

SAMSI UG WORKSHOP  
NC STATE UNIVERSITY

# Characterizing and Classifying Alzheimer's Diagnosis

Frederick Donahey, Khoa Huynh, Linda Pan,  
Xin Tan, Lynn Zhu



# Overview

1. Background
2. Objectives
3. Exploratory Data Analysis
4. Classification Methods and Results
5. Characterization of PET Scans
6. Conclusions

# Background

## Alzheimer's disease - Neurodegenerative disease

### Prevalence <sup>[1]</sup>

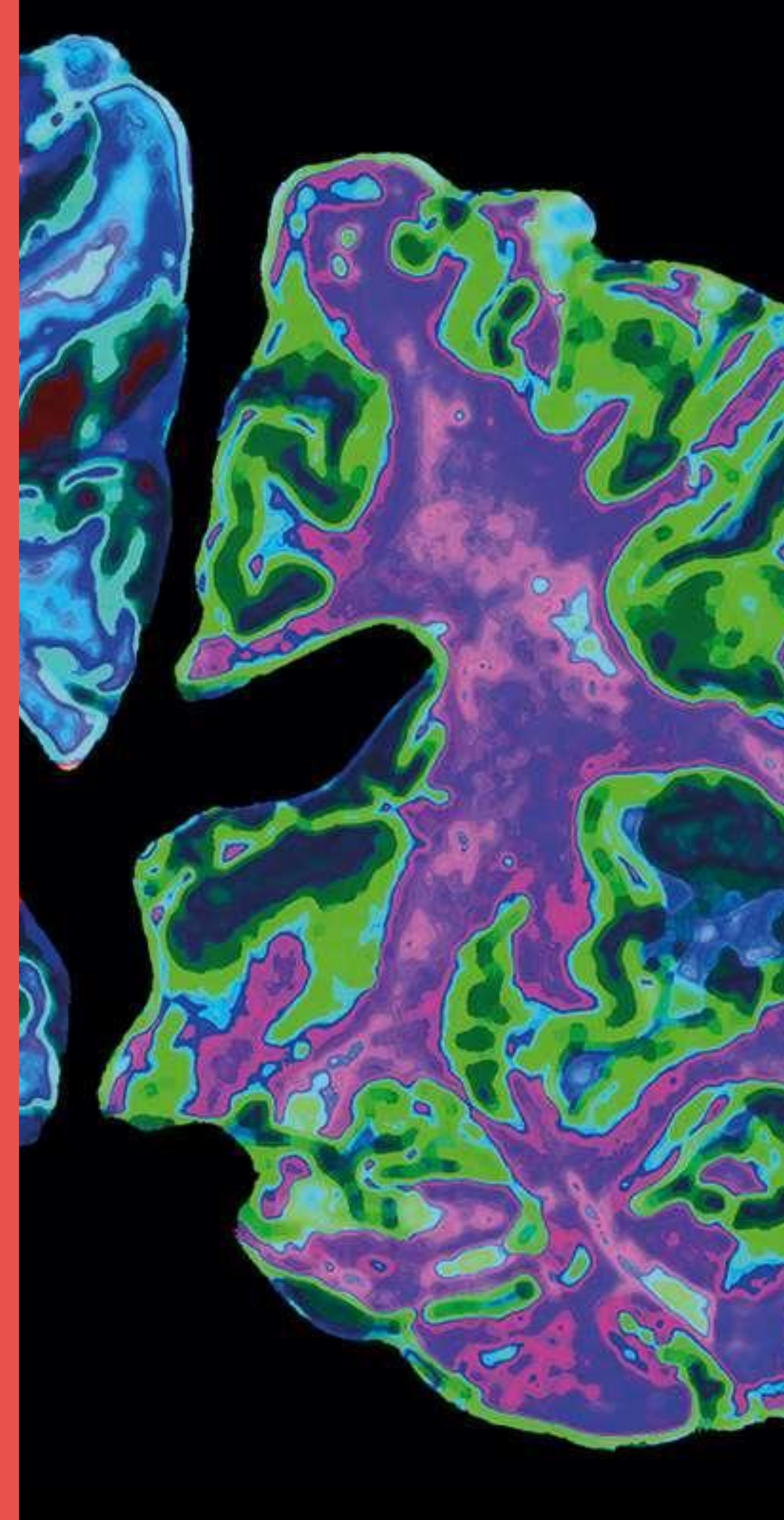
- 6th leading cause of death in the US
- Affects more than 5.8 millions Americans.

### Pathology

- Extracellular beta amyloid plaque + intracellular neurofibrillary tangles
- Neuronal damage and brain region death

