Cross Atlas Remapping via Optimal Transport (CAROT)

Javid Dadashkarimi

Advisors: Dustin Scheinost and Amin Karbasi

Committee members: Xenophon Papademetris and James Duncan

Computer Science Department
Yale University

Martinos Center for Biomedical Imaging
Massachusetts General Hospital
Boston, MA

March, 2023

Contact information:

javid.dadashkarimi@yale.edu

Biography

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

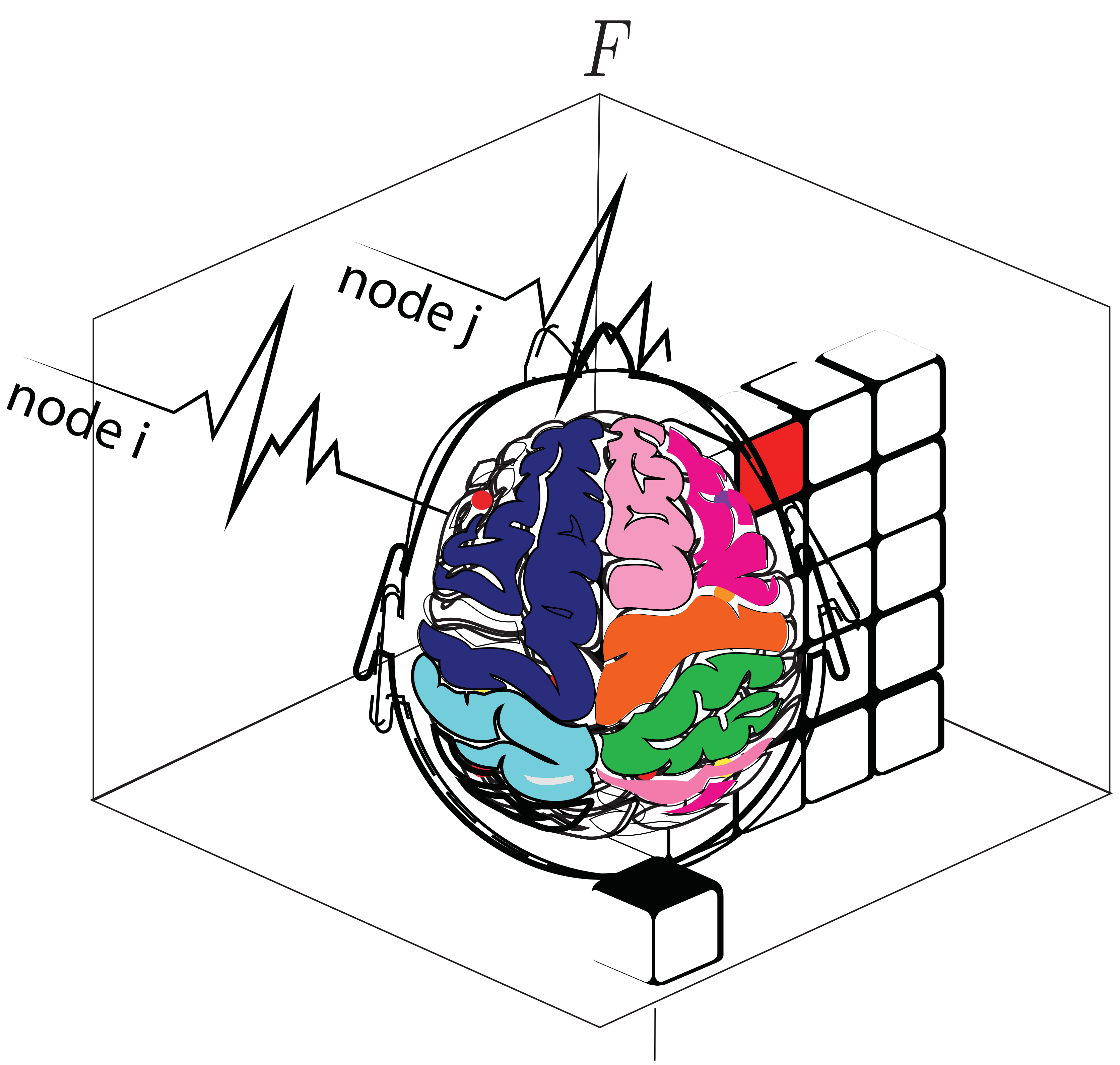
University of Tehran – Tehran, Iran
MEng in Software Engineering
BEng in Software Engineering
Mentors: Azadeh Shakery and Heshaam Faili
Thesis: *Dictionary-based Cross-lingual Information Retrieval*

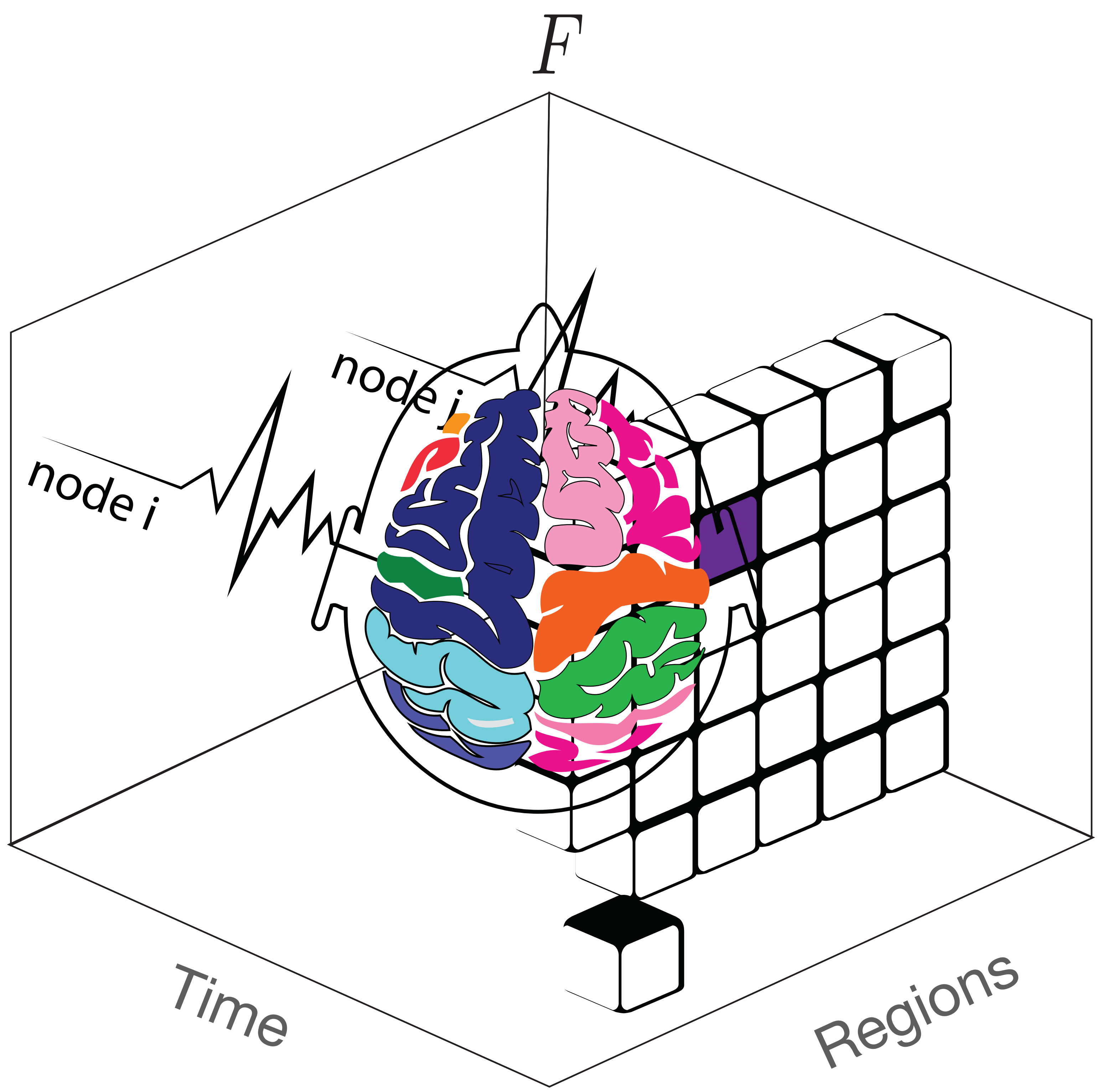
Functional Connectivity

Widely used in neuroscience to understand the functional organization of the brain.

1. What are a connectomes
2. How to make functional connectivity
3. Applications in neuroscience

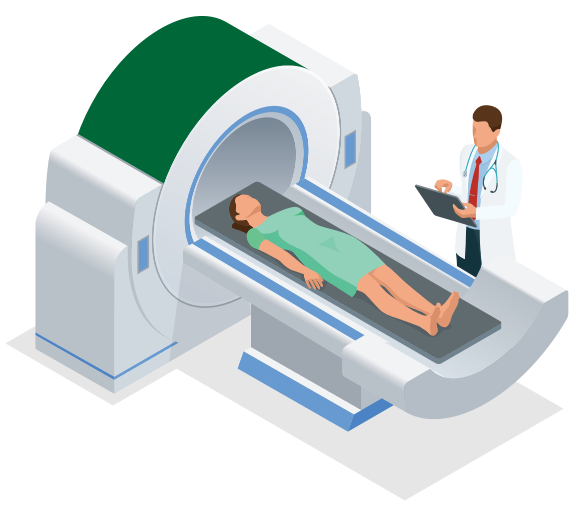
- **Definition:** A connectome—a matrix describing the connectivity between any pair of brain regions
  - Is a popular approach in neuroscience to study the brain's functional organization.
- **How to make**: They are created by parcellating the brain into distinct areas using an atlas and estimating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



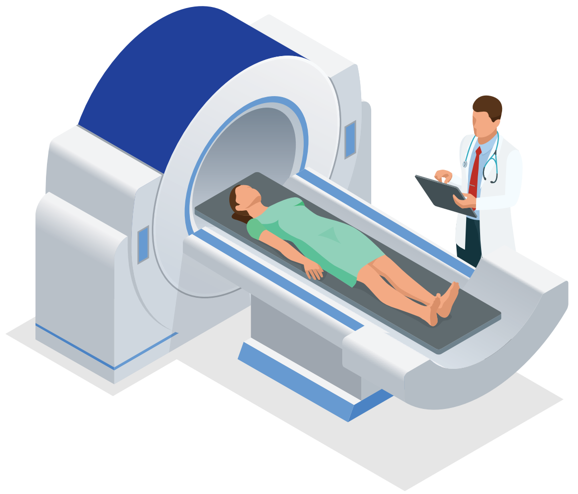


- **Steps** motion correction, skull stripping, registering to common space, and registering to the anatomical image are common steps before parcellations.

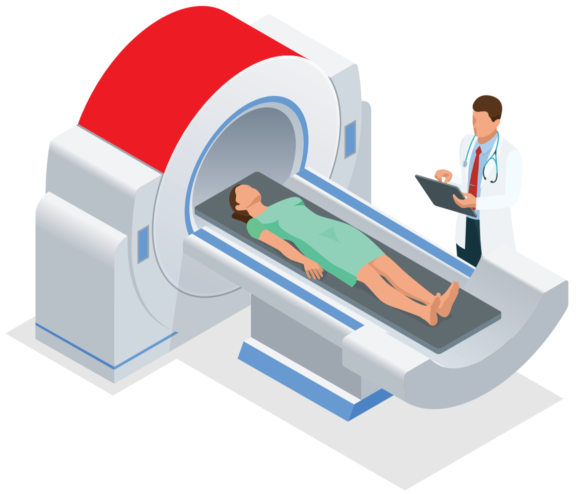
- The need for an atlas to create a connectome hinders comparisons across studies.
- Different atlases divide the brain into different regions of varying size and topology.
- Thus connectomes created from different atlases are not directly comparable.



- **Generalizability**:
  - Currently, no solutions exist to extend previous results to a connectome generated from a different atlas.
  - This prevents these datasets from being combined without reprocessing data.



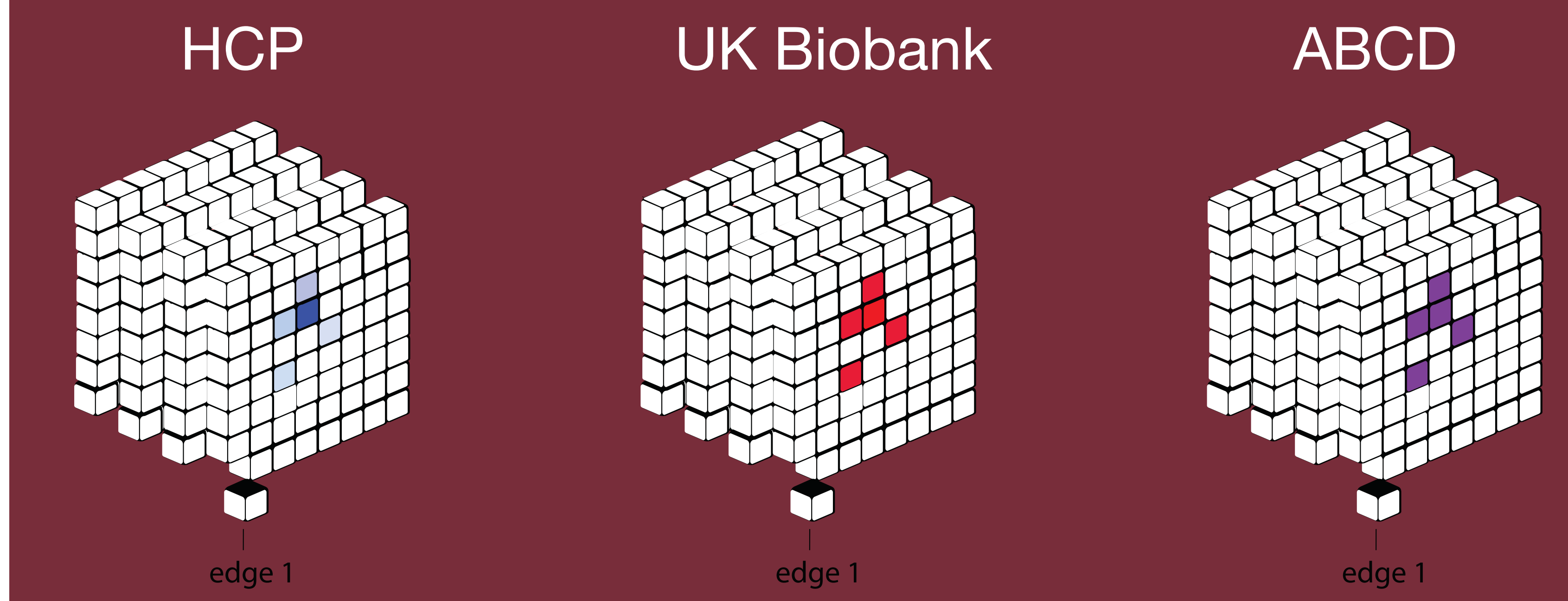
- **Storage concerns**:
  - Smaller labs might not have the resources to store and reprocess these data from scratch.
- **Privacy concerns**:
  - Due to privacy, some datasets are only released as fully processed parcellations.
  - Critically, in this case, it is not possible to go to the data to create connectomes from another atlas.



