, and mounted on a microscope slide.

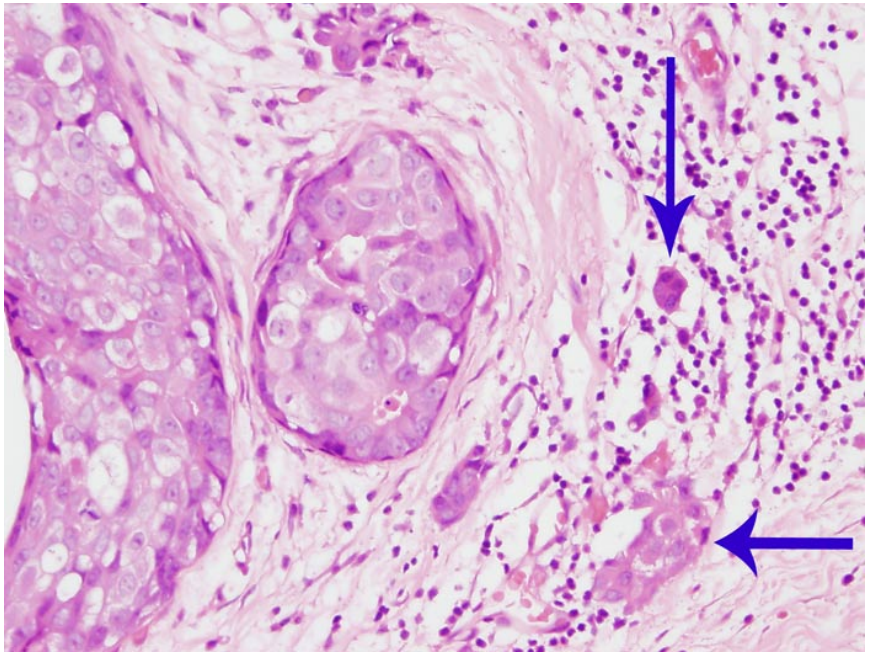
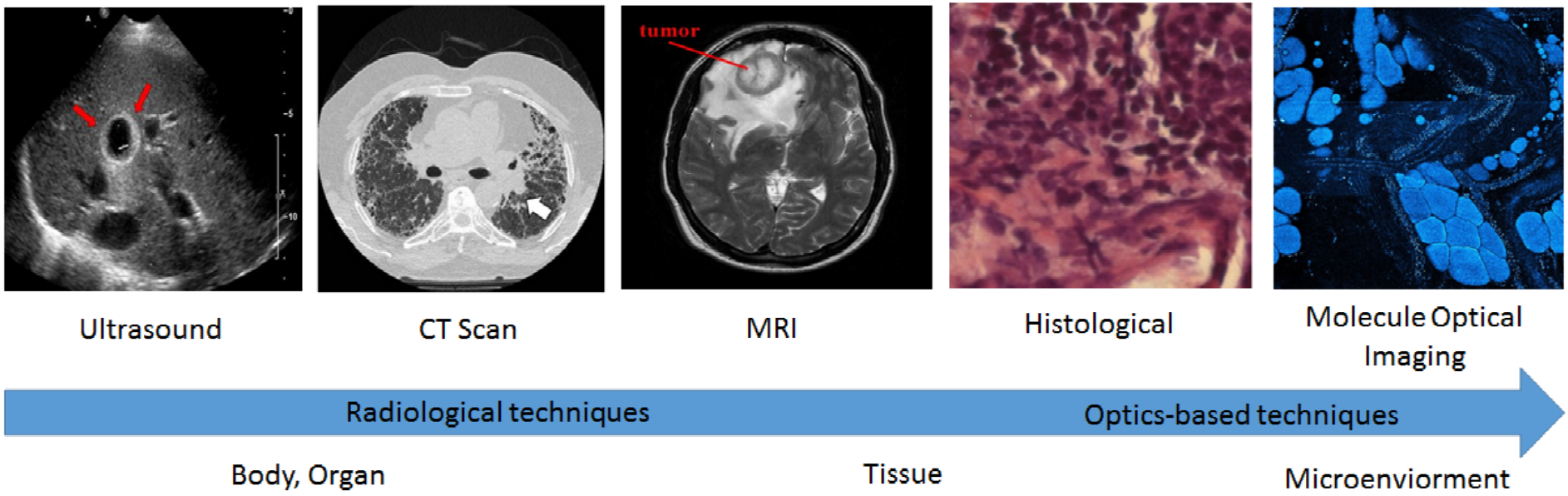


Figure: High grade of ductal carcinoma in situ (DCIS) for breast cancer

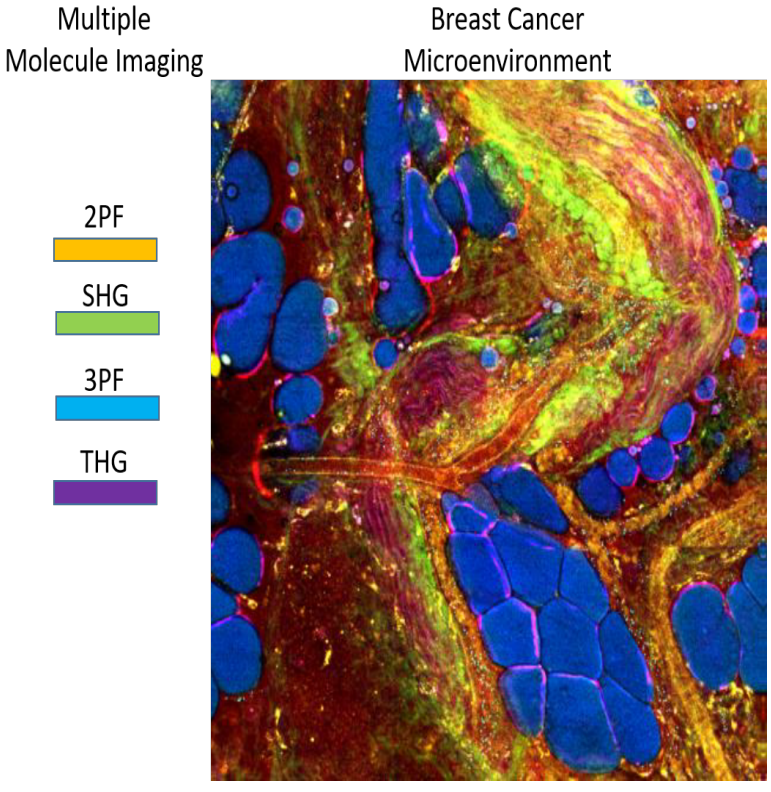
Biomedical Imaging for Breast Cancer Diagnosis



- **Mammograph**: Not effective to distinguish high-risk from indolent cancer cases
- **MRI**: Lacks accuracy in early diagnosis at molecule level
- **Histological**: external labeling (staining) could lead tissue distortion

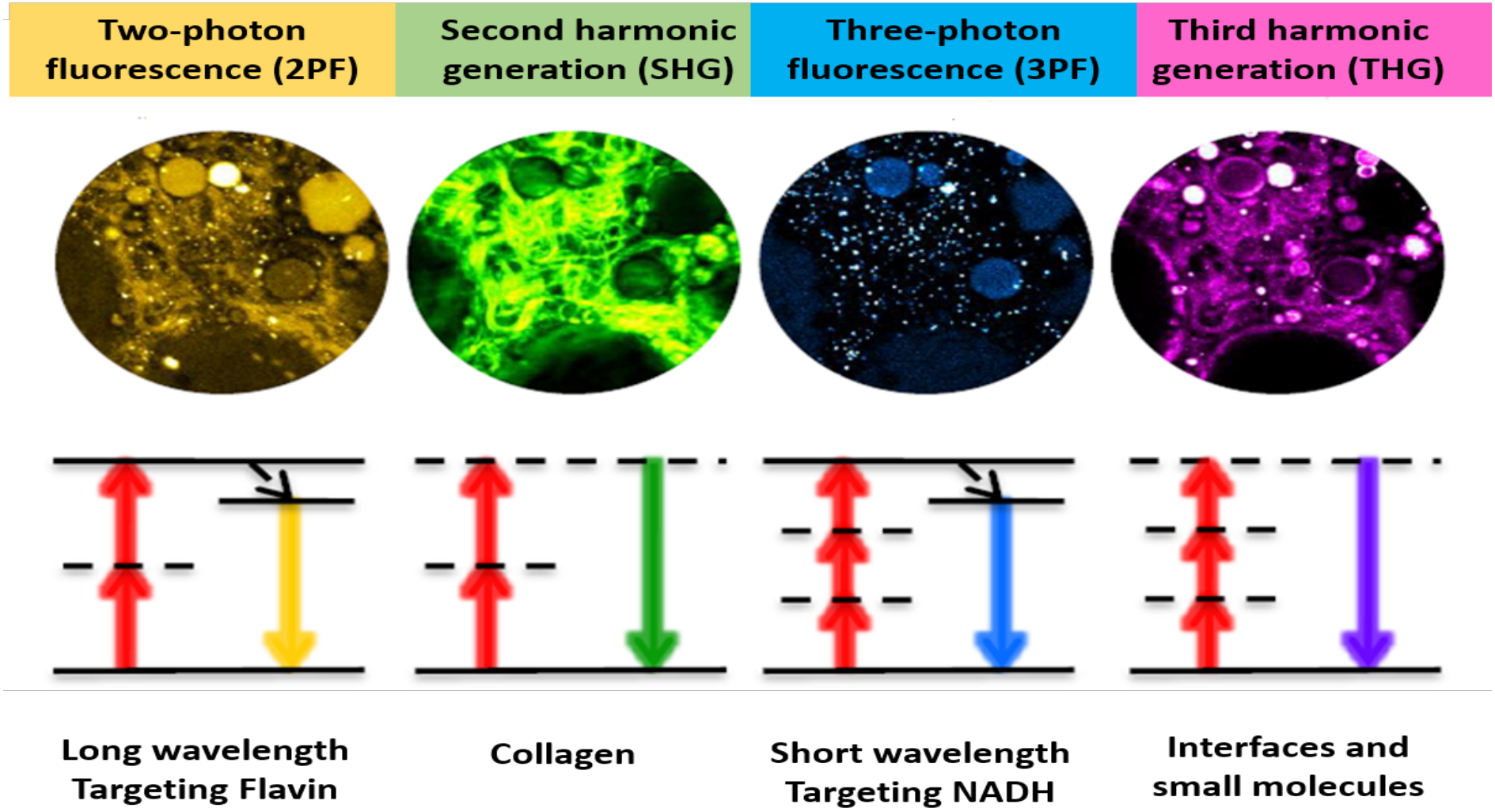
Optical Imaging: Multiphoton Molecule Optical Imaging

- Intrinsic **molecule contrast**
- **Deeper** penetration
- **Label free**: reduce photo damage
- **Non-invasive** real-time **in vivo** imaging



Multi-contrast by Multiphoton Nonlinear Imaging

- Different contrasts at **tissue**, **cellular** and **molecular** levels



Tumor-associated Microvesicles for Breast Cancer

Detect **molecular changes** during early stages of **cancer development**

- **Tumor-associated microvesicles**: optical biomarkers (Tu et al., 2016)
- More frequently observed from non-tumor tissues among cancer patients

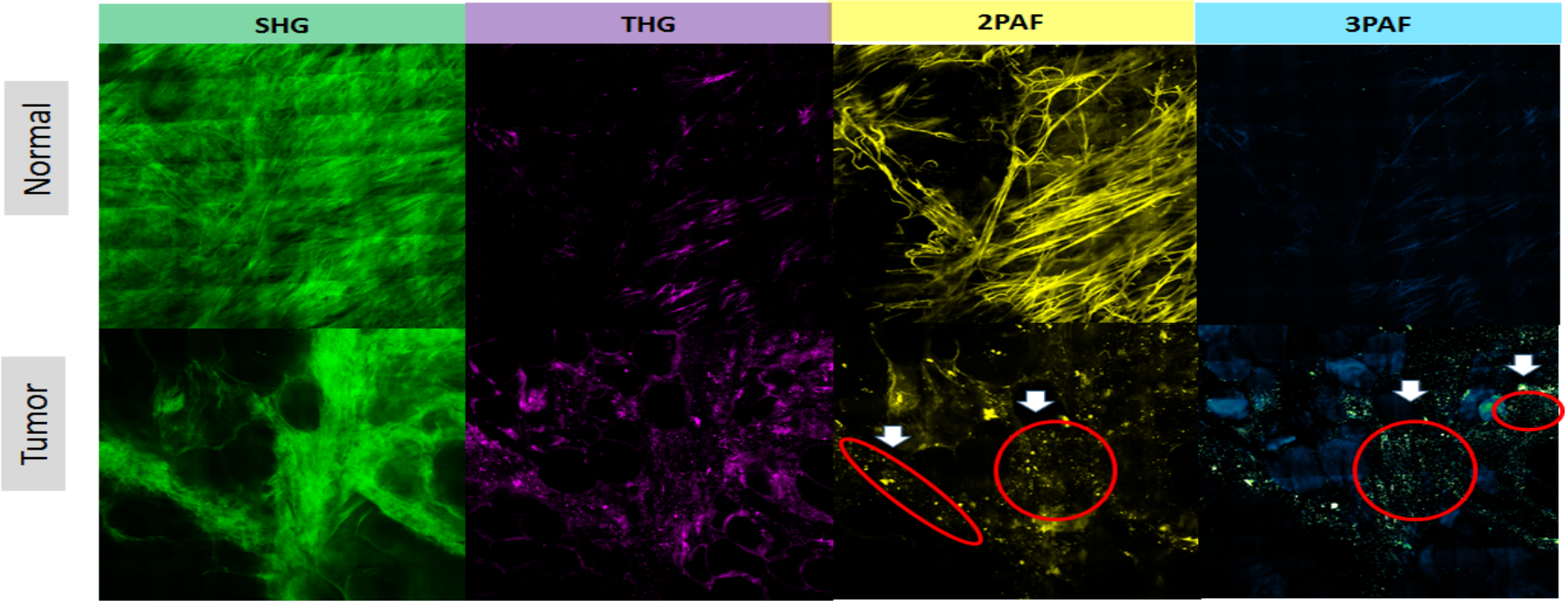


Figure: Four-modality imaging data from a healthy human tissue and a cancerous human tissue, image size of $4,000 \times 4,000$ pixels, each pixel of $0.5\mu m$.

Challenges in Imaging Analysis

- **Non-vector data**: 2D image (matrix), 3D (over time), even higher order (multiple modalities)
- **Ultrahigh dimensionality**: curse of dimensionality
- **Spatial information**: neighbouring pixels are highly correlated
- **Integration from multiple sources**: images from different modalities and time points
- Require computationally efficient analysis procedures and algorithms

Generalized Linear Model or Support Vector Machine Model

- Generalized linear model:

$$g(\boldsymbol{\mu}_i) = \boldsymbol{\beta}^T \boldsymbol{x_i}$$

- Support vector machine model:

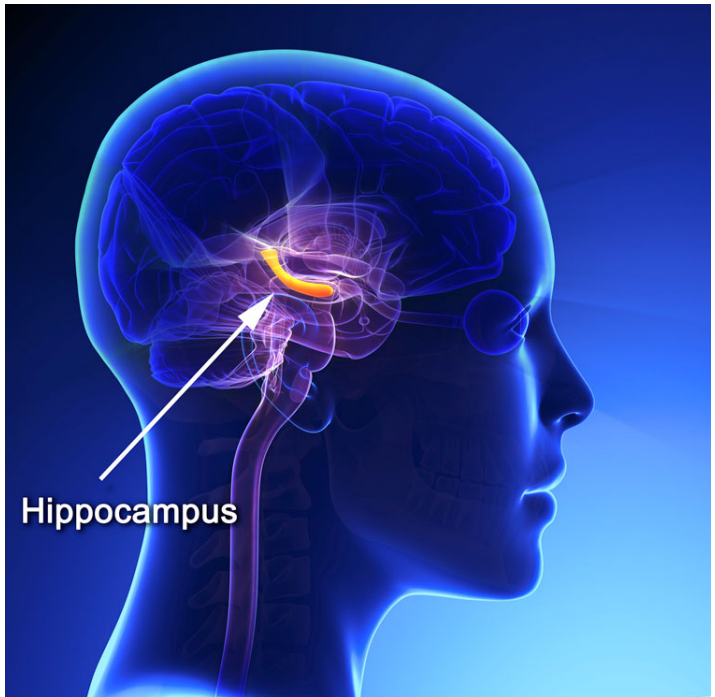
$$\min \sum_{i=1}^{n} \left(1 - \boldsymbol{y}_i(\boldsymbol{\beta}^T \boldsymbol{x}_i)\right)_+$$

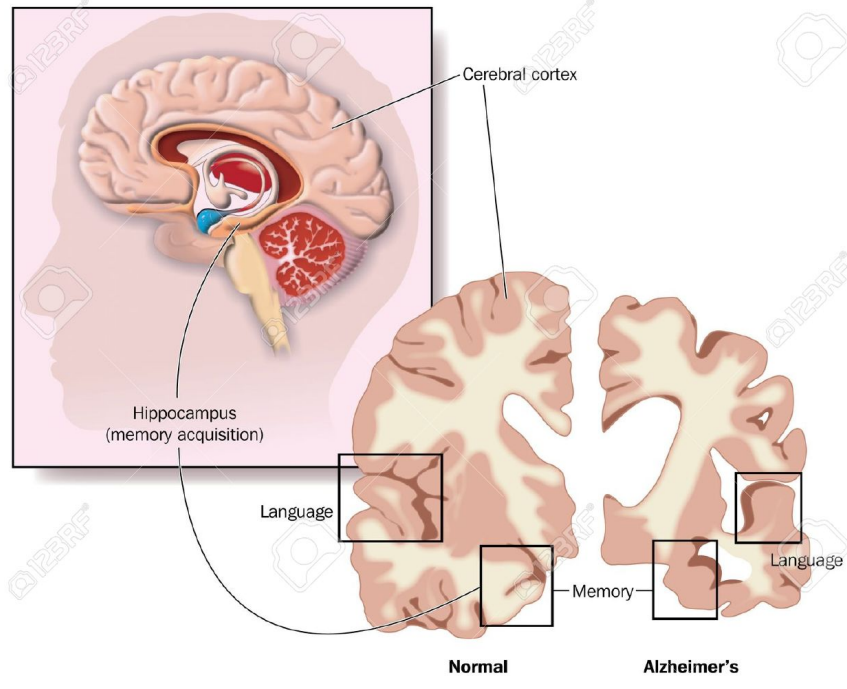
- $\boldsymbol{y}_i$ is the response outcome, $\boldsymbol{x}_i$ are covariates associated with imaging, treatment or other demographical information.

- $g(\cdot)$ is a link function, $\mu_i = \mathbf{E}(\boldsymbol{y}_i)$

Example: Hippocampal Volume (HCV)

- Use HCV as response: HCV $\sim$ Age + Gender+ Diagnosis
- Use HCV as a preditor: Diagnosis$\sim$ Age + Gender+ HCV
- Diagnosis: normal control (NC) or Alzheimer's disease (AD).
- For Alzheimer's disease, shrinkage is especially severe in the hippocampus, located in the medial temporal lobe of the brain that plays a key role in formation of new memories.



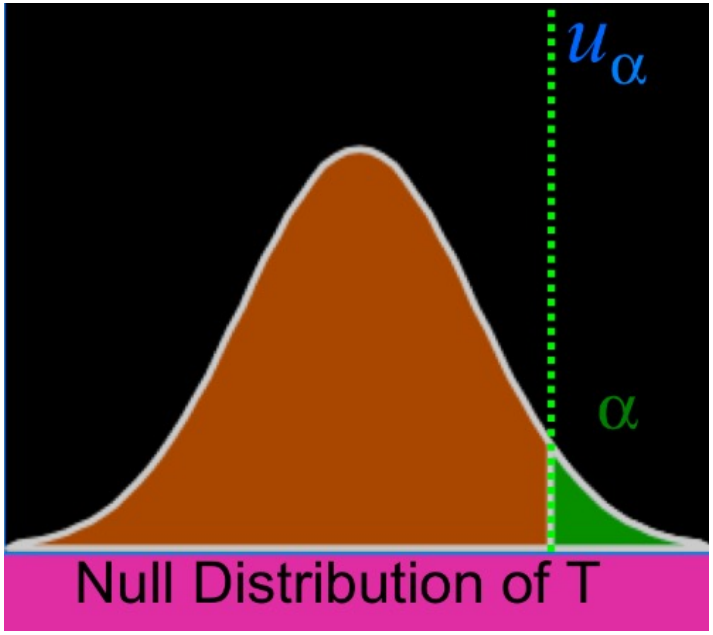


Spatial Hypotheses

- How do we extend from standard univariate hypothesis testings to answering spatially motivated questions?
- Not easy: high dimensionality.
- Testing millions of voxels instead of one HCV.
- False positive rate could be very high.

Hypothesis Testing

- Null Hypothesis H_0
- Test statistic T
- α-level
  - Acceptable false positive risk
  - Level $\alpha = \Pr(T > u_\alpha | H_0)$
  - Threshold u_α controls false positive risk at level α



Different Threshold Values

- Sensitivity: true positive.
- Specificity: true negative.

