Explainable Projections:Application to PREDICT-HD

FINAL PROJECT DESCRIPTION

Sungmin Hong
sh4174@nyu.edu, N16303443

Project page (on Github):
https://github.com/NYU-CS6313-SPRING2016/Group-16-Explainable-Projections
Video: https://vimeo.com/167444745
Working demo:
https://github.com/NYU-CS6313-SPRING2016/Group-16-Explainable-Projections
(No public access allowed)

What is the problem you want to solve and who has this problem?

[Problem Statement]
Huntington's disease(HD) is a fatal neurodegenerative disease which is caused by a known genetic pathology in Huntington's protein. Clinical symptoms of HD include motor, cognitive and psychiatric abilities of patients. HD diagnosis is determined based on unequivocal movement disorder. Despite of an utmost demand of identifying features which can be used as a biomarker of the HD progression prior to disease manifestation, significant features have not yet been discovered. Most of previous HD studies relied on brain sub-region volumes, especially in a striatum, to show the correlation of the degeneration of brain and HD progression. Although the degeneration is most significantly observed in a striatum, the other parts of brain also degenerate while HD progresses. Also, the brain sub-regions are not independent from each other since they are all connected and adjacent. HD does not only affect the particular sub-regions of brains as already discovered, but it might also affect the other parts of brain. Thus, the relationship of changes in brain sub-regions and HD progression is complicated since there are numerous sub-regions which results in high dimensional features. CAP score, which is the product of age and a gene status, is suggested as the estimated HD progression score in clinical studies. Prodromal HD patients are classified into different risk groups (controls, low, medium, and high groups) based on their CAP scores. Although a few recent studies have been published to support the relationship between CAP

scores and prodromal HD progression, solid proofs have not been suggested. Thus, the clinical risk group analysis based on CAP score needs more supports from data to prove the correlation to prodromal HD progression.

[Analytical Questions]

1. Are brain subregion volumes related with Huntington's disease? If so, which brain subregions are related with disease progress?

There are 135 brain subregion volumes in the data. Each brain subregion represent a part of our brains, such as, putamen, white matters, accumbens, and etc. We want to analyze the Relationship between brain subregion volume changes and HD progress. It is known that basal ganglia of our brains(putamens, caudates, and ventricles) is strongly related with HD progress. However, how it is related to HD is not clearly identified. Moreover, relationships between other brain sugregions and HD are completely unknown.

2. How are brain sub-region volumes correlated?

The brain subregions are not independent to each other. For example, when cerebellum expands, it pushes the nearby basal ganglia regions. Thus, although the dataset has all brain subregion volumes some might be redundant for data-analysis because it might not be independent enough to be included as features. So we want to analyse the correlation of brain subregions by volumes to select the most effective brain subregions for analysis.

4. How are risk groups projected on different brain subregion spaces?

Risk groups are classified by disease progress, low, medium, and high groups and a control group. The correlation score is one measure to show the relationship of brain subregion volume changes and disease progress. However, there are numerous hidden information and relationship which cannot be represented by a single measure. Those hidden information and relationship can be revealed by exhaustive research with right direction. Exploratory analyses on patients data with different perspectives will help users understand what directions are hidden in the data.

3. How brain subregions volume change over ages with respect to risk groups?

Aging is one of most significant cause of brain volume degeneration. Even people without neurodegenerative diseases have brain volume degeneration for many brain subregions. Also, CAP score(risk score for premanifest HD patients) is also defined by the product of age and genetic impairment rate in a specific gene. Thus, there is a need to investigate the effect of aging with respect to brain volume changes and see how different risk groups have different trends in brain subregion volume changes on ages. Unlike most of neurodegenerative diseases, HD is manifested on patients on various ages, which means there is no obvious relationship with HD and ages, such as, it does not likely to manifest to young patients. Therefore, a visualization framework needs to show how different risk groups are distributed on ages for different brain subregions.

What does your data look like? Where does it come from? What real-world phenomena does it capture?

PREDICT-HD is a NIH funded project which aims to document the natural history of prodromal HD progression in the largest cohort ever studied and to develop neurobiological predictors of HD progression. There are 321 participants with 857 MRI T1 scans (64% female, 36% male). The number of scans of a control group (Non-HD) is 243, and the number of risk group scans is 614. Most of data attributes are quantitative, including age, brain sub-region volumes, and CAP score. CAP group is an ordinal attribute, which is derived from CAP score.

Attribute Names	**Attribute Types**	**Explanation**
Age	Quantitative	Age of patients
Brain Subregion Volumes	Quantitative	Volumes of brain subregions of patients. Total 135 brain subregions are segmented and manually cleaned.
CAP Score	Quantitative	Clinically estimated disease burden score by genetic impairment of age.
CAP Group	Ordinal (Derived from CAP Score)	Clustered risk groups by thresholding CAP scores into 4 groups(Control, low, medium and high groups)

Other than given data from PREDICT-HD database, we derive the correlation between volumetric features by conventional correlation measures, such as, Pearson correlation to show how features are related and affect each other. Also, the correlation between brain subregion volume changes and disease burden (CAP score) is also derived from brain subregion volumes and CAP score to show how subregion volumes are related with disease risk burden.

Attribute Names	**Attribute Types**	**Explanation**

Correlation between Brain Volume Features	Quantitative	The correlation between brain sub-region volumetric features.
Correlation to CAP score	Quantitative	Correlation of brain subregion volume changes and CAP scores

What have others done to solve this or related problems?