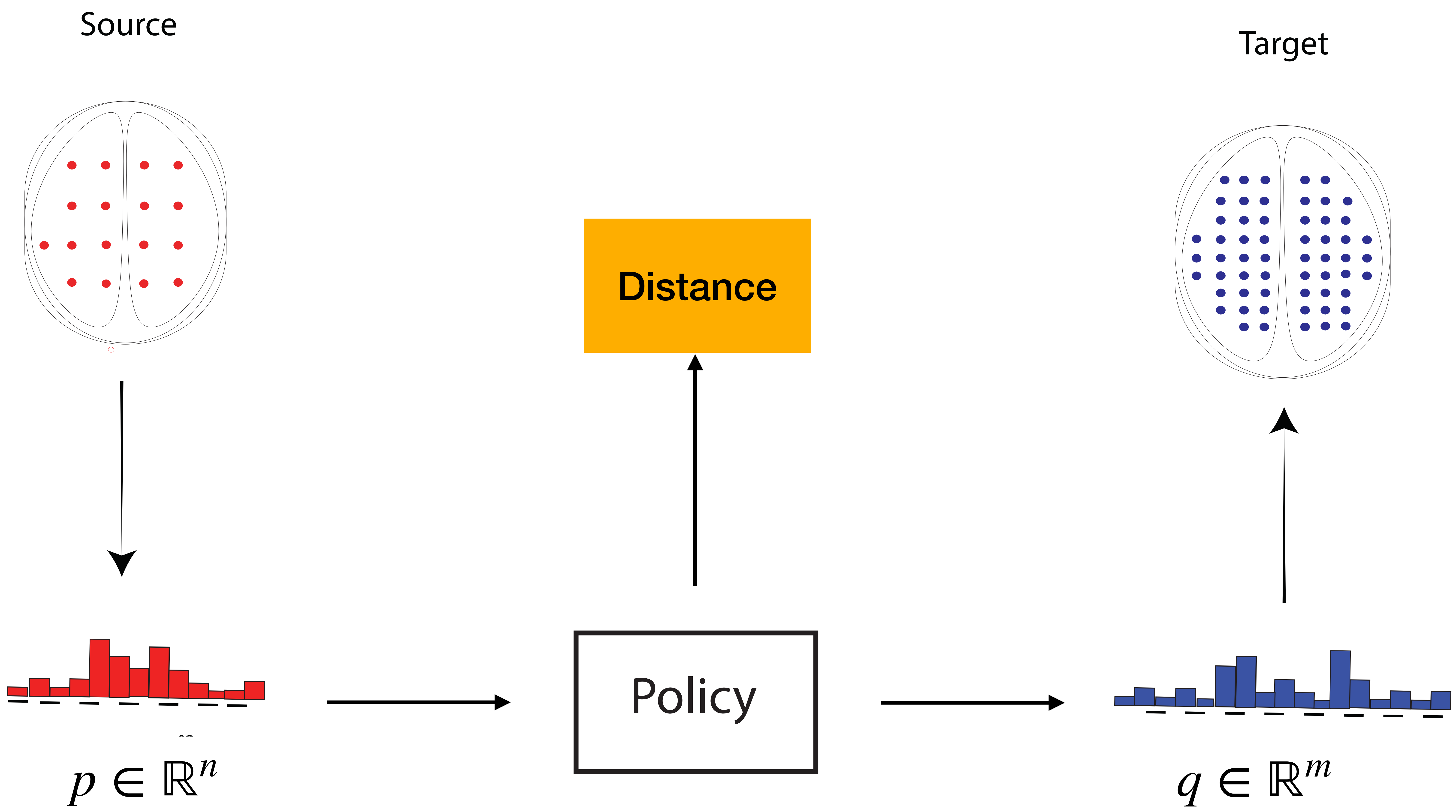
Real-world challenges

Different studies have different standards and limitations

1. Generalizability
2. Storage concerns
3. Privacy concerns



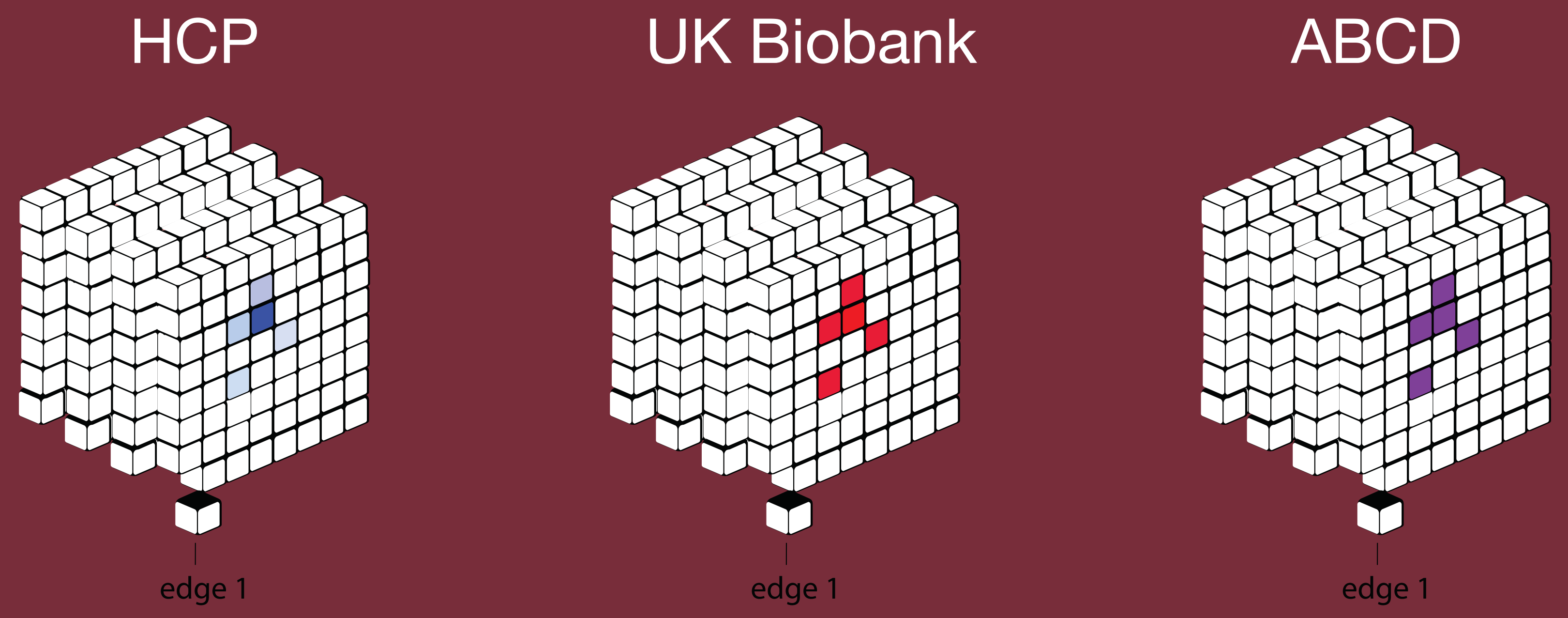
A moment-to-moment transportation method



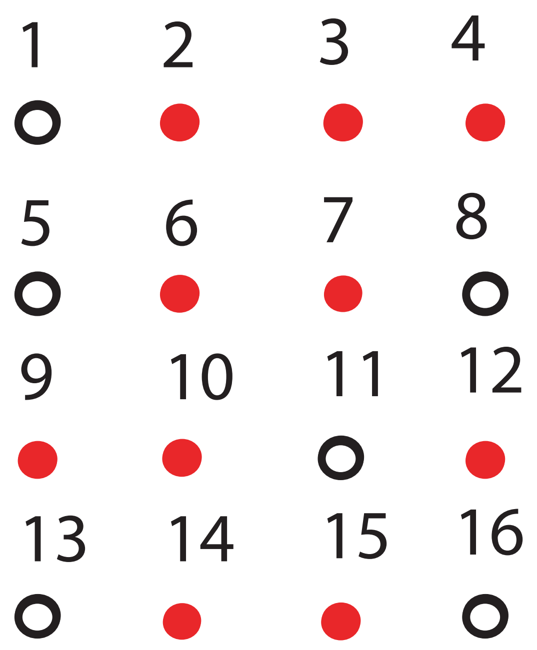
Our solution: dataset harmonization

Estimating connectomes in a missing form

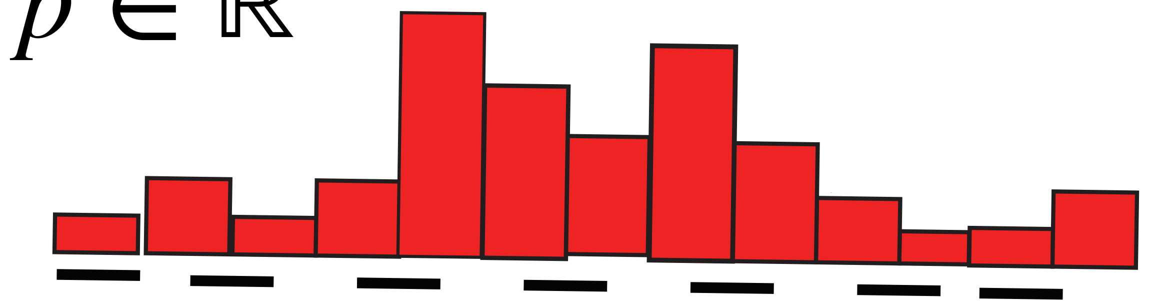
1. Time series-based approach
2. Transforming distribution of ROIs across atlases



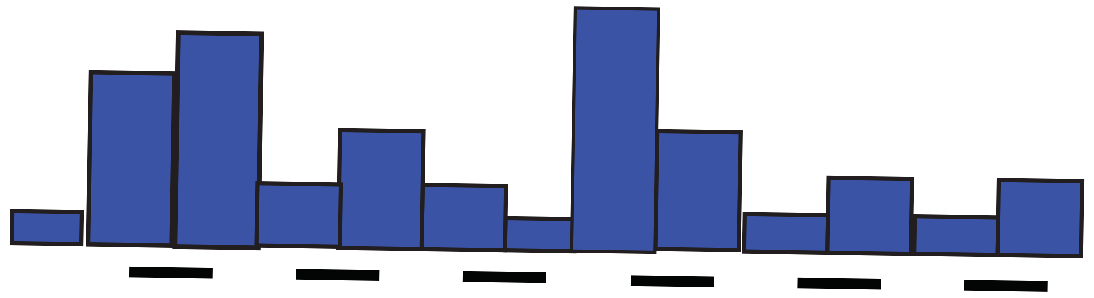
$m = n$



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



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



$$(1-p_1)p_2p_3p_4(1-p_5)p_6p_7p_8(1-p_8)p_9p_{10}(1-p_{11})p_{12}(1-p_{13})p_{14}p_{15}(1-p_{16})$$

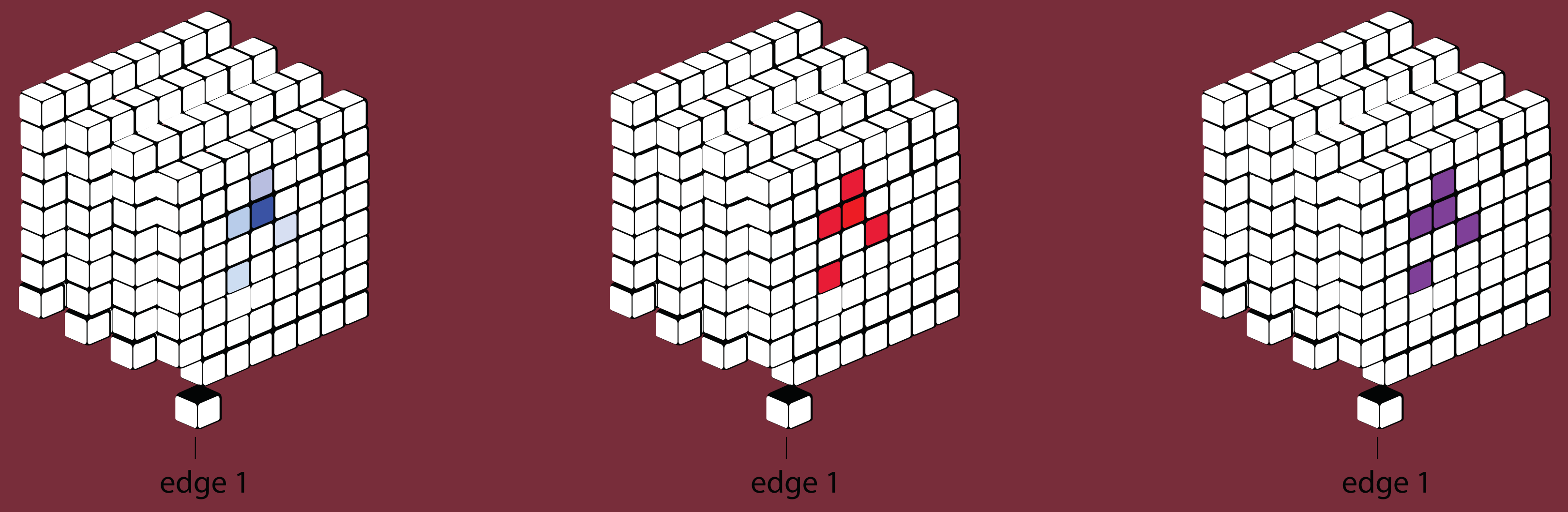
$$(1-q_1)q_2q_3q_4(1-q_5)q_6q_7q_8(1-q_8)q_9p_{10}(1-q_{11})q_{12}(1-q_{13})q_{14}q_{15}(1-q_{16})$$

Our solution: dataset harmonization

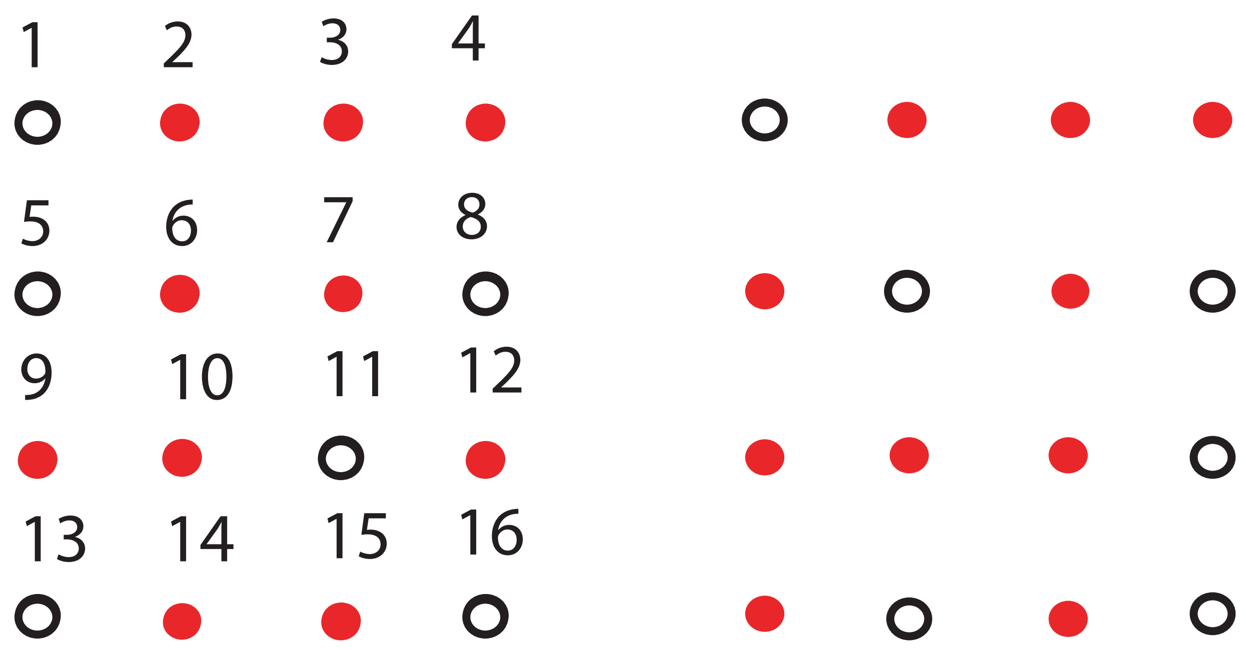
Estimating connectomes in a missing form

1. Time series-based approach
2. Transforming distribution of ROIs across atlases

HCP UK Biobank ABCD



$$m = n$$



$$(1-p_1)p_2p_3p_4(1-p_5)p_6p_7p_8(1-p_8)p_9p_{10}(1-p_{11})p_{12}(1-p_{13})p_{14}p_{15}(1-p_{16})$$

$$(1-p_1)p_2p_3p_4p_5(1-p_6)p_7(1-p_8)p_9p_{10}p_{11}(1-p_{12})p_{13}(1-p_{14})p_{15}(1-p_{16})$$

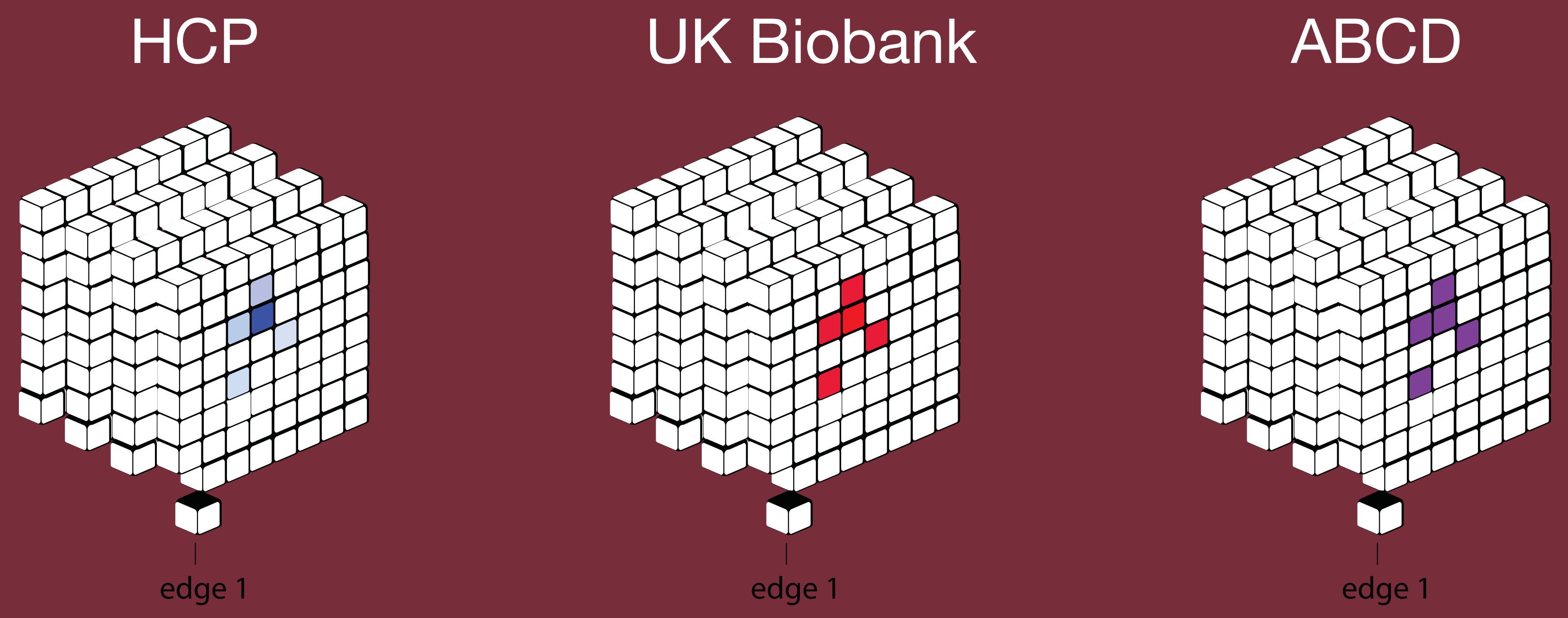
$$(1-q_1)q_2q_3q_4(1-q_5)q_6q_7q_8(1-q_8)q_9p_{10}(1-q_{11})q_{12}(1-q_{13})q_{14}q_{15}(1-q_{16})$$

$$(1-q_1)q_2q_3q_4q_5(1-q_6)q_7(1-q_8)q_9q_{10}q_{11}(1-q_{12})q_{13}(1-q_{14})q_{15}(1-q_{16})$$

Our solution: dataset harmonization

Estimating connectomes in a missing form

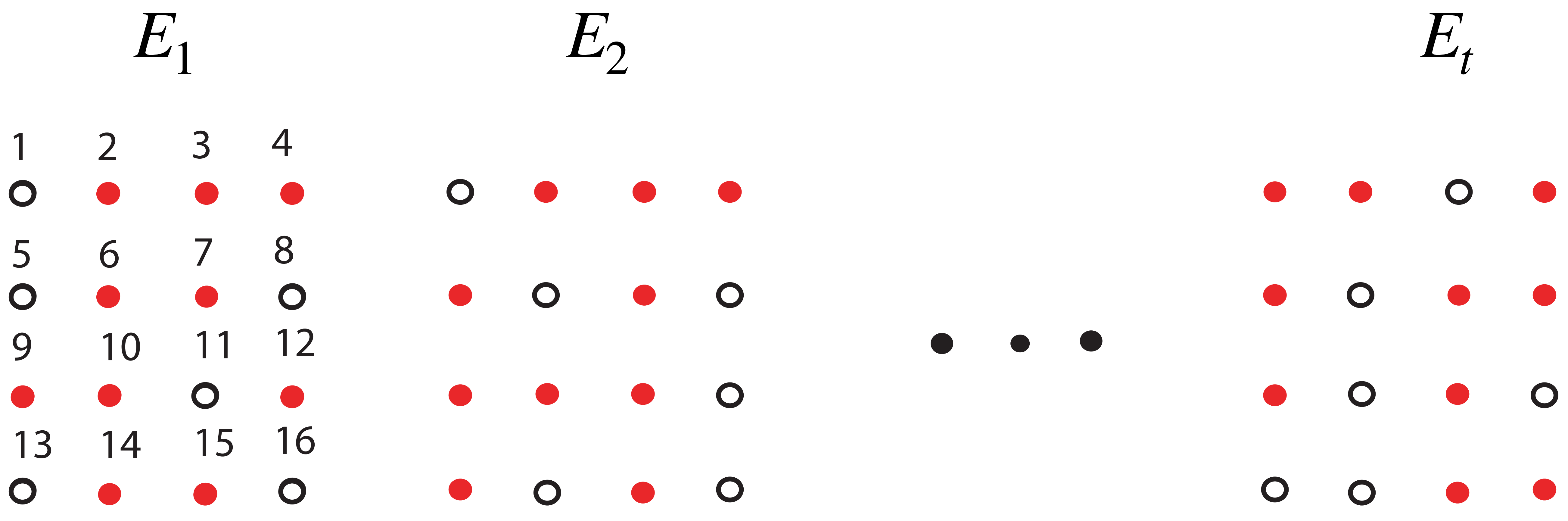
1. Time series-based approach
2. Transforming distribution of ROIs across atlases



