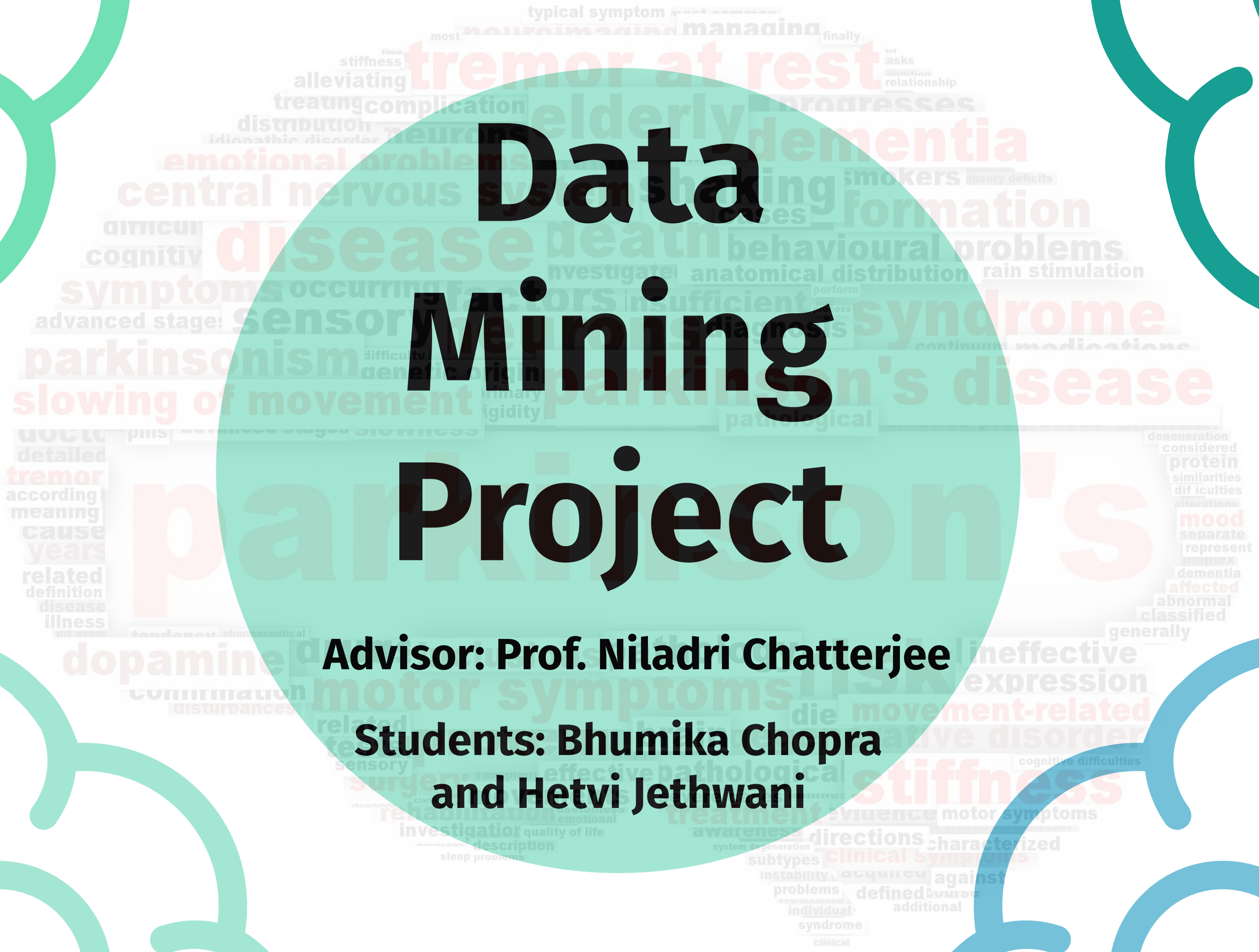
# Data Mining Project

**Advisor: Prof. Niladri Chatterjee**

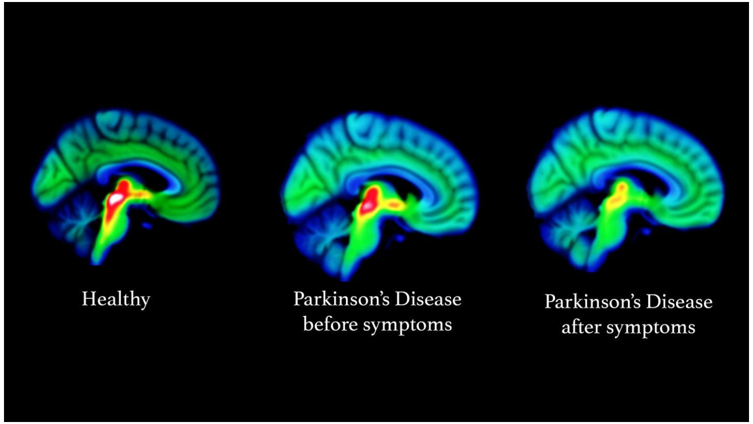
**Students: Bhumika Chopra and Hetvi Jethwani**



# Early-stage detection of Parkinson's using ML

# About Parkinson's Disease

- It is a progressive nervous system disorder which causes abnormal brain activity.
- Symptoms include tremors, shakiness, stiffness, difficulty in talking, slowed movement and coordination.
- The cause is unknown. In certain cases genetic mutations and exposure to certain toxins (environmental conditions) play a role in identification.
- Currently there are no blood or laboratory tests available to detect non-genetic cases of Parkinson's disease. Diagnosis is based on a person's medical history and neurological exam.


Healthy
Parkinson's Disease before symptoms
Parkinson's Disease after symptoms

# Project milestones

01 Literature Review

02 Modelling the problem

03 Data cleaning and preprocessing

04 **Exploratory analysis**

Use unsupervised learning algorithms to visualize the dataset

05 **Baseline models**

Apply existing algorithms on individual datasets, compare with SOTA

