Machine learning application to human brain network studies

Yulia Dodonova[1,2] , **Leonid Zhukov**[2]

[1] Institute for Information Transmission Problems
[2] National Research University Higher School of Economics

Structure

0. Overview

1. Machine learning tasks in neuroscience

2. Why networks?

3. Classification of brain networks

    - General notes and pitfalls
    - Preprocessing
    - Kernel approach
    - Spectral approach

4. Some results and further directions

Machine Learning on Neuroimaging Data

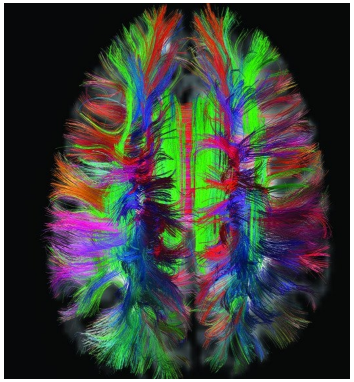
Classify normal and pathological brain structures

Alzheimer's disease, Parkinson disease, Autism Spectral Disorders, etc.

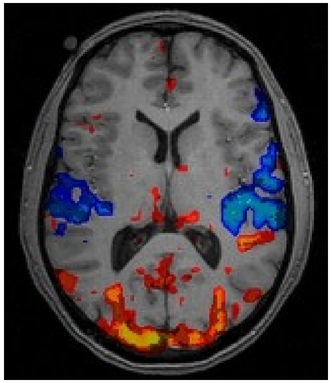
Predict treatment outcome

Huntington disease, stroke, etc.

