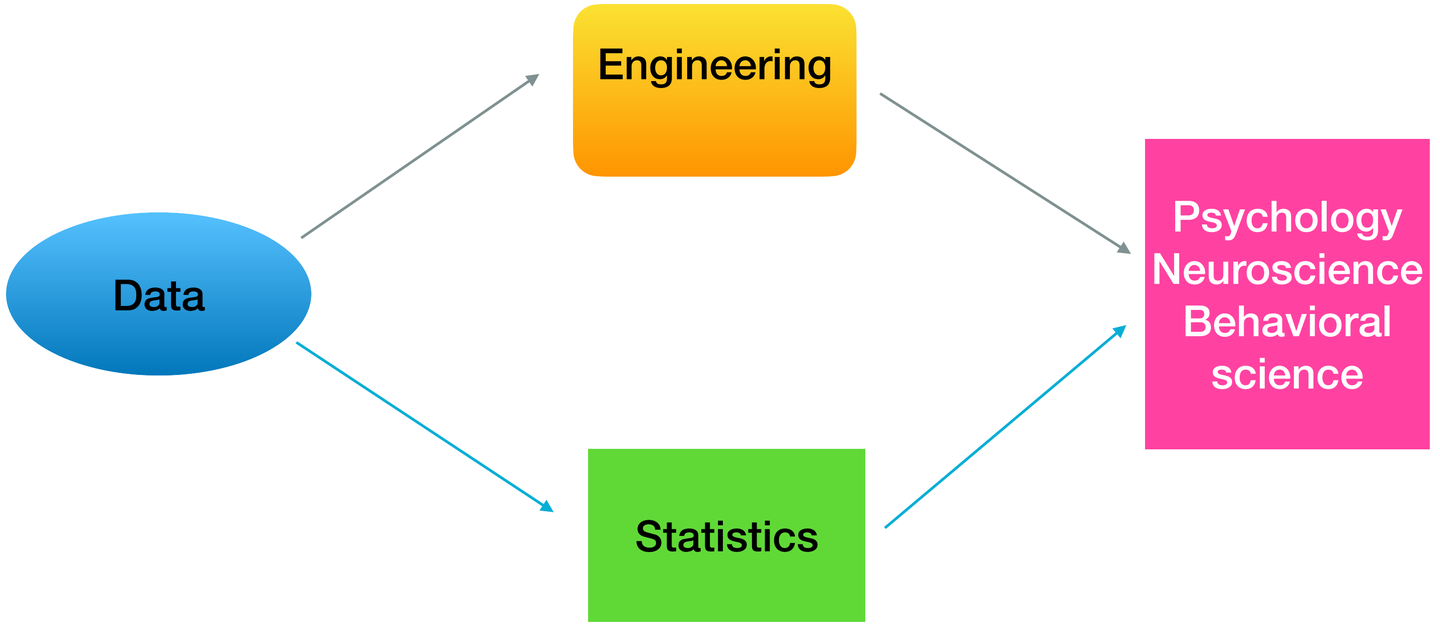
NIH T32 Research Presentation

On statistics, data science, and neuroscience

Yuan Yuan, July 24, 2024

About me

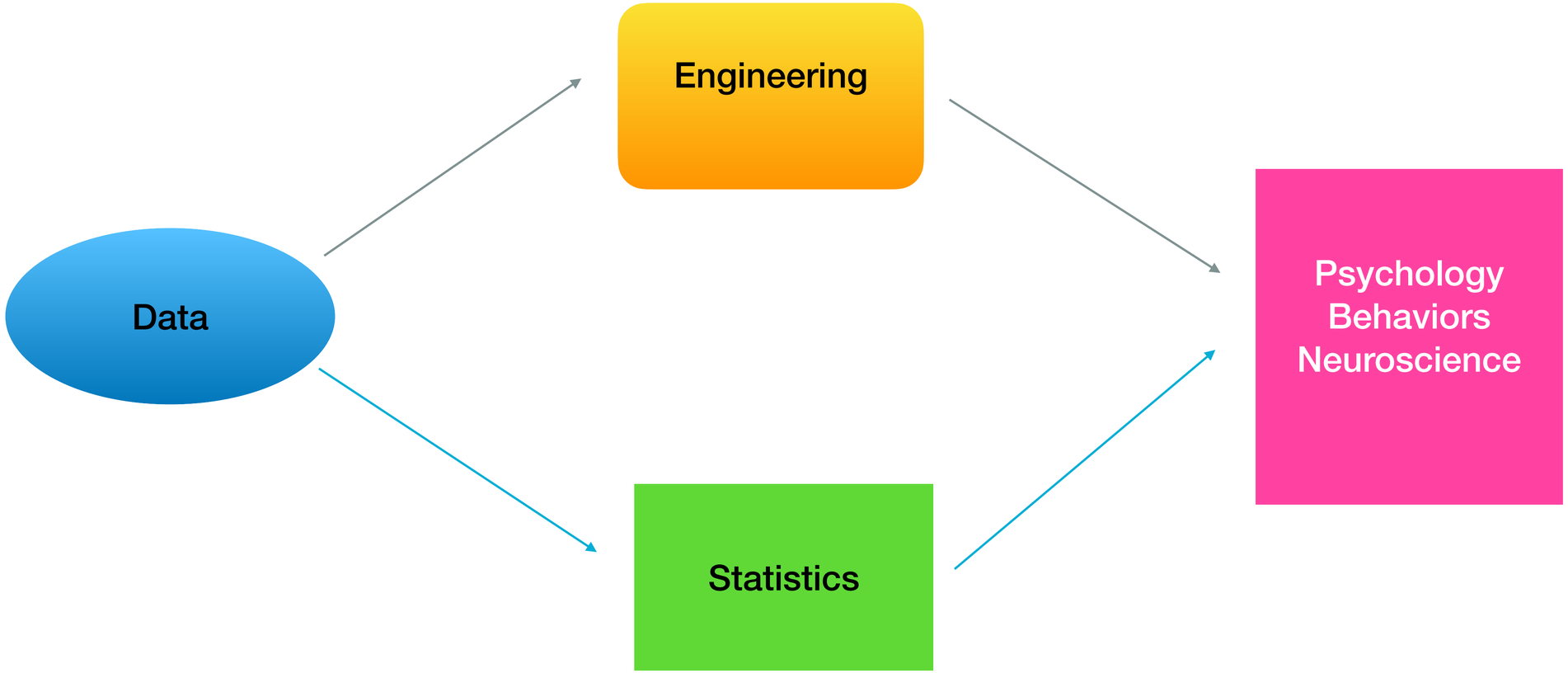
- Yuan Yuan
  - Ph.D. in Statistics, Auburn University, 2021
  - MS. in Psychology, Auburn University, 2015
  - B.S. in EE, HUST, 2010
- Research Interests
  - high-dimensional statistics,
  - data science,
  - neuroscience,
  - behavioral science.



Outline

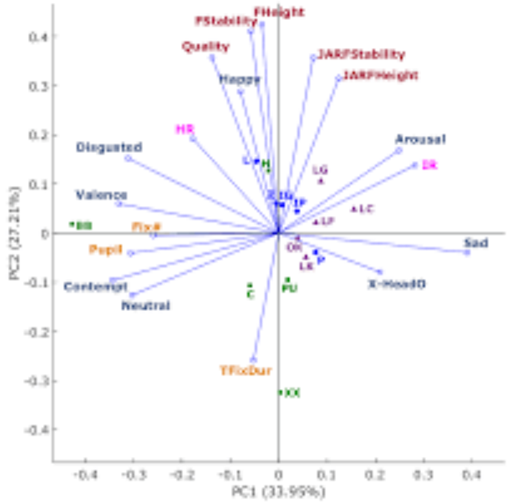
- Introduction
- Past findings
  - Study 1
  - Study 2
- Future plan

