

Data-driven Methods for Functional Connectomes Using Optimal Transport

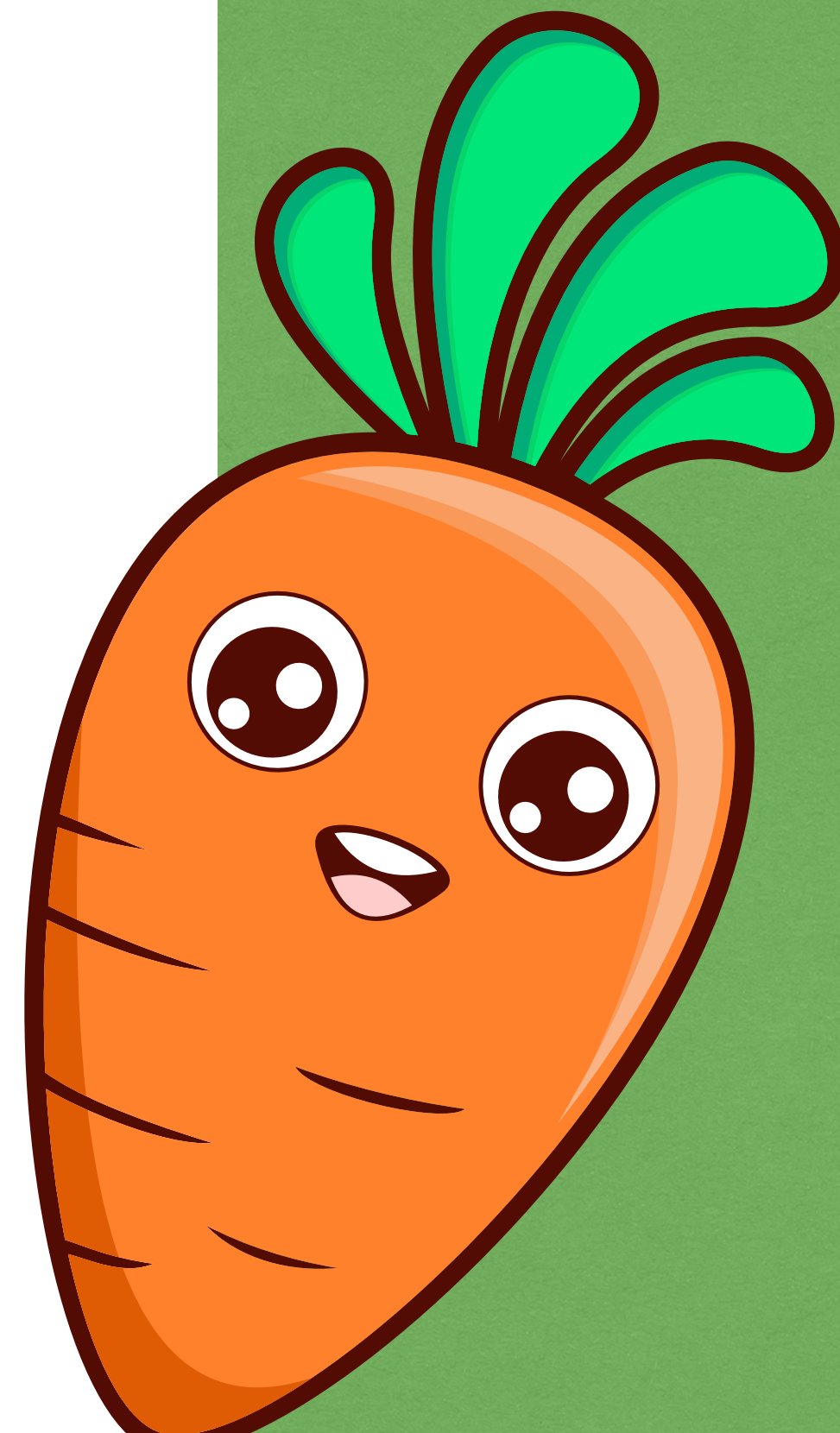
Javid Dadashkarimi

Advisor: Dustin Scheinost, Amin Karbasi
Committee: Xenophon Papademetris,
James Duncan

javid.dadashkarimi@yale.edu



**Computer Science Department
Yale University**



2017-present



Yale University – New Haven, US

Ph.D. in Computer Science

En Route MSc, MPhil (2019)

Mentors: Dustin Scheinost and Amin Karbasi

Thesis: *Data-driven mappings between functional connectomes using optimal transport*

2008-2015



University of Tehran – Tehran, Iran

MEng in Software Engineering

BEng in Software Engineering

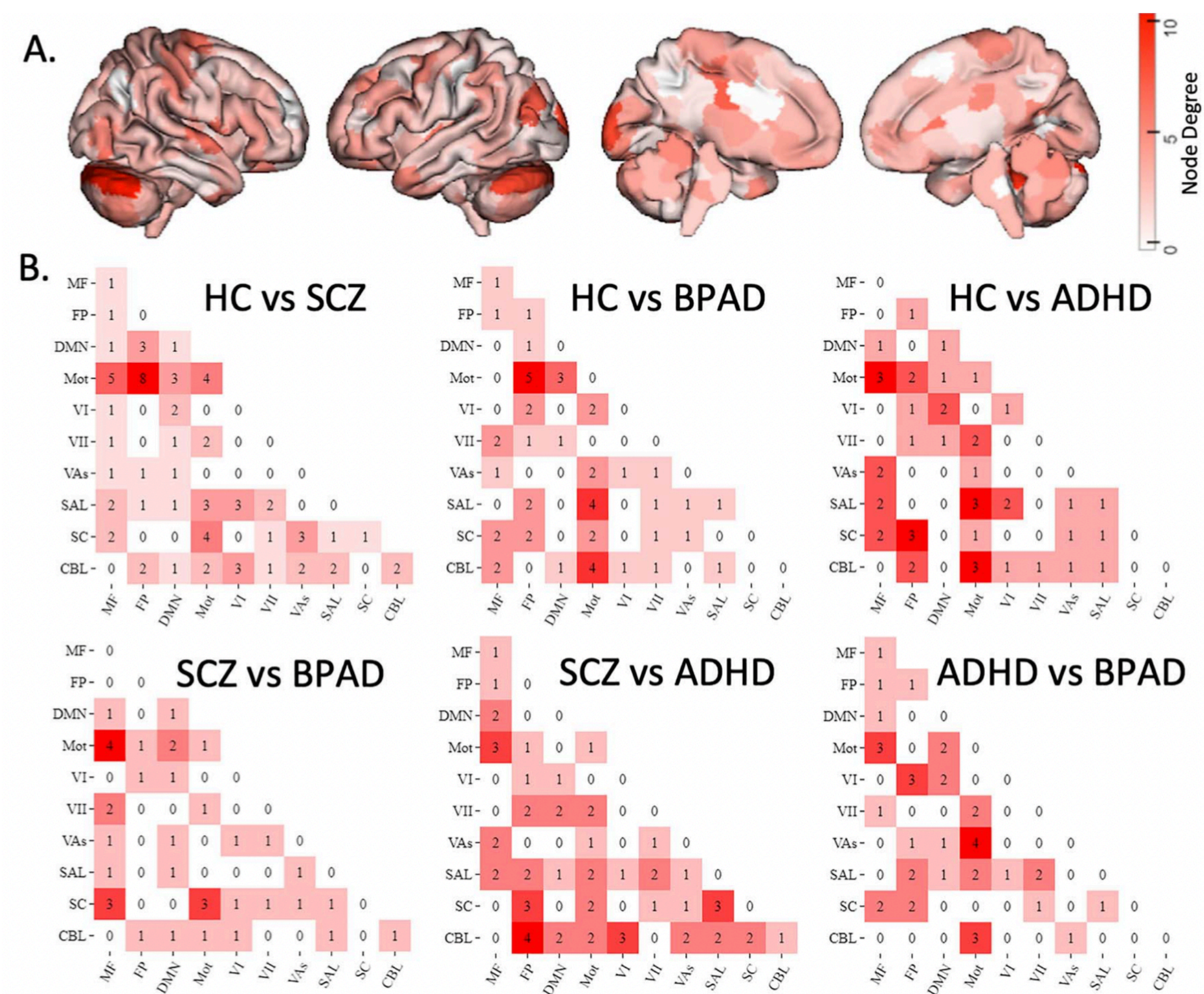
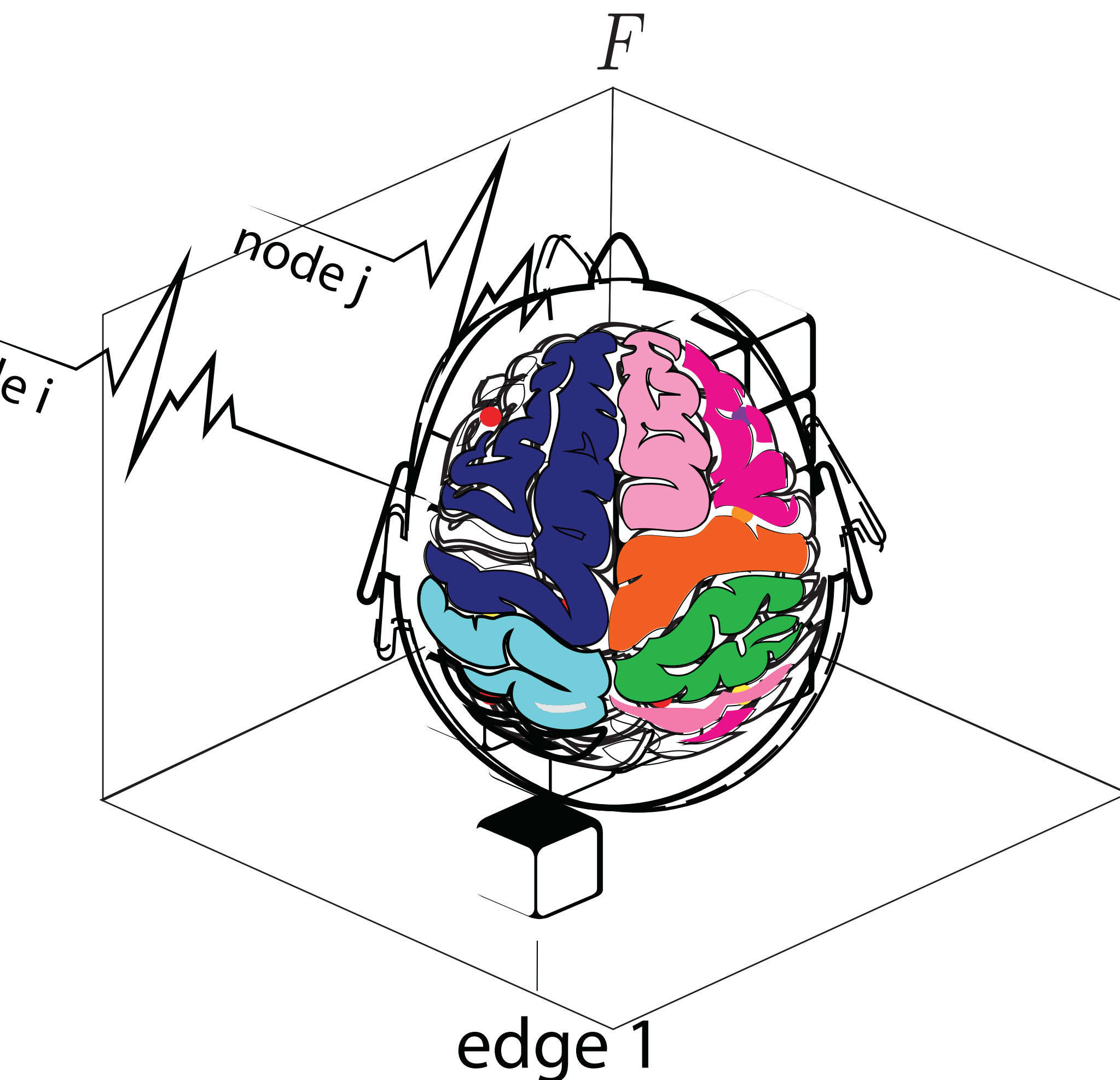
Mentors: Azadeh Shakery and Heshaam Faili

Thesis: *Dictionary-based Cross-lingual Information Retrieval*

Short Bio

- A connectome—a matrix describing the connectivity between any pair of brain regions—is a popular approach in neuroscience to study the functional organization of the brain.
- They are created by parcellating the brain into distinct areas using an atlas and estimating the connections between these regions.
- Applications: To study individual differences in brain function, associating brain and behavior, and understanding brain alterations in neuropsychiatric disorders.

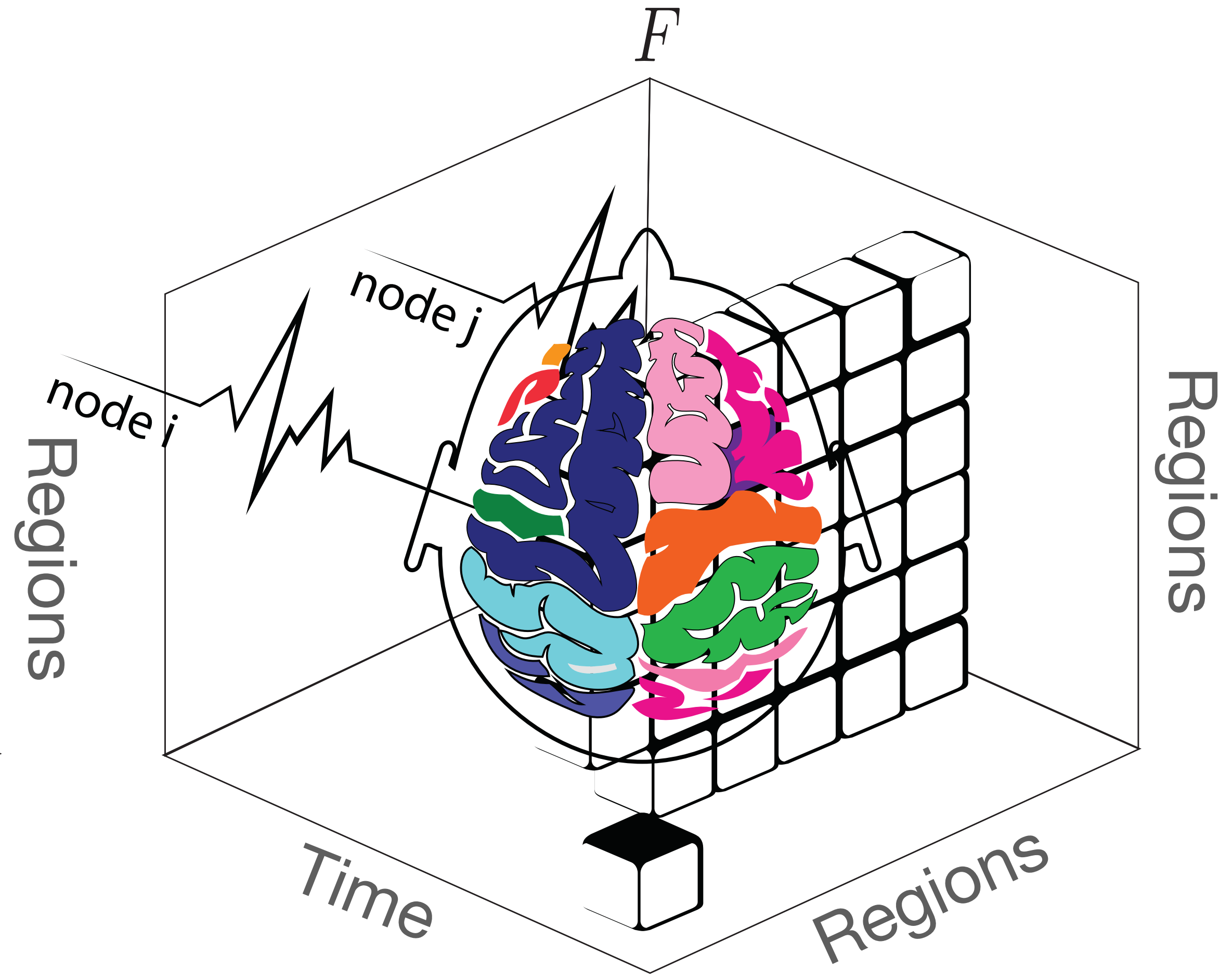
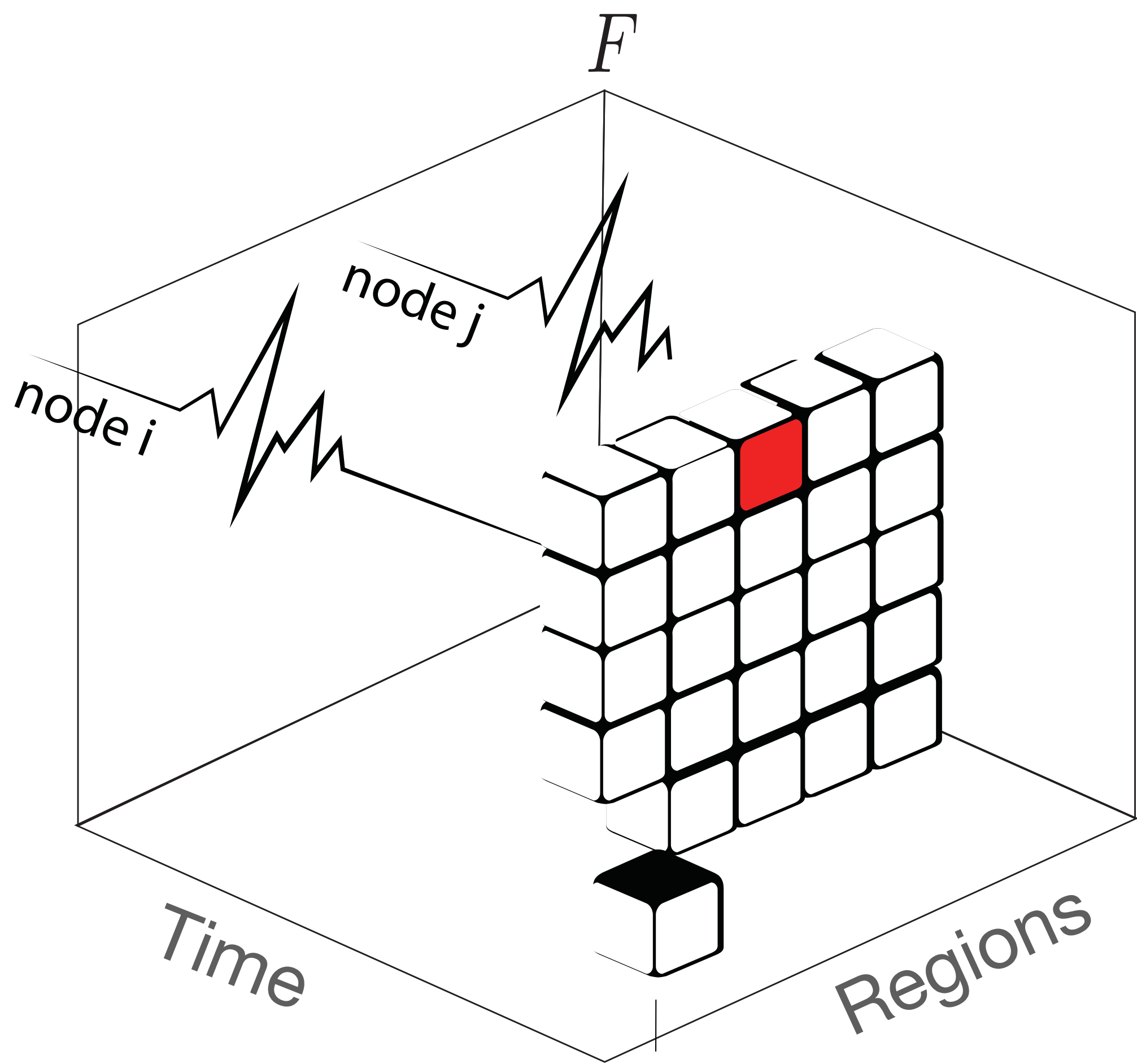
Functional Connectome

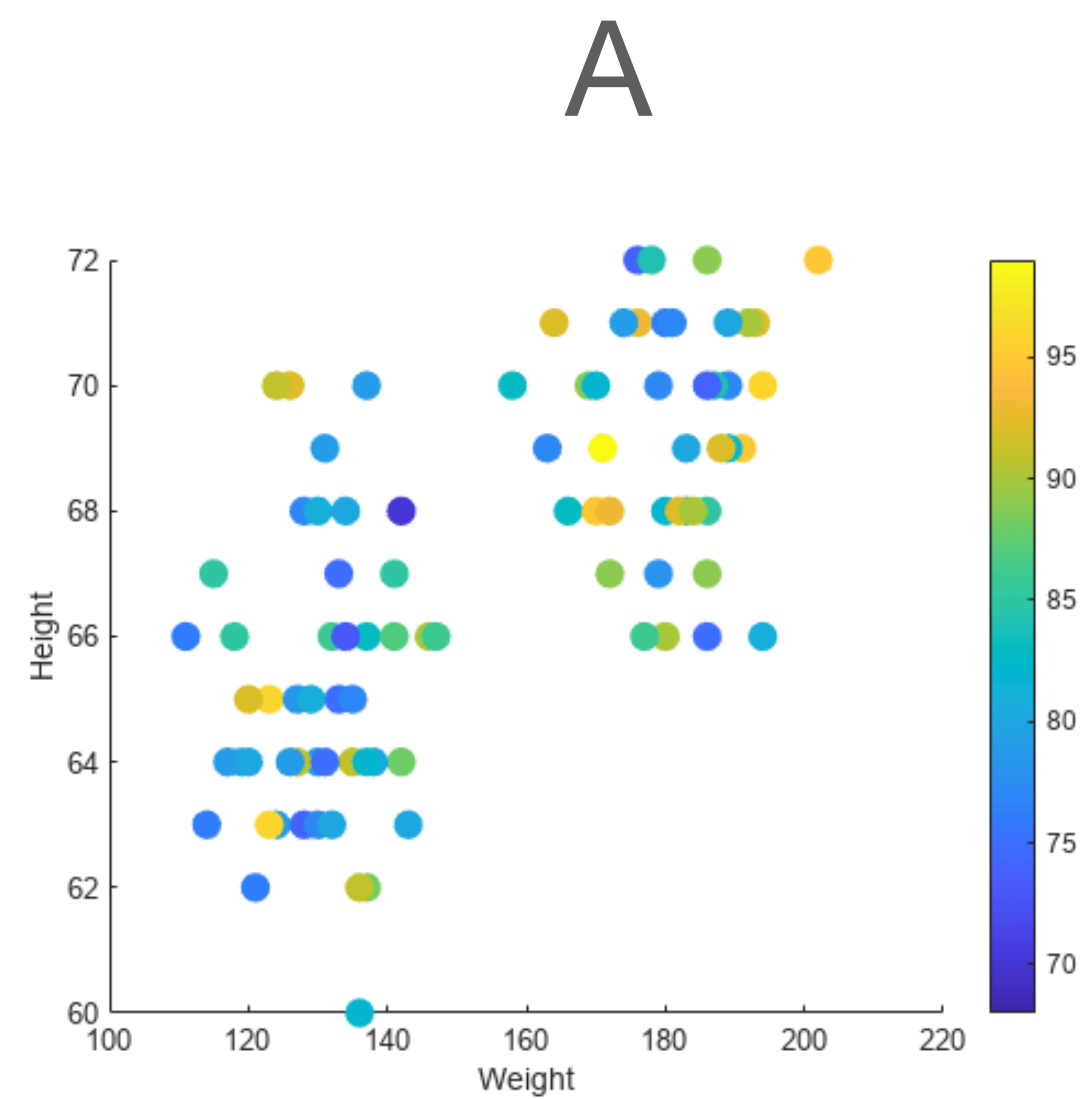


Mass multivariate analysis of disease group differences in brain network structure across all tasks. (A) Surface illustration of nodes where edges (network connections) significantly differ across all clinical groups, as measured with Hotelling's T2. (B) illustrates significant network-to-network edges.