Our solution: dataset harmonization

Estimating connectomes in a missing form

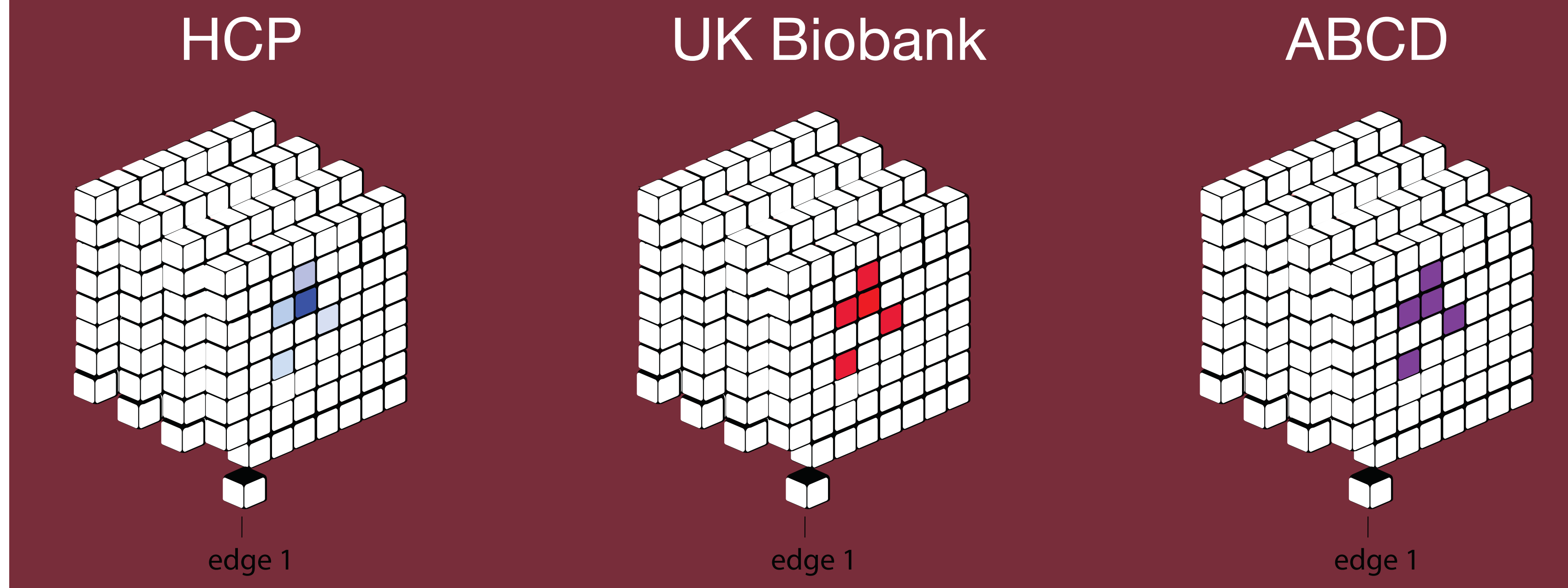
1. Time series-based approach
2. Transforming distribution of ROIs across atlases



$$p_1^{N_A}(1-p_1)^{N_I}\ p_2^{N_A}(1-p_2)^{N_I}\ \cdots\ p_{16}^{N_A}(1-p_{16})^{N_I}$$

$$q_1^{N_A}(1-q_1)^{N_I}\ q_2^{N_A}(1-q_2)^{N_I}\ \cdots\ q_{16}^{N_A}(1-q_{16})^{N_I}$$

$$\frac{P(\text{observation}\,|\,p)}{P(\text{observation}\,|\,q)} = \frac{p_1^{N_A^1}(1-p_1)^{N_I^1}\ p_2^{N_A^1}(1-p_2)^{N_I^1}\cdots p_{16}^{N_A^1}(1-p_{16})^{N_I^1}}{q_1^{N_A^1}(1-q_1)^{N_I^1}\ q_2^{N_A^1}(1-q_2)^{N_I^1}\cdots q_{16}^{N_A^1}(1-q_{16})^{N_I^1}}$$



$$\text{normalized relative likelihood} = \left(\frac{p_1^{N_A^1} p_2^{N_A^2} \ldots (1-p_{16})^{N_I}}{q_1^{N_A^1} q_2^{N_A^2} \ldots (1-q_{16})^{N_I}}\right)^{\frac{1}{N}}$$

$$= \frac{1}{N} \log\left(\frac{p_1^{N_A} p_2^{N_A} \ldots (1-p_{16})^{N_I}}{q_1^{N_A} q_2^{N_A} \ldots (1-q_{16})^{N_I}}\right)$$

$$= \frac{1}{N} \log p_1^{N_A^1} + \quad \frac{1}{N} \log(1-p_1)^{N_I^1} \quad \ldots \quad -\frac{1}{N} \log q_{16}^{N_A} \quad -\frac{1}{N} \log(1-q_{16})^{N_I}$$

$$= \frac{N_A^1}{N} \log p_1 + \quad \frac{N_I^1}{N} \log(1-p_1) \quad \ldots \quad -\frac{N_A}{N} \log q_{16} \quad -\frac{N_I}{N} \log(1-q_{16})$$

$$= p_1 \log p_1 \; + (1-p_1)\log(1-p_1) \cdots -q_{16} \log q_{16}$$

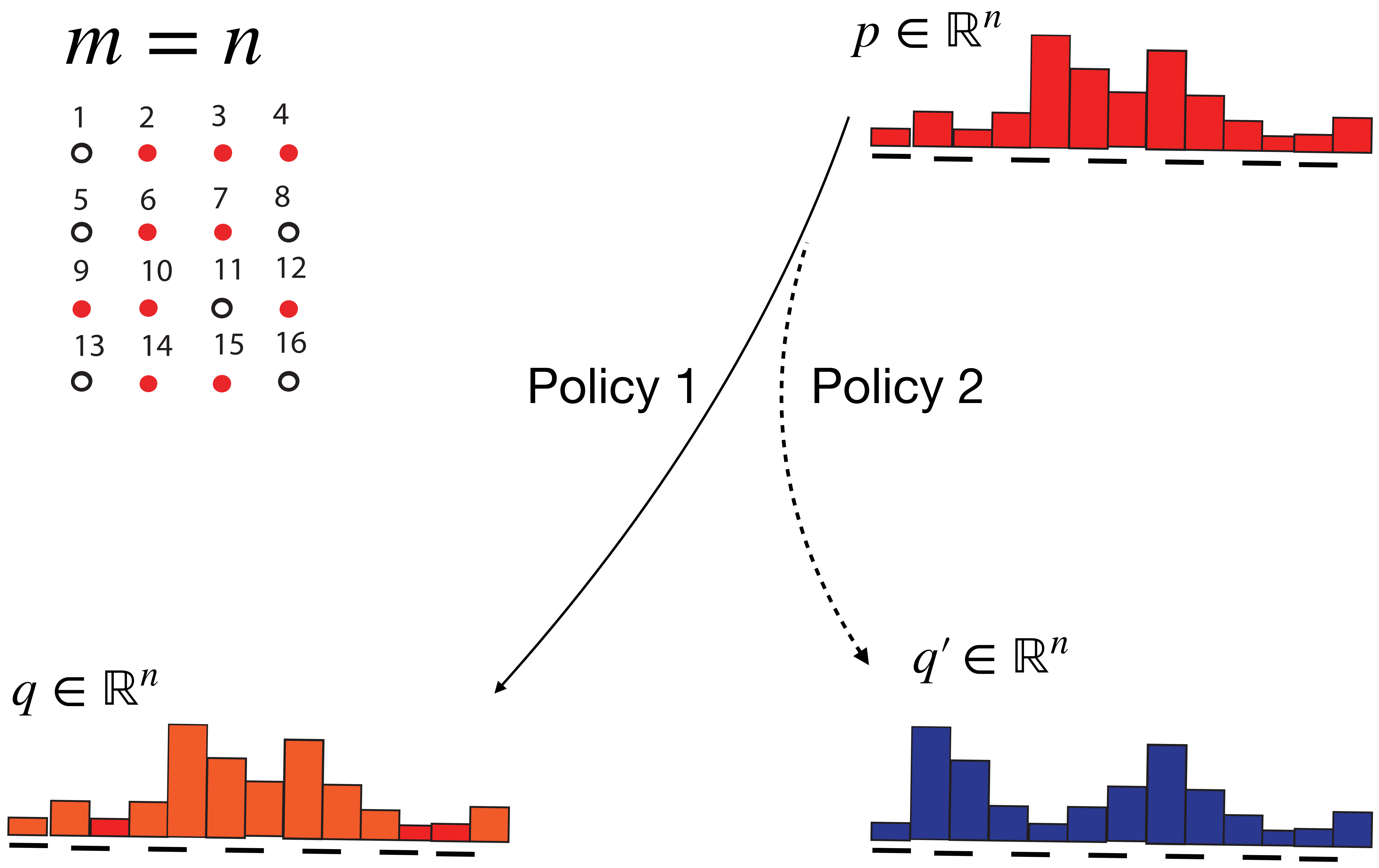
$$-(1-q_{16})\log(1-q_{16})$$

$$= D_{KL}(p\,||\,q) = \sum_{x \in \mathcal{X}} p(x) \log\left(\frac{p(x)}{q(x)}\right)$$

Kullback–Leibler divergence

Measures exactly the same thing

1. Log properties, product to addition, division to subtraction
2. How likely $q(x)$ would generate samples from $p(x)$

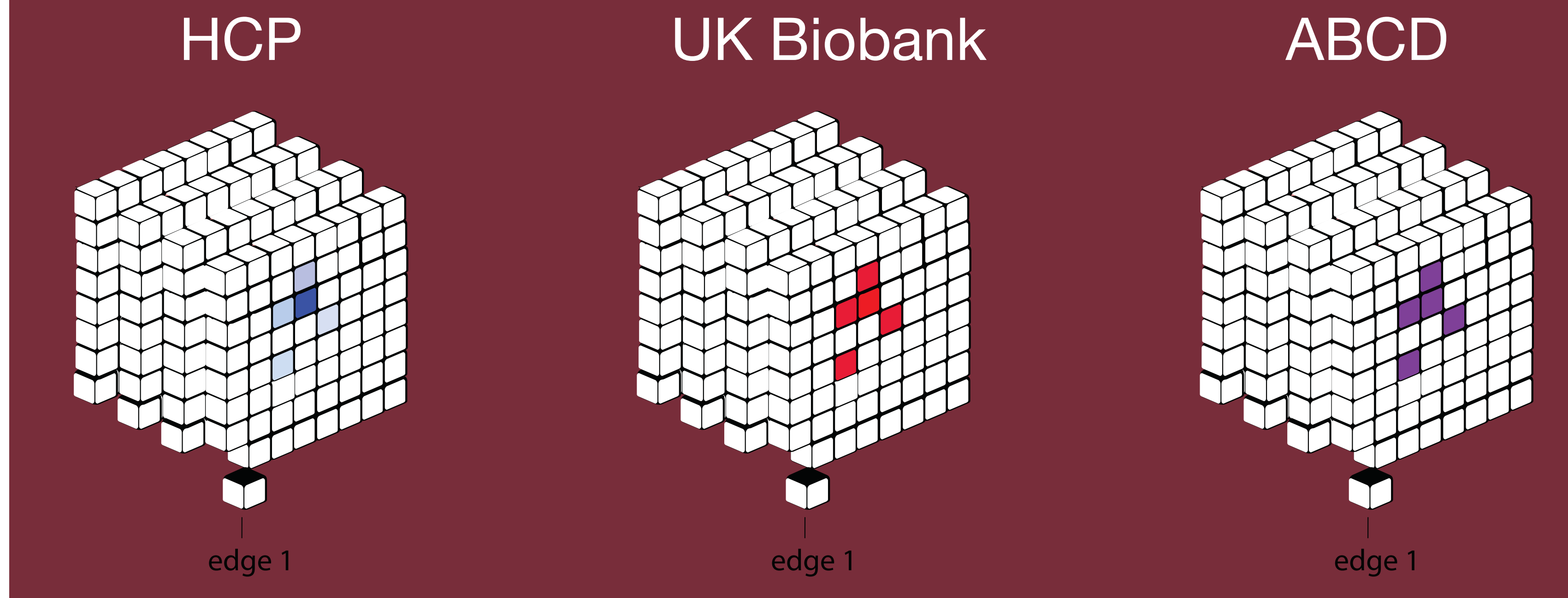


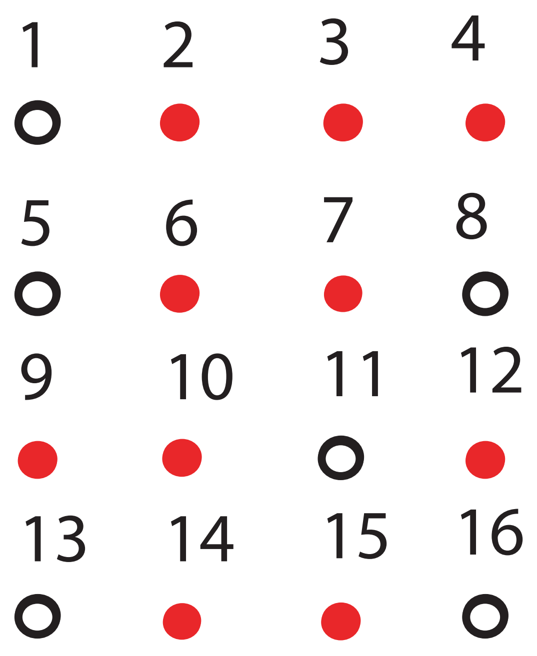
$$KL(p\,||\,q) < KL(p\,||\,q')$$

Our solution: dataset harmonization

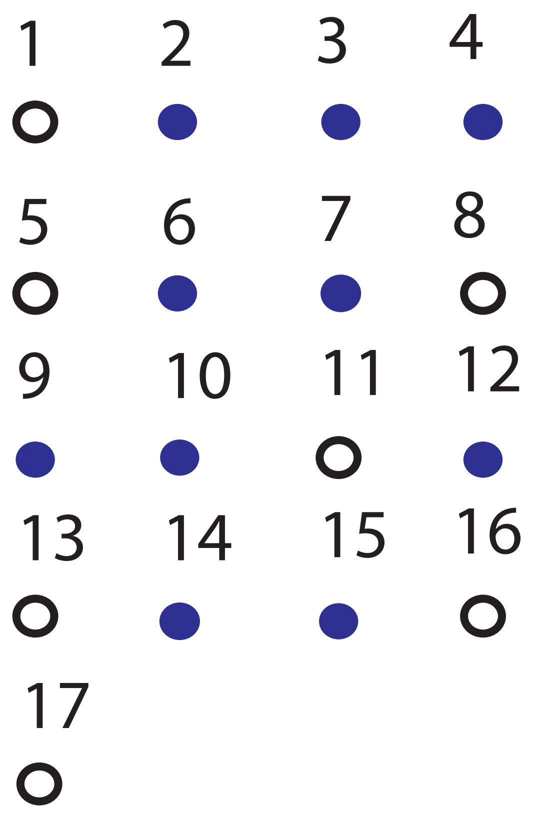
Estimating connectomes in a missing form

1. Time series-based approach
2. Transforming distribution of ROIs across atlases





How about when $m \neq n$



$$\frac{P(\text{observation} \mid p)}{P(\text{observation} \mid q)} = \frac{p_1^{N_A}(1-p_1)^{N_I}\, p_2^{N_A}(1-p_2)^{N_I} \cdots p_{16}^{N_A}(1-p_{16})^{N_I} \times 0}{q_1^{N_A}(1-q_1)^{N_I}\, q_2^{N_A}(1-q_2)^{N_I} \cdots q_{17}^{N_A}(1-q_{17})^{N_I}}$$

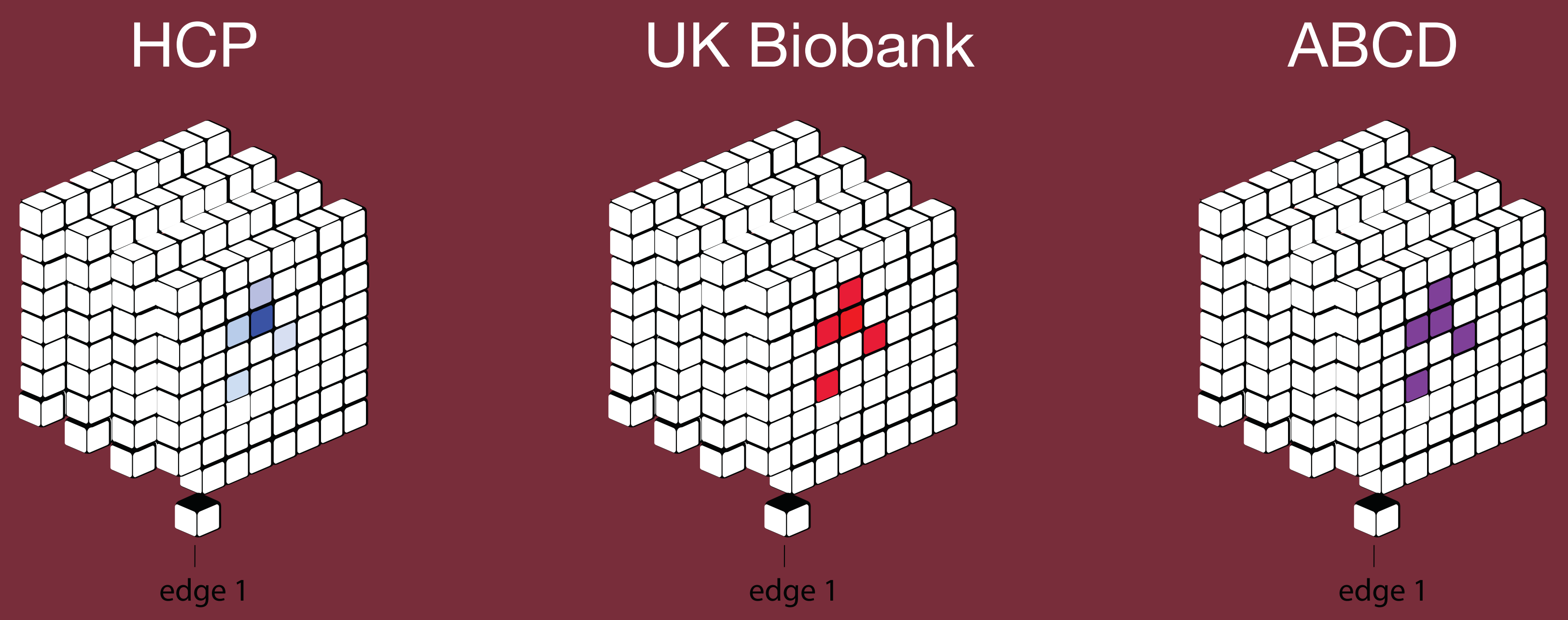
$$= 0$$

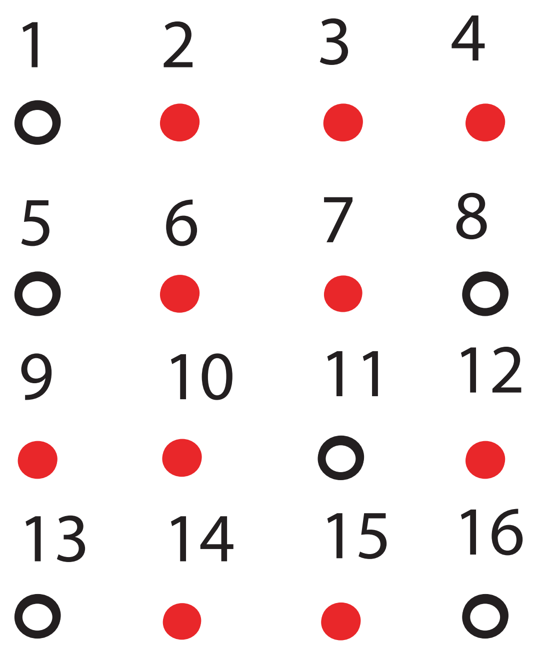
KL divergence fails in this scenario.