Introduction

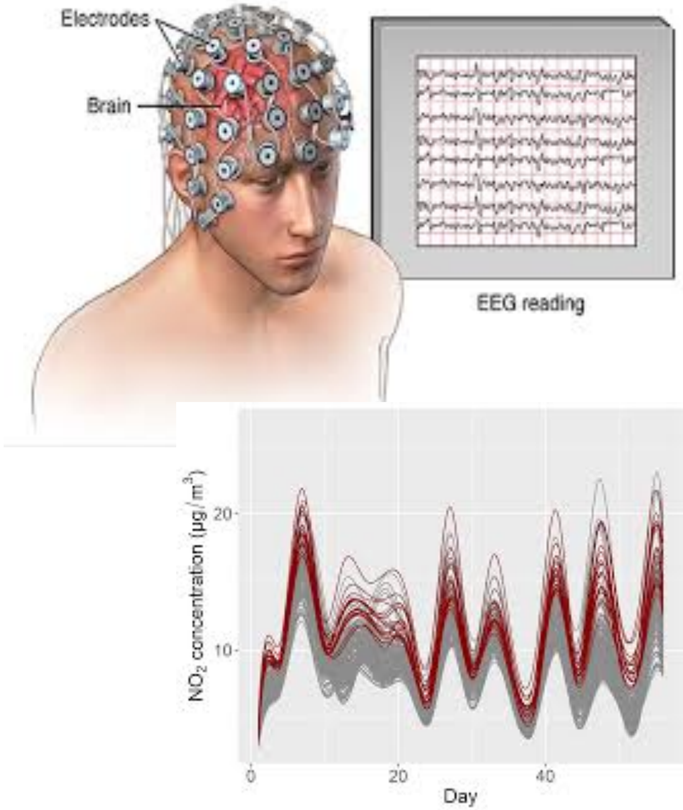


Introduction

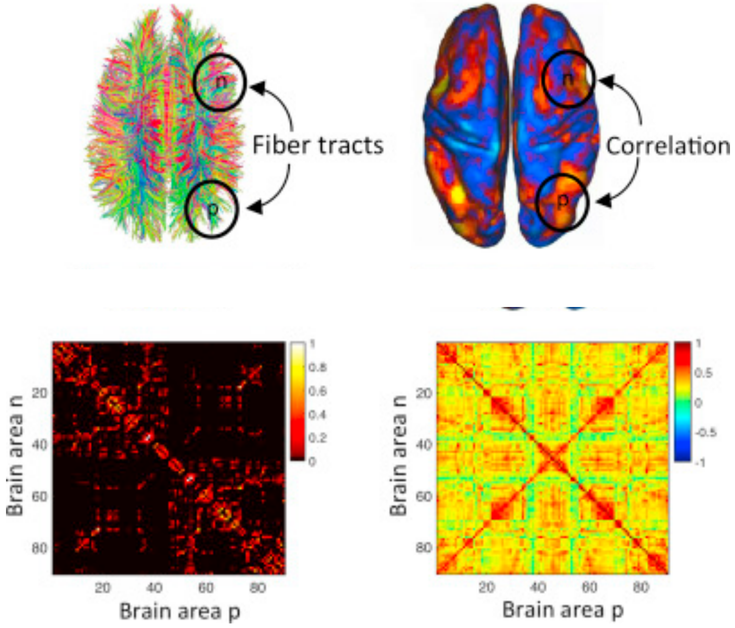
- **Multivariate**



- **Functional**

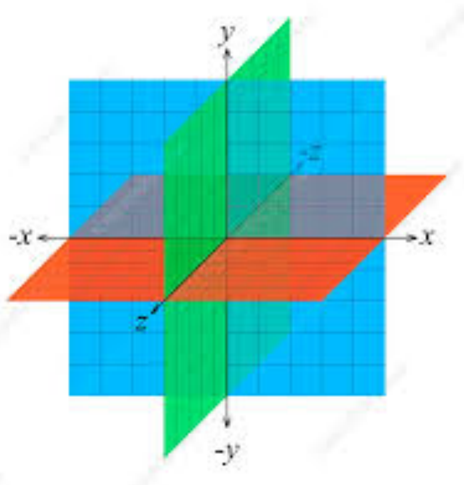


- **Manifolds**

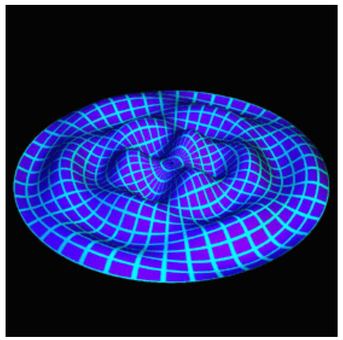


Statistical methods

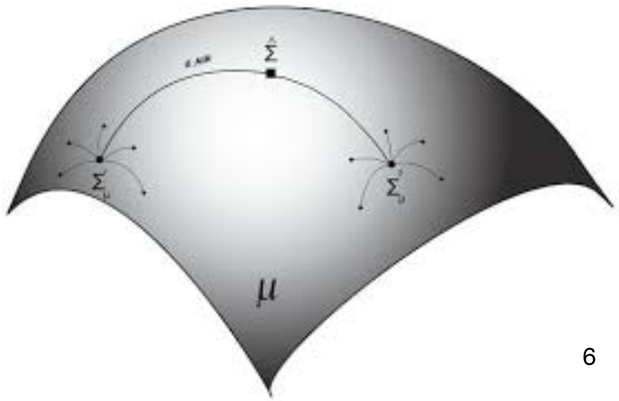
- **Multivariate**



- **Functional**

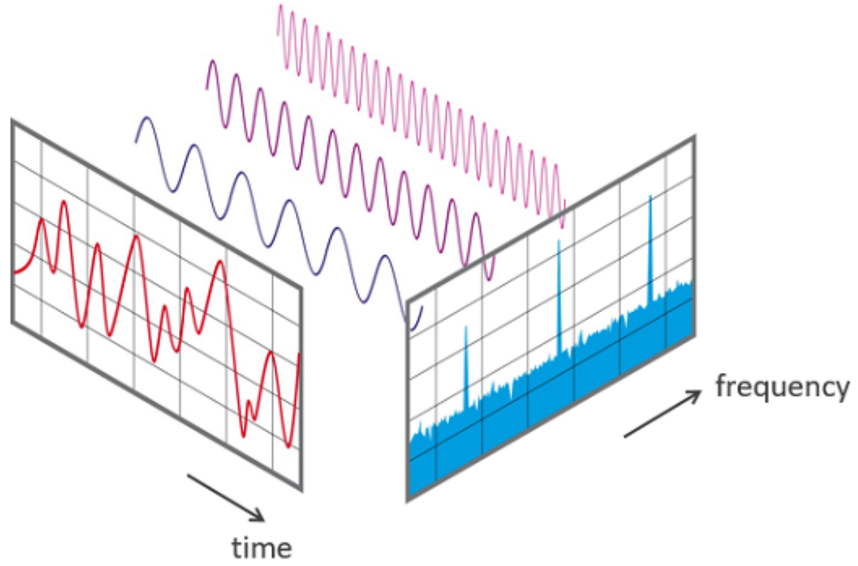


- **Manifold**



Engineering methods

- **Data driven models**



Outline

- Introduction
- Past findings
  - Study 1: statistic method
  - Study 2: data engineering method
- Future plan

Study1

- SURE Independent Screening for Functional Regression Model

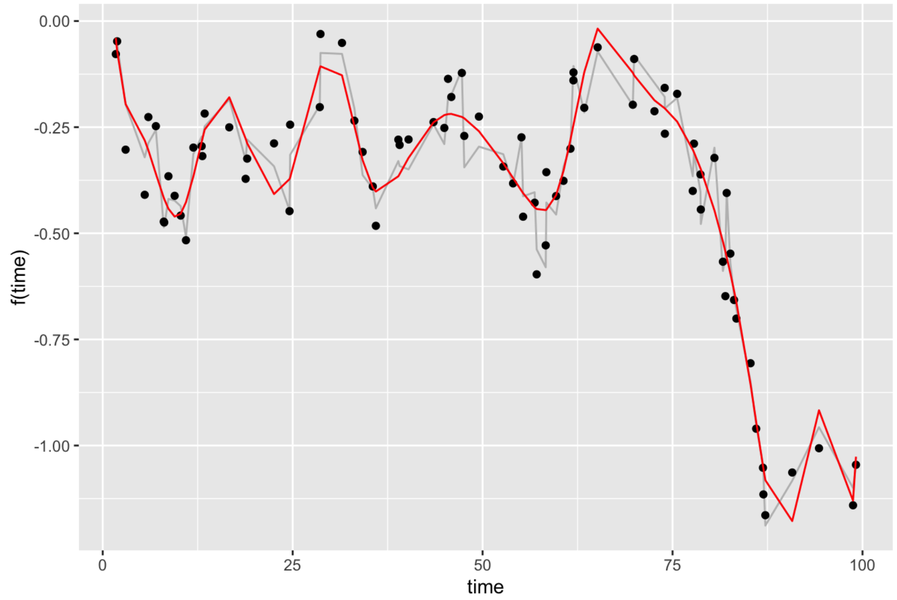
- **Purpose**
  - Develop methods for high-dimensional functional data regression.
    - Variables selection and parameter estimation simultaneously;
    - Adopted for ultra-high dimensional functional dataset.

Functional Variables

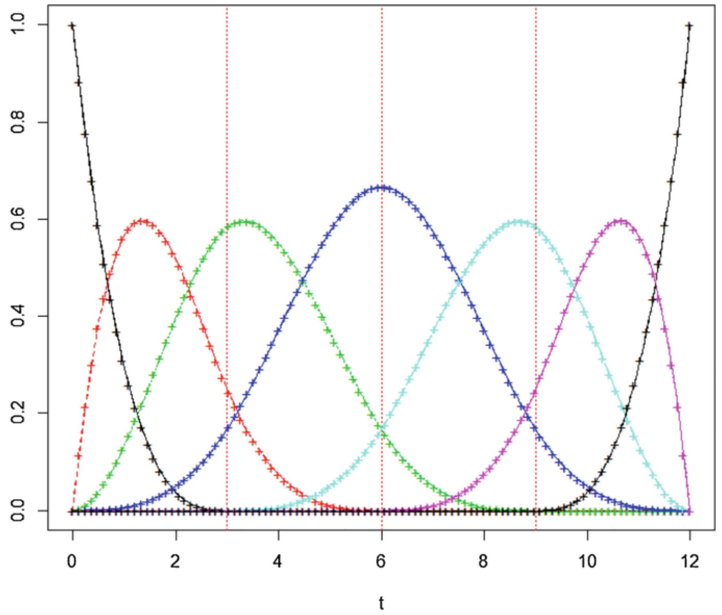
- Functional variables, are smooth, square integrable, random functions observed on a continuum (e.g., time, length, width, etc.), with minimal time-related autocorrelations.

- In order to fit the scatter points into random functions, the functional predictors X(t), are often estimated via smoothing, which usually involves using of basis-functions.

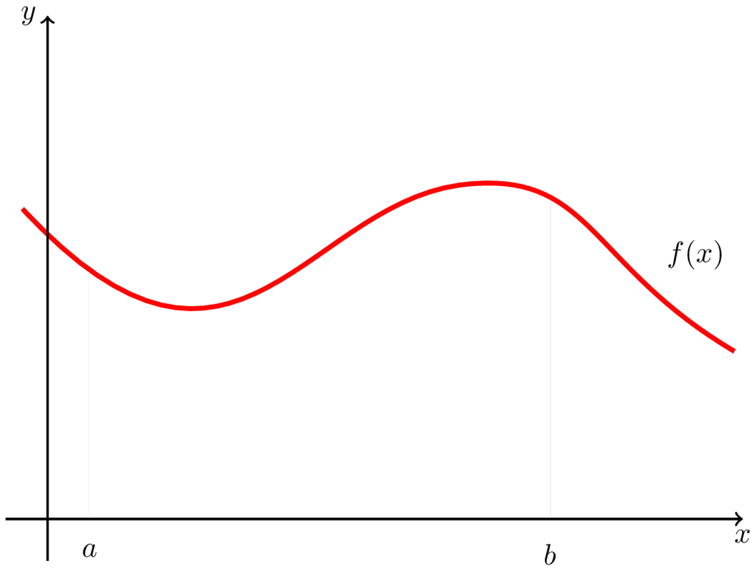
Discrete scatter points **X**



Smoothing basis $\Phi(t)$

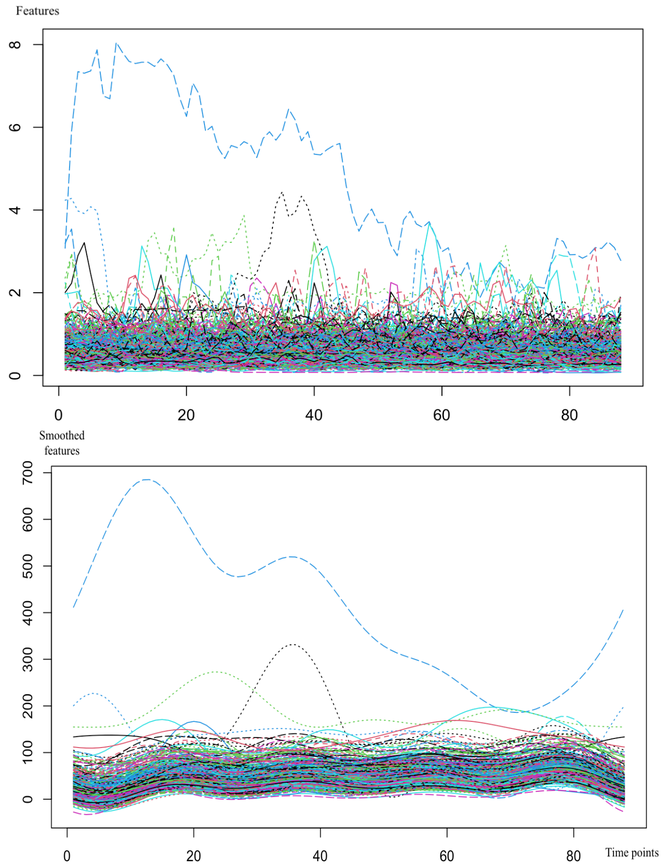


Random function $X_i(t) + \epsilon_i$



Functional regression model

The functional regression model find relationships between a functional variable, X(t), and the scalar response Y.