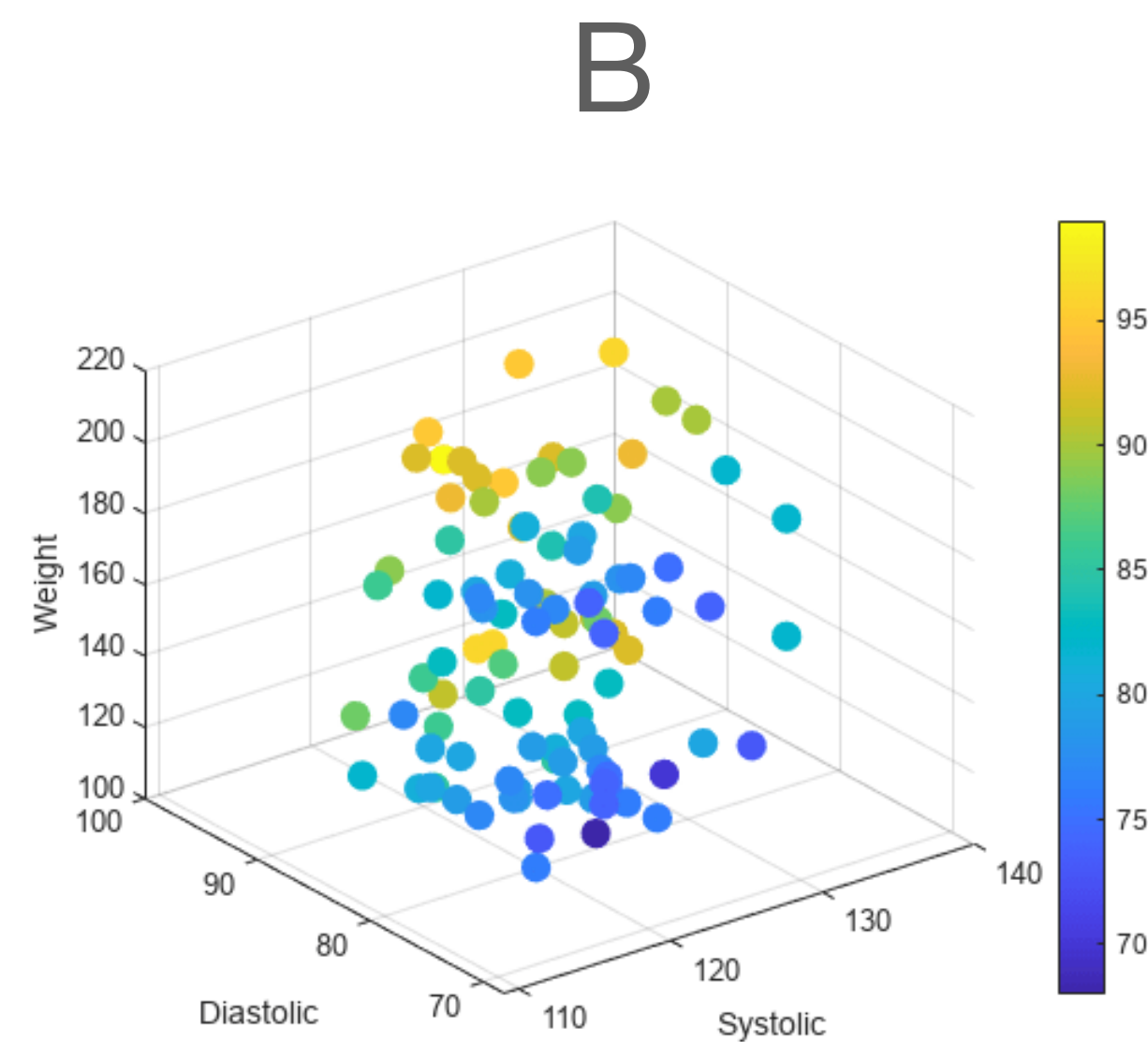
- The need for an atlas to create a connectome hinders comparisons across studies and replication and generalization efforts.
- Different atlases divide the brain into different regions of varying size and topology.
- Thus connectomes created from different atlases are not directly comparable.
- Further, several atlases exist with no gold standards, and more are being developed yearly.

Limitations in Open Science





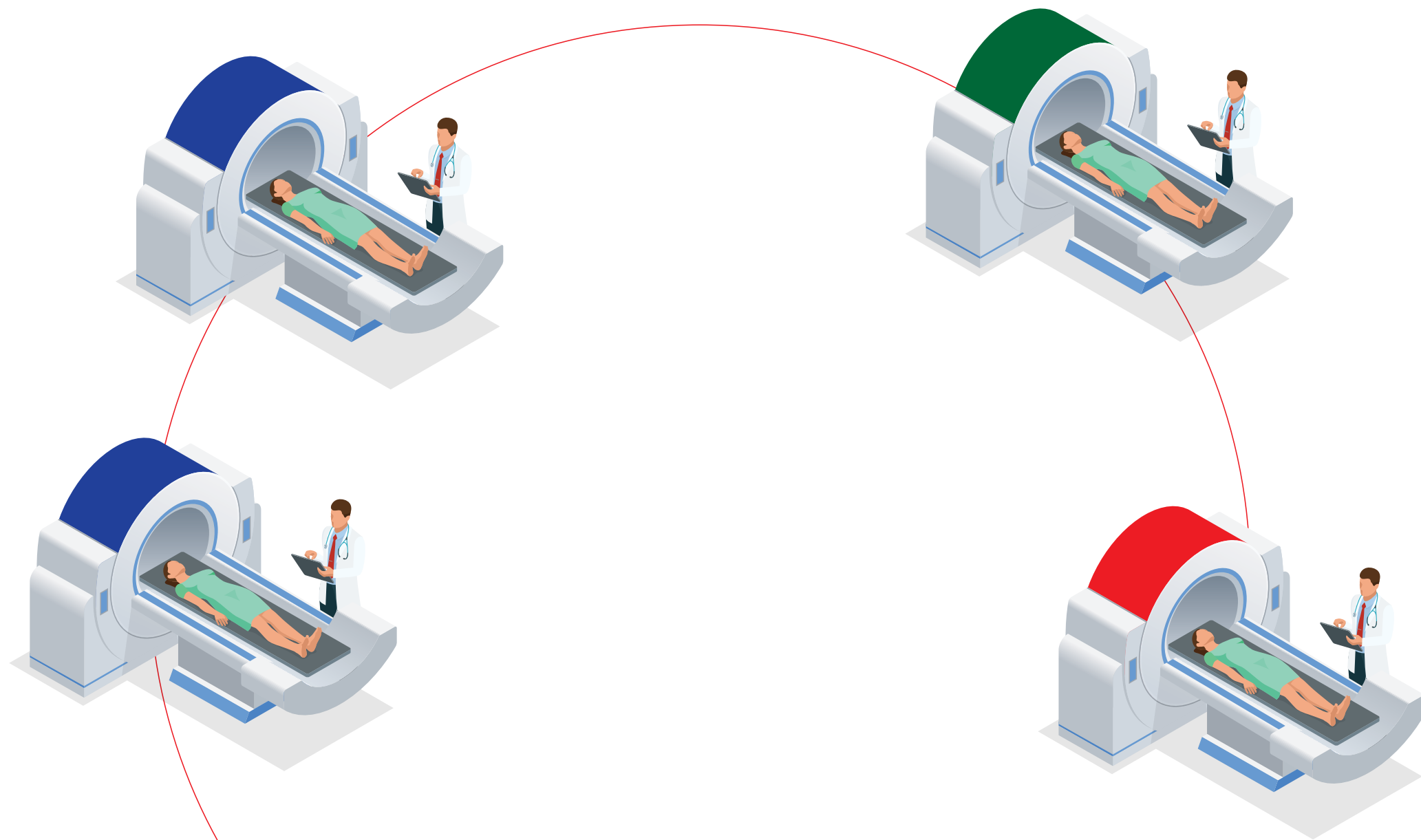
$$X, \beta \in \mathbb{R}^2$$



$$X \in \mathbb{R}^3$$

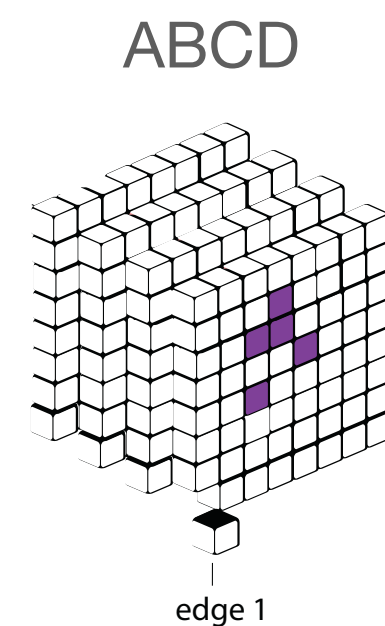
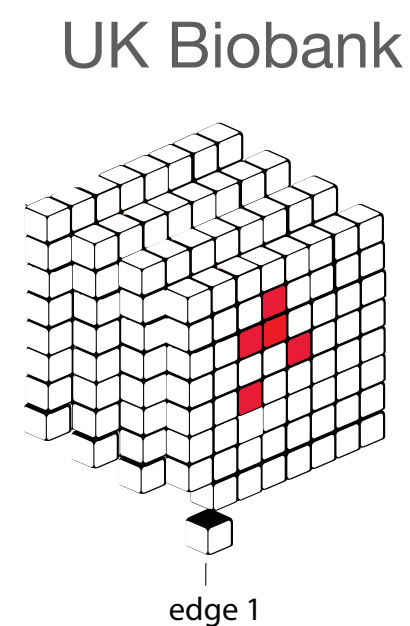
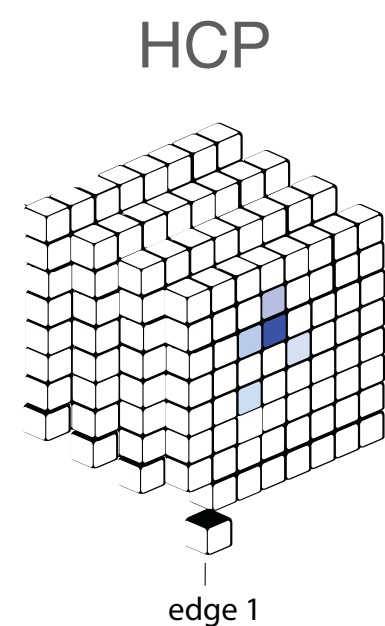
- Even in predictive modeling, feature space across all data points should be consistent.
 - It's impractical to train a model on A and test on B: $Y = X^T \beta + \epsilon$.
- Other techniques include meta-learning, transfer learning, and federated learning.

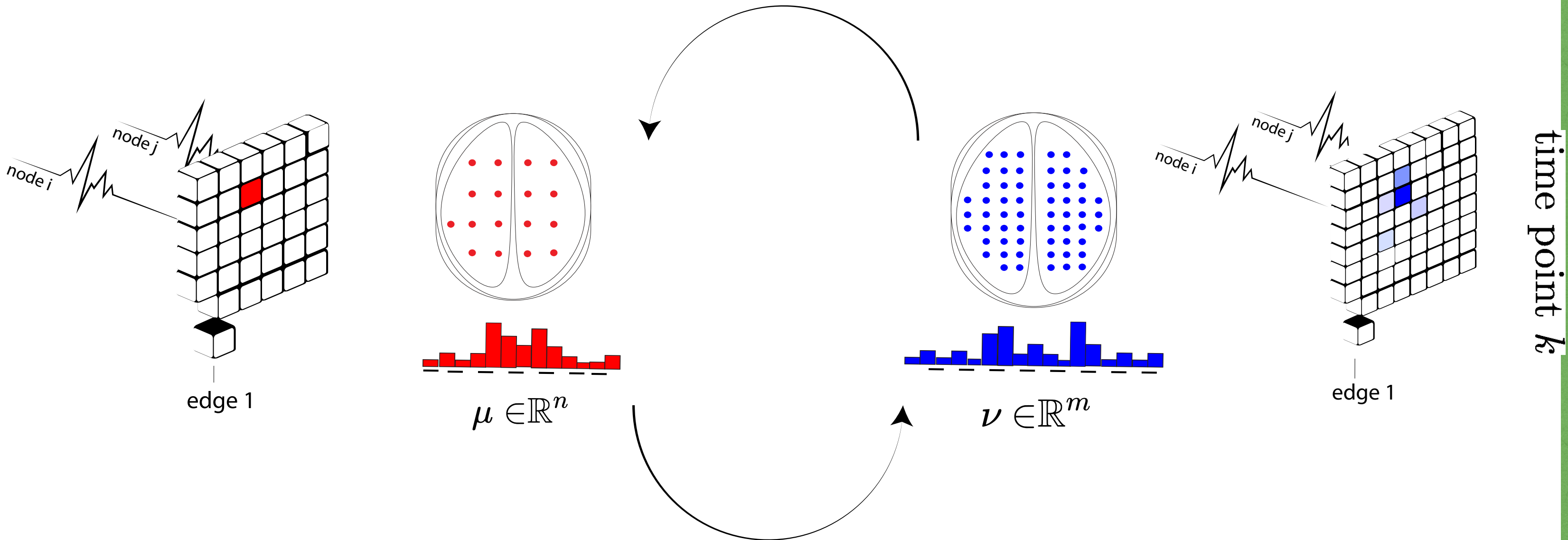
Small Studies



- Smaller labs might not have the resources to store and reprocess these data from scratch.
- Finally, due to privacy concerns of being able to identify a participant based on unprocessed data, some datasets are only released as fully processed connectomes.
 - Critically, in this case, it is not possible to go to the data to create connectomes from another atlas.

Large Scale Projects





Classical methods have limitations

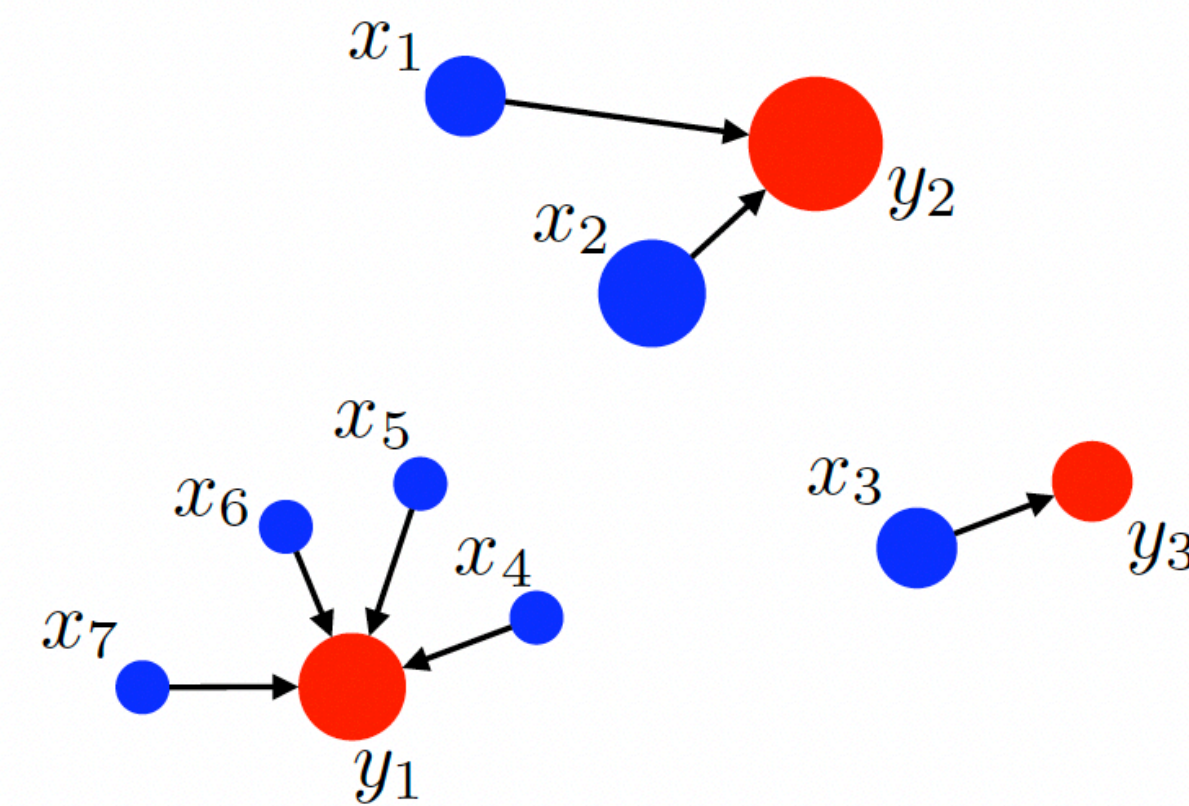
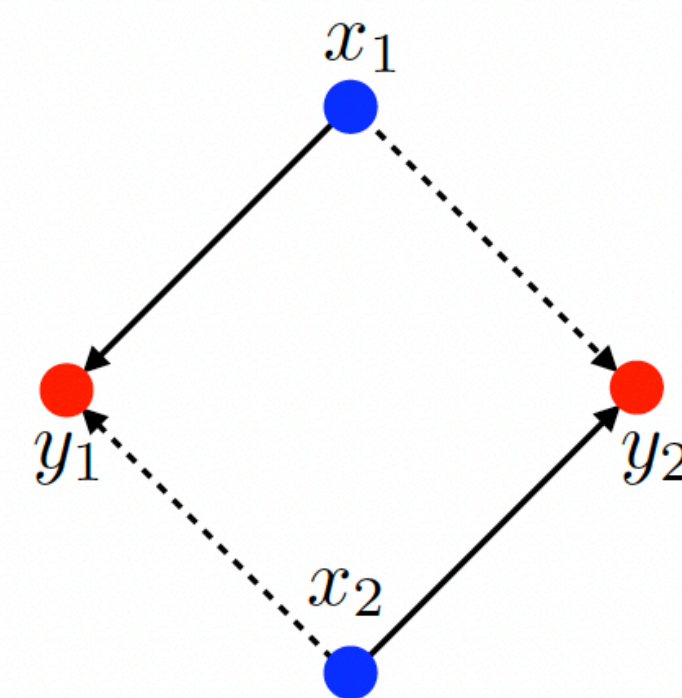
- Thus, algorithms to map and transform connectomes have applications to improve the generalizability of scientific findings.
- Classical algorithm either depends on having an equal number of supports or don't capture the geometry of space (e.g., KL divergence)

~~$$D_{KL}(\mu || \nu) = \sum_{x \in \mathcal{X}} \mu(x) \log \left(\frac{\mu(x)}{\nu(x)} \right)$$~~

Background

Optimal transport

Monge [1781]



A mapping between locations x and y

$$T : \{x_1, \dots, x_n\} \rightarrow \{y_1, \dots, y_n\}$$

must verify

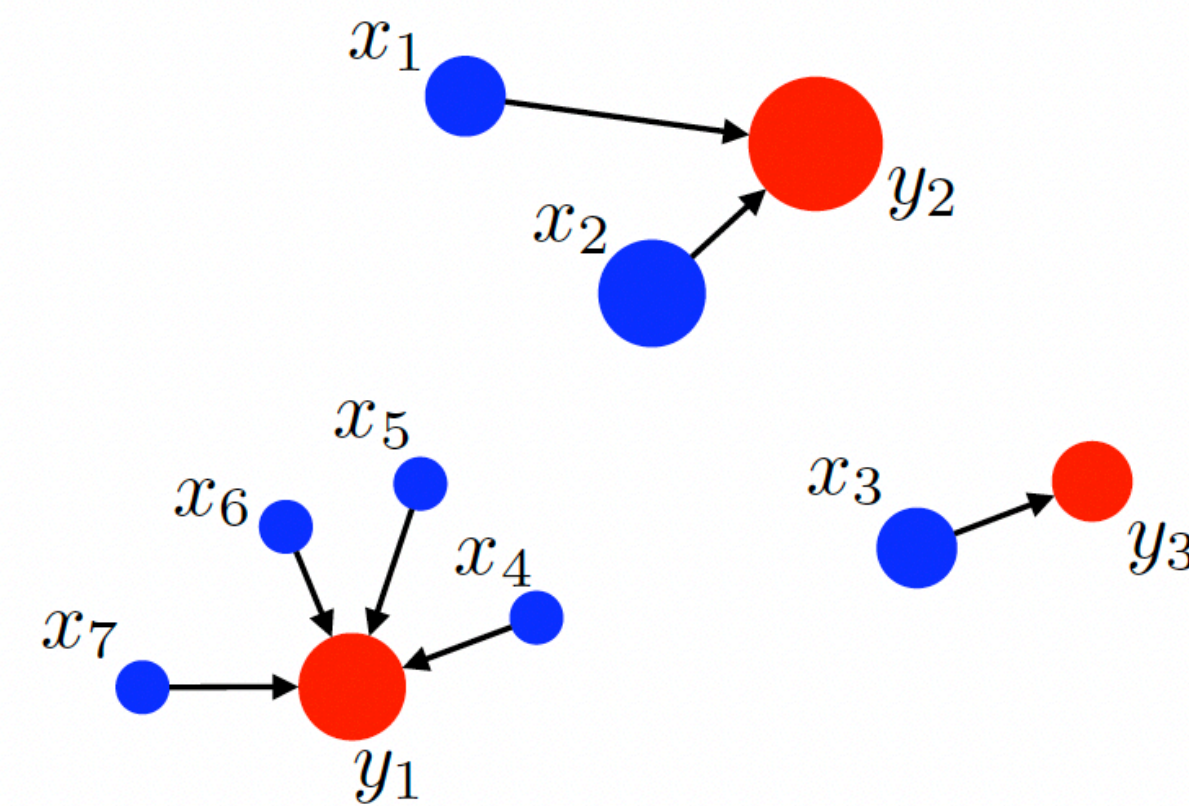
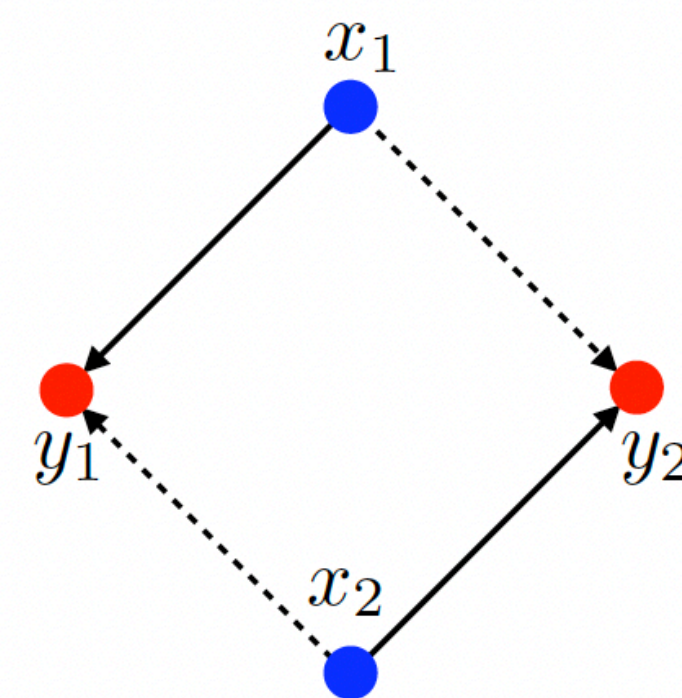
$$b_j = \sum_{i:T(x_i)=y_j} a_i$$

The only criterion here is to make sure we transfer all mass into some location y_j

Background

Optimal transport

Monge [1781]