There are no directly related works published in my limited knowledge. Paulsen et al. showed that the statistical significance of the correlation of volume changes in basal ganglia regions and HD progression [1]. Like Paulsen et al., the most of HD studies have focused on revealing the relationship between HD progression and an individual feature or a few selected features rather than using dimensionality reduction techniques, or change feature selections. Stahnke et al. suggested an interactive framework which aims to interpret dimensionality reduced data [2]. The suggested framework allows users to select data and look into the features with respect to selected data more closely to understand why the data are clustered together, or why they have small or large distance in the given projection space. Although the feature correlation and the axes of the projection space were not well defined, Stahnke's work shows comprehensive ideas how high dimensional data are projected on 2D space.

Initial Mockup

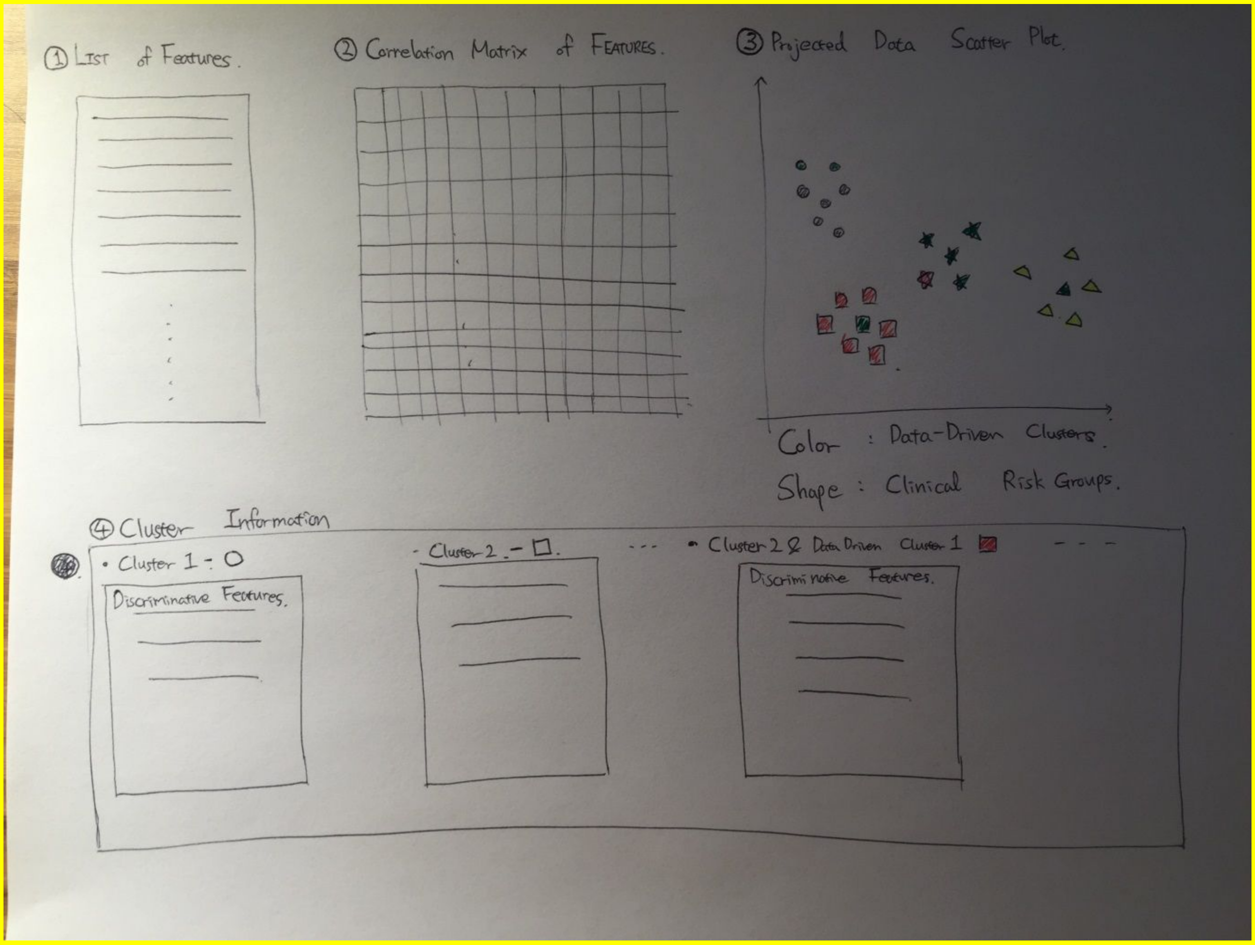

① List of Features.
② Correlation Matrix of Features.
③ Projected Data Scatter Plot.
Color : Data-Driven Clusters.
Shape : Clinical Risk Groups.
④ Cluster Information
• Cluster 1 - ○
Discriminative Features.
- Cluster 2 - □.
• Cluster 2 & Data Driven Cluster 1
Discriminative Features.

Figure shows the mock-up of the suggested framework.

(1) List of the sub-region volume features in the predefined order.
- The list can be sorted by individual feature importance or the volume values.

