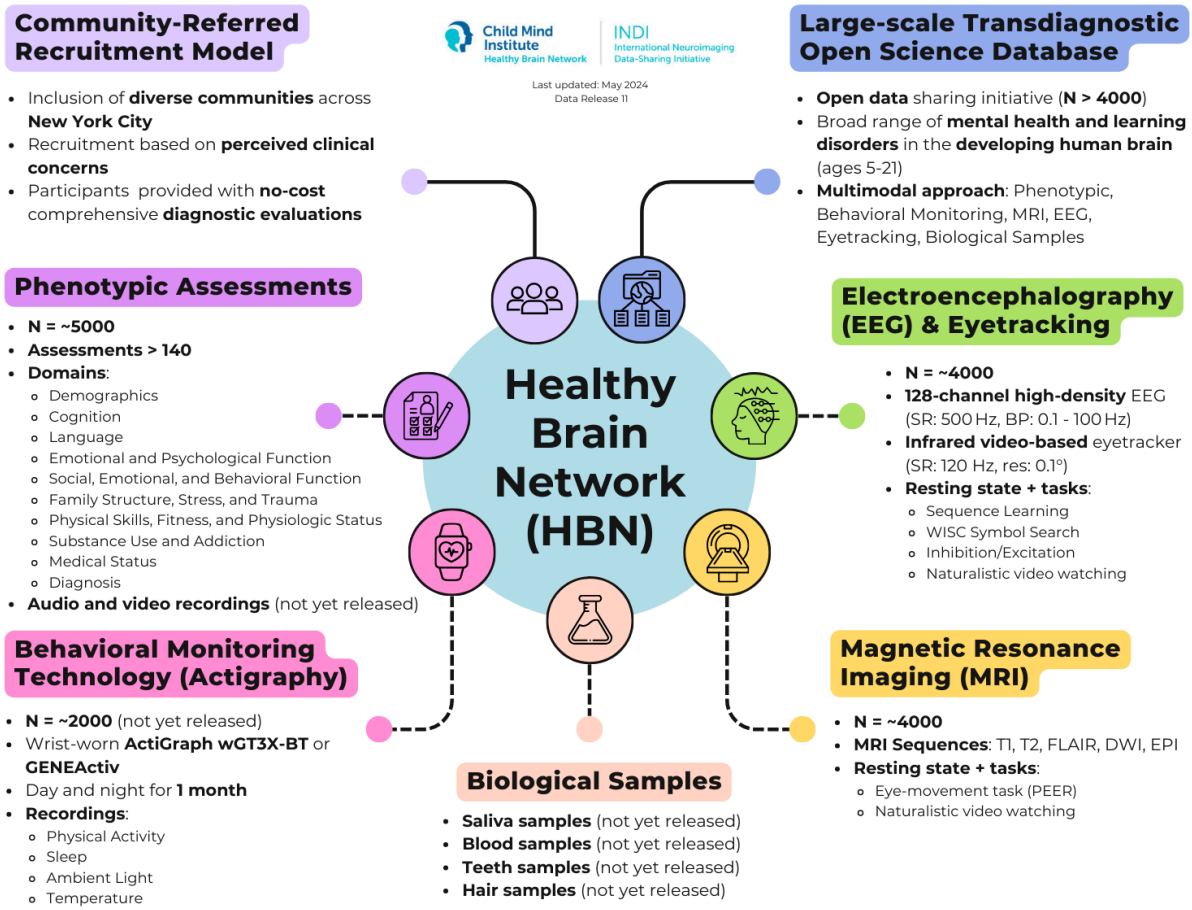
The Healthy Brain Network

All datasets for this University challenge are provided by the Healthy Brain Network, in collaboration with the Reproducible Brain Charts project.