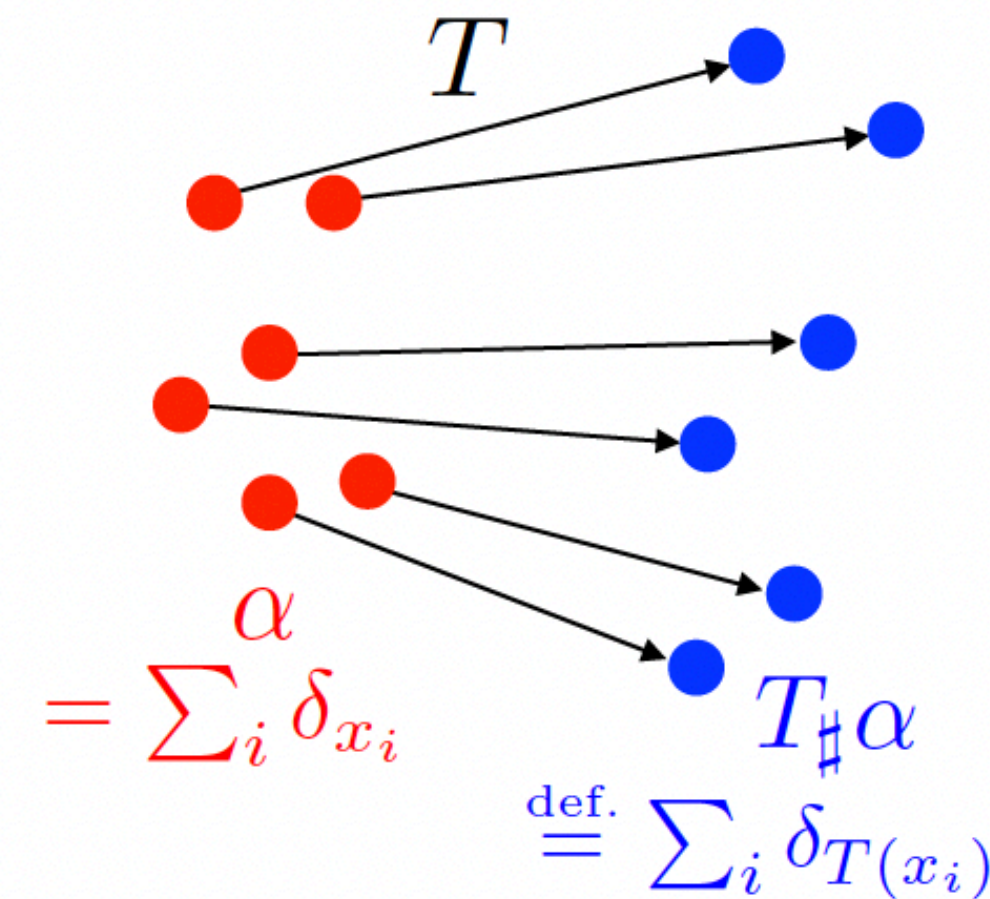
This map should minimize some transportation cost, which is parameterized by a cost function C

$$\min_T \left\{ \sum_i C(x_i, T(x_i)) : T_{\#}\alpha = \beta \right\},$$

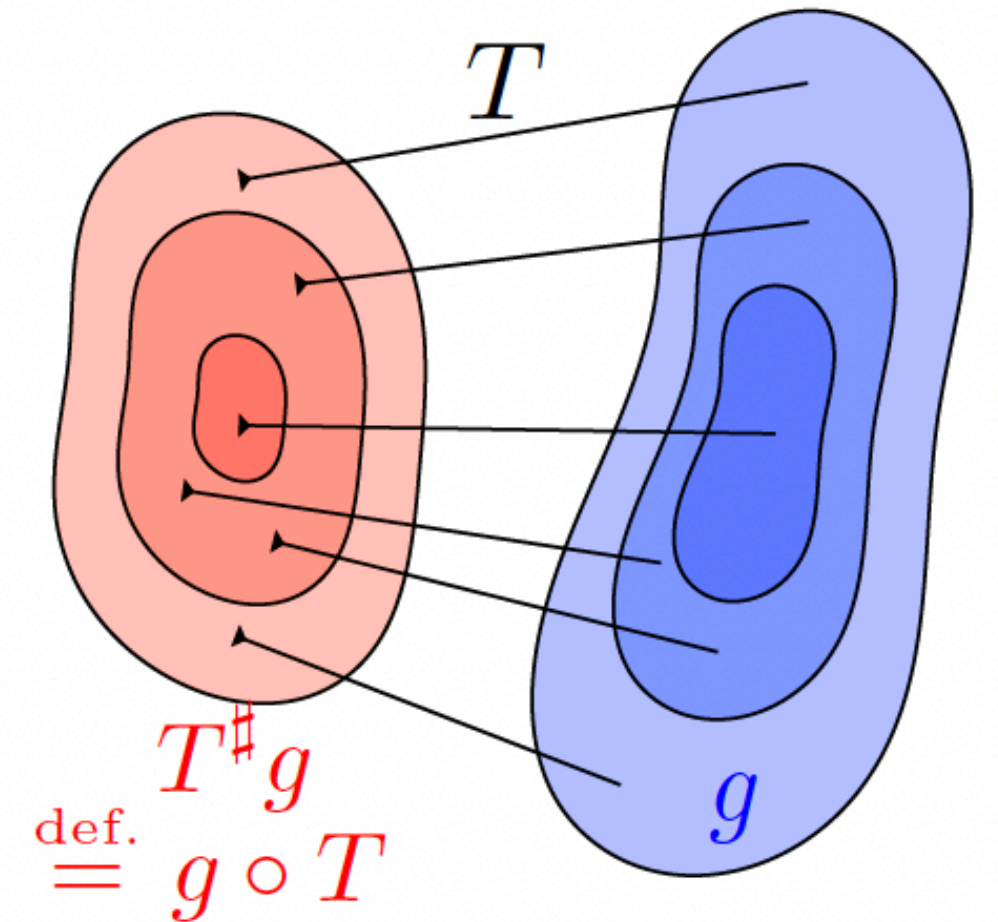
Background

Optimal transport

Monge [1781]



Push-forward of measures



Pull-back of functions

Kantorovich Relaxation [1942]

$$\mathbf{U}(\mathbf{a}, \mathbf{b}) \stackrel{\text{def.}}{=} \left\{ \mathbf{P} \in \mathbb{R}_+^{n \times m} : \mathbf{P} \mathbf{1}_m = \mathbf{a} \quad \text{and} \quad \mathbf{P}^T \mathbf{1}_n = \mathbf{b} \right\},$$

$$\mathbf{P} \mathbf{1}_m = \left(\sum_j \mathbf{P}_{i,j} \right)_i \in \mathbb{R}^n \quad \text{and} \quad \mathbf{P}^T \mathbf{1}_n = \left(\sum_i \mathbf{P}_{i,j} \right)_j \in \mathbb{R}^m.$$

Admissible Couplings

Kantorovich
[1942]

Background

Optimal transport

Monge [1781]

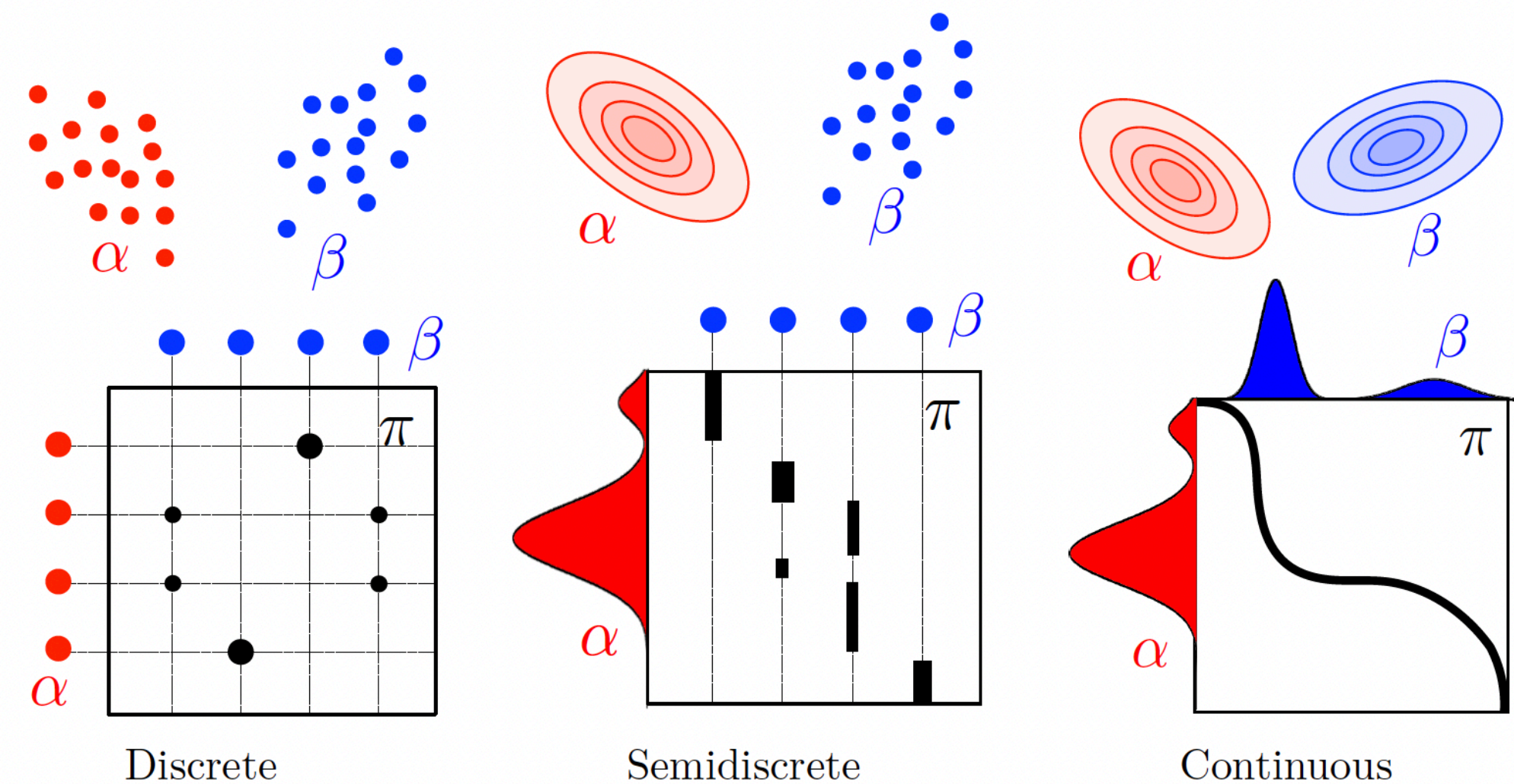
Kantorovich
[1942]

Kantorovich Relaxation is symmetric

$$P \in U(a, b) \Leftrightarrow P^T \in U(b, a)$$

Kantorovich's optimal transport problem now reads

$$L_C(\mathbf{a}, \mathbf{b}) \stackrel{\text{def.}}{=} \min_{\mathbf{P} \in \mathbf{U}(\mathbf{a}, \mathbf{b})} \langle \mathbf{C}, \mathbf{P} \rangle \stackrel{\text{def.}}{=} \sum_{i,j} C_{i,j} P_{i,j}.$$



Background

Optimal transport

Monge [1781]

Hitchcock
[1941]

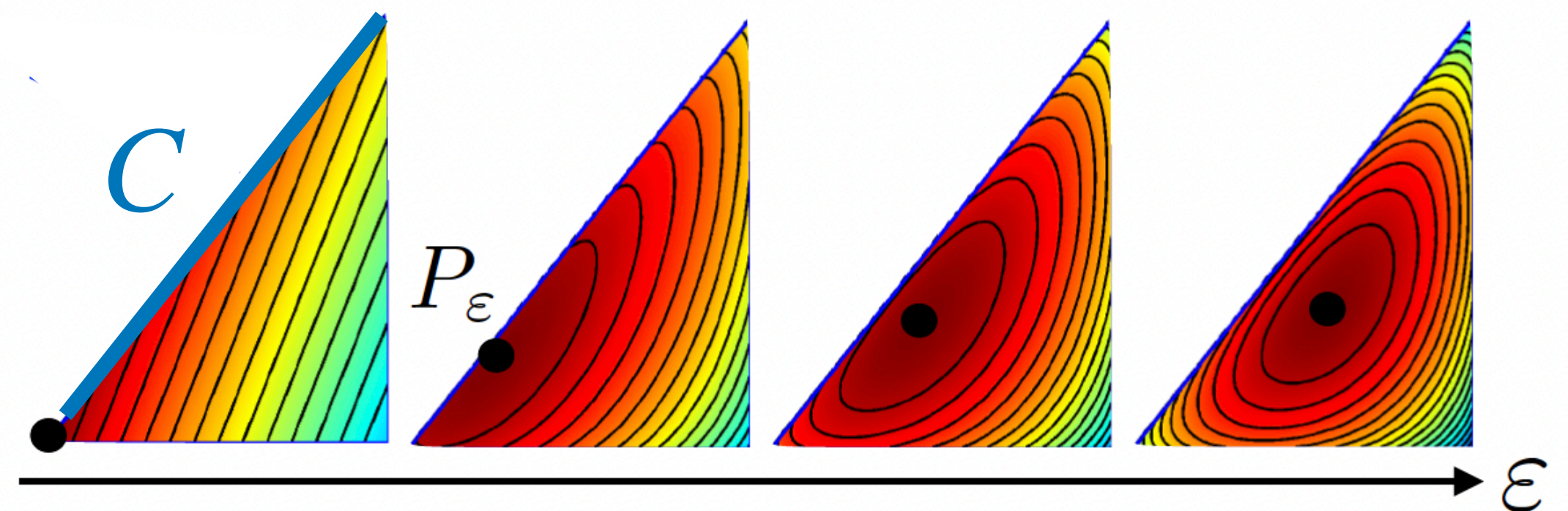
Kantorovich
[1942]

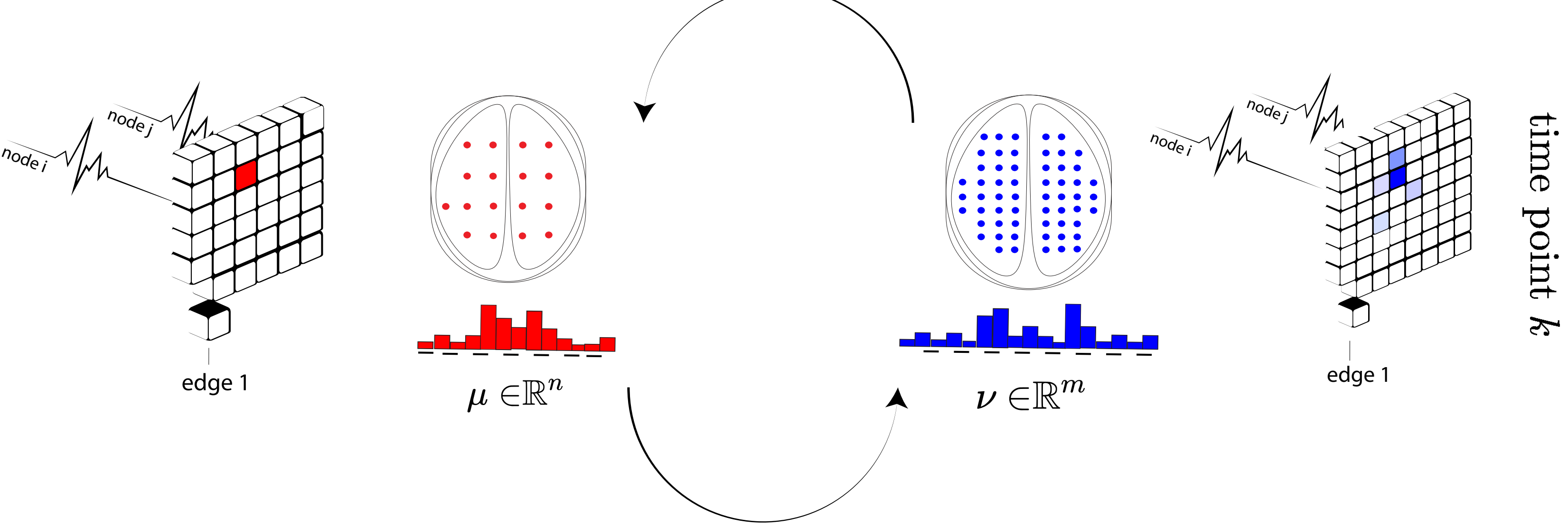
This is a linear program and can not be solved in polynomial time

Entropy regularization: An approximation solution

$$L_C^\varepsilon(\mathbf{a}, \mathbf{b}) \stackrel{\text{def.}}{=} \min_{\mathbf{P} \in \mathbf{U}(\mathbf{a}, \mathbf{b})} \langle \mathbf{P}, \mathbf{C} \rangle - \varepsilon \mathbf{H}(\mathbf{P}).$$

$$\mathcal{O}(n^2 \log(n) \eta^{-3}) \text{ for } \varepsilon = \frac{4 \log(n)}{\eta}$$



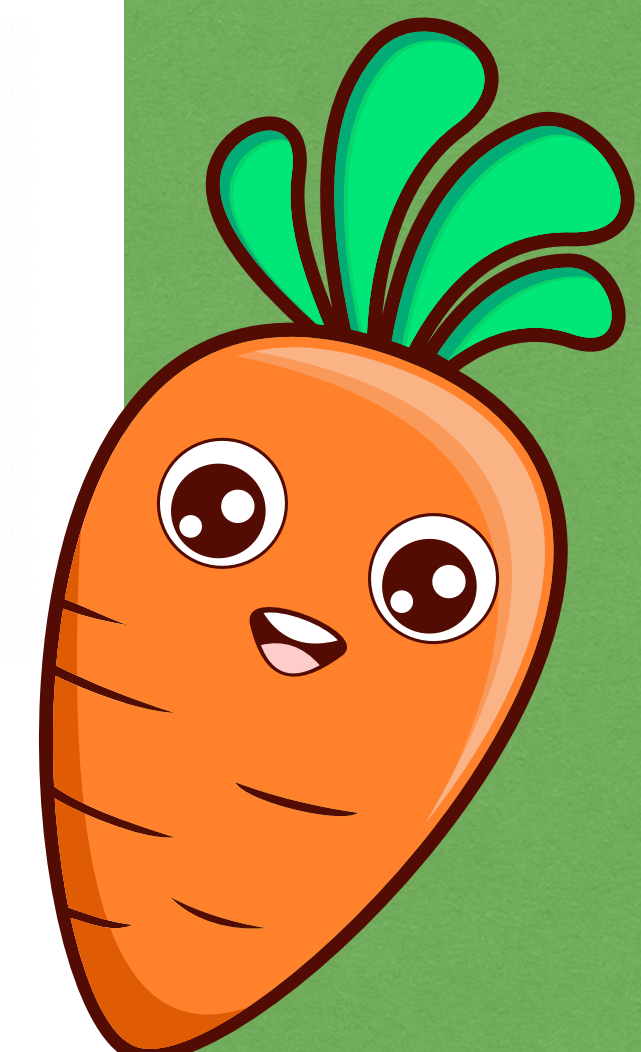


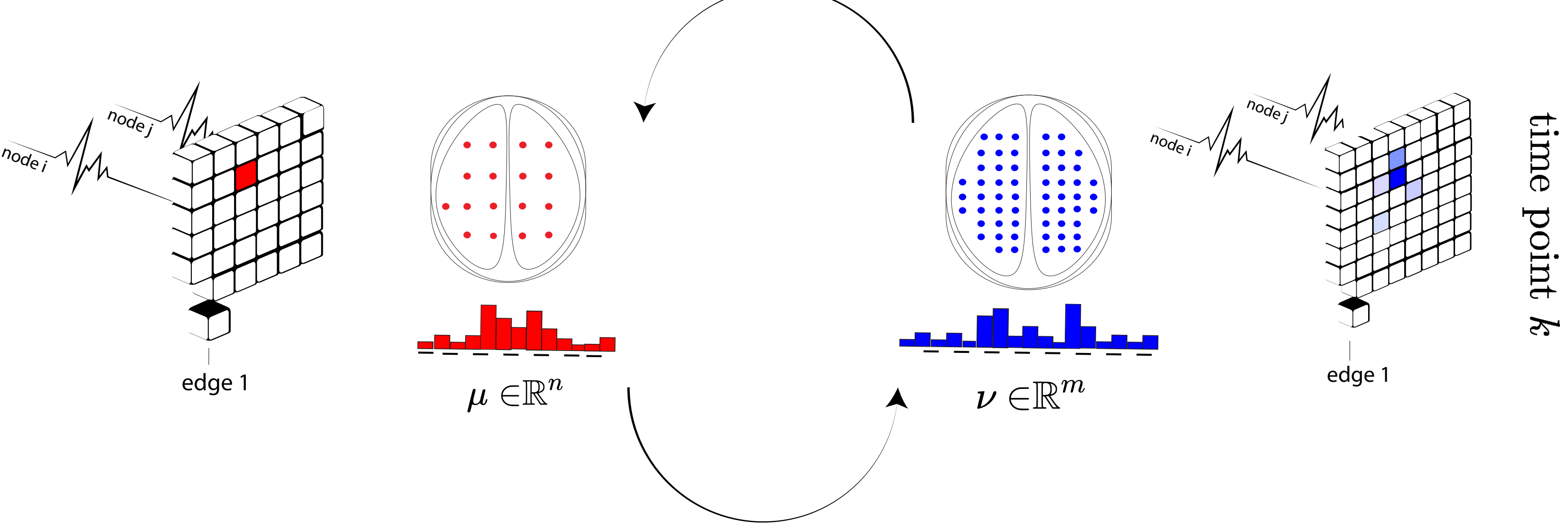
Cross Atlas Remapping via Optimal Transport (CAROT)

$$L_c(\mu_t, \nu_t) = \min_{\mathcal{T}} C^T \mathcal{T} - \epsilon H(\mathcal{T}) \text{ s.t. } A \underline{\mathcal{T}} = \begin{bmatrix} \mu_t \\ \nu_t \end{bmatrix}$$

$$C = \begin{pmatrix} C_{1,1} & \dots & C_{1,m} \\ \vdots & \ddots & \vdots \\ C_{n,1} & \dots & C_{n,m} \end{pmatrix} \in \mathbb{R}^{n \times m}$$

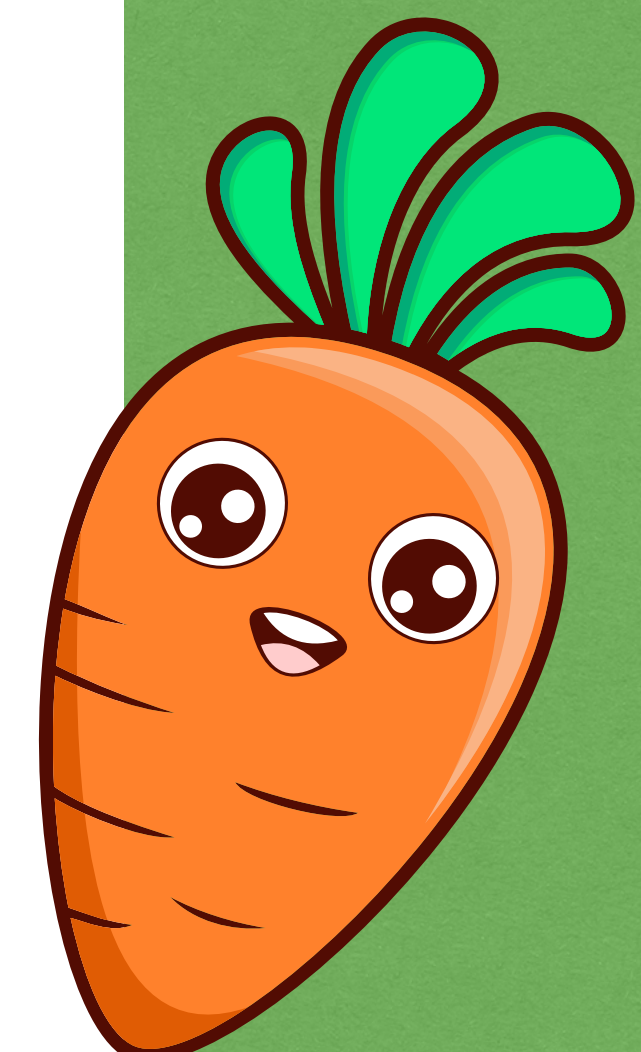
$$A = \begin{matrix} & \begin{matrix} 1 & 2 & \dots & n \end{matrix} \\ \begin{matrix} m \\ n \end{matrix} & \begin{pmatrix} \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \\ 1 & 1 & \dots & 1 \end{pmatrix} & \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \\ 1 & 1 & \dots & 1 \end{pmatrix} & \dots & \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \\ 1 & 1 & \dots & 1 \end{pmatrix} \end{pmatrix} \end{matrix}$$