MRI



Diffusion MRI

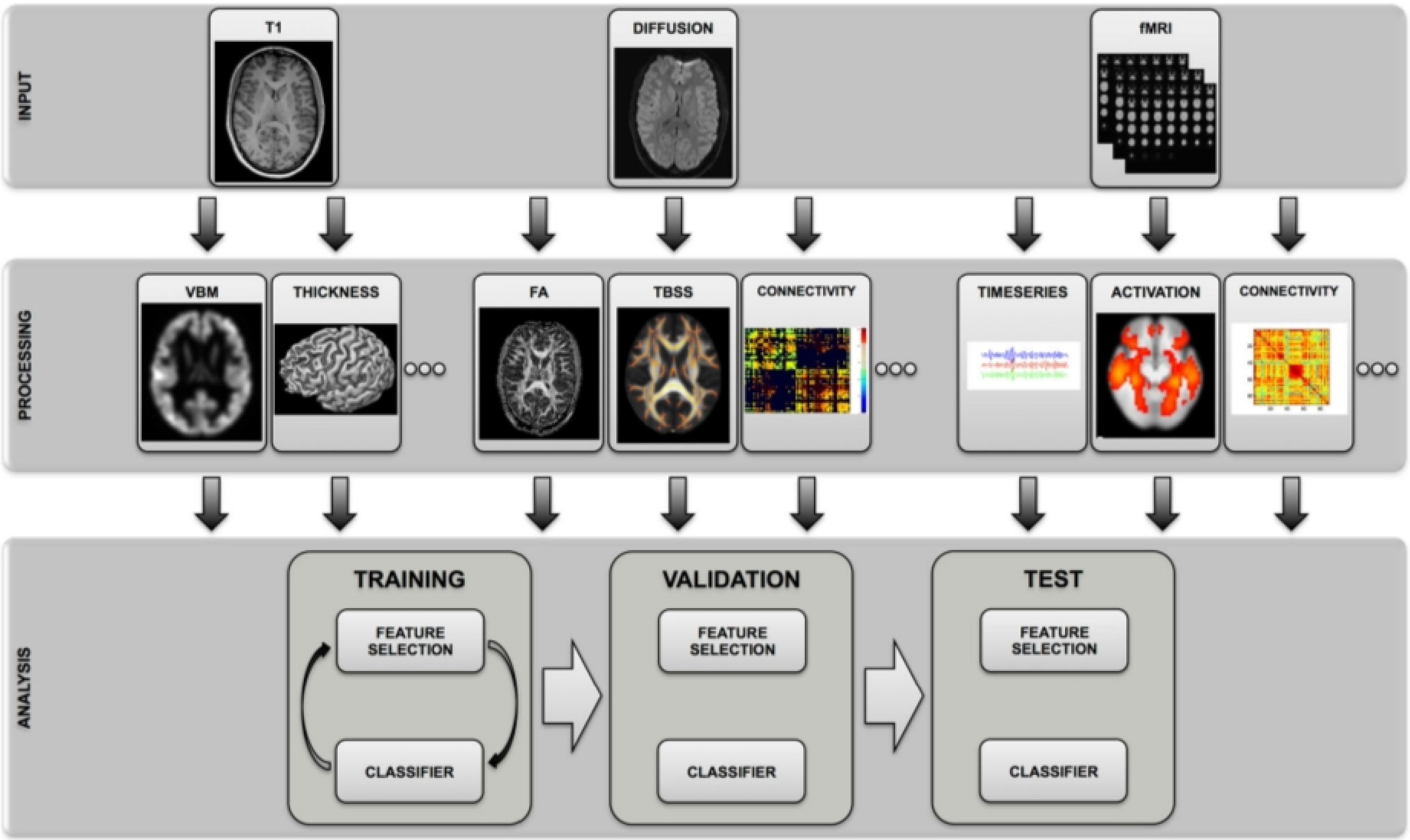


Functional MRI

MRI, dMRI, fMRI

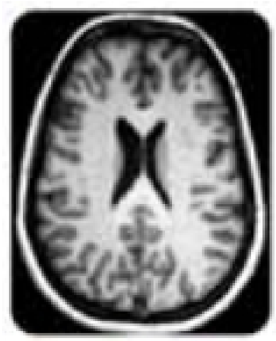
High spatial (~1 mm) resolution,
static or low (seconds) time resolution

General Pipeline



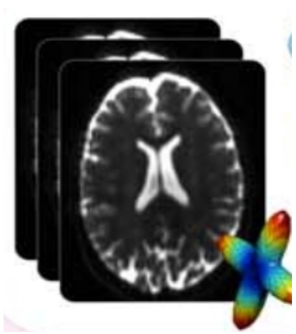
Haller et al. (2014)

Connectomes: Brain Networks



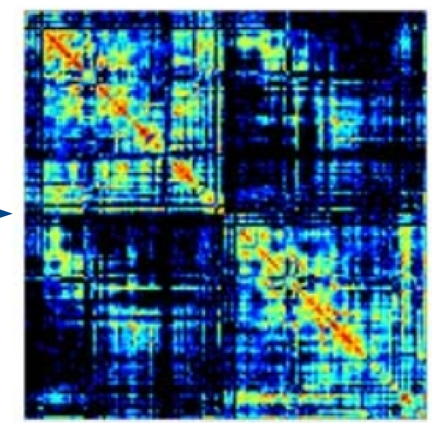
MRI

Zone parcellation – vertices



DTI

Tractography – edges



Adjacency matrices

The term "Connectome" introduced by O.Sporns and P.Hagmann in 2005

Connectomes: general

- Small graphs (~100 vertices)
- Undirected (symmetric adjacency matrices)
- Connected

- Each vertice is uniquely labeled
- A set of labels is the same across networks
- Vertices have 3D coordinates

- Sparse (~10% density)
- Weighted edges (weights are proportional to the number of streamlines between the brain regions)