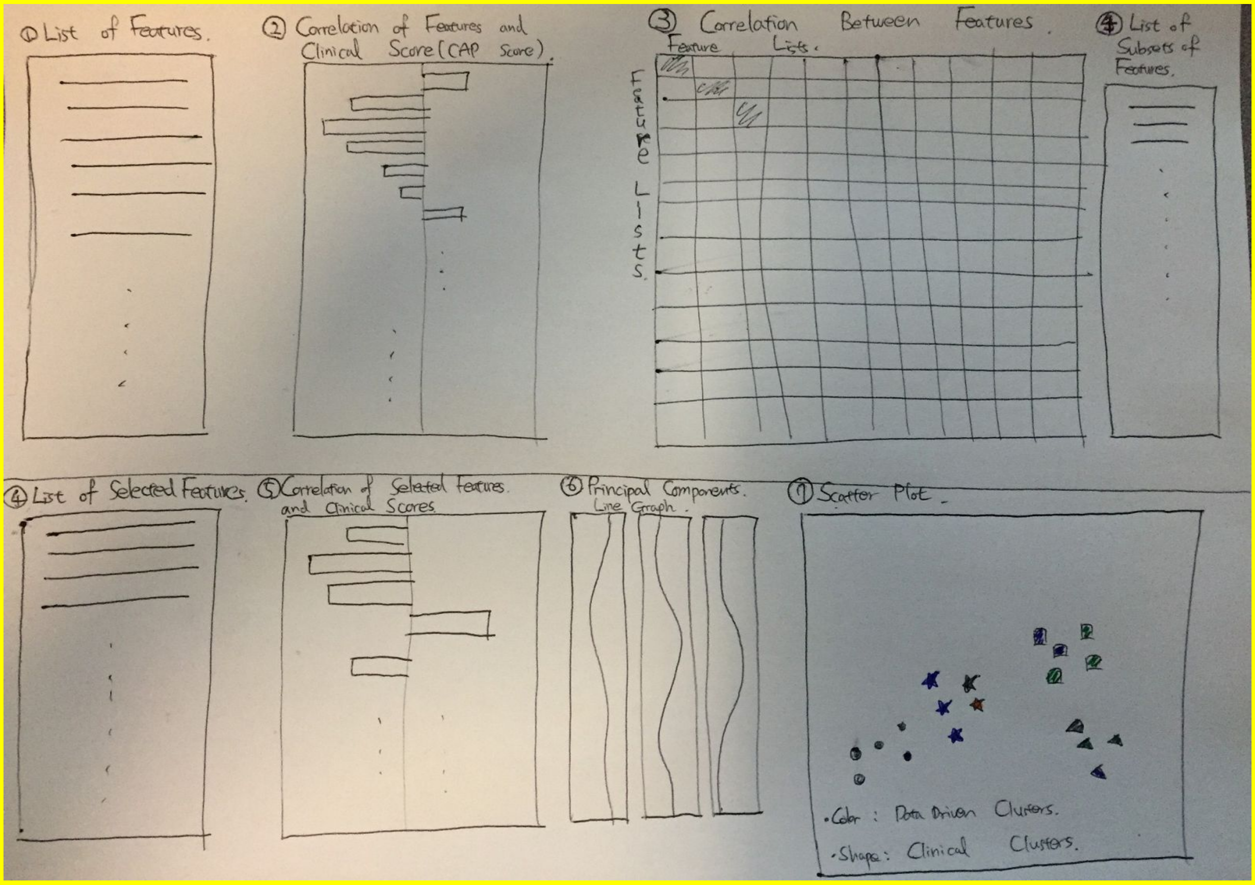
(2) Heatmap of Correlation between features
- plot the correlation matrix of volumetric features by ordinal color.

(3) Scatter plot of projected data on 2D space and data-driven clusters and clinical risk groups.
- The data driven clusters will be represented by categorical colors since the number of clusters can vary. The clinical risk groups will be displayed by shapes since it only has four different groups. The colored shapes will display how the data-driven clusters and the clinically decided clusters are coincided, or, at least, related with given data points.

(4) Cluster information from either data-driven clusters, clinical clusters, or both.
- The most discriminative features will be displayed for each cluster to show what features devote most to determine the cluster. Also, by selecting a few features in the

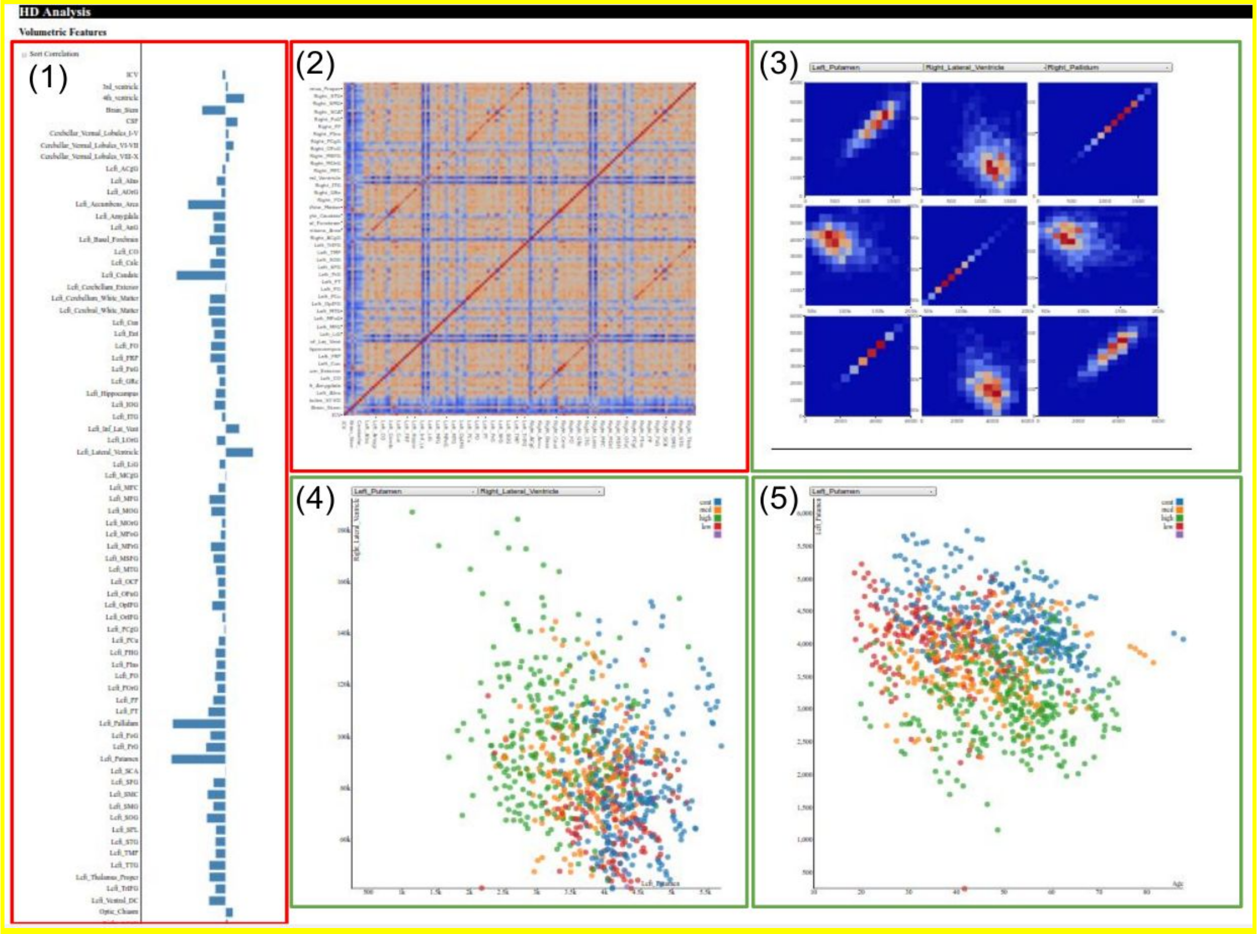
first list or the second correlation matrix, we will do the clustering again only with the selected features to see the effect of selected features