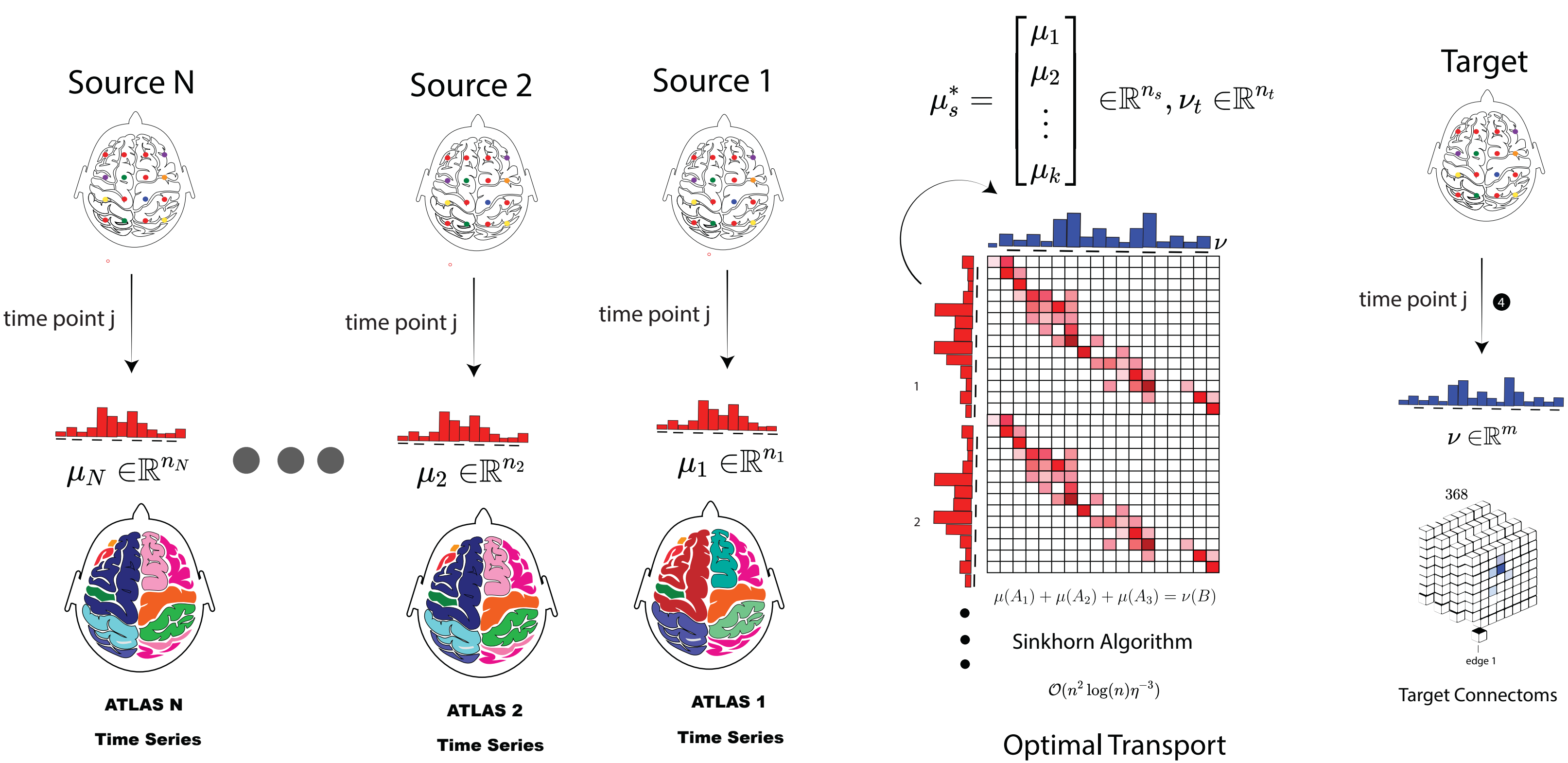




Cross Atlas Remapping via Optimal Transport (CAROT)

- Once we have trained the mapping, we can estimate the target Connectome by $\nu = \mu T$
- Sometimes, the large-scale studies release their data in multiple atlases (e.g., HCP, UK Biobank, Rest MDD)
- Next, we want to expand the current framework into a more dynamic architecture





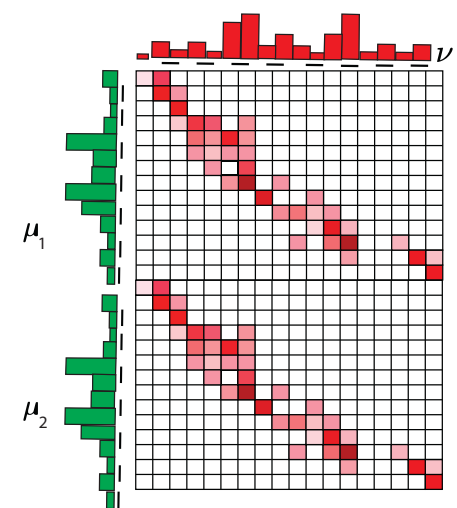
Cross Atlas Remapping via Optimal Transport (CAROT)

$$L_c(\mu_t^*, \nu_t^*) = \min_{\mathcal{T}} C^T \mathcal{T} - \epsilon H(\mathcal{T}) \text{ s.t. } A \mathcal{T} = \begin{bmatrix} \mu_t^* \\ \nu_t^* \end{bmatrix}.$$

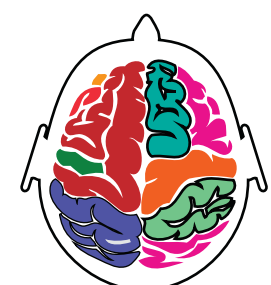
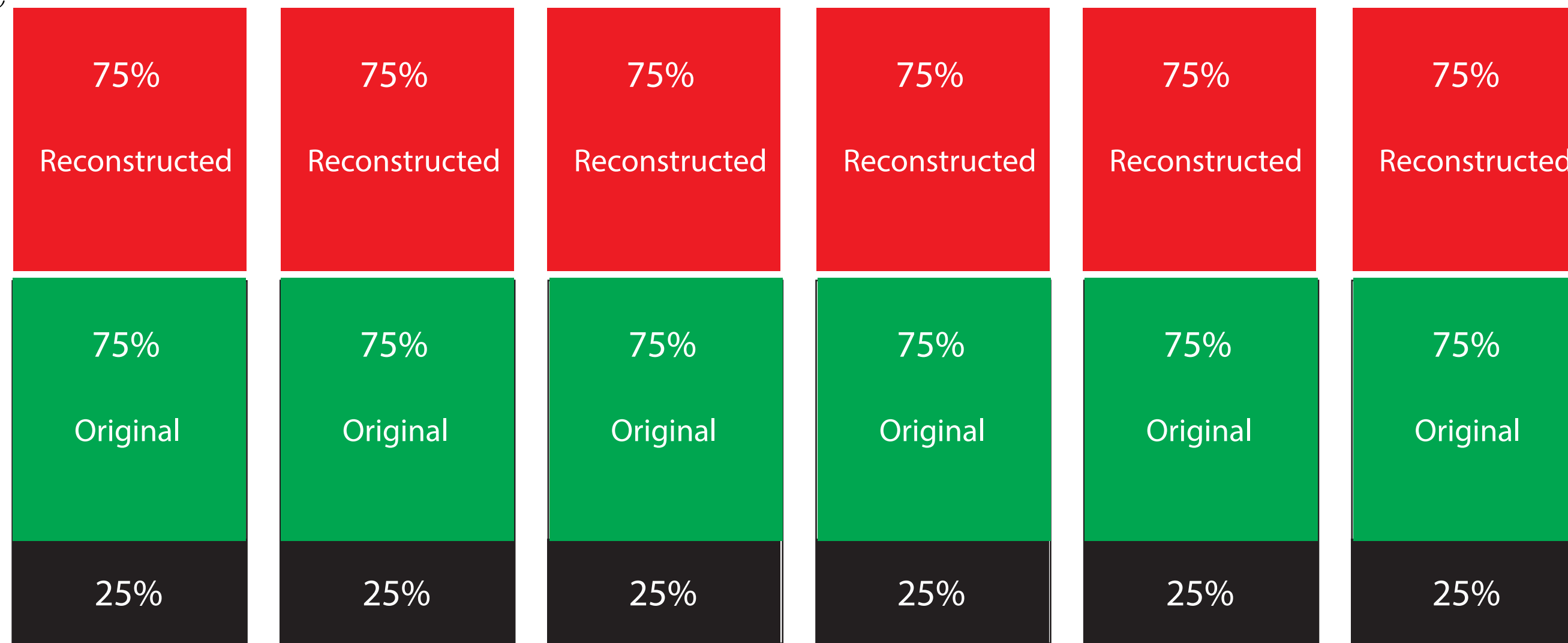
$$\mu_s^* = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_k \end{bmatrix} \in \mathbb{R}^{n_s}, \nu_t \in \mathbb{R}^{n_t}, C^* = \begin{pmatrix} C_{1,1} & \dots & C_{1,m} \\ \vdots & \ddots & \vdots \\ C_{n_s,1} & \dots & C_{n,m} \end{pmatrix} \in \mathbb{R}^{n_s \times m}$$



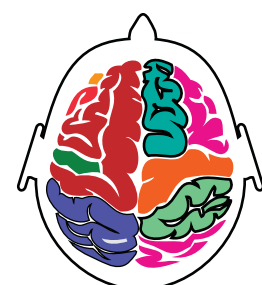
The Human Connectome project is used for training mappings, intrinsic analysis, and for some downstream analysis



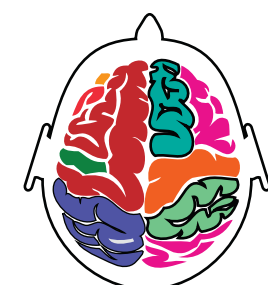
apply CAROT for a given target atlas



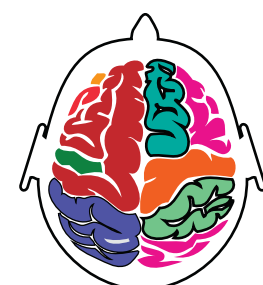
Shen Atlas



Craddock



Dosenbach



Schaefer



Brainnetom



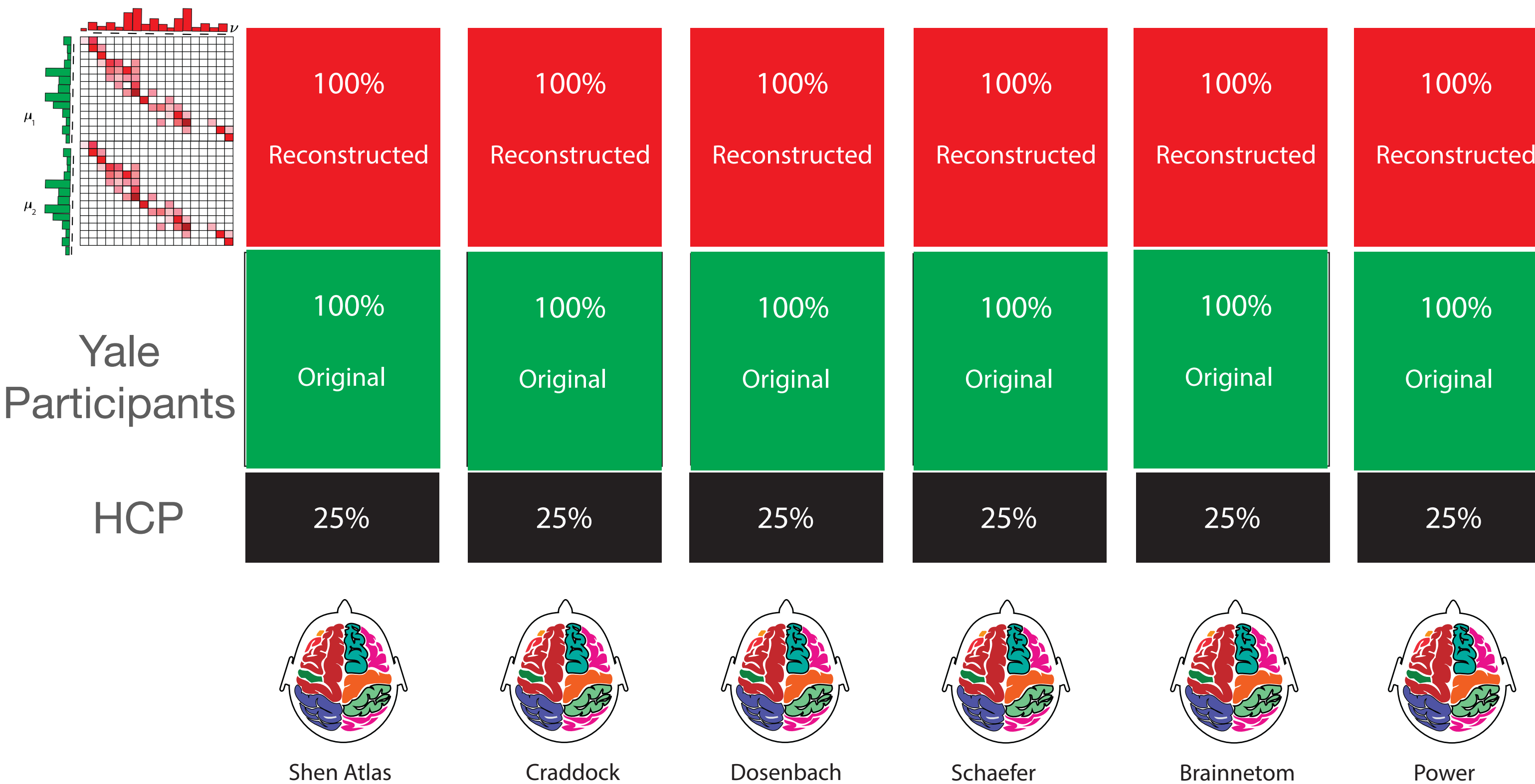
Power

$$\binom{6}{2} + 6 = 21 \text{ transportation policies}$$

Human Connectome Project

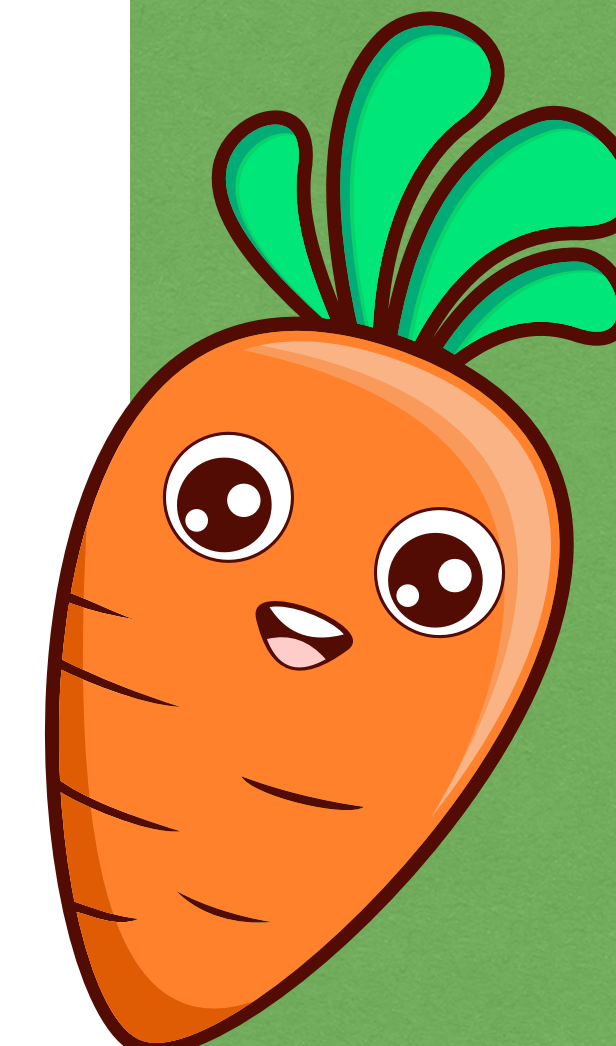


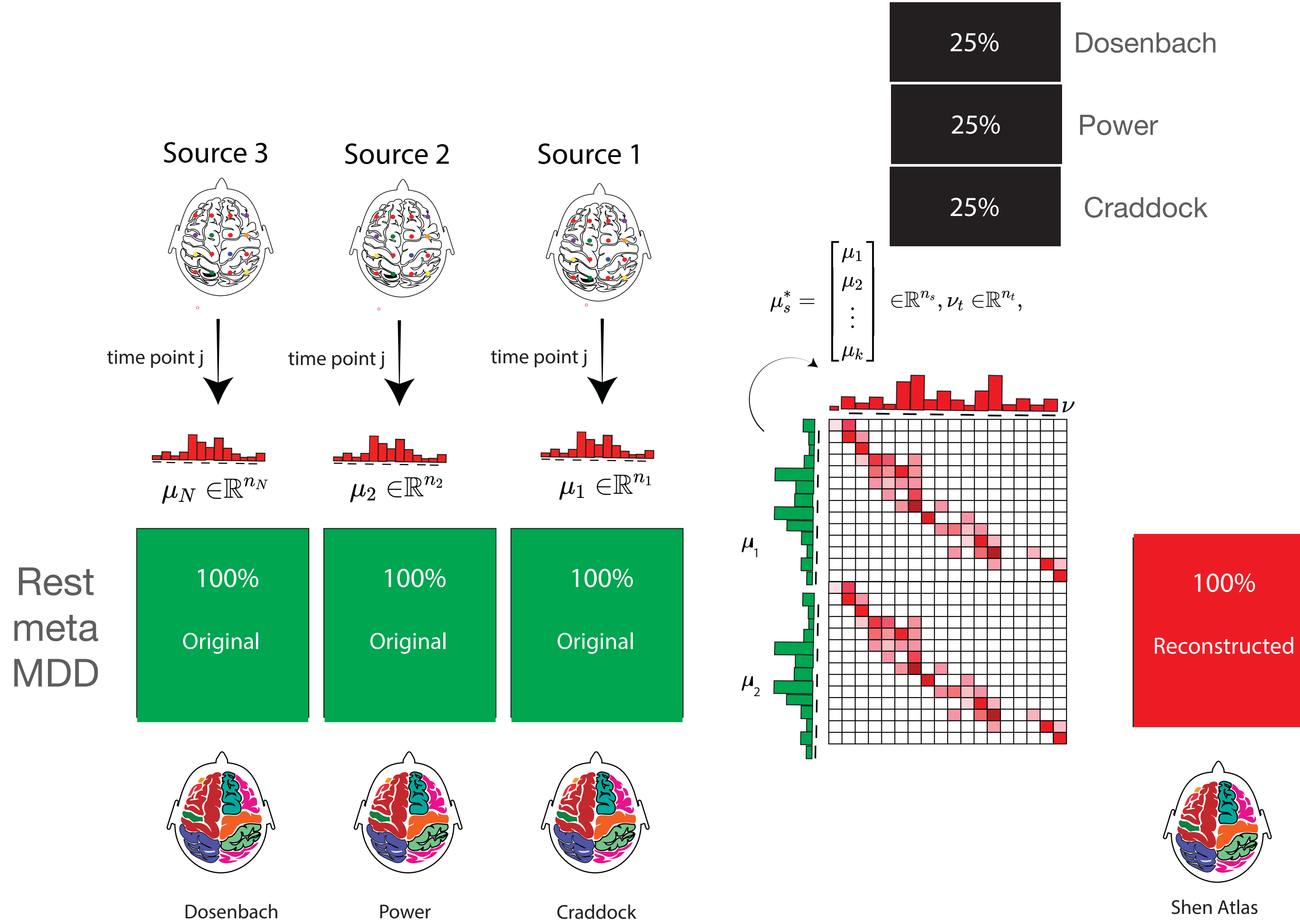
Cross-dataset analysis: We used resting-state data collected from 100 participants at the Yale School of Medicine.



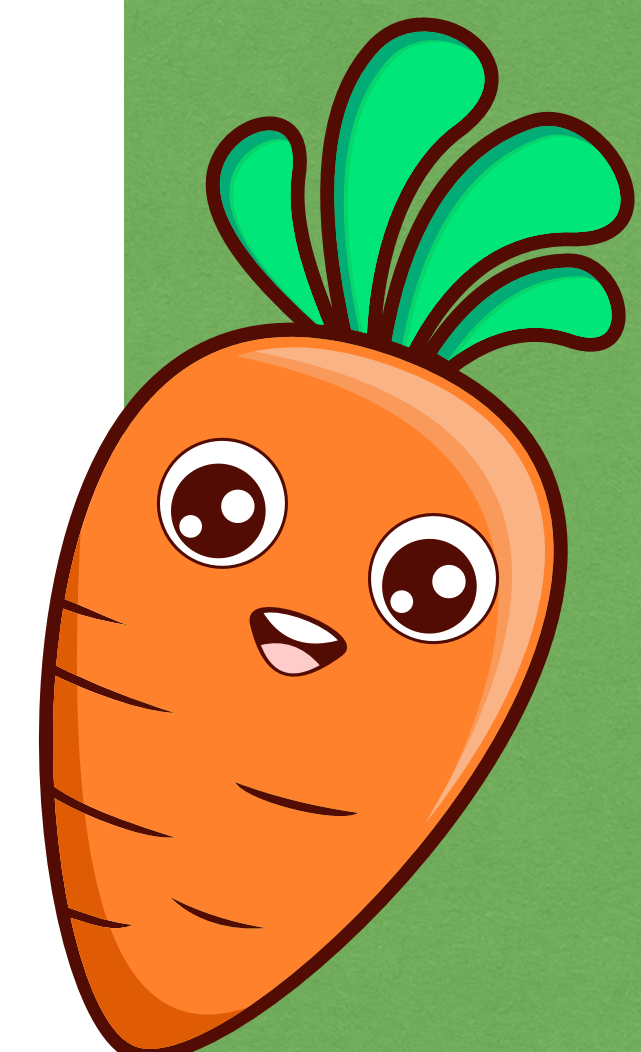
A Second Dataset: Yale Participants

This dataset included 50 females (age=33) and 50 males (age=34.9) with eight functional scans (48 minutes total).