- $$Y_i = \alpha + \int_\tau X_i(t)\beta(t)dt + \epsilon_i, \quad i = 1,...,n,\ \tau = (t_1, \ldots, t_p),$$

Functional predictor: $$X_i(t) \approx \sum_{k=1}^{l} c_{i_k}(t)\phi_k(t) = c_i\Phi(t)$$

Regression coefficients: $$\beta(t) \approx \sum_{k=1}^{l} b_k\phi_k(t) = b\Phi(t)$$

Discretized functional regression model: $$Y_i \approx \sum_{k=1}^{l} c_{ik}\Phi(t)\Phi(t)^T b_k^T + \epsilon_i = z_i b^T + \epsilon_i$$

We can then estimate the regression coefficients of the discretized functional regression.

Multiple Functional Regression Model

Consider a multiple functional regression model with J functional predictors and a scalar response:

$$Y_i = \alpha + \sum_{j=1}^{J} \int_{\tau} X_{ij}(t)\beta_j(t)dt + \epsilon_i, \quad i = 1, ..., n \tag{1}$$

J = dimensionality; n = sample size.

Smooth $X_j(t)$ and $\beta_j(t)$ using basis functions, and use the Riemann integration techniques and reformulate the problem into typical linear form:

$$Y = Z_j^T b_j + \epsilon, \quad j = 1, \ldots, J \tag{2}$$

$Y = (Y_1, \ldots, Y_n)^T$, $Z_j = (z_1, ..., z_n)^T$, $z_i = (1, c_{i_1}^T J_{\phi_1}, ..., c_{i_J}^T J_{\phi_J})^T$, $b_j = (b_1, b_2, ..., b_{l_j})^T$, $J_{\Phi_j} = \int_{\tau} \phi_j(t)\phi_j^T(t)dt$.

Dimensionality in Functional Regression Model

$$Y_i = \alpha + \sum_{j=1}^{J} \int_{\tau} X_{ij}(t)\beta_j(t)dt + \epsilon_i, \quad i = 1, ..., n \tag{1}$$

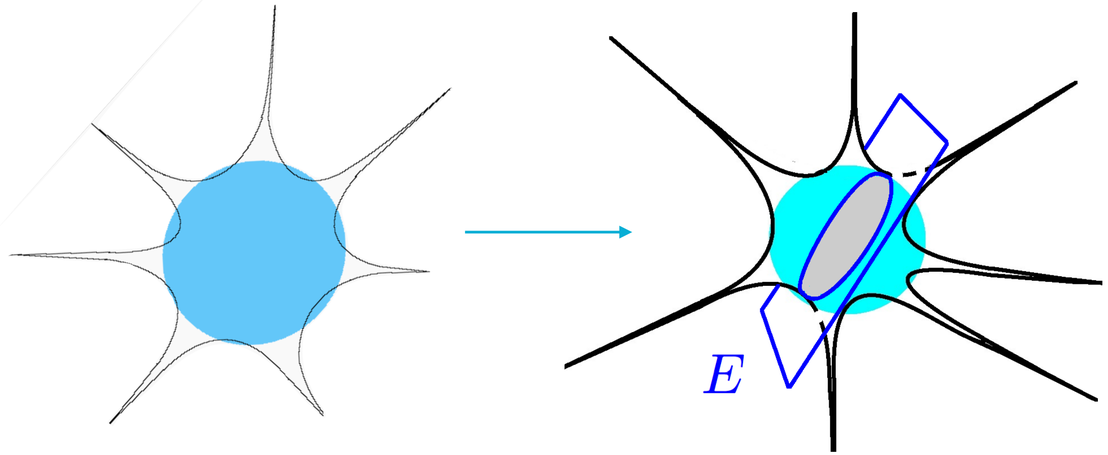
In typical regression problem, we do require n > J. In more tough conditions, we have n ~ J, which could be handled with penalized regression (Lasso, SCAD).

We would like to fit the above model under following assumptions:

- Ultrahigh dimensionality: $J = \exp\{O(n^{\xi})\}$ for some $\xi > 0$.
- Sparsity: $M_*^f = \{X_j(t) : ||\beta_j(t)||_1 \neq 0, 1 \leq j \leq J\}$be the true model. Let $s = |M_*^f|$, we have that $s \ll J$.

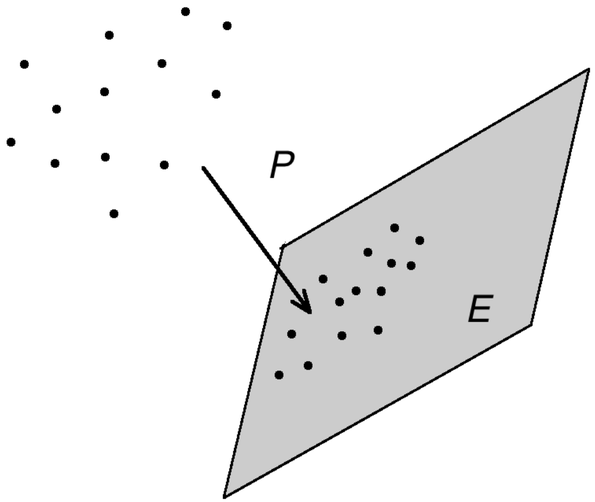
High Dimensional Statistic Modeling: Curse of dimensionality

J > n or J >> n, the data distribution is sparse; direct reduce dimensionality becomes difficult.



J < n, the data distribution is dense,

we may reduce dimension via projection.

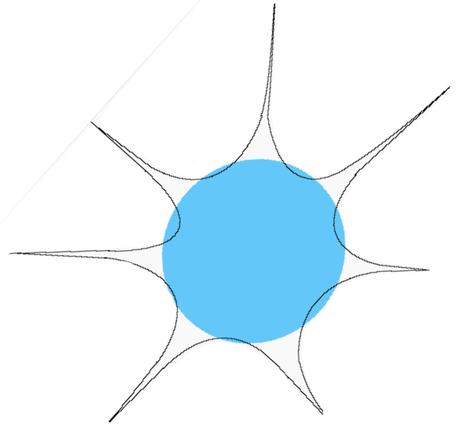


An intuitive, hyperbolic picture of a convex body in R^n. The bulk is a round ball that contains most of the volume.

(Figures from the *HDP book*, Vershynin, 2019)

The SURE Independent Screening for Multivariate Variables (Fan and Lv, 2008)

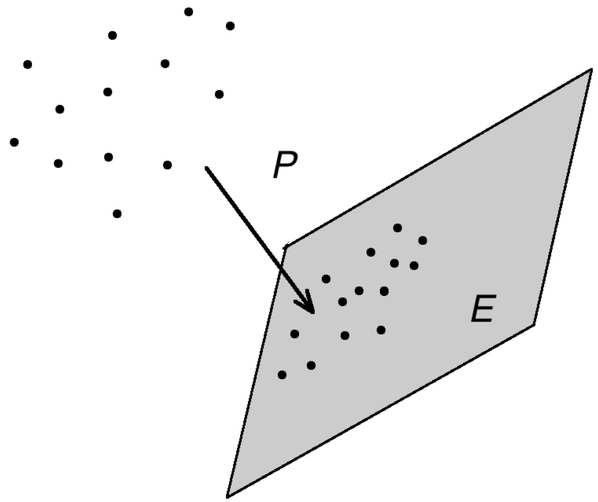
Step 1: Correlation Learning *



compute $\omega_j = X_j^T y,$ and rank the correlations in discreasing order

$$M_d = \{1 \leq j \leq J : |\omega_j| \text{ is among the first } d \text{ largest of all}\}.$$

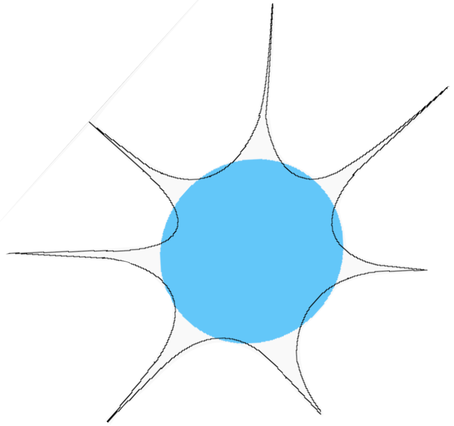
Step 2: Penalized multiple regression
(e.g., LASSO, SCAD)



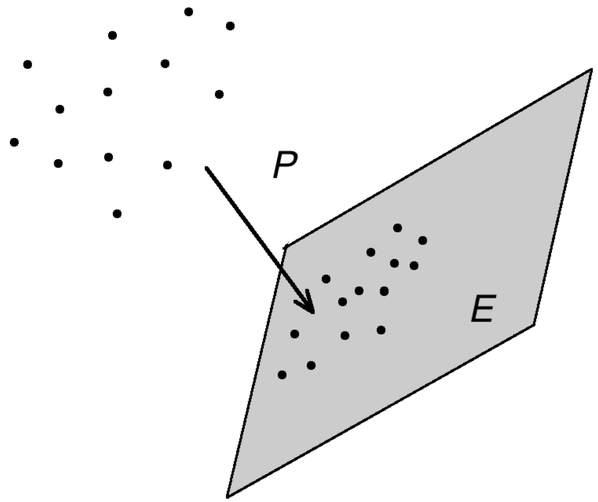
* : The Sure independent screening method has been shown to enjoy sure screening property and oracle property (Fan and Lv, 2008).