### Treatment

- Currently there is no cure for the disease and no way to reverse its progress

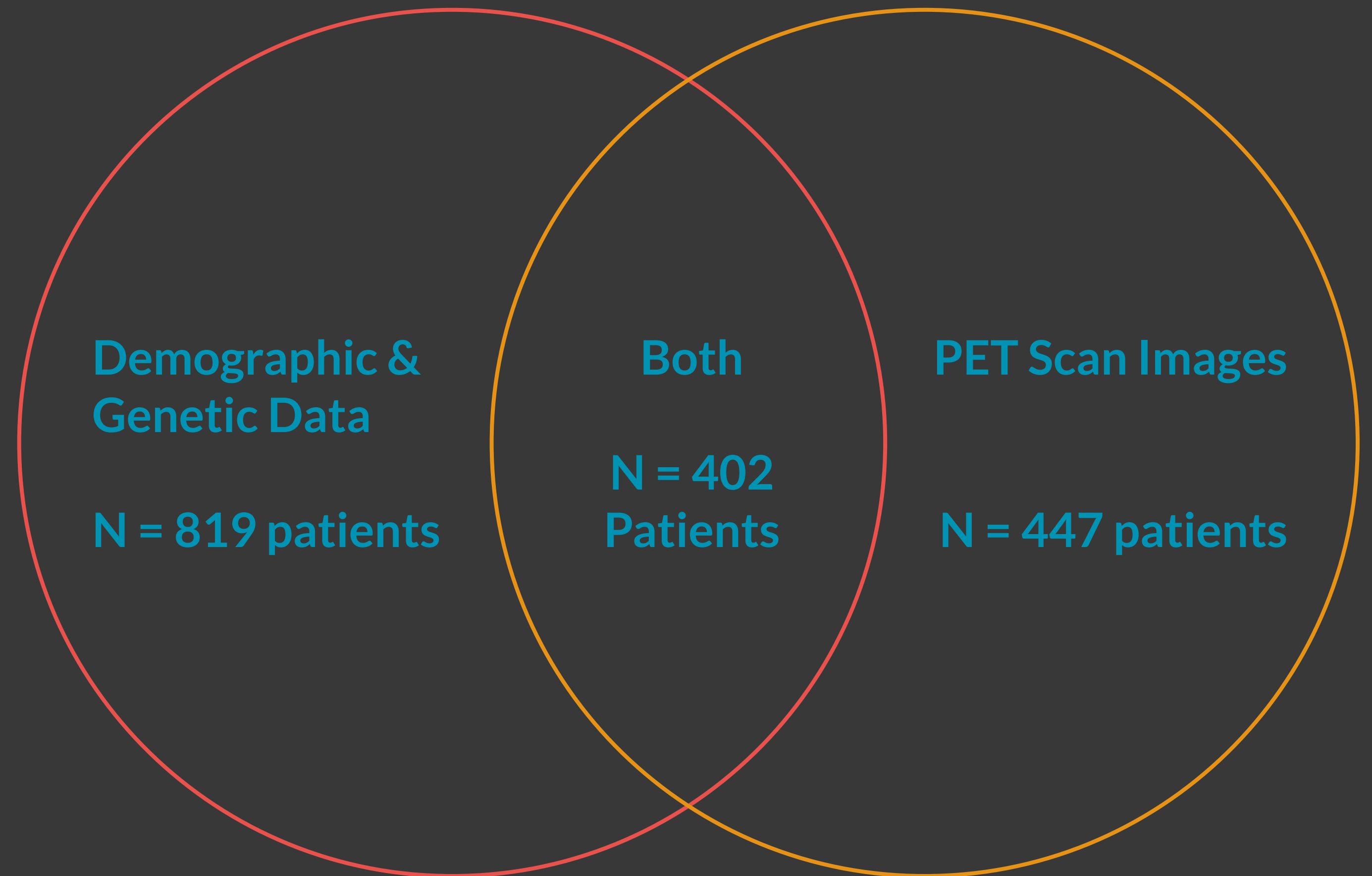




# Our Data

## Demographic and Genetic Data:

- Age
- Sex
- **MMSE Score**
- **APOE4**
- Site
- Ethnicity
- Race
- Marriage Status



Data Source: ADNI (<http://adni.loni.usc.edu/>)