Project Update


① List of Features.
② Correlation of Features and Clinical Score (CAP Score).
③ Correlation Between Features.
Feature Lists.
Feature Lists.
④ List of Subsets of Features.
④ List of Selected Features.
⑤ Correlation of Selected Features and Clinical Scores.
⑥ Principal Components. Line Graph.
⑦ Scatter Plot.
• Color : Data Driven Clusters.
• Shape : Clinical Clusters.

* (4) Cluster information section in the proposal is removed since what we really want to see is the effect of brain subregions for the overall clustering results not a single cluster. However, it can be added if we can see some significant information or differences between clusters.
* (2) Bar graphs which visualizes the correlation of features and clinical scores is added to show what brain subregions are more correlated with clinical scores. The correlation coefficient is calculated by Pearson correlation method.
* (8) Lists of sets of selected features are added in the updated mock up. As previously described, we will pre-calculate the subsets of features based on the correlation of features and clinical scores and the correlation in between features or low rank approximation methods.
* (4) Lists of selected features are the list of brain subregions which are selected from the upper section.
* (6) The line graph of principal components will visualize predefined K principal components to show how it is composed of by brain subregions.

Final Visualization



In the proposed framework, we want to offer users to have an idea how brain subregion volume changes are related with HD risk burdens by projecting patients' data to a space with user selected subregions. We aim to let users explore patients' data by projecting data to a selected 2D space with brain subregions and patients' age. The proposed framework offers the information of the correlation between brain subregions and HD risk score(CAP score) and the correlation in between brain subregion volume changes to help them to identify interesting brain subregions they want to investigate. The proposed framework can be divided into five subplots.

(1) Correlation of volume subregion volumes and CAP scores

- It visualizes the correlation between brain subregion volumes and patients CAP scores. The correlation is calculated by Pearson correlation metric. The visualization framework offers sorting via the absolute value of correlation to show which features are more correlated with respect to the correlation value. The correlation value is ranged from -1 to 1. The magnitude of correlation value (absolute value) represents how much subregion volumes are correlated with CAP scores. Negative correlation means that a brain subregion volume is negatively correlated to CAP score, which means as CAP score increases, the volume decreases. Positive correlation is vice versa.

(2) Correlation between volume subregion volume changes

- The heatmap visualizes the correlation between brain features. It displays the correlation value with ordinal color map from blue to red for -1 to 1 correlation value range. The correlation in between features supplement Plot (1) which only shows the correlation between brain subregion volume changes and CAP scores. For example, when we the two most correlated brain subregions are left putamen and right putamen, which are counterparts of each side of our brains. The putamens are almost perfectly correlated which may be less interesting to plot data on.

(3) Joint histogram of data distribution

- The joint histogram displays the comprehensive views of data distribution with multiple projected feature spaces which users select. The comprehensive view gives users to identify what brain features that they might be interested to select for more detailed view.

(4) Detailed scatter plot of data distribution on two selected brain subregions

- The scatter plot displays the detailed view of data distribution on an user selected space with two brain subregions. Different risk groups are colored with categorical colors to show how different risk groups are distributed on a selected space

(5) Scatter plot of data on a selected brain subregion with age axis

- The scatter plot of data on age axis shows that how brain volume changes over ages. Since patients data are all in different time frame (different ages), there is no well defined time trend which can be displayed as line graph or histograms. The scatter plot on all data on age axis for a selected brain subregion visualizes what possible trend in volume changes and ages can exist and be revealed from patients data.

- Major difference from previous project proposal and project update is exclusion of data-driven clustering analysis. A few data driven analysis has been tried, such as K-Means and PCA, but it was hard to see any relationship to clinical risk groups and to explain what actual meaning is for a data driven analysis. Thus, we omitted the data driven analysis and added more exploratory data analysis by projecting patients data on different spaces which users can select.

Data Analysis

(1) Which brain subregions are related with disease risk burdens?