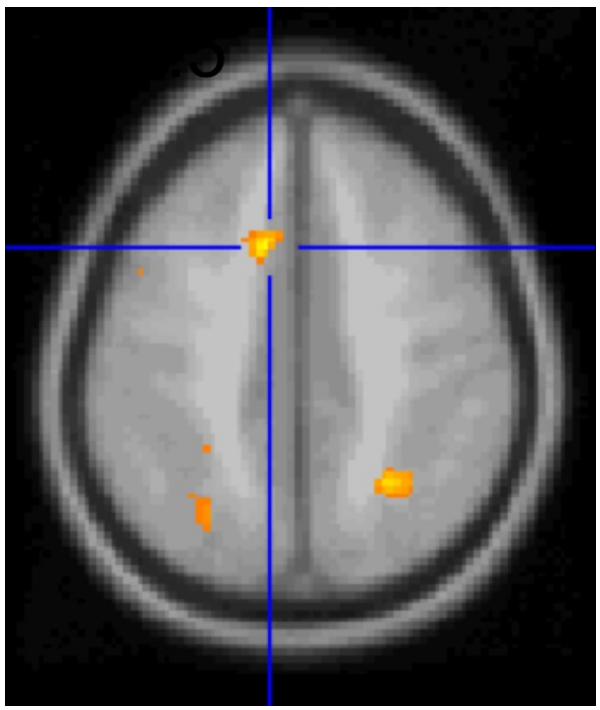


Figure: High threshold, good specificity, poor sensitivity

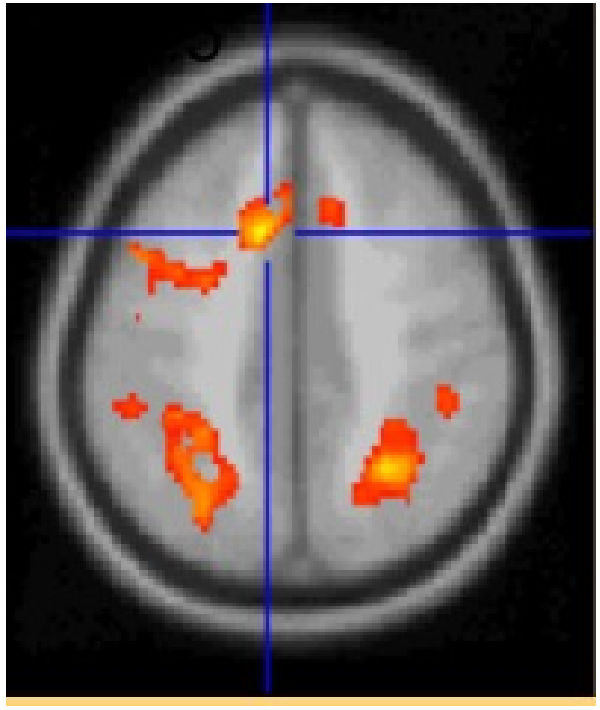


Figure: Medium threshold, a good balance of sensitivity and specificity

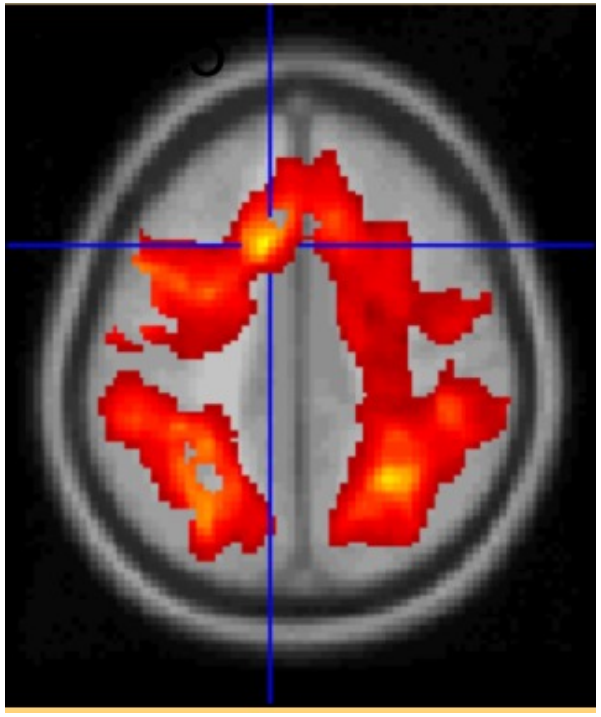


Figure: Low threshold, good sensitivity, poor specificity

Multiple Comparisons

- Family-wise Error Rate (FWER): the probability of making at least one false discoveries (type I errors) for multiple hypotheses testings.
- False discovery rate (FDR)
  - FDR $= E[FP/(TP + FP)]$
  - TP+FP: voxels are tested positive
  - Realized false discovery rate: FP/(TP+FP)
- Bonferroni Correction: Conservative
  - FWE: α
  - For N independent voxels: $\alpha = N\nu$ (ν = voxel-wise error rate)
  - To control FWE, set $\nu = \alpha/N$

False Discovery Rate

- For any threshold, all voxels can be cross-classified:

	H_0 Retained	H_0 Rejected	Total
H_0 True	TN	FP	T_0
H_0 False	FN	TP	T_1
Total	M	P	N

- Realized FDR

$$rFDR = FP/(TP + FP) = FP/P$$

  - Special case: if $P = 0$, rFDR $= 0$
- We can only observe M, P and N, do not know TP or FP

Control False Discovery Rate

- Control FDR

$$FDR = E(rFDR) = E(\frac{FP}{P}) \leq \alpha$$

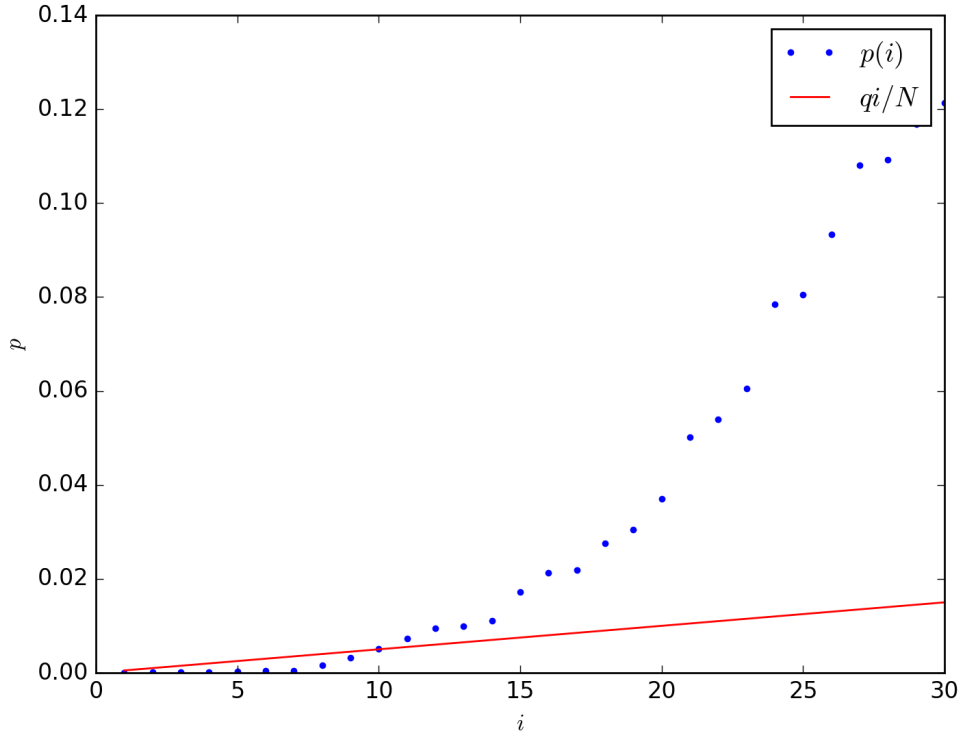
- Benjamini and Hochberg (1995) (BH) introduced the FDR procedure.
- Define p-values $(P_1, \ldots, P_N)$ for the N tests.
- Let $P_{(0)} = 0$, and order the p-values:

$$P_{(0)} = 0 < P_{(1)} < \ldots < P_{(N)}.$$

Benjamini and Hochberg (1995)

- Given a pre-specified false discovery rate $0 < \alpha < 1$, the BH threshold value is

$$T_{BH} = \max_i \left\{ P_{(i)} : P_{(i)} \leq \alpha \frac{i}{N}, 0 \leq i \leq N \right\}$$



Benjamini and Hochberg (1995)

- Let r be the largest i such that $P_{(i)} \leq \alpha \frac{i}{N}$
- Reject all hypotheses corresponding to $P_{(1)}, \ldots, P_{(r)}$
- BH (1995) show (for independent tests) that this procedure guarantees

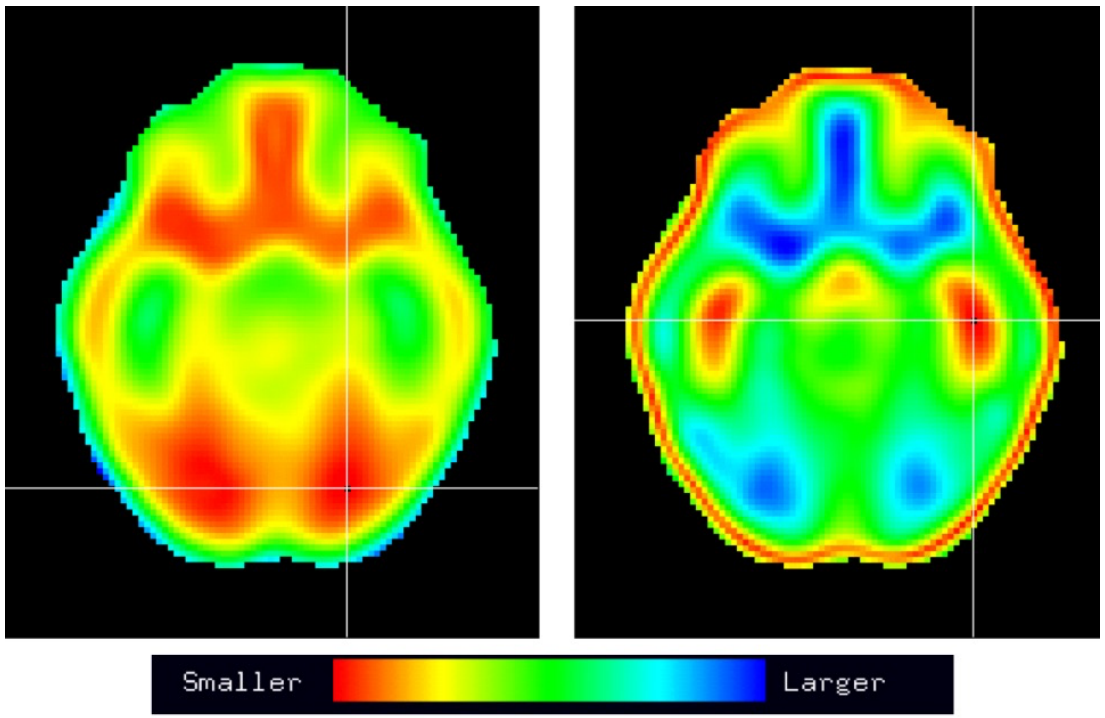
$$FDR = E\big(rFDR(T_{BH})\big) \leq \frac{T_0}{N}\alpha \leq \alpha,$$

where T_0 is the number of true H_0.

Spatial Modeling

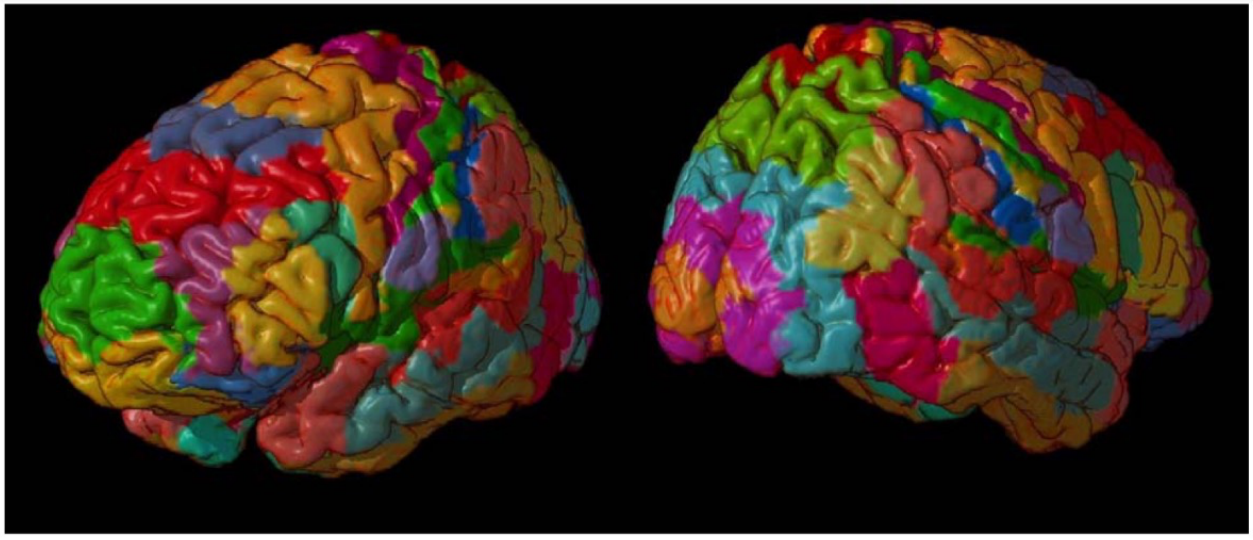
- Treat correlation as a function of distance.
- Define distance function between activity in voxels i and j:

$$d_{ij} = \{(\mu_i - \mu_j)'(\mu_i - \mu_j)\}^{1/2}, \mu_i \text{ is a summary statistics.}$$



Regions of interests

- Regions of interests using neuroanatomical templates

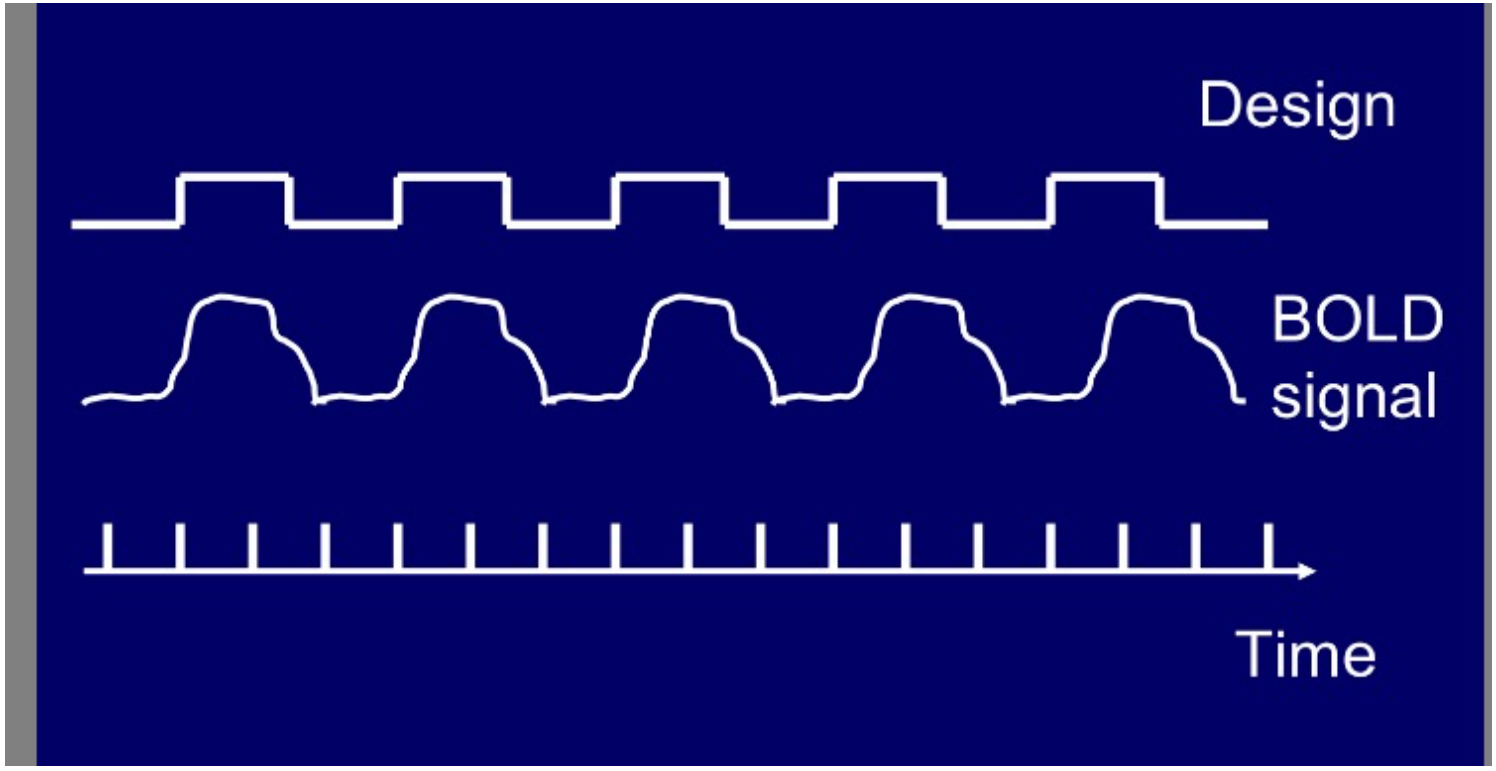


$G = 116$ (Anatomical regions); [Tzourio-Mazoyer et al., 2002]

$G = 52$ (Brodmann regions) [Brodmann, 1909]]

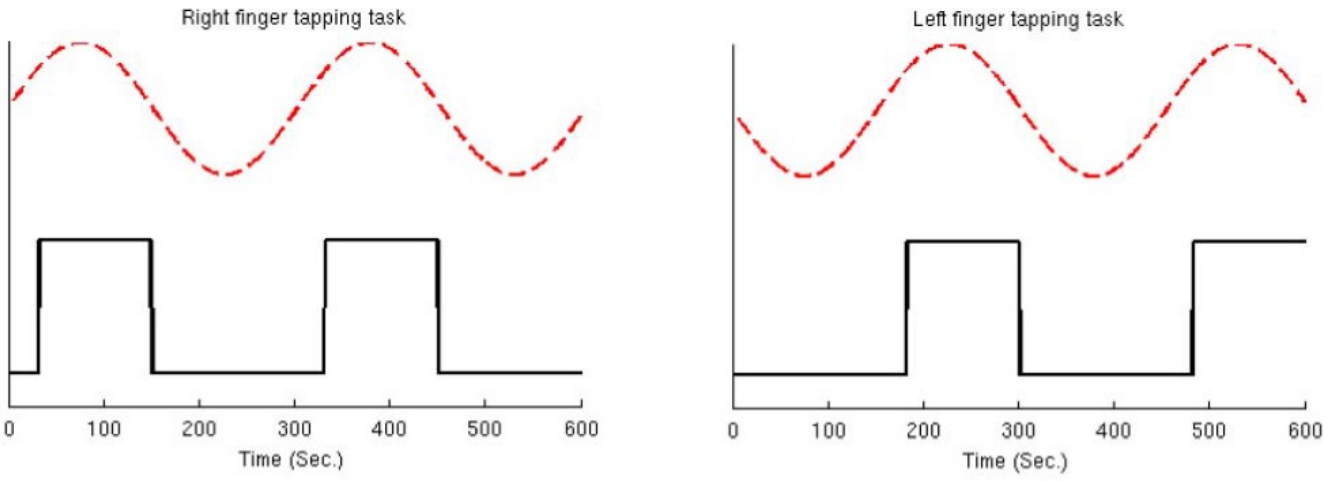
- Alternatively, partition V voxels into G regions using data - driven clustering methods based on the functional distances (e.g., Bowman et al., 2004; Bowman and Patel, 2004; Bowman, 2005)

Idealized fMRI Block Design Experiment



Left and Right Finger Tapping

(a) task paradigm



(b) Task Functions of right and left hand

Lee et al. (2012)

Spatial Modeling Approaches (Bowman, Caffo, Bassett, Kilts (2008, NeuroImage)

- Spatial model:

$$\boldsymbol{y}_{ig}(v) = \boldsymbol{x}_{igv}\boldsymbol{\beta}_{ig}(v) + \boldsymbol{H}_{iv}\boldsymbol{\nu}_i(v) + \boldsymbol{\varepsilon}_{ig}(v),$$

$$i = 1, \ldots, s, g = 1, \ldots, G, v = 1, \ldots, V$$

where

$\boldsymbol{y}_{ig}$	$s \times 1$	serial brain activity at location (e.g. BOLD) (within cluster g)
$\boldsymbol{x}_{igv}$	$s \times q$	design matrix (common to all v) for cluster g
$\boldsymbol{\beta}_{ig}$	$q \times 1$	parameter vector containing individualized effects
$\boldsymbol{\varepsilon}_{ig}$	$s \times q$	random error
$\boldsymbol{H}_{iv}$	$s \times m$	Other covariates (e.g., high-pass filtering matrix to remove unwanted low-frequency trends from the data)
$\boldsymbol{\nu}_i$	$m \times 1$	parameters corresponding to other coavariates

- s is the number of scans, q is the number of stimulants, V is the number of voxels, and G is the number of regions

Spatial Bayesian Hierarchical Modeling

- Spatial model: j represents the jth stimulant (task)

$$
\begin{aligned}
\boldsymbol{\beta}_{igj}(v) \mid (\mu_{gj}(v), \alpha_{igj}, \sigma^2_{gj}) &\sim N(\mu_{gj}(v) + \alpha_{igj}, \sigma^2_{gj}), \\
\mu_{gj}(v) \mid \lambda^2_{gj} &\sim N(\mu_{0gj}, \lambda^2_{gj}), \text{ voxel means}, \\
{\sigma_{gj}}^{-2} &\sim \mathit{Gamma}(a_0, b_0), \text{ common voxel variance}, \\
\boldsymbol{\alpha}_{ij} \mid \boldsymbol{\Gamma}_j &\sim N(0, \boldsymbol{\Gamma}_j), \text{ subject-specific random effects} \\
\lambda_{gj}^{-2} &\sim \mathit{Gamma}(c_0, d_0), \text{ variances of voxel means}, \\
\boldsymbol{\Gamma}_j^{-1} &\sim \mathit{Wishart}((h_0 \boldsymbol{H}_{0j})^{-1}, h_0), \text{ spatial covariance matrix.}
\end{aligned}
$$

- $\boldsymbol{\alpha}_{ij} = (\alpha_{i1j}, \ldots, \alpha_{iGj})'$.
- Priors:
  - $(a_0, b_0) = (0.1, 0.005), (c_0, d_0) = (0.1, 0.01)$
  - $h_0 = G$ ($h_0 \geq G$ to ensure proper posterior)
  - $\boldsymbol{H}_{0j} = \hat{\boldsymbol{\Gamma}}_j$ (sample covariance matrix)

Estimation

- MCMC: Gibbs Sampler
- Draw samples from the joint posterior distribution of all model parameters.
- Estimate functions of the model parameters from the joint posterior samples.
- E.g., Intra-regional task-related functional connectivity is given by

$$\rho_{gj} = \frac{\gamma_{gg}^{(j)}}{\gamma_{gg}^{(j)} + \sigma_{gj}^2}$$

- $\gamma_{gg}^{(j)}$ is the diagonal component of $\mathbf{\Gamma}_j$ corresponding to the gth region.
- Provide similarity in brain function between voxels within anatomical regions.
- Applying conjugate model leads to fast estimation: posterior distributions are in the same family as the prior distribution

Example: Cocaine Addicts (Bowman et al., 2007)

fMRI Study of Inhibitory Control in Cocaine Addicts

- $n = 28$ subjects ($n_p = 12$ cocaine addicts, $n_c = 16$ controls)
- Study Conditions: Inhibitory Control
  - ▷ Correctly inhibiting a prepotent response
    - ○ Response primed with frequent occurrence of "go cues"
    - ○ "Stop signal" is an auditory tone signaled after a "go cue"
- Sessions: Two Sessions (170 scans per subject in each session)
  - ▷ Addicts: Pre- and Post-Treatment
  - ▷ Controls: Baseline and Follow-up
- 176 million measurements per subject! 4.9 billion for all subjects!!
- Objective: Treatment-emergent neural processing changes related to response inhibition in cocaine-addicts

Example: Cocaine Addicts (Bowman et al., 2007)

Results

Treatment-emergent changes in activity related to inhibitory control for patients relative to controls

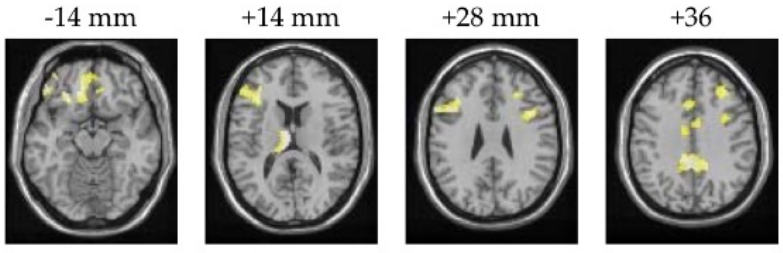
$$A = [(\text{Post-tx} - \text{Pre-tx in Patients}) - (\text{Follow up} - \text{BL in Controls})]$$