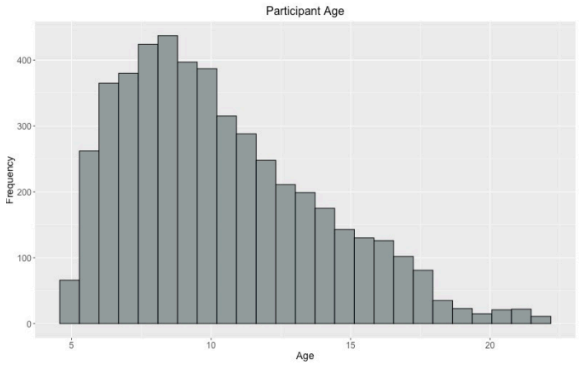


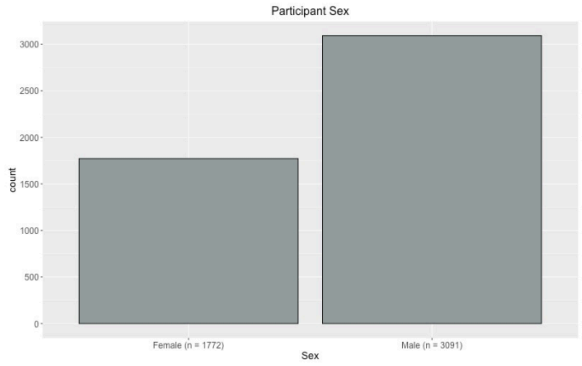
The Healthy Brain Network is a **community-based research initiative of the Child Mind Institute**, an independent, international non-profit dedicated to transforming the lives of children and families struggling with mental health and learning disorders.

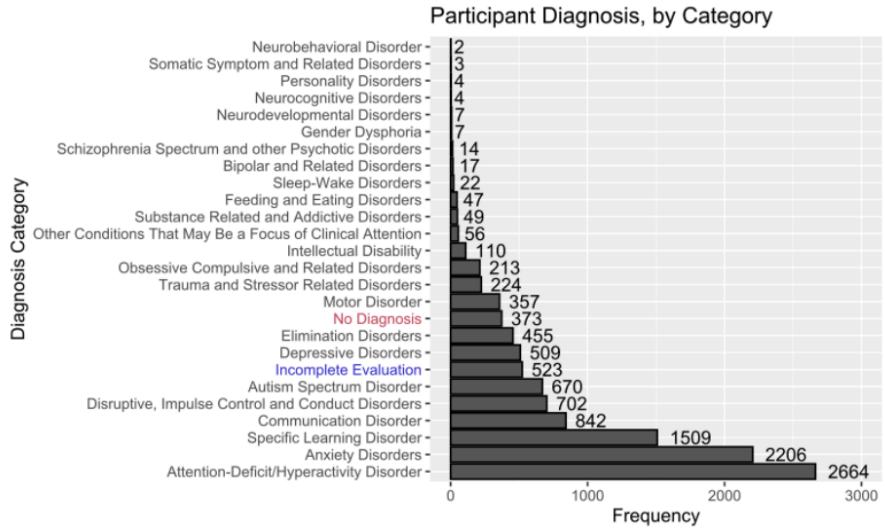
HBN utilizes a **community-referred recruitment model** by encouraging participation of families who have concerns about mental health in their children and by offering **no-cost diagnostic evaluations**. As part of the CMI open science effort, HBN releases multimodal datasets capturing a broad range of clinical psychopathology in children and adolescents.

Between 2015 and 2024, HBN collected and openly shared behavioral-monitoring, magnetic resonance imaging, electroencephalography, and eye-tracking datasets from **over 5000 children and adolescents, along with data from over 140 phenotypic and diagnostic**

assessments, encompassing a wide variety of domains, including cognition, language, emotion, and fitness. This extensive transdiagnostic dataset has contributed to the work of hundreds of researchers and clinicians, and HBN data has been used in over 350 scientific publications.