Connectomes: Machine learning

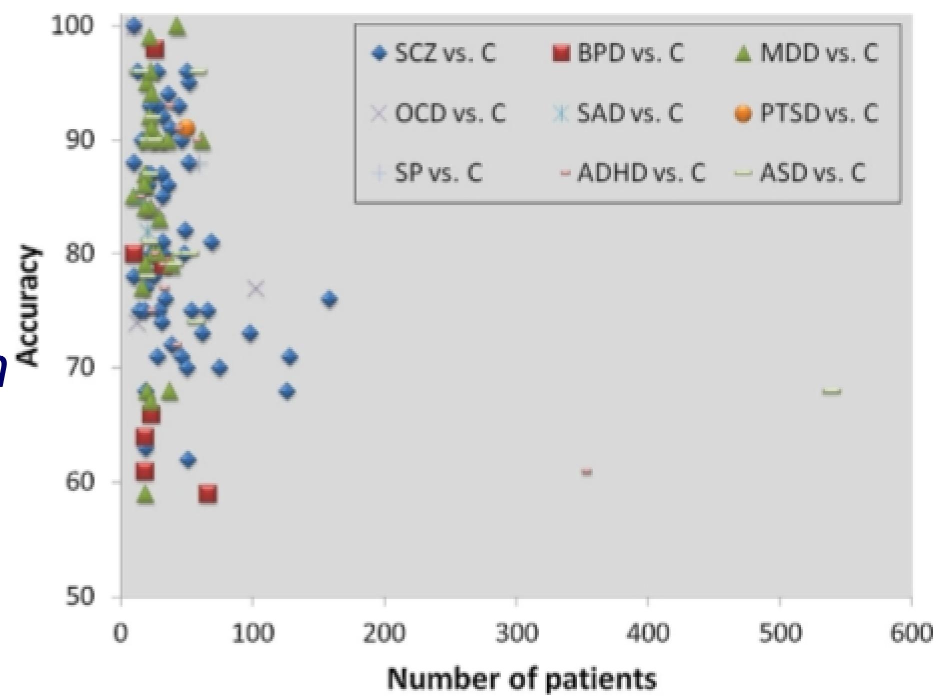
What are the feature vectors?

- "Bag-of-edges" (huge dimensionality)
- Vectors of local metrics (better, but still a lot)
- Vectors of global metrics (too global?)

Graph clustering coefficient, graph characteristic path length, small-worldness, modularity, etc.

What are the sample sizes?

A figure from a review of 118 pattern recognition studies in neuroimaging, Wolfers et al. (2015)



Connectomes: Normalization

Geometric Normalization

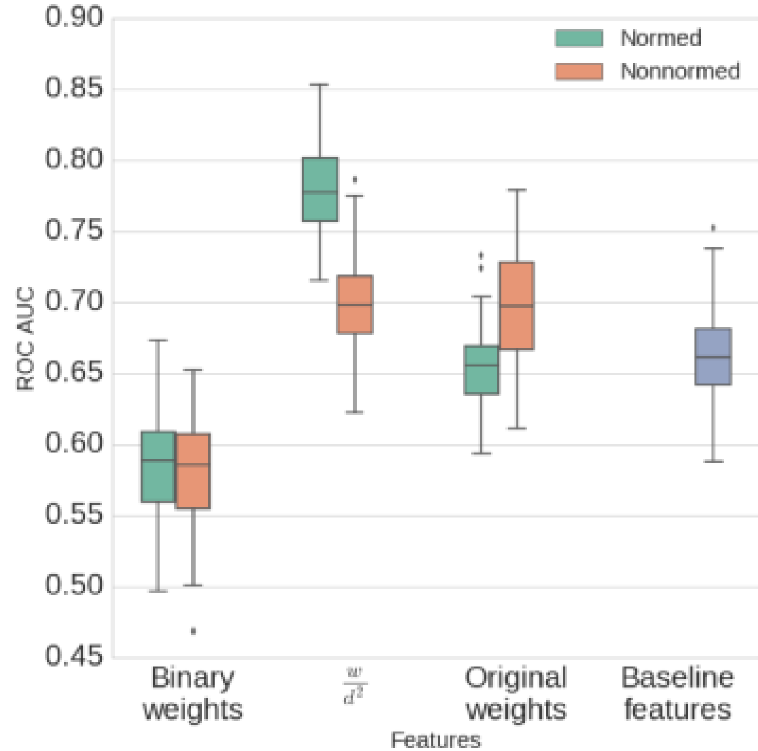
By physical sizes of brain regions, distances between regions, etc.

$$a_{ij}^{weighted} = \frac{a_{ij}}{l_{ij}^2}$$

Topological Normalization

By the maximum weight, sum of weights, etc.

$$w_{ij}^{normed} = \frac{w_{ij}}{\sqrt{d_i d_j}}$$



Petrov, Dodonova, Zhukov (2016) Boosting Connectome Classification via Combination of Geometric and Topological Normalizations.