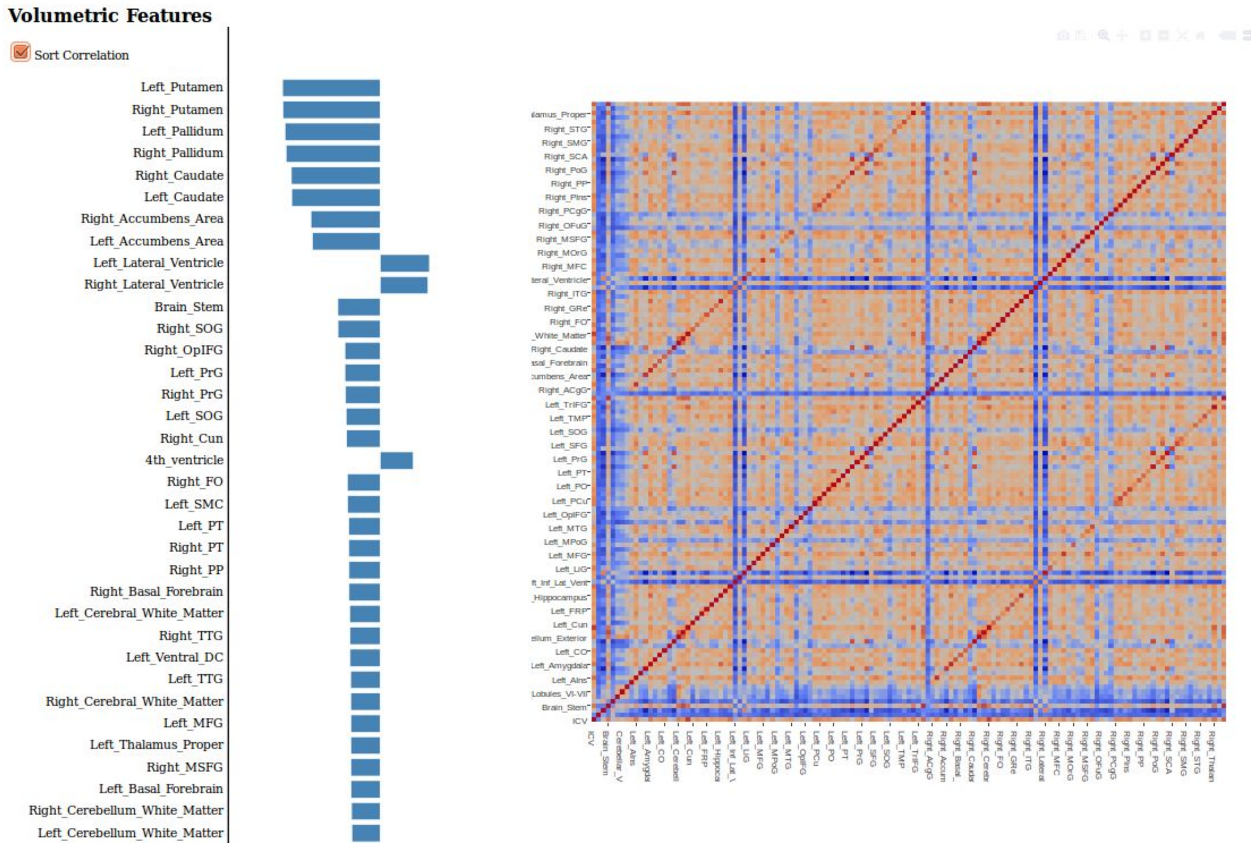


Figure DA1. List of brain subregion volume features and their correlation scores to CAP scores and correlation between brain subregions.

* When we sort the correlation value, we can observe that putamens, pallidums and caudates are most correlated regions to disease burden scores. This observation is coincided with the clinical findings, basal ganglia complex, which is composed of caudate nucleus, putamen, nucleus accumbens and pallidums.
* We can observe that the lateral ventricle areas are positively related with disease burden scores, which means as disease progresses, lateral ventricles become larger. Also, 4th ventricle area is also positively correlated.

(2) How brain subregion volumes are correlated?