Target

- Voxel-Specific Activation Maps
- Regional Activation Maps
- Within-Region Task-Related FC
- Inter-Regional Task-Related FC

Example: Cocaine Addicts (Bowman et al., 2007)

fMRI Voxel-Specific Activation Maps



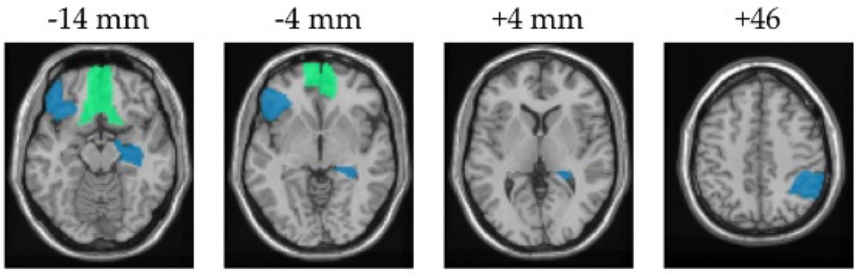
$$\Pr(A(v) > 0) \geq 0.70$$

Treatment-related changes are probable in

- -14 mm: Medial orbital frontal cortex (OFC) (BA 11)
- +14 mm: Left thalamus and left middle frontal gyrus (BA 46)
- +28 mm: Left and right middle frontal gyrus (BA 9)
- +36 mm: Cingulate gyrus, with most spatially extensive in the posterior cingulate gyrus (BA 31, 32)

Example: Cocaine Addicts (Bowman et al., 2007)

fMRI Regional Activation/Deactivation Maps



$\Pr(|A(g)| > 0) \geq 0.80$

Treatment-related changes are probable in

Increases:

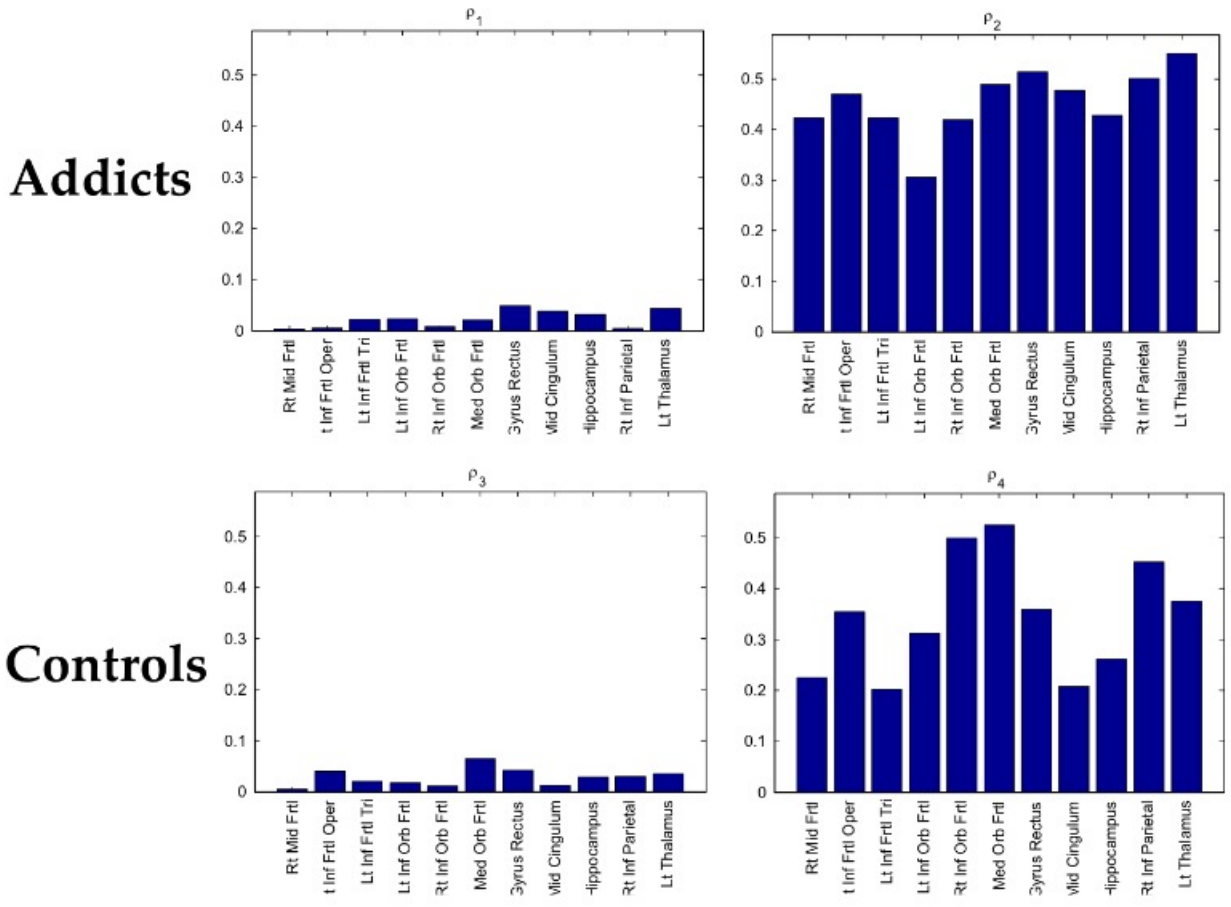
- -14 mm:Gyrus rectus and medial OFC
- -4 mm: Medial OFC

Decreases:

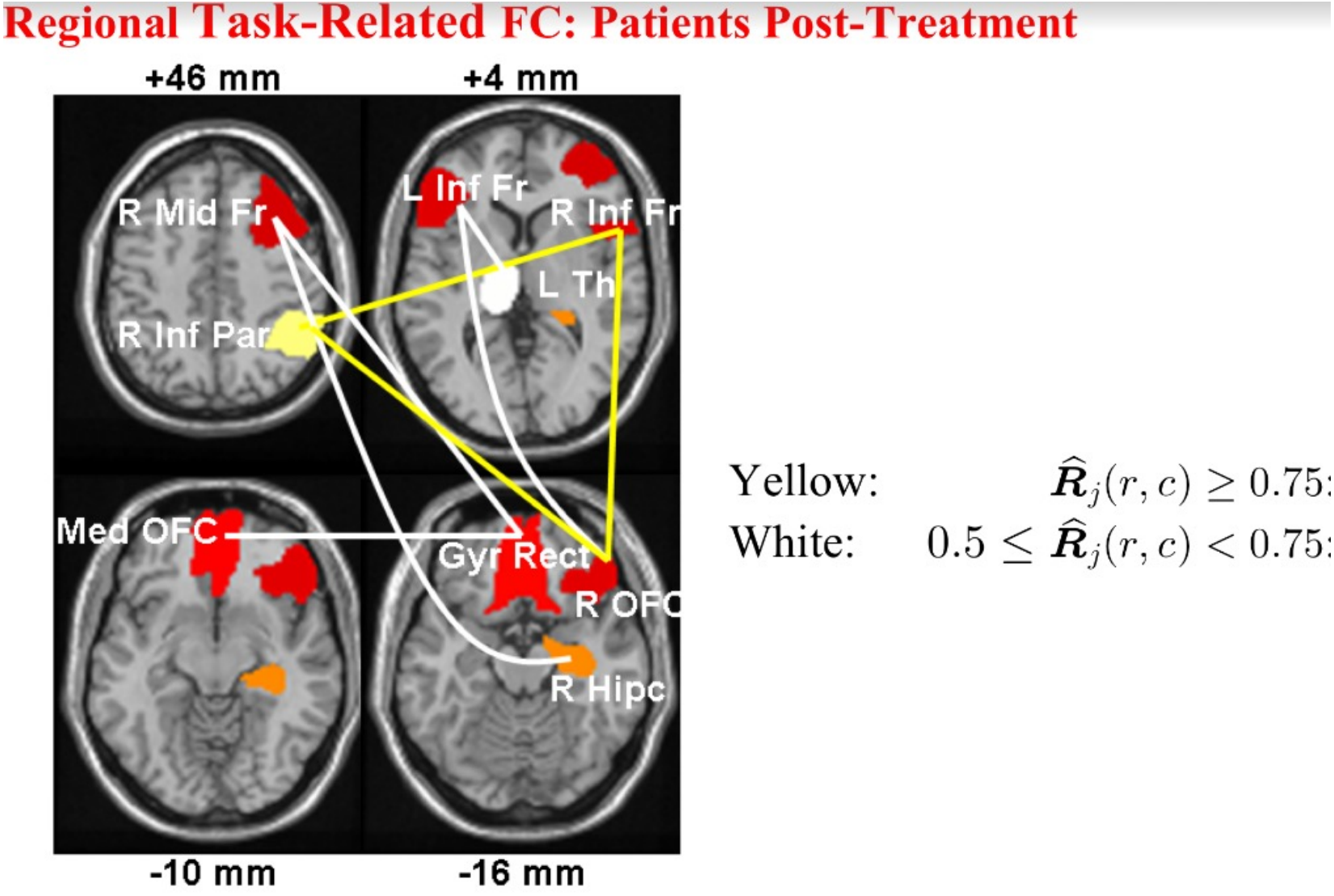
- -14 mm: Left inferior OFC
- -4 mm: Right hippocampus (from -14 to +4) and left inferior OFC
- +46 mm: Right inferior parietal region

Example: Within-region Brain Function Connectivity

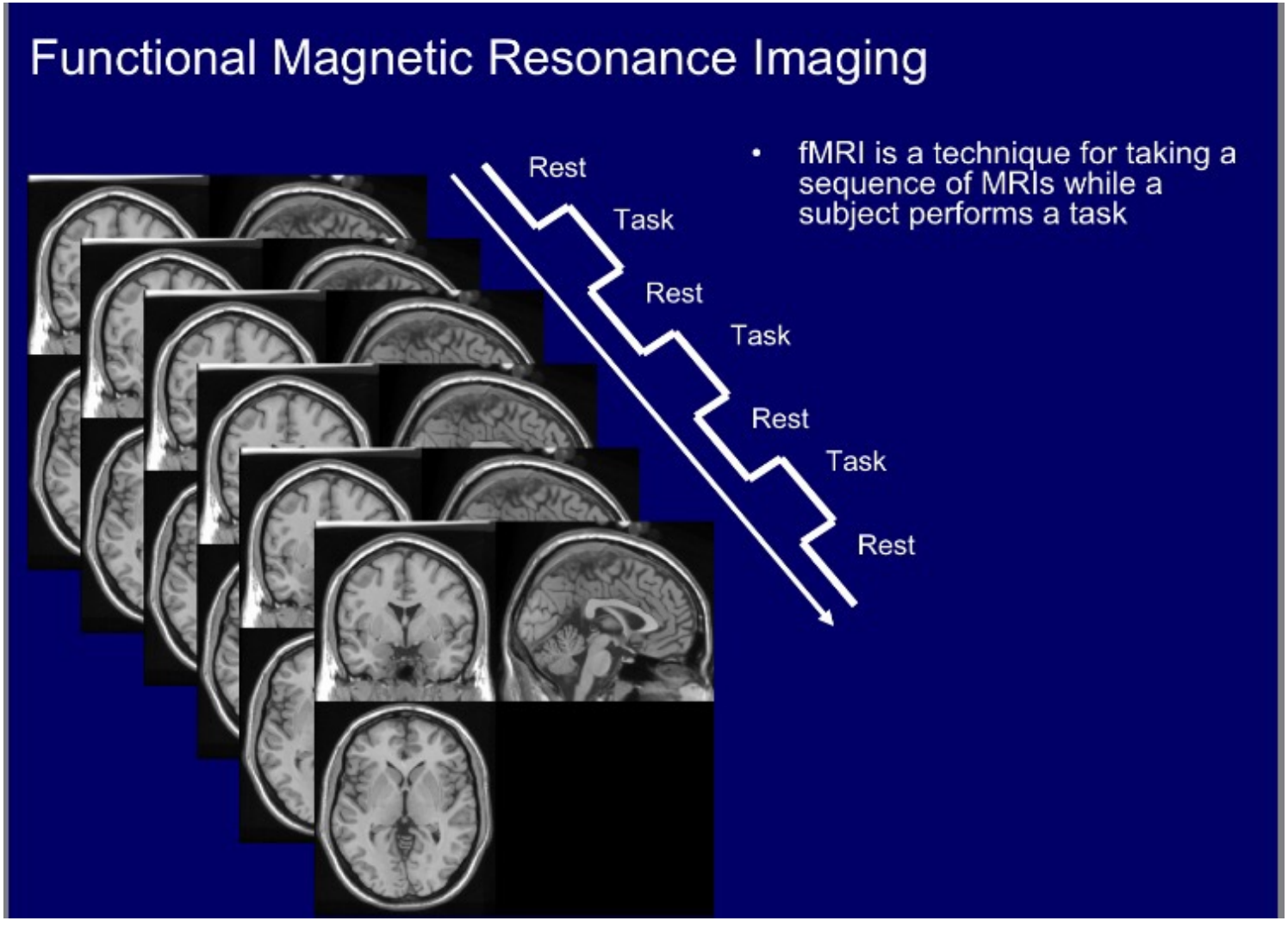
Within-Region Task-Related FC: $\rho_{gj} = \gamma_{gg}^{(j)} \Big/ \left(\gamma_{gg}^{(j)} + \sigma_{gj}^2\right)$



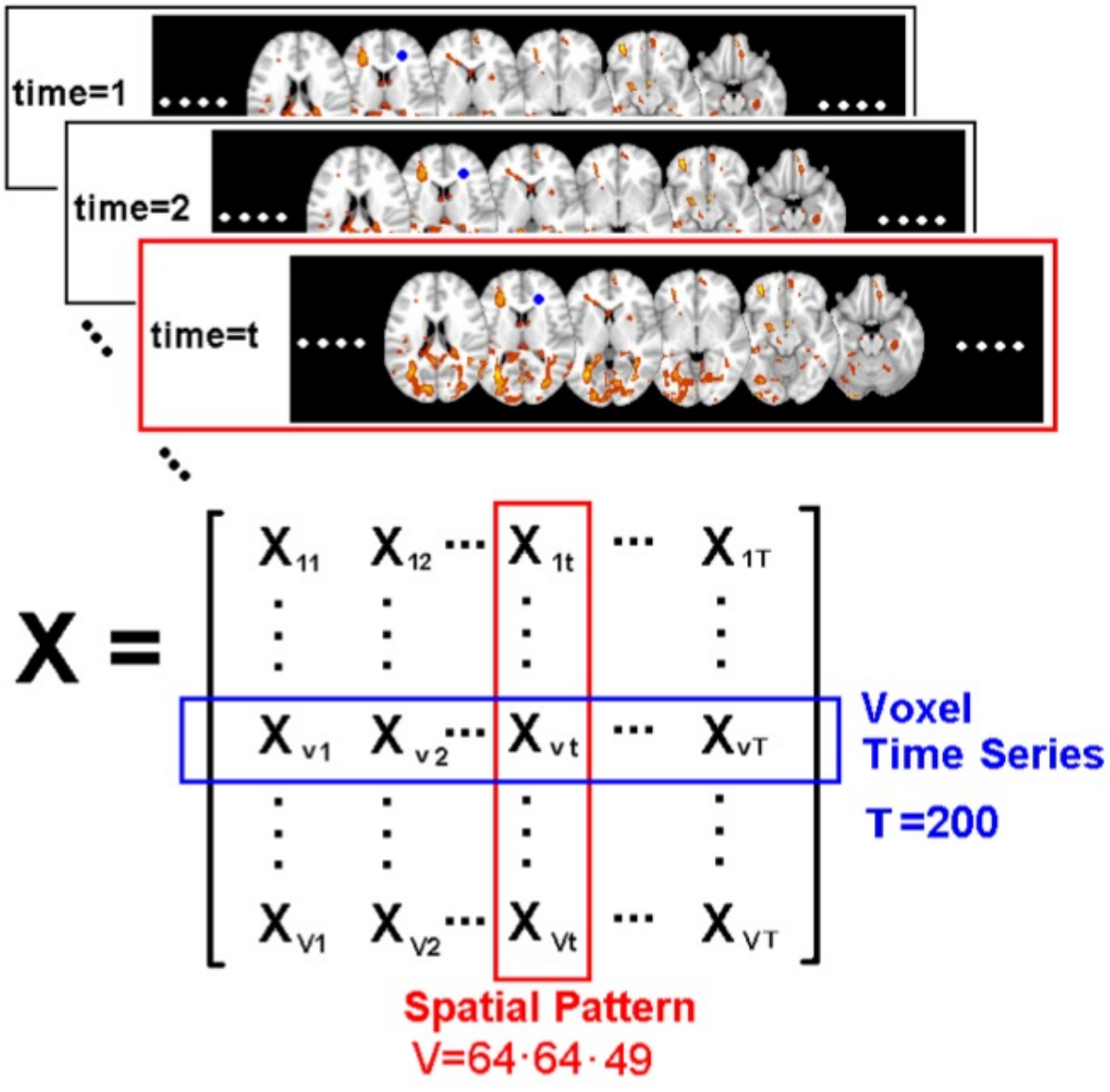
Example: Between-region Brain Function Connectivity



fMRI Over Time

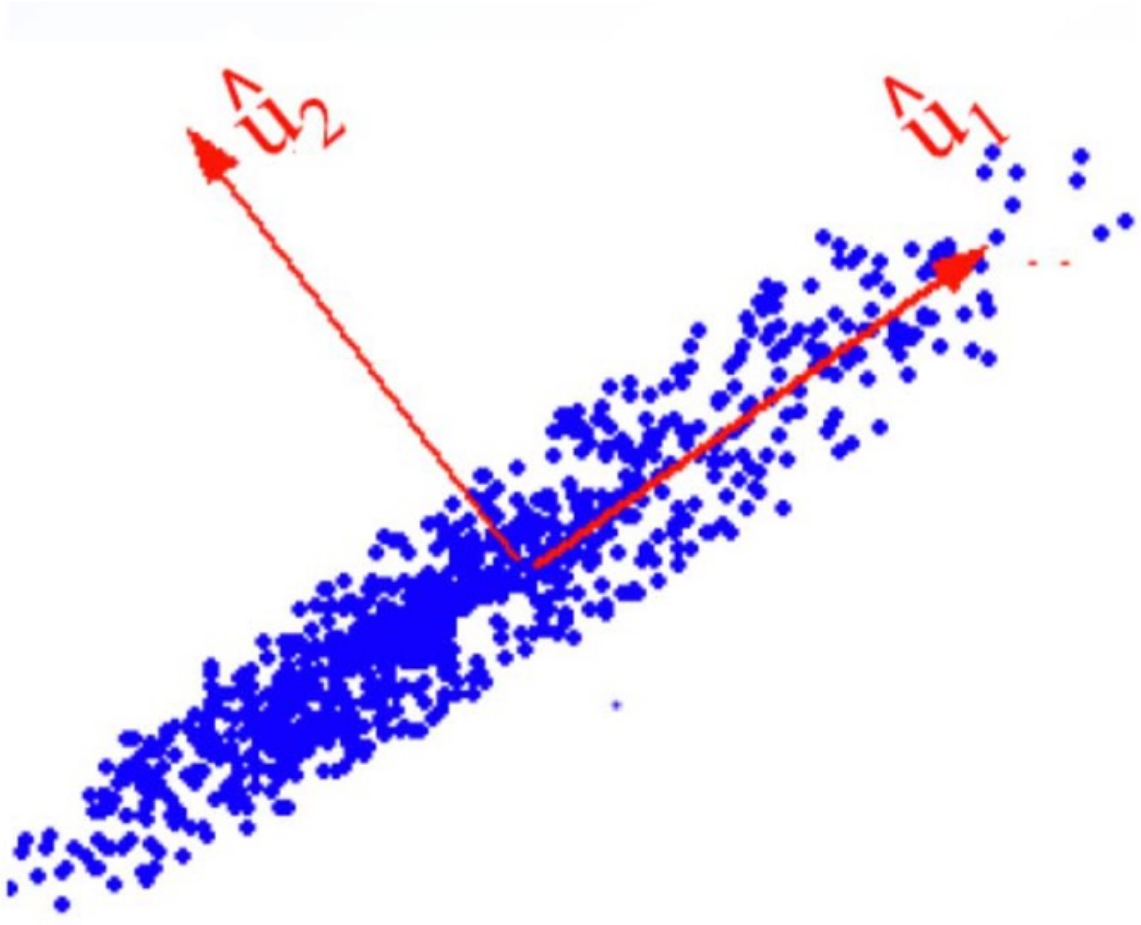


Spatial-Temporal Data Matrix



Principle Component Analysis (PCA)

- Idea: Obtain a few linear combination of the raw variables to explain the majority of the data variation.
- Effective dimension reduction tool.