The SURE Independent Screening for Multiple Functional Regression

Step 1: Functional correlation learning *



Step 2: Penalized multiple functional regression
(e.g., Functional LASSO (flasso))



compute $\omega_j^f = Z_j^T y,$ and choose a submodel of size d :

$$M_d = \{1 \le j \le J : |\omega_j^f| \text{ is among the first } d \text{ largest of all}\}.$$

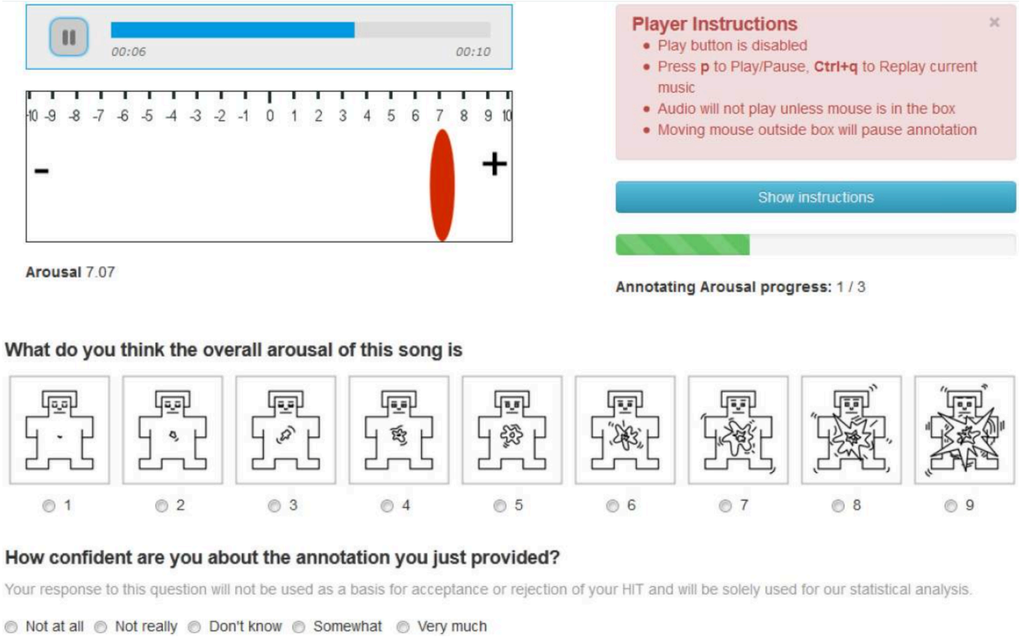
* : variations of functional correlation learning method in Yuan and Billor, (2024).

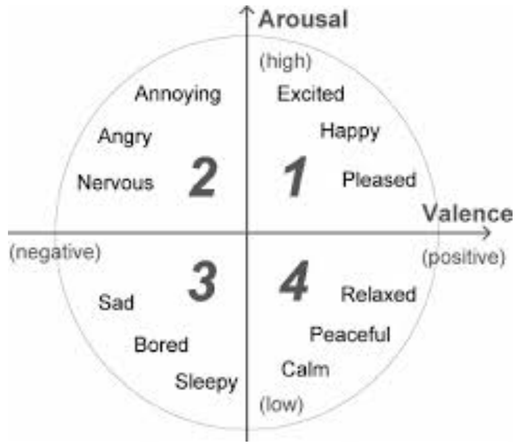
Data Application: the DEAM dataset

(MediaEval Database for Emotional Analysis in Music)

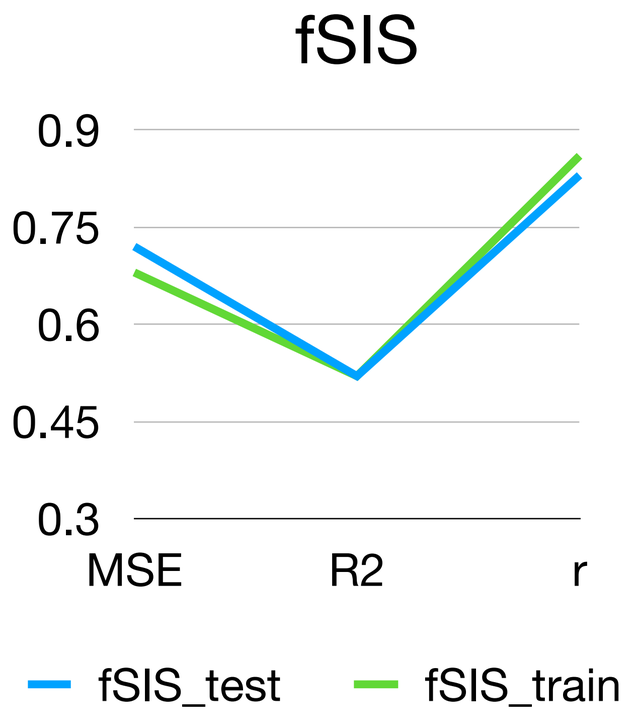
- Containing more than 1800 songs and their emotional annotations.

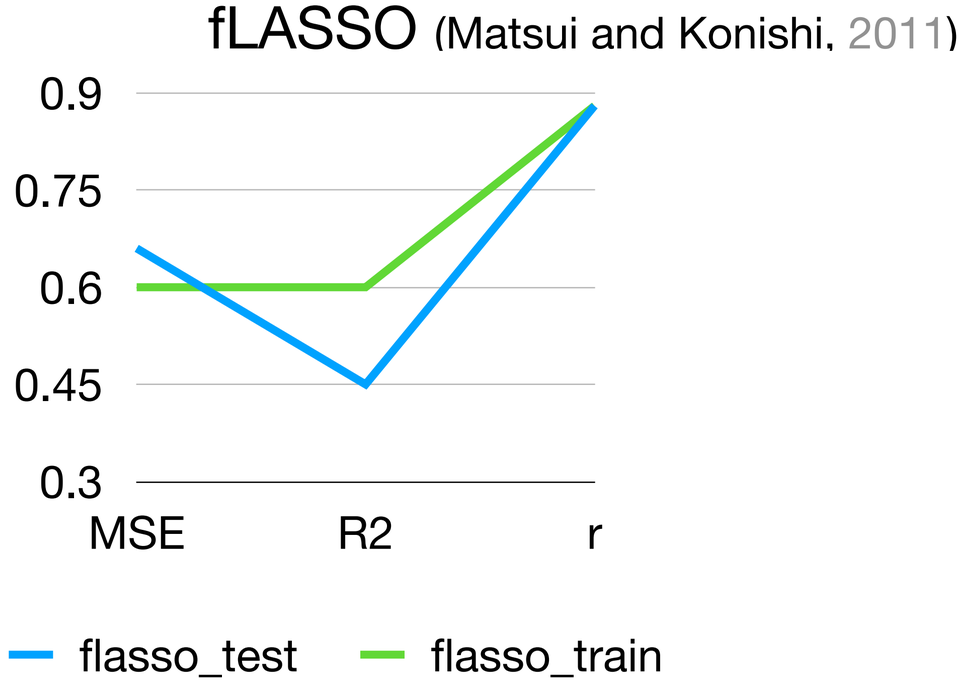
- Predictors:
  - 260 functional features extracted for each song from the OpenSMILE.
- Response:
  - Human rated emotional response (i.e., arousal and valance), for each song.
- Training set: MTurk rated, n = 485.
- Testing set: Researcher rated, n = 58.





Results: predicting Arousal



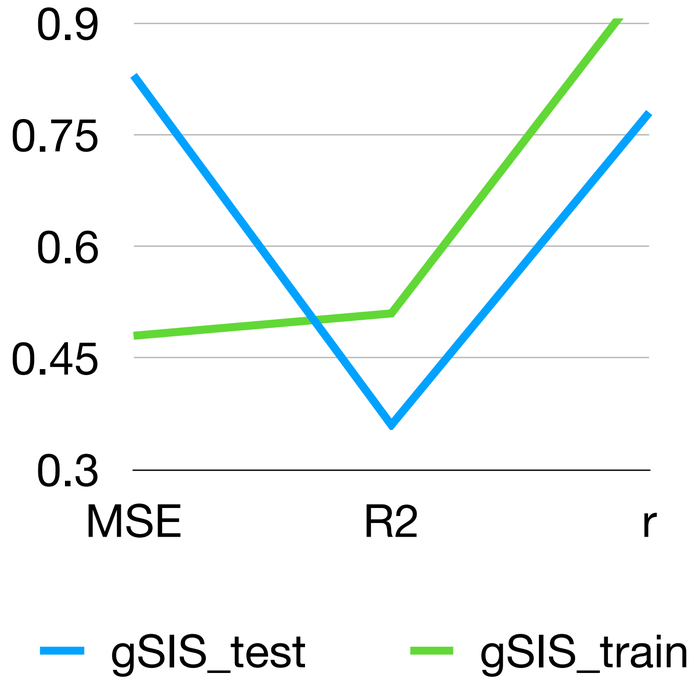


- Yuan.Y., Billor, N. (2024). SURE Independent Screening for Functional Regression Model. Communications in Statistics - Simulation and Computation,1-20.