# Diagnosis of Alzheimer's disease

## Mental Status Test - Mini-Mental State Examination (MMSE)

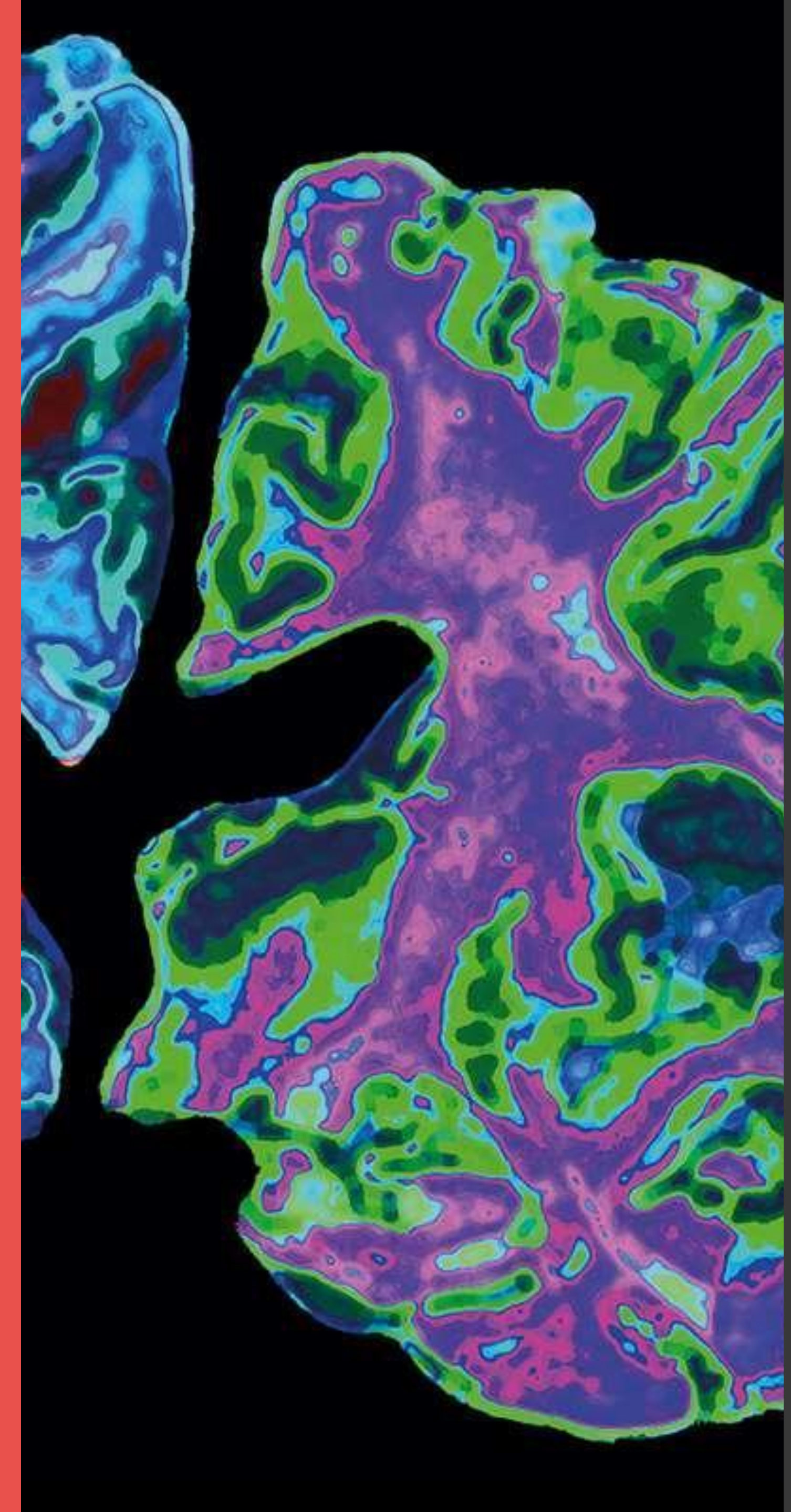
Maximum score of 30

- 20-24 = Mild dementia
- 13-20 = Moderate dementia
- <12 = Severe dementia

Alzheimer's patients decrease on average 2-4 points each year

## Brain Imaging - Positron Emission Tomography (PET) Scan

- Capture the image of the activity (metabolic level) of brain.
- Used to rule out other conditions such as brain tumor, Louis body dementia and Parkinson's disease with Louis body





# Genetic Risk Factor: ApoE4

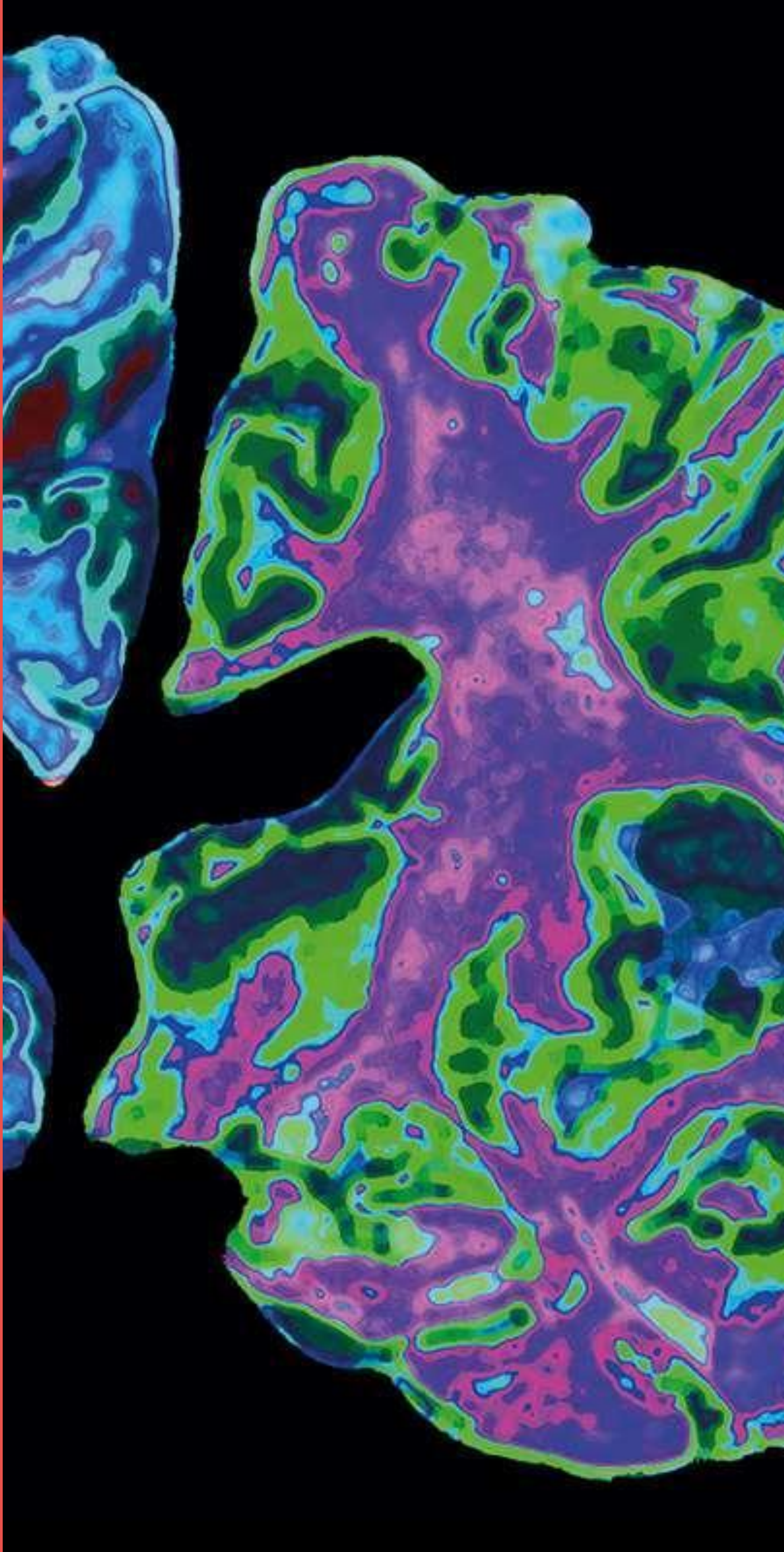
There exist **three alleles** for the ApoE gene:

- ApoE2
- ApoE3
- ApoE4

The presence of the ApoE4 allele **increases** the risk of developing Alzheimer’s disease

[2]

Genotype	E2/E2	E2/E3	E2/E4	E3/E3	E3/E4	E4/E4
Disease Risk	40% less likely	40% less likely	2.6 times more likely	Average risk	3.2 times more likely	14.9 times more likely





# Objectives and Methods

## 1. Characterize diagnoses

- Healthy
  - Mild cognitive impairment (MCI)
  - Alzheimer's disease (AD)
- in PET scan images

### Methods:

- PCA
- Tensor-on-scalar Regression

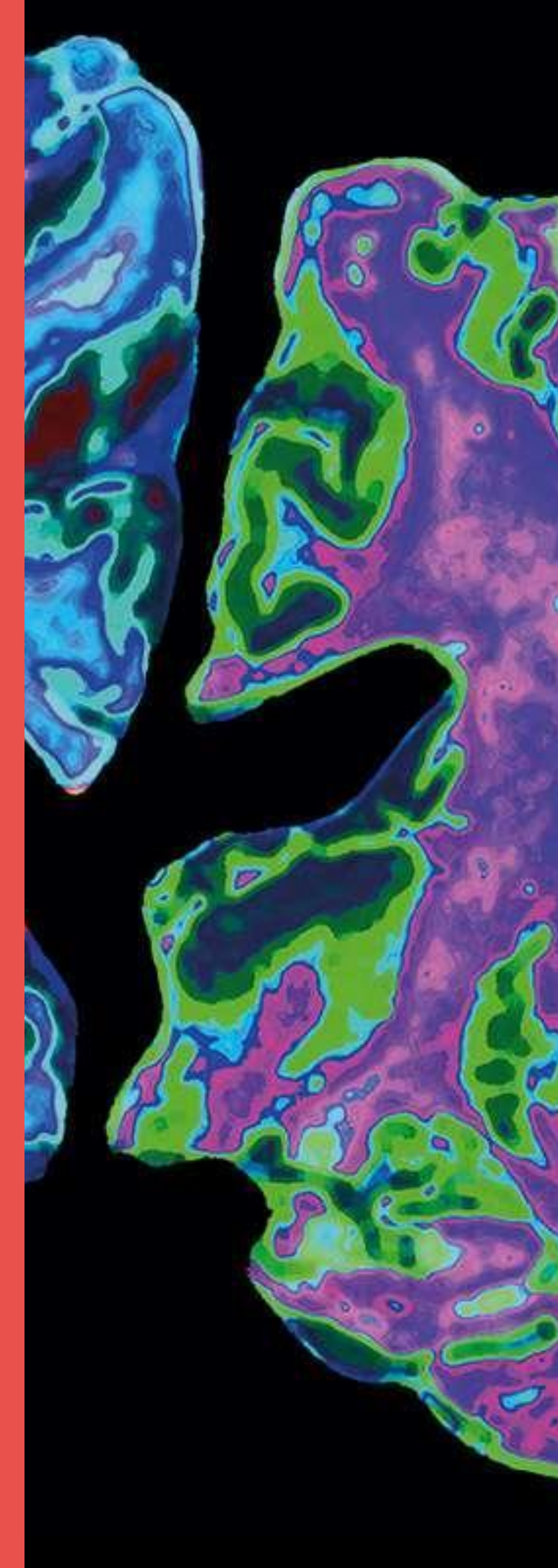
## 2. Classify patients

Patients into the **three diagnoses** based on:

- demographic data
- genetic characteristics

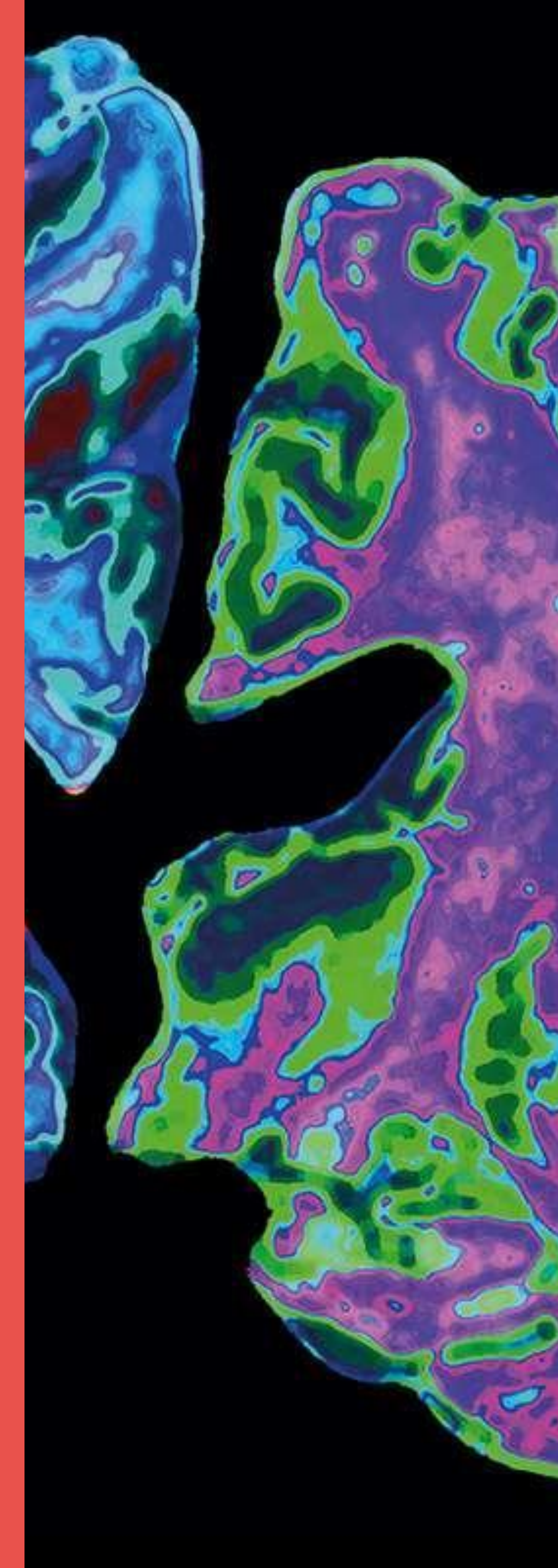
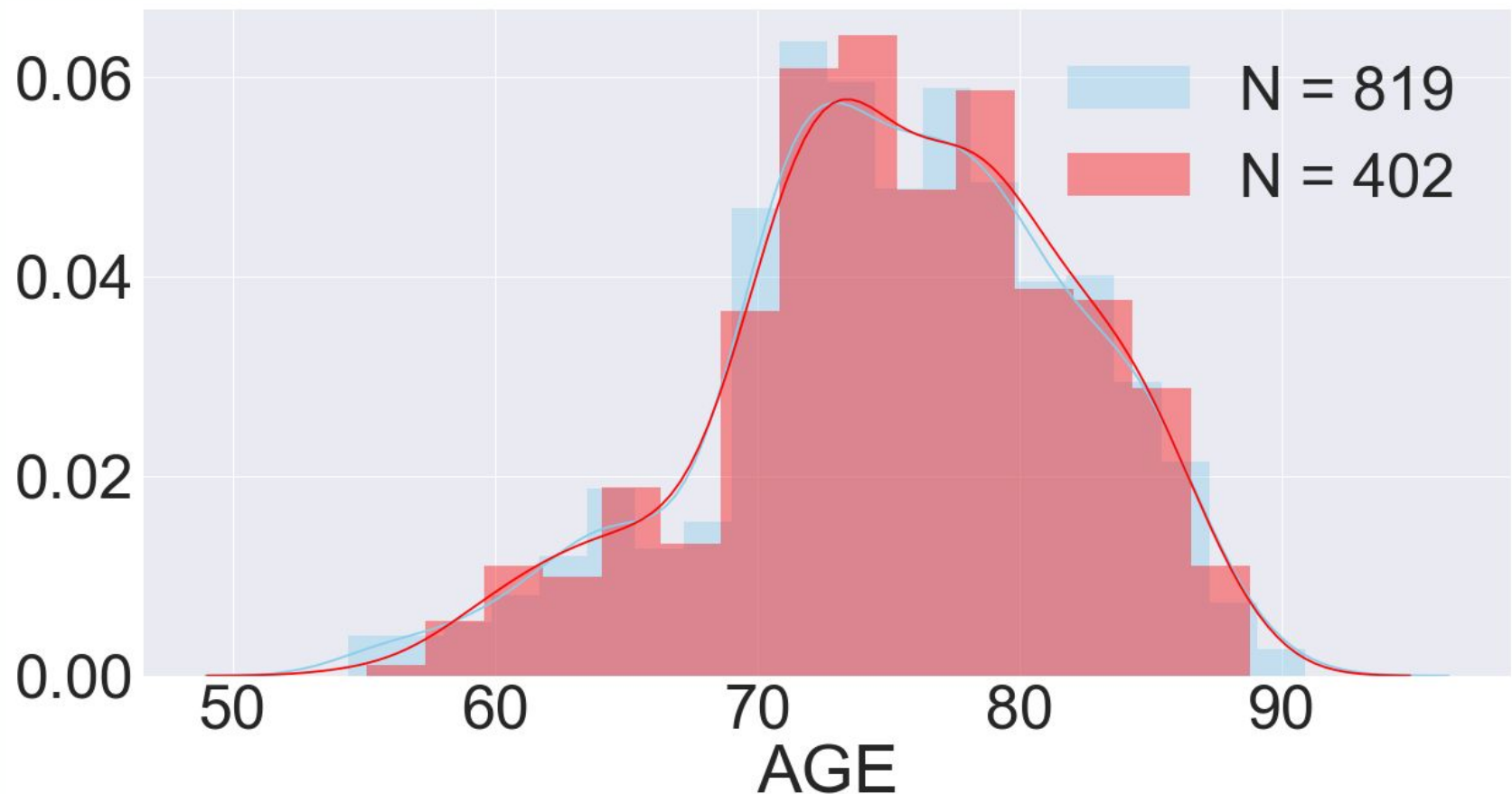