Two Perspectives of PCA

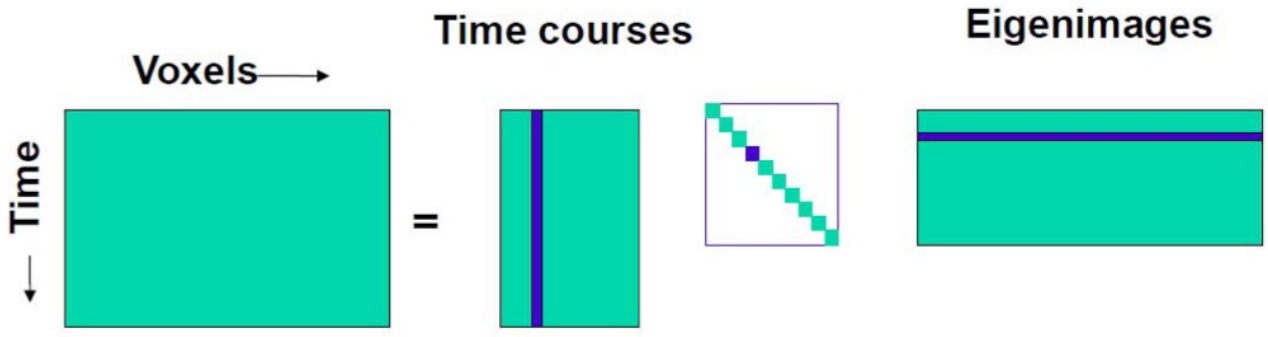
- Perspective 1: maximizing variance.

$$\boldsymbol{u}_j = \text{argmax}_{\{\boldsymbol{u}:\|\boldsymbol{u}\|=1\}} \text{Var}\,(\boldsymbol{X}'\boldsymbol{u}),\ \text{s.t.}\ \boldsymbol{u}_j'\boldsymbol{u}_{j'} = 0$$

- $\boldsymbol{u}_j$: jth PC loading vector
- Perspective 2: low-rank approximation.

$$\text{argmax}_{\{\boldsymbol{u}_k,\boldsymbol{v}_k:\|\boldsymbol{u}_k\|=1\}} \|\boldsymbol{X} - \sum_{k=1}^{r} \boldsymbol{u}_k \boldsymbol{v}_k'\|_F^2,\ \text{s.t.}\ \boldsymbol{u}_j'\boldsymbol{u}_{j'} = 0, \boldsymbol{v}_{j'}'\boldsymbol{v}_{j'} = 0$$

Singular Value Decomposition

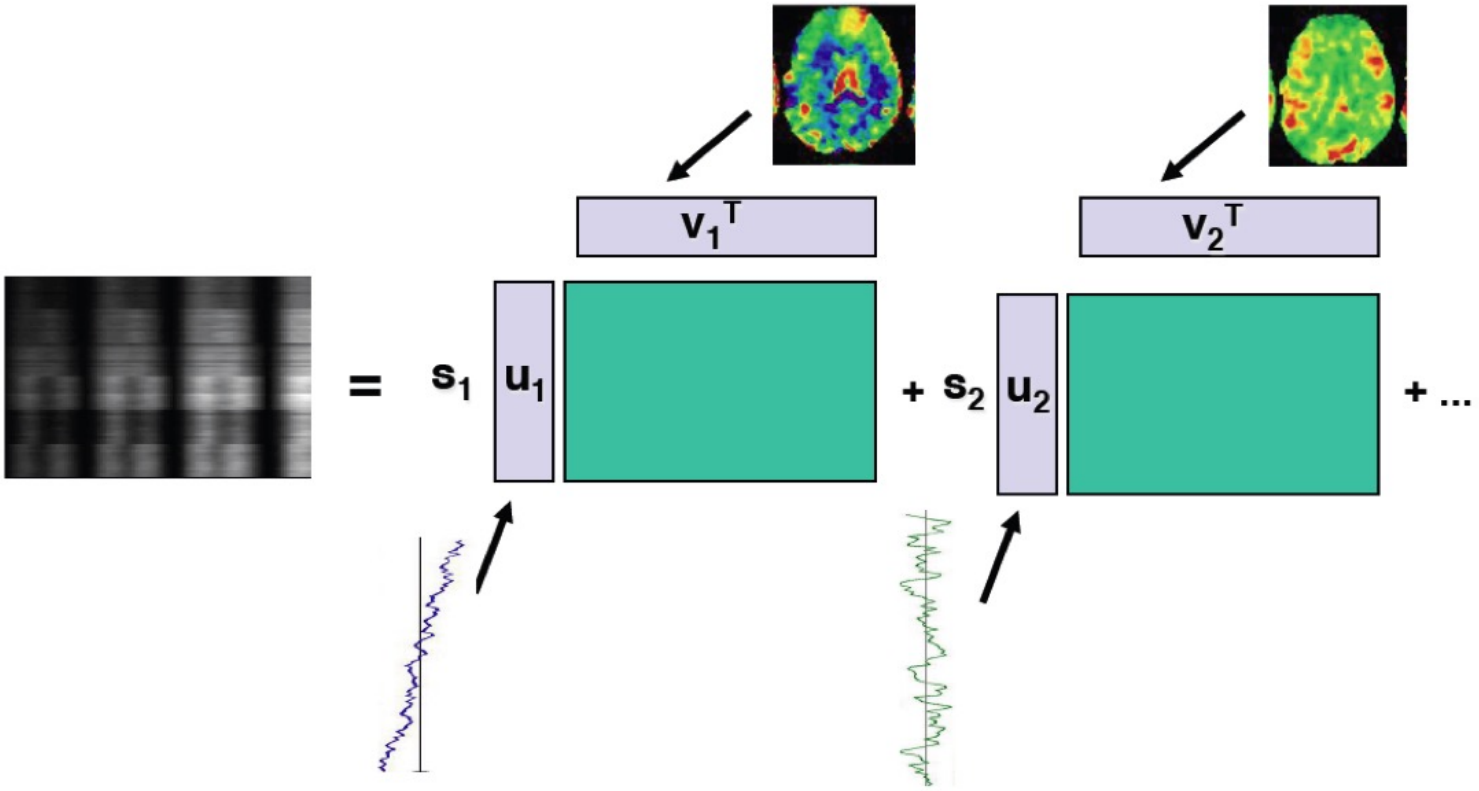


$$\mathbf{X} = \mathbf{USV}^T$$

$$\mathbf{X} = s_1\mathbf{u}_1\mathbf{v}_1^T + s_2\mathbf{u}_2\mathbf{v}_2^T + \ldots + s_N\mathbf{u}_N\mathbf{v}_N^T$$

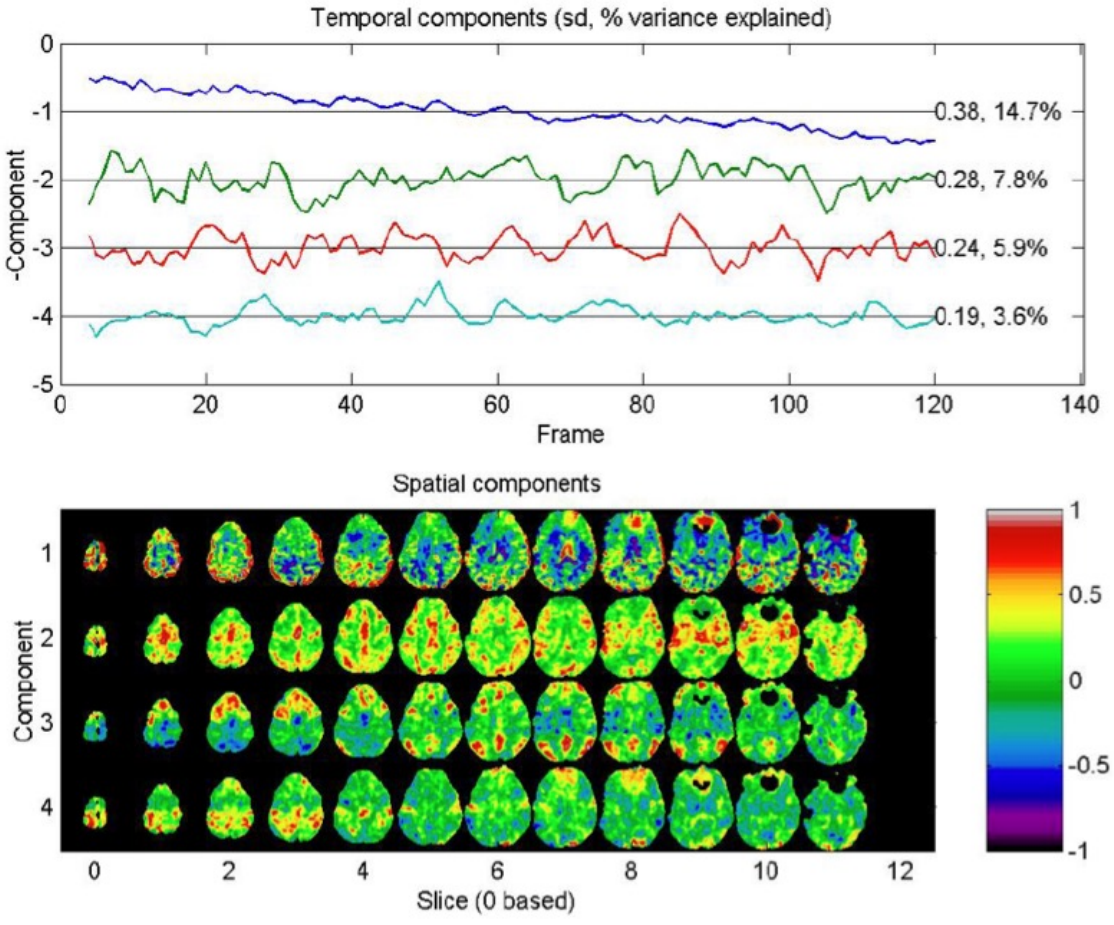
Lindquist (2013)

SVD for Imaging over Time



Lindquist (2013)

Example: Spatial-temporal components



Lindquist (2013)

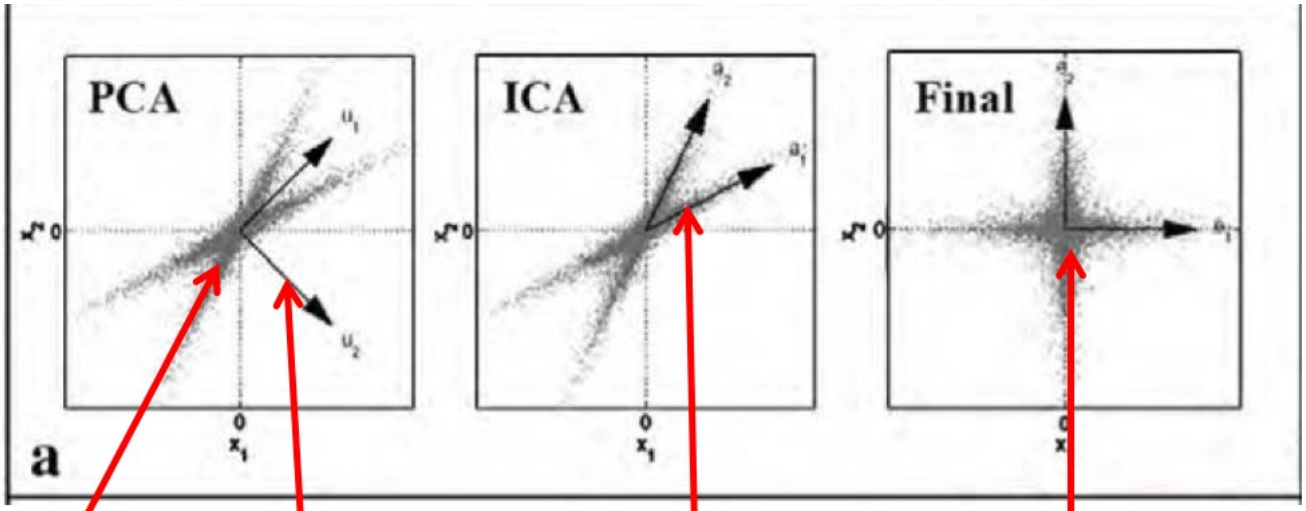
Independent Component Analysis (ICA)

$$\boldsymbol{X} = \boldsymbol{A}\boldsymbol{S}$$

- Observed:
  - $\boldsymbol{X}$, random $M \times T$ matrix, T time points.
- Unobserved:
  - $\boldsymbol{S}$: random $M \times T$ matrix where each row is mutually independent.
  - $\boldsymbol{A}$: non-random full-rank $M \times M$ mixing matrix.
- Estimate $\boldsymbol{W} = \boldsymbol{A}^{-1}$
- Estimate the independent source $\boldsymbol{S}$ by

$$\hat{\boldsymbol{S}} = \hat{\boldsymbol{W}}\boldsymbol{X}.$$

PCA vs ICA



Observed mixed outputs
a1*s1 and a2*s2

PCA, maximize variance
u1 and u2, uncorrelated

ICA, higher order statistics
a1 and a2, independent

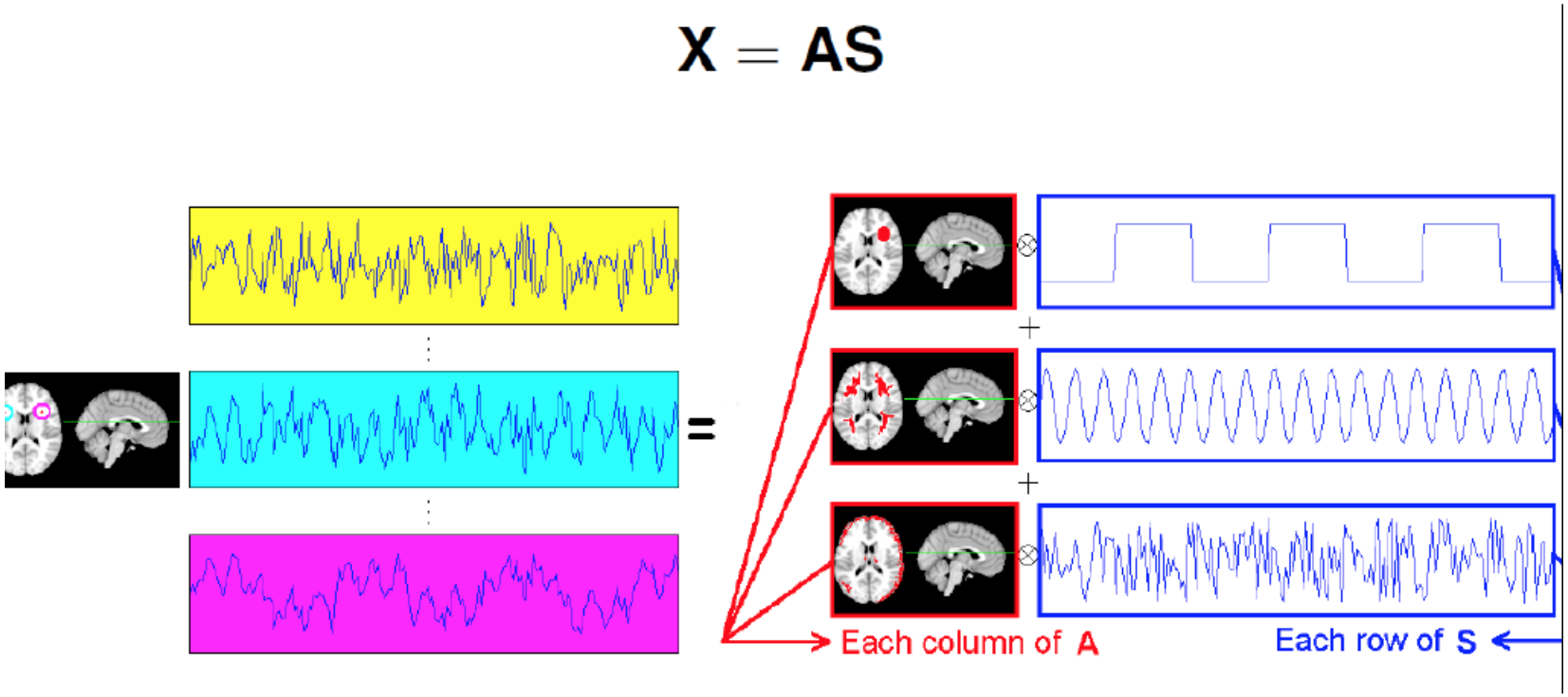
Hidden independent sources
s1 and s2

Calhoun et al. (2009)

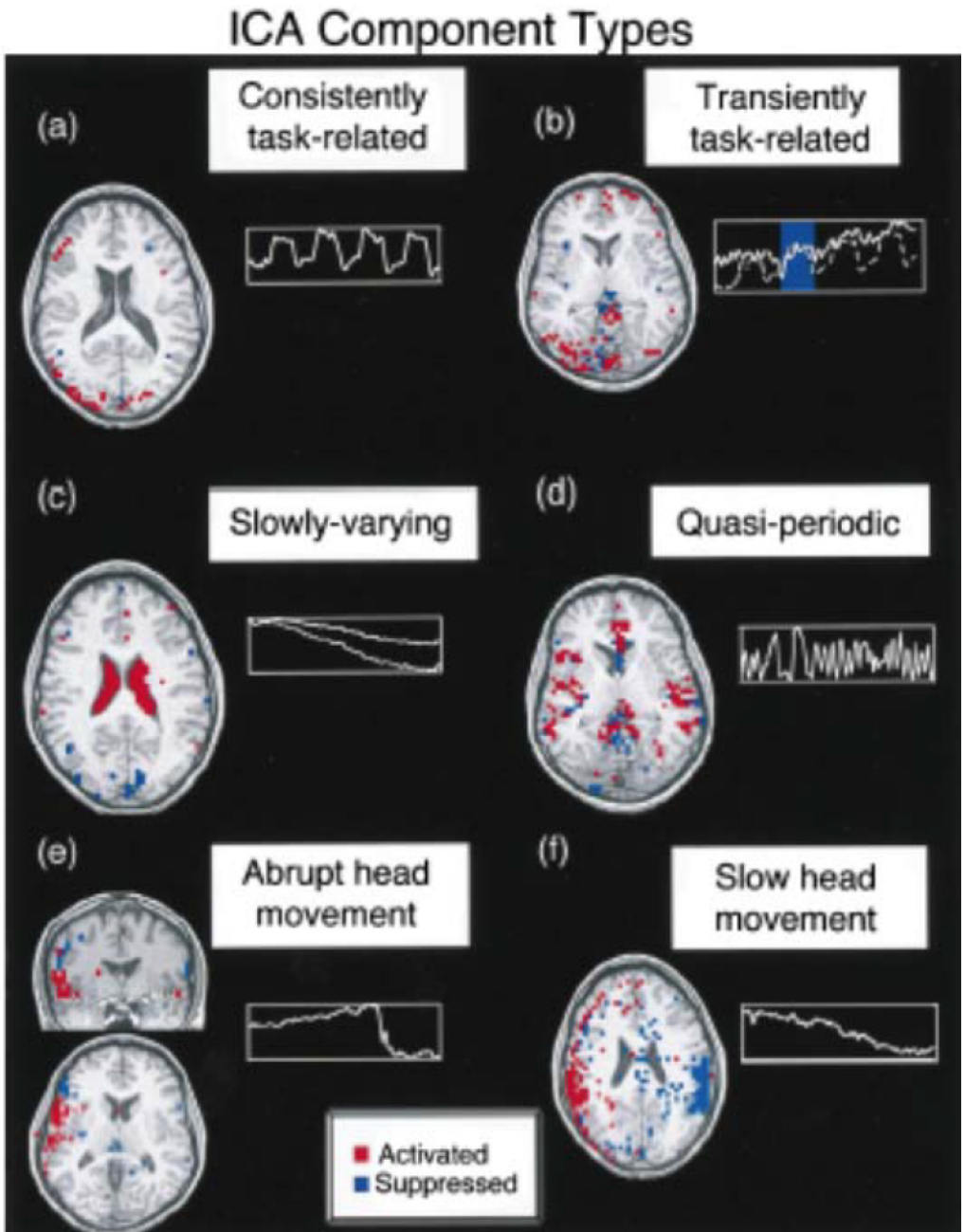
ICA Preprocessing

- In many cases, the number of rows (voxels numbers) is bigger than the number of sources (M).
- $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}' \approx \tilde{\boldsymbol{U}}_{V\times M}\tilde{\boldsymbol{D}}_{M\times M}\tilde{\boldsymbol{V}}'_{M\times T}$.
- Apply ICA to $\tilde{\boldsymbol{V}}_{M\times T} = \tilde{\boldsymbol{W}}^{-1}\boldsymbol{S}$.
- The mixing matrix $\boldsymbol{A} = \tilde{\boldsymbol{U}}\tilde{\boldsymbol{D}}\tilde{\boldsymbol{W}}^{-1}$.
- Free Matlab software: `http://research.ics.aalto.fi/ica/fastica/`

ICA for fMRI Data (Lee et al., 2012)

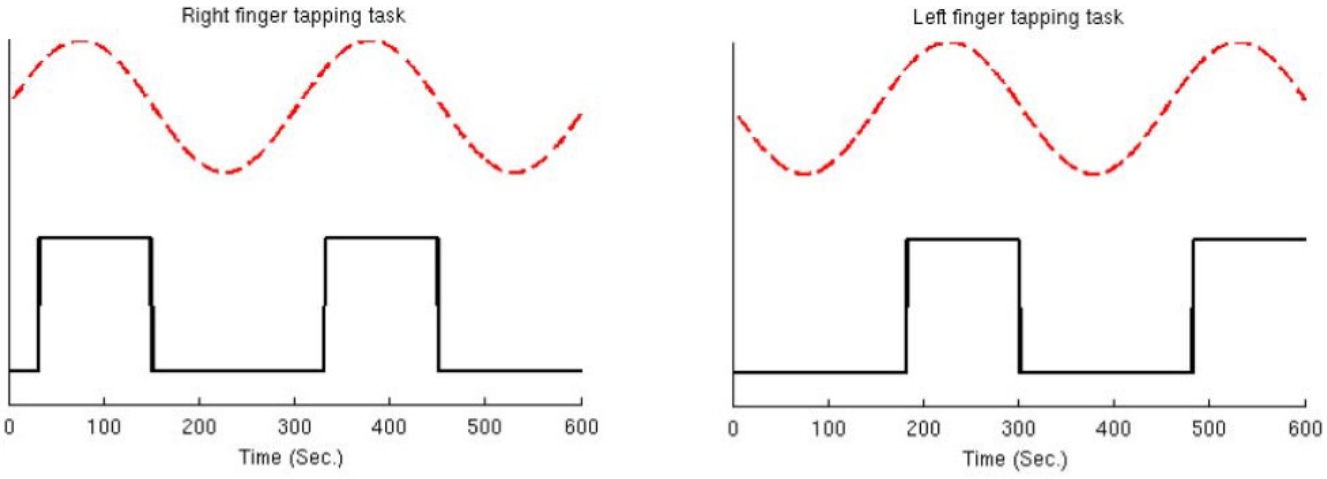


ICA Component Types (McKeown et al., 1998)



Example: Left and Right Finger Tapping (Lee et al., 2012)