06 **Multimodal Learning**

Data fusion and integration

# Modelling the problem

- Classification Task-
  - Given a new sample corresponding to a person, classify if the person is having parkinson's or not
- Datasets-
  - Too many but too few samples!!
  - Combination of speech data, hand-drawings, pose estimation data, brain scan parameters and genetic data
- All existing work is done on these small specialized datasets
- Novelty-
  - We aim to apply multimodal learning on these "non-parallel" datasets to improve existing results
  - We will be using speech data [which has extracted features], image data [which has to be processed], and typing data [which has to be processed]

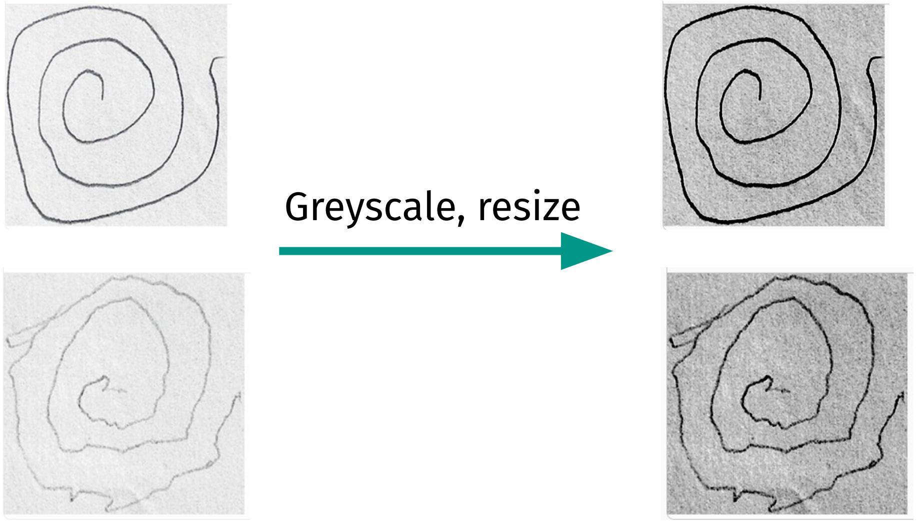
# Illustrating Data Preprocessing: speech data

- Speech data:
  - Columns in the table contain subject number, subject age, subject gender, time interval from baseline recruitment date, motor (Unified Parkinson's Disease Rating Scale) UPDRS, total UPDRS, and 16 biomedical voice measures
  - Voice measures include- jitter, shimmer, noise-to-harmonics ratio, etc.

- We have feature vectors we just normalize/standardize them.