Our solution: dataset harmonization

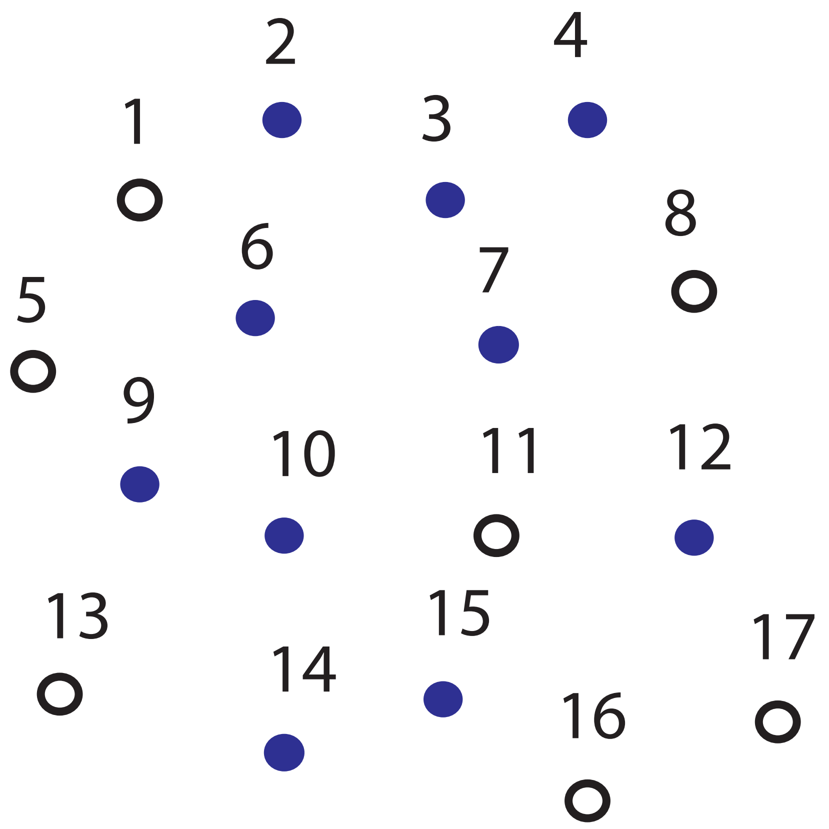
Estimating connectomes in a missing form

1. Time series-based approach
2. Transforming distribution of ROIs across atlases





$$\mathcal{M}_1 \neq \mathcal{M}_2$$

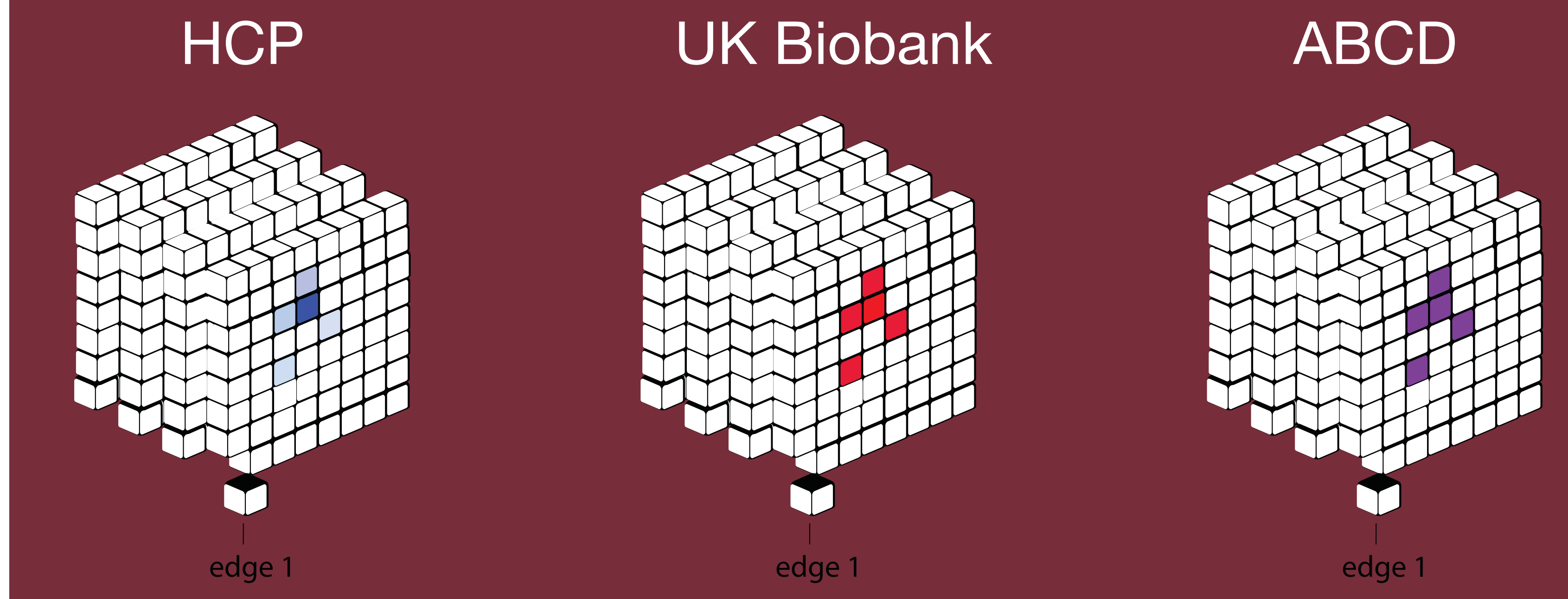


1. How about when the two distributions are defined in completely different spaces?
2. Optimal Transport captures both geometry and inconsistency of dimensions between p and q.

Our solution: dataset harmonization

Estimating connectomes in a missing form

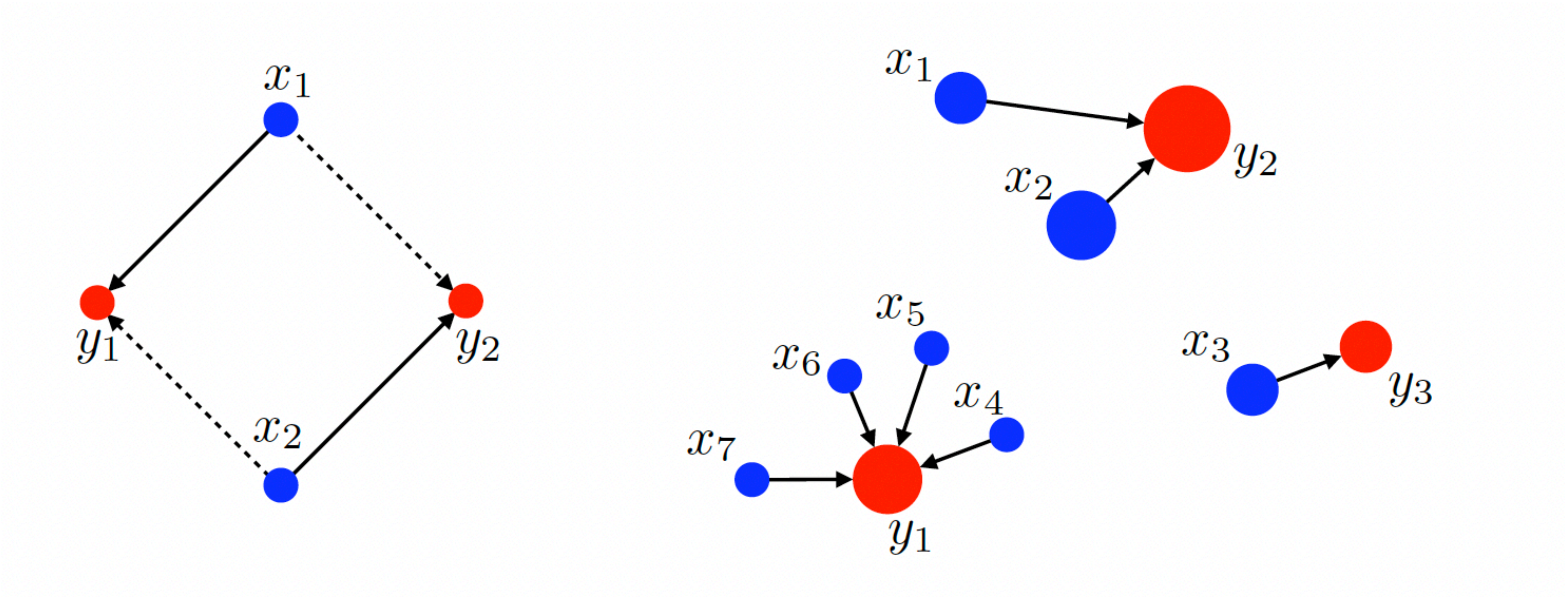
1. Time series-based approach
2. Transforming distribution of ROIs across atlases



Monge [1781]

Background

Optimal transport



A mapping between locations x and y

$$T : \{x_1, .., x_n\} \to \{y_1, .., y_n\}$$

must verify

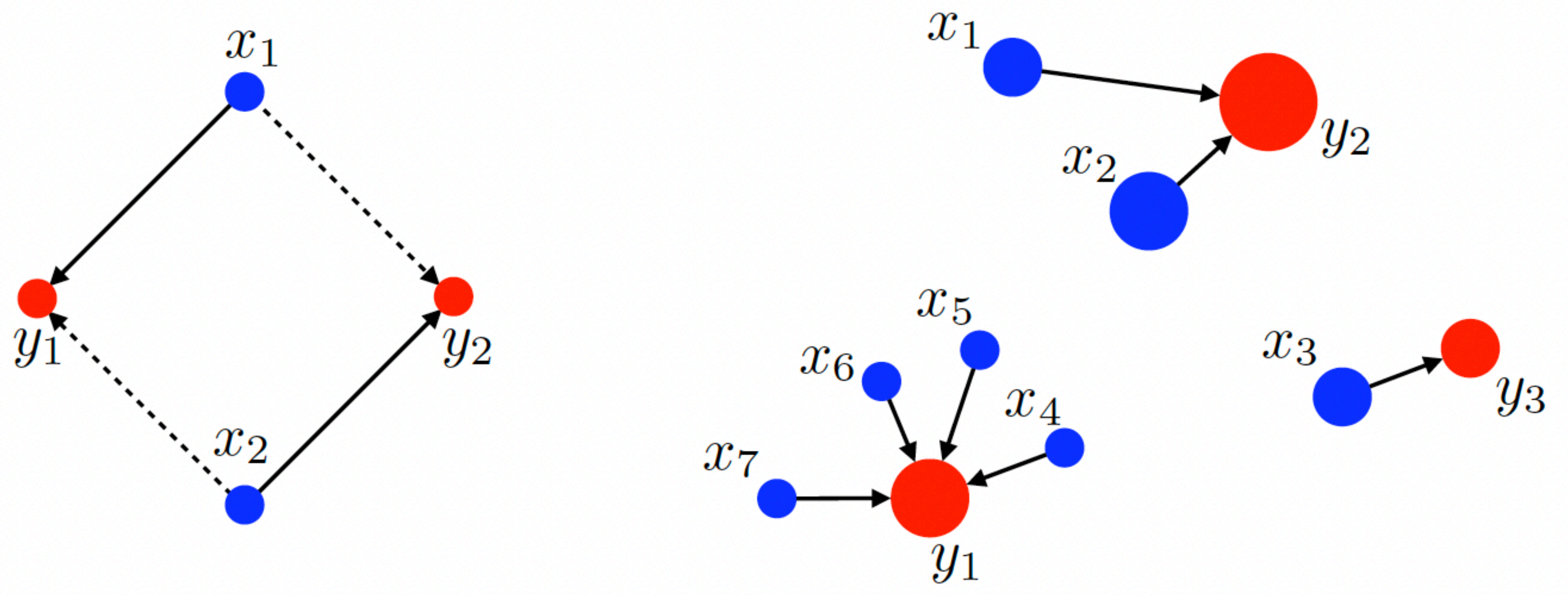
$$b_j = \sum_{i:T(x_i)=y_j} a_i$$

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

Monge [1781]

Background

Optimal transport



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

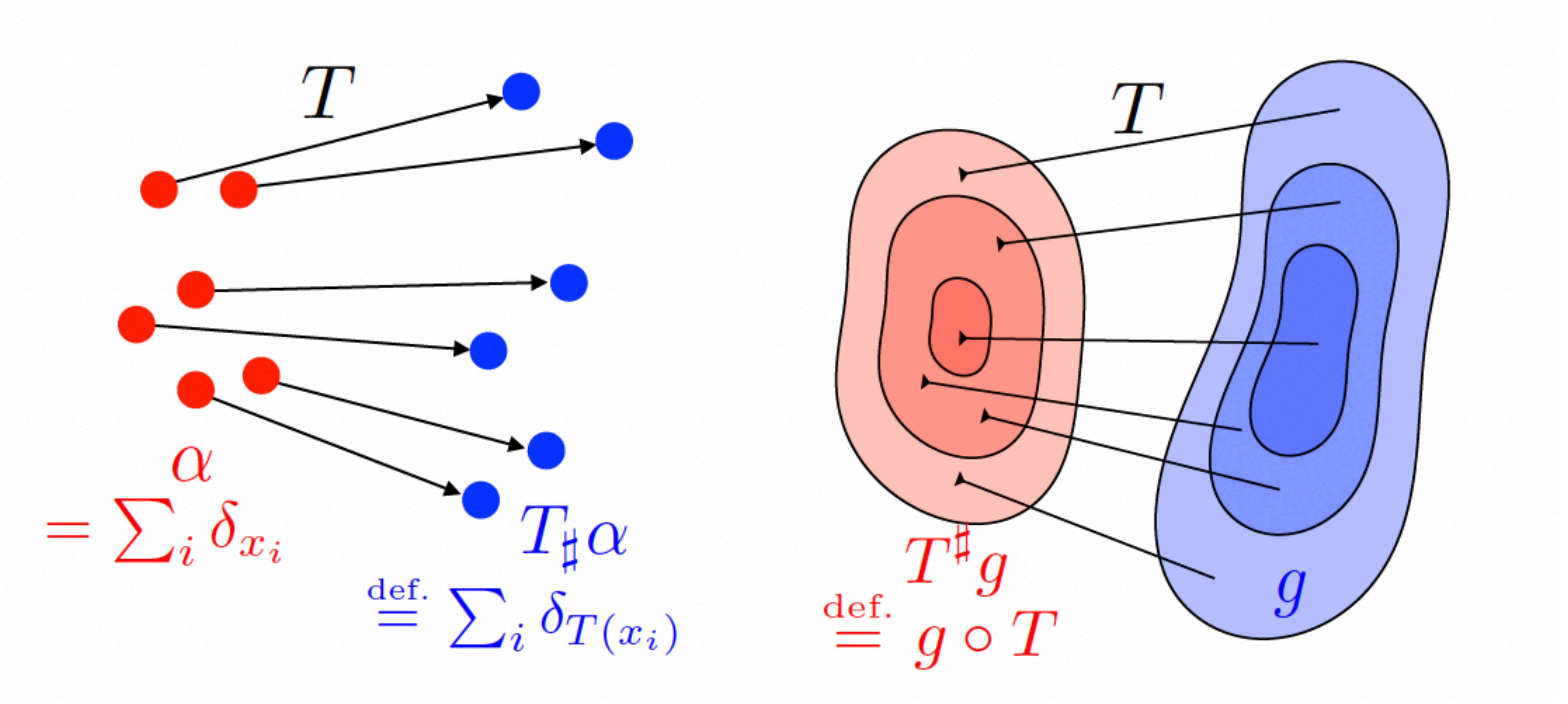
$$\min_T \left\{ \sum_i C(x_i, T(x_i)) : T_\sharp \alpha = \beta \right\},$$

Monge [1781]

Background

Optimal transport

Kantorovich [1942]



Push-forward of measures

Pull-back of functions

Kantorovich Relaxation [1942]

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

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

Admissible Couplings

Background

Optimal transport

Monge [1781]