(a) task paradigm

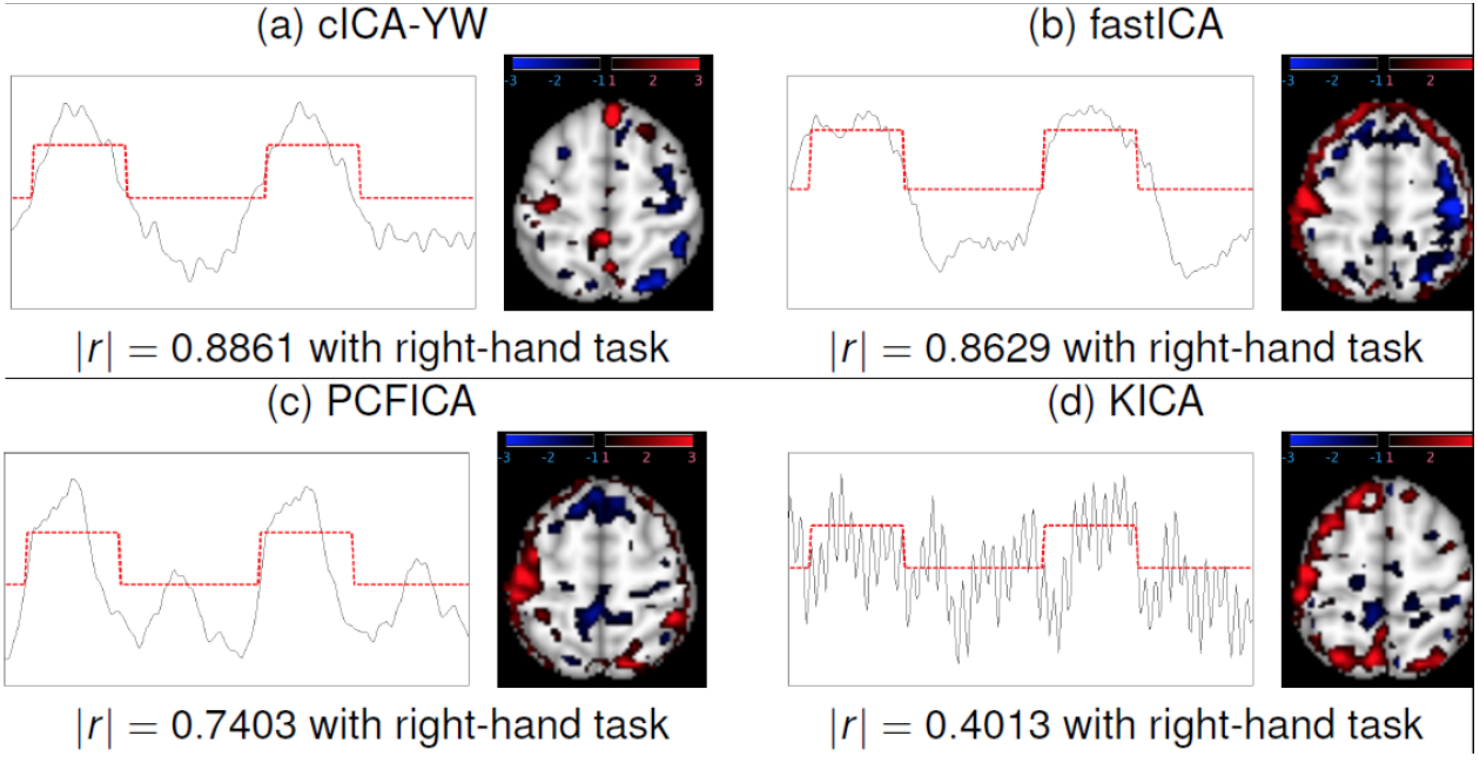


(b) Task Functions of right and left hand

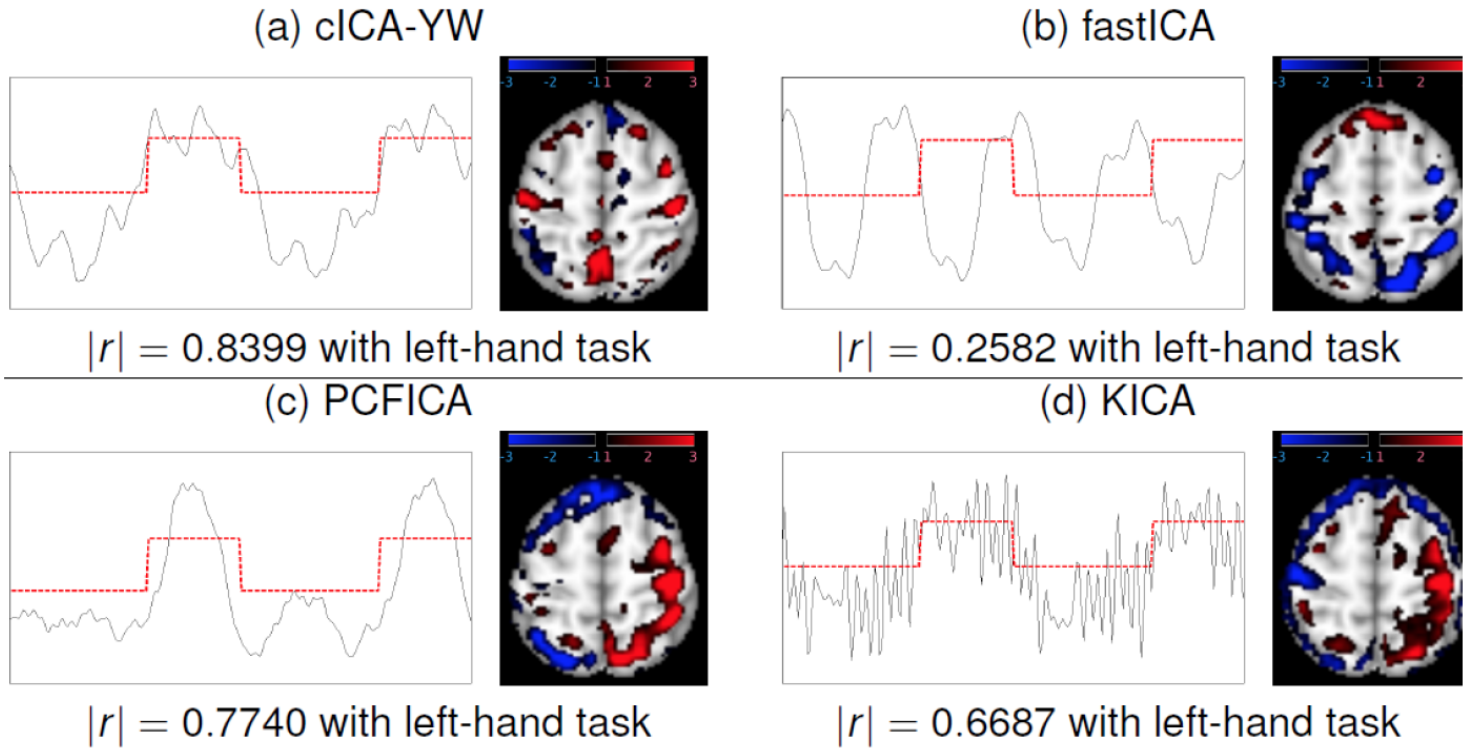
Lee et al. (2012)

Temporal independent components (ICs) and corresponding spatial maps, (Lee et al., 2012)

- Activated voxels with $z < -1$ were colored according to a blue-black color gradient
- Activated voxels with $z > 1$ were colored according to a red-black color gradient.
- A darker color represents less activation and a brighter color represents higher activation.
- The areas with $z < -3$ or $z > 3$ are colored blue and red, respectively.



Temporal independent components (ICs) and corresponding spatial maps (Lee et al., 2012)



(a) cICA-YW: Yule-Walker ICA (Brockwell and Davis, 1991) (b) fastICA: Hyvainen, Karhunen, and Oja (2001) (c) PCFICA: Pre-whitening for characteristic function based ICA (Chen and Bickel 2005) (d) KICA: kernel ICA (Bach and Jordan 2003). The comparison suggests that cICA-YW can recover the task-related signals of interest more accurately, and detect the regions activated by the tasks more sensitively.

A Regularization Framework to Modify PCA

- Low-rank approximation:

$$\min_{\{\boldsymbol{u},\boldsymbol{v}\}} \|\boldsymbol{X} - \boldsymbol{u}\boldsymbol{v}'\|_F^2$$

- $\boldsymbol{u}$: principal components (scores); $\boldsymbol{v}$: principal (loading) directions
- Minimize a penalized version: principal components

$$\|\boldsymbol{X} - \boldsymbol{u}\boldsymbol{v}'\|_F^2 + P_{\lambda_1,\lambda_2}(\boldsymbol{u}, \boldsymbol{v})$$

- Allow two regularization parameters
- Sparse-induced penalty or smoothness-induced penalty or hybrid

Sparse Regularization (Lee et al. 2010)

- Sparseness $\Leftrightarrow$ variable selection
- E.g., find $(\boldsymbol{u}, \boldsymbol{v})$ to minimize

$$\|\boldsymbol{X} - \boldsymbol{u}\boldsymbol{v}'\|_F^2 + \lambda_1 P_1(\boldsymbol{u}) + \lambda_2 P_2(\boldsymbol{v})$$

  - $P_1(\cdot)$ and $P_1(\cdot)$: adaptive Lasso (Zou, 2006)
  - $\lambda_2 = 0$: sparse PCA (Shen and Huang, 2008)
- Efficient computing:
  - Alternating adaptive LASSO regression
  - Coordinate-descent algorithm

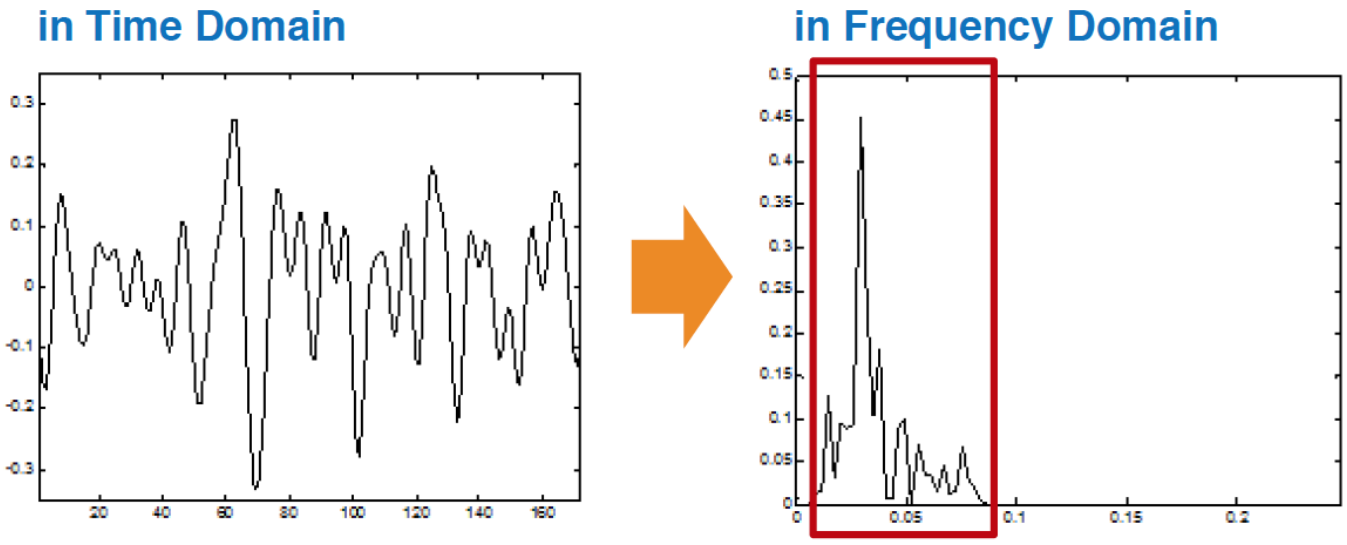
A Sparse Reduced Rank Group-wise Functional Connectivity Analysis(Ahn et al. 2013)

- Motivating data: The ADHD-200 Sample
- Global competition in 2011 to develop a predictive tool for ADHD diagnosis based on fMRI of brain
- Data were collected from 8 institutions around the world
- There are 4 groups of subjects:
  - Typically Developing Children (488)
  - ADHD Combined Type (158)
  - ADHD Hyperactive/Impulsive Type (11)
  - ADHD Inattentive Type (110)
- Goal:
  - Find group difference using time courses data
  - Find brain regions and time frequencies where ADHD patients show different signals, compared with controls

ADHD-200 Data

- Resting-state fMRI: Participants were asked simply to remain still, close their eyes, think of nothing systematically and not fall asleep
- A black screen was presented to them
- Use AAL atlas and NYU sample
- NYU data:
  - # of time points is the same for all subjects: 172
  - # of subjects per group is large enough (Total: 215)
  - Controls: 98
  - ADHD combined type: 72
  - Remove ADHD Hyperactive/Impulsive type: 2 (too small)
  - ADHD Inattentive type: 43
  - Total scan time =6mins

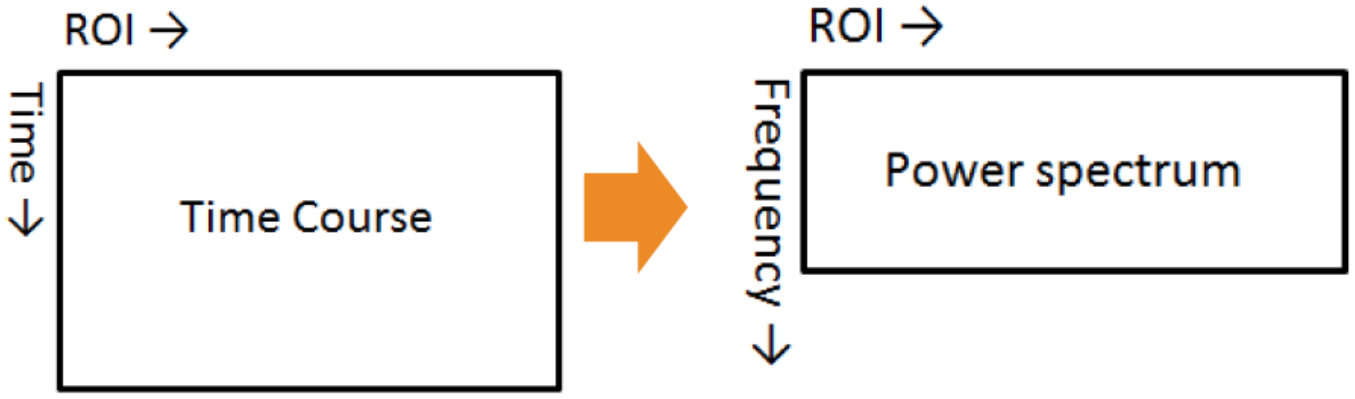
Data Structure



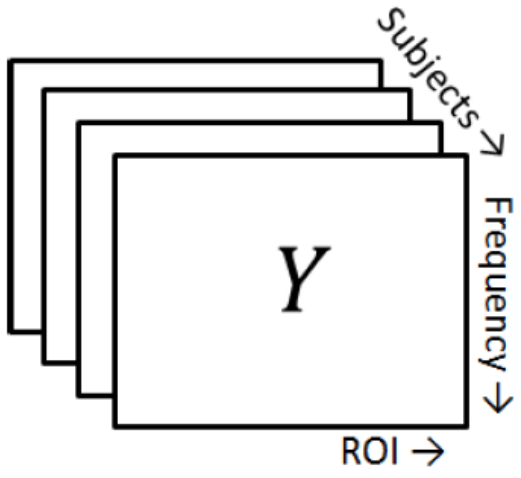
- Most of functional imaging data show significant fluctuations at certain range of frequencies
- In resting state fMRI data, there exist significant fluctuations at low frequencies such as below 0.1Hz
- Change the time-series data into the frequency domain data

Data Structure

in Matrix Form



Now we have a dataset :

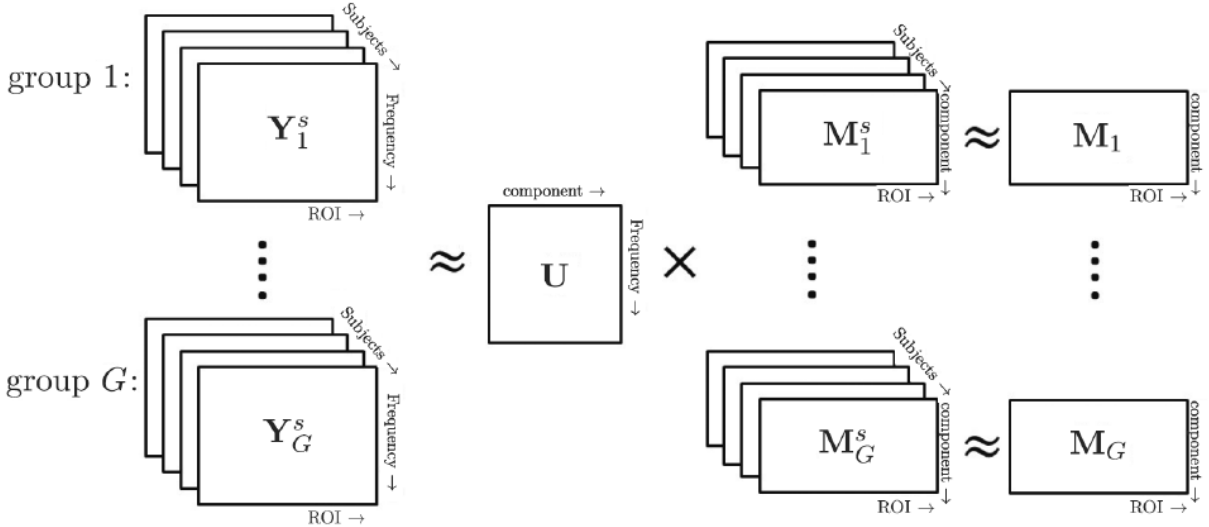


Model Structure

- For subject s of group g, consider a multi-group low-rank spatial-temporal model:

$$\boldsymbol{Y}_g^s = \boldsymbol{U}\boldsymbol{M}_g^s + \boldsymbol{E}_g^s$$

  - $\boldsymbol{U}$: frequency factor matrix which is common across groups
  - $\boldsymbol{M}_g^s$ subject-specific spatial factor matrix
  - $\boldsymbol{E}_g^s$: the subject-specific error matrix



Model Structure

- A group-specific spatial factor model

$$\boldsymbol{M}_g^s = \boldsymbol{M}_g + \boldsymbol{F}_g^s$$

  - $\boldsymbol{M}_g$: spatial factor matrix specific to the gth group
  - $\boldsymbol{F}_g^s$ is error matrix following a normal distribution with mean 0, and independent variance matrix

- If other characteristics $\boldsymbol{X}$ such as gender or age are available, consider a regression model:

$$\boldsymbol{M}_g^s = \boldsymbol{B}\boldsymbol{X} + \boldsymbol{F}_g^s$$

Estimation

- Minimize the residual sum of squares:

$$\sum_{g=1}^{G}\sum_{s=1}^{S_g} \| Y_g^s - \boldsymbol{U}\boldsymbol{M}_g^s \|_F^2, \text{ s.t. } \boldsymbol{U}'\boldsymbol{U} = I$$

- With initial $\hat{\boldsymbol{U}}$ and $\hat{\boldsymbol{M}}_g^s$, impose sparse structure on $\boldsymbol{U}$ by minimizing

$$\sum_{g=1}^{G}\sum_{s=1}^{S_g} \| Y_g^s - \boldsymbol{U}\boldsymbol{M}_g^s \|_F^2 + \lambda \| U \|_1$$

  which leads to the identification of important frequencies (with high power spectra)
- Estimate each component of $\boldsymbol{U}$ and $\boldsymbol{M}_g^s$ sequentially

Testing Group Difference

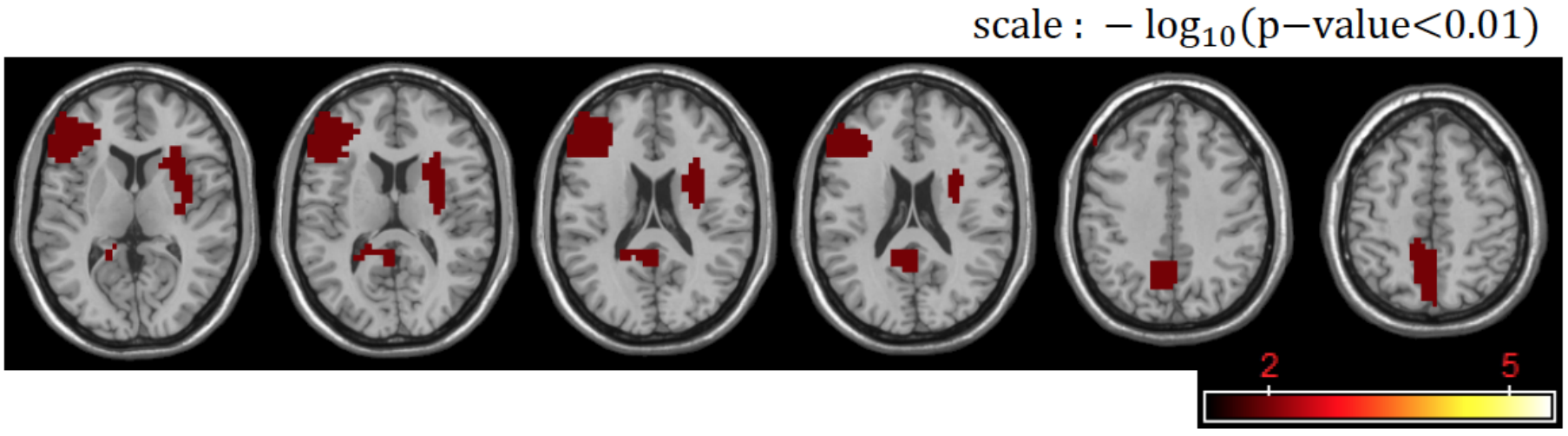
- Based on the model $\boldsymbol{M}_g^s = \boldsymbol{M}_g + \boldsymbol{F}_g^s$, for every ith component and jth region, $i = 1, \ldots, q$, $j = 1, \ldots, R$. T is the number of distinct frequencies, R is the number of regions, $q = \min\{T, R\}$
- Test at least one group is different from the others

$$H_0 : \boldsymbol{M}_1(i,j) = \ldots = \boldsymbol{M}_G(i,j)$$

- Test group difference for each pair of groups:

$$H_0 : \boldsymbol{M}_{g_1}(i,j) = \boldsymbol{M}_{g_2}(i,j)$$

Testing differences among 3 groups

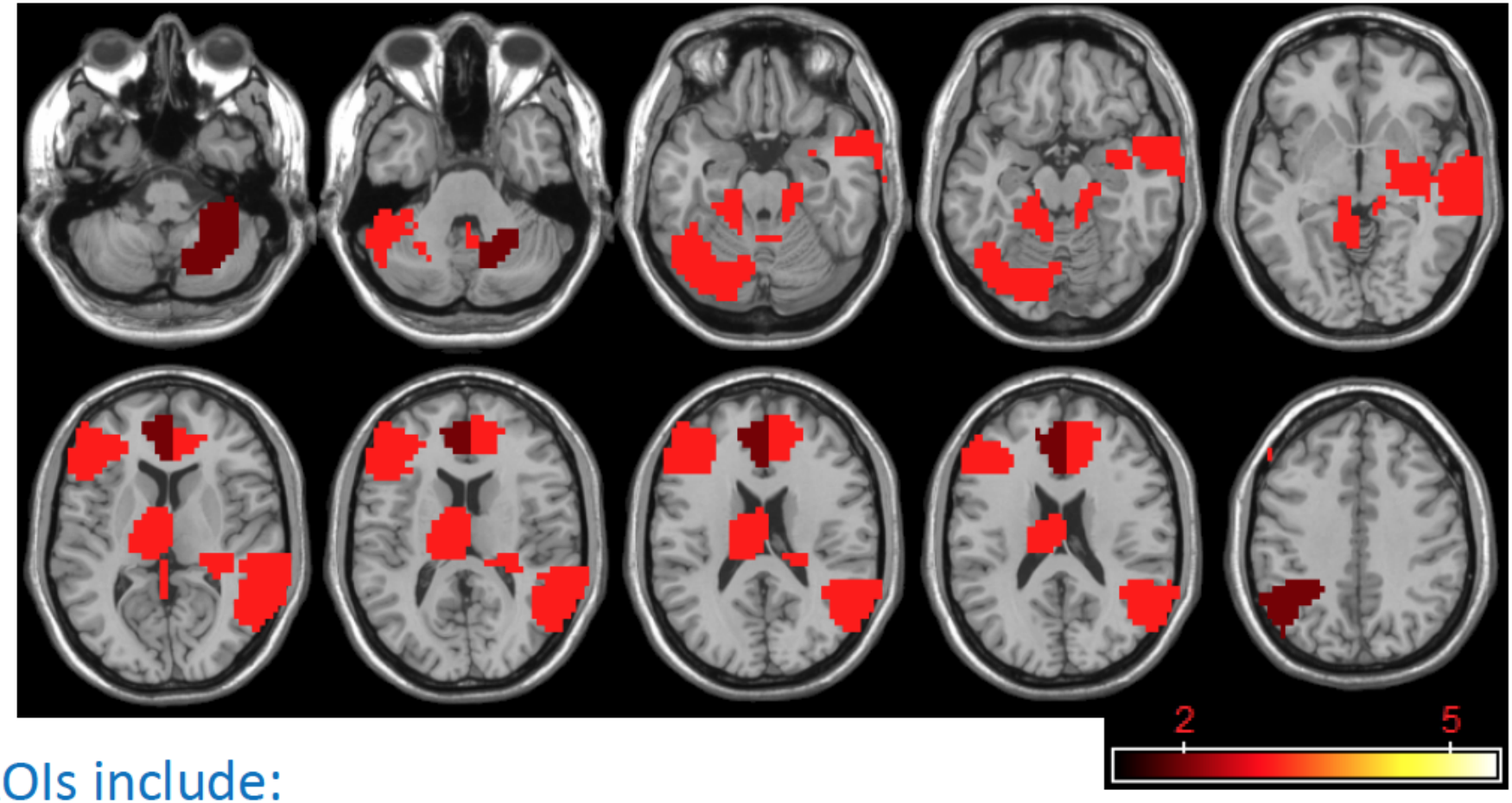


- 13th ROI : left triangular part of inferior frontal gyrus
  language

- 74th ROI : right putamen
  movement regulation and some forms of learning

Controls vs ADHD combined type

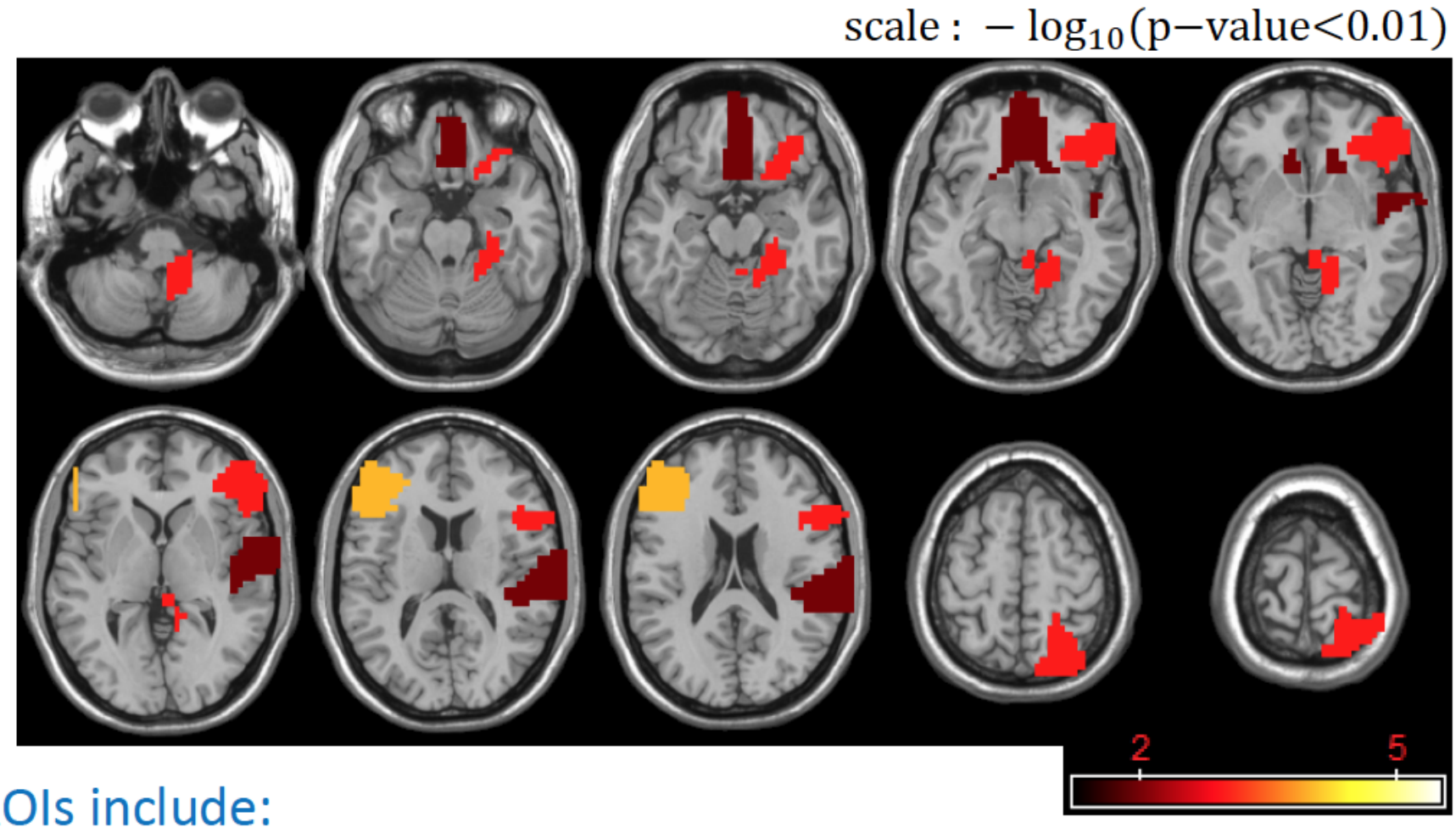
scale : $-\log_{10}(\text{p-value}<0.01)$



ROIs include:

Inferior Frontal gyrus, Anterior Cingulate, Hippocampus, Angular gyrus, Thalamus, Middle Temporal gyrus , and some regions in Cerebellum

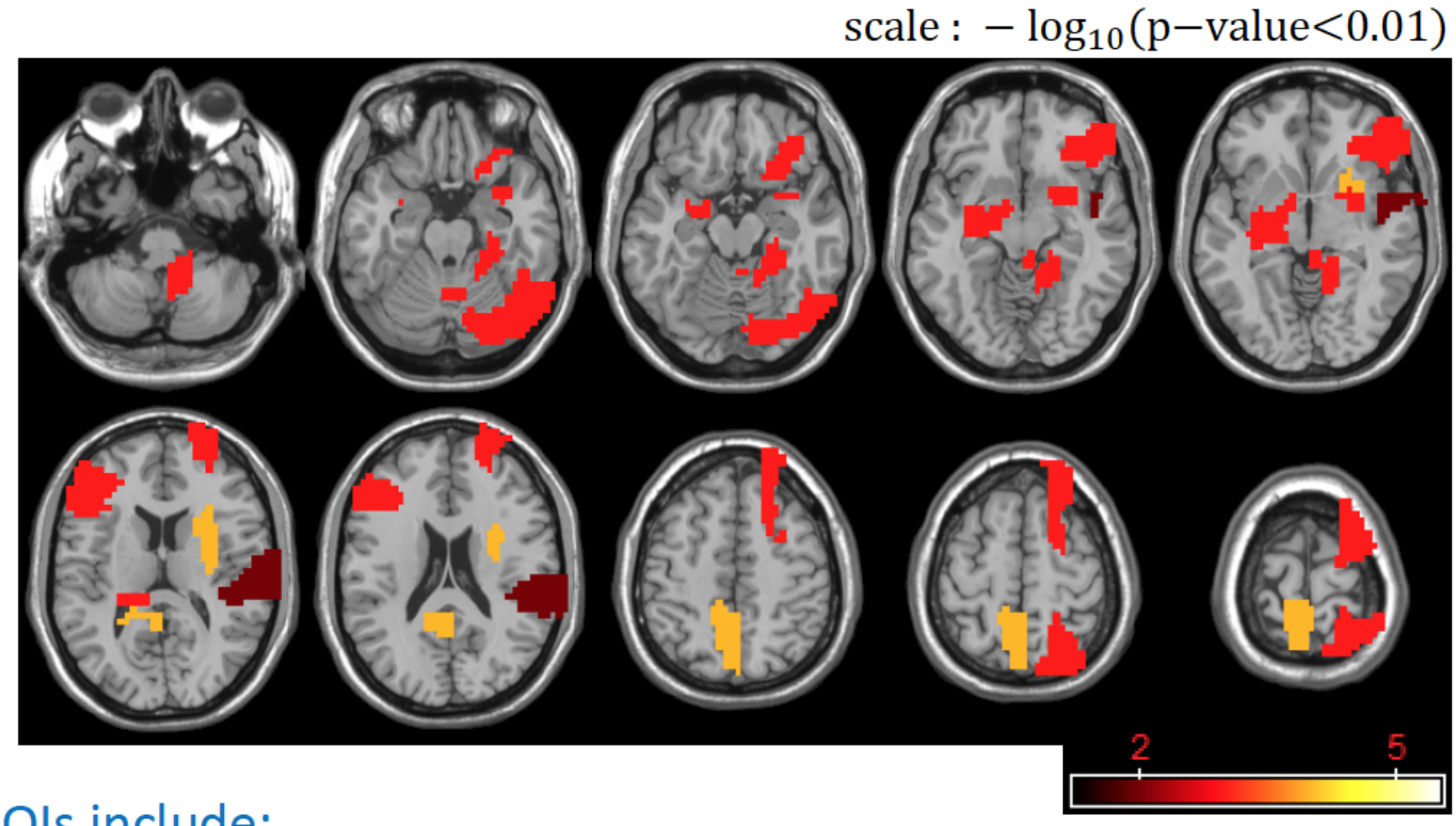
Controls vs ADHD inattentive type



ROIs include:

Inferior Frontal gyrus, Rectus, Superior Parietal gyrus, Superior Temporal gyrus, and some regions in Cerebellum

ADHD combined vs inattentive types



ROIs include:

Inferior Frontal gyrus, Superior Frontal gyrus, Hippocampus, Amygdala, Superior Parietal gyrus, Precuneus, Putamen, Superior Temporal gyrus, and some regions in Cerebellum

Significant ROIs

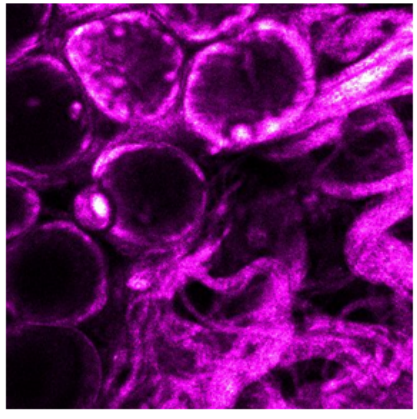
- The prefrontal cortex is an important brain region in ADHD studies
- ADHD patients show different activation pattern in the temporal lobe associated with language and verbal memory
- The cerebellum is responsible for motor control and cognitive functions
- The parietal lobe is related to attention, memory, and cognitive process
- The insula plays a role in consciousness related to emotions as well as perception, motor control, and self-awareness
- The cingulate gyrus is mainly associated with cognitive process that is linked to the signs of ADHD

Tensor Structure and Notation

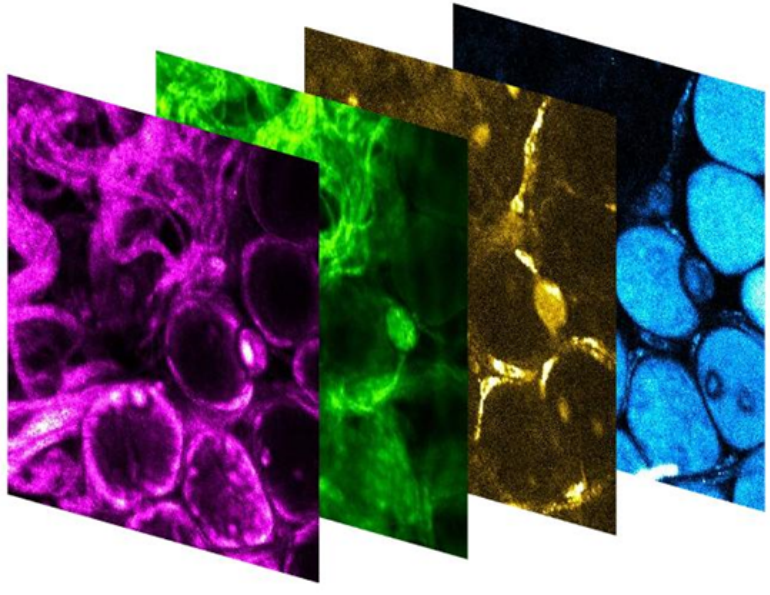
- Tensor is a multi-dimensional array: A Dth-order (D-way) tensor

$$\boldsymbol{\mathcal{X}} \in \mathbb{R}^{p_1 \times p_2 \times \ldots \times p_D}, \quad \boldsymbol{\mathcal{X}} = (x_{i_1 i_2 \ldots i_D})$$

- p_d is the marginal dimension for the dth mode ($d = 1, \ldots, D$)



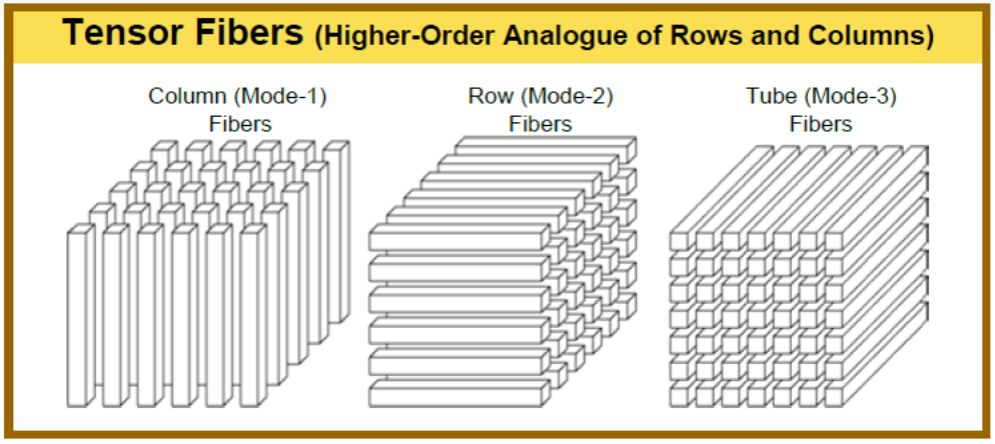
2-D tensor (matrix) $\in R^{375 \times 375}$



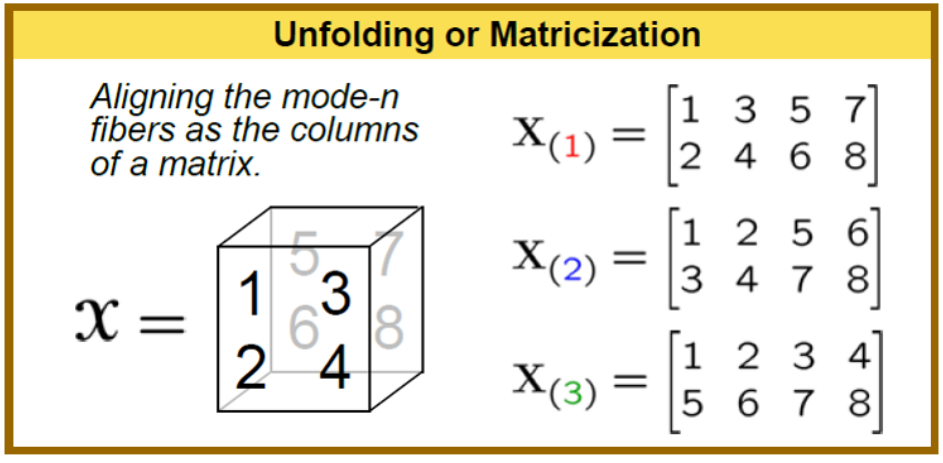
3-D tensor $\in R^{375 \times 375 \times 4}$

Tensor Structure and Notation (Kolda et al., 2009)

- **Tensor fiber**: a vector by fixing every index but one in a tensor



- **Mode-d matricization**: align all mode-d fibers as columns of a matrix



Tensor Operations

- **Inner product** $\langle \cdot, \cdot \rangle$ of two tensors with the same dimension

$$\langle \boldsymbol{\mathcal{A}}, \boldsymbol{\mathcal{B}} \rangle = \langle \text{vec}(\boldsymbol{\mathcal{A}}), \text{vec}(\boldsymbol{\mathcal{B}}) \rangle = \sum_{i_1=1}^{p_1} \sum_{i_2=1}^{p_2} \cdots \sum_{i_D=1}^{p_D} a_{i_1 i_2 \ldots i_D} b_{i_1 i_2 \ldots i_D}$$

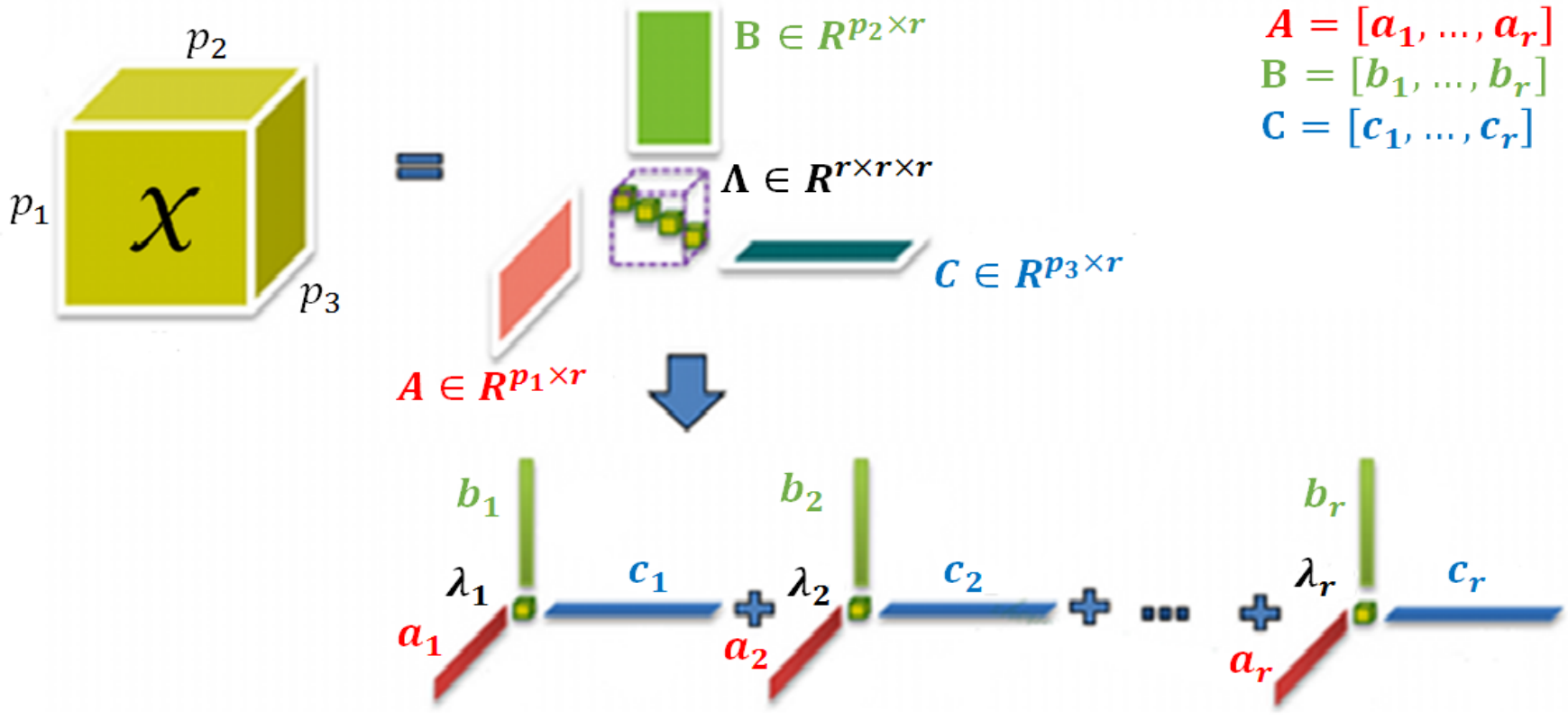
- **Outer product** "$\circ$" operating on D vectors $\boldsymbol{b}^{(d)} \in \mathbb{R}^{p_d}, d = 1 \ldots, D$

$$\boldsymbol{\mathcal{X}} = \boldsymbol{b}^{(1)} \circ \boldsymbol{b}^{(2)} \circ \ldots \circ \boldsymbol{b}^{(D)}, \quad \text{where} \quad x_{i_1, \ldots, i_D} = b^{(1)}_{i_1} b^{(2)}_{i_2} \ldots b^{(D)}_{i_D}$$

Low-rank CANDECOMP/PARAFAC (CP) Decomposition

- A rank-r CP decomposition for a 3-way tensor $\mathcal{X} \in R^{p_1 \times p_2 \times p_3}$:

$$\mathcal{X} = \sum_{k=1}^{r} \lambda_k \boldsymbol{a}_k \circ \boldsymbol{b}_k \circ \boldsymbol{c}_k = \left[\Lambda; \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}\right]$$



Low-rank CANDECOMP/PARAFAC (CP) Decomposition

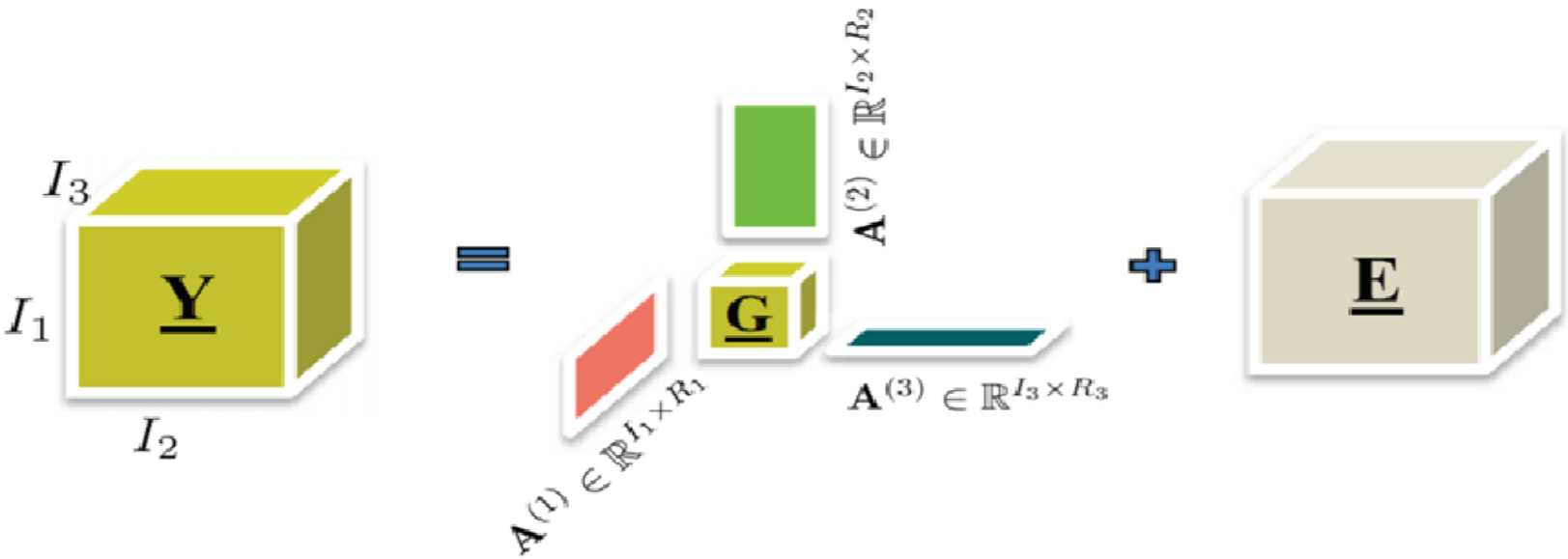
- $\boldsymbol{a}_k$'s, $\boldsymbol{b}_k$'s and $\boldsymbol{c}_k$'s are normalized loading vectors
- **Core tensor** $\boldsymbol{\Lambda}$ is diagonal of $\lambda_1, \ldots, \lambda_r$
- $\boldsymbol{A}$, $\boldsymbol{B}$ and $\boldsymbol{C}$ are loading matrices, usually not orthogonal
- **Advantage**: the rank is well defined, useful in dimension reduction

Tucker Decomposition

- The Tucker decomposition is defined as

$$\boldsymbol{X} \approx \sum_{i_1, i_2, \ldots, i_D} \lambda_{i_1 i_2 \ldots i_D} \boldsymbol{b}^{(1)}_{i_1} \circ \boldsymbol{b}^{(2)}_{i_2} \circ \ldots \circ \boldsymbol{b}^{(D)}_{i_D}$$

- $\| \boldsymbol{b}^{(d)}_{i_d} \|_F = 1$ and $\boldsymbol{b}^{(d)}_{i_d}$'s are orthogonal within each mode
- The decomposition is unique, but the rank is less clear
- Much more complicated than CP decomposition since the core tensor $\boldsymbol{G}$ is not diagonal



Tensor Regression for Neuroimaging Analysis (Zhou, Li and Zhu, 2013)

- Tensor predictor regression:
  - Aim at regression of a scalar response (continuous, binary, count) on an image predictor (MRI, PET, EEG) plus additional vector of covariates (age, gender) — by contrast the classical regression deals with vector-valued covariates
  - Key idea: introduce a low rank tensor decomposition to achieve substantial dimension reduction
- Regularized tensor predictor regression:
  - Couple tensor regression with a variety of penalty functions
  - Consider a particular form of regularized tensor regression to select regions in an image that are relevant to the clinical outcome — region selection in tensor regression corresponds to variable selection in the classical vector-valued regression
- Multi-modality tensor regression; multi-response tensor regression; longitudinal tensor GEE ...