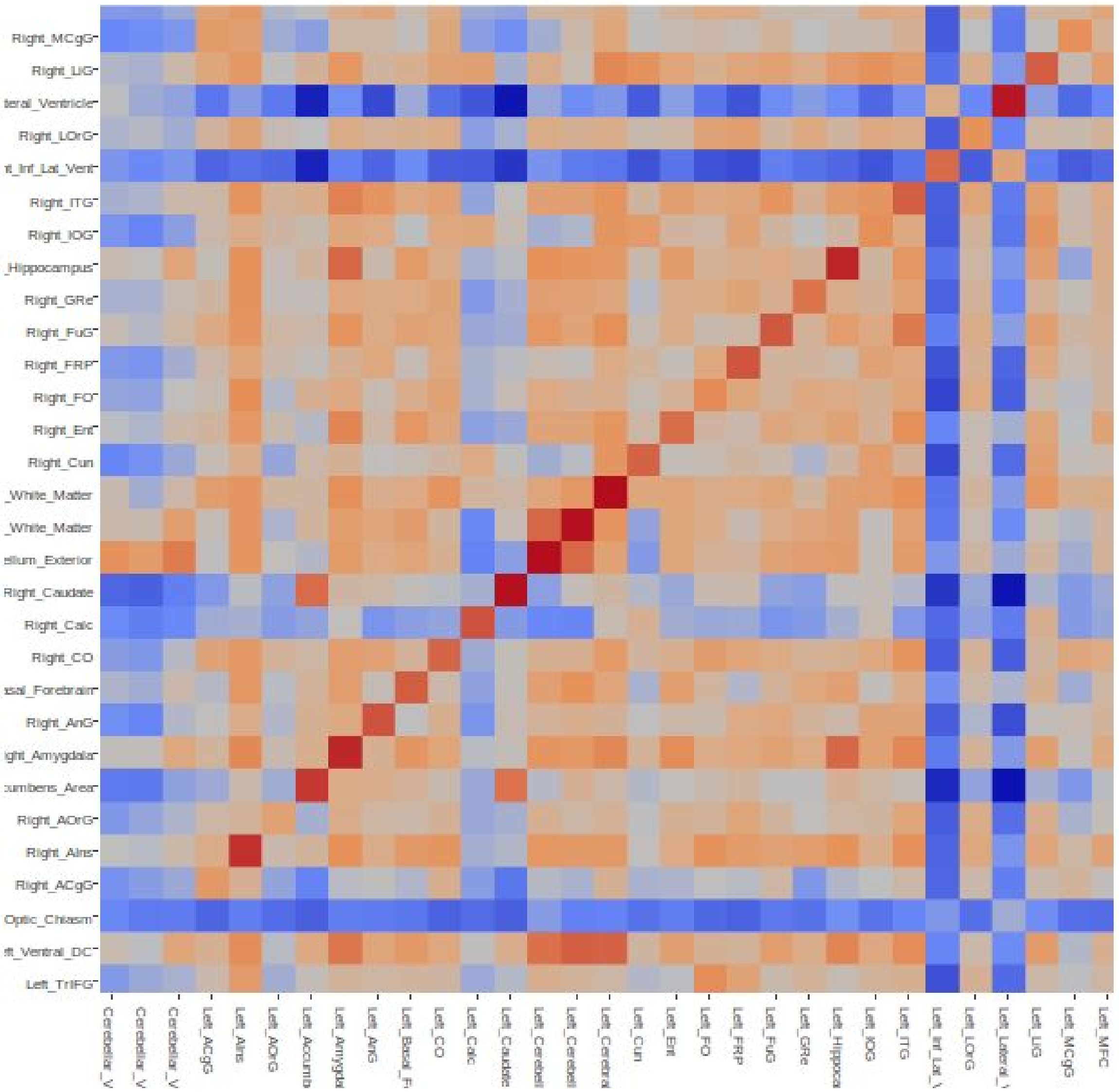


Figure DA2. Zoomed view of the heatmap of correlation between brain subregion volume changes.

* The volume subregion list in Figure DA1 also shows that the most of counterparts in each side of a brain have similar correlation values to each other. Figure DA2 shows that the correlation between volume changes in the counterparts of each brain subregion, such as, left and right putamens, left and right white matters are highly correlated. High correlation between counterparts of brain subregions may indicate that the development of counterparts may be identical, thus the correlation between volume changes are very high and the correlations to disease burden score are similar.

(3) There are some brain subregions which their volumes are increased while disease progresses. Also, they are negatively correlated with most of other brain subregions.