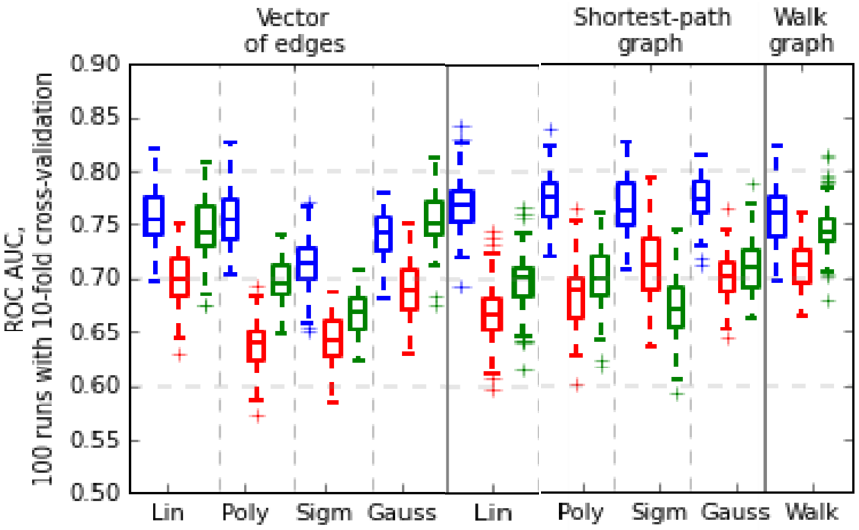
Autism spectrum disorder versus typical development

Connectomes: Graph kernels

Walk kernel:

$$K_{walk\,graph}(G,G') = \sum_{i,j=1}^{|V_*|} [\sum_{k=0}^{\infty} \alpha_k A_*^k]_{ij}$$

$$= \sum_{i,j=1}^{|V_*|} [\sum_{k=0}^{\infty} \alpha^k A_*^k]_{ij} = \sum_{i,j=1}^{|V_*|} [(I-\alpha A_*)^{-1}]_{ij}$$



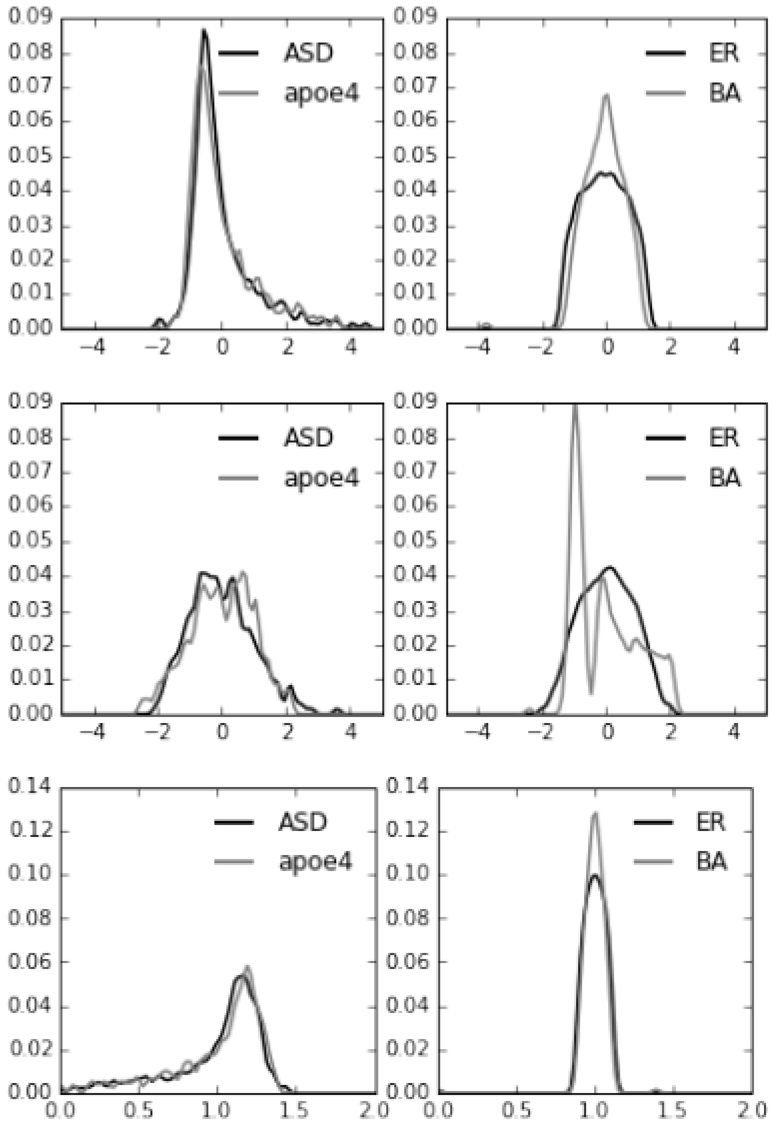
Shortest path kernel:

$$K_{path\,graph}(D,D') = \sum_{\substack{d_{ij}\in D \\ d'_{ij}\in D'}} K_1(d_{ij}, d'_{ij})$$

Dodonova, Petrov, Zhukov (2015) Comparing effectiveness of SVM with different kernels for gender classification based on brain networks (in Russian)

Males versus females

Connectomes: Graph spectra



Spectra of the adjacency matrices

Spectra of the Laplacian matrices

$$L = D - A$$

Spectra of the normalized Laplacians

$$\mathcal{L} = D^{-1/2} L D^{-1/2}$$

See also De Lange et al., 2014 for the very similar plots

Connectomes: Graph spectra

Use spectra as
feature vectors...

Features	Original weights	Binarized weights	Weights by l^2
Edges	0.552±0.038	0.558±0.028	0.550±0.000
	0.570±0.062	0.563±0.062	0.520±0.072
Degrees	0.550±0.000	0.550±0.000	0.631±0.034
	0.548±0.062	0.545±0.079	0.704±0.049
A spectra	0.550±0.000	0.550±0.000	0.550±0.000
	0.717±0.034	0.780±0.039	0.533±0.066
L spectra	0.550±0.000	0.550±0.000	0.604±0.011
	0.525±0.065	0.638±0.041	0.512±0.057
$\mathcal{L}$ spectra	0.550±0.000	0.689±0.043	0.592±0.031
	0.584±0.065	0.647±0.039	0.506±0.067

(a) UCLA APOE-4 dataset

Table 1. Best results of the linear (top row in each cell) and tree-based (bottom row) models.

Carriers versus non-carriers of an allele associated with an increased risk of Alzheimer's

Connectomes: Graph spectra

Use spectra as feature vectors...

Features	Original weights	Binarized weights	Weights by l^2
Edges	0.552±0.038 0.570±0.062	0.558±0.028 0.563±0.062	0.550±0.000 0.520±0.072
Degrees	0.550±0.000 0.548±0.062	0.550±0.000 0.545±0.079	0.631±0.034 0.704±0.049
A spectra	0.550±0.000 0.717±0.034	0.550±0.000 0.780±0.039	0.550±0.000 0.533±0.066
L spectra	0.550±0.000 0.525±0.065	0.550±0.000 0.638±0.041	0.604±0.011 0.512±0.057
$\mathcal{L}$ spectra	0.550±0.000 0.584±0.065	0.689±0.043 0.647±0.039	0.592±0.031 0.506±0.067

(a) UCLA APOE-4 dataset

Table 1. Best results of the linear (top row in each cell) and tree-based (bottom row) models.