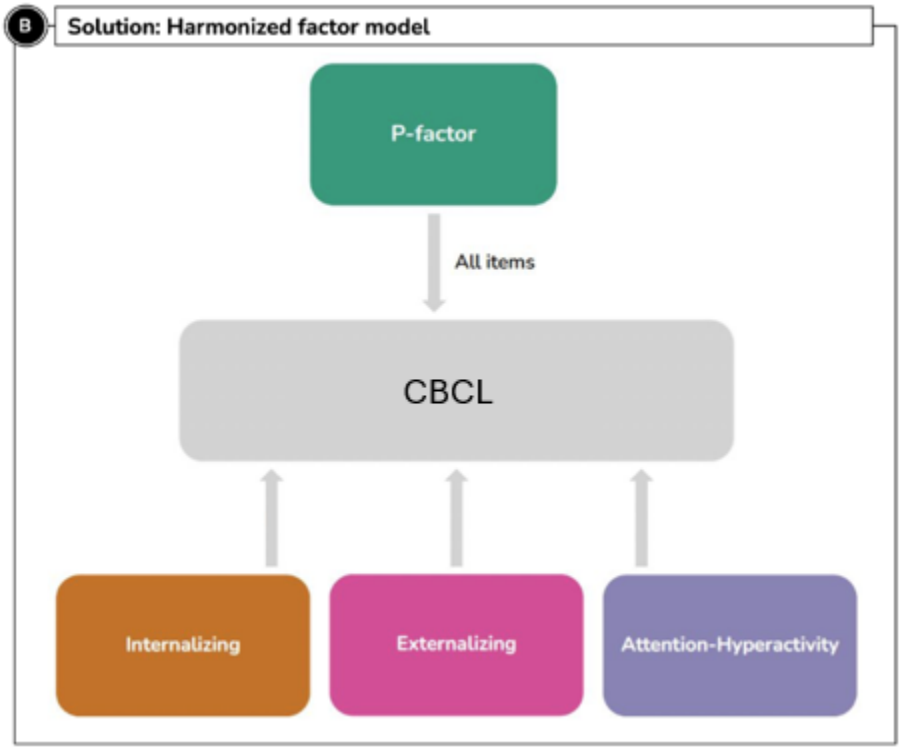


Participants in the HBN study **range in age from 5 to 22 years**, with the majority falling between 7 and 12 years old. In the latest HBN data release (which is release 11), there are **1,772 female participants and 3,091 male participants** (in the dataset, 1=female, 0=male). Please note that this information corresponds to sex assigned at birth and not gender identity. Finally, our data indicate that the majority of HBN participants have received a diagnosis. **Only around 8% of all participants received no diagnosis, including 10% of the female participants and 6% of the male participants**. The majority of children and adolescents who participated in the HBN study received a diagnosis for **Attention-Deficit/Hyperactivity Disorder (or ADHD) or Anxiety Disorders**. Some participants can receive more than one diagnosis.

In this challenge, you will work with a subset of the whole HBN phenotypic dataset, and in particular with demographic data, such as age and sex, factor data derived from the Child

Behavior Checklist (or CBCL), and Magnetic Resonance Imaging (or MRI) data. MRI data were collected in 4 different sites: Staten Island, RUBIC, CBIC, and CUNY. As part of the HBN study, multiple MRI datasets were collected, such as structural (or T1-weighted), diffusion-weighted imaging (or DWI), and functional MRI (or fMRI). fMRI data include both resting state data and naturalistic video watching data. Here, you will work with **functional connectivity matrices derived from resting state fMRI data**.

P-factor and bifactor analysis



Phenotypic assessments can be analyzed in several different ways, either through analyzing itemized responses, through a weighted average of the responses, or the use of **more advanced methods**. Performing a factor analysis is one of the advanced methods to obtain a summary of phenotypic data.

Factor data such as **p-factor, Internalizing, Externalizing, and Attention-Hyperactivity factors,** derive from the modeling of the HBN **Child Behavior Checklist (CBCL)** using a **bifactor model**. CBCL is a parent-reported assessment of problem behaviors and competencies in a variety of domains across children and adolescents aged 6 to 18 over the past 6 months (Achenbach, 1991). For instance, parents can be asked whether the child can't concentrate, whether they are too fearful or anxious, or whether they seem unhappy, sad, or depressed. For a complete list of CBCL questions, you can consult the CBCL data dictionary. Scores of CBCL include Anxiety/Depression problems, Social Problems, Attention Problems, among others.