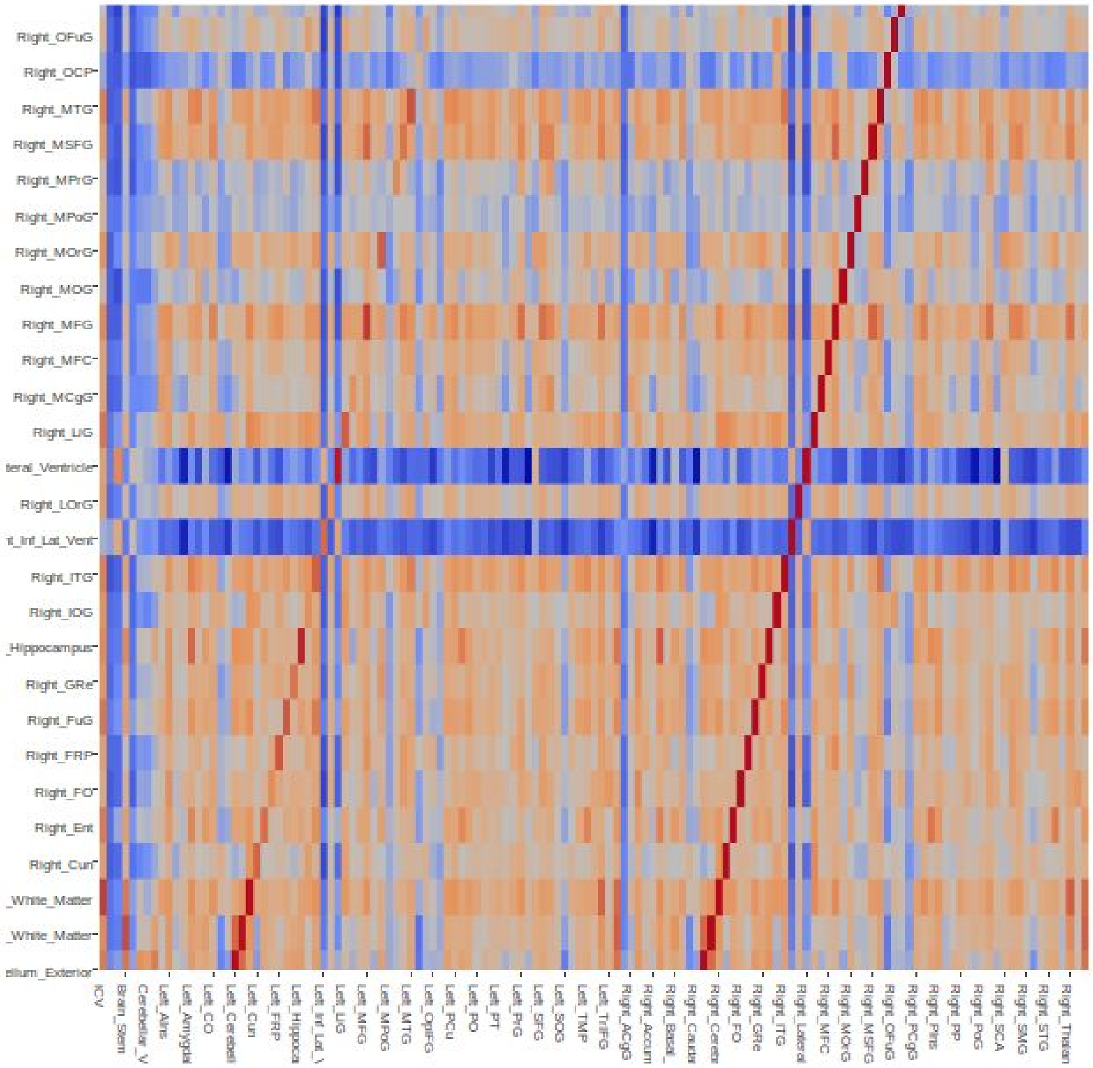
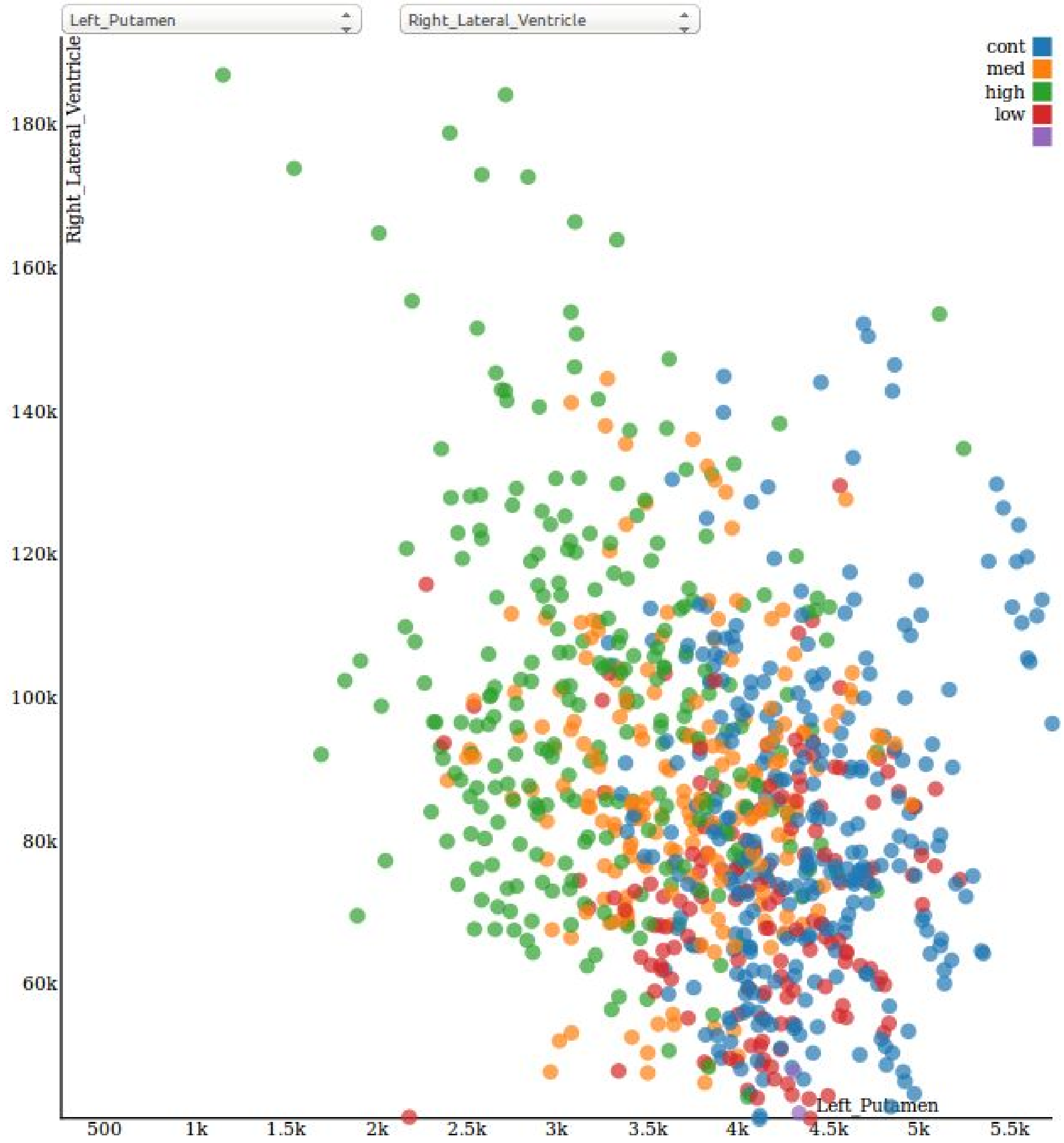


Figure DA3. Enlarged heatmap of correlation between features.

