Task-based fMRI Timeseries Modeling for Alzheimer's Disease Diagnosis

Javier Salazar Cavazos

Department of Electrical Engineering and Computer Science
University of Michigan

April 2nd, 2025

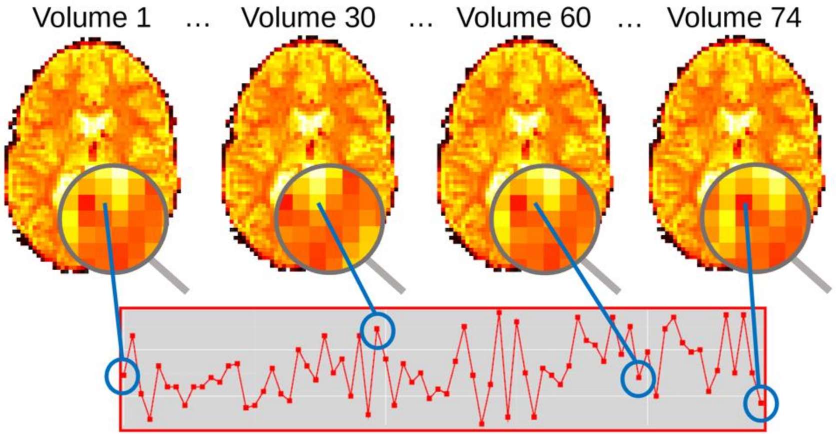
Background on Functional MRI & Alzheimer’s Disease

Background on functional MRI

https://www.monash.edu/researchinfrastructure/mbi/facilities/human/3t-mri

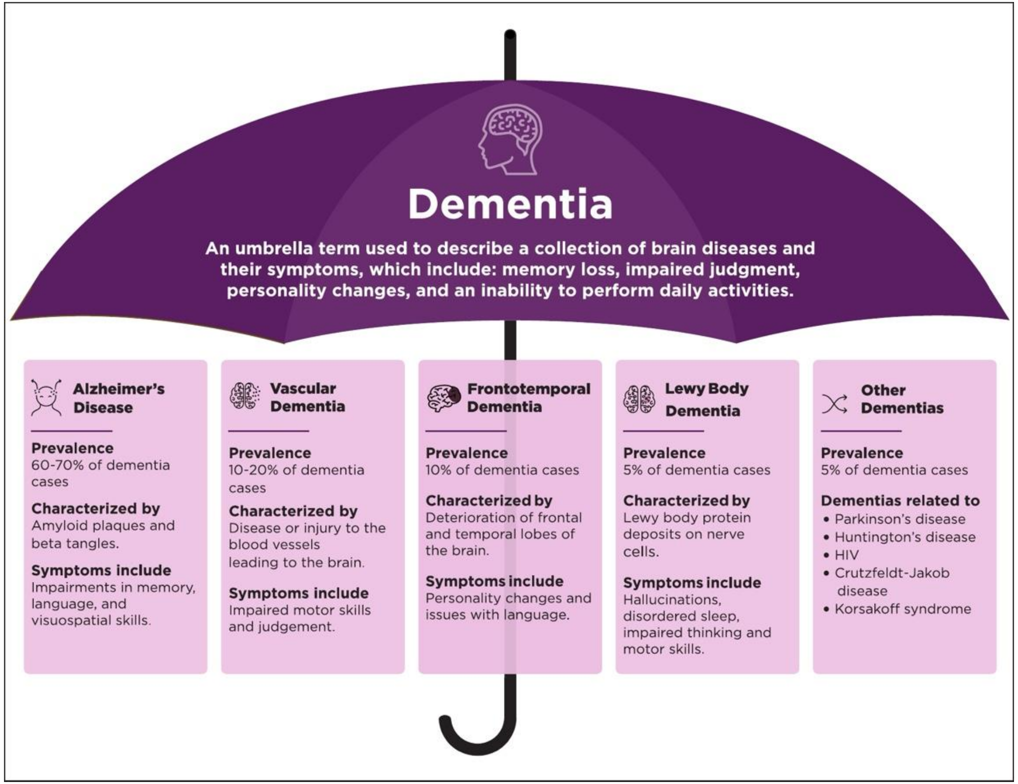
MRI Scanner

https://martynmcfarquhar.github.io/NCCN-IA-fMRIPreProcessing/1.fmri-structure.html



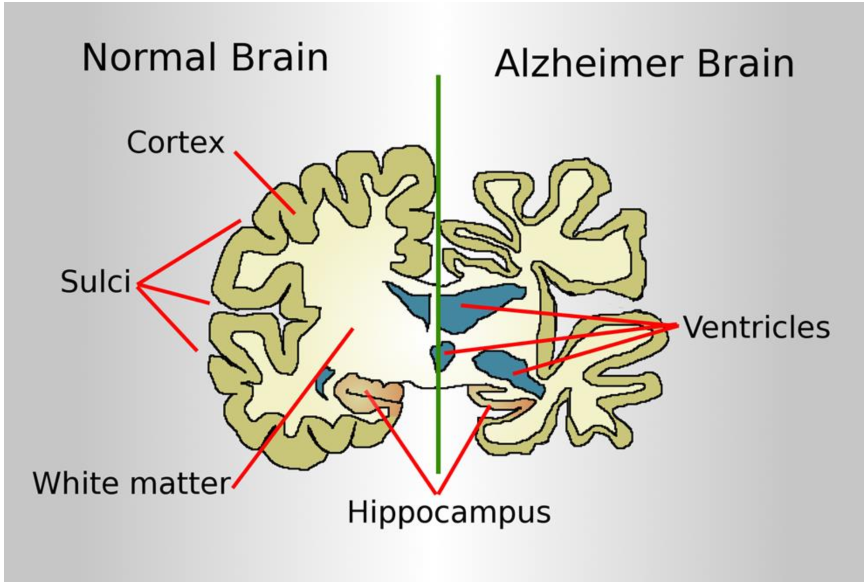
- 3D MRI volume acquired over time to form timeseries for each voxel region
- BOLD data measures blood concentration in gray matter regions
- We use BOLD data as a proxy for neural activity

Background on AD



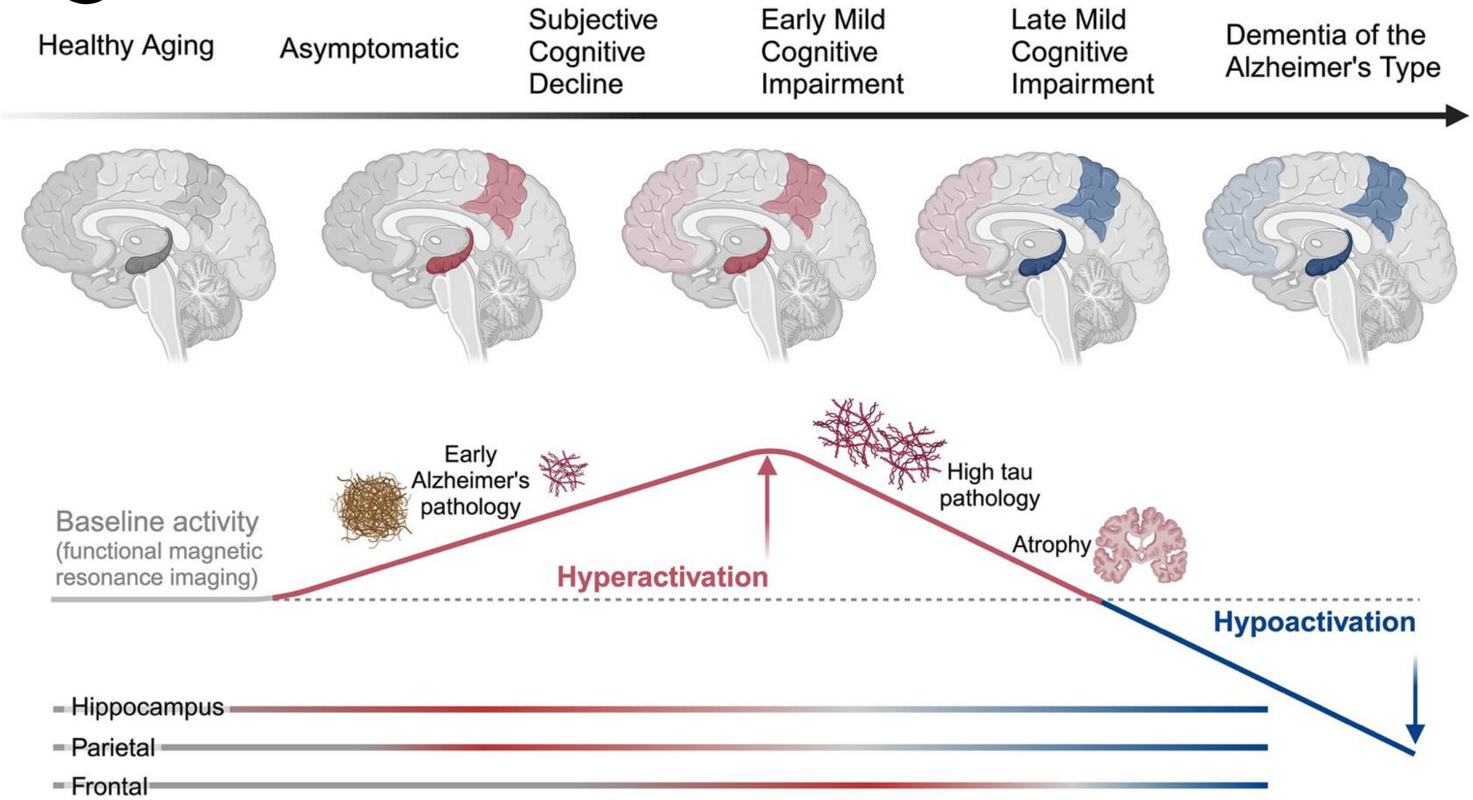
https://www.summerfieldredlands.com/understanding-the-umbrella-of-dementia/