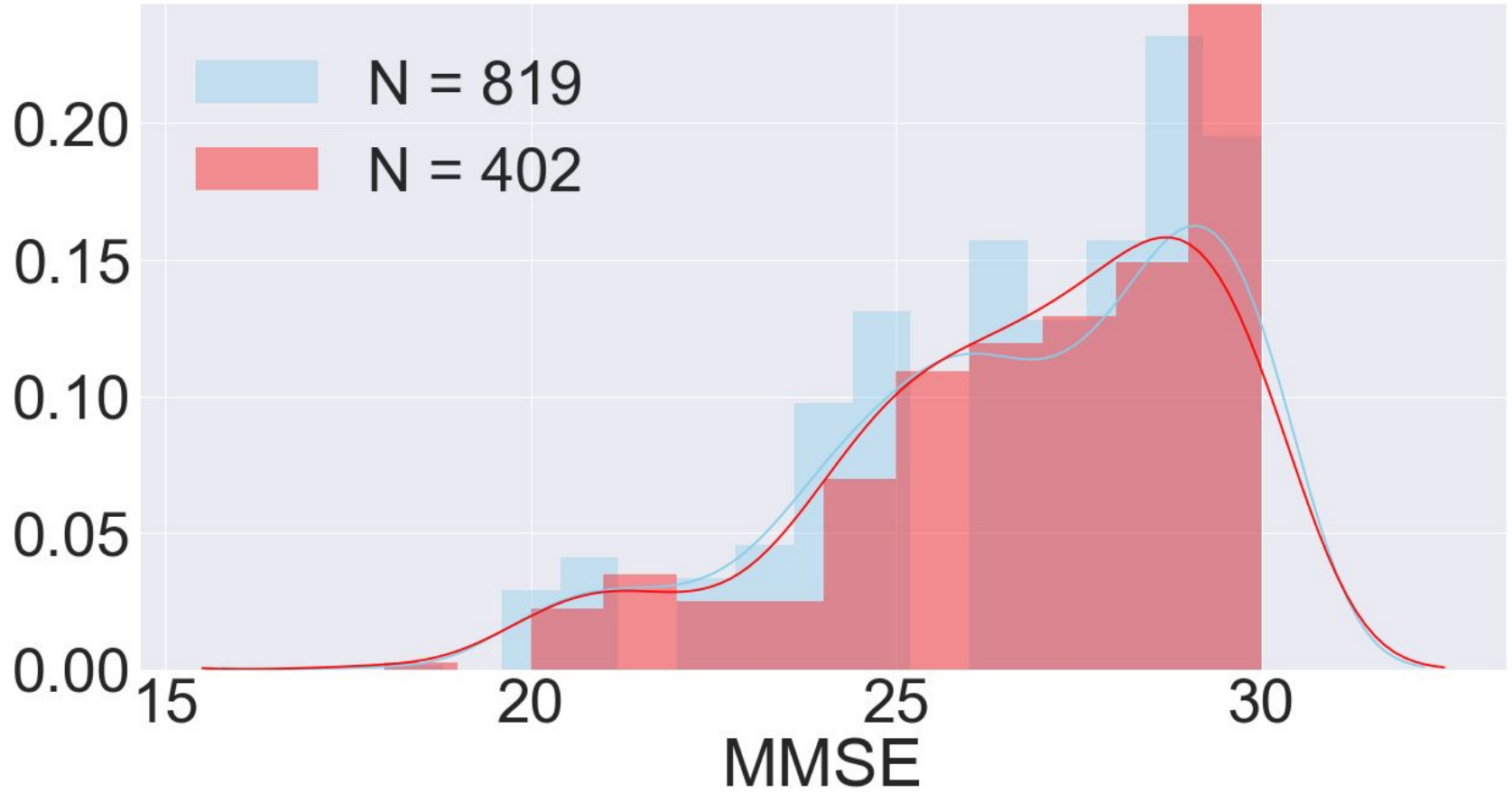
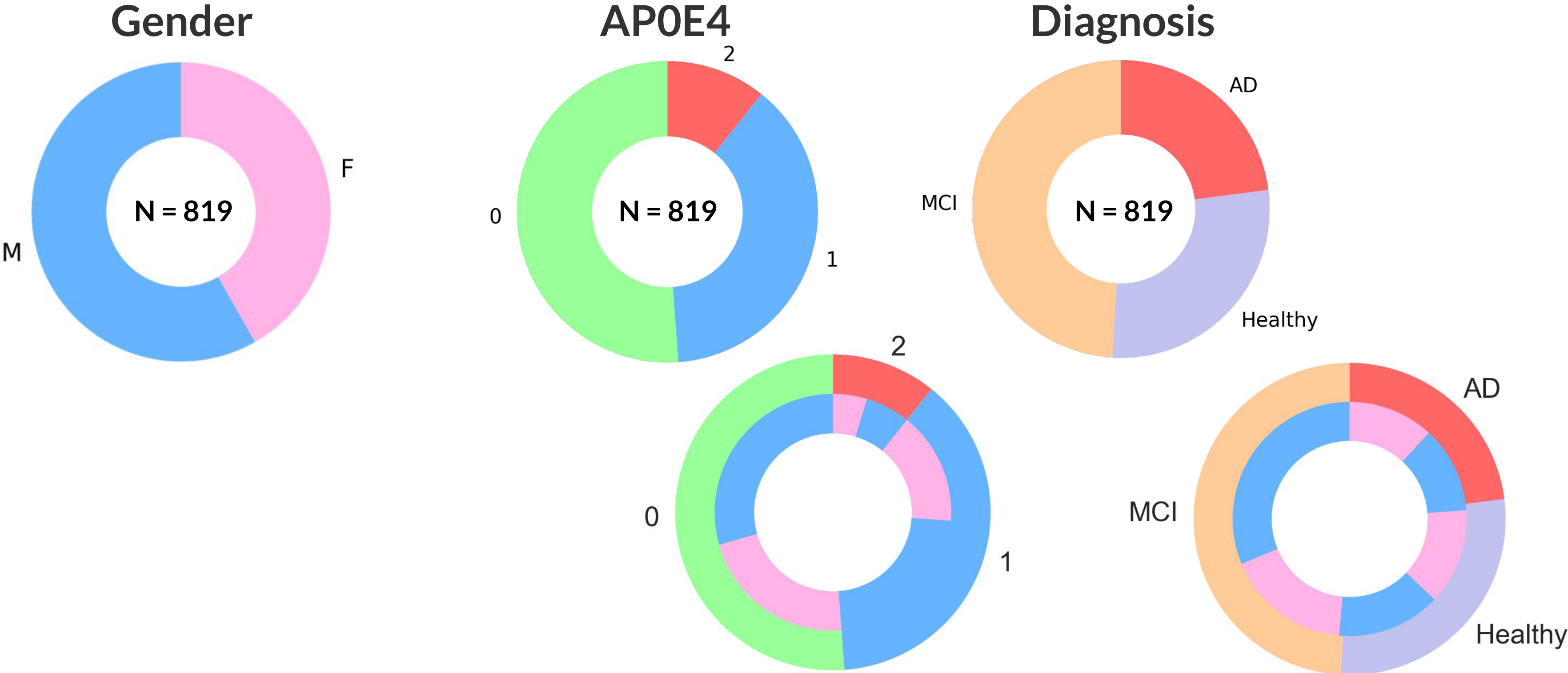
### Methods:

- Decision Tree (DT)
- Random Forest (RF)
- Neural Network (NN)
- Gradient Boosting Machine (GBM)





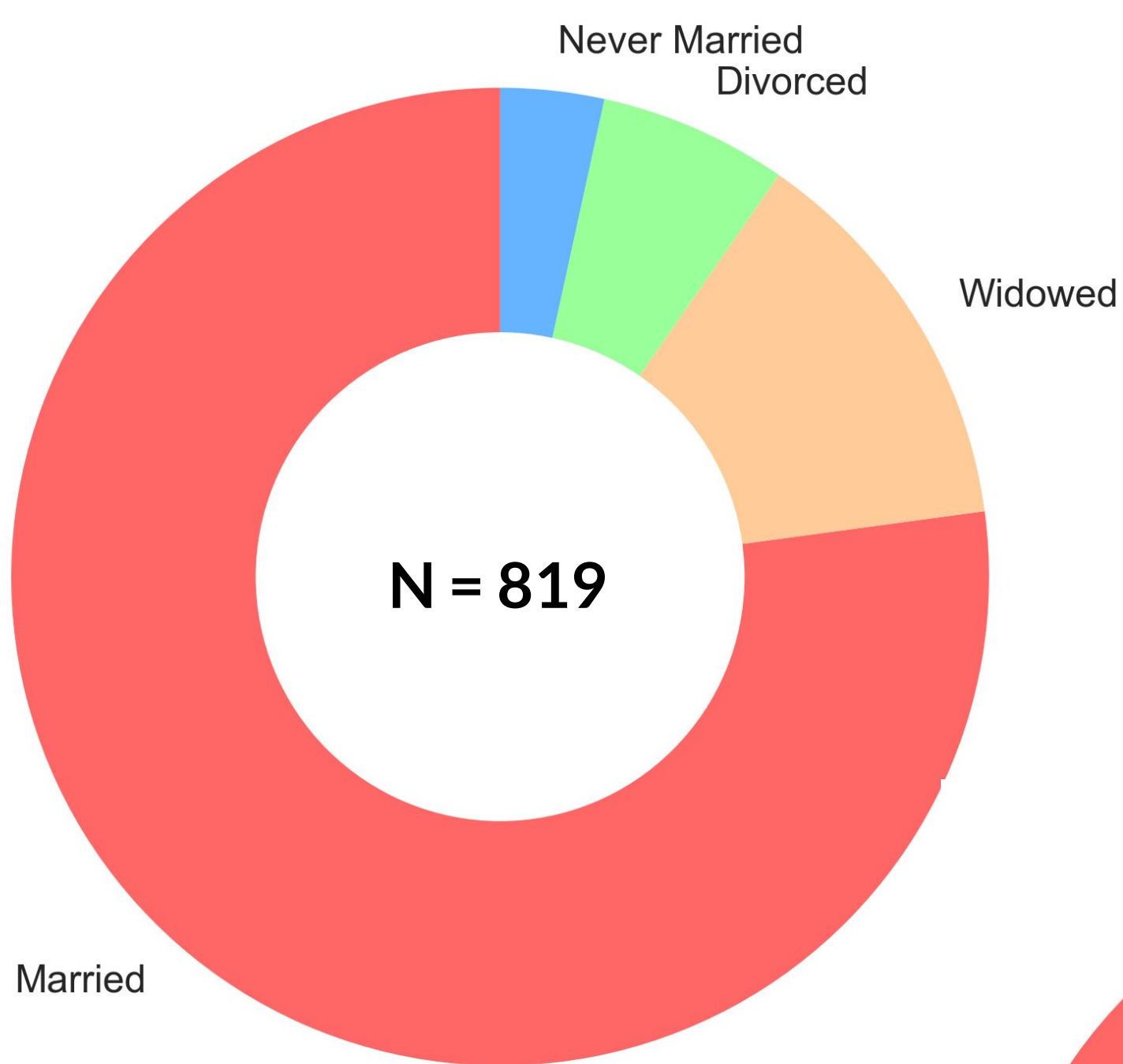