Carriers versus non-carriers of an allele associated with increased risk of Alzheimer's

… or produce kernels based on spectral distributions

The Kullback-Leibler kernel:

$$KL(p||q) = \int_{-\infty}^{\infty} p(x) log \frac{p(x)}{q(x)} dx$$

$$K_{KL}(p,q) = e^{-\alpha(KL(p||q)+KL(q||p))}$$

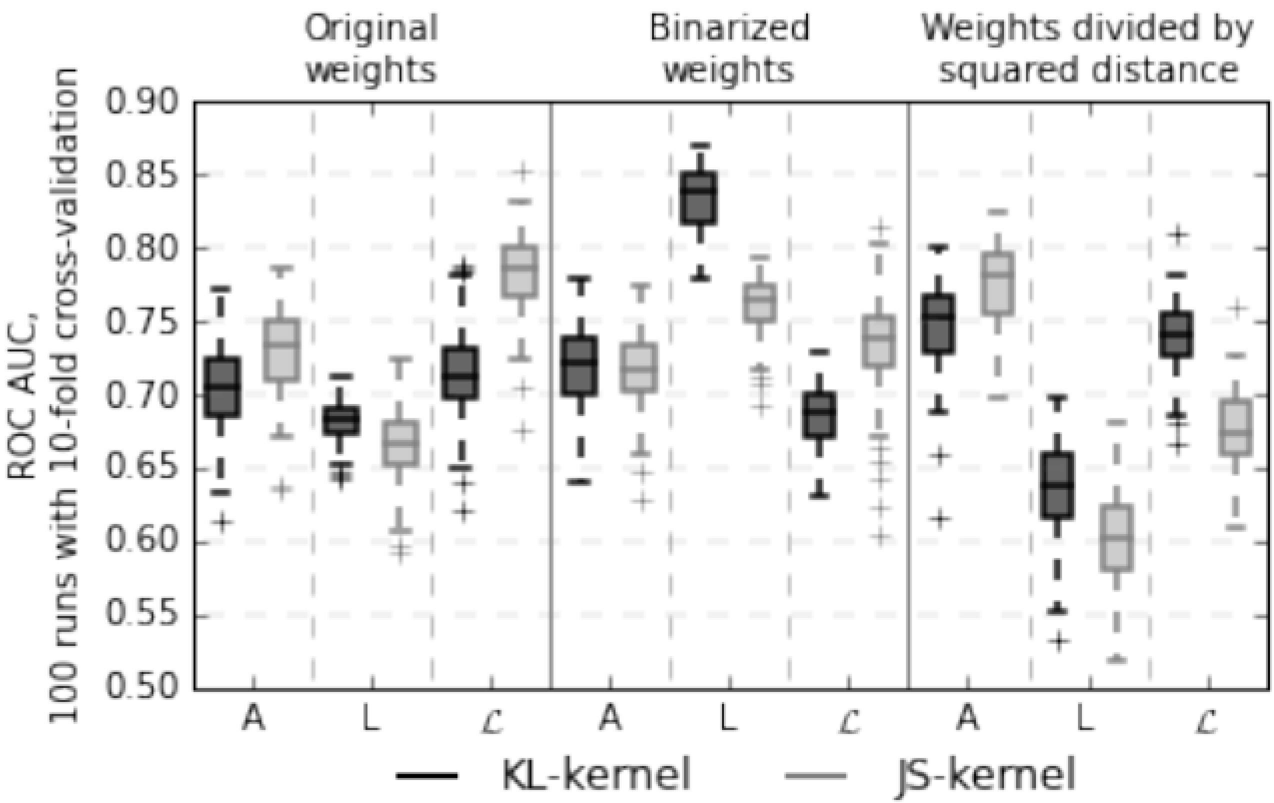
The Jensen-Shannon kernel:

$$JS(p||q) = \frac{1}{2}(KL(p||r) + KL(q||r))$$

$$r(x) = \tfrac{1}{2}(p(x) + q(x))$$

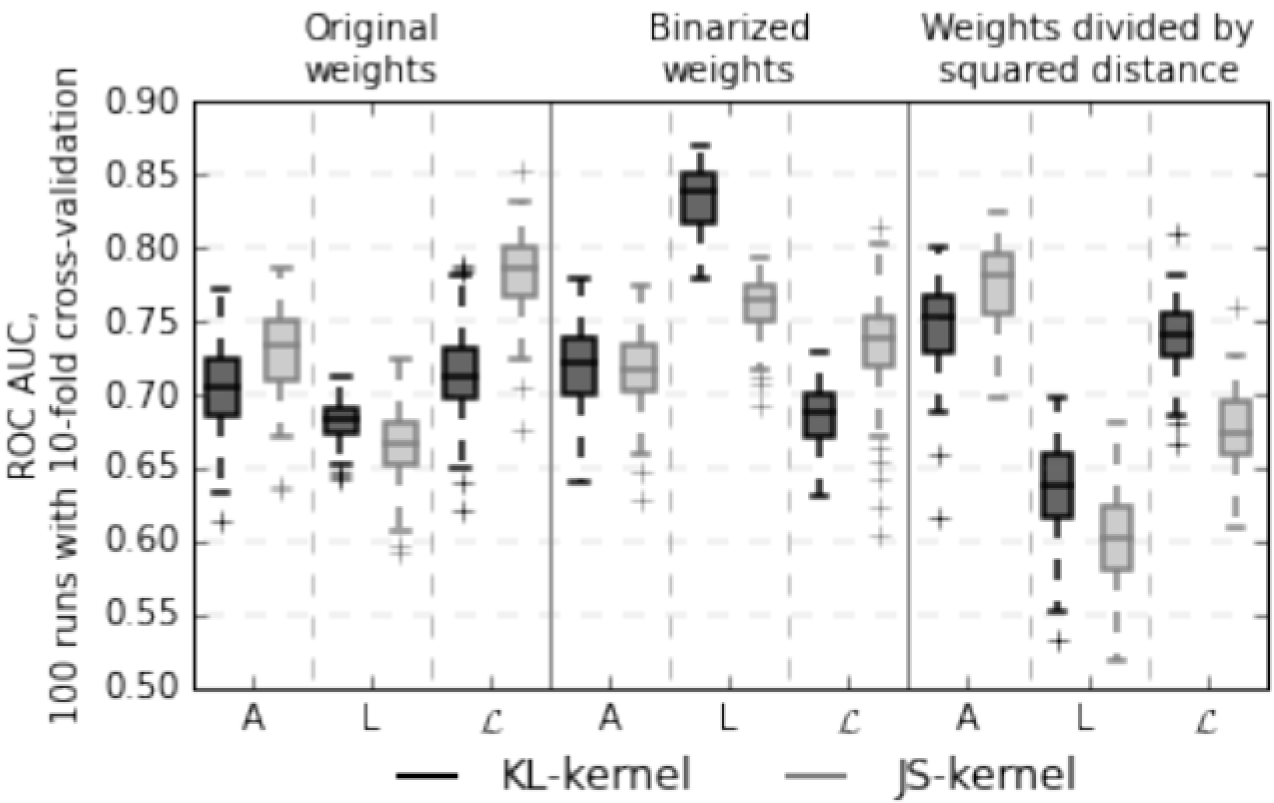
$$K_{JS}(p,q) = e^{-\alpha\sqrt{JS(p||q)}}$$

