Setup We have a scalar response variable, say Y, which is the IQ score (a scalar). Then we have two different predictors, say X_1 and X_2, which are two different connectivity matrices: a functional one (fcMRI) and a structural one (DTI). Thus, we observe both the functional and the structural connectivity matrices for each individual, and we'd like to predict his IQ score based on this information.

Files Usually, 8 files per patient: 4 DTI and 4 fcMRI. fcMRI can be either GSR or noGSR (we prefer the latter; will provide info about this upon request). For each type (DTI or fcMRI), there are 1 numerical file containing the connectivity matrix and 3 other files containing the labels (see below).

Data 207 patients. Only 196 have DTI. Only 171 have noGSR fcMRI. All patients who have noGSR fcMRI also have DTI. Among these 171 patients, only 159 also have the response WASI_FULL4 (the IQ score). Therefore, we have 159 observations in total.

Graphs The connectivity matrices are all undirected and weighted. The latter could be a problem for the graph-coeff case since we use k-cores, and not s-cores (the generalization to weighted graphs). Also, it could be the main reason for the not-so-satisfying results that Gabriel achieved with these data.

Labels So, for each individual we have a single numerical file, which contains the values of the connectivity. Also, we have three label files, which contain the headers for the connectivity matrix. They are three because they contain different kind of labels, but they all refer to the same numerical file. The three kind of labels are:

1. the (full-name) zone of the brain, as in the files "xxx_region_names_full_file.txt";
2. the abbreviated zone of the brain, as in the files "xxx_region_names_abbrev_file.txt";
3. the exact position wrt the xyz space, as in the files "xxx_region_xyz_centers_file.txt".

However, 1 and 2 should be equivalent.

Problem 1 (labels) There is no one-to-one correspondence between 1/2 and 3, meaning that more than one xyz may belong to the same region (i.e. there are duplicated labels in 1 and 2). Possible solutions: **using directly 3**, which is certainly more precise, **or instead using 1/2** (sticking to the "official" convention), maybe averaging among all the data points that share the same region label (?).

Problem 2 (labels) The ordering of the labels seemed to be different between individuals at a first glance, but then it turned out that only 1 is disordered (wrt to the others, and also between patients). Therefore, it is probably mistaken. **We can use 2 or 3 then**, ignoring 1 with no remorse since it should contain the same information of 2.
Warning: if this is correct, then why should we need a distinct label file for each patient?

Conclusion (labels) Let me begin by saying that the labels are just a matter of interpretation, so from a computational point of view we can merely take the numerical files and use them. We just need to make sure the ordering of the rows/columns is preserved, so that afterwards we can understand to what region/point a single line is referred. That being said, we can use 2 if we want to stick to the official convention, but then we would have to deal in some way with the duplicated labels, or 3.