Psychopathology often involves **comorbid disorders**, which is the co-occurrence of psychiatric disorders such as depression and anxiety. To better understand co-occurrences of psychiatric disorders, a bifactor model can be used to parse comorbid conditions across independent dimensions of psychopathology. **These models can differentiate between psychopathology features that are common across multiple disorders and those that are specific to**

individual disorders. The factor that is common across disorders is called **p-factor**. You can think of it as a common thread that links various mental health disorders together. This helps us understand that different mental disorders (like anxiety, depression) might share something in common and that someone showing a specific mental health disorder might be more likely to show another because they are all connected through this general factor. Conversely, **specific factors** are characteristics that make each disorder different from others, e.g., depression different from anxiety. This approach has gained popularity in psychology and psychiatry research as a method to **reach a complete understanding of psychopathology as well as to identify causes, neurobiological substrates, and behavioral outcomes** (Bornovalova, 2020).

Let's get into the methodological details of how factor data were extracted:

A bifactor structural model specifies that the covariance among a set of item responses can be accounted for by a **single general factor called p-factor**, that reflects the common variance running among all scale items, and **specific factors**, that reflect residual variance among clusters of items. This reduces the number of variables into a smaller set of factors that explain the patterns of correlations among the observed variables.

To estimate p-factor and specific factors, a **confirmatory bifactor model** was used:

- As previously mentioned, a **bifactor** analysis is a type of model where general and specific latent factors are hypothesized to influence the response of items, which are the observed variables.

- **Latent factors** are factors that are not directly observed but are inferred from the relationships between observed variables.

- The p-factor is the general factor that captures the **broad psychopathological risk**; the **Internalizing, Externalizing, and Attention-Hyperactivity factors** are the specific factors that capture different types of symptoms or traits once the general factor is out.

- **Internalizing factors** reflect emotional or mood problems such as depressive and/or anxiety symptoms, withdrawal, and somatic complaints; **externalizing factors** reflect dysregulations in behavior including conduct problems, aggressive and rule-breaking behavior; **attention/hyperactivity factors** reflect inappropriate levels of inattention, as well as hyperactivity and impulsivity.

- Importantly, specific factors are **orthogonal** to the p-factor and among each other, and potentially explain item response variance that is not already accounted for by the p-factor. In other words, the p-factor explains the variance shared by all psychiatric symptoms/disorders, with additional variance accounted for by Internalizing, Externalizing, and Attention-Hyperactivity factors.

- A **confirmatory analysis** is an analysis that, rather than exploring the underlying factor structure, tests whether predefined general and specific factors adequately explain the relationships between observed responses (here, HBN CBCL responses). In particular, the model described in McElroy et. al (2018) was tested.

- Here, p-factor scores were **z-scored** and a higher p-factor score indicates that a child is experiencing a **broader and more generalized level of psychological difficulties**.

In conclusion, general and specific factors are not a simple summary statistics of the CBCL responses, but rather represents a latent variable, derived from a confirmatory factor analysis, that reflects the **interrelationships among the emotional and behavioral symptoms extracted from the HBN CBCL responses** (Hoffman et al., 2022, Hoffmann et al., 2023).

The factor dataset for this challenge includes **a participant identifier column and four columns with numerical values for the following factors**: (1) p-factor, (2) Internalizing factor, (3) Externalizing factor, and (4) Attention-Hyperactivity factor. The factor values have been z-scored, which standardizes the data by quantifying how many standard deviations each data point is from the mean of the group.

Functional connectivity matrices processed through RBC

Functional connectivity, which is a measure that attempts to summarize how different brain regions may interact with each other, is a vital part of brain imaging. Functional connectivity graphs or matrices were created from resting-state functional Magnetic Resonance Imaging (or rs-fMRI) data.