Tensor regression model

- Formulation: regression with image (tensor) covariates
  - Y = univariate response, e.g., cognitive score, disease status
  - $\boldsymbol{Z} \in \mathbb{R}^q$ = conventional covariate vector containing age, gender, etc
  - $\boldsymbol{X} \in \mathbb{R}^{p_1 \times \ldots \times p_D}$ = D-dimensional array-valued predictor
    - EEG data: $D = 2$, where one dimension is time and the other is channel
    - MRI, $D = 3$, representing the 3D structure of an image
    - fMRI, $D = 4$, with an additional time dimension

- Challenges:
  - Ultrahigh-dimensionality: MRI, $256 \times 256 \times 256$, ~ 16 million voxels; fMRI, $\sim 100,000$ voxels, acquired between 200 and 1000 times;
  - Complex spatial and temporal structures
  - The sample size used to be small ($\sim 10 - 100$)
  - large-scale neuroimaging data: Alzheimer's Disease Neuroimaging Initiative (ADNI); Attention Deficit Hyperactivity Disorder Sample Initiative (ADHD); Autism Disorder (ABIDE)

Tensor regression model

- Model development:
  - Start with a vector-valued $\boldsymbol{X}$ and absorb $\boldsymbol{Z}$: the classical GLM,

$$p(y|\theta,\phi) = \exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right\}$$

    with mean $E(Y) = b'(\theta) = \mu$ and variance $(Y) = b''(\theta)a(\phi)$
  - Link function:

$$g(\mu) = \alpha + \boldsymbol{\beta}^{\mathsf{T}}\boldsymbol{X}$$

  - For an array $\boldsymbol{X}$ and a vector $\boldsymbol{Z}$, the link function becomes:

$$g(\mu) = \alpha + \boldsymbol{\gamma}^{\mathsf{T}}\boldsymbol{Z} + \langle \boldsymbol{B}, \boldsymbol{X}\rangle$$

    where the inner product $\langle \boldsymbol{B}, \boldsymbol{X}\rangle = \langle \mathrm{vec}\boldsymbol{B}, \mathrm{vec}\boldsymbol{X}\rangle$
  - This model is prohibitive, if no further constraint, as the number of parameters is $1 + p_0 + \prod_{d=1}^{D} p_d$

Tensor regression model

- Key idea: impose a low rank decomposition of $\boldsymbol{B}$
  - Definition: an array $\boldsymbol{B} \in \mathbb{R}^{p_1 \times \cdots \times p_D}$ admits a rank-R decomposition if

$$\boldsymbol{B} = \sum_{r=1}^{R} \beta_1^{(r)} \circ \cdots \circ \beta_D^{(r)} = [\![\boldsymbol{B}_1, \ldots, \boldsymbol{B}_D]\!]$$

    where $\beta_d^{(r)} \in \mathbb{R}^{p_d}, d = 1, \ldots, D, r = 1, \ldots, R$, are all column vectors, $\circ$ denotes an outer product, and $\boldsymbol{B}_d = [\beta_d^{(1)} \ldots \beta_d^{(R)}] \in \mathbb{R}^{p_d \times R}$

  - For $D = 2$, $R = 1$, $\boldsymbol{B} = [\![\boldsymbol{B}_1, \boldsymbol{B}_2]\!]$, $\boldsymbol{B}_1 = \beta_1$, $\boldsymbol{B}_2 = \beta_2$,

$$\boldsymbol{B} = \beta_1 \circ \beta_2$$

  - For $D = 2$, $R = 2$, $\boldsymbol{B} = [\![\boldsymbol{B}_1, \boldsymbol{B}_2]\!]$, $\boldsymbol{B}_1 = [\beta_1^{(1)}, \beta_1^{(2)}]$, $\boldsymbol{B}_2 = [\beta_2^{(1)}, \beta_2^{(2)}]$,

$$\boldsymbol{B} = \beta_1^{(1)} \circ \beta_2^{(1)} + \beta_1^{(2)} \circ \beta_2^{(2)}$$

Tensor regression model

- Rank-R tensor regression model:
  - Link function:

$$g(\mu) = \alpha + \boldsymbol{\gamma}^{\mathsf{T}}\boldsymbol{Z} + \langle \sum_{r=1}^{R} \boldsymbol{\beta}_1^{(r)} \circ \boldsymbol{\beta}_2^{(r)} \circ \cdots \circ \boldsymbol{\beta}_D^{(r)}, \boldsymbol{X} \rangle$$

  - For $D = 2$, $R = 1$, $\boldsymbol{B} = \boldsymbol{\beta}_1 \circ \boldsymbol{\beta}_2$

$$\begin{aligned} g(\mu) &= \alpha + \boldsymbol{\gamma}^{\mathsf{T}}\boldsymbol{Z} + (\boldsymbol{\beta}_2 \otimes \boldsymbol{\beta}_1)^{\mathsf{T}} \mathrm{vec}\boldsymbol{X} \\ &= \alpha + \boldsymbol{\gamma}^{\mathsf{T}}\boldsymbol{Z} + \boldsymbol{\beta}_1^{\mathsf{T}}\boldsymbol{X}\boldsymbol{\beta}_2 \end{aligned}$$

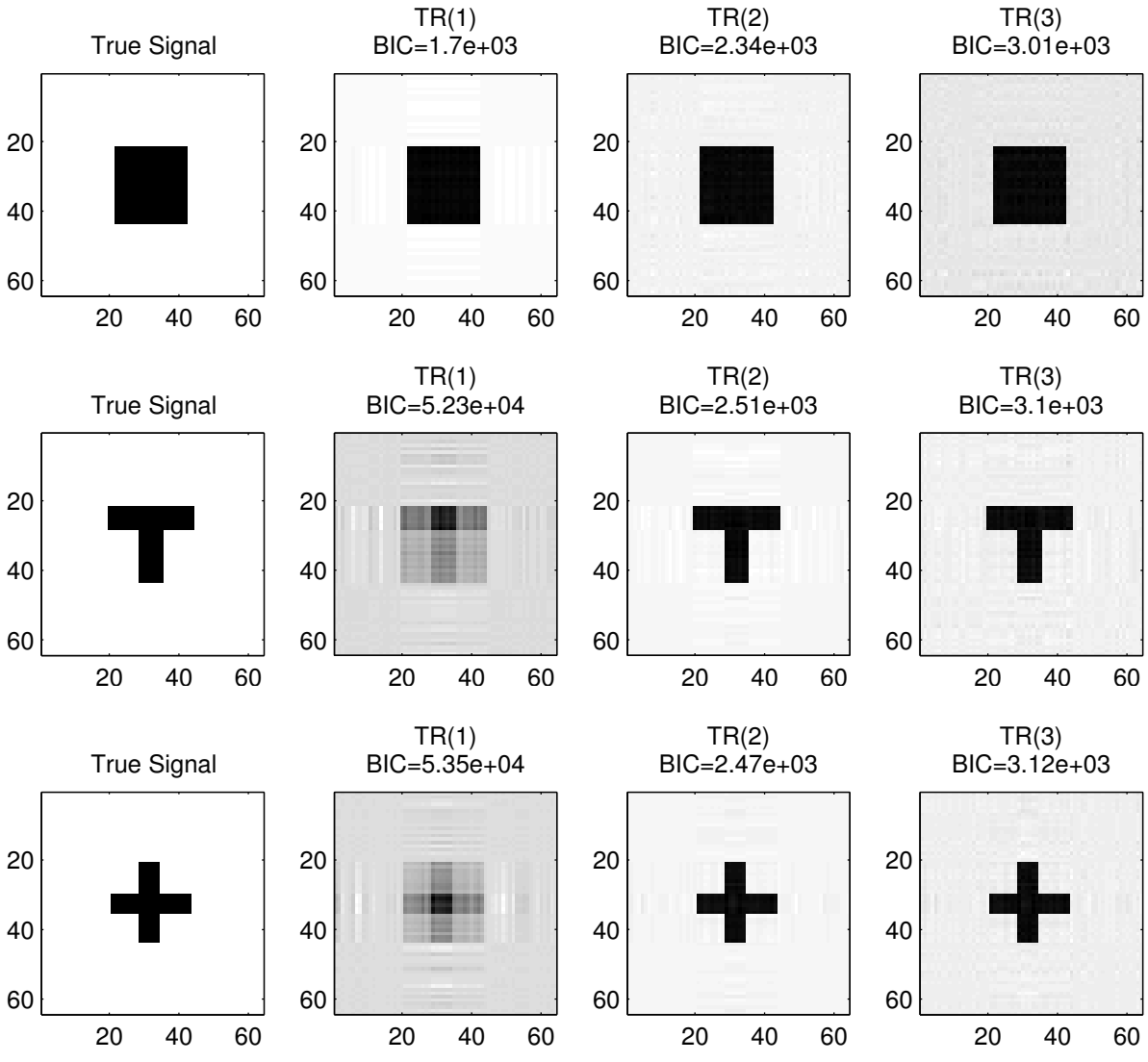
  - For $D = 2$, $R = 2$, $\boldsymbol{B} = \boldsymbol{\beta}_1^{(1)} \circ \boldsymbol{\beta}_2^{(1)} + \boldsymbol{\beta}_1^{(2)} \circ \boldsymbol{\beta}_2^{(2)}$

$$g(\mu) = \alpha + \boldsymbol{\gamma}^{\mathsf{T}}\boldsymbol{Z} + \langle \boldsymbol{\beta}_2^{(1)} \otimes \boldsymbol{\beta}_1^{(1)} + \boldsymbol{\beta}_2^{(2)} \otimes \boldsymbol{\beta}_1^{(2)}, \mathrm{vec}\boldsymbol{X} \rangle$$

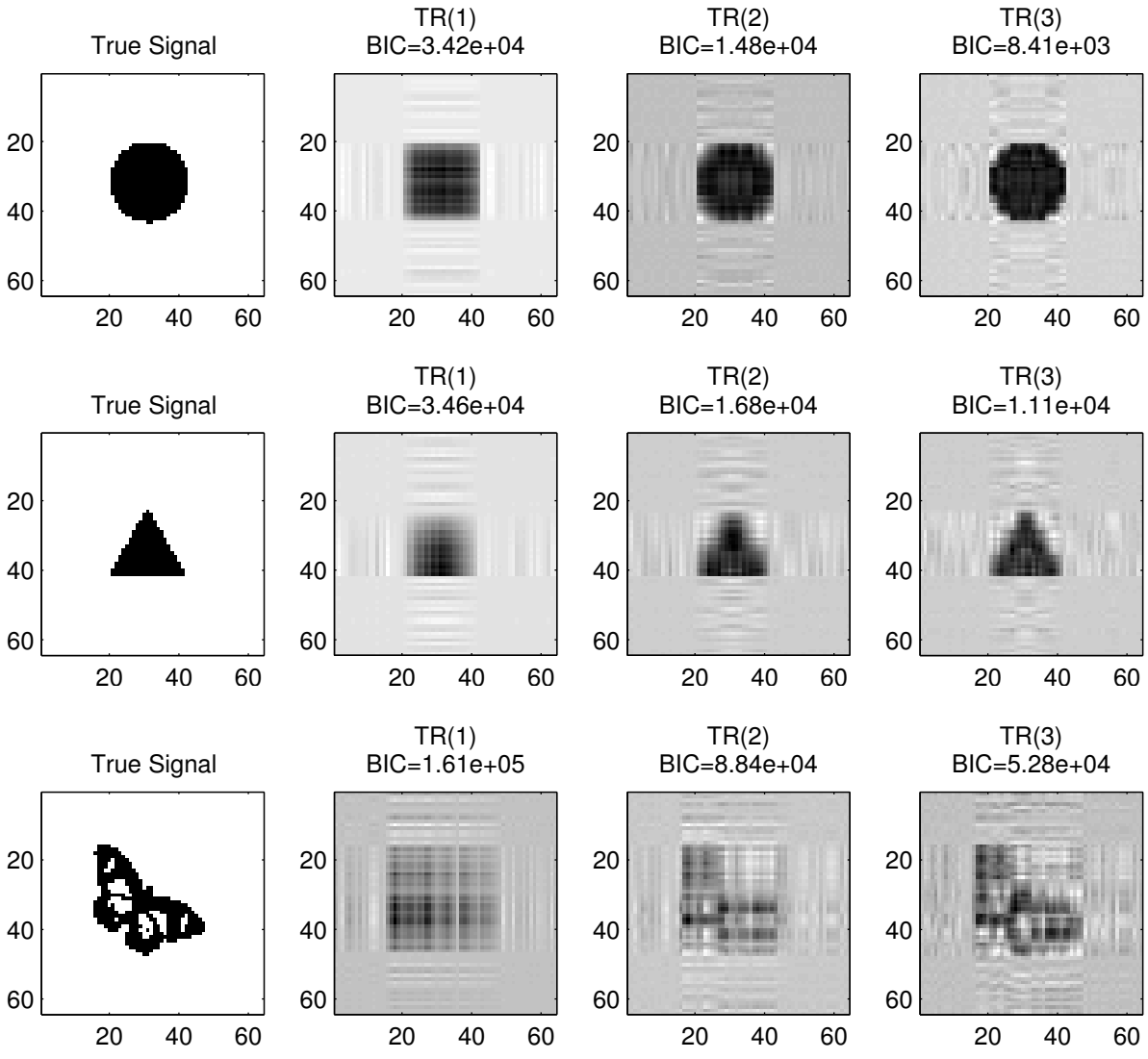
Tensor regression model

- Dimension reduction: the imposed structure substantially reduces the dimensionality from the order of $p_1 \times \ldots \times p_D$ to $R \times (p_1 + \ldots + p_D)$
- For an 256-dim MRI image, it reduces from $256 \times 256 \times 256 = 16,777,216$ to $256 + 256 + 256 = 768$-dimensional for a rank-1 model, and to 2,304 for a rank-3 model
- The model is identifiable up to scaling and permutation under minor conditions
- Can be viewed as a supervised version of the classical CP
- Data-driven low-rank approximation of the signal image

Signal recovery: $Y = \alpha + \gamma^{\mathsf{T}} \boldsymbol{Z} + \langle \boldsymbol{B}, \boldsymbol{X} \rangle + \varepsilon$



Signal recovery: $Y = \alpha + \gamma^{\mathsf{T}} \boldsymbol{Z} + \langle \boldsymbol{B}, \boldsymbol{X} \rangle + \varepsilon$



Estimation and asymptotics

- Maximum likelihood estimation:
  - Key observation: $g(\mu)$ is not linear in $(\boldsymbol{B}_1, \ldots, \boldsymbol{B}_D)$ jointly, but is linear in $\boldsymbol{B}_d$ individually
  - A block-relaxation algorithm: alternating update of $\boldsymbol{B}_d$
  - Simple to implement; numerically stable; guaranteed convergence

- Rank estimation (model selection): BIC $= -2\ell(\boldsymbol{\theta}) + \log(n)p_e$, p_e is the degrees of freedom of parameters.

- Asymptotics:
  - Consistency: the MLE is consistent, i.e., $\hat{\boldsymbol{B}}_n$ converges to $\boldsymbol{B}_0$ in probability
  - The MLE is consistently estimating the best rank-R approximation of $\boldsymbol{B}_{\text{true}}$ in the Kullback-Leibler distance
  - Asymptotic normality: $\sqrt{n}\left\{\text{vec}(\hat{\boldsymbol{B}}_{n1}, \ldots, \hat{\boldsymbol{B}}_{nD}) - \text{vec}(\boldsymbol{B}_{01}, \ldots, \boldsymbol{B}_{0D})\right\}$ converges in distribution to a normal with mean zero and covariance matrix $\boldsymbol{I}^{-1}(\boldsymbol{B}_{01}, \ldots, \boldsymbol{B}_{0D})$, where $\boldsymbol{I}$ is the information matrix

Regularization

- Why regularization:
  - $p > n$ is a rule rather than an exception in neuroimaging analysis; e.g., for a 256-dim image, rank-1 model ~ 768, and rank-3 model $\sim 2,304$, whereas the sample size n (~ 100)
  - In general, regularization is useful for stabilizing the estimates, improving the risk property, and incorporating prior knowledge

- Why new regularization:
  - In principle, regularization can be directly applied to the vectorized coefficient array $\boldsymbol{B}$ instead of vectorizing the array $\boldsymbol{X}$

- Region selection:
  - Identify brain regions / activity patterns relevant to clinical outcome
  - Corresponds to variable selection in the traditional vector-valued covariates regression

Hard thresholding

- Hard thresholding:
  - Sparsity principle: the true signal is sparse in terms of the ℓ_0 norm of the tensor parameter
  - The ℓ_1 norm is a convex relaxation of the ℓ_0 norm
  - The rank R is fixed, a priori

- Example: matrix covariate and lasso penalty

$$\min_{\boldsymbol{B}} \frac{1}{2}\sum_{i=1}^{n}(y_i - \alpha - \boldsymbol{\gamma}^{\mathsf{T}}\boldsymbol{z}_i - \langle \boldsymbol{B}, \boldsymbol{x}_i \rangle)^2 + \rho \left\| \mathrm{vec}\boldsymbol{B} \right\|_1$$

Hard thresholding

- More generally,

$$-\ell(\alpha, \gamma, \boldsymbol{B}_1, \ldots, \boldsymbol{B}_D) + \sum_{d=1}^{D} \sum_{r=1}^{R} \sum_{i=1}^{p_d} P_\lambda(|\beta_{di}^{(r)}|, \rho),$$

  - where P is a scalar penalty function, ρ is the penalty tuning parameter, and λ is an index for the penalty family.

- A range of penalty functions:
  - Power family: $P_\lambda(|\omega|, \rho) = \rho|\omega|^\lambda$, $\lambda \in (0, 2]$; lasso: $\lambda = 1$; ridge: $\lambda = 2$
  - Elastic net: $P_\lambda(|\omega|, \rho) = \rho[(\lambda - 1)\omega^2/2 + (2 - \lambda)|\omega|], \lambda \in [1, 2]$
  - SCAD: $\partial/\partial|\omega| P_\lambda(|\omega|, \rho) = \rho \left\{ 1_{\{|\omega| \leq \rho\}} + \frac{(\lambda\rho - |\omega|)_+}{(\lambda - 1)\rho 1_{\{|\omega| > \rho\}}} \right\}$, $\lambda > 2$

- Computation: in each updating step, fit a penalized GLM

ADHD revisited

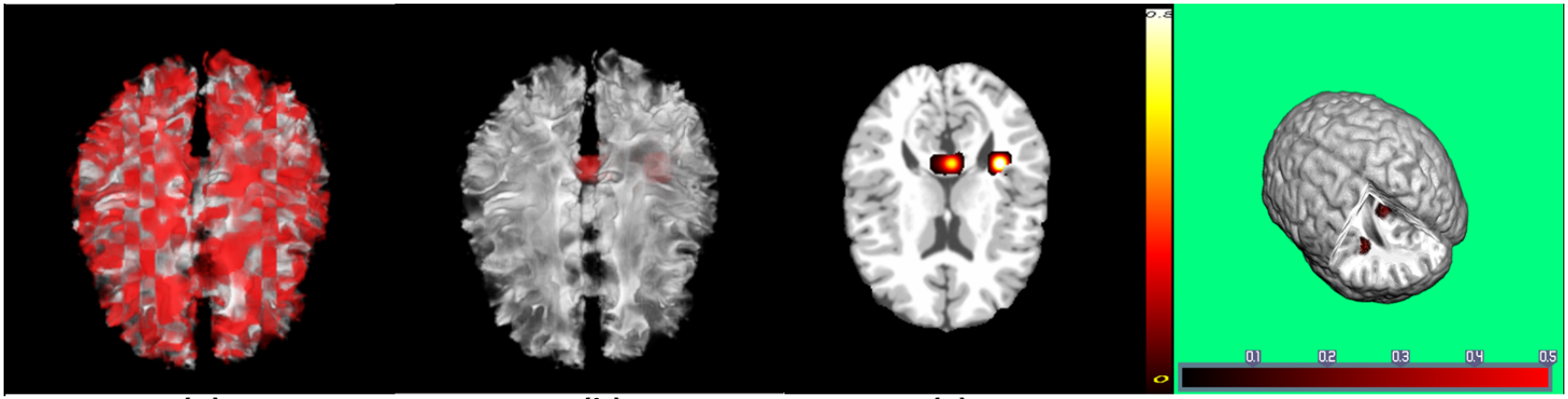
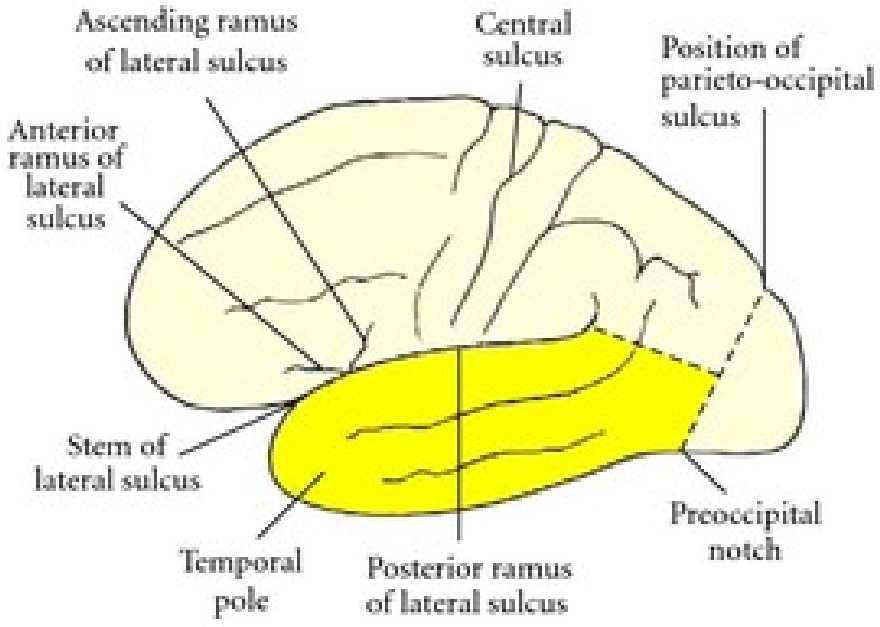
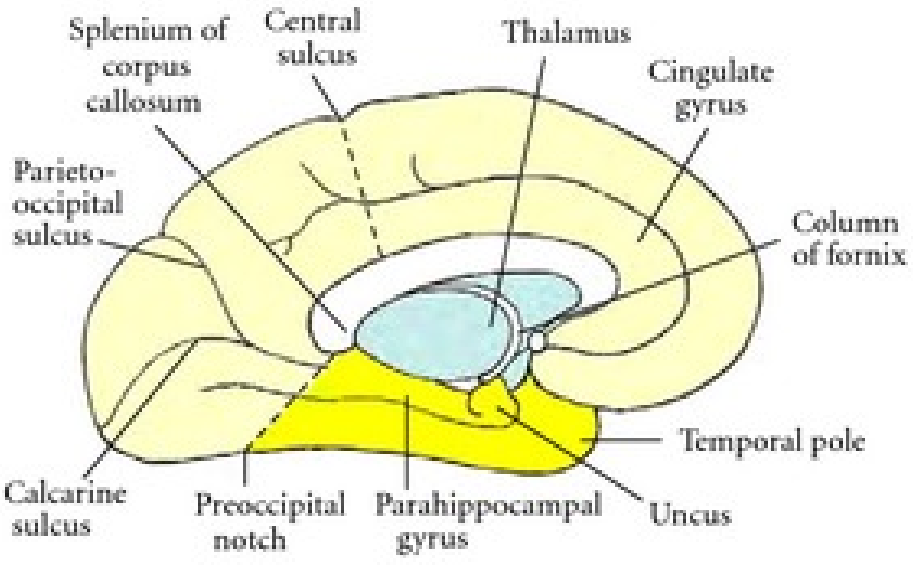


Figure: Panel (a) is the unpenalized estimate overlaid on a randomly selected subject; (b) is the regularized estimate; (c) is a selected slice of the regularized estimate overlaid on the template; and (d) is a 3D rendering of the regularized estimate.