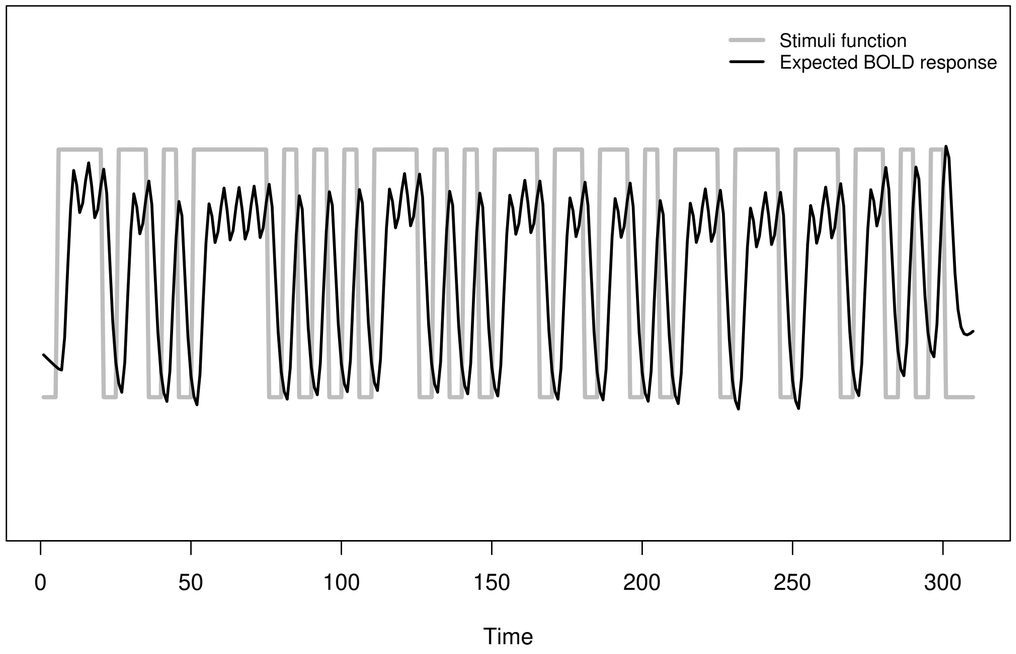
Results: predicting Arousal

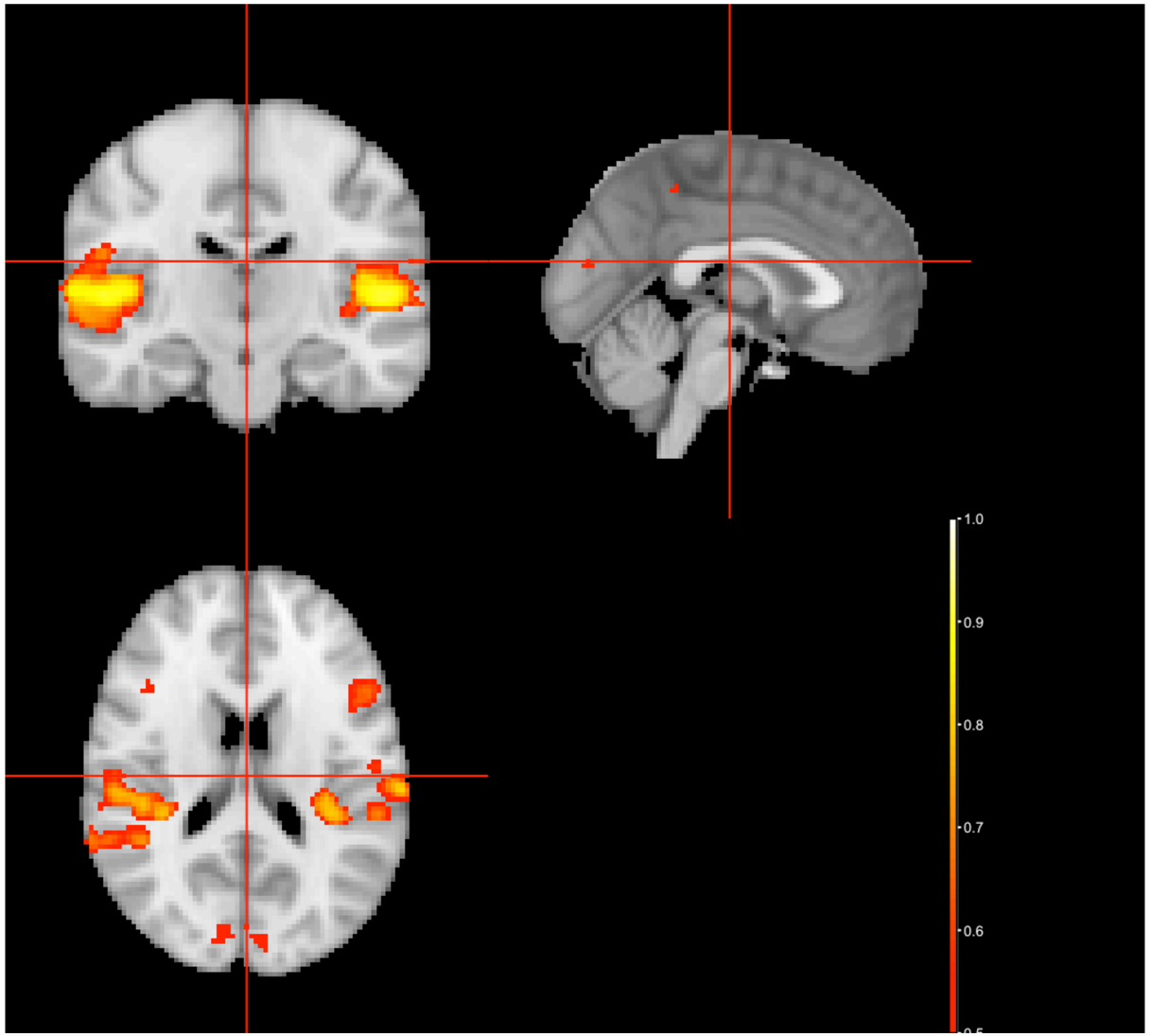
gLASSO (Yuan and Lin, 2006)

gSIS (Qiu, 2019; Qiu and Ahn, 2020)



Data application II: fSIS for fMRI dataset





Regions of activation of human voices vs. non voice sounds achieved for a single subject, achieved with the ***functional correlation learning*** approach. Data from the “voice localizer” study in *OpenNeuro* (Pernet et al., 2015; Gorgolewski et al., 2017).

Reference

- Aljanaki, A., Yang, Y. H., and Soleymani, M. (2017). Developing a benchmark for emotional analysis of music. PloS one, 12(3)
- Fan J, Lv J. Sure independence screening for ultrahigh dimensional feature space. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2008;70(5):849–911.
- Matsui, H., and Konishi, S. (2011). Variable selection for functional regression models via the L1regularization. Computational Statistics and Data Analysis, 55(12), 3304-3310
- Qiu D, Ahn J. Grouped variable screening for ultra-high dimensional data for linear model. Computational Statistics & Data Analysis. 2020;144:106894
- Qiu D. Grouped variable screening for ultrahigh dimensional data under linear model[dissertation]. University of Georgia; 2016.
- Vershynin, R. (2018). High-dimensional probability: An introduction with applications in data science (Vol. 47). Cambridge university press.
- Yuan, M., & Lin, Y. (2006). Model selection and estimation in regression with grouped variables. Journal of the Royal Statistical Society Series B: Statistical Methodology, 68(1), 49-67.
- Cyril R Pernet, Phil McAleer, Marianne Latinus, Krzysztof J Gorgolewski, Ian Charest, Patricia EG Bestelmeyer, Rebecca H Watson, David Fleming, Frances Crabbe, Mitchell Valdes-Sosa, et al. The human voice areas: Spatial organization and inter-individual variability in temporal and extra-temporal cortices. Neuroimage, 119:164–174, 2015.
- Krzysztof Gorgolewski, Oscar Esteban, Gunnar Schaefer, Brian Wandell, and Russell Poldrack. Openneuro- a free online platform for sharing and analysis of neuroimaging data. Organization for Human Brain Mapping. Vancouver, Canada, 1677, 2017.

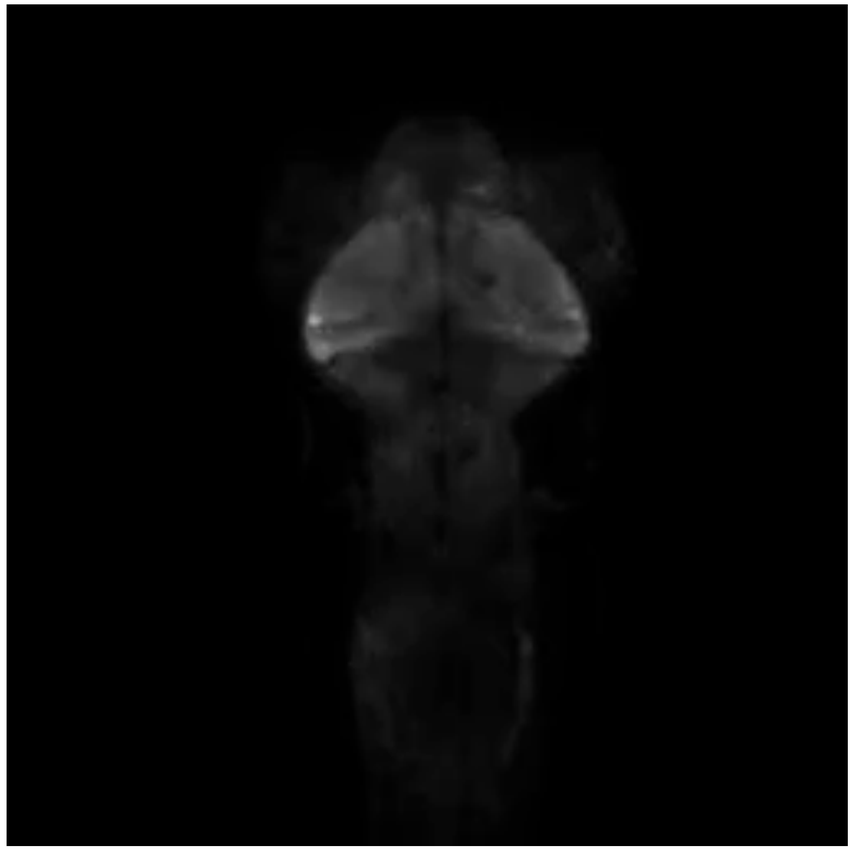
Study2

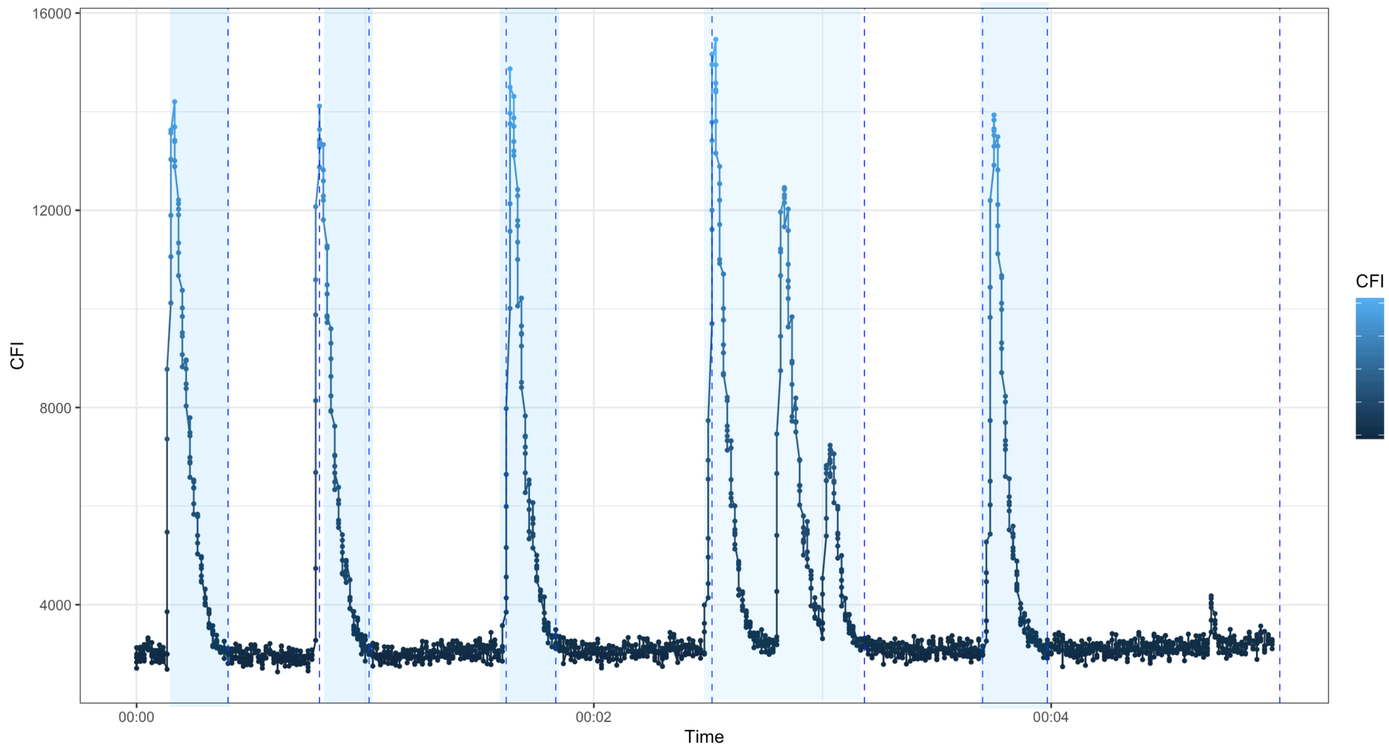
- Computational feature extraction for zebrafish larvae Calcium Imaging video during PTZ induced epileptic seizure