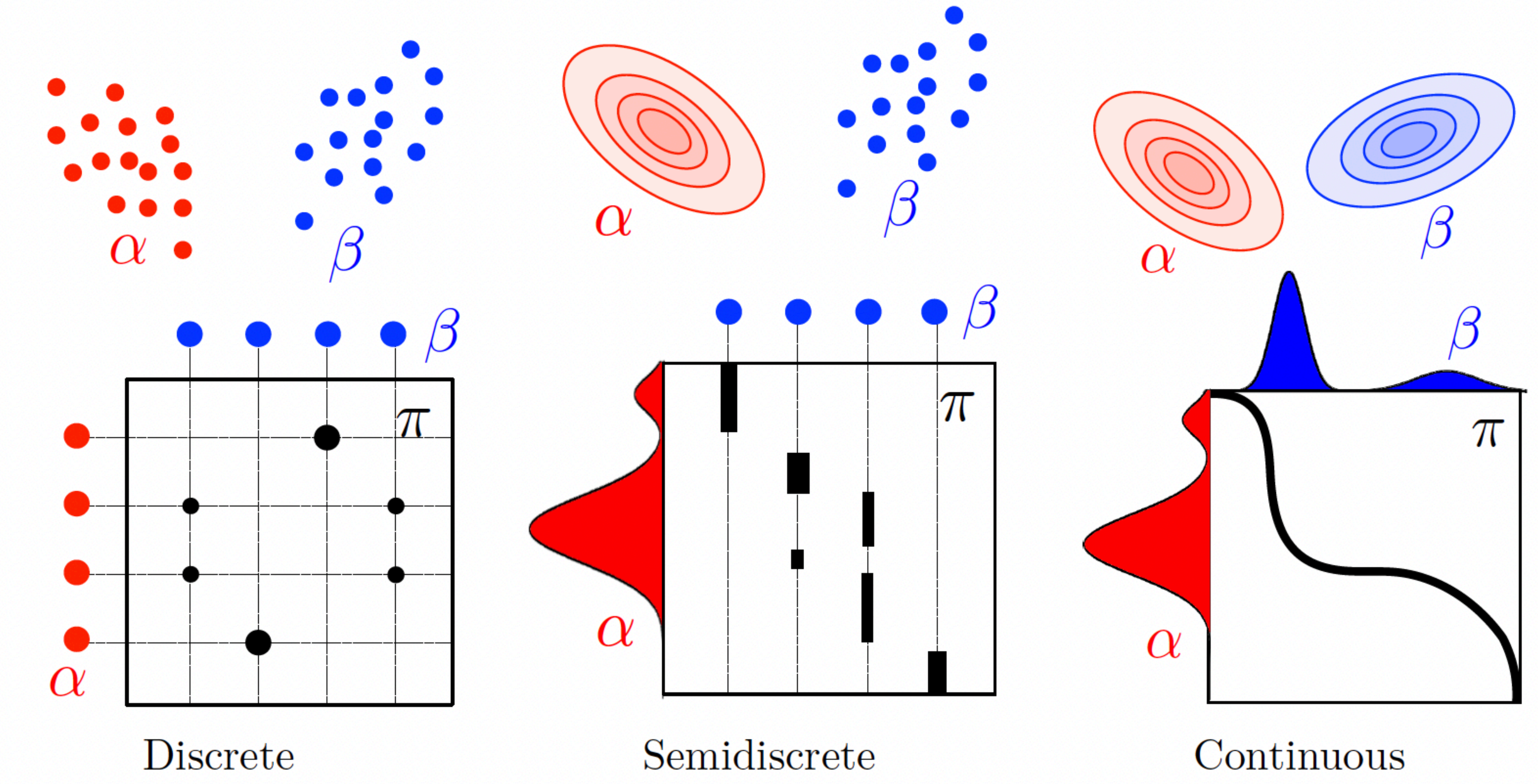
Kantorovich [1942]

$$A = \begin{array}{c} m \\ \\ n \end{array} \begin{pmatrix} \overset{1}{\begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \\ (1 & 1 & \cdots & 1) \end{pmatrix}} & \overset{2}{\begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \\ (1 & 1 & \cdots & 1) \end{pmatrix}} & \cdots & \overset{n}{\begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \\ (1 & 1 & \cdots & 1) \end{pmatrix}} \\ \vdots & & & \\ (1 \; 1 \; \cdots \; 1) & \cdots & & \cdots (1 \; 1 \; \cdots \; 1) \end{pmatrix}$$

$$L_c(\mu_t, \nu_t) = \min_T C^T T \quad \text{s.t, } A\underline{T} = \begin{bmatrix} \mu_t \\ \nu_t \end{bmatrix}.$$

Kantorovich's optimal transport problem now reads

$$\mathrm{L}_{\mathbf{C}}(\mathbf{a}, \mathbf{b}) \overset{\text{def.}}{=} \min_{\mathbf{P} \in \mathbf{U}(\mathbf{a},\mathbf{b})} \langle \mathbf{C}, \mathbf{P} \rangle \overset{\text{def.}}{=} \sum_{i,j} \mathbf{C}_{i,j} \mathbf{P}_{i,j}.$$



Kantorovich Relaxation is symmetric

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

Background

Optimal transport

Monge [1781]

Hitchcock [1941]

Kantorovich [1942]

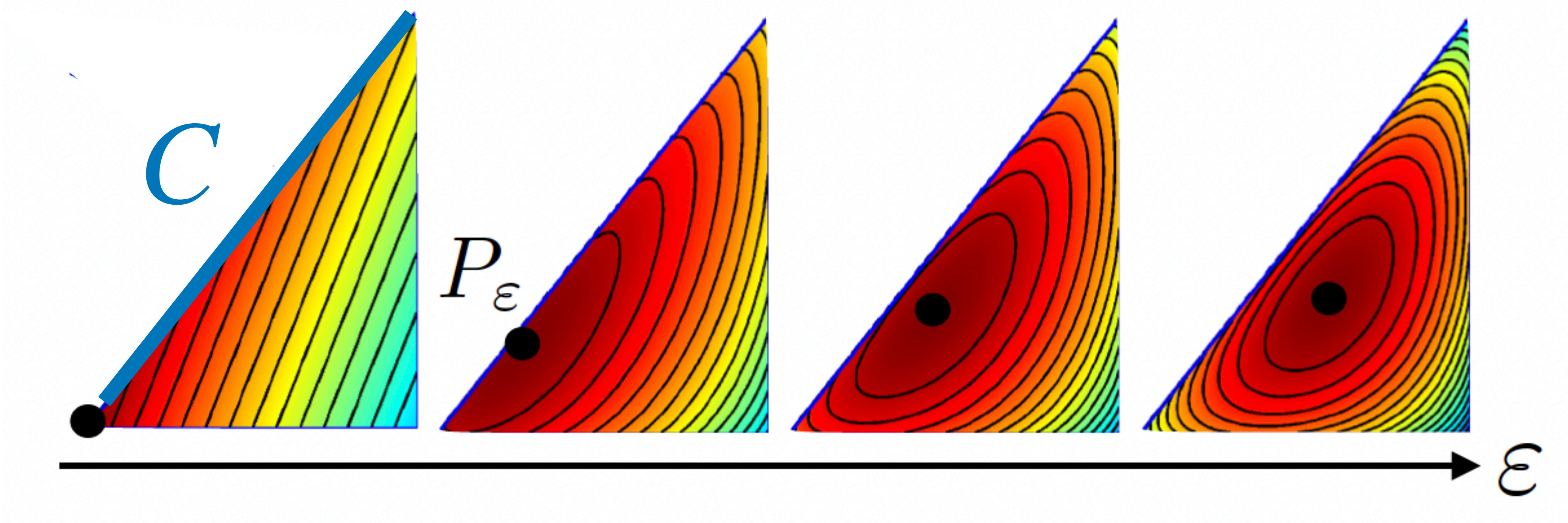
Entropy regularization: An approximation solution

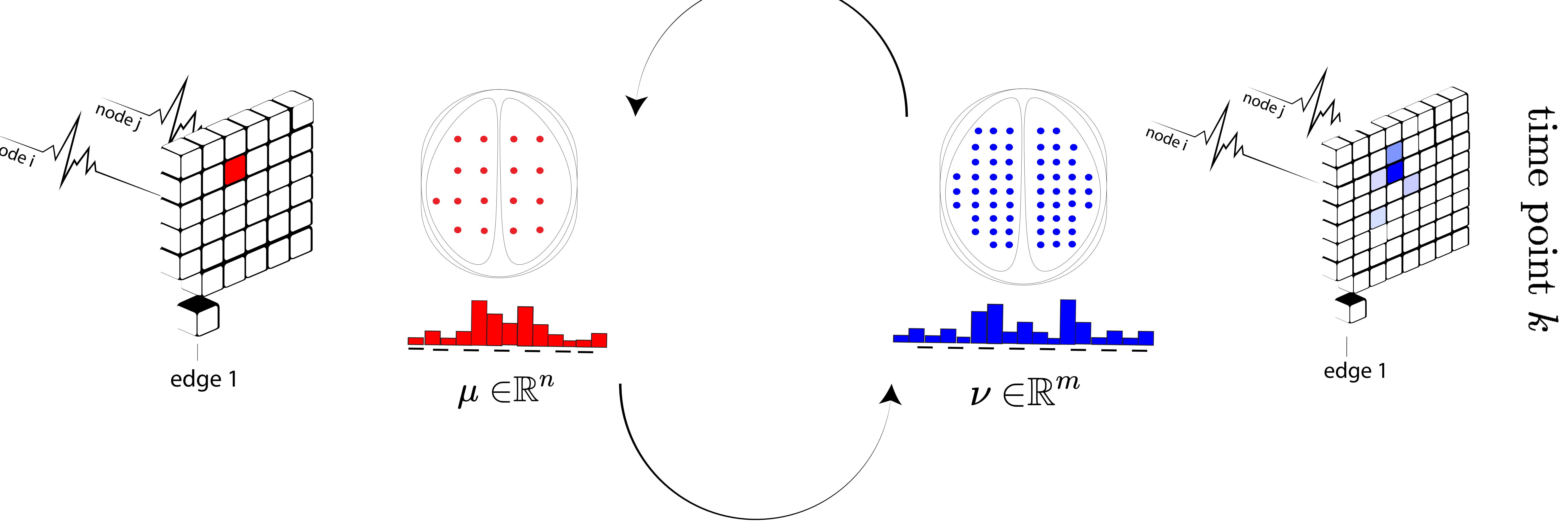
$$\mathrm{L}_{\mathbf{C}}^{\varepsilon}(\mathbf{a},\mathbf{b}) \stackrel{\text{def.}}{=} \min_{\mathbf{P}\in\mathbf{U}(\mathbf{a},\mathbf{b})} \langle \mathbf{P},\ \mathbf{C}\rangle - \varepsilon \mathbf{H}(\mathbf{P}).$$

Iterative solutions: Sinkhorn algorithm

$$C_a^1 = \{P, P1 = a\} \qquad C_b^2 = \{P, P1 = b\}$$

$$P^{(l+1)} = \mathrm{Proj}_{C_a^1}^{\mathrm{KL}} P(l) \qquad P^{(l+2)} = \mathrm{Proj}_{C_b^2}^{\mathrm{KL}} P(l+1)$$





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

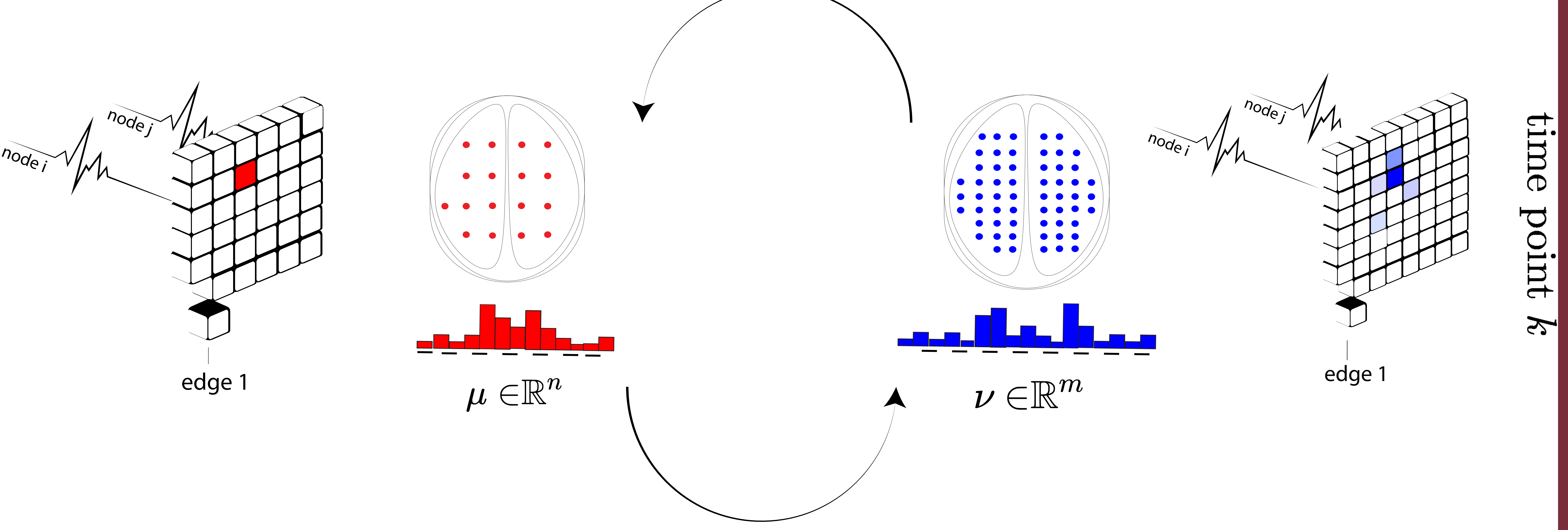
$$A = \begin{matrix} m \\ \\ n \end{matrix} \overset{\begin{matrix} 1 & \qquad\qquad\qquad 2 & \qquad\qquad\qquad n \end{matrix}}{\begin{pmatrix} \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{pmatrix} & \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{pmatrix} & \dots & \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{pmatrix} \\ (1 \;\; 1 \;\; \dots \;\; 1) & (1 \;\; 1 \;\; \dots \;\; 1) & \dots & (1 \;\; 1 \;\; \dots \;\; 1) \\ \vdots & & & \\ (1 \;\; 1 \;\; \dots \;\; 1) & \dots & & \dots (1 \;\; 1 \;\; \dots \;\; 1) \end{pmatrix}}$$

$$C = \begin{pmatrix} C_{1,1} & .. & C_{1,m} \\ .. & & .. \\ C_{n,1} & .. & C_{n,m} \end{pmatrix} \in \mathbb{R}^{n \times m} \qquad C_{i,j} = \text{Euclidean distance}$$

Cross Atlas Remapping via Optimal Transport (CAROT)

A data-driven method to measure the distance and find a policy to transform connectomes

1. Translating each time frame to a vector
2. Cost matrix
3. Loss function



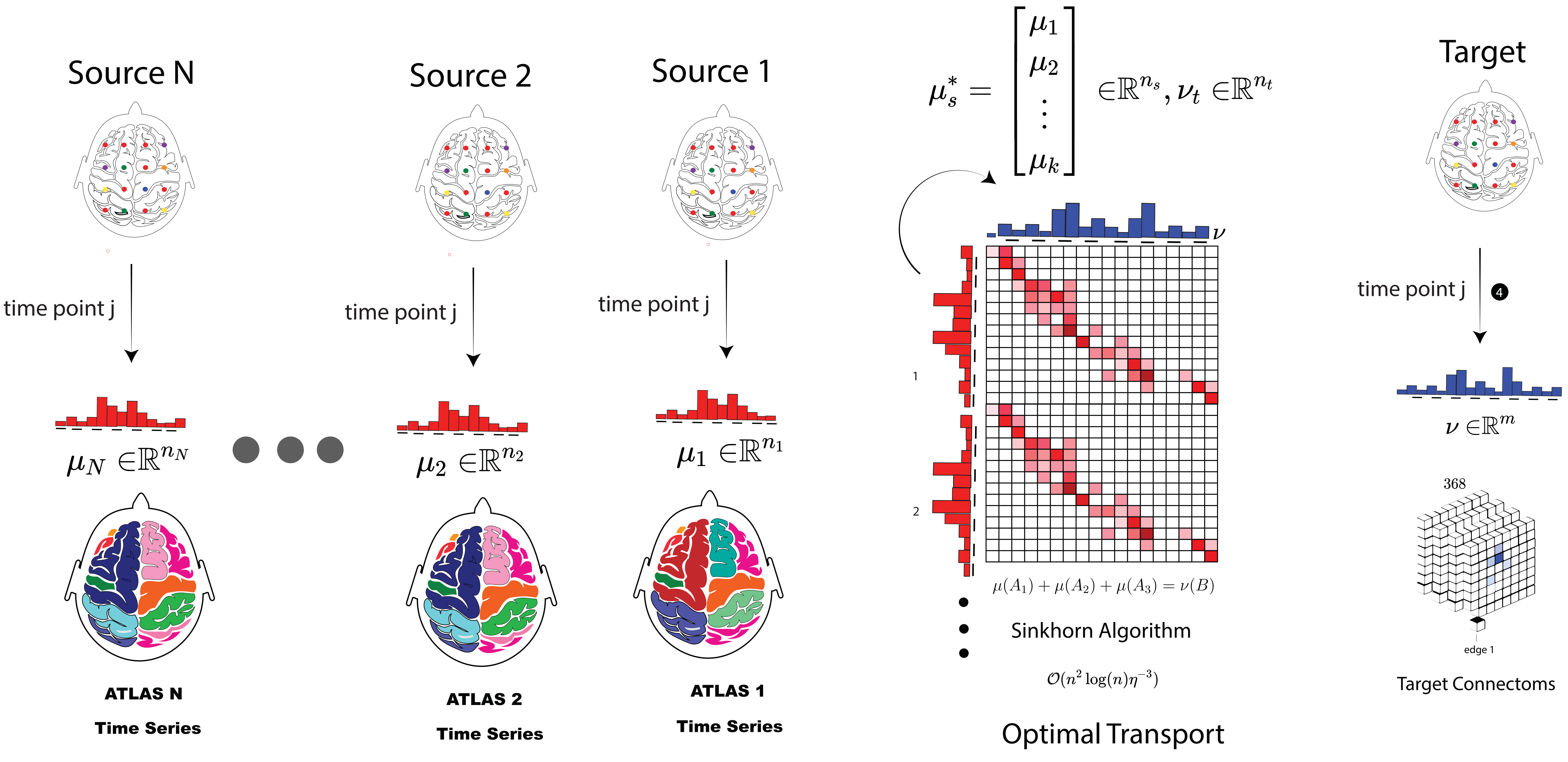
test data point $\nu = \mu T$

What if multiple parcellations for each individual are available?