Effects of AD on structural brain



https://www.emf.ethz.ch/en/knowledge/topics/health/neurodegenerative-diseases/

Background on AD in fMRI

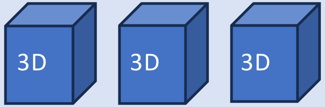


Fischer, Larissa, et al. "Precuneus activity during retrieval is positively associated with amyloid burden in cognitively normal older APOE4 carriers." *Journal of Neuroscience* 45.6 (2025).

Background - fMRI Data "Views"

4D Spatial-Temporal Data

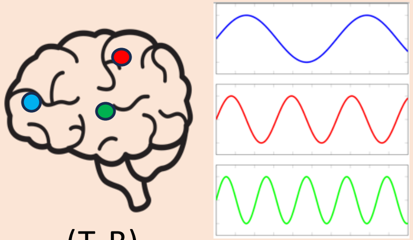
t=0 t=1 t=2



(T, X, Y, Z)

- High-dimensional
- Difficult to train models
- Harder to draw insights
- No information is lost since it is the raw data

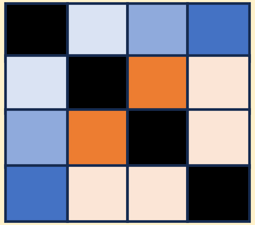
Timeseries Data



(T, B)

- Low-dimensional
- Highly interpretable
- Easier to train
- Some info lost (especially if B is small)

Functional Connectivity Data



(B, B)

- Super Low-dimensional
- Looks at relationships rather than BOLD activity
- Time information lost entirely

Chp. 5 – 4D CNN w/ rs-fMRI

Chp. 8 – Task fMRI

Timeseries Extraction

Timeseries Data

Atlas

Predefined coordinate areas in MNI space

264 ROIs(for Power et al)
Example ROI #62:
(65,-33,20) --> auditory

Data-driven Brain Parcellation

Clustering brain voxels based on {BOLD, FC, …}

fMRI-specific clustering methods for this

Anatomy-driven Brain Parcellation

Group voxels by known anatomical regions

Average timeseries for anatomical area

For our work, we use a modified "Power" atlas that contains a few extra ROIs (cerebellum)

Task-based fMRI Timeseries Modeling for Alzheimer's Disease Diagnosis

Preliminaries

- Previous work (ISMRM 2025) was on resting-state fMRI data that is plentiful due to ADNI consortium
- This ongoing work is on task-based fMRI data that is severely more constrained (zero public datasets for AD)
- Because of this, the 4D CNN model (ISMRM 2025) is not appropriate and not very interpretable!
- We will require a new approach for this problem

Preliminaries

- Many meta-analysis papers on MRI/PET/fMRI/… in AD
- Plentiful publications on AD + ML in resting-state fMRI
- Zero published works on AD + ML in task-based fMRI
- There are a few papers that do manual data analysis on task-based fMRI to draw conclusions
- Interesting topic since we believe that task-based fMRI can better “stress” the brain networks relative to resting

MADC Data (private dataset)

Resting (rs-fMRI)		
CN	MCI	DAT
256	51	41
Object Localization (tb-fMRI)		
CN	MCI	DAT
92	14	10
Face Name Association (tb-fMRI)		
CN	MCI	DAT
183	39	26

Numbers are # of subjects. CN = cognitively normal, MCI = mild cognitive impairment, DAT = Dementia of the Alzheimer's type

- Spatial: 2.4 mm
- Temporal: 0.8s TR, ~5 minute session
- Multi-Band Single-Echo EPI
- 3 scores based on written exams for each subject

For MCI & DAT classes:
We are severely data constrained for task-based fMRI

Goals

- This is not so much about automated detection
- Nothing to compare against in task-based fMRI
- More about comparing resting-state vs task-based fMRI as modalities useful in AD characterization
- Gaining insights as to what ROIs and subnetworks are important for diagnosis in task-based fMRI

Problem

- Explore the potential of task-based fMRI as a biomarker
- Severely data constrained in MCI/DAT subjects
- Direct classification models may not be ideal
- “Forced” to use healthy controls only for training
- We will look at alternative problem formulations

Proposed Ideas

Pretrain + Finetune

- Use rs-fMRI from ADNI & MADC for classification
- Finetune on task data
- Worth trying due to simplicity

Data Augmentation

- VAE-style generation
- BioDiffusion; synthesize multivariate class-specific samples
- Not enough samples for BioDiffusion and VAE gives poor resolution
- Does the small training dist. generalize?

ICA-style Learning

- Learn mixing coefficients to be used for a simple SVM/MLP classifier
- Use or develop a "Deep" ICA (nonlinear mixing) method
- Nonlinear mixing might be better suited for Task-based fMRI data?

Outlier Detection

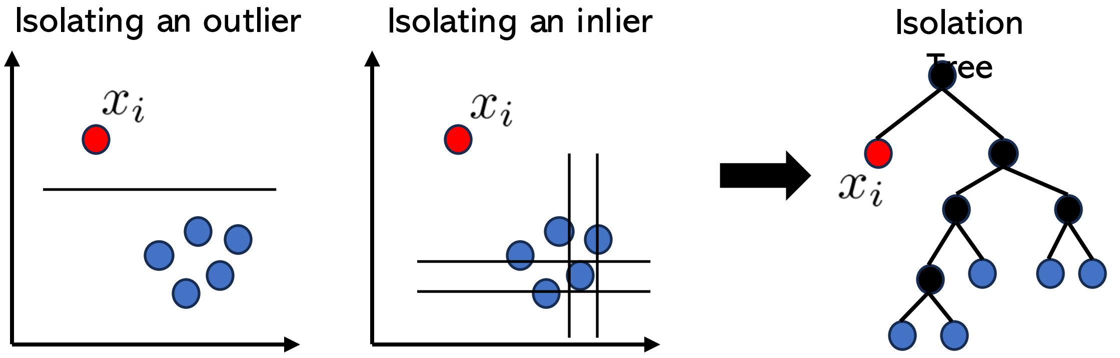
- We have plenty of healthy rs & tb fMRI data
- Train a model to do forecasting/masking/recon/regression
- Use outlier scores to predict severity of disease
- Use behavior score regression since it correlates with disease

We think outlier detection is a good approach so we can use MCI/DAT data for testing only

Outlier Detection

Outlier Detection

Isolation Forests are one of the most well-known and successful methods used in practice



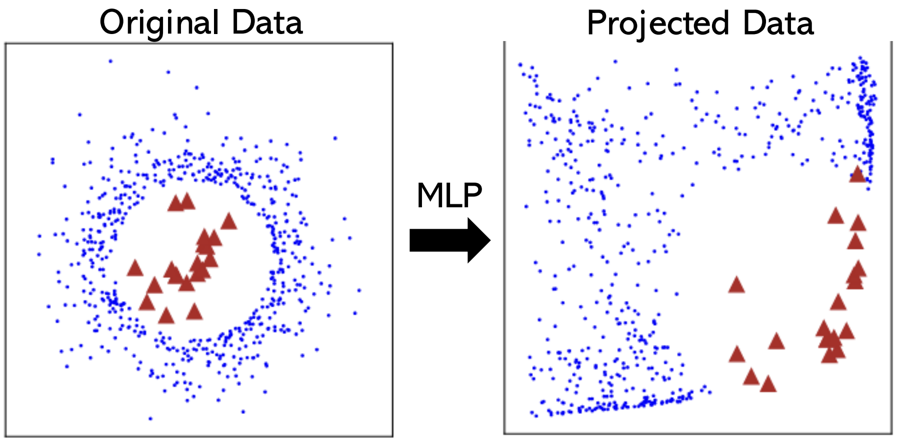
Isolation Forest

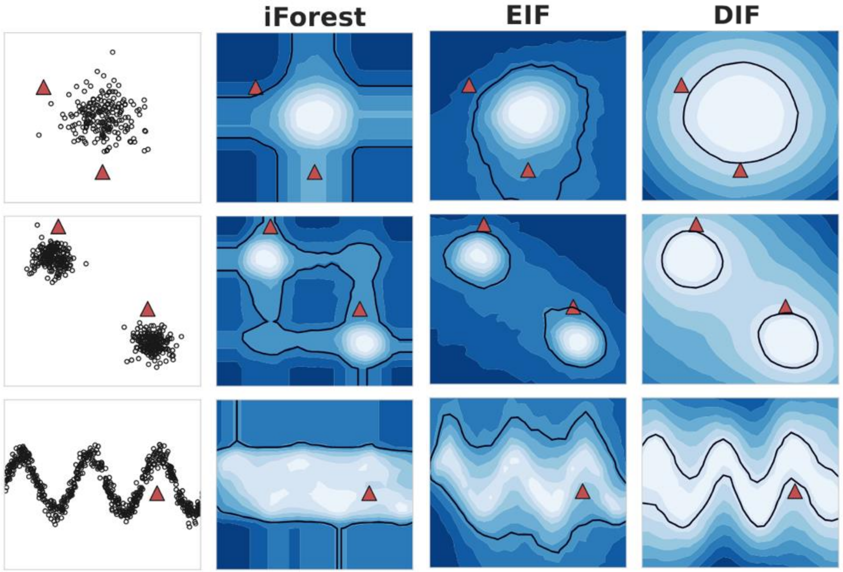
$$s(x_i) \propto \frac{\mathbb{E}[p(x_i)]}{\mathbb{E}[\mathbb{E}[p(x)]]}$$