- **Purpose**
  - Image/video reconstruction
  - Data driven modeling

Data Description

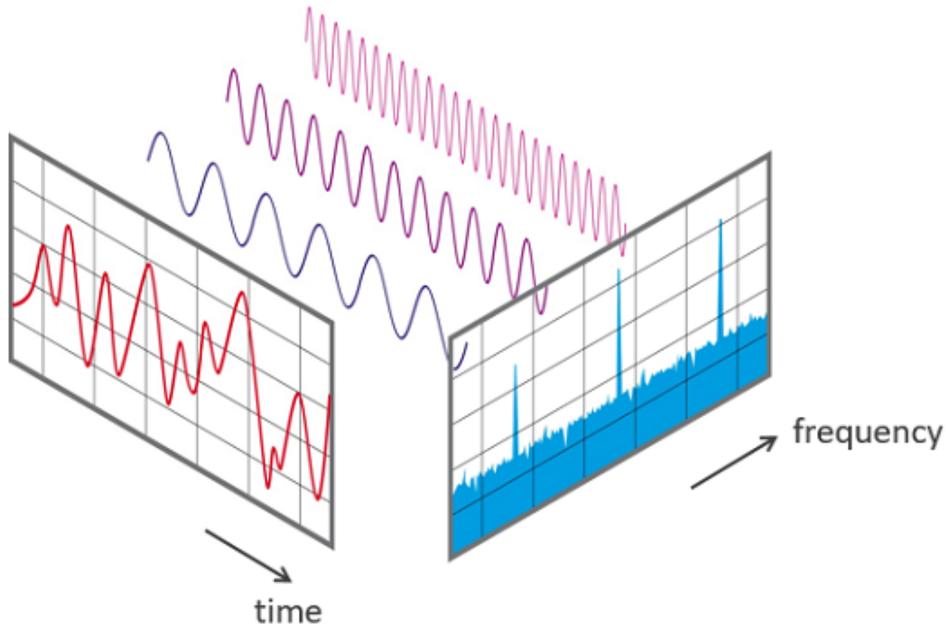
- Calcium imaging video of zebrafish larvae during PTZ induced seizure (Zheng et al., 2019).
- The resolution of the original video data is 256x256, with a total of 2^16 pixels per frame. The data was recorded in N=2000 time frames with a sampling frequency of fs=6.7.



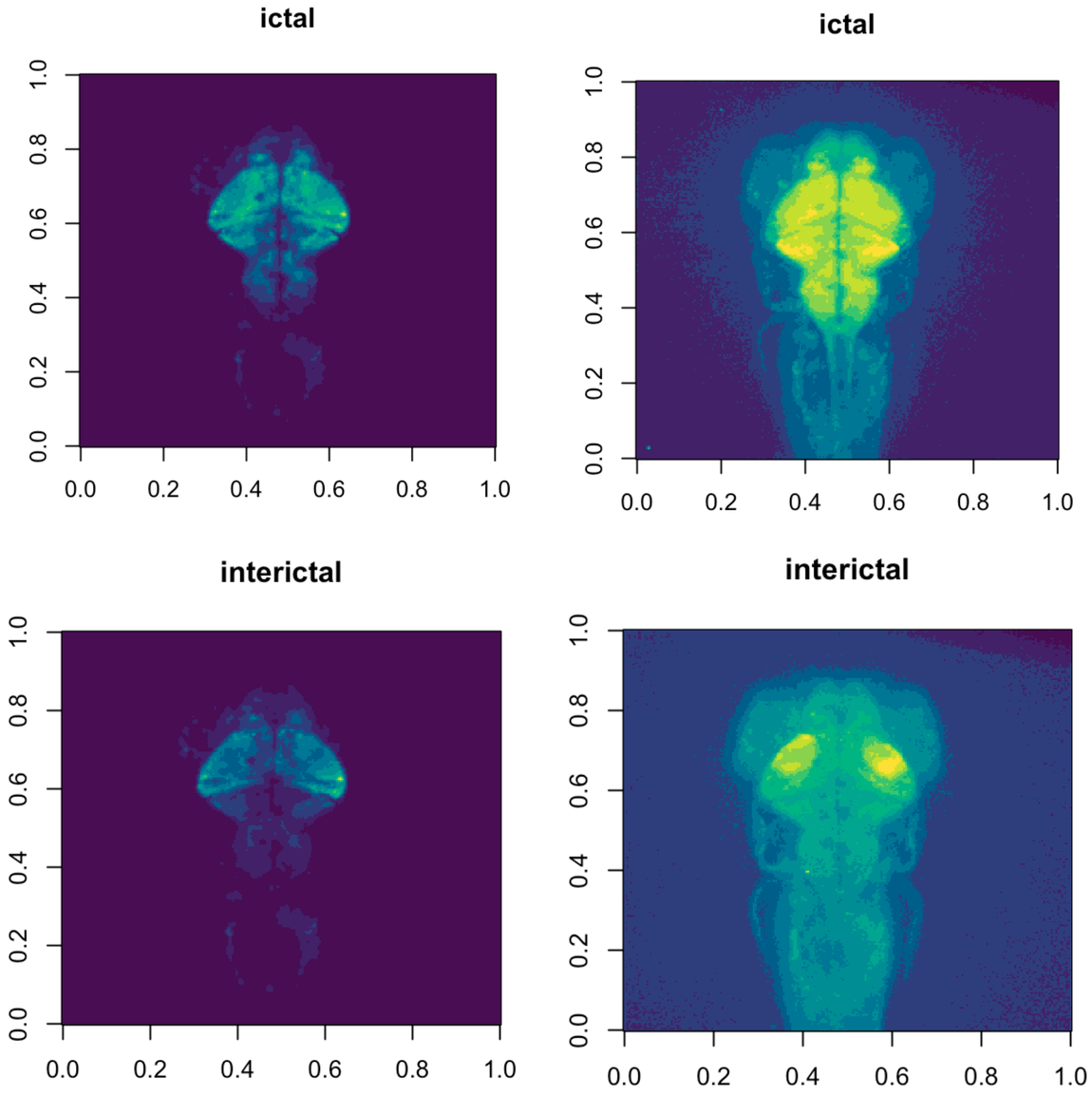


Why feature extracting?

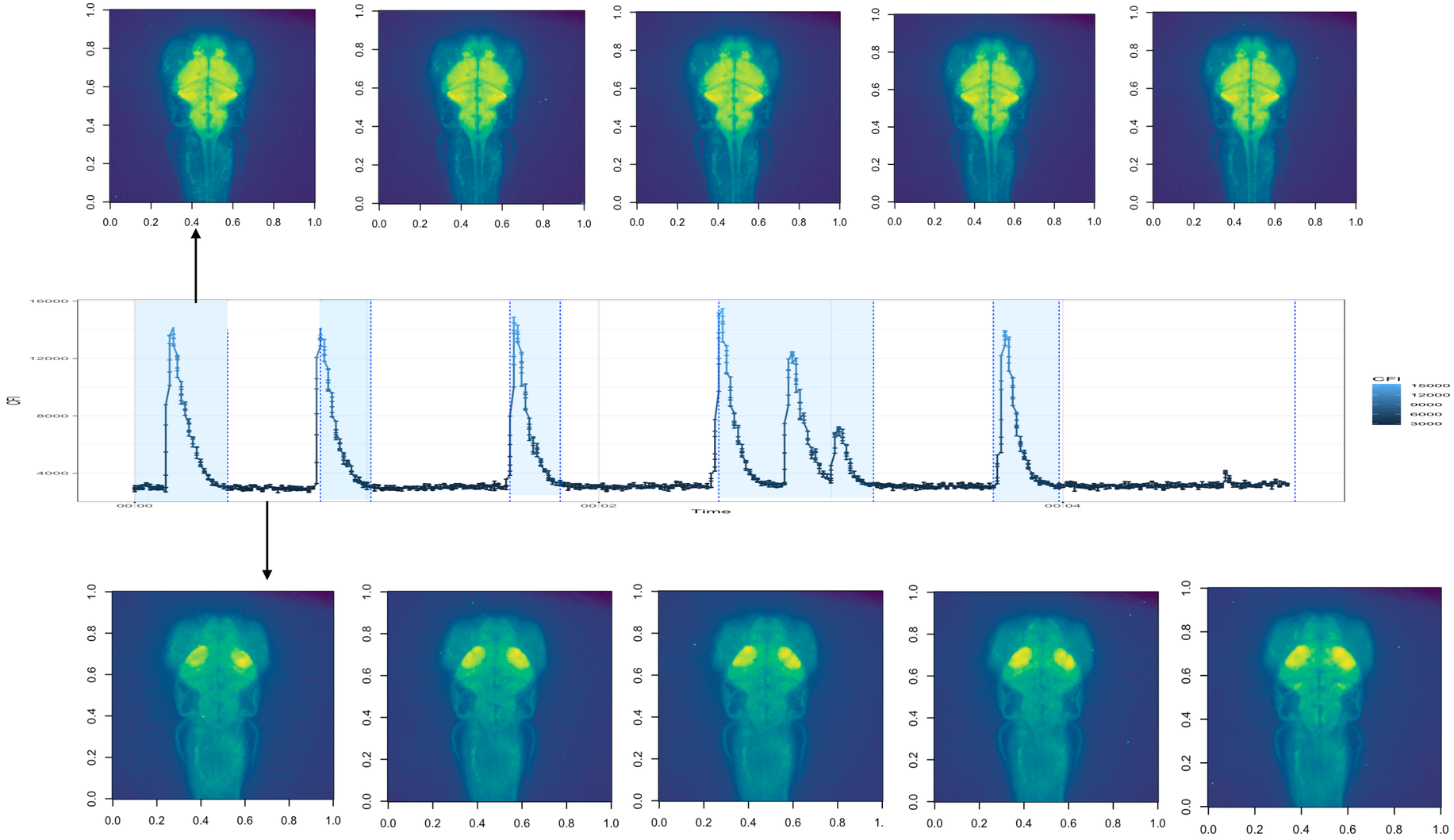
- Non stationary, dynamic.
- Not ready for statistical modeling.
- Features: statistical characteristics of time-frequency distribution of neural signals.



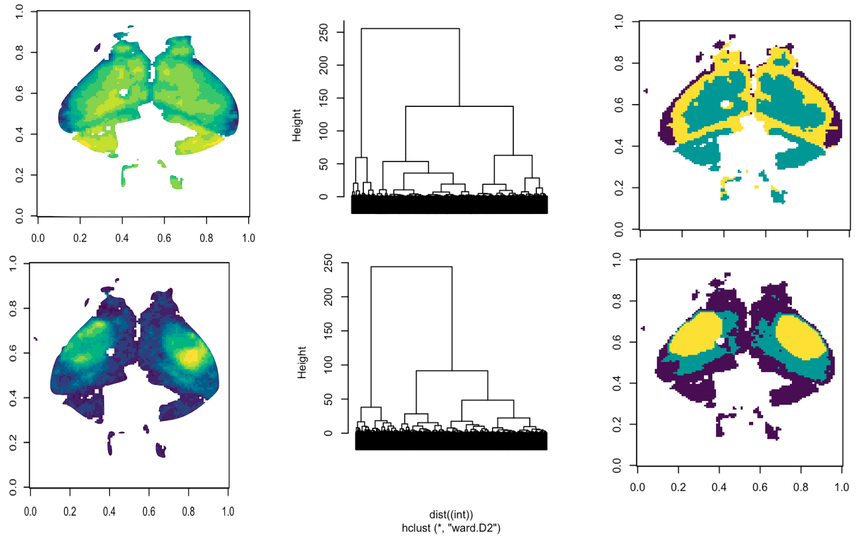
Feature extracting: Average vs. Fluctuation (i.e., measure of dispersion.)