# Exploratory Data Analysis



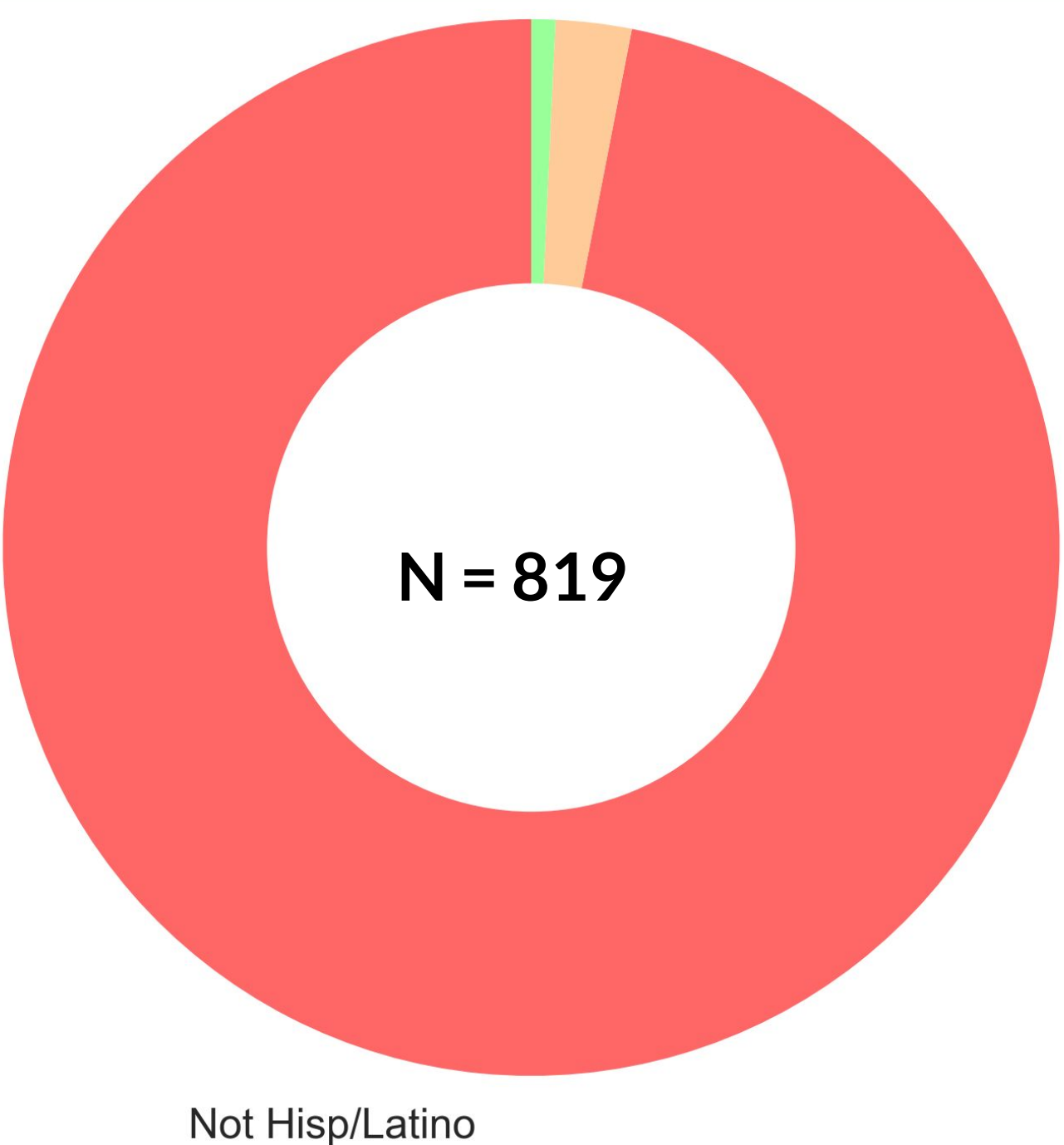


# Exploratory Data Analysis

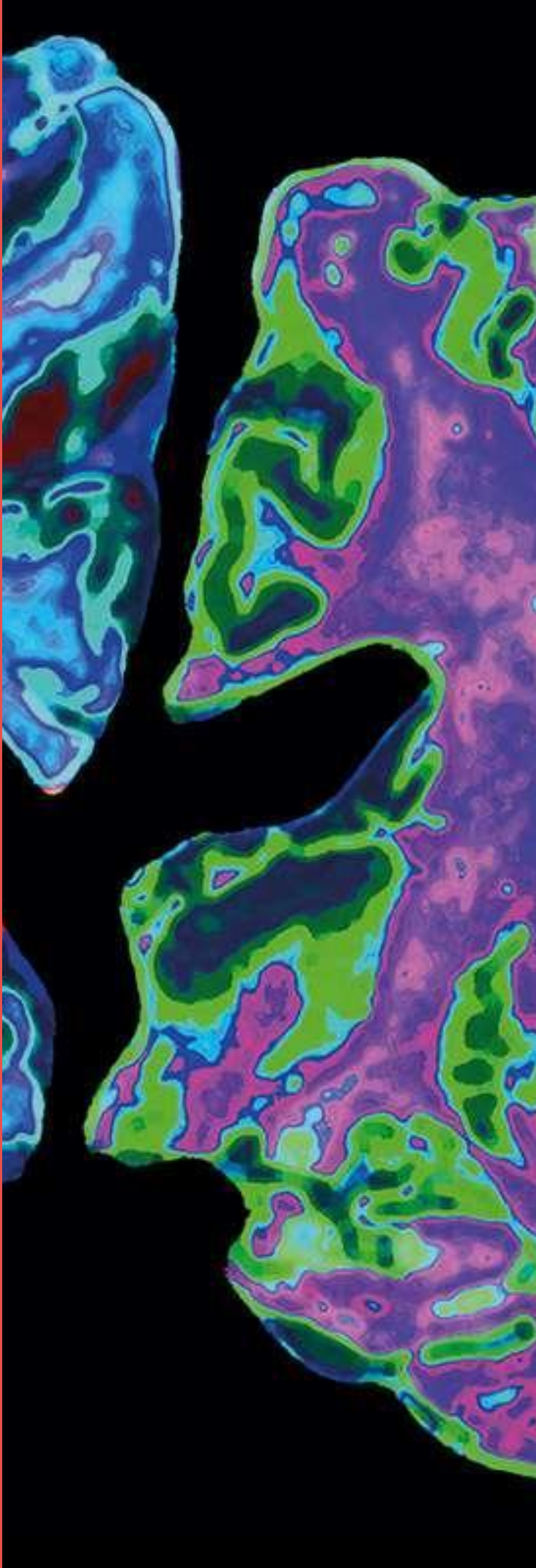