This is an effective method but limited since it is only vertical/horizontal lines. Is there something better?

Deep Isolation Forests

- Despite the name, no training!
- Just do isolation forest in random "subspaces"
- Forest = ensemble of trees w/ different projections



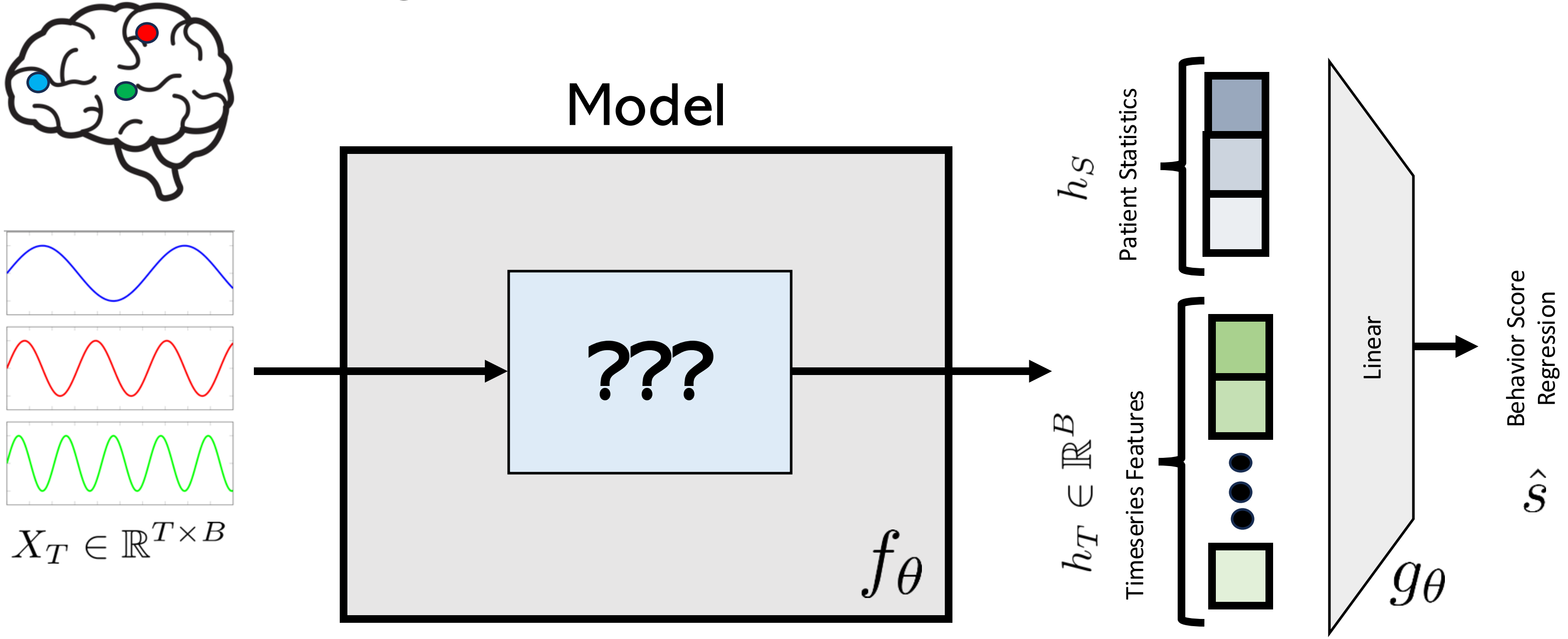


Xu, Hongzuo, et al. "Deep isolation forest for anomaly detection." *IEEE Transactions on Knowledge and Data Engineering* 35.12 (2023): 12591-12604.

Outlier Detection

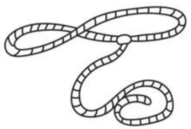
- We hope that our MCI/DAT data is well-separated
- If not, Deep Isolation Forest is a more robust approach
- The question becomes what features to use in DIF?
- The BOLD timeseries data? Many downsides to this
- Better to extract features from some model

Score Regression



$$\min_{\theta} \|s - g_\theta(f_\theta(X_T), h_s)\|_2^2 + \lambda \|f_\theta(X_T)\|_1$$

Data Fidelity   Sparse ROIs



Lasso-style

Deep State Space Models (SSMs)

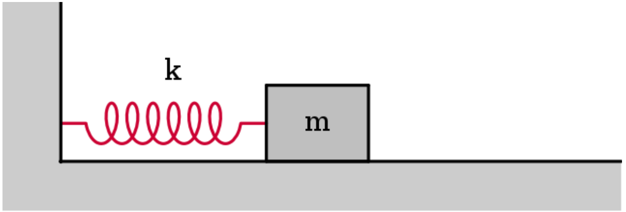
State Space Models (SSMs)

Given input signal u(t), generate output sequence y(t) using state x(t)

$$\left.\begin{aligned} x'(t) &= Ax(t) + Bu(t) \\ y(t) &= Cx(t) + Du(t) \end{aligned}\right\} \text{This is an LTI system}$$

A = State Matrix
B = Input Matrix
C = Output Matrix
D = Feedthrough Matrix

Toy Example



Consider a mass attached to the wall with a spring.

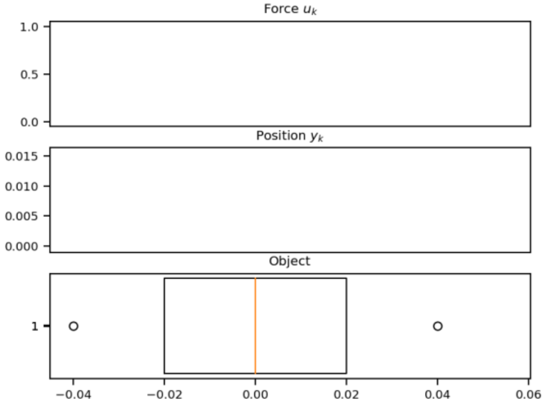
Let m = mass, k = spring constant, and b = friction constant

Recall that u(t) is force applied to spring and y(t) is position of object

$$my''(t) = u(t) - by'(t) - ky(t)$$

$$\Downarrow$$

$$A = \begin{bmatrix} 0 & 1 \\ -\frac{k}{m} & -\frac{b}{m} \end{bmatrix}, B = \begin{bmatrix} 0 \\ \frac{1}{m} \end{bmatrix}, C = \begin{bmatrix} 1 & 0 \end{bmatrix}$$



Of course, in contrast, Deep SSMs learn {A,B,C} for complex systems

Discretization*

So far, we have focused on continuous time SSMs, but sampled data is discrete!

Recall differential equations (Euler's method):

$$x'(t) = \lim_{\Delta \to 0} \frac{x(t+\Delta) - x(t)}{\Delta} \Rightarrow \Delta x'(t) = x(t+\Delta) - x(t)$$

Using this we can say:

$$x(t+\Delta) = \Delta x'(t) + x(t) \quad \text{and recall} \quad x'(t) = Ax(t) + Bu(t)$$

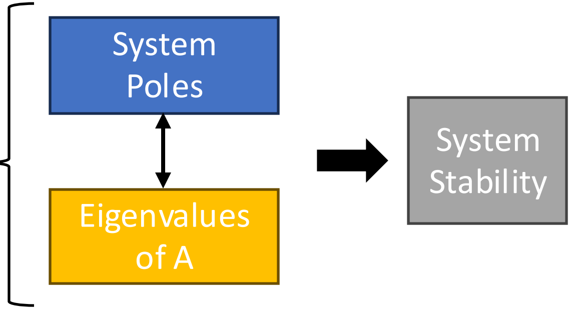
If we combine both from above:

$$x(t+\Delta) = \underbrace{(\Delta A + I)}_{\bar{A}} x(t) + \underbrace{(\Delta B)}_{\bar{B}} u(t)$$

* Better methods for this exist like zero-order hold (ZOH)

A: System Matrix

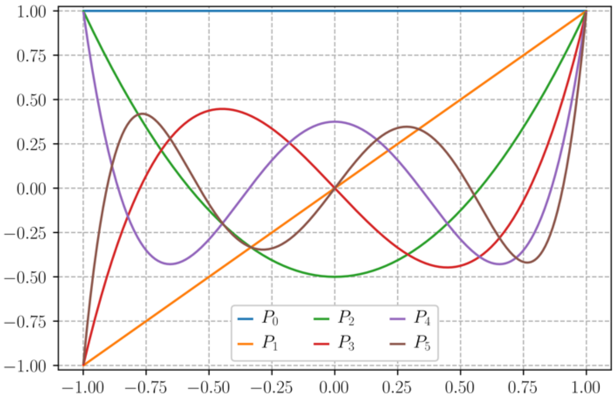
Turns out choice of A/B matrices matter greatly for performance (not surprising!)