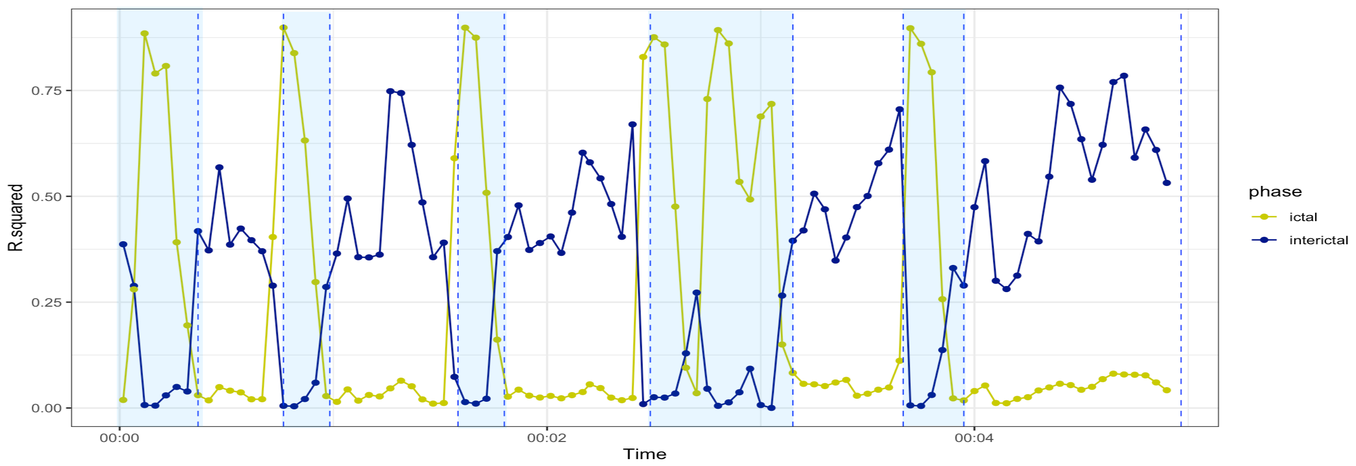


The fluctuation across seizure: the patterns



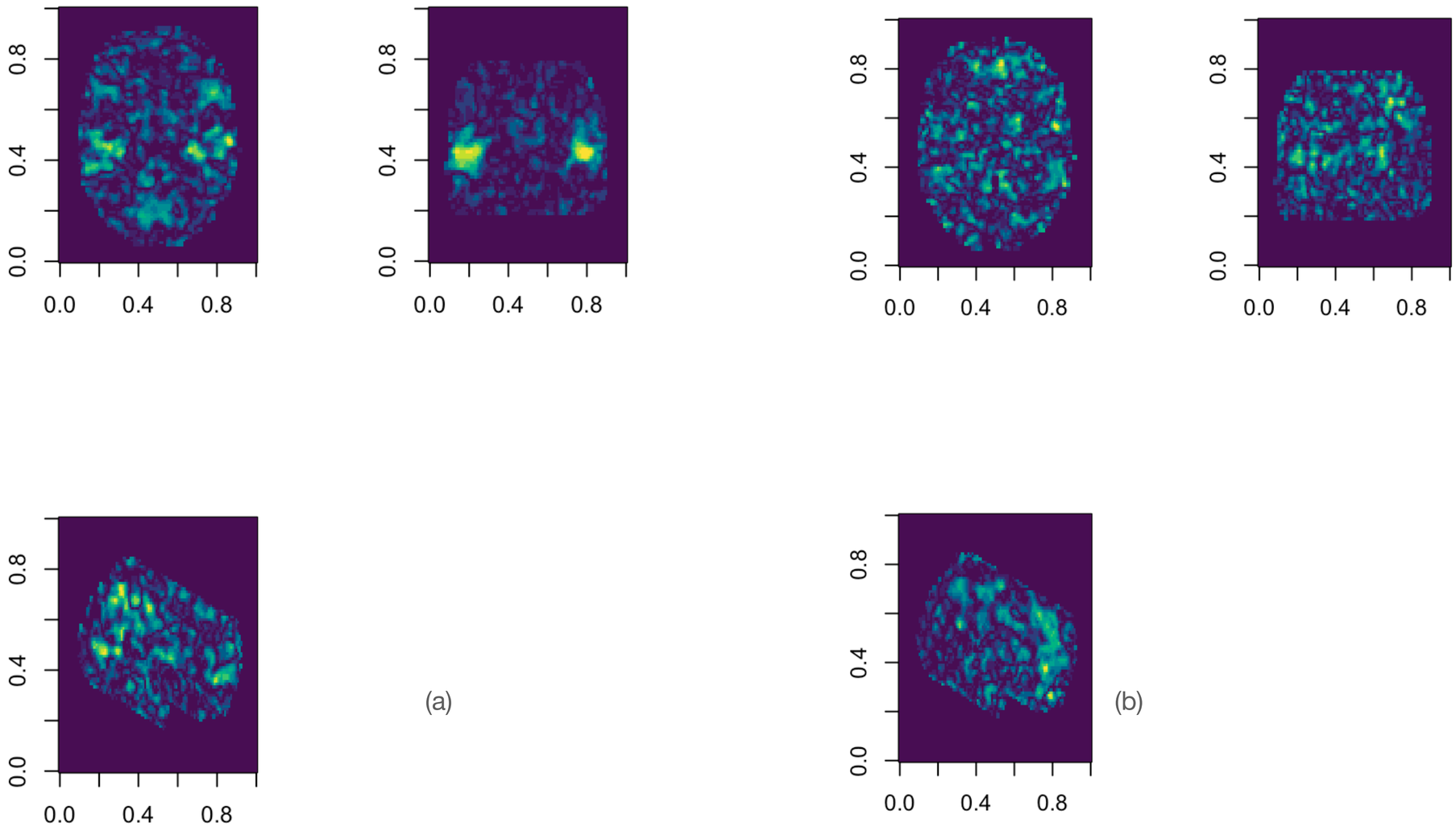
The patterns and anti-correlational network





There exist distinct neuronal clusters across ictal and inverictal periods,
With recurrent, stable organization of patterns.

Data Example II: Image Reconstruction in fMRI Data



The same feature reconstruction technology applied for fMRI data, in voices condition (a) and non-voices sounds condition (b). Data from OpenNeuro (Pernet et al., 2015; Gorgolewski et al., 2017).

References

- Jingyi Zheng, Fushing Hsieh, and Linqiang Ge. “A data-driven approachto predict and classify epileptic seizures from brain-wide calcium imaging video data”. In: IEEE/ACM transactions on computational biology and bioinformatics 17.6 (2019), pp. 1858–1870.

- Yuan.Y. On Functional Sure Independence Screening (fSIS) and Dynamic Spectral Featuring Extraction for Neuroscience Data [dissertation]. Auburn University; 2021.

- Cyril R Pernet, Phil McAleer, Marianne Latinus, Krzysztof J Gorgolewski, Ian Charest, Patricia EG Bestelmeyer, Rebecca H Watson, David Fleming, Frances Crabbe, Mitchell Valdes-Sosa, et al. The human voice areas: Spatial organization and inter-individual variability in temporal and extra-temporal cortices. Neuroimage, 119:164–174, 2015.

- Krzysztof Gorgolewski, Oscar Esteban, Gunnar Schaefer, Brian Wandell, and Russell Poldrack. Openneuro- a free online platform for sharing and analysis of neuroimaging data. Organization for Human Brain Mapping. Vancouver, Canada, 1677, 2017.

Outline

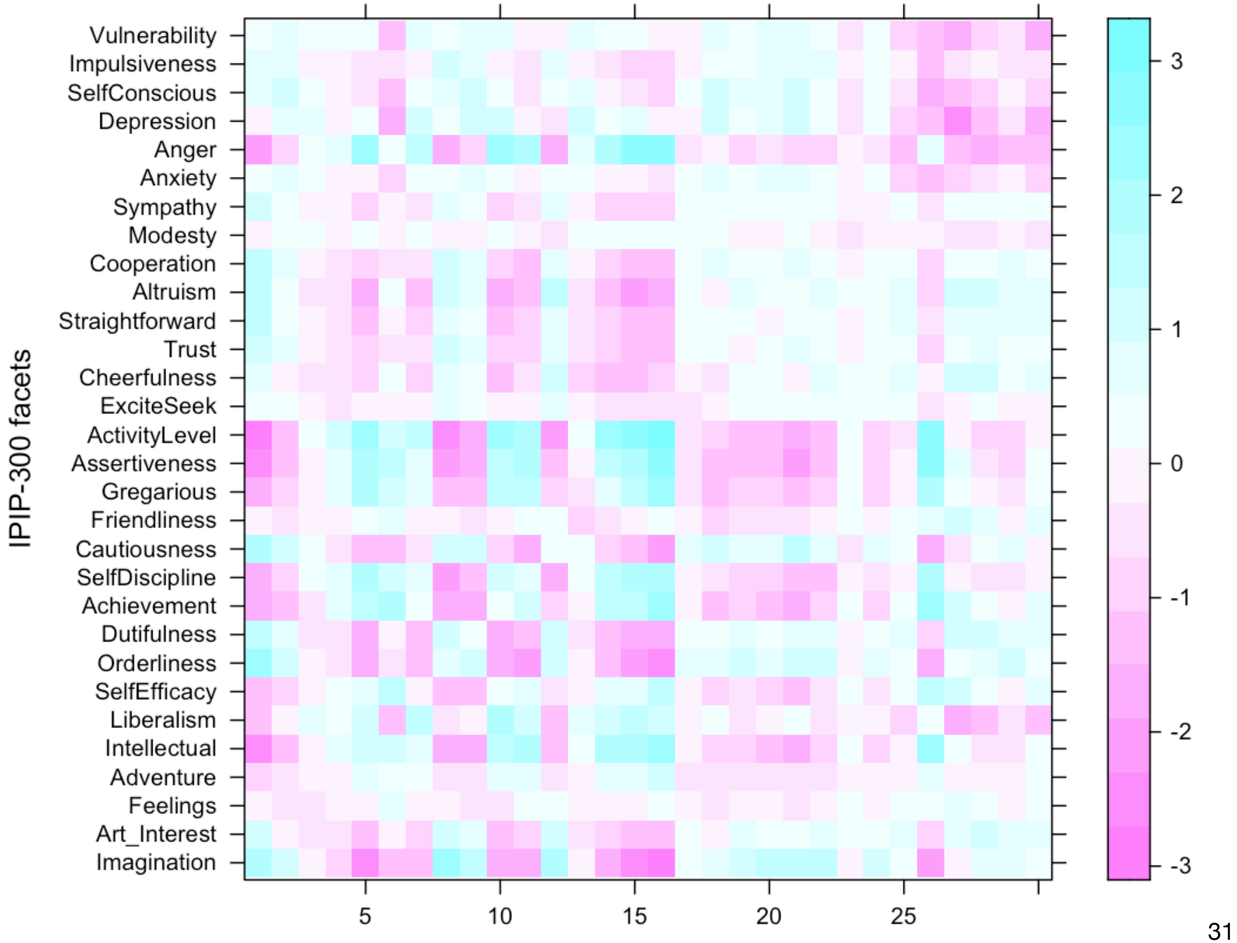
- Introduction
- Past findings
  - Study 1: statistic method
  - Study 2: data engineering method
- Future plan
  - Neuroscience research: clinical neural-behavioral research

Towards an integrated neural-behavioral science

- Group-level behavioral science vs. individual-level neuroscience
- Exploring individualized behavioral sicence
    - Interview, case study, natural observation, ….
    - Quantitative within-person modeling

Pilot study: functional connectivity in psychometric modeling

Multiple Regression → School Behaviors



Variable selection/screening

Penalized regression → School Behaviors

Whose GPA would be higher?