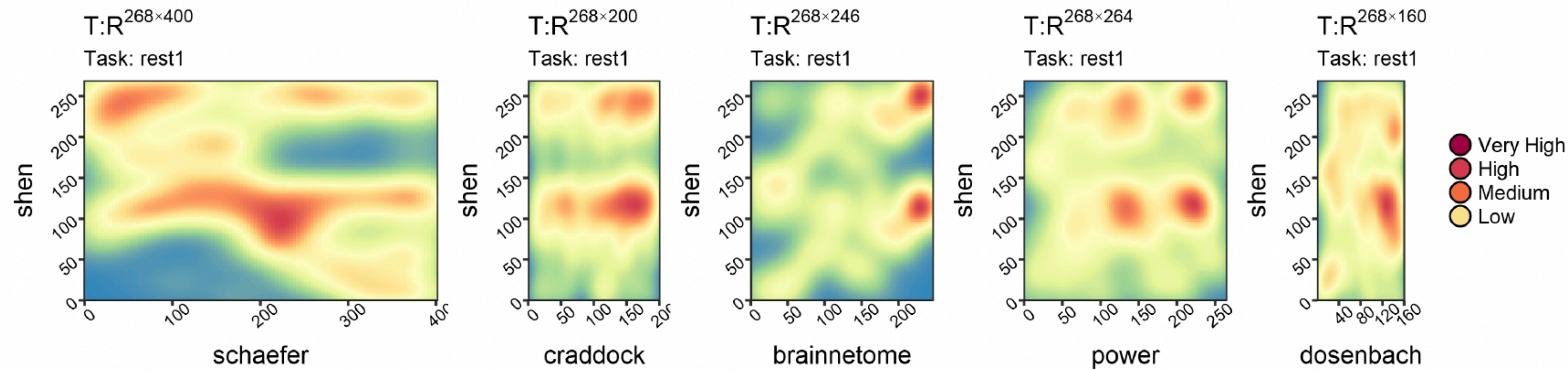




Third Dataset: Sex classification trained on Yale

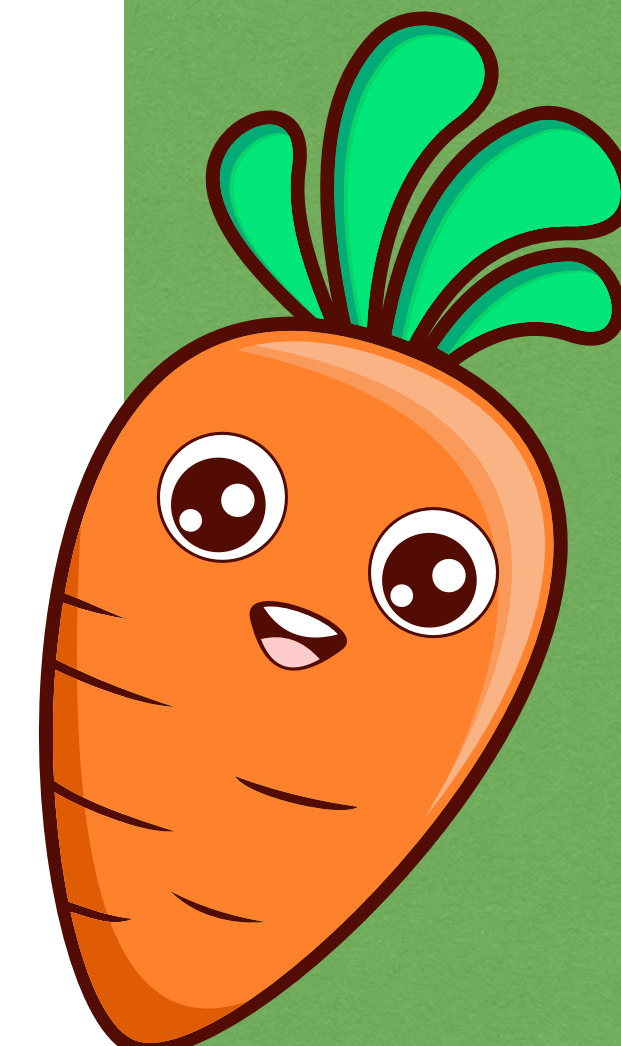


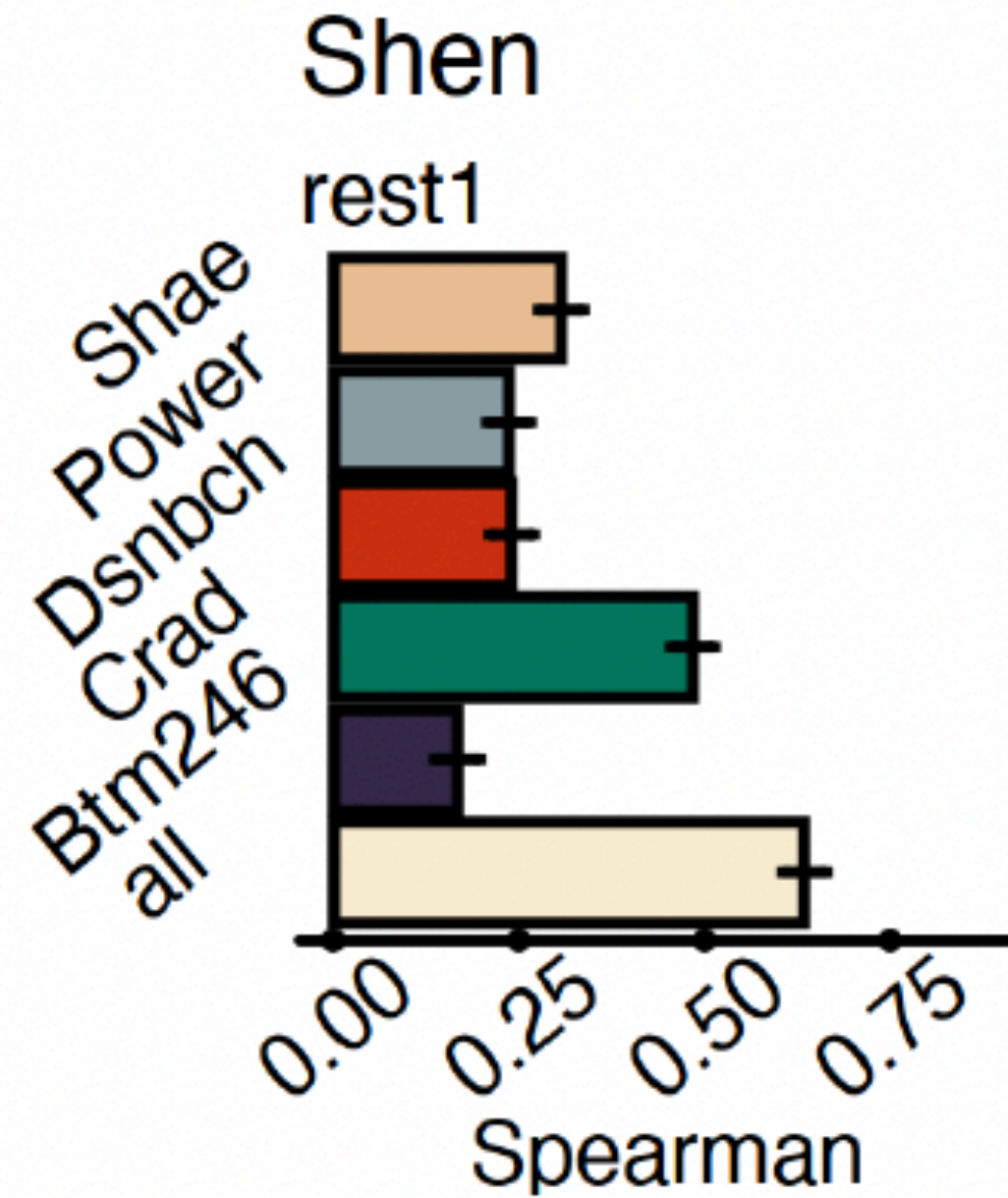
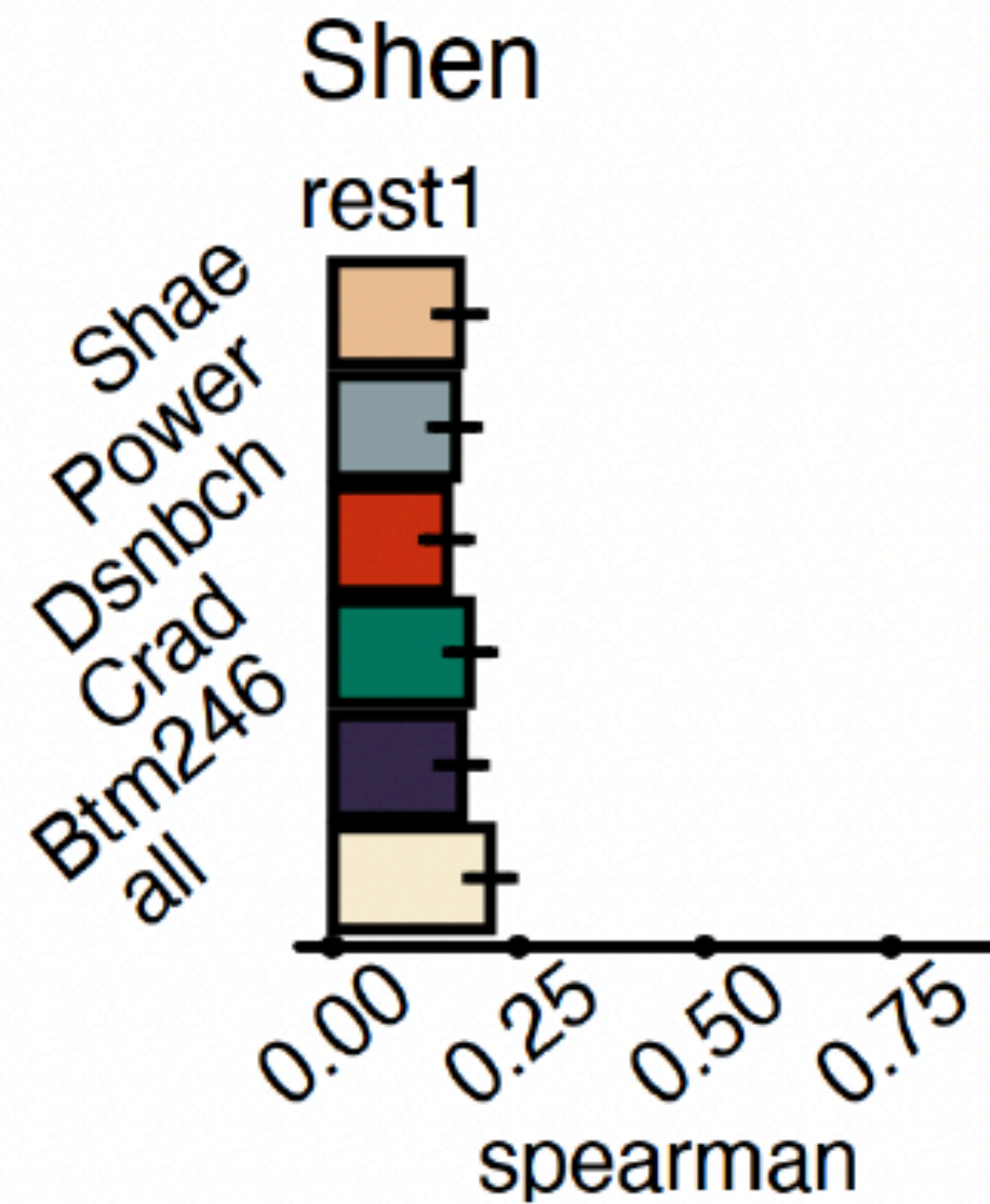
We generalize a sex classification model (using Yale School of Medicine and created with the Shen atlas) to the REST-Meta-MDD dataset, which only provides preprocessed timeseries data from the Dosenbach, Power, and Craddock atlases



- You can see that some spots are more intense than others indicating higher transformation between regions.
- This emphasizes some of these topological differences between atlases.
 - The horizontal line between Schaefer and Shen is belonging to areas that are missing in Schaefer

How does a mapping look like?





What should we choose as a cost matrix?

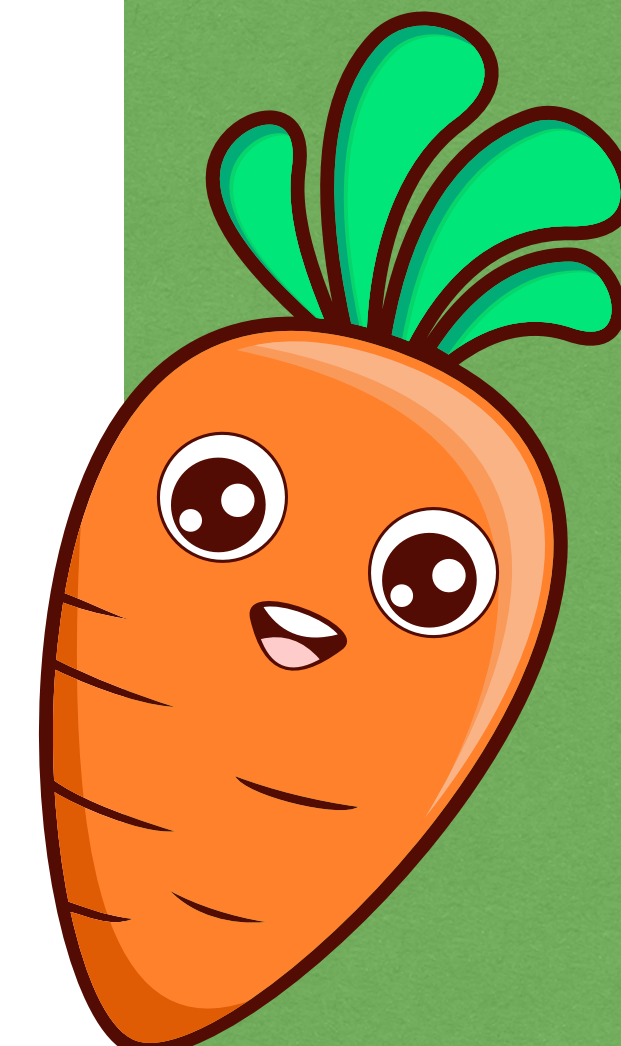
$$C^* = \begin{pmatrix} C_{1,1} & \dots & C_{1,m} \\ \vdots & \ddots & \vdots \\ C_{n_s,1} & \dots & C_{n,m} \end{pmatrix}$$

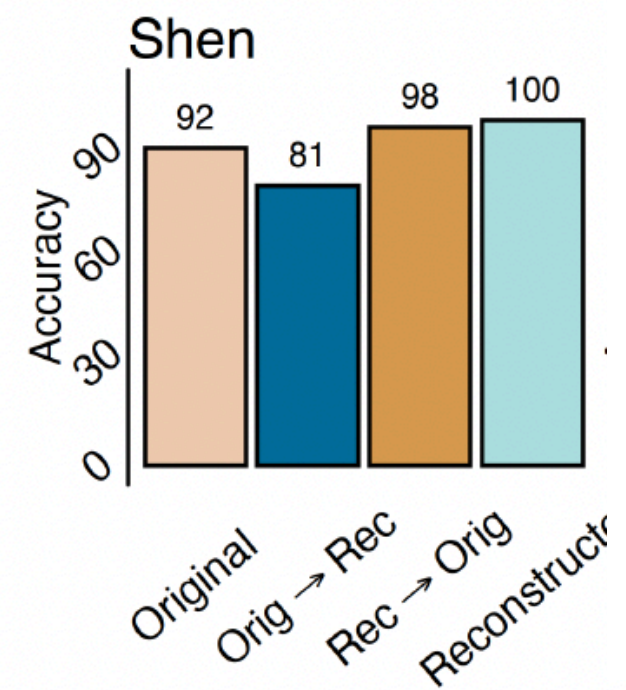
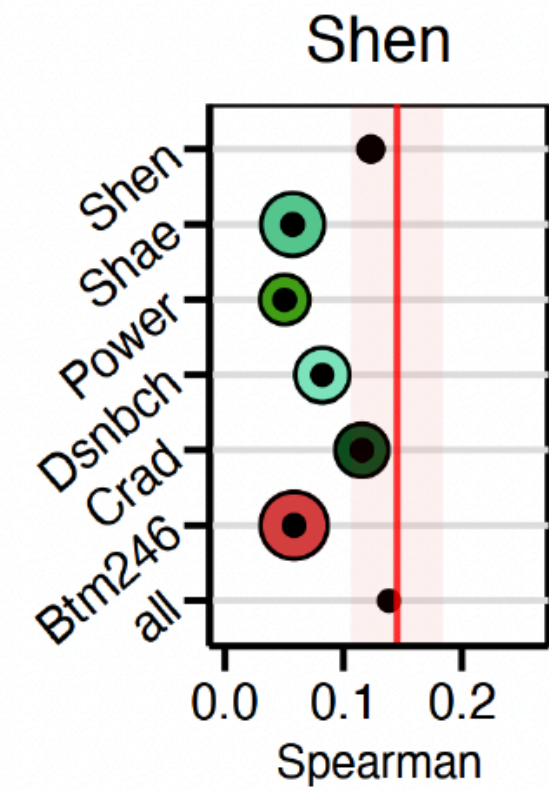
$$C^* = 1 - \begin{pmatrix} \rho_{1,1} & \dots & \rho_{1,m} \\ \vdots & \ddots & \vdots \\ \rho_{n_s,1} & \dots & \rho_{n,m} \end{pmatrix} \in \mathbb{R}^{n_s \times m}$$

$$C_{\text{euc}}(p, q) = \sqrt{\sum_{i=1}^3 (q_i - p_i)^2}$$

$$\rho_{i,j} = \text{Spearman}(\text{ROI}_1, \text{ROI}_2)$$

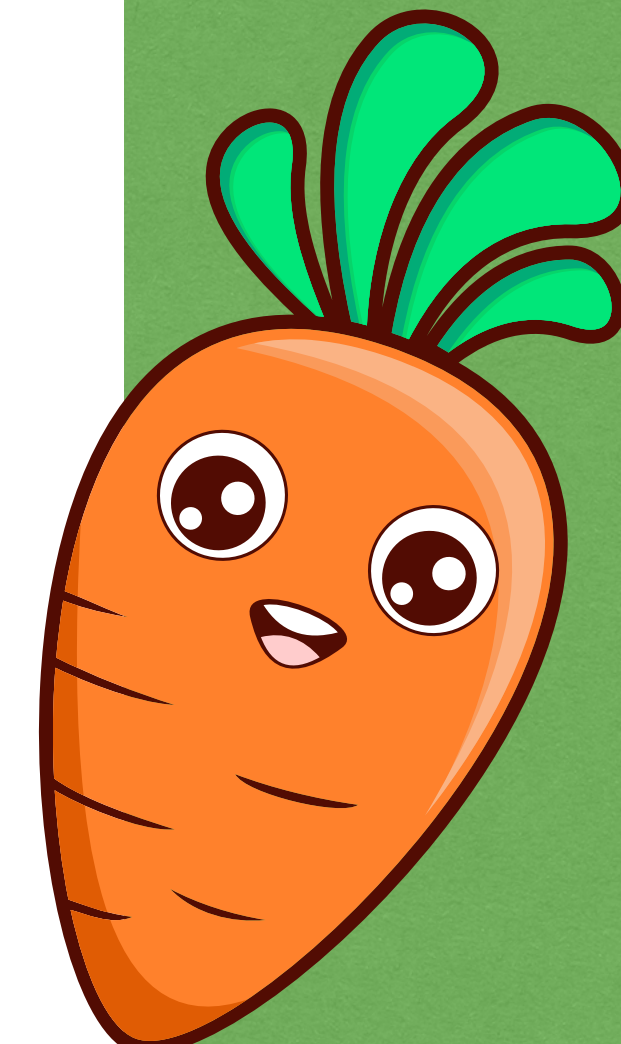
- The performance with a single source is quite sensible. But we could do better than that.





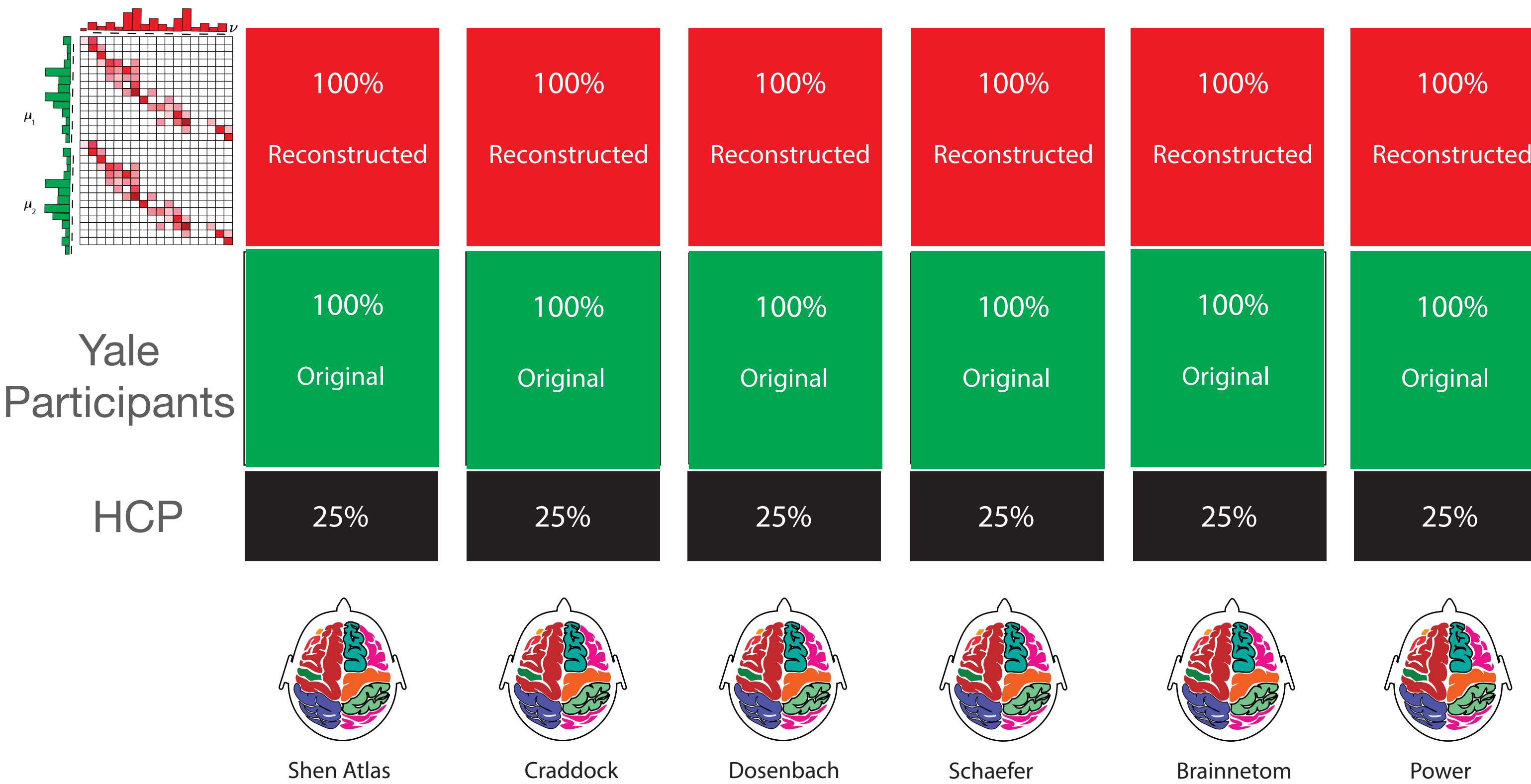
- The correlation as a function of k is linearly increasing.
- There are differences among various runs and targets
 - Topologically similar atlases reproduced more similar connectomes
- We can predict behavior (e.g., fluid intelligence) and can identify individuals across different runs

Intrinsic evaluation and downstream analysis in HCP



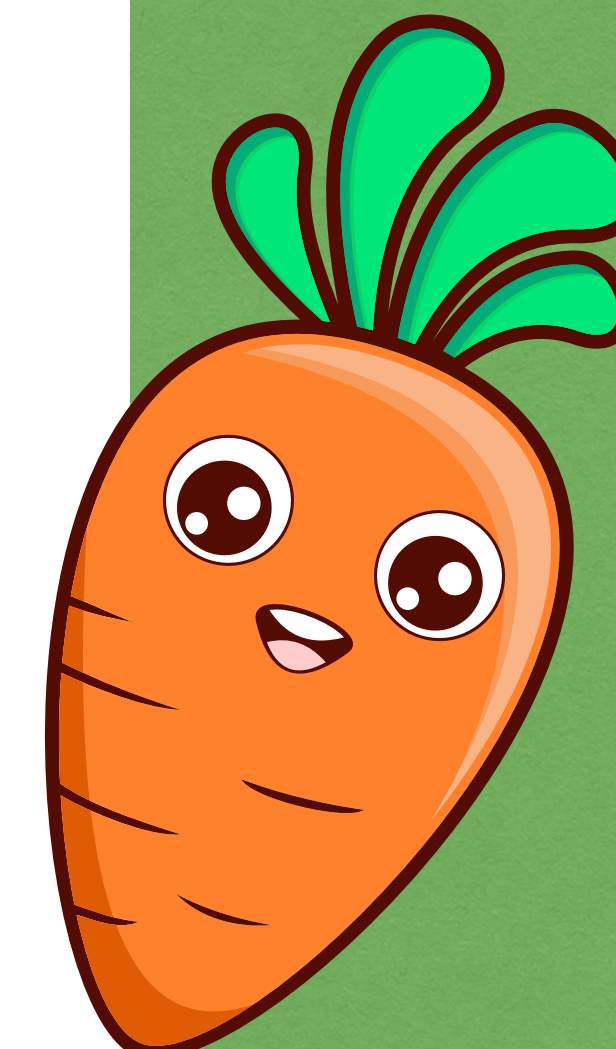
Can we generalize these mappings into a different dataset?