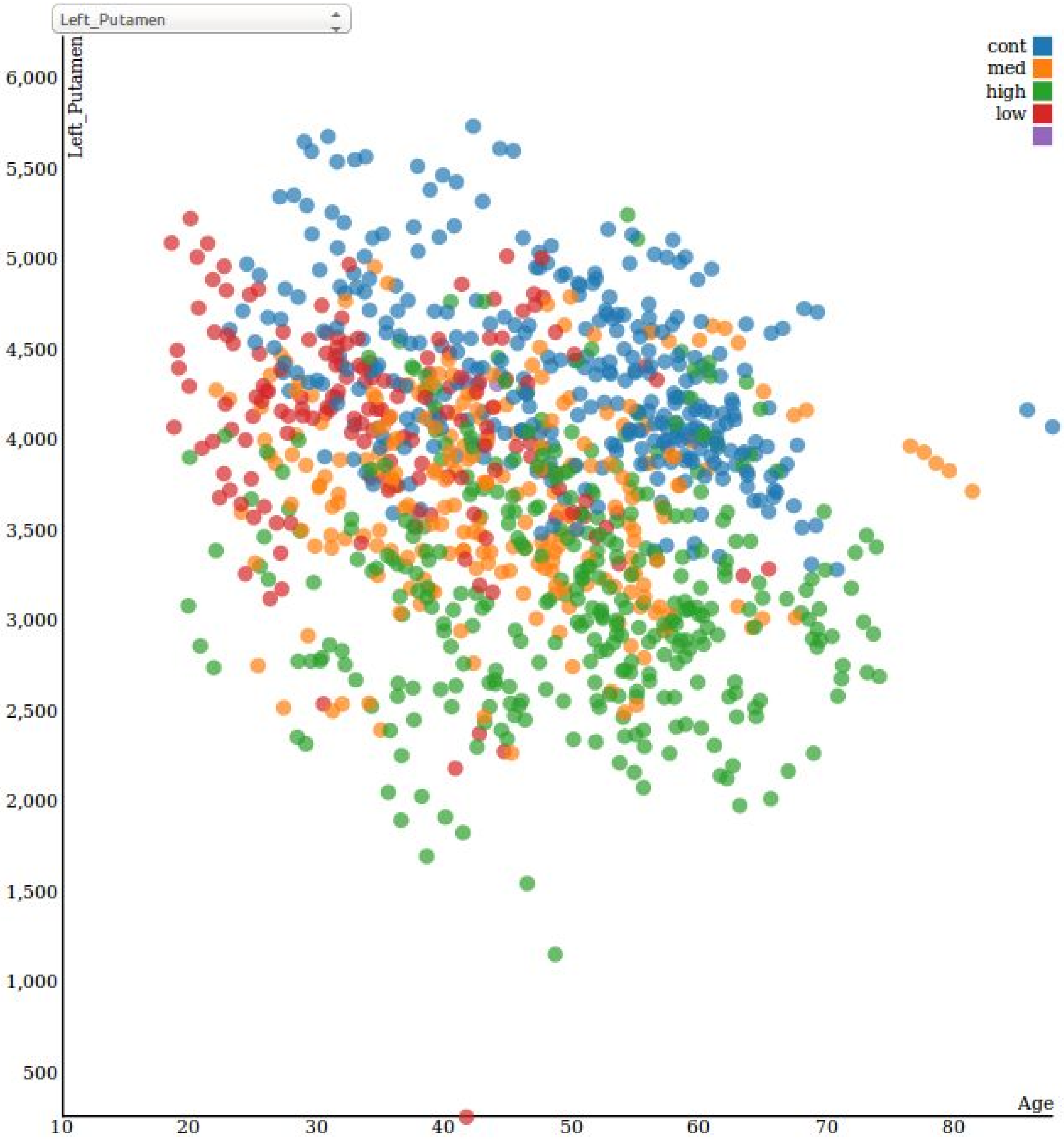
* Lateral ventricle areas are negatively correlated with most of other brain subregions. Based on the correlation list visualization, most of brain subregion volumes decrease while disease progresses. Therefore, the ventricle areas widens while disease progresses. It might be caused to fill empty space because of degeneration of other parts, or some other reasons which are worth to investigate.

(4) High Risk groups tend to have smaller left putamen volumes and larger right lateral ventricle volumes, but the variance of right lateral ventricle volumes is very large. Thus, although right ventricle volumes have big positive correlation value to CAP scores, it might not be statistically valid to assume that HD patients have larger right lateral ventricle volumes. However, it seems significant that HD patients have smaller putamen volumes than control groups.



The observation can be made to various other features like right putamen, left caudate, and right caudate which HD high risk group patients tend to have smaller volumes.

(5) High risk group patients' left putamen volume degenerates faster over ages than people in a control group. The scatter plot of data with age axis and a selected brain subregion shows that not only volumes of left putamen area is smaller for HD high risk group patients, it also degenerates faster over ages. Thus, it might indicate the possibility that the rate of volume degeneration can be a good indicator of HD along with volume itself.



Limitations and Future Works

We proposed the visualization framework to visualize the information of high dimensional volumetric subregion features and data projection to user selected spaces to reveal hidden information in clinical high dimensional data. The interaction between plots, such as the list of correlation to CAP score and the heatmap of correlation between brain subregions is not implemented which might cause the problem for users to understand the relationship between the visualization easily. The interactions can be added to 1) between correlation list and heat map, 2) between correlation list and joint histograms, 2) joint histograms and enlarged scatter plot with two selected brain subregions, 3) joint histograms and scatter plot of a selected brain subregion on age axis and 4) the enlarged scatter plot and the scatter plot of a selected brain subregion on age axis. In my limited knowledge, I think the interaction should be implemented in front propagation, which means that the interaction in the more general view should affect the detailed view, but the interaction on more detailed view does not affect the more general view. For example, if I interact with list of correlation, then the interaction should be applied to joint histograms or the enlarged scatter plot, not other way around. But this needs to be considered more thoroughly for the future directions.