In previous work, it was found that using Legendre polynomials for initialization is a good strategy

$$\underbrace{A_{nk}}_{\text{HiPPO}} = \begin{cases} (2n+1)^{\frac{1}{2}}(2k+1)^{\frac{1}{2}} & n > k \\ n+1 & n = k \\ 0 & n < k \end{cases}$$

Gu, Albert, et al. "How to train your hippo: State space models with generalized orthogonal basis projections." *arXiv preprint arXiv:2206.12037* (2022).



https://en.wikipedia.org/wiki/Legendre_polynomials

Deep SSM + Structured A = S4

Gu, Albert, Karan Goel, and Christopher Ré. "Efficiently modeling long sequences with structured state spaces." *arXiv preprint arXiv:2111.00396* (2021).

Structured State Space for Sequence Modeling (S4)

Inference:

$$x_k = \bar{A} x_{k-1} + \bar{B} u_k, \quad y_k = \bar{C} x_k$$

Training:

$$y_k = \bar{C}\bar{A}^k \bar{B} u_0 + \bar{C}\bar{A}^{k-1}\bar{B} u_1 + \ldots + \bar{C}\bar{B} u_k$$

$$\bar{K} = (\underbrace{\bar{C}\bar{B}, \bar{C}\bar{A}\bar{B}, \ldots, \bar{C}\bar{A}^k\bar{B}}_{\text{1D elements}}, \ldots) \Rightarrow y = u * \bar{K}$$

Recall that we have an LTI system:

$$y = u * \bar{K} \Leftrightarrow \tilde{Y} = \tilde{U}\tilde{K}$$

We can easily parallelize things for fast training!
In practice, K is not actually constructed. There are some tricks for this.

Why do SSMs matter?

	sCIFAR
Transformer	62.2
LSTM	63.01
r-LSTM	72.2
UR-LSTM	71.00
UR-GRU	74.4
HiPPO-RNN	61.1
LMU-FFT	-
LipschitzRNN	64.2
TCN	-
TrellisNet	73.42
CKConv	63.74
LSSL	84.65
S4	**91.13**

Vectorized image data.

	RAW	0.5×
Transformer	✗	✗
Performer	30.77	30.68
ODE-RNN	✗	✗
NRDE	16.49	15.12
ExpRNN	11.6	10.8
LipschitzRNN	✗	✗
CKConv	71.66	65.96
WaveGAN-D	96.25	✗
LSSL	✗	✗
S4	**98.32**	**96.30**

Speech classification.

- Performs well for sequence data
- Vanilla CNNs & Transformers perform poorly with timeseries
- Discretized matrices mean flexibility to sampling rate effects
- Trains like a transformer but does inference like an RNN

S4 → Mamba

LTI systems are great but there are limitations to what we can learn, for example, noun selection:

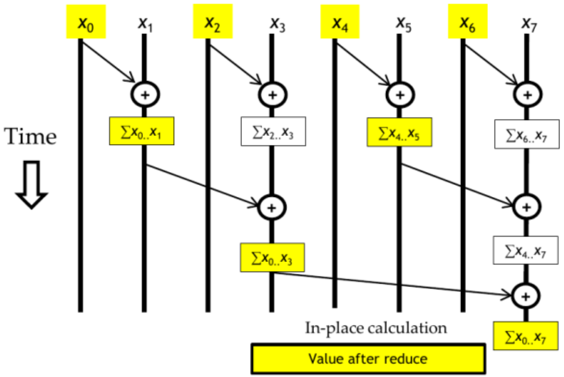
fMRI	members	really	like	cats

→

fMRI	members	cats

A is kept the same (we want state to be static) but we let B & C be time-varying so SSM is content-aware

No longer LTI system, but surprisingly, you can still compute things quickly with parallel scanning:



Source: NVIDIA

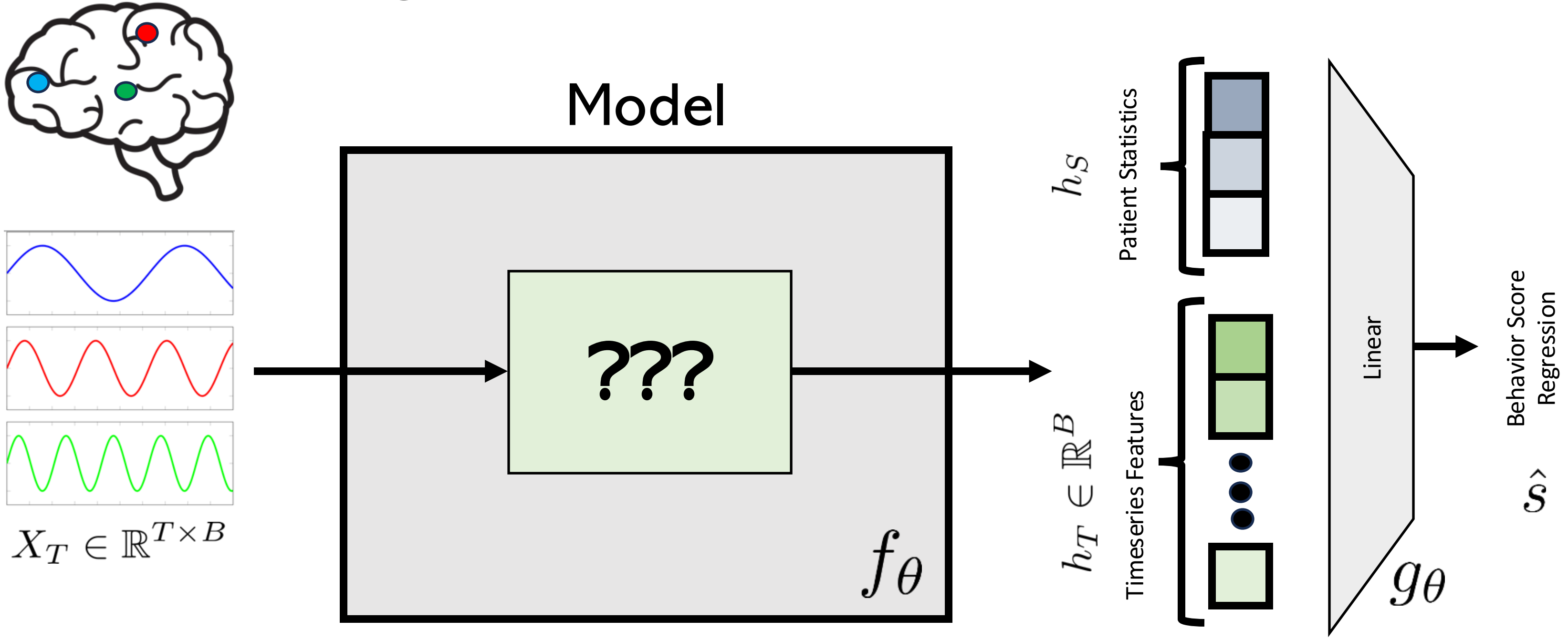
Mamba (S6) = S4 + Selective Scanning

- Not LTI so more powerful model for complex tasks
- No easy convolution, must use recurrence equation
- Still fast thanks to parallel scanning

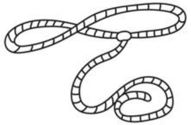
Gu, Albert, and Tri Dao. "Mamba: Linear-time sequence modeling with selective state spaces." *arXiv preprint arXiv:2312.00752* (2023).