Cross Atlas Remapping via Optimal Transport (CAROT)

Test data point available in the source atlas

1. Applying the trained policies T
2. Some of large scale projects release data in multiple atlases
3. A need for an advance version



Cross Atlas Remapping via Optimal Transport (CAROT)

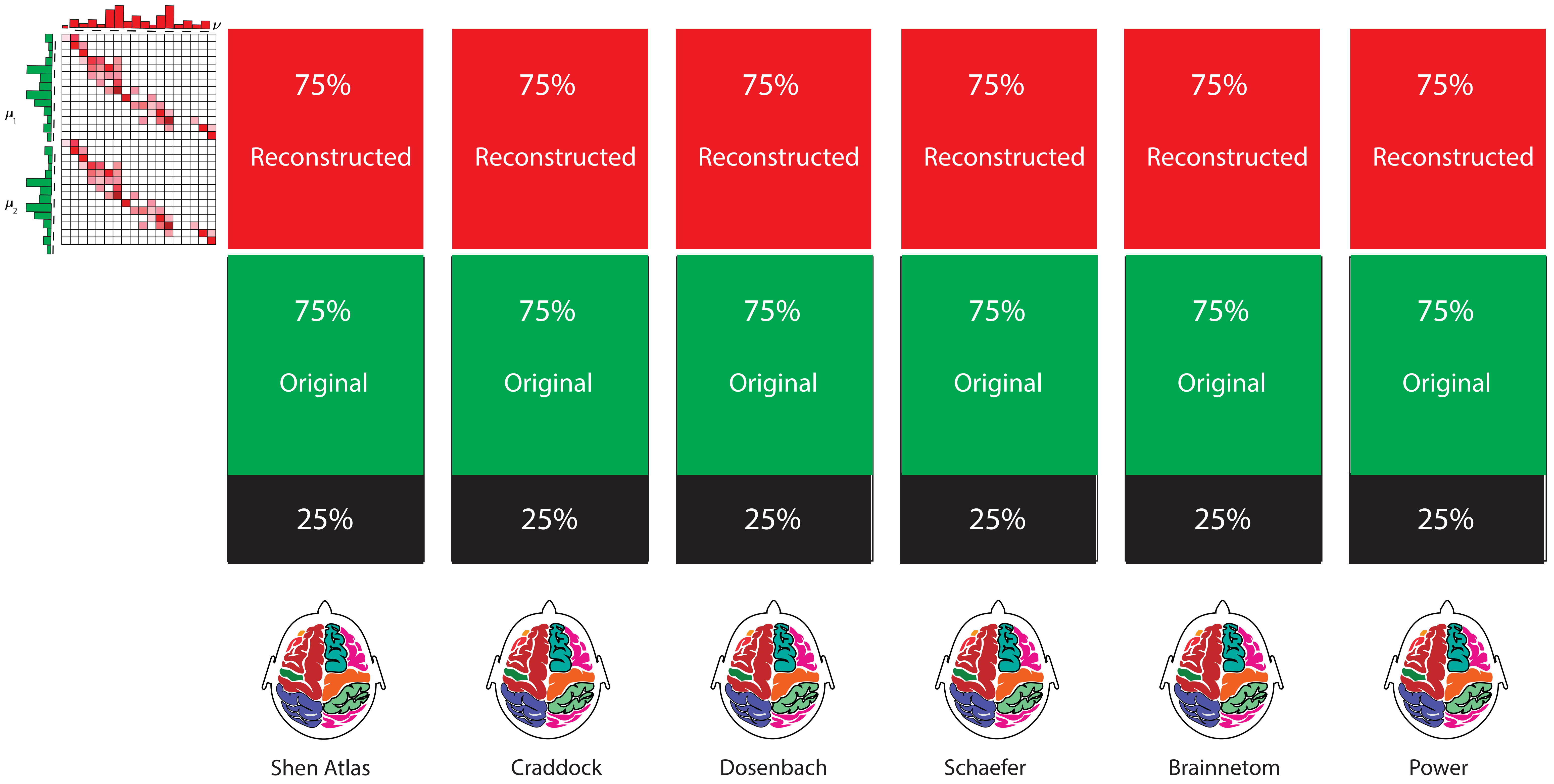
An advanced version when multiple parcellations are available

1. Incorporating multiple time series
2. Bigger cost matrix
3. Bigger policy

$$L_c(\mu^*_t, \nu^*_t) = \min_T C^T T - \epsilon H(T) \text{ s.t, } A\underline{T} = \begin{bmatrix}\mu^*_t \\ \nu^*_t\end{bmatrix}.$$

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

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



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

Experiments:

Human connectome projects

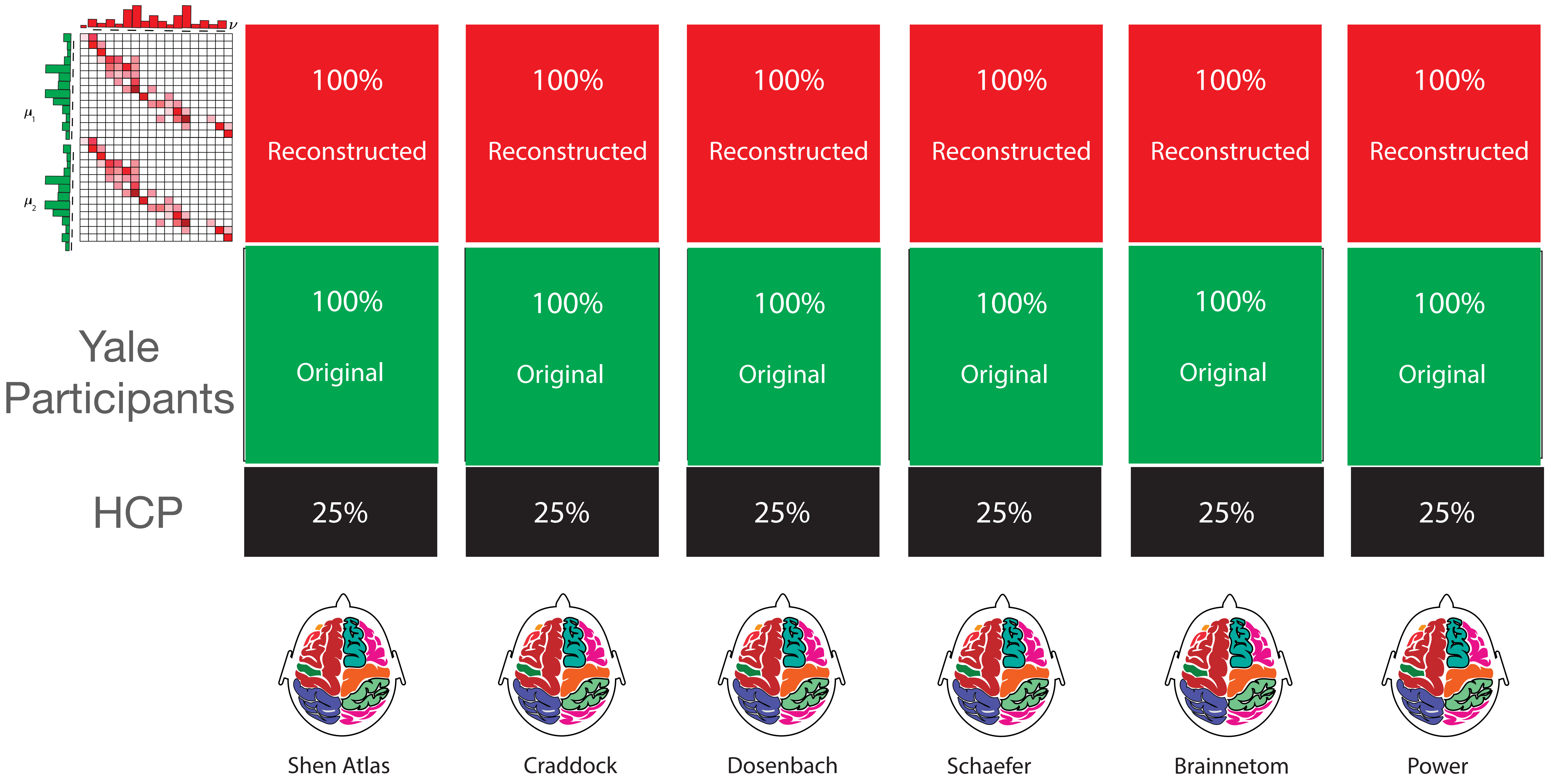
1. Train-test split
2. 25% for policy training
3. 75% for testing
4. 10 fold CV

Experiments:

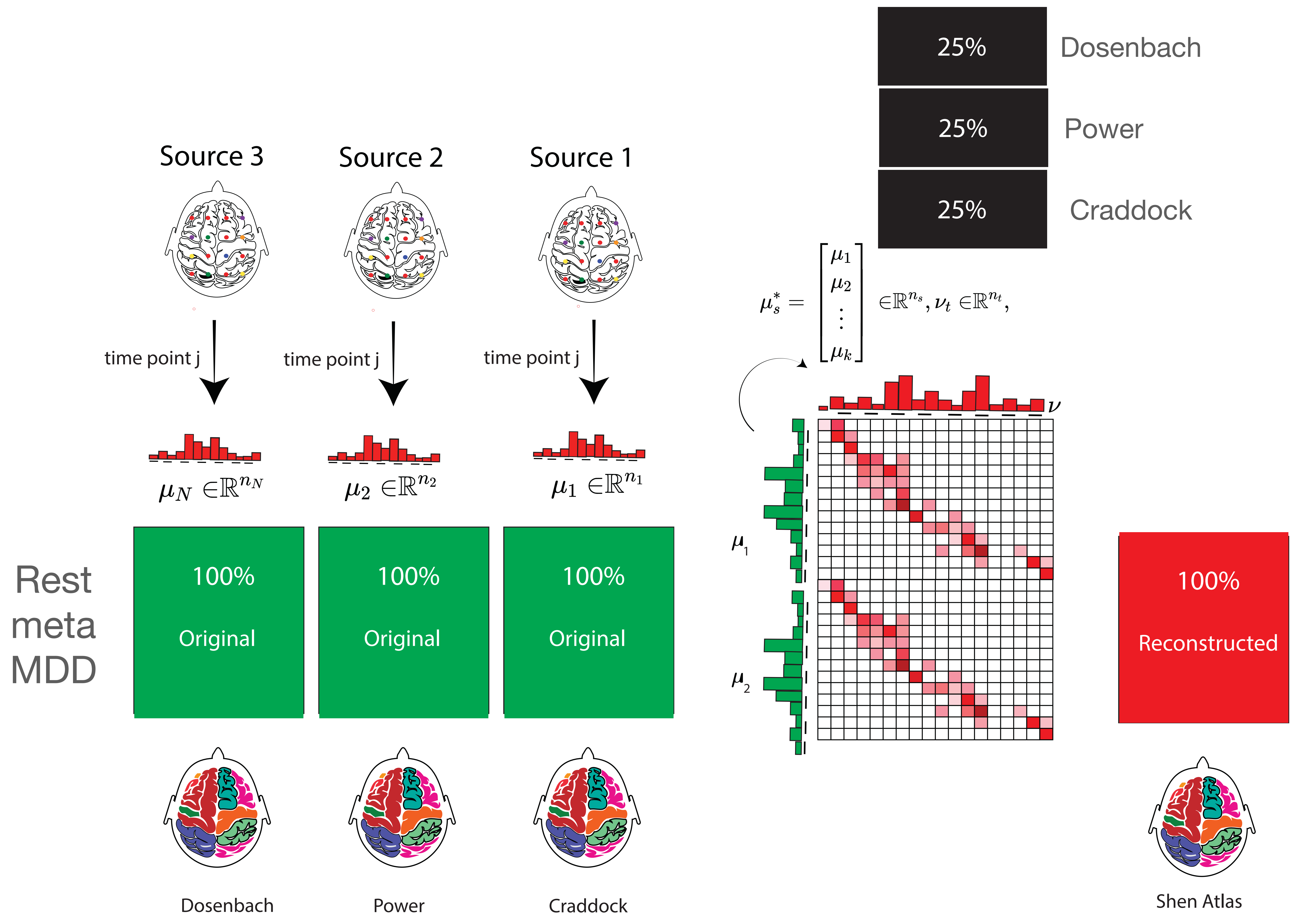
Yale participant

1. Testing generalization of policies
2. 50 male and 50 female Yale participant dataset

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



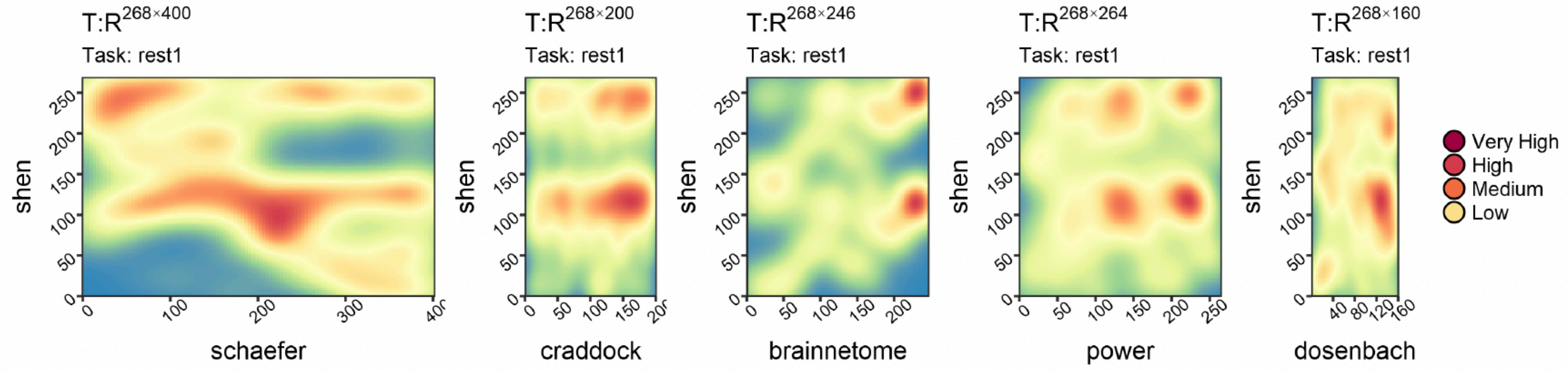
This dataset included 50 females (age=33) and 50 males (age=34.9) with eight functional scans (48 minutes total).



Experiments:

Rest-meta MDD dataset

1. Testing generalization of models
2. A dataset that is not released on Shen 268
3. A model that performs the best on Shen 268

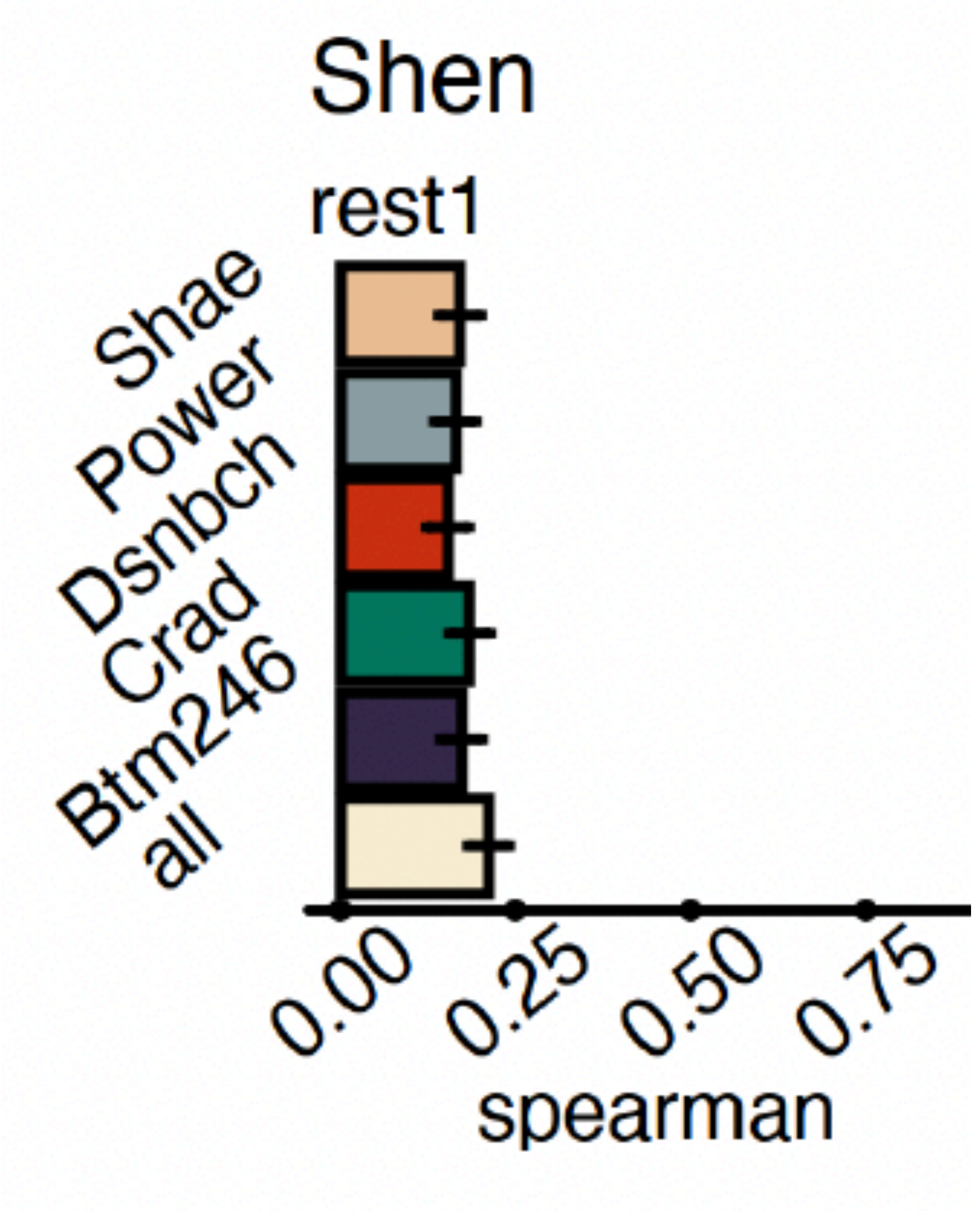


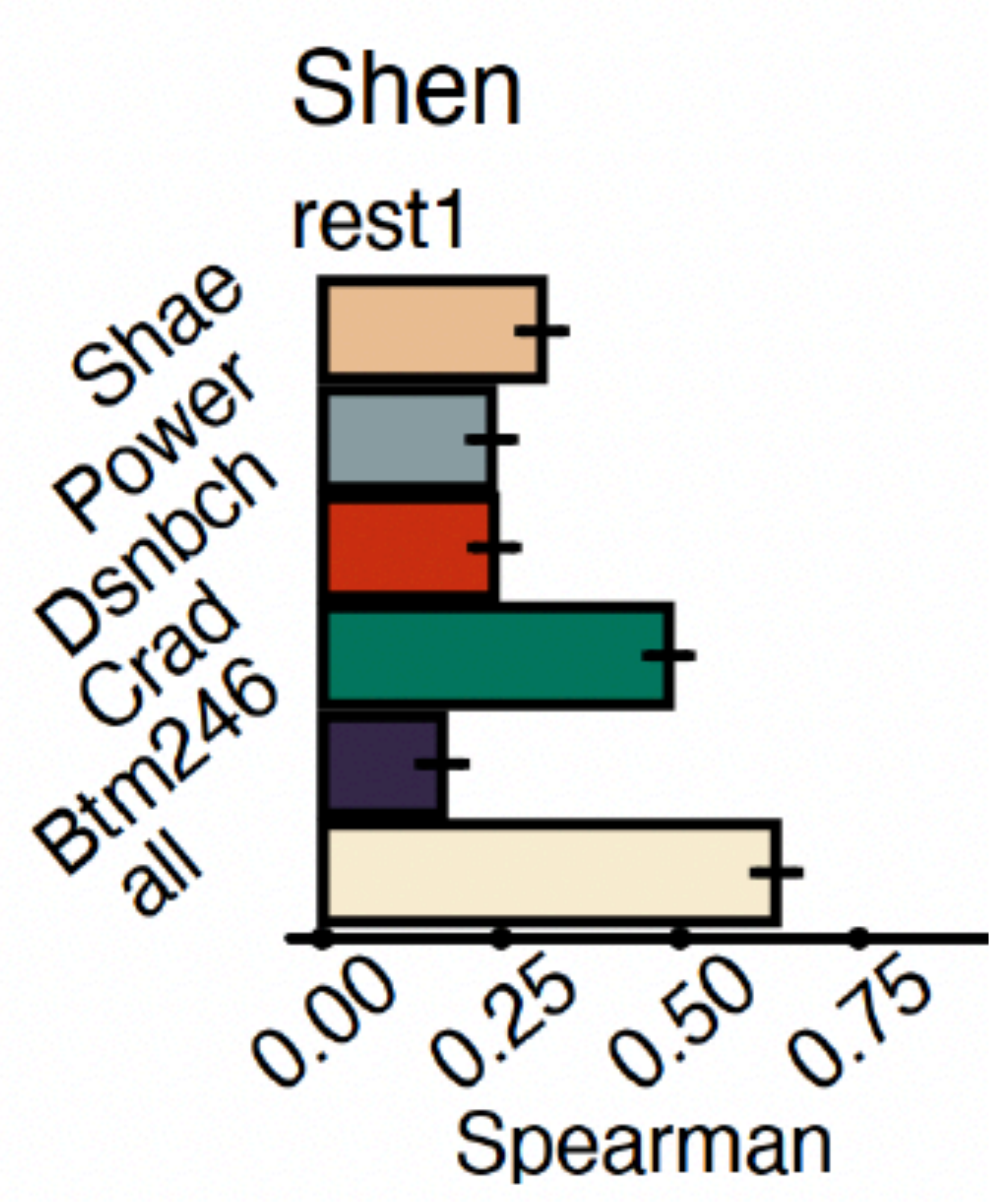
- You can see that some spots are more intense than others indicating higher transformation between regions.
- This emphasizes some of the structural differences between atlases:
  - The horizontal line between Schaefer and Shen is belonging to areas that are missing in Schaefer