# Illustrating Data Preprocessing: image data

- Image data:
    - Raw data is available.
    - We will use CNNs to train on image data
    - We will observe the features extracted by different layers of the network



Healthy:

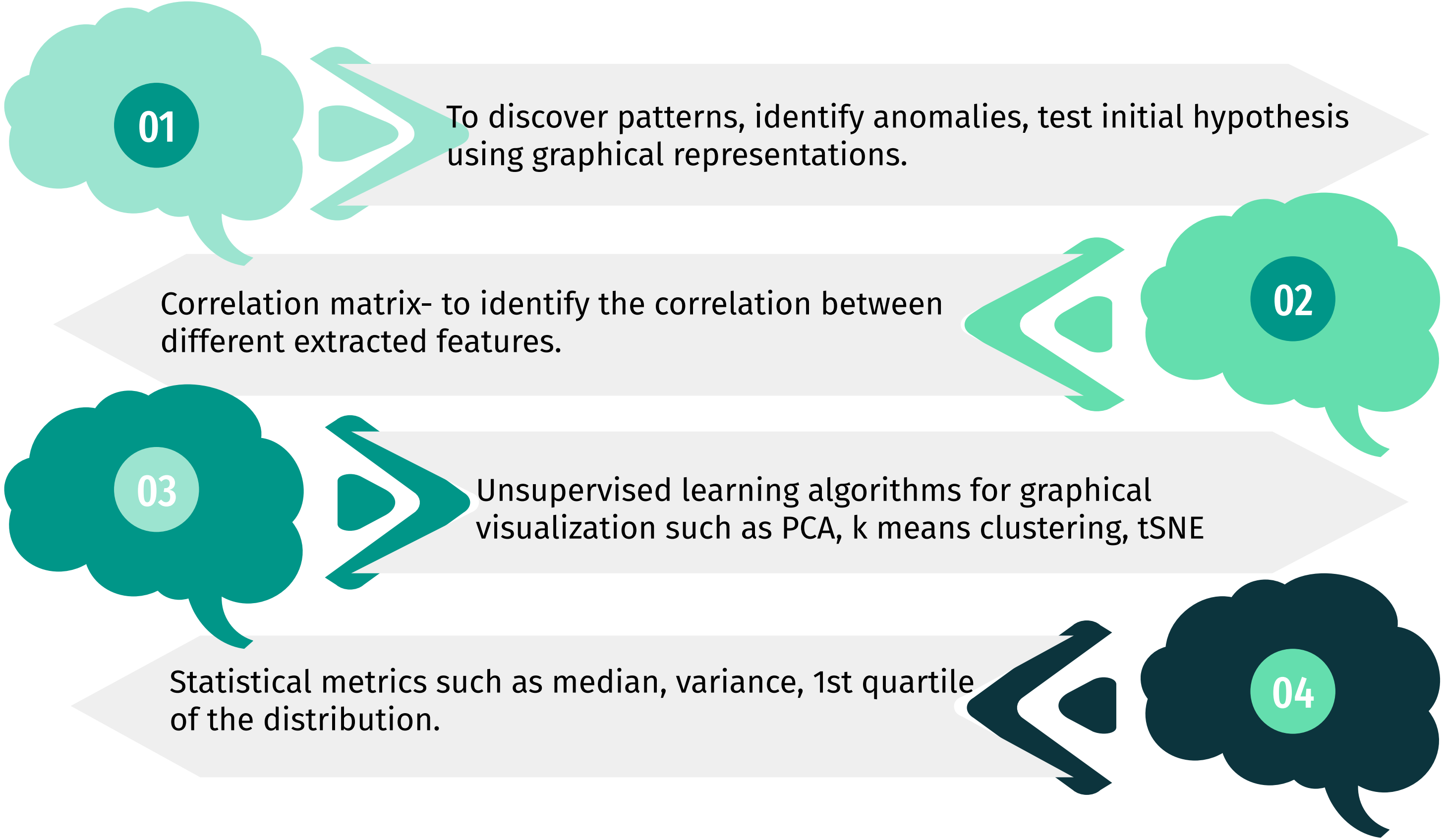
PD patient:

Greyscale, resize

# Illustrating Data Preprocessing: Keystroke data

- Keystroke data-
  - Includes timing information from typing activity as the participants used various Windows applications such as email, word processing, web searches, etc
  - Dataset contains
    - User data - gender, Parkinson’s affected or not, tremors, diagnosis year, Levadopa presence, MAOB (enzymes) presence, etc
    - Tappy data - key press and release time stamps data
  - Remove keystrokes with negative hold/latency times (error) and with very long hold/latency times (more likely to be deliberate).
  - Calculate-
    - Mean, SD, skew, kurtosis for Left hold time, right hold time, LL/LR/RL/RR - transition time
    - Mean difference between L/R hold time, LR/RL latency, LL/RR latency
  - Concatenate calculated values to get per-user feature vectors

# Exploratory data analysis



01

To discover patterns, identify anomalies, test initial hypothesis using graphical representations.

02

Correlation matrix- to identify the correlation between different extracted features.

03

Unsupervised learning algorithms for graphical visualization such as PCA, k means clustering, tSNE

04

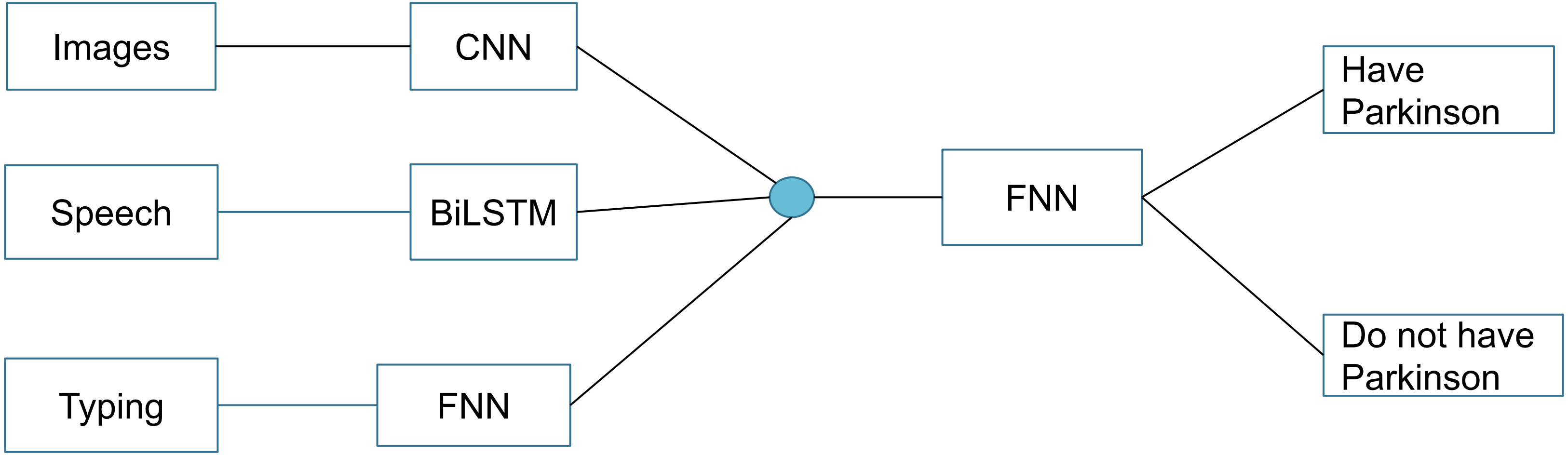
Statistical metrics such as median, variance, 1st quartile of the distribution.

# State of the art Models

- On UCI-Oxford Dataset-
  - XGBoost (eXtreme Gradient Boosting) Classifier
- On Handwritten Spirals Dataset-
  - CNN with dropout regularization

# Multimodal Learning

- Models up till now have focused solely on image or video or audio data
- We plan to combine data from different modes - speech (audio), keystrokes, hand-drawing (images) to generate better models



**Challenge?**
Combining the extracted features, assigning weights (importance) to different modes.

# Multimodal Learning

- We plan to use transfer learning to resolve this
- First we apply PCA or an unsupervised feature extraction technique to get the vectors from image and speech data to the same representation space
- Then we train a ML model.
- We will train the model in 2 stages-
  - First train on image data and then use this trained model to learn speech data.
  - First train on speech data and then use this trained model to learn image data.

# Attempt to relate these 3 different approaches to training and testing

1

Simultaneous training and testing on speech and image data

2

Training on speech data and testing on image data

3

Training on image data and testing on speech data