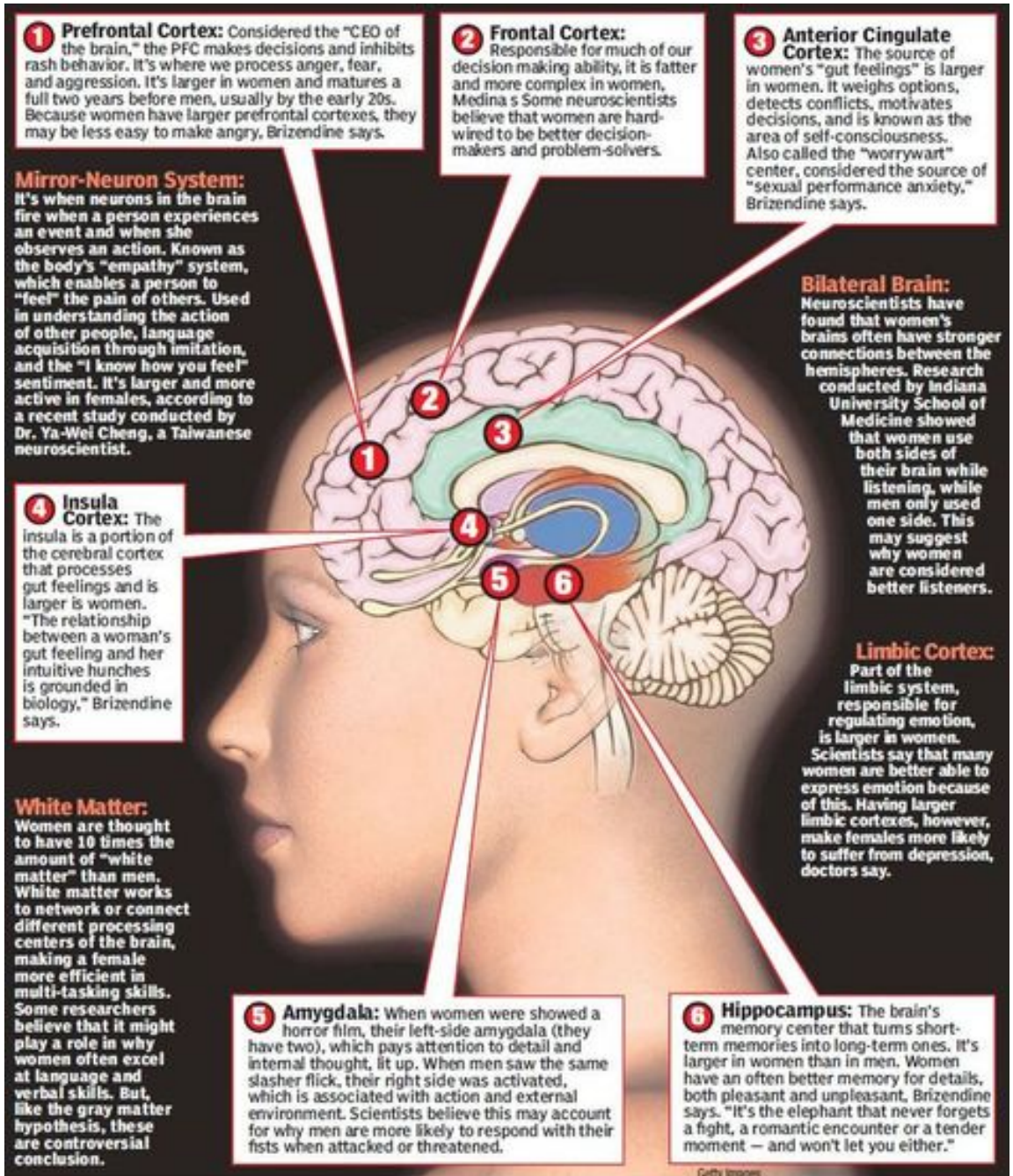
- Findings of two regions of interest: left temporal lobe white matter and the splenium in the corpus callosum

Brain ROI's and Functions





Brain ROI's and Functions



Longitudinal Imaging Analysis (Zhang et al., 2014)

- Scientific background:
  - Alzheimer's disease (AD) and normal aging
  - AD: characterized by progressive impairment of cognitive and memory functions; an irreversible neurodegenerative disorder; and the leading form of dementia in the elderly subjects
  - The number of affected people is rapidly increasing, and is projected to be 1 in 85 worldwide by the year 2050
  - Imperative to understand, diagnose, and treat AD
  - Equally important to understand normal aging

- Broad scientific questions of interest:
  - How the brain structures and functions correlate with cognitive behavior, all of which are expected to change progressively
  - Where in the brain to find the highly correlated changes
  - How to differentiate between AD and normal aging in terms of progressive decline in brain structures, functions, and cognition

Motivation

- Alzheimer's Disease Neuroimaging Initiative (ADNI):
  - 88 subjects with mild cognitive impairment (MCI), a prodromal stage
  - Anatomical magnetic resonance imaging (MRI): preprocessed
  - 5 different time points: baseline, 6-mo, 12-mo, 18-mo and 24-mo
  - Cognitive score: the Mini-Mental State Examination (MMSE), which examines orientation to time and place, immediate and delayed recall of three words, attention and calculation, language and visuo-constructional functions

Tensor generalized estimation equations

- Longitudinal anatomical MRI modeling:
  - Understand association between cognitive behavior and the structural brain atrophy
  - Predict future cognition given the MRI scan and other information
  - Identify brain regions that are most relevant to cognitive decline

- Existing literature:
  - Longitudinal images based classification (Misra et al., 2009; Davatzikos et al., 2009; McEvoy et al., 2011; Hinrichs et al., 2011), and cognitive score prediction (Zhang et al., 2012)
  - Regression of longitudinal images on covariates (Skup et al., 2012; Li et al., 2013)
  - Functional principal components to quantify longitudinal images (Shinohara et al., 2011)
  - Lack of effective solutions for longitudinal imaging analysis

Tensor generalized estimation equations (GEE)

- Tensor GEE
  - Longitudinal regression of a scalar response (e.g., cognitive score) on a tensor predictor (MRI image)
  - Integrate low rank tensor decomposition and generalized estimating equations (GEE):
    CP decomposition: preserves the spatial structure of the image
    GEE: accommodates longitudinal correlation of the data
  - Lasso/SCAD penalized tensor GEE for selection of brain regions that are highly relevant to the clinical outcome – variable selection in the tensor context
  - Develop an efficient and scalable computational algorithm
  - Establish the asymptotic properties of the tensor GEE solution: the tensor GEE estimator inherits the robustness feature of the classical GEE estimator $\Rightarrow$ estimate is consistent even if the working correlation structure is misspecified

Model

- Notations:
  - $i = 1, \ldots, n$ subject; $j = 1, \ldots, m$ time points
  - Y_{ij} = univariate response; e.g., MMSE cognitive score
  - $\boldsymbol{X}_{ij} \in \mathbb{R}^{p_1 \times \cdots \times p_D}$ = order-D tensor predictor; e.g., $D = 3$ for MRI
  - Some additional vector of covariates; omitted here for simplicity
- Model:
  - Y_{ij} follows an exponential family, with mean $\mu(\theta_{ij})$, variance $\sigma_{ij}^2 = \mu'(\theta_{ij})$, a canonical link function, and the systematic part $\theta_{ij} = \langle \boldsymbol{B}, \boldsymbol{X}_{ij} \rangle$
  - $\boldsymbol{B}$ is the coefficient tensor of the same size as $\boldsymbol{X}_{ij}$: effects of every array element of $\boldsymbol{X}_{ij}$ on $\boldsymbol{Y}_{ij}$
  - The GEE estimator of $\boldsymbol{B}$ is the solution of

$$\sum_{i=1}^{n} \frac{\partial \boldsymbol{\mu}_i(\boldsymbol{B})}{\partial \mathrm{vec}(\boldsymbol{B})} \boldsymbol{V}_i^{-1} \{ \boldsymbol{Y}_i - \boldsymbol{\mu}_i(\boldsymbol{B}) \} = \boldsymbol{0},$$

where $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{im})^{\mathsf{T}}$, $\boldsymbol{\mu}_i(\boldsymbol{B}) = [\mu_{i1}(\boldsymbol{B}), \ldots, \mu_{im}(\boldsymbol{B})]^{\mathsf{T}}$, and $\boldsymbol{V}_i = \mathrm{cov}(\boldsymbol{Y}_i)$ is the response covariance matrix of the i-th subject

Model Development

- There are totally $\prod_d p_d$ estimating equations to solve, since the derivative is with respect to the vector $\mathrm{vec}(\boldsymbol{B}) \in \mathbb{R}^{\prod_d p_d}$; e.g., for $32 \times 32 \times 32$ MRI means $32^3 = 32,768$ equations
- Impose a low rank decomposition of $\boldsymbol{B}$: rank-R CP decomposition

$$\boldsymbol{B} = \sum_{r=1}^{R} \boldsymbol{\beta}_1^{(r)} \circ \cdots \circ \boldsymbol{\beta}_D^{(r)} = [\![\boldsymbol{B}_1, \ldots, \boldsymbol{B}_D]\!]$$

where $\boldsymbol{\beta}_d^{(r)} \in \mathbb{R}^{p_d}, d = 1, \ldots, D, r = 1, \ldots, R$, are all column vectors, $\circ$ denotes an outer product, and $\boldsymbol{B}_d = [\boldsymbol{\beta}_d^{(1)} \ldots \boldsymbol{\beta}_d^{(R)}] \in \mathbb{R}^{p_d \times R}$

Tensor generalized estimation equations

- Define the tensor GEE estimator of $\boldsymbol{B}$ as the solution of

$$\sum_{i=1}^{n} \frac{\partial \boldsymbol{\mu}_i(\boldsymbol{B})}{\partial \boldsymbol{\beta}_{\boldsymbol{B}}} \boldsymbol{V}_i^{-1} \{ \boldsymbol{Y}_i - \boldsymbol{\mu}_i(\boldsymbol{B}) \} = \boldsymbol{0},$$

  where $\boldsymbol{\beta}_{\boldsymbol{B}} = \mathrm{vec}(\boldsymbol{B}_1, \ldots, \boldsymbol{B}_D)$
- The derivative is now with respect to $\boldsymbol{\beta}_{\boldsymbol{B}} \in \mathbb{R}^{R \sum_d p_d}$, and thus the number of estimating equations is reduced from the exponential order $\prod_d p_d$ to the linear order $R \sum_d p_d$
- Substantial reduction in dimensionality is the key to enable effective estimation and inference under a limited sample size
- Any two elements $\beta_{i_1 \ldots i_d}$ and $\beta_{j_1 \ldots j_d}$ in $\boldsymbol{B}$ share common parameters if $i_d = j_d$ for any $d = 1, \ldots, D$; so the coefficients are correlated if they share the same spatial locations along any one of the tensor modes
- Incorporates the spatial structure of the tensor coefficient implicitly

Tensor generalized estimation equations

- Estimate the intra-subject covariance structure $\boldsymbol{V}_i$ by

$$\boldsymbol{V}_i = \boldsymbol{A}_i^{1/2}(\boldsymbol{B})\boldsymbol{R}\boldsymbol{A}_i^{1/2}(\boldsymbol{B}),$$

  where $\boldsymbol{A}_i(\boldsymbol{B}) \in \mathbb{R}^{m\times m}$ is a diagonal matrix with $\sigma_{ij}^2(\boldsymbol{B})$ on the diagonal, and $\boldsymbol{R} \in \mathbb{R}^{m\times m}$ is the working correlation matrix

- Common choice of $\boldsymbol{R}$: independence, autocorrelation (AR), compound symmetry, and unstructured correlation
- Plug in an estimate of $\boldsymbol{R}$ and evaluate the derivate explicitly,

$$\boldsymbol{s}(\boldsymbol{B}) \equiv \sum_{i=1}^{n}[\boldsymbol{J}_1,\ \boldsymbol{J}_2, \ldots, \boldsymbol{J}_D]^{\mathsf{T}}\mathrm{vec}(\boldsymbol{X}_i)\boldsymbol{A}_i^{1/2}(\boldsymbol{B})\widehat{\mathbf{R}}^{-1}\boldsymbol{A}_i^{-1/2}(\boldsymbol{B})\{\boldsymbol{Y}_i - \boldsymbol{\mu}_i(\boldsymbol{B})\} = \mathbf{0},$$

  where $\boldsymbol{J}_d$ is the $\prod_{d=1}^{D} p_d \times Rp_d$ Jacobian matrix

Estimation and Asymptotics

- Estimation: a block relaxation algorithm
  - Estimate $\boldsymbol{B}_1, \ldots, \boldsymbol{B}_D$ one at a time
  - Each step is a classical GEE, with Rp_d equations to solve
- Rank estimation: a quasi-likelihood independence model criterion

$$\mathrm{BIC}(R) = -2\ell(\widehat{\boldsymbol{B}}(R); \boldsymbol{I}_m) + \log(n)p_e,$$

  - Working correlation matrix: independence structure $\boldsymbol{I}_m$
  - Effective number of parameters: $p_e = R(p_1 + p_2) - R^2$ for $D = 2$, and $p_e = R(\sum_d p_d - D + 1)$ for $D > 2$
- Asymptotics:
  - Consistency: the GEE solution is consistent even under a misspecified working correlation structure
  - Asymptotic normality
  - Rank selection consistency: the estimated rank converges to the true rank in probability

Regularization and Asymptotics

- Regularization:

$$n^{-1}\boldsymbol{s}(\boldsymbol{B}) - \begin{pmatrix} \partial_{\beta_{\mathbf{11}}^{(\mathbf{1})}} P_\lambda(|\beta_{11}^{(1)}|, \rho_n) \\ \vdots \\ \partial_{\beta_{di}^{(r)}} P_\lambda(|\beta_{di}^{(r)}|, \rho_n) \\ \vdots \\ \partial_{\beta_{Dp_D}^{(R)}} P_\lambda(|\beta_{Dp_D}^{(R)}|, \rho_n) \end{pmatrix} = \mathbf{0},$$

  - Lasso: $P_\lambda(|\beta|, \rho_n) = \rho_n|\beta|$
  - SCAD: $\partial/\partial|\beta| P_\lambda(|\beta|, \rho_n) = \rho_n \left\{1_{\{|\beta| \le \rho_n\}} + (\lambda\rho_n - |\beta|)_+/(\lambda - 1)1_{\{|\beta| > \rho_n\}}\right\}$
  - ρ_n is the penalty tuning parameter

- Asymptotics:
  - Region selection consistency: the support of true tensor coefficient can be recovered with probability approaching one under the SCAD penalty

Simulation

n	m	Working Correlation	Bias^2	Variance	MSE
50	10	Exchangeable	122.0	383.6	**505.6(7.9)**
		AR-1	139.1	530.0	669.1(15.8)
		Independence	119.1	393.9	513.0(11.0)
100	10	Exchangeable	85.8	128.9	**214.7(2.2)**
		AR-1	88.0	159.1	247.1(3.0)
		Independence	93.0	141.2	234.2(2.8)
150	10	Exchangeable	86.1	51.3	**137.2(0.6)**
		AR-1	85.6	56.0	141.6(0.6)
		Independence	84.9	62.3	147.2(0.9)

Table: Bias, variance, and MSE of the tensor GEE estimates under various working correlation structures. Reported are the average out of 100 simulation replicates. The true intra-subject correlation is equicorrelated with $\rho_n = 0.8$.

ADNI revisit

	RMSE: $\{\sum_{i=1}^{n} n^{-1}(Y_{im} - \hat{Y}_{im})^2\}^{1/2}$			
Working Correlation	Independence	Equicorrelated	AR(1)	Unstructured
regularization (Lasso)	2.460	2.349	2.270	2.570
regularization (SCAD)	2.324	2.202	**2.147**	2.674
no regularization	2.526	2.427	2.429	2.628
	Correlation: $\mathrm{Corr}(Y_{im}, \hat{Y}_{im})$			
Working Correlation	Independence	Equicorrelated	AR(1)	Unstructured
regularization (Lasso)	0.705	0.733	0.747	0.700
regularization (SCAD)	0.742	0.767	**0.781**	0.658
no regularization	0.701	0.716	0.725	0.693

Table: Prediction of future clinical MMSE scores using tensor GEE.

ADNI revisit

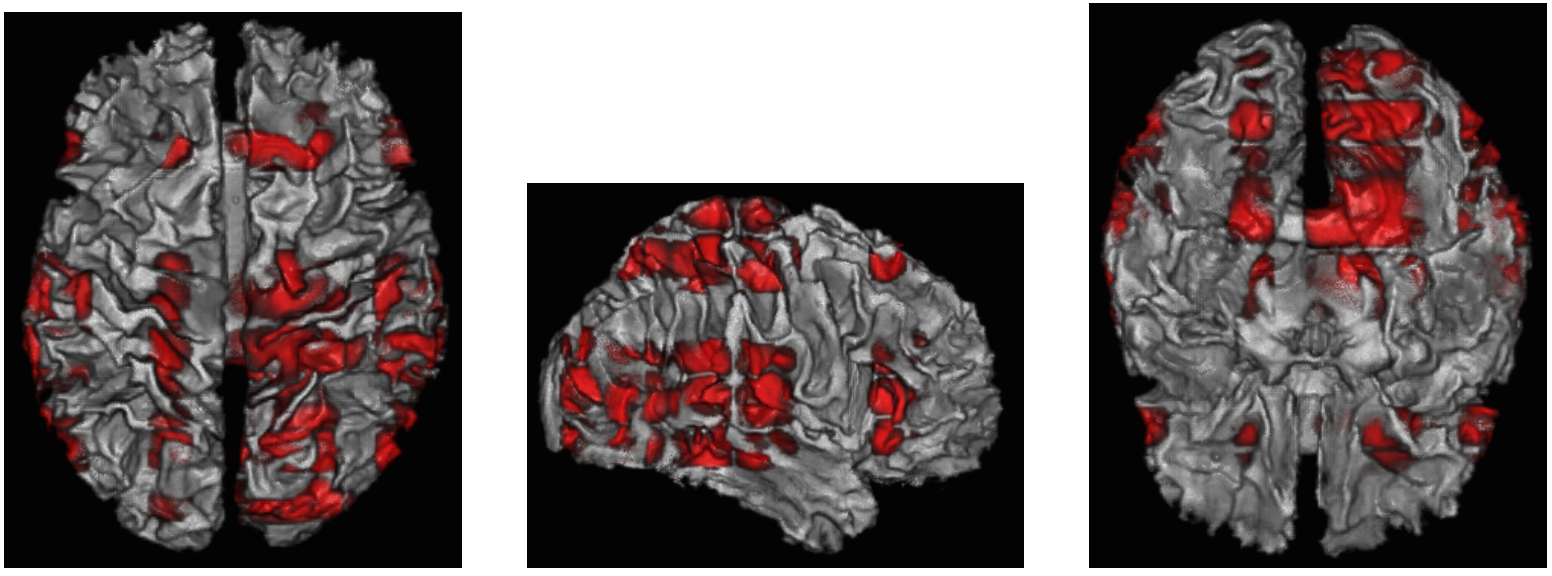


Figure: Regularized estimate overlaid on a random subject, with top, side and bottom views.

- Findings:
  - Identified regions: cerebral cortex, part of temporal lobe, parietal lobe, and frontal lobe
  - Involved in controlling language (Broca's area), reasoning (superior and inferior frontal gyri), part of sensory area (primary auditory cortex, olfactory cortex, insula, and operculum), somatosensory association area, memory loss (hippocampus), and motor function

Summary and Concluding Remarks

- Medical imaging: ultrahigh dimension
- Bayesian hierarchical modeling: Could be computationally intensive
- Hypothesis testing: control FDR
- Dimension Reduction and regularization method
- Tensor decomposition: more than 1 or 2 arrays, e.g., additional dimension for multimodality data
- Temporal and spatial imaging data
- Scalable computation
- Will cover brain network in Network Data Analysis
- Potential project topics: imaging classification, Bayesian modeling incorporating time and spatial information, heterogeneity imaging, time-varying feature