Generalization of the mappings on Yale participants

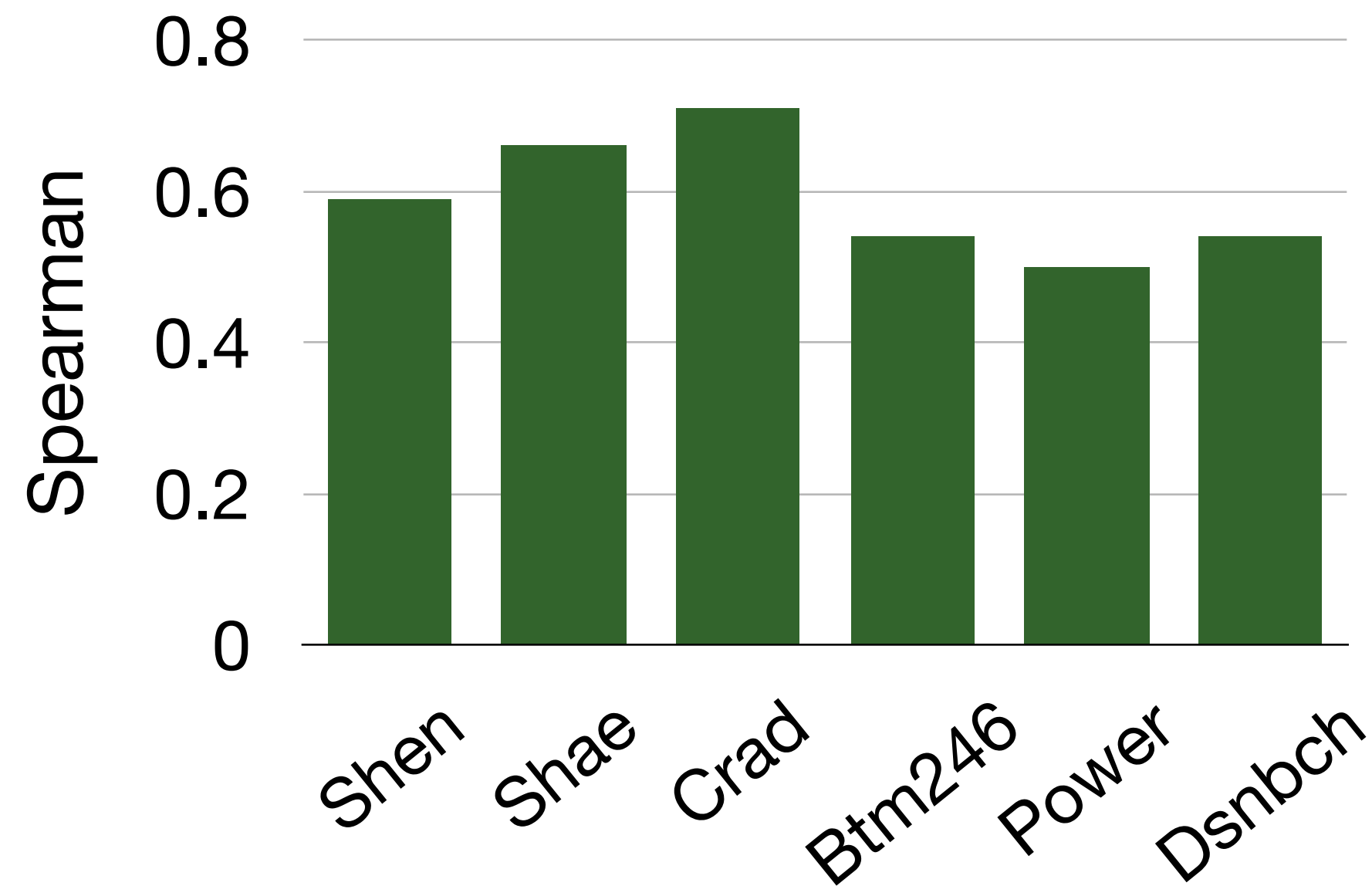


Review

A Second Dataset: Yale Participants

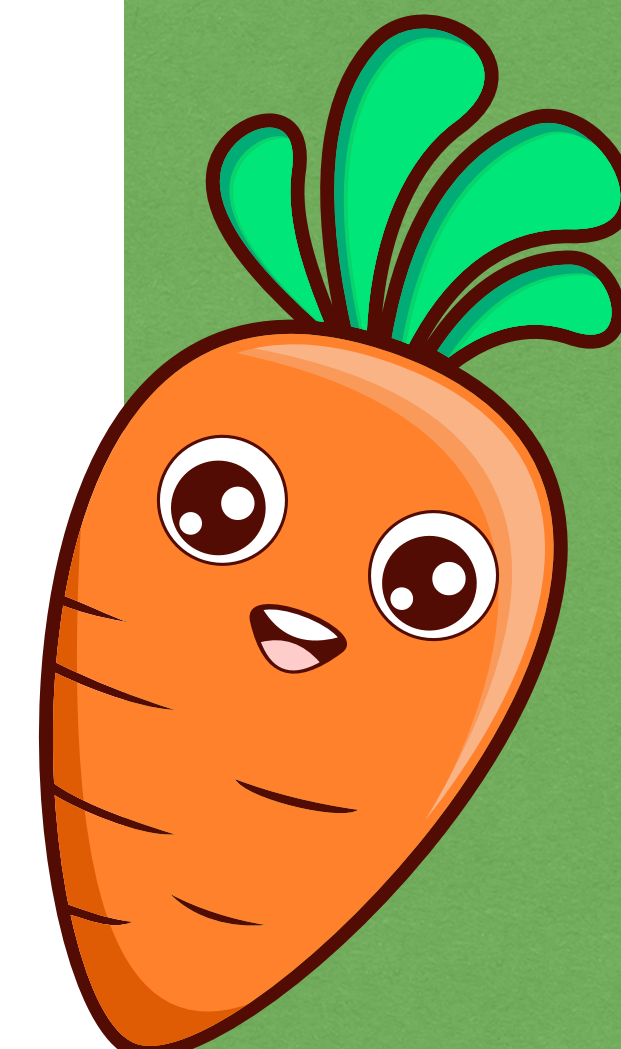


Generalization of the mappings on Yale participants

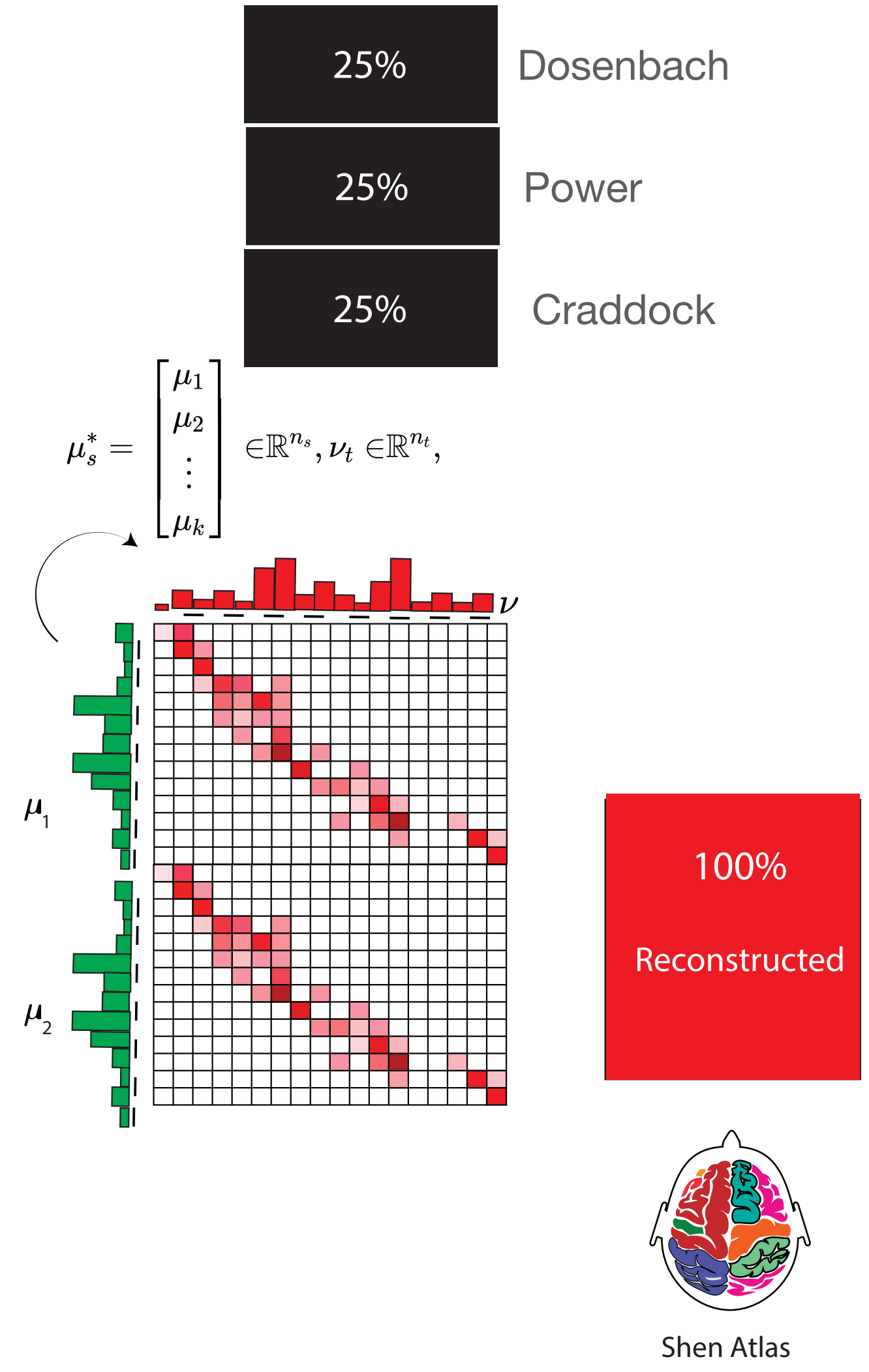
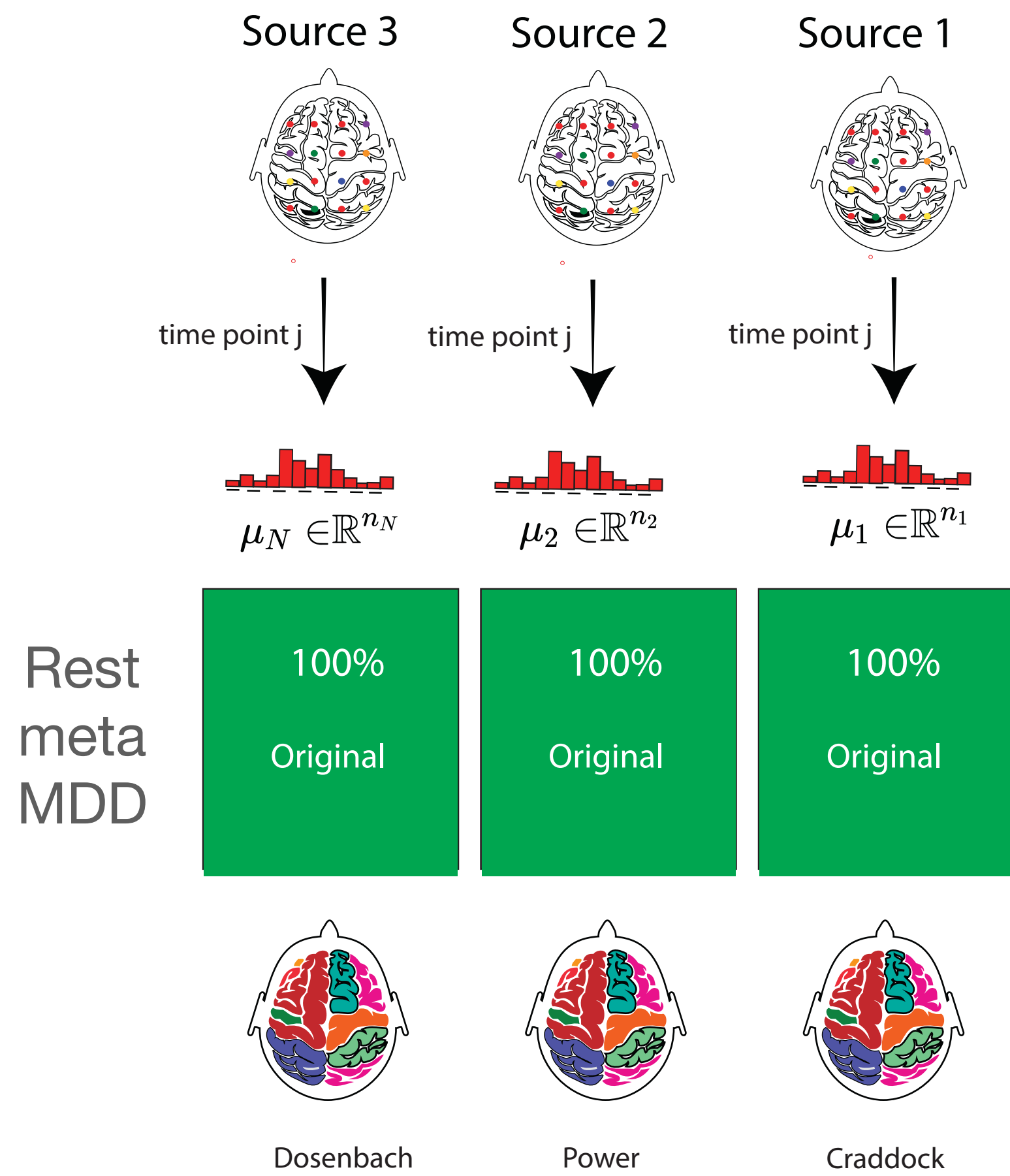


- We investigated if CAROT mappings trained in one dataset generalize to other datasets.
- We applied the mappings trained on HCP and reconstructed connectomes using the Yale dataset using these mappings.
- Spearman's rank correlation between the upper triangles of the connectomes was used to assess the similarity between the reconstructed and original connectomes.

A Second Dataset: Yale Participants

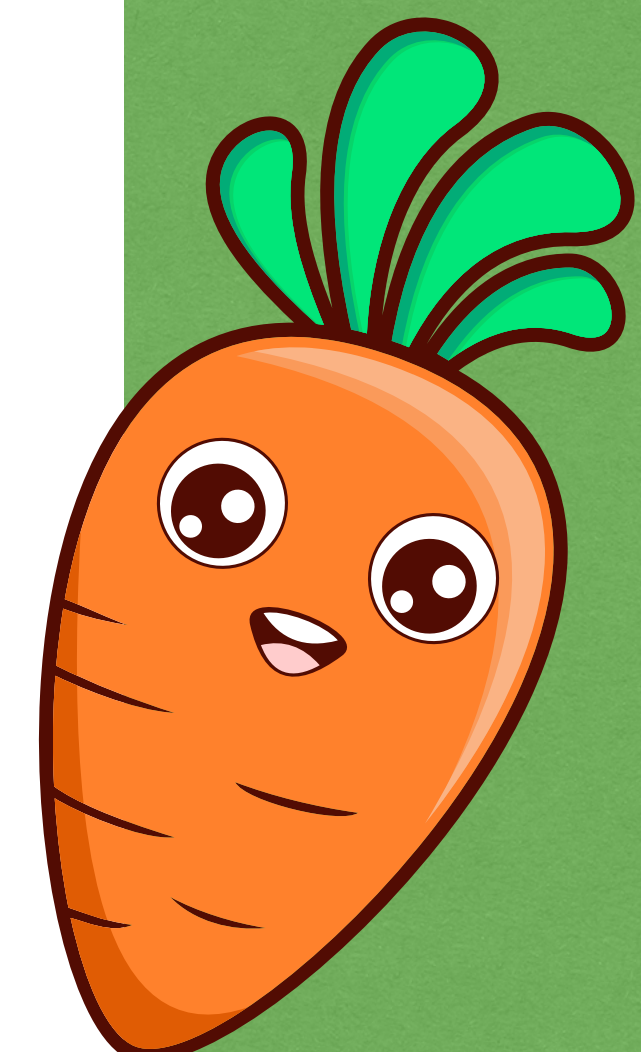


Can we test a model trained on Shen and try it on a large-scale project for which Shen is unavailable?

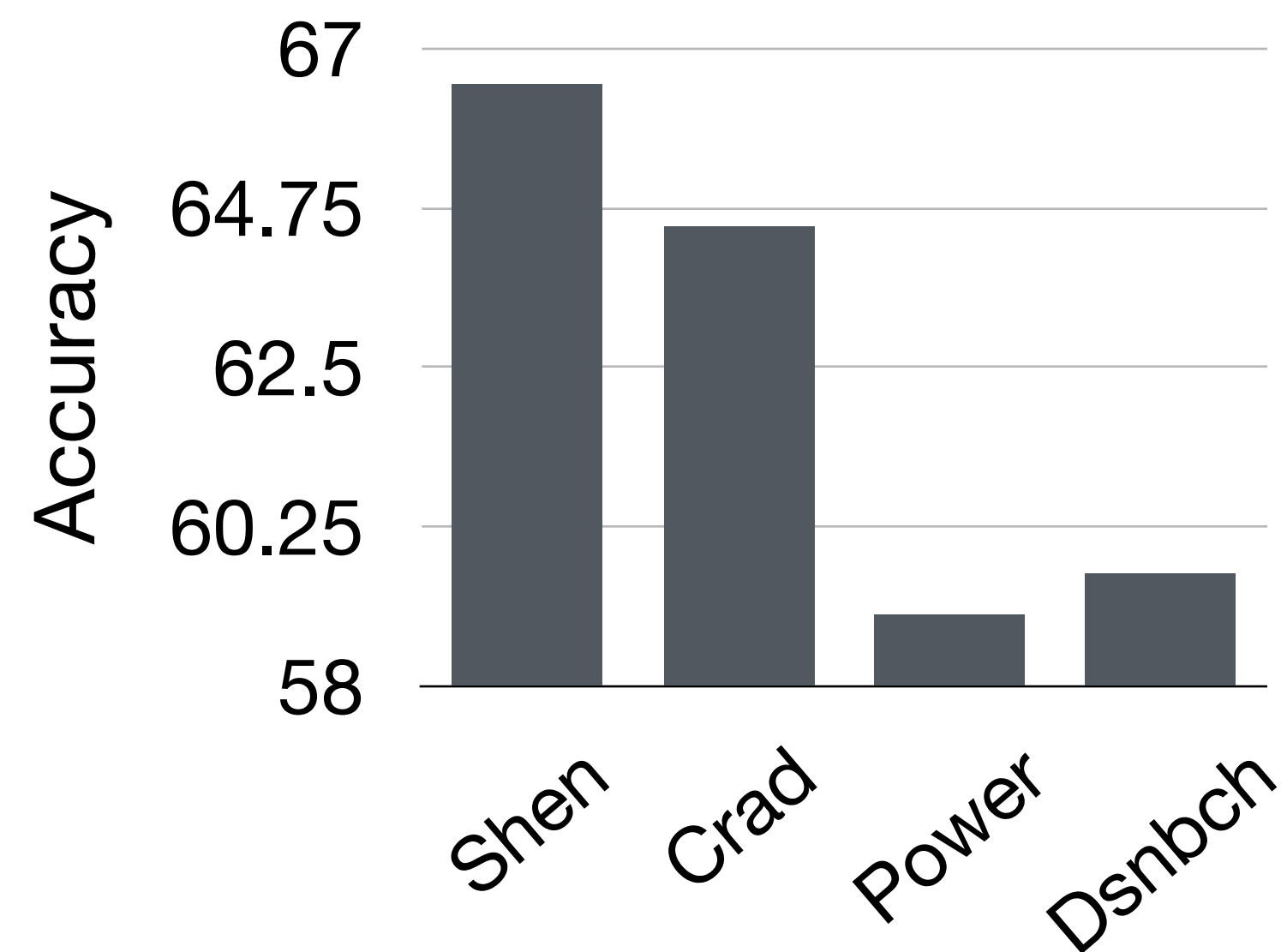


Review

Third Dataset: Sex classification trained on Yale

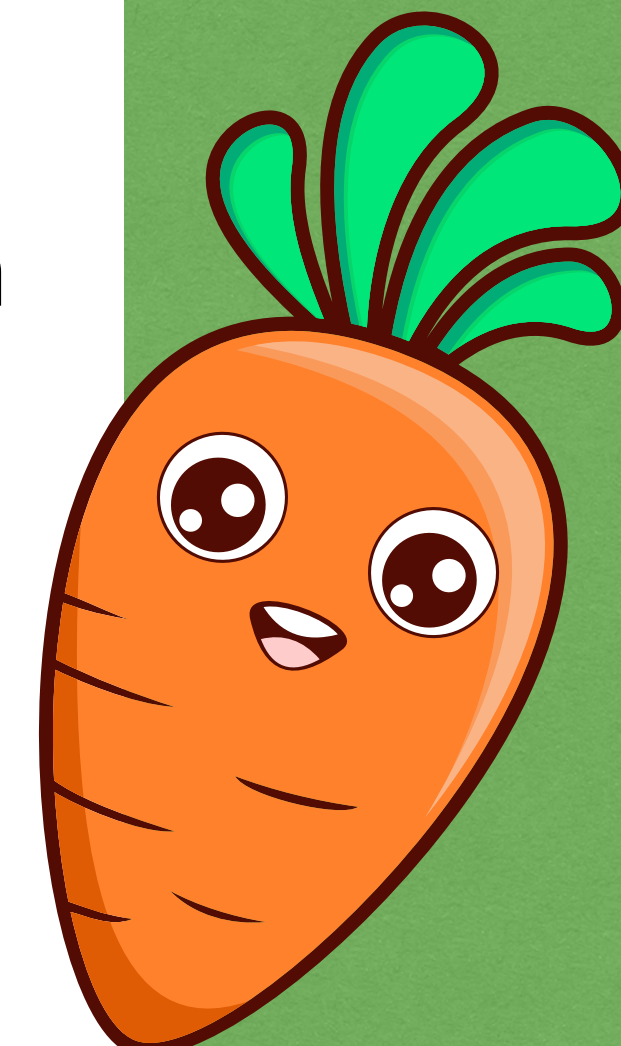


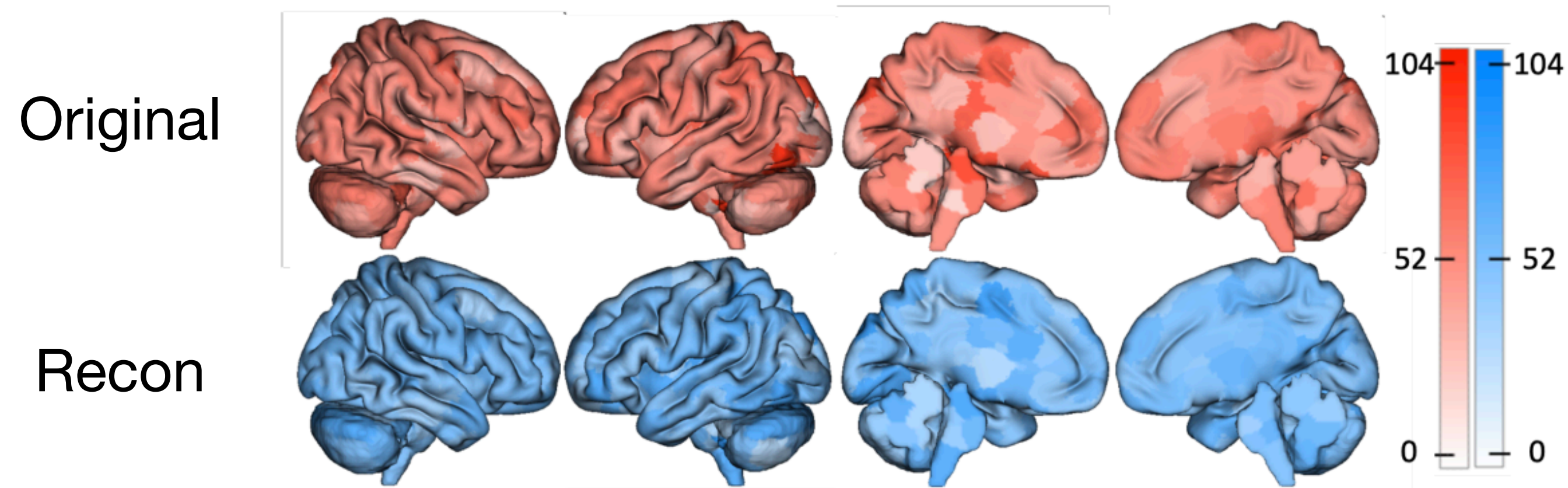
Generalization of the model on REST-MDD depression dataset



- In this evaluation, we generalize a sex classification model on Yale data:
 - The REST-Meta-MDD dataset (Yan et al., 2016) only provides preprocessed timeseries data from the Dosenbach, Power, and Craddock atlases.
- Overall, the sex classification model demonstrated significant classification accuracy in the Yale dataset (Accuracy=60.5% ; Naive model accuracy=50%;).
- Next, the sex classification model performed significantly better than chance in the REST-Meta-MDD dataset when using the reconstructed connectomes.