Score Regression

Score Regression

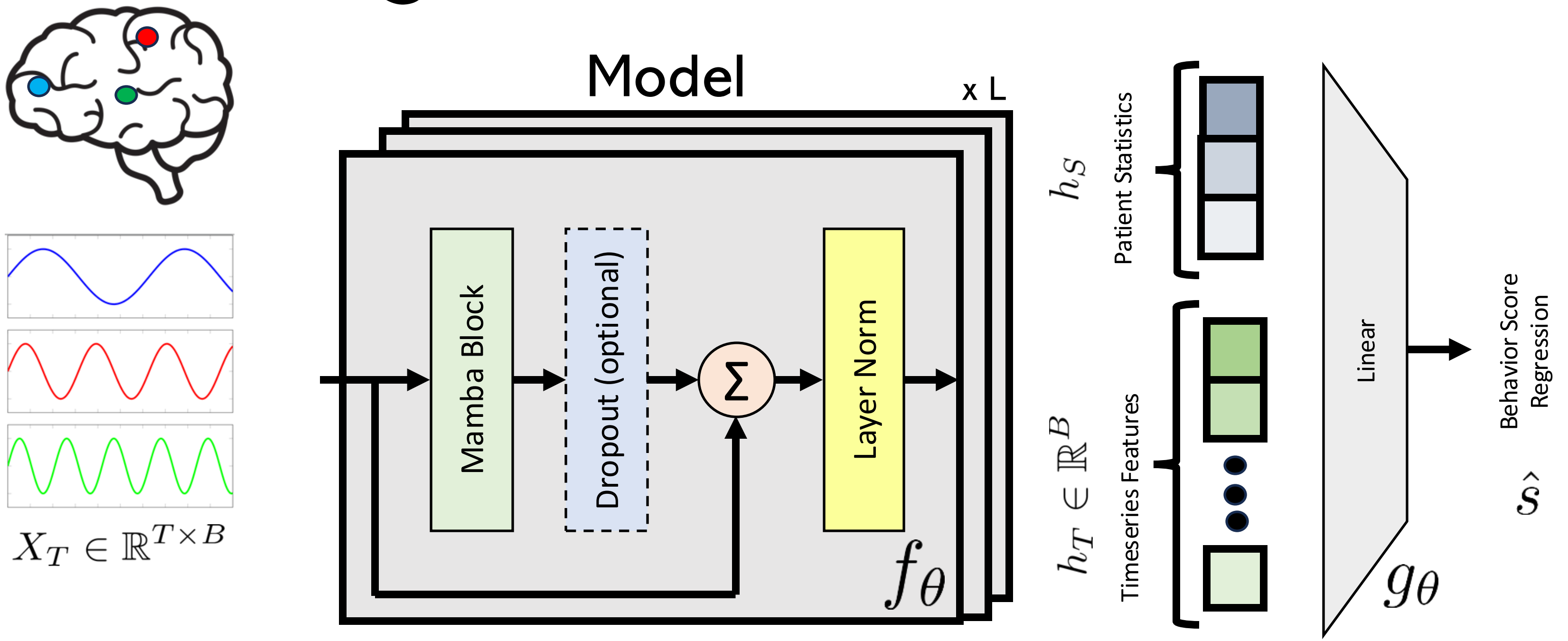


$$\min_\theta \underbrace{\|s - g_\theta(f_\theta(X_T), h_s)\|_2^2}_{\text{Data Fidelity}} + \underbrace{\lambda \|f_\theta(X_T)\|_1}_{\text{Sparse ROIs}}$$

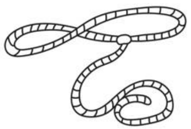


Lasso-style

Score Regression



$$\min_{\theta} \underbrace{\|s - g_\theta(f_\theta(X_T), h_s)\|_2^2}_{\text{Data Fidelity}} + \underbrace{\lambda\|f_\theta(X_T)\|_1}_{\text{Sparse ROIs}}$$



Lasso-style

The "Limitation" of Mamba

Univariate Data

$$x'(t) = Ax(t) + Bu(t)$$
$$y(t) = Cx(t) + Du(t)$$

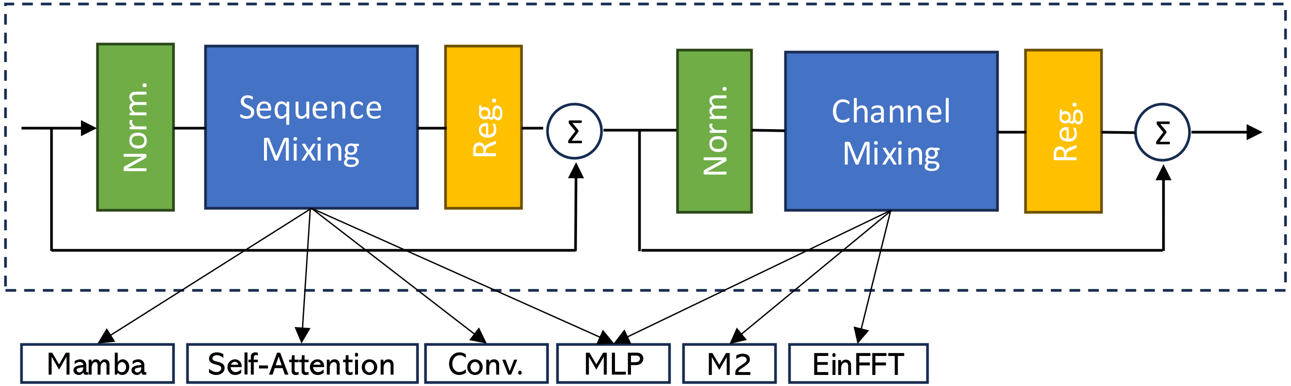
Multivariate Data

$$\mathbf{x}'(t) = \mathbf{A}\mathbf{x}(t) + \mathbf{B}\mathbf{u}(t)$$
$$\mathbf{y}(t) = \mathbf{C}\mathbf{x}(t) + \mathbf{D}\mathbf{u}(t)$$

Linearity; Separate SSMs for each feature dim.

We are capturing intra-variable features (temporal patterns) but not inter-variable features (interactions)!

Modern Model Backbone

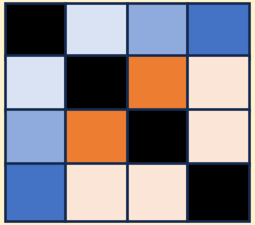


- Transformer = Self-Attention + MLP
- CNN = Convolution + MLP
- MLP-Mixer = MLP + MLP
- Mamba Model = Mamba + MLP

We can use this to learn ROI interactions but at the huge expense of interpretability!

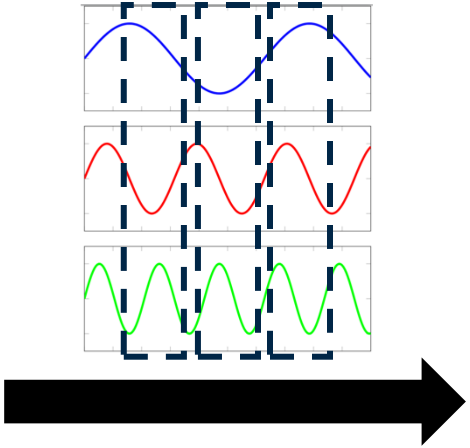
Learning spatial interactions

Functional Connectivity



(B, B)

- Super Low-dimensional
- Looks at relationships rather than BOLD activity
- Time information lost entirely

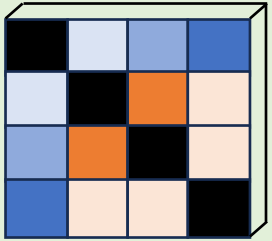


Note:

B = 272 so a vectorized graph is quite large!