Policies

How does a policy look like

1. Topological differences are clear
2. Schaefer doesn't include some areas





$$C^* = \begin{pmatrix} C_{1,1} & .. & C_{1,m} \\ .. & .. & .. \\ C_{n_s,1} & .. & C_{n,m} \end{pmatrix} \in \mathbb{R}^{n_s \times m}$$

$$C = 1 - \begin{pmatrix} \rho(U_{1,.}, N_{1,.}) & .. & \rho(U_{1,.}, N_{n,.}) \\ .. & .. & .. \\ \rho(U_{m,.}, N_{1,.}) & .. & \rho(U_{m,.}, N_{n,.}) \end{pmatrix} \in \mathbb{R}^{m \times n}$$

$C_{i,j}$ = Euclidean distance

ρ_{U_i,N_j} = Spearman correlation

What should we choose as a cost matrix?

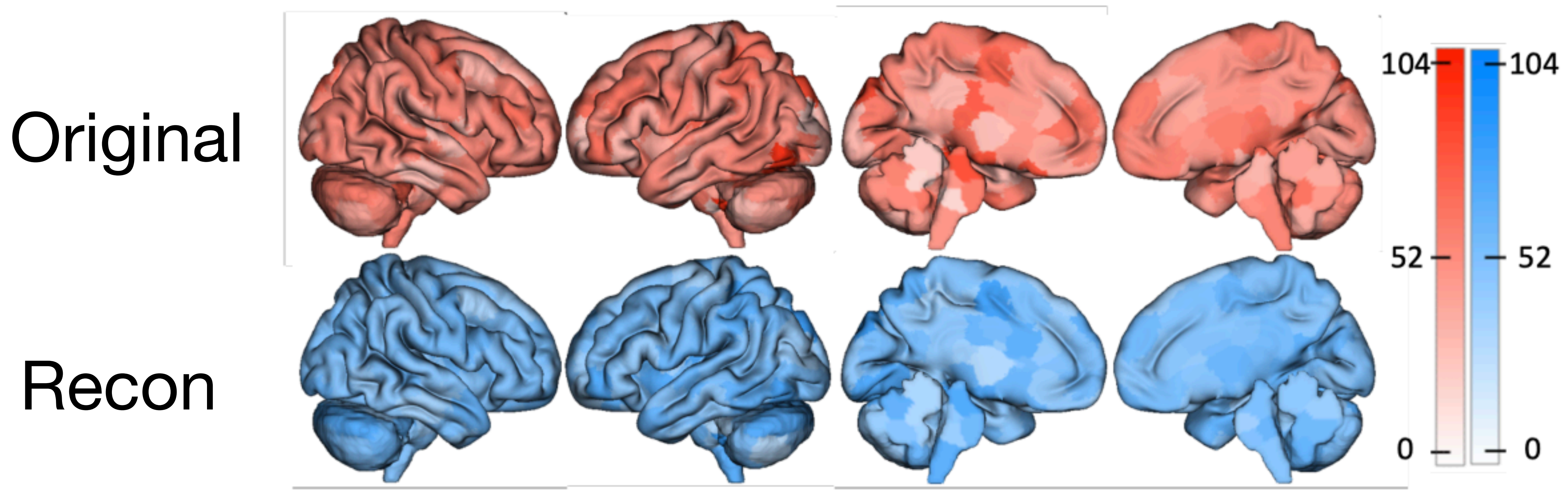
Functional distance vs Euclidean distance

1. Euclidean distance
2. Functional distance

Experimental results

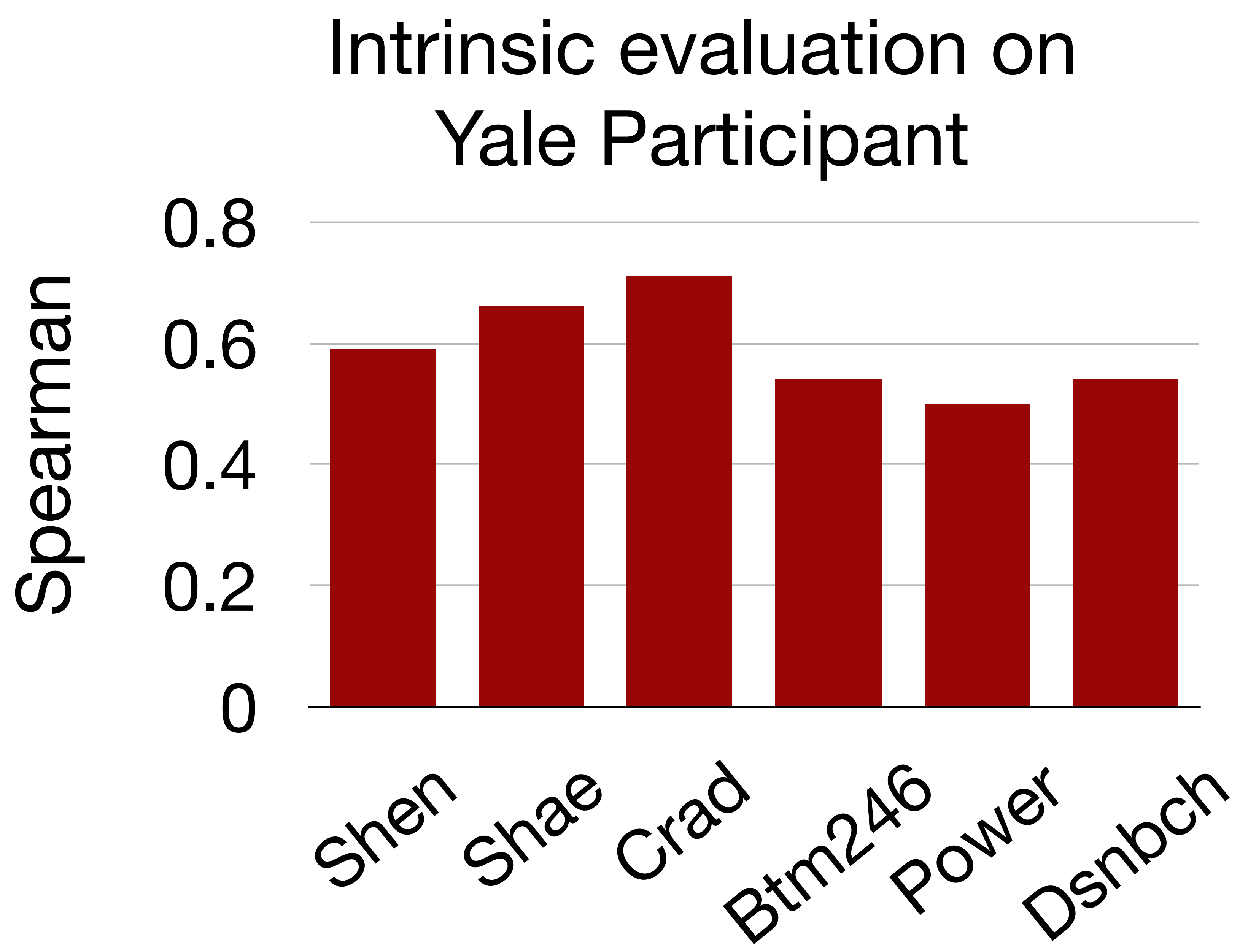
HCP dataset, resting scan connectomes

1. Intrinsic evaluation; correlation with original counterparts
2. Downstream analysis, results on predicting IQ
3. Fingerprinting, two resting sessions



Reconstructed connectomes give similar aging results as the original connectomes. (Top) the nodes with the largest number of edges are significantly associated with age for original connectomes from the HCP using Shen. (bottom) reconstructed Shen connectomes (r=0.61)

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

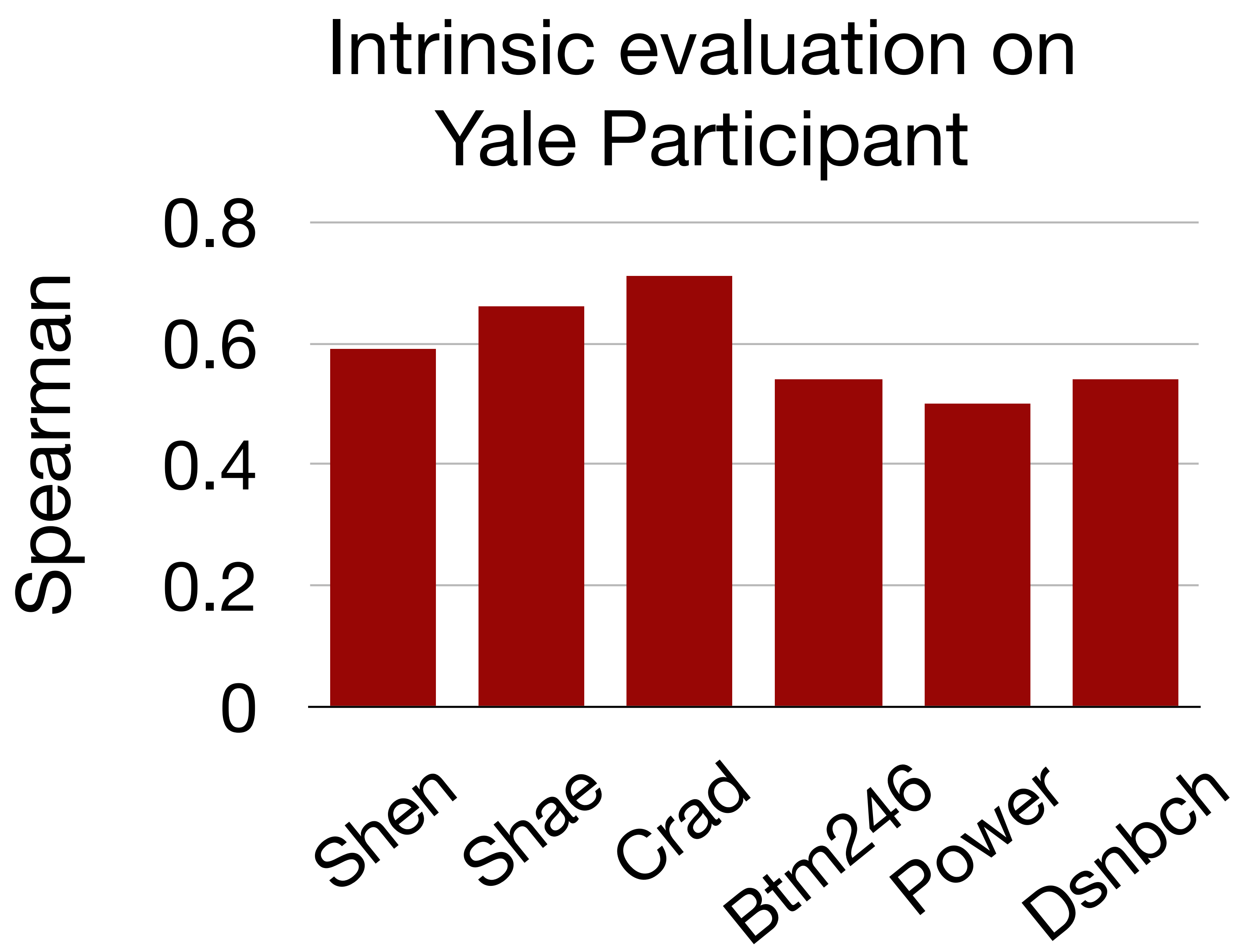


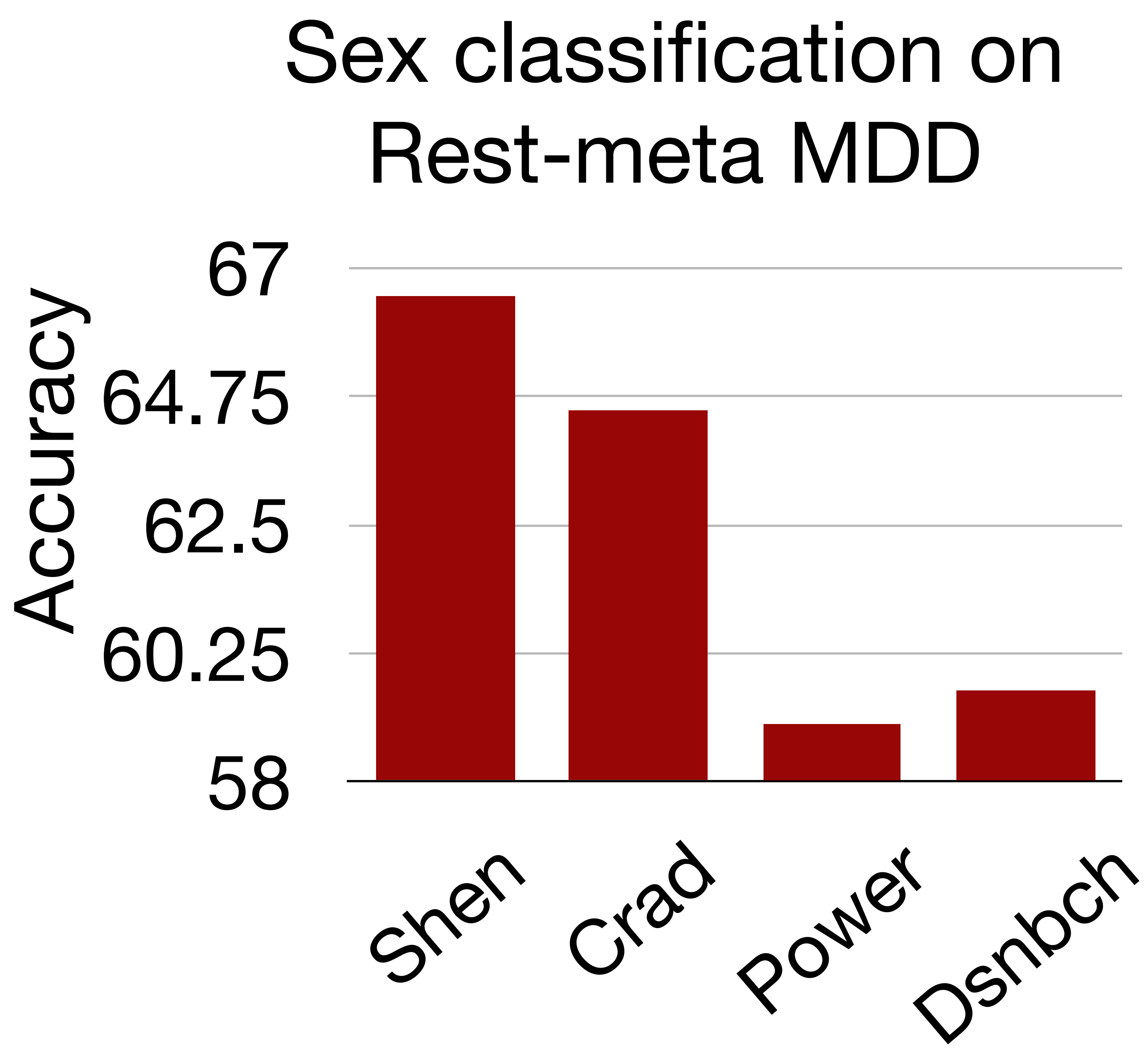
- We investigated if CAROT mappings trained in one dataset generalize to other datasets.
- We applied the mappings trained on HCP and reconstructed connectomes using the Yale dataset using these mappings.
- Spearman's rank correlation between the upper triangles of the connectomes was used to assess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

Real-world example

Testing a model on Rest-meta MDD

1. Generalization on Yale participant
2. Sex classification on Rest-meta MDD





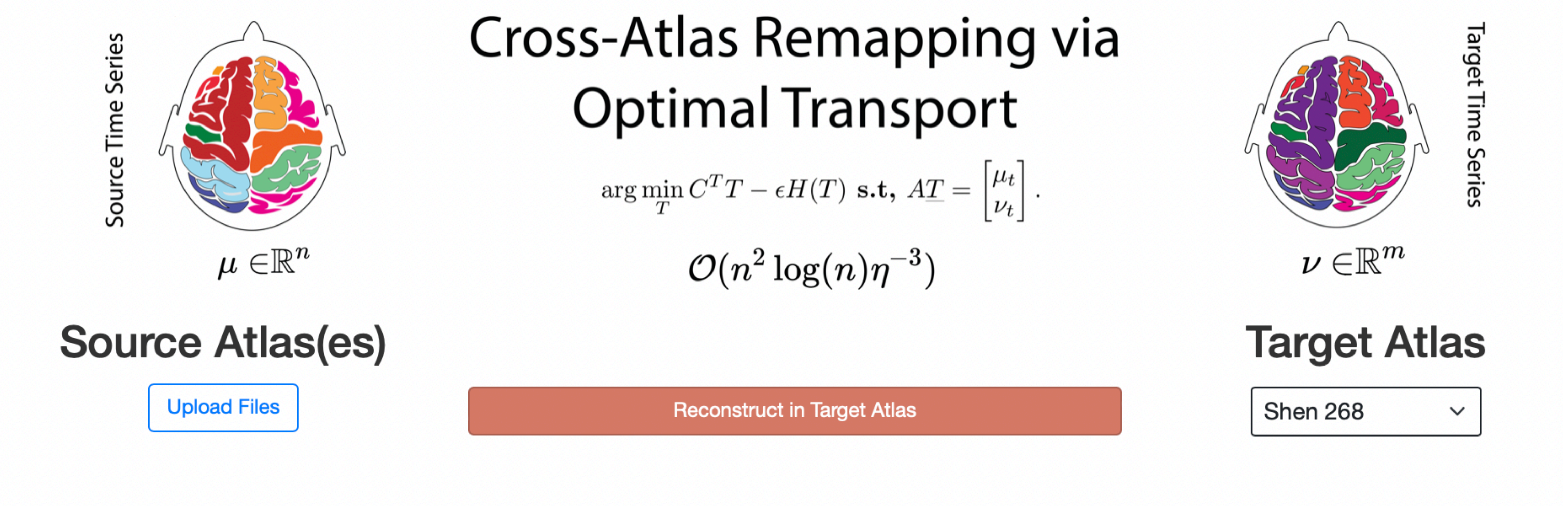
- In this evaluation, we generalize a sex classification model on Yale dataset:
  - The REST-Meta-MDD dataset (Yan et al., 2016) only provides preprocessed timeseries data from the Dosenbach, Power, and Craddock atlases.
- Overall, the sex classification model demonstrated significant classification accuracy (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.