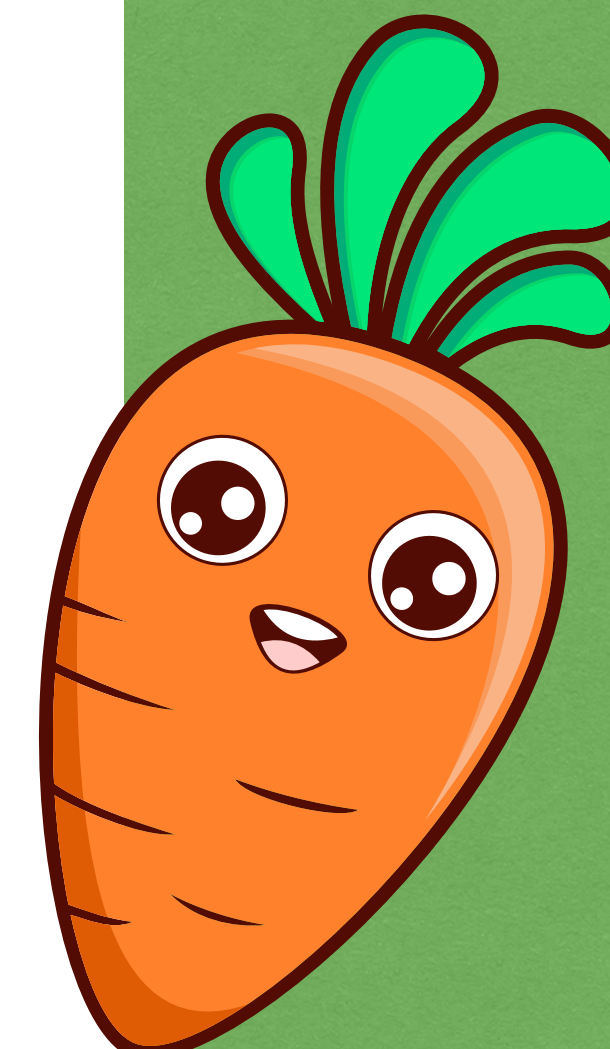
Third Dataset: Sex classification trained on Yale

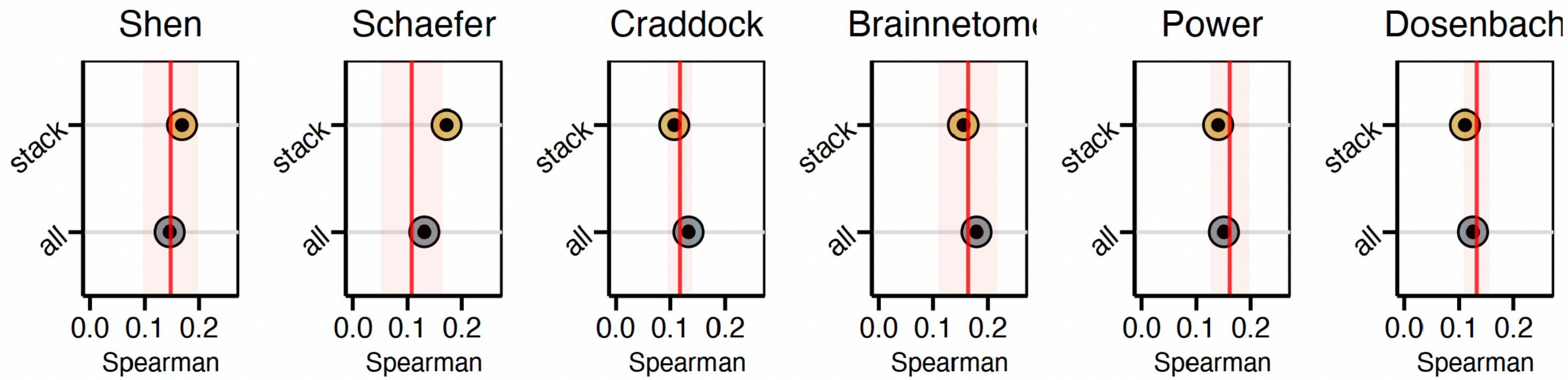
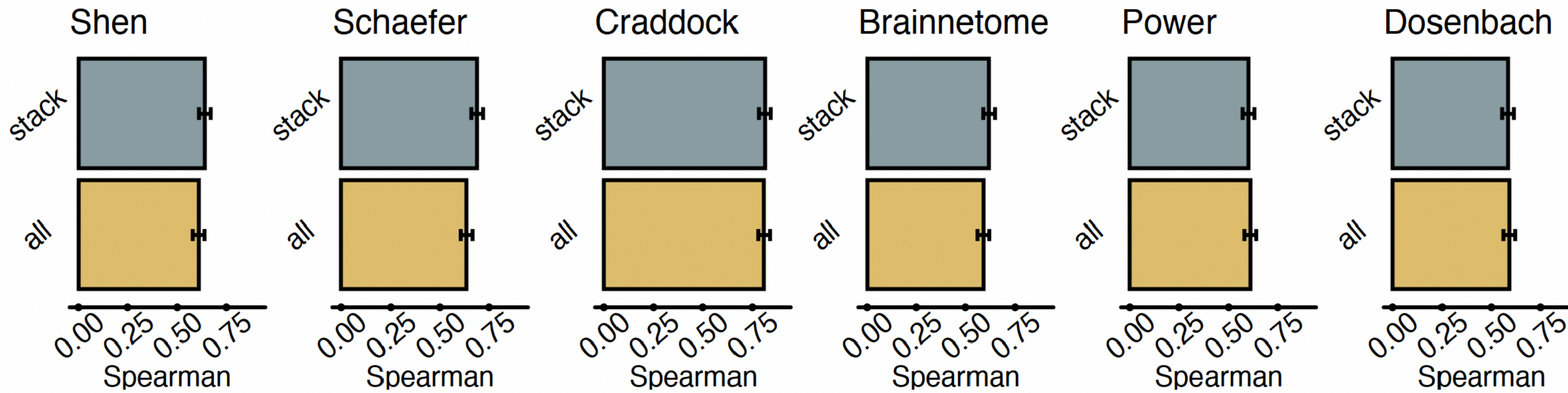




- Reconstructed connectomes give similar aging results as the original connectomes.
- These spatial maps correlate at $r = 0.61$, suggesting that analyses with the reconstructed connectomes produce the same neuroscientific insights as analyses with the original connectomes.

Explanatory Analysis: Age Differences in HCP



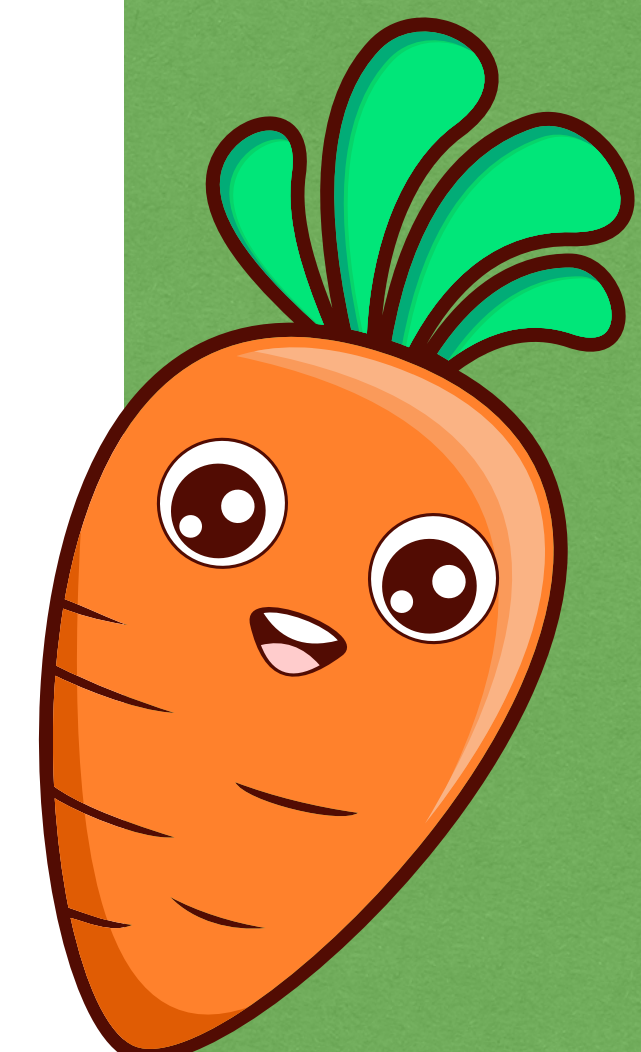


The number of possibilities to train T in CAROT:


$$\binom{n-1}{1} + \binom{n-1}{2} + \dots + \binom{n-1}{n-1} = 2^{n-1} - 1,$$

Equals the number of subsets of a set of size n

**One limitation,
Stacking
CAROT**



Source Time Series



$\mu \in \mathbb{R}^n$

Source Atlas(es)


[Upload Files](#)

Cross-Atlas Remapping via Optimal Transport

$$\arg \min_T C^T T - \epsilon H(T) \text{ s.t. } AT = \begin{bmatrix} \mu_t \\ \nu_t \end{bmatrix}$$

$$\mathcal{O}(n^2 \log(n) \eta^{-3})$$

Target Time Series



$\nu \in \mathbb{R}^m$

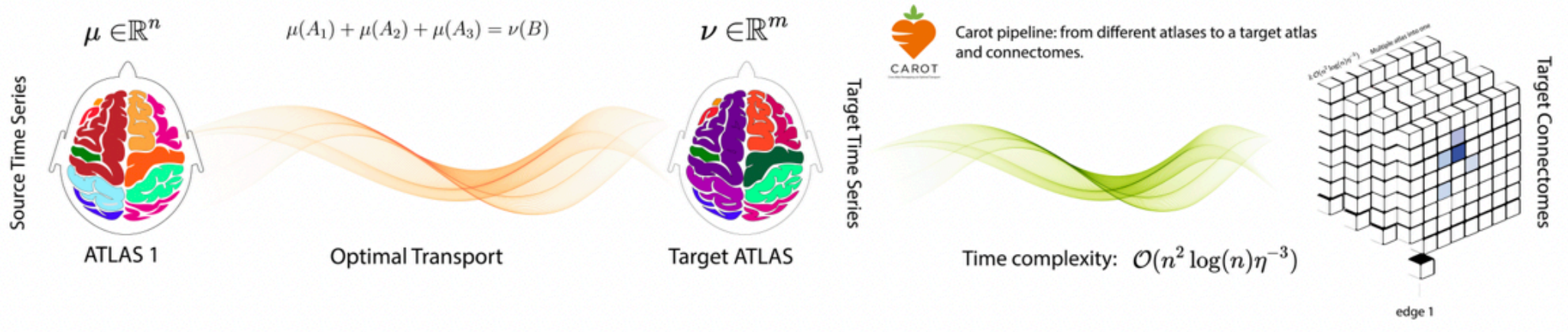
Target Atlas

Shen 268

[Reconstruct in Target Atlas](#)

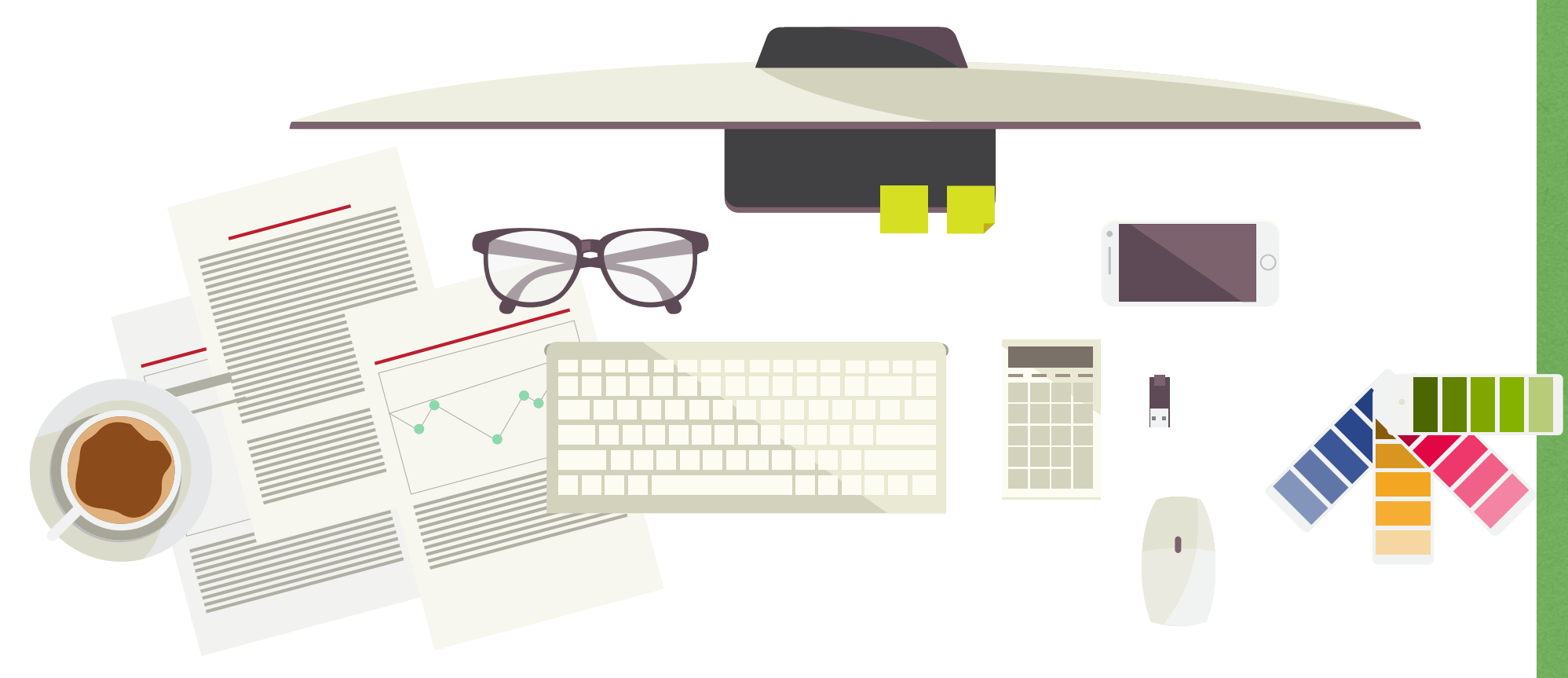
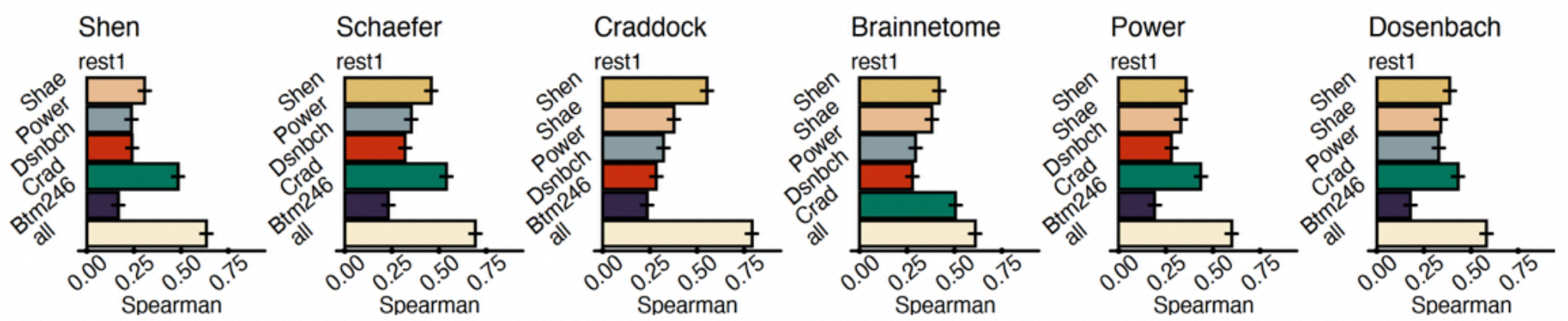
How it works?

Whether using large-scale projects---like the Human Connectome Project (HCP), the Adolescent Brain Cognitive Development (ABCD) study, Healthy Brain Network (HBN), and the UK Biobank---or pooling together several smaller studies, open-source, publicly available datasets allow for unrepresented sample sizes and promote generalization efforts. Overall, releasing preprocessing data can enhance participant privacy, democratize science, and lead to unique scientific discoveries. But releasing preprocessed data also limits the choices available to the end-user. For connectomics, this is especially true as connectomes created from different atlases (i.e., ways of dividing the brain into distinct regions) are not directly comparable. In addition, there exist several atlases with no gold standards, and more being developed yearly, making it unrealistic to have processed, open-source data available from all atlases. To address these limitations, we propose Cross Atlas Remapping via Optimal Transport (CAROT) to find a mapping between two atlases, allowing data processed from one atlas to be directly transformed into a connectome based on another atlas without needing raw data.



Quality of final connectomes

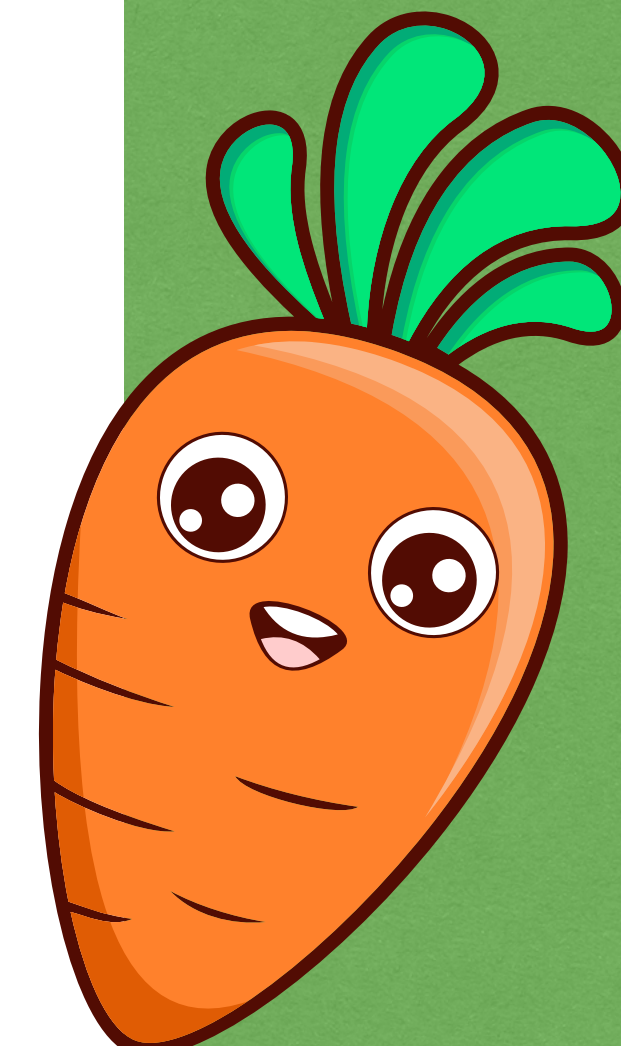
To validate CAROT, we compare reconstructed connectomes against their original counterparts (i.e., connectomes generated directly from an atlas), demonstrate the utility of transformed connectomes in downstream analyses, and show how a connectome-based predictive model can be generalized to publicly available processed data that was processed with different atlases. Overall, CAROT can reconstruct connectomes from an extensive set of atlases---without ever needing the raw data---allowing already processed connectomes to be easily reused in a wide-range of analyses while eliminating wasted and duplicate processing efforts. Using multiple source atlases improves the similarity of reconstructed connectomes. In the following figure the Spearman's rank correlation between of the reconstructed connectomes and connectomes generated directly with the target atlases are shown for each pair of source and target atlas as well reconstructed connectomes using all of the source atlases. For each of the target atlases, using all source atlases produces higher quality reconstructed connectomes. Error bars are generated from 100 iterations of randomly splitting the data into 25% for training and 75% for testing.



1. Our GitHub repository contains all the code necessary for specifying cost matrix, building mappings, and recreating functional connectivity for a given atlas: <https://github.com/dadashkarimi/carot>
2. The online demo supports six different atlases and entirely operates on a browser via javascript: <https://www.carotproject.com>

- In sum, CAROT allows a connectome generated based on one atlas to be directly transformed into a connectome based on another without needing raw data.
- These reconstructed connectomes are similar to and, in downstream analyses, behave like the original connectomes created from the raw data.
- Using CAROT on preprocessed open-source data will increase its utility, accelerate the use of big data, and help make a generalization and replication efforts easier.

Summary



1. Javid Dadashkarimi, Amin Karbasi, Qinghao Liang, Matthew Rosenblatt, Stephanie Noble, Maya Foster, Raimundo Rodriguez, Brendan Adkinson, Jean Ye, Huili Sun, Chris Camp, Michael Farruggia, Link Tejavibulya, Wei Dai, Rongtao Jiang, Angeliki Pollatou, and Dustin Scheinost, (2022)

[Cross Atlas Remapping via Optimal Transport \(CAROT\): Creating connectomes for any atlas when raw data is not available](#), **under review**

2. Javid Dadashkarimi, Amin Karbasi, and Dustin Scheinost, (2022)

[Combining multiple atlases to estimate data-driven mappings between functional connectomes using optimal transport](#), **MICCAI**

3. Qinghao Liang, Javid Dadashkarimi, Wei Dai, Amin Karbasi, Joseph Chang, Harrison H. Zhou, and Dustin Scheinost, (2022)

[Transforming connectomes to any parcellation via graph matching](#), **Best Paper in Graphs in Biomedical Image Analysis**

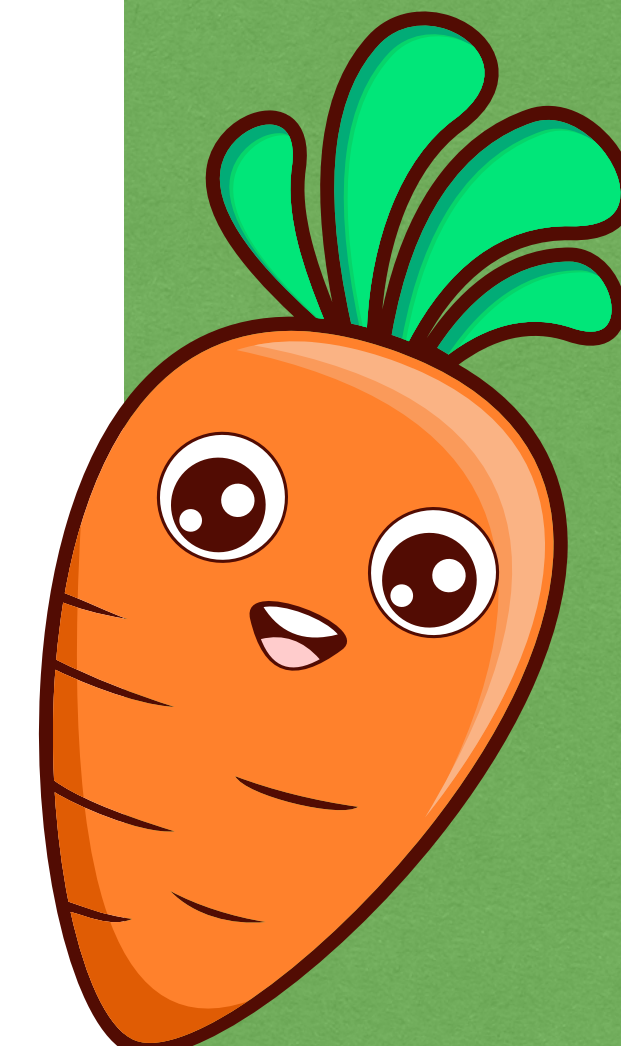
4. Javid Dadashkarimi, Amin Karbasi, and Dustin Scheinost, (2021)

[Data-driven mapping between functional connectomes using optimal transport](#), **MICCAI**

5. Javid Dadashkarimi, Siyuan Gao, Erin Yeagle, Stephanie Noble, Dustin Scheinost, (2019)

[A mass multivariate edge-wise approach for combining multiple connectomes to improve the detection of group differences](#), **Best Poster** in Connectomics in NeuroImage at **MICCAI**