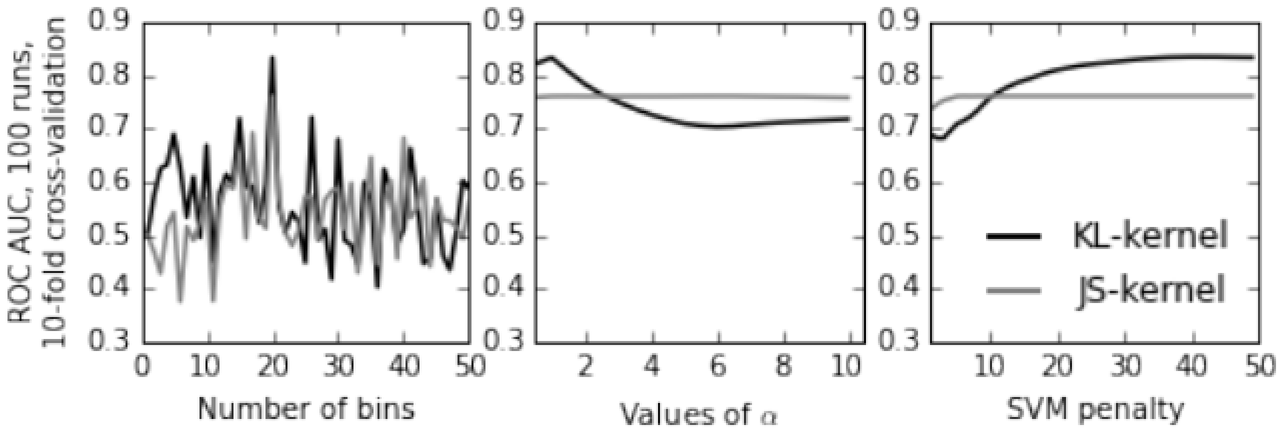
Information divergence kernels on spectra



Carriers versus non-carriers of an allele associated with increased risk of Alzheimer's

Information divergence kernels on spectra





Work in progress:

Avoid explicit density reconstruction

For example,
Earth Mover's distance seems to work

Challenges

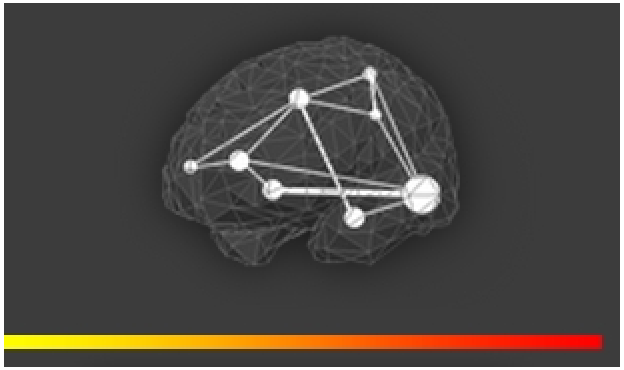
- Normal Anatomical **Inter-individual Variability**
- Dysbalance Between **Number of Features** and Number of Subjects, Data Reduction and Over-Fitting
- **Variability** Related to Patient Selection, Inter-scanner Variability and Data Preprocessing
- The relative **class frequencies** in the training set are often different from the test set or target application domain
- In many cases, the question to be answered will not be related to distinguishing patients from controls; rather, **distinction between different disorders in the same population will be needed**

Connectomes: Data

The Human Connectome Project

The Enigma Collaboration

UCLA Multimodal Connectivity Database and other sources

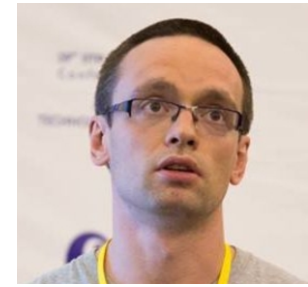


Team

Leonid Zhukov

Mikhail Belyaev

Yulia Dodonova



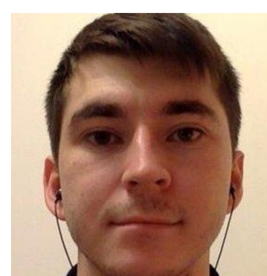
Dmitry Petrov

Kharkevich Institute for Information Transmission Problems, Data Analysis in Neuroscience Group

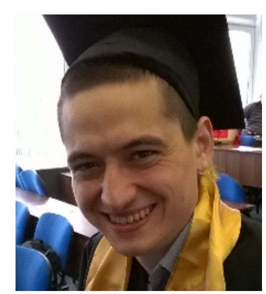
NATIONAL RESEARCH UNIVERSITY

Faculty of Computer Science Machine Learning on Neuroimaging Data Group

Anna Tkachev



Amir Safiullin



Anvar Kurmukov

Q?

Yulia Dodonova
ya.dodonova@mail.ru

Leonid Zhukov
lzhukov@hse.ru