0.5*B*(B-1) = ~37k

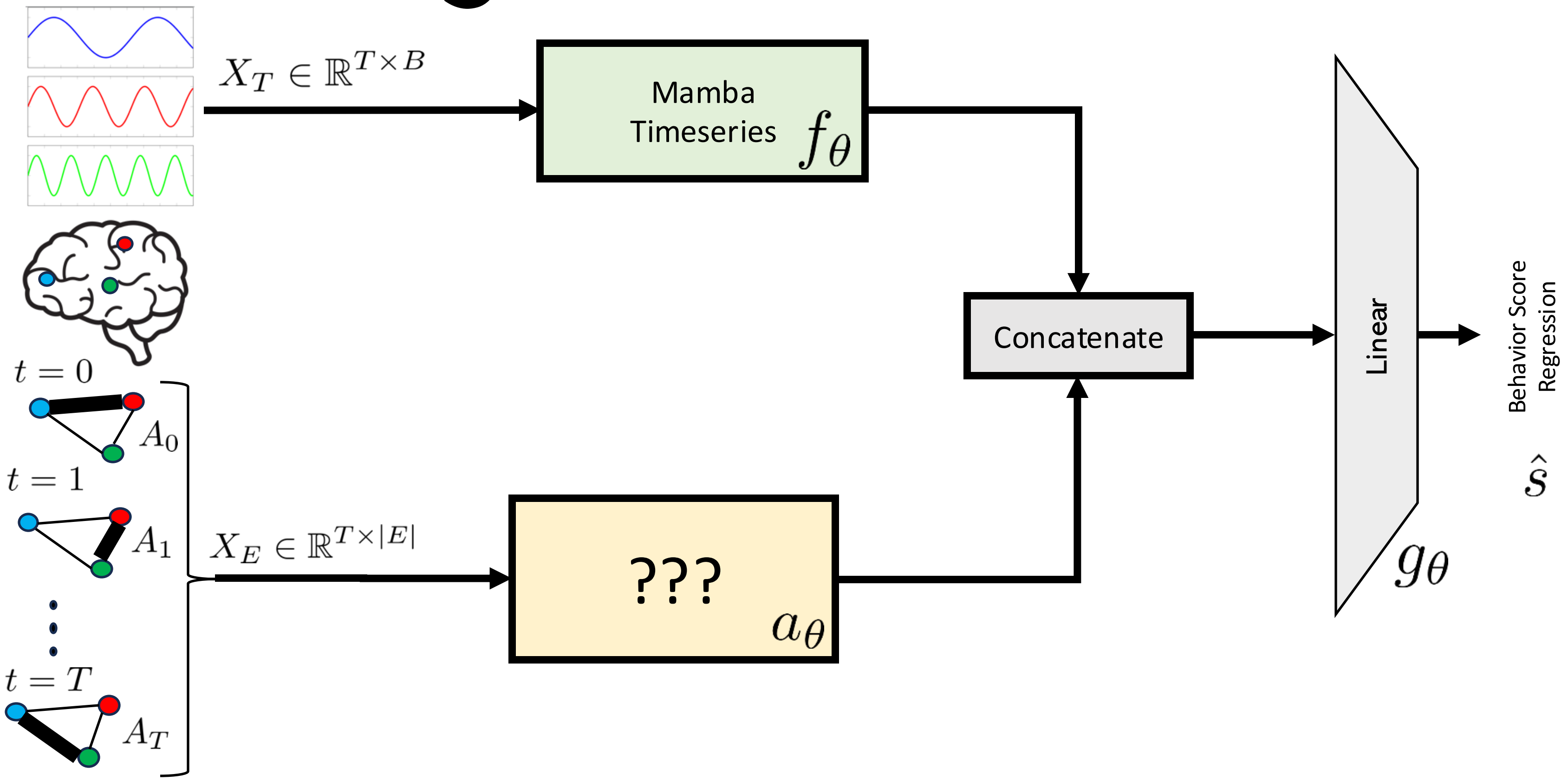
Dynamic Functional Connectivity



(T, B, B)

- Looks at relationships rather than BOLD activity
- Captures ROI interaction dynamics over time
- Useful for task-based fMRI

Score Regression



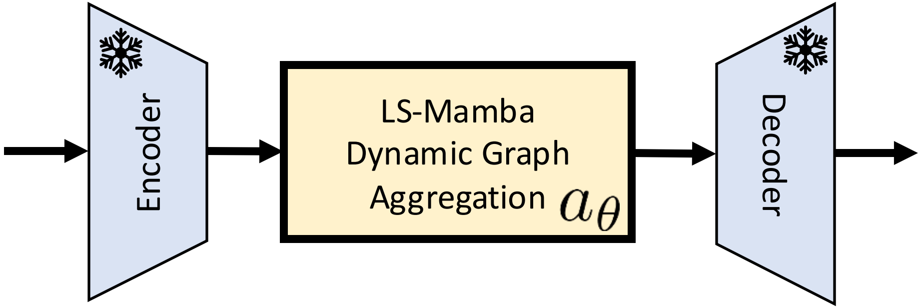
Dynamic Graph Aggregation

Lets "summarize" all ROI interactions to one useful FC for score regression:

$$X_E \in \mathbb{R}^{T\times|E|} \longrightarrow a_\theta(X_E) \in \mathbb{R}^{|E|}$$

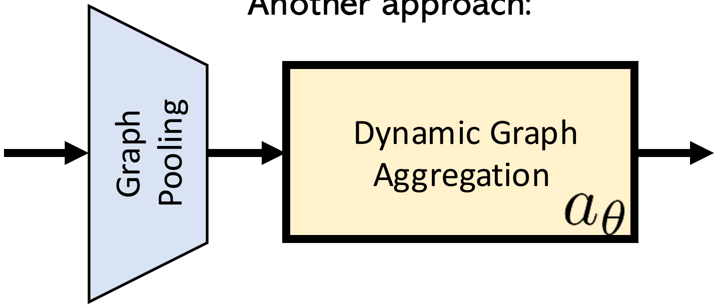
Remember that |E| = 37k which is very large to model directly!

One possible approach:



- Learn ROI interactions in a compressed space
- Good assumption since nearby ROIs likely have similar activity, i.e., fMRI data is correlated

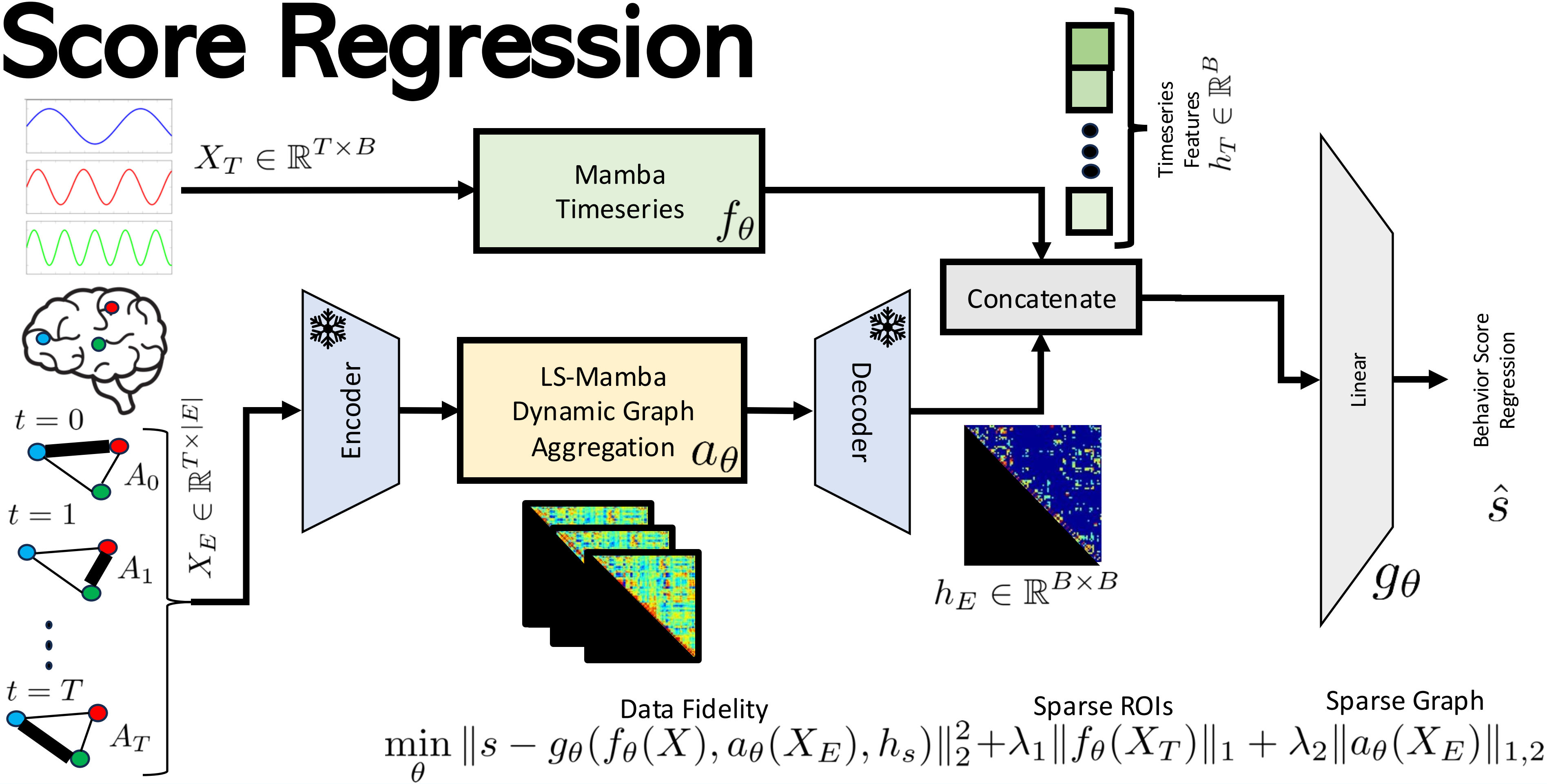
Another approach:



- Top-k pooling can learn to pick the k most important nodes (e.g., drop visual cortex)
- DiffPool can cluster similar nodes together (e.g., merge all ROIs in parietal lobe)

Cangea, Cătălina, et al. "Towards sparse hierarchical graph classifiers." *arXiv preprint arXiv:1811.01287* (2018).
Ying, Zhitao, et al. "Hierarchical graph representation learning with differentiable pooling." *Advances in neural information processing systems* 31 (2018).