Publications

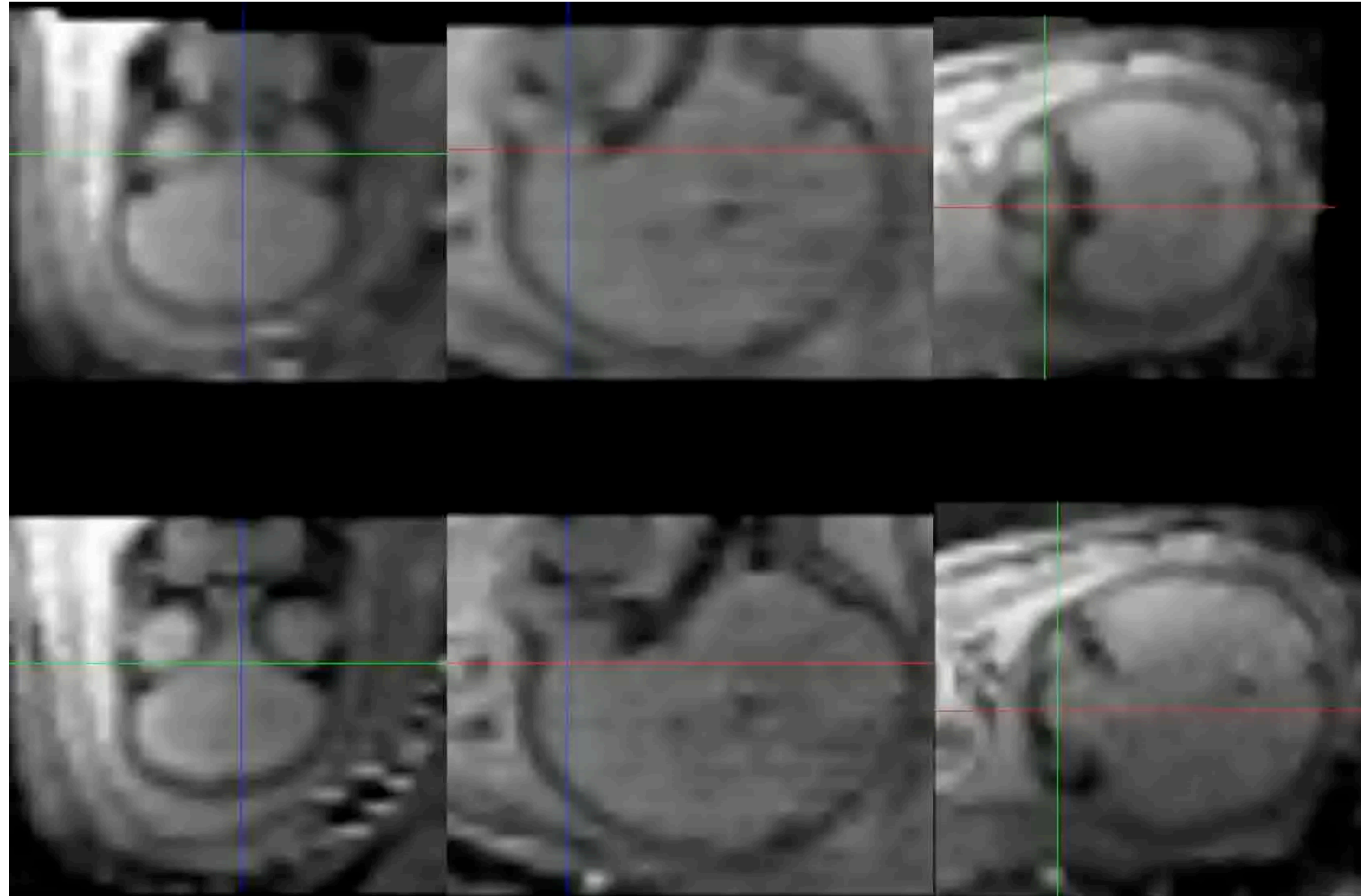


Thank you so much: MINDS lab and IID lab

- Dustin Scheinost
- Amin Karbasi
- Qinghao Liang
- Matthew Rosenblatt
- Stephanie Noble
- Raimundo Rodriguez
- Brendan Adkinson
- Huili Sun
- Jean Ye
- Maya Foster
- Chris Camp
- Michael Farruggia
- Link Tejavibulya
- Wei Dai
- Raina Vin
- AJ Simon
- Camille Duan
- Rongtao Jiang
- Angeliki Pollatou

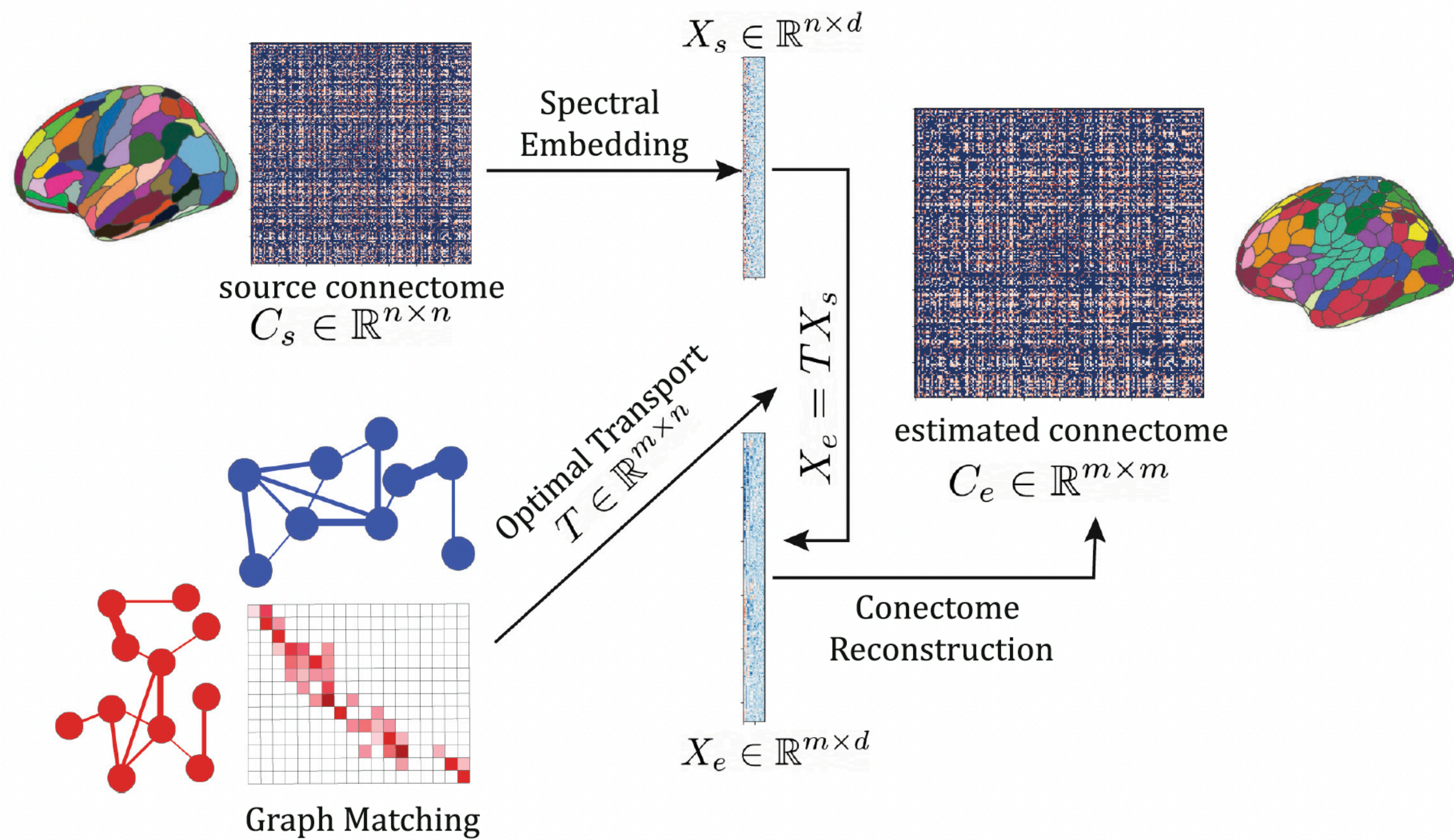


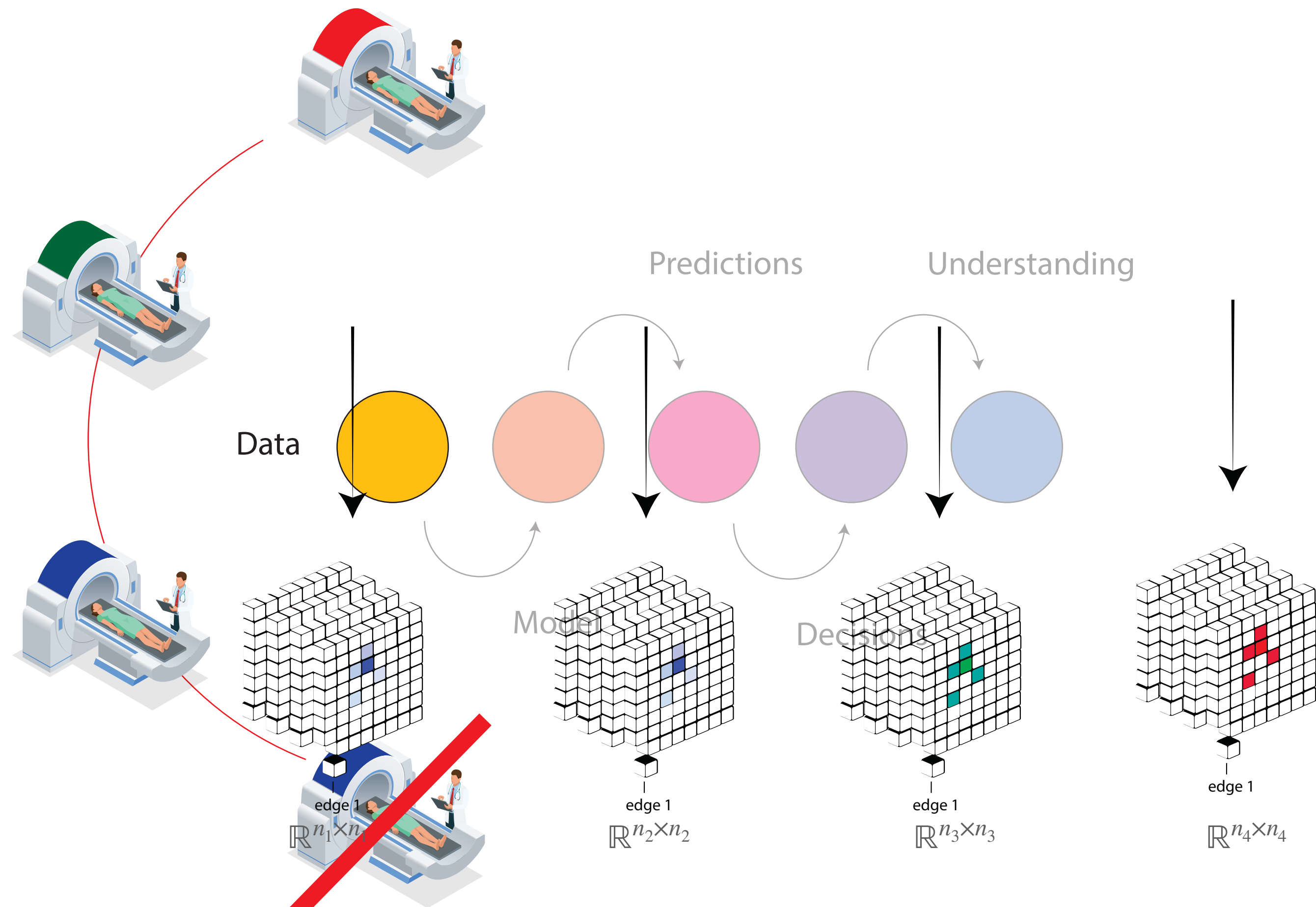
Functional MRI of 30 week fetus



Transforming Connectomes to “Any” Parcellation via Graph Matching

Qinghao Liang^{1(✉)}, Javid Dadashkarimi², Wei Dai³, Amin Karbasi^{2,4},
Joseph Chang⁵, Harrison H. Zhou⁵, and Dustin Scheinost^{1,5,6(✉)}



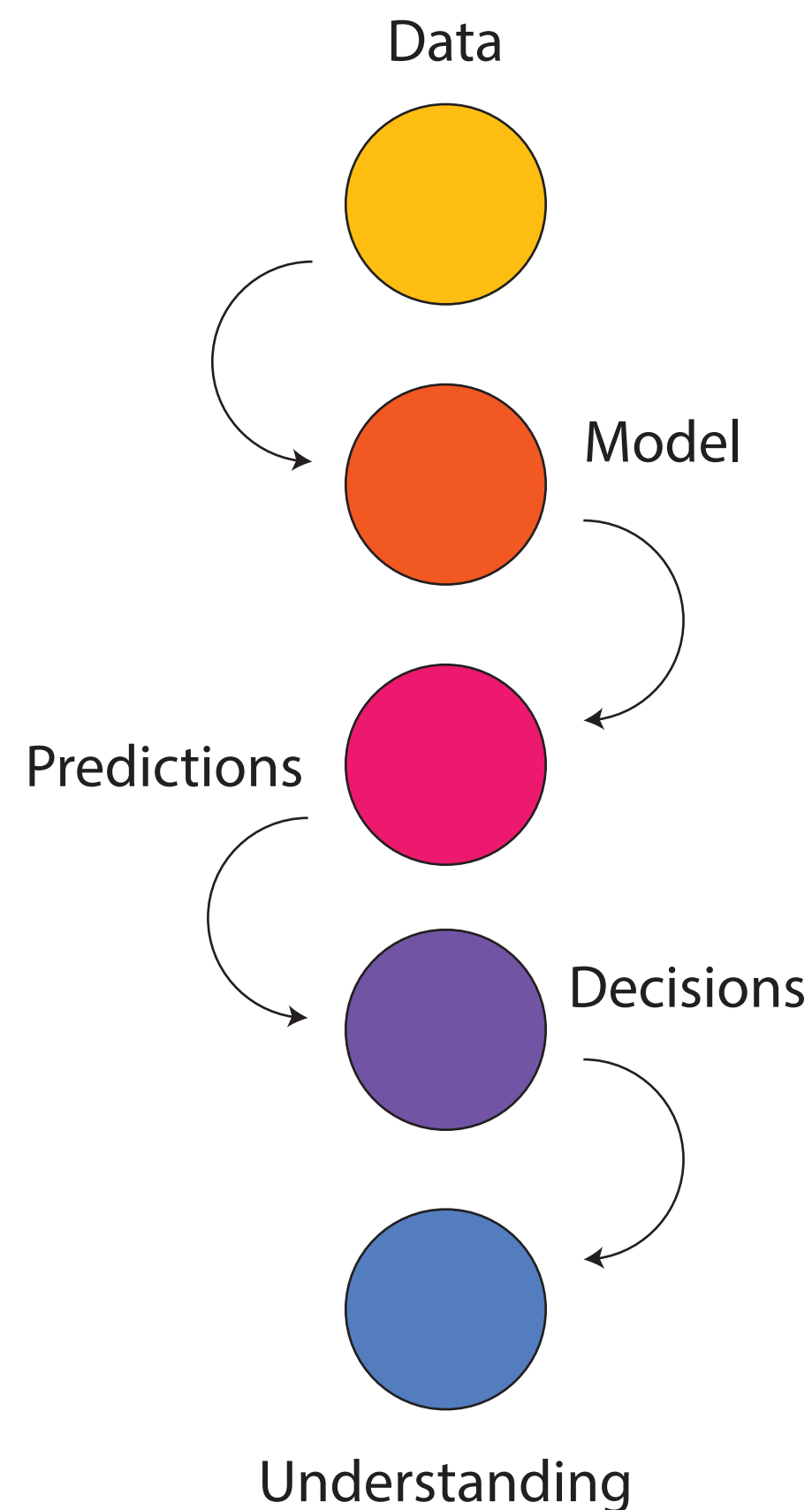


No gold atlas leads different functional connectomes across sites

Decentralized Neuroimaging Data



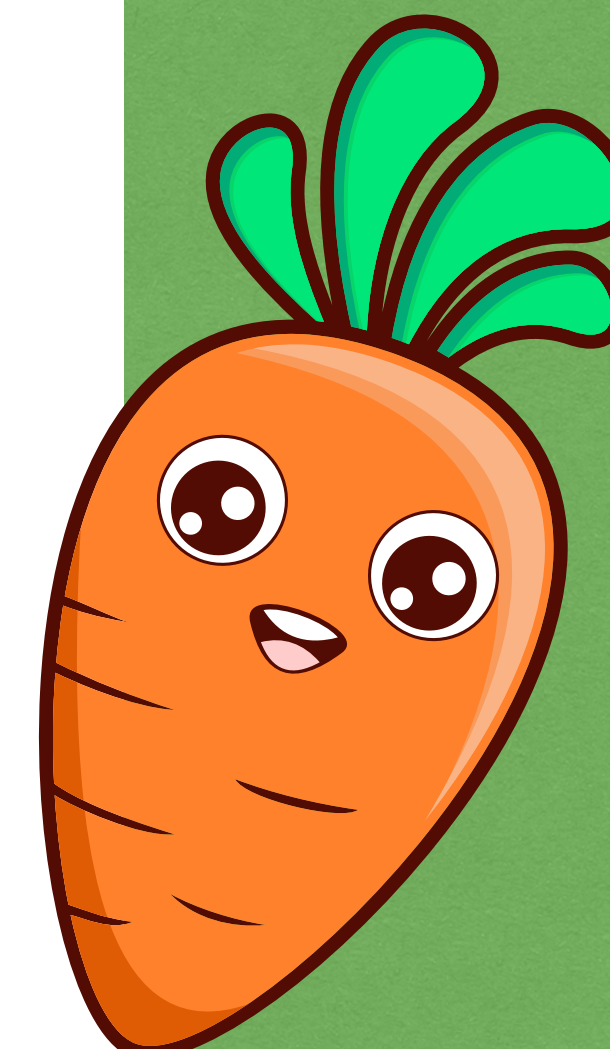
- Since several atlases exist with no gold standards, it is unrealistic to have processed, open-source data available from all atlases.
- These limitations directly inhibit the potential benefits of open-source neuroimaging data. Therefore we have this vastly large decentralized collection of data. Some of the with privacy concerns that are released in some limited set of atlases. Something that has been heavily neglected in our community.

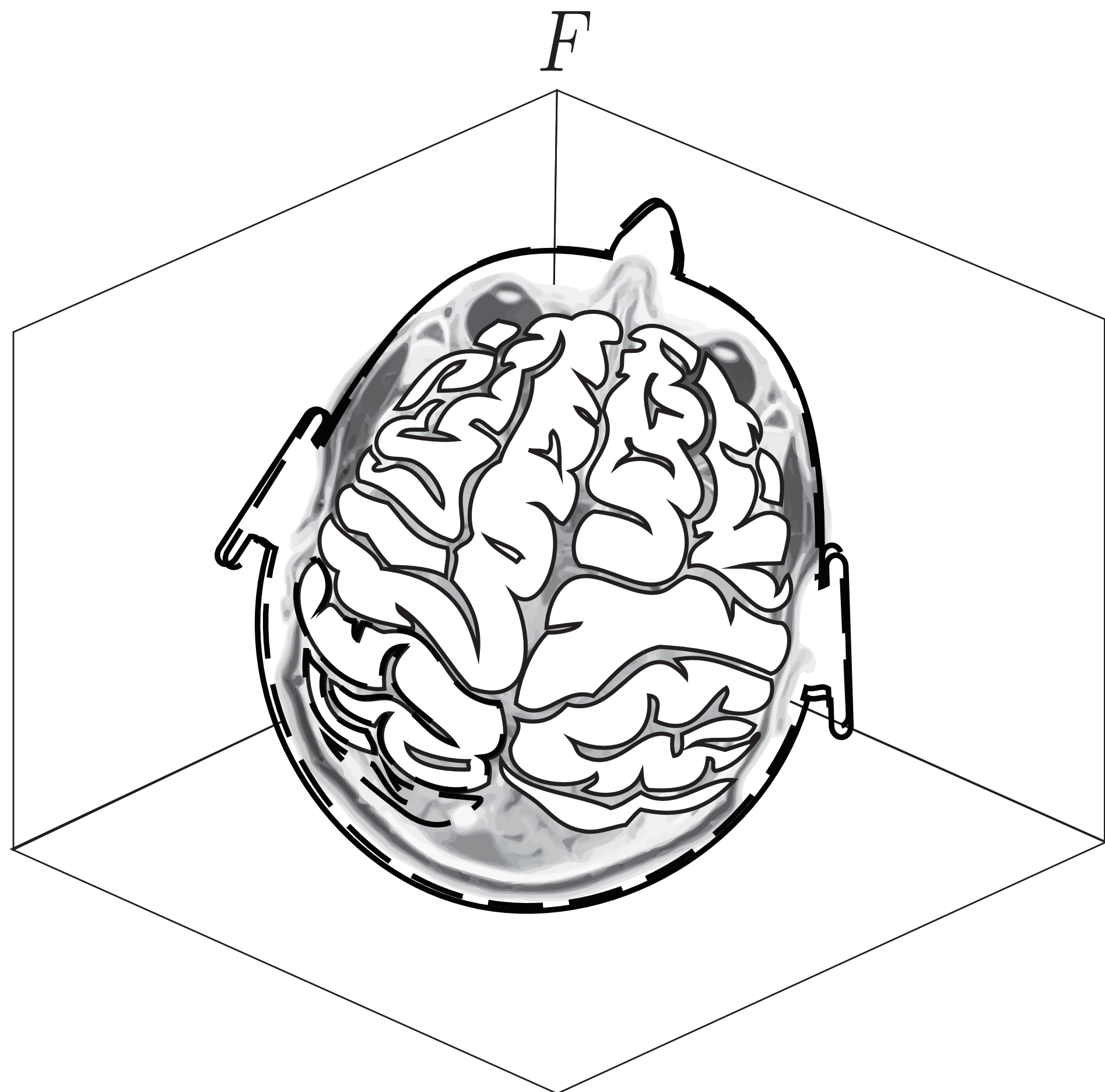


- Key Ingredients of data science: Data, model, predictions, decisions, and understanding
- Beyond data, everything else has uncertainty
- A model is the description of data that one can observe from a system.
 - There are all sorts of models in machine learning, but they vary in complexity, interpretability, and performance.
 - Depending on the application, one may prefer one over another
- High-risk decision-making systems are established in a way that is intelligible to no experts
- Eventually, we want to extend our understanding

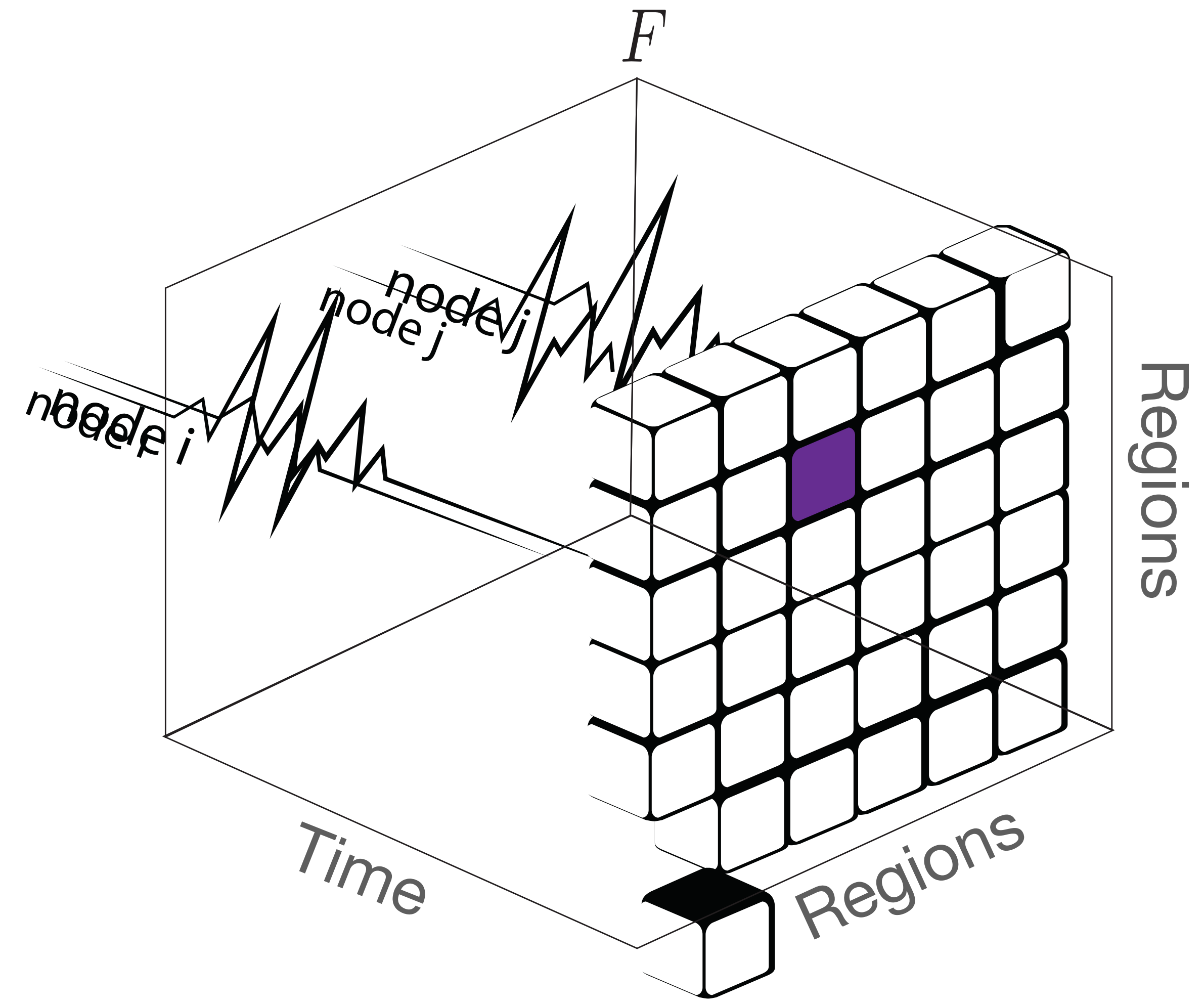
Key ingredients in data science, machine learning, and artificial

Data Science: key Ingredients of artificial





Anatomical Image



Functional Image