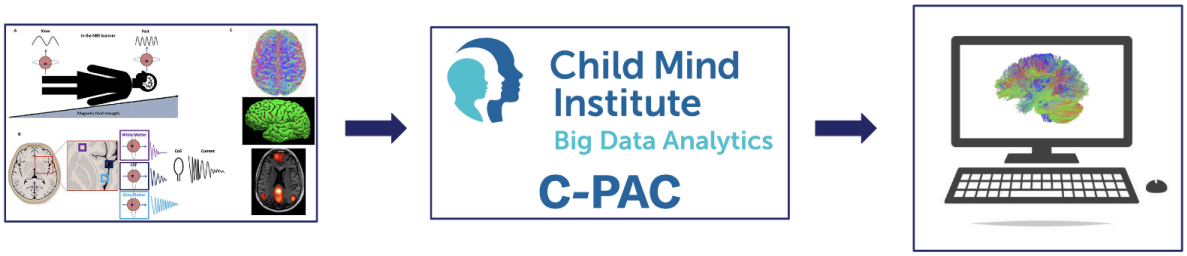
In brain imaging, the **MRI technique** involves the use of radio waves and a strong magnetic field to generate images of a subject's brain and to record brain activity. During an MRI scan, the subject lies down inside a large, tube-shaped magnet. Radio Frequency electromagnetic pulses are then introduced, and the machine "listens" to how the molecules realign themselves, giving us a sense of the sorts of tissue or activity that are present in each region.

Functional MRI (fMRI) uses a principle called Blood Oxygen Level Dependency (BOLD) to measure brain activity. When a brain area is more active it consumes more oxygen and to meet this increased demand blood flow is directed to the active area. Oxygen is transported by hemoglobin in the blood, and the magnetic properties of hemoglobin change depending on whether it is carrying oxygen or not. BOLD fMRI detects these changes in magnetic properties, allowing researchers to indirectly measure neural activity by observing changes in blood flow.

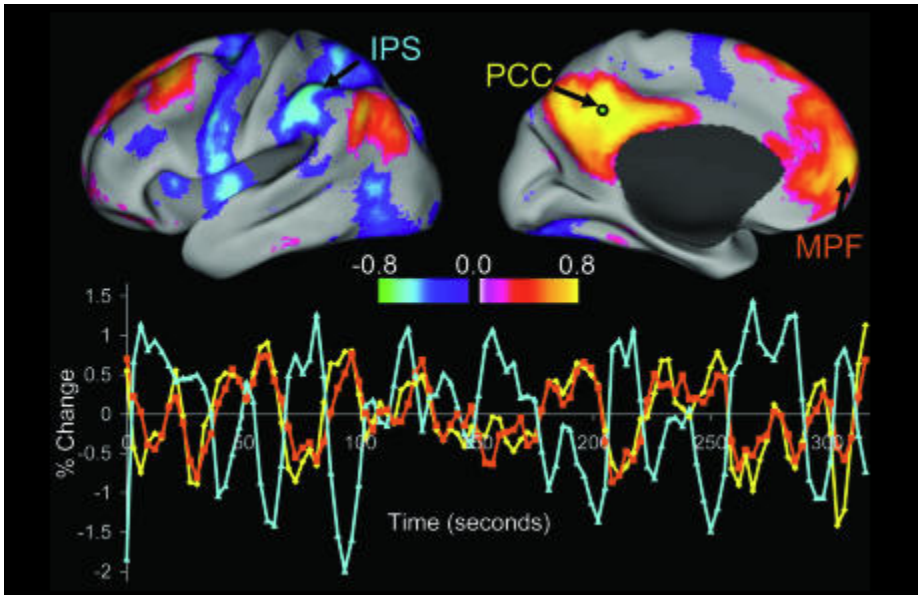
Functional connectivity data is used in neuroscience research to understand the relationships between different regions of the brain, and chart how they each experience change over time. By having a measure of this change, researchers can understand how the brain works in response to specific stimuli and tasks, or even during resting state. **Resting-state** refers to any period of time in which the brain of a subject is awake and alert - not actively engaged in a task, but not asleep. You can also think of this state as a sort of mind-wandering.