Score Regression



Data Fidelity
Sparse ROIs
Sparse Graph

$$\min_\theta \|s - g_\theta(f_\theta(X), a_\theta(X_E), h_s)\|_2^2 + \lambda_1\|f_\theta(X_T)\|_1 + \lambda_2\|a_\theta(X_E)\|_{1,2}$$

Cost Function

Data Fidelity — Sparse ROIs — Sparse Graph

$$\min_{\theta} \|s - g_\theta(f_\theta(X), a_\theta(X_E), h_s)\|_2^2 + \lambda_1 \|f_\theta(X_T)\|_1 + \lambda_2 \|a_\theta(X_E)\|_{1,2}$$

Sparse ROIs

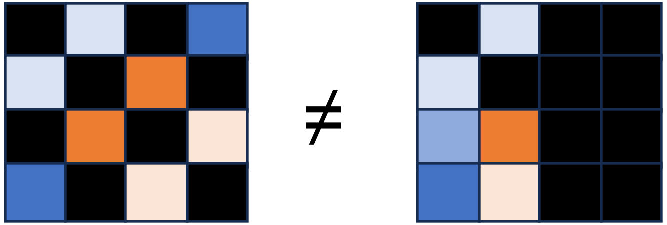
$$\lambda_1 \|f_\theta(X_T)\|_1$$

Visual cortex temporal features probably not useful

Sparse Graph

$$\lambda_2 \|a_\theta(X_E)\|_{1,2}$$

$$\|\text{vec}(\cdot)\|_1$$



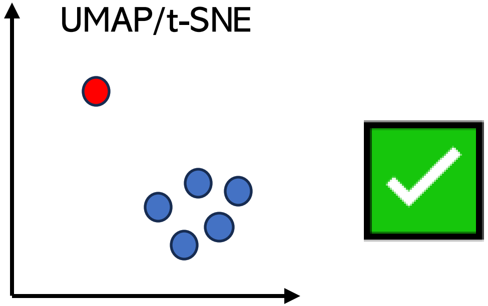
We are removing useless nodes in the graph as opposed to having sparse edges to find important fMRI subnetworks!

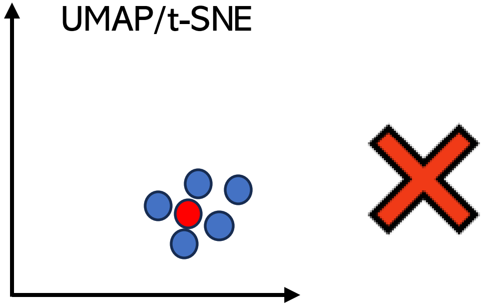
Post-training

Use extracted features for any of the following tasks:

- Outlier detection (Deep Isolation Forest)
- Clustering (K-means, PET-TURTLE (mine), etc…)
- Classification (MLP, SVM, K-nearest neighbor)

But firstly, we should visualize the extracted features to ensure these approaches are viable:

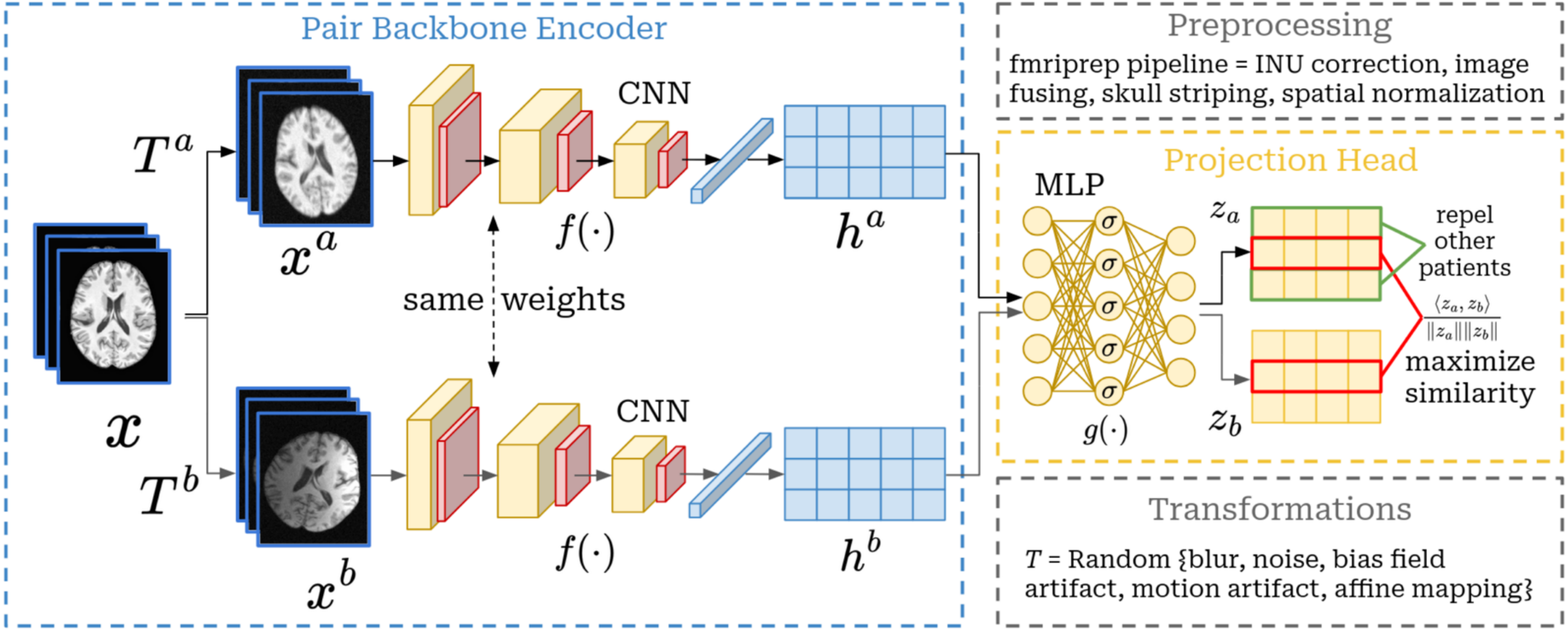




Possible Alternative to Score Regression (if time allows)

Contrastive Learning

Structural MRI ISMRM submission involved contrastive learning
Could we do something similar in fMRI for unsupervised learning?
We can explore gender prediction, score regression, etc…



Contrastive Learning

Method	Accuracy	Sensitivity	Specificity
2 class setting (CN/DAT) ref. classification 0.86 ACC (3D CNN)			
CL + LP	**0.82**	**0.91**	**0.73**
VAE + MLP	0.76	0.88	0.62
3 class setting (CN/MCI/DAT) ref. classification 0.58 ACC (3D CNN)			
CL + LP	**0.56**	**0.57**	**0.61**
VAE + MLP	0.47	0.48	0.58

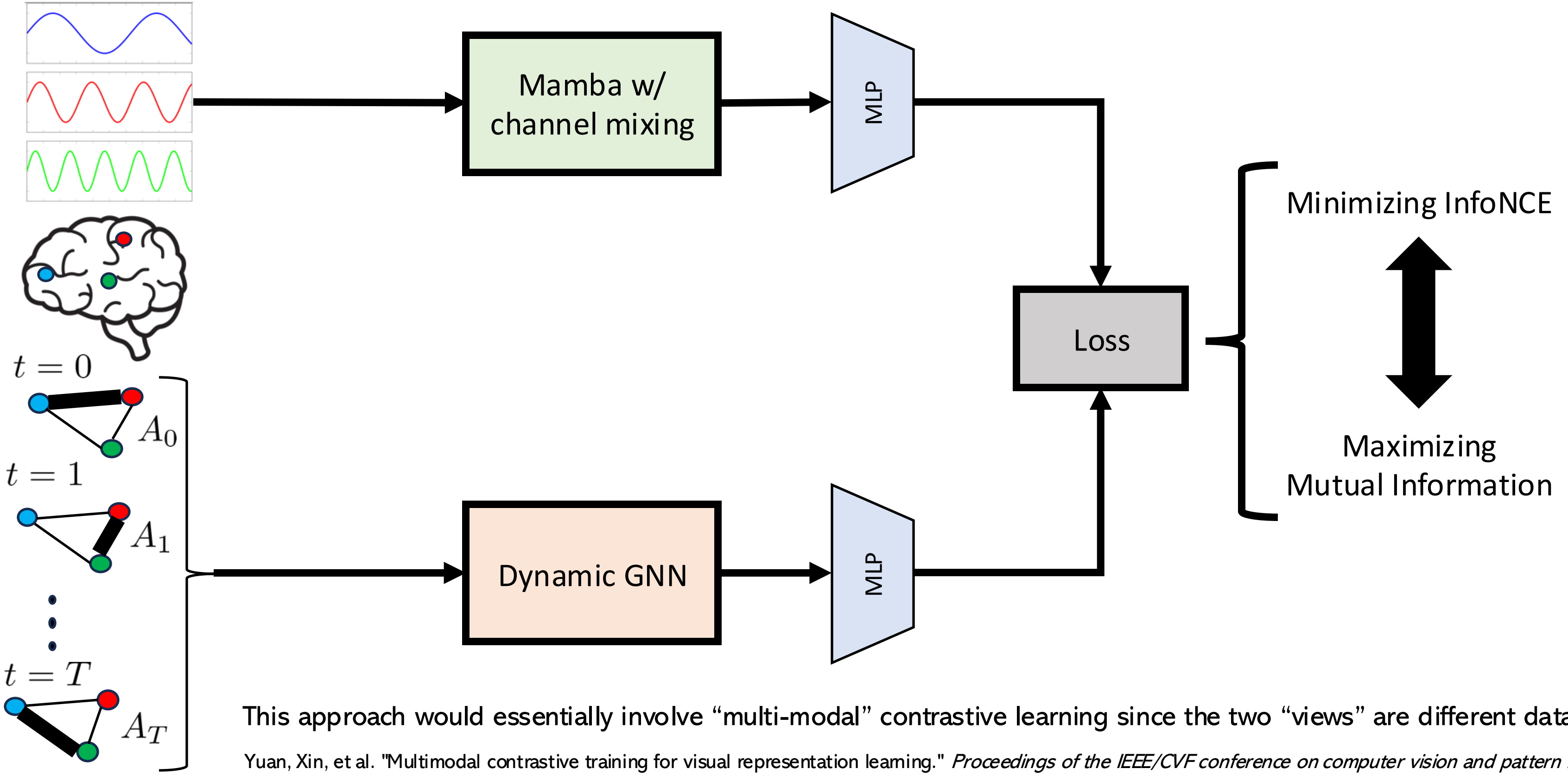
From my sMRI experiments, I found that CL is very competitive (close to end-to-end classification model) relative to VAE approach