Real-world example

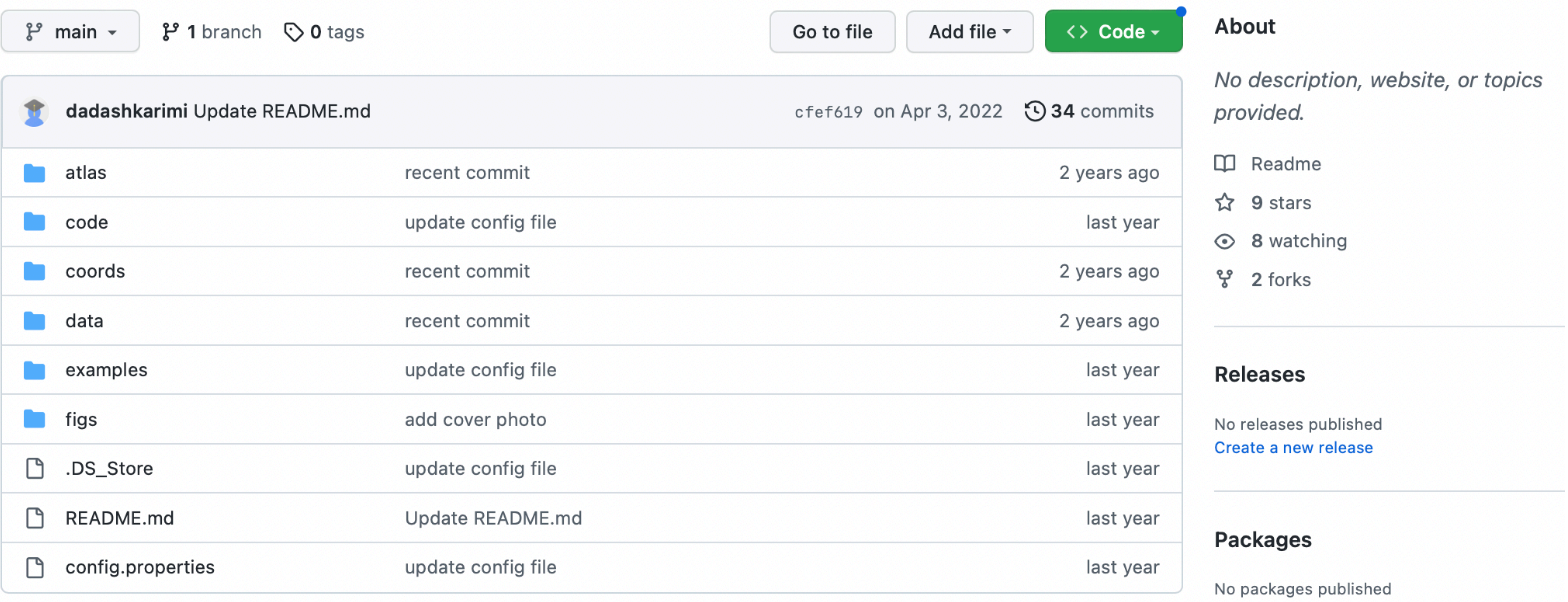
Testing a model on Rest-meta MDD

1. Generalization on Yale participant
2. Sex classification on Rest-meta MDD

carotproject.com



https://github.com/dadashkarimi/carot



Software

GitHub and live demo

1. Live demo for some atlases
2. GitHub repository for all types of data

- 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

CAROT encourages open science in connectomics

1. CAROT helps overcome multiple atlas problem
2. CAROT brings good quality
3. Policies are generalizable over datasets

Publications

Journal and conference papers

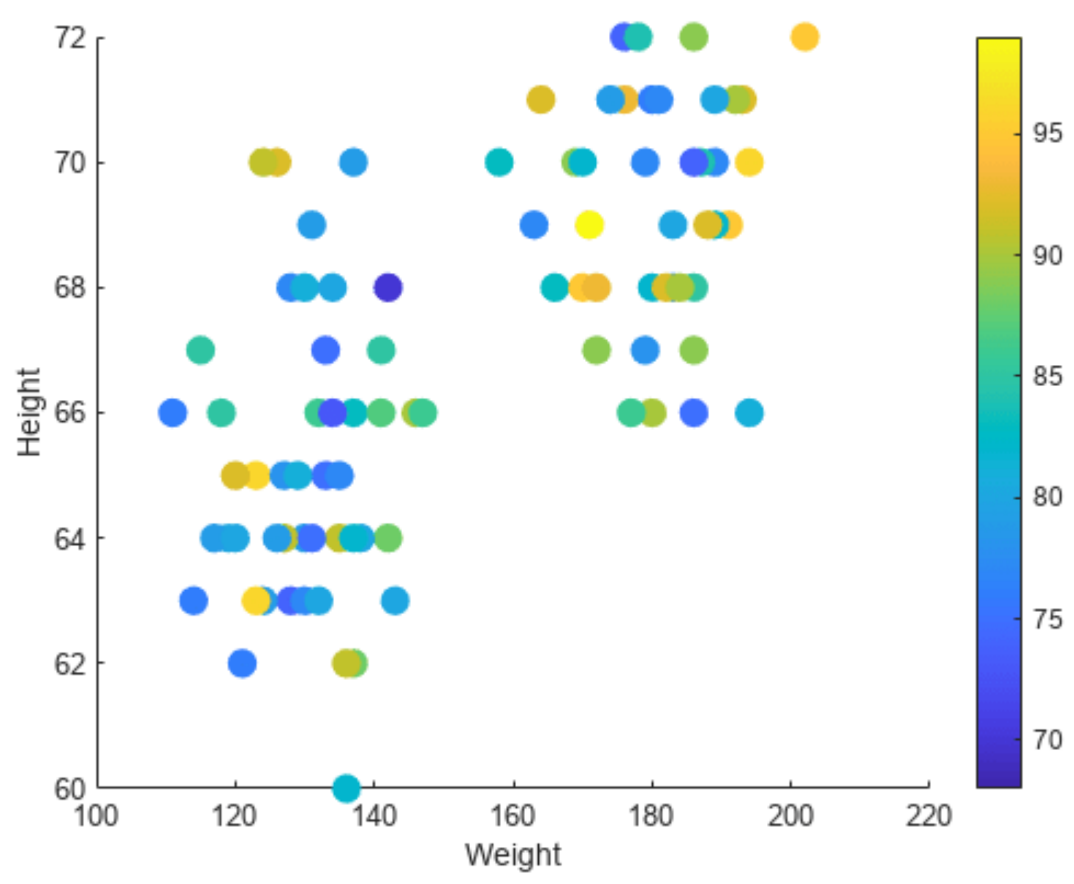
1. IEEE Information Sciences and Systems 2023
2. MICCAI 2022
3. MICCAI 2021
4. MICCAI 2019

1. Javid Dadashkarimi, Matthew Rosenblatt, Amin Karbasi, and Dustin Scheinost, (2023) Stacking multiple optimal transport policies to map functional connectomes, **CISS**
2. 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**
3. Javid Dadashkarimi, Amin Karbasi, and Dustin Scheinost, (2022) Combining multiple atlases to estimate data-driven mappings between functional connectomes using optimal transport, **MICCAI**
4. 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**
5. Javid Dadashkarimi, Amin Karbasi, and Dustin Scheinost, (2021) Data-driven mapping between functional connectomes using optimal transport, **MICCAI**
6. 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**

Thank you so much: MINDS lab and IID lab

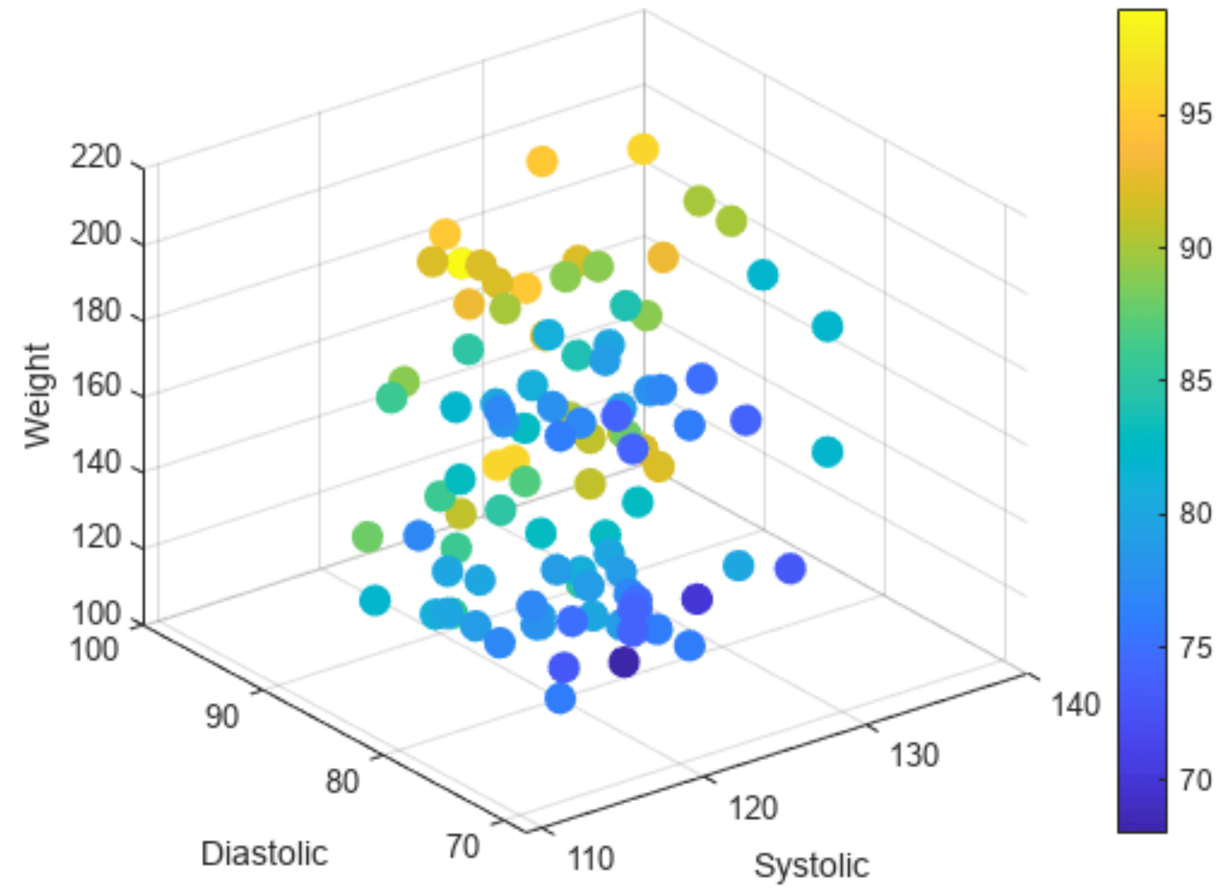
- Dustin Scheinost
- Amin Karbasi
- Qinghao Liang
- Matthew Rosenblatt
- Stephanie Noble
- Raimundo Rodriguez
- Brendan Adkinson
- Huili Sun
- Jean Ye
- Maya Foster
- Chris Camp
- Michael Farruggia
- Link Tejavibulya
- Wei Dai
- Raina Vin
- AJ Simon
- Camille Duan
- Rongtao Jiang
- Angeliki Pollatou

A



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

B



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

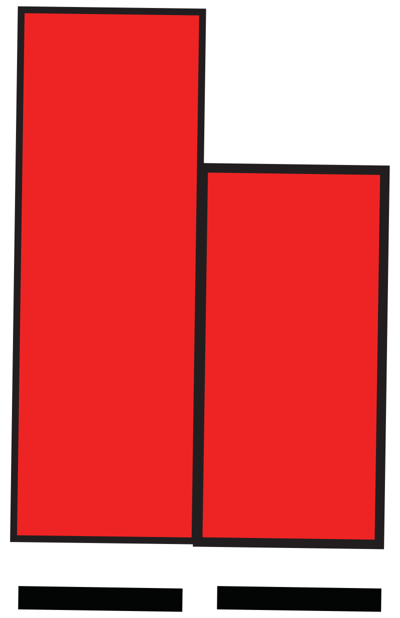
- **Example 1**: In predictive modeling, the feature space between train and test should match:
  - It's impractical to train a model on A and test on B: $Y = X^T\beta + \epsilon$.
  - Other techniques include meta-learning, transfer learning, and federated learning.

Theoretical Concerns

Examples from predictive modeling and explanatory analysis

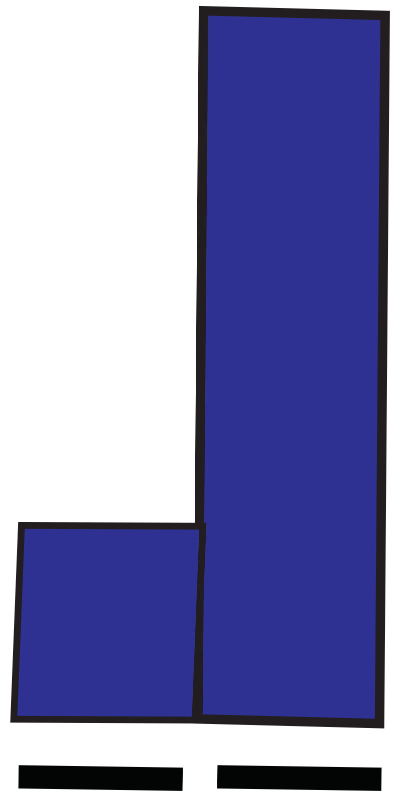
1. Predictive modeling
2. Explanatory analysis

Observation A A | A A | A A A A A | A A | A A A



$p_1 \quad p_2$

$p_1\ p_1\ p_2\ p_1\ p_1\ p_2\ p_1\ p_1\ p_1 p_1\ p_1\ p_2\ p_1\ p_1\ p_2\ p_1\ p_1\ p_1$



$q_1 \quad q_2$

$q_1\ q_1\ q_2\ q_1\ q_1\ q_2\ q_1\ q_1\ q_1 q_1\ q_1\ q_2\ q_1\ q_1\ q_2\ q_1\ q_1\ q_1$

$$\frac{P(\text{observation} \mid \text{True Coin})}{P(\text{observation} \mid \text{Coin 2})} = \frac{p_1^{N_H} \cdot p_2^{N_T}}{q_1^{N_H} \quad q_2^{N_T}}$$

An overview on information theory

Classical methods to compare distributions

1. KL divergence
2. JS divergence
3. Having the same number of supports is equired

$$\text{normalized relative likelihood} = \left(\frac{p_1^{N_H} p_2^{N_H}}{q_1^{N_H} q_2^{N_T}}\right)^{\frac{1}{N}}$$

$$= \frac{1}{N} \log\left(\frac{p_1^{N_H} p_2^{N_H}}{q_1^{N_H} q_2^{N_T}}\right)$$

$$= \frac{1}{N} \log p_1^{N_H} + \frac{1}{N} \log p_2^{N_T} - \frac{1}{N} \log q_1^{N_H} - \frac{1}{N} \log q_2^{N_T}$$

$$= \frac{1}{N} \log p_1^{N_H} - \frac{1}{N} \log q_1^{N_H} + \frac{1}{N} \log p_2^{N_T} - \frac{1}{N} \log q_2^{N_T}$$

$$= \frac{N_H}{N} \log p_1 - \frac{N_H}{N} \log q_1 + \frac{N_T}{N} \log p_2 - \frac{N_T}{N} \log q_2$$

$$= p_1 \log p_1 - q_1 \log q_1 + p_2 \log p_2 - q_2 \log q_2$$

$$= D_{KL}(p \,||\, q) = \sum_{x \in \mathcal{X}} p(x) \log\left(\frac{p(x)}{q(x)}\right)$$

Kullback–Leibler divergence

Measures exactly the same thing

1. Log properties, product to addition, division to subtraction
2. How likely $q(x)$ would generate samples from $p(x)$

Jensen-Shannon divergence $= \frac{1}{2}D_{KL}\left(p\,||\,\frac{p+q}{2}\right) + \frac{1}{2}D_{KL}\left(p\,||\,\frac{p+q}{2}\right)$

squared Hellinger distance $= 2\sum_x \left(\sqrt{p(x)} - \sqrt{q(x)}\right)^2$

α – divergence $= \frac{4}{1-\alpha}\left(1 - \sum p(x)^{\frac{1-\alpha}{2}} q(x)^{\frac{1+\alpha}{2}}\right)$

chi-squared divergence $= \sum_x \frac{\left(p(x) - q(x)\right)^2}{p(x)}$

What if p and q have different number of classes?

Jensen–Shannon divergence

A symmetric version of KL-divergence

1. Symmetric
2. Other divergence methods