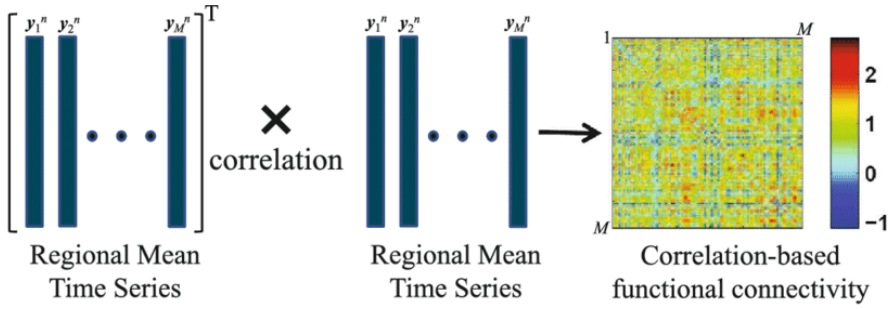
The functional connectivity matrices included in this dataset were created by a software designed specifically for processing resting-state MRI data. This software tool (known as **C-PAC, or the Configurable Pipeline for the Analysis of Connectomes**), was used to process the resting state MRI data from the Healthy Brain Network. FC data were created as part of the **Reproducible Brain Charts project (RBC)**, which is an initiative that aims to aggregate several of the largest studies of brain development in youth as publicly available data resources for the scientific community.

You can think of C-PAC as a tool that is used to distill the signal within raw data collected by an fMRI scan. The data undergo a variety of transformations and realignments, ensuring that the signals can be compared across individuals. Subsequently, you can think of **functional connectivity as the measure of how different parts of the brain interact with each other**, as it is the correlation between activity in each pair of regions**.** Functional connectivity data is often understood in the form of a **matrix or a graph**, in which each cell of the matrix captures each pair of correlations.