Other works such as deep clustering methods using contrastive loss have found the same:

Dataset (ACC)	VAE	DAC = VAE + K-means Loss	Contrastive Clustering (CC)
CIFAR-10	0.291	0.522	**0.790**
CIFAR-100	0.152	0.238	**0.429**
STL-10	0.282	0.470	**0.850**
ImageNet-10	0.334	0.527	**0.893**

Li, Yunfan, et al. "Contrastive clustering." *Proceedings of the AAAI conference on artificial intelligence*. Vol. 35. No. 10. 2021.

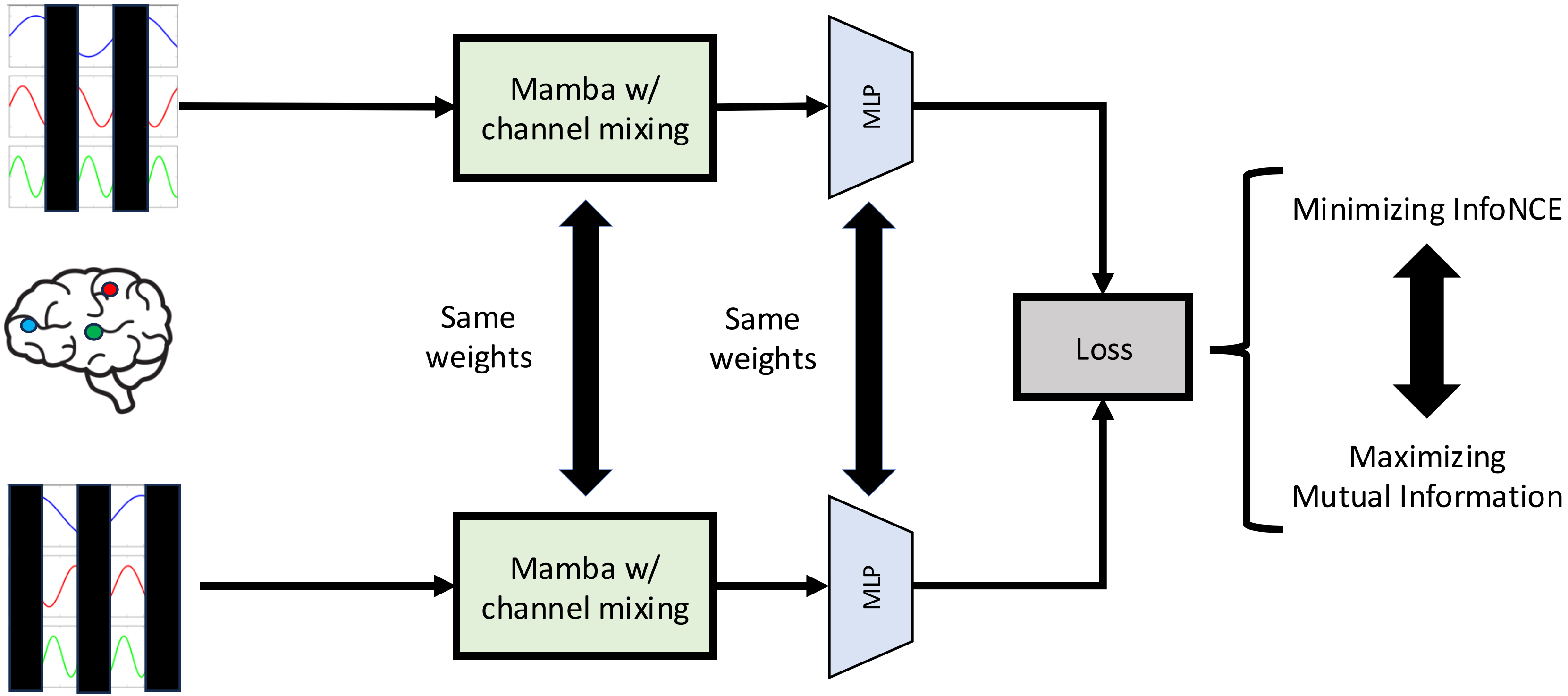
Possible idea #1



This approach would essentially involve "multi-modal" contrastive learning since the two "views" are different data sources:

Yuan, Xin, et al. "Multimodal contrastive training for visual representation learning." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2021.

Possible idea #2



This approach would involve timeseries contrastive learning with some mechanism to create “views” such as masking:

Pöppelbaum, Johannes, Gavneet Singh Chadha, and Andreas Schwung. "Contrastive learning based self-supervised time-series analysis." *Applied Soft Computing* 117 (2022): 108397.

Conclusion

- Task-based fMRI data could be a promising modality
- Limited data size --> alternative problem statements
- Score Regression + Outlier Detection
- We separately model ROI timeseries + spatial interactions in an interpretable way
- Possible to infer biological insights about what subnetworks are important for AD

Thank you for your attention!

Javier Salazar Cavazos

Advisors: Jeffrey A. Fessler, Laura Balzano, and Scott Peltier

Department of Electrical Engineering and Computer Science

University of Michigan

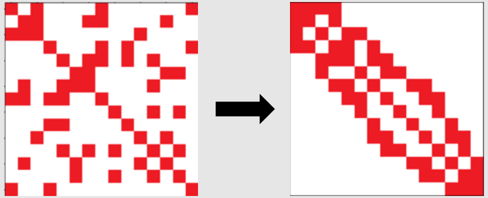
April 2^{nd}, 2025

My Questions

Q1: Block Diagonal Reordering

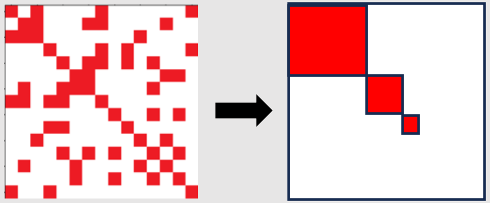
$$\min_{\theta} \overbrace{\|s - g_\theta(f_\theta(X), a_\theta(X_E), h_s)\|_2^2}^{\text{Data Fidelity}} + \lambda_1 \overbrace{\|f_\theta(X_T)\|_1}^{\text{Sparse ROIs}} + \lambda_2 \overbrace{\|\underbrace{a_\theta(X_E)}_{\text{This matrix will not be structured!}}\|_{1,2}}^{\text{Sparse Graph}}$$

Cuthill-Mckee Algo.



Is there something similar for graphs?

???



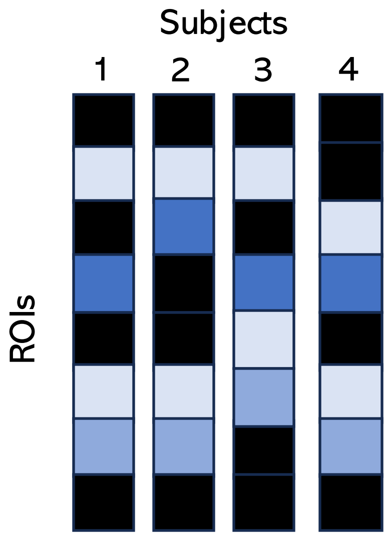
I expect there will be false connections

My thoughts:

- Convert correlation matrix to affinity matrix ([-1, 1] -> [0, bound] , critical to think about function!)
- Apply spectral clustering to find minimum “cuts” in graph
- Use labels to reorder symmetric matrix
- Optionally, look at the eigenspectrum of the graph Laplacian to find # of clusters (unknown here!)

Q2: Population Variance Control

Prototype enforcement?



How to ensure each subject has same ROIs selected?

My thoughts:

Learnable projection!

$$k = \mathrm{softmax}(h_T \odot p)$$

$$i = k_i > \lambda \in [0, 1]$$

$$\tilde{h_T} = (h_T \odot k)_i$$

Learnable prototype that works across the healthy population (removes L1 norm in cost function)