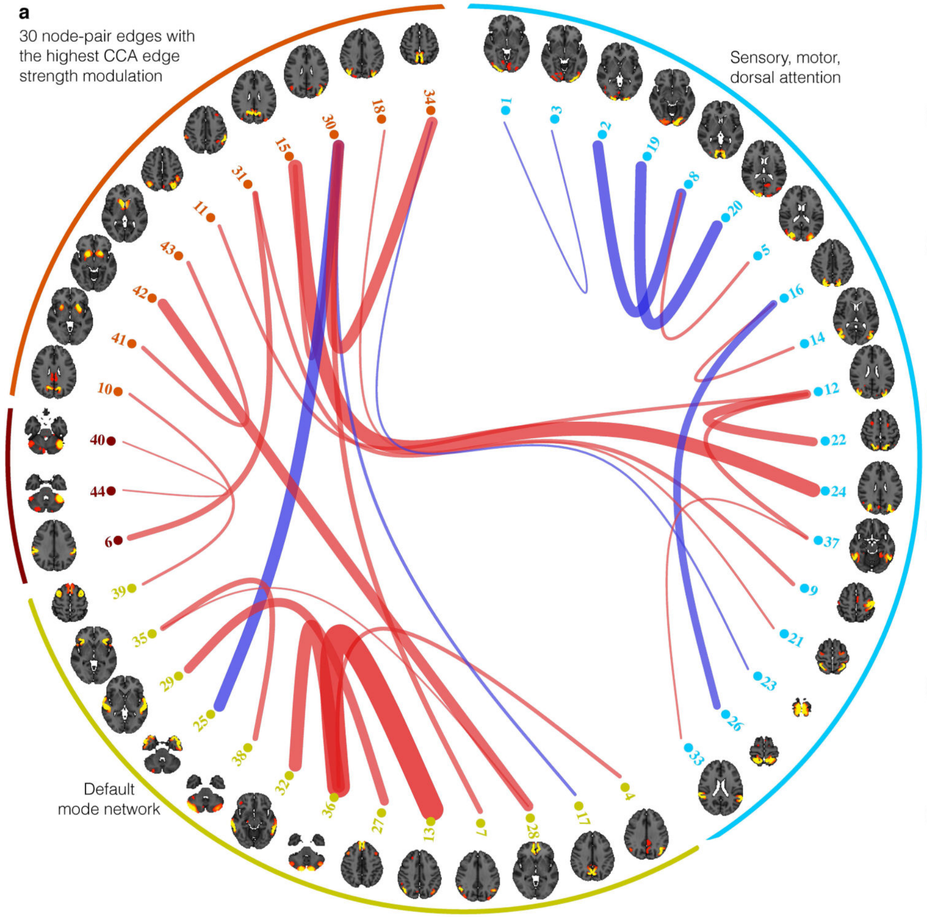
- *Someone who believe they could obtain more achievement via cooperation.*
- *Someone who believe they will be more self-conscious if they are also Imaginative.*
- *Someone who thought they could be more friendly by having art interests.*

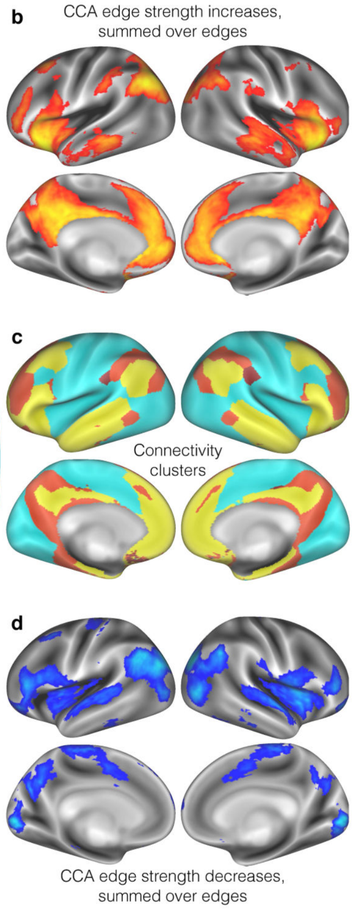
Whose Friends rate them higher?

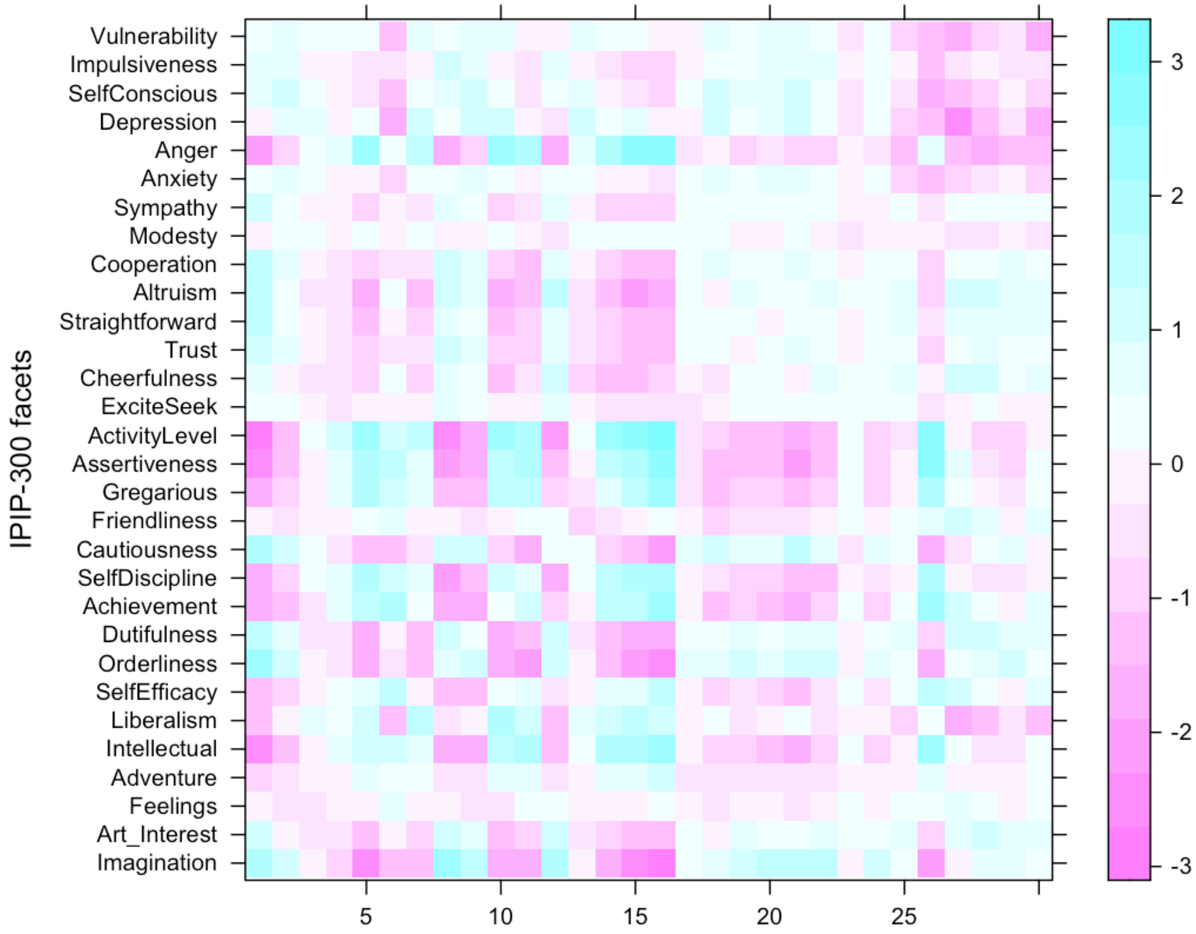
- Show emotions while at the same time cautiousness;
- Achieving while keep cheerfulness.
- Friendly and show sympathy.

Towards neural-behavioral science: can we do more?

- Smith.S et al., 2015







References

- Beaty, R. E., Kaufman, S. B., Benedek, M., Jung, R. E., Kenett, Y. N., Jauk, E., Neubauer, A. C., & Silvia, P. J. (2016). Personality and complex brain networks: The role of openness to experience in default network efficiency. Human brain mapping, 37(2), 773–779.

- Goyal, N., Moraczewski, D., Bandettini, P. A., Finn, E. S., & Thomas, A. G. (2022). The positive–negative mode link between brain connectivity, demographics and behaviour: A pre-registered replication of smith et al.(2015). Royal Society Open Science, 9(2), 201090.

- Simon, S. S., Varangis, E., & Stern, Y. (2020). Associations between person- ality and whole-brain functional connectivity at rest: Evidence across the adult lifespan. Brain and behavior, 10(2), e01515.

- Smith, S. M., Nichols, T. E., Vidaurre, D., Winkler, A. M., Behrens, T. E., Glasser, M. F., Ugurbil, K., Barch, D. M., Van Essen, D. C., & Miller, K. L. (2015). A positive-negative mode of population covariation links brain connectivity, demographics and behavior. Nature neuroscience, 18(11), 1565–1567.

- Toschi, N., Riccelli, R., Indovina, I., Terracciano, A., & Passamonti, L. (2018). Functional connectome of the five-factor model of person- ality. Personality Neuroscience, 1, e2.

- Van Schuerbeek, P., Baeken, C., & De Mey, J. (2016). The heterogeneity in retrieved relations between the personality trait ‘harm avoidance’and gray matter volumes due to variations in the vbm and roi labeling processing settings. PloS one, 11(4), e0153865.

- Thank you!