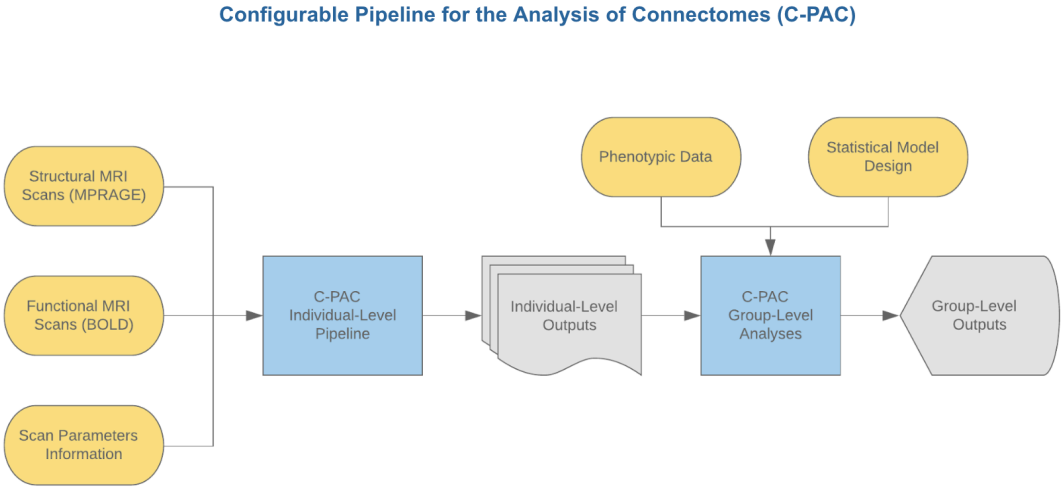


Acquisition (MRI) Preprocessing (C-PAC) Connectome





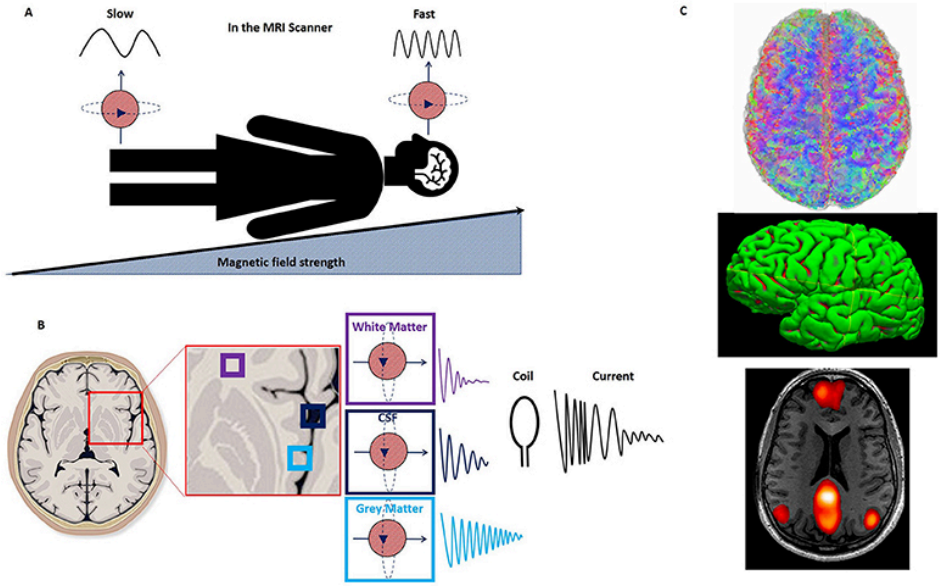
Now, let's get into the **methodological details of C-PAC** and how functional connectivity data was computed:



C-PAC builds upon a robust set of existing neuroimaging software packages and makes it easy for both novice users and experts to **clean and explore their data using a wide array of processing tools**. Users define analysis pipelines by specifying a combination of options and analyses to be run. Results can then be compared across groups and individual subjects.

Group results are advantageous in the sense that they typically have large participant numbers and can often be generalized to a specific population. Single-subject results, which are not as generalizable, offer unique insights about the intricacies of one person's brain functionality. For example, it could be interesting to see the difference in functional connectivity for one person when they are presented with any given task A as compared to the functional connectivity for the same person when they are presented with any given task B.

C-PAC relies on both structural and functional neuroimaging data, where structural scans capture high-resolution images of brain anatomy, and functional scans capture ongoing variation in the brain activity. Magnetic Resonance images are captured by introducing a magnetic field and perturbing pulses to the tissue, and recording the ways in which each voxel (small cube) of tissue responds. In the case of structural images, often measured by T1-weighted (T1w) contrast, white areas indicate connective (white-matter) tissue, gray areas predominantly contain the nerves of the brain itself, and black areas are largely cerebrospinal fluid. For functional imaging, as mentioned earlier, this depends on the BOLD signal. A typical C-PAC analysis pipeline will take in one T1-weighted image and one functional **BOLD** image.



While **structural imaging data** can be analyzed on its own to test certain hypotheses, **functional imaging data** requires the analysis of structural data, as well. For example, if a researcher was interested in measuring differences in gray matter volume as a function of age, they could process T1w images and make estimates of this change. Conversely, if a researcher was interested in exploring the brain areas involved in finger movement, they could investigate the BOLD activity in different regions of interest in the brain at the time of a finger-movement task, but to do so would require the analysis of structural data as well, so that they would have a precise measurement of the regions themselves.

The processing of these images together involves many interconnected stages, and one such workflow from C-PAC processing is shown in the figure below. The structural images are first aligned with anatomical templates that are used to compare images across individuals. Then, structural and functional images are aligned together to accurately map brain activity onto anatomical structures. Finally, functional connectivity matrices are generated by **correlating BOLD activity in different regions of the brain**. In this case, a functional connectivity matrix would typically be a Pearson correlation of the timeseries of a pair of **voxels** or regions of interest (**ROIs**) defined prior to the study. Functional connectivity succinctly tells us which

regions of the brain have the same patterns of activity, opposite patterns of activity, or little to no relationship.

Functional connectivity matrices are always **square and symmetric**, corresponding to the number of regions of interest being used in a given anatomical atlas. For instance, using a 100 region atlas, the functional connectivity matrix would have a shape of 100x100, where each value corresponds to the connectivity among that individual pair. For this challenge, you will be provided with 200 x 200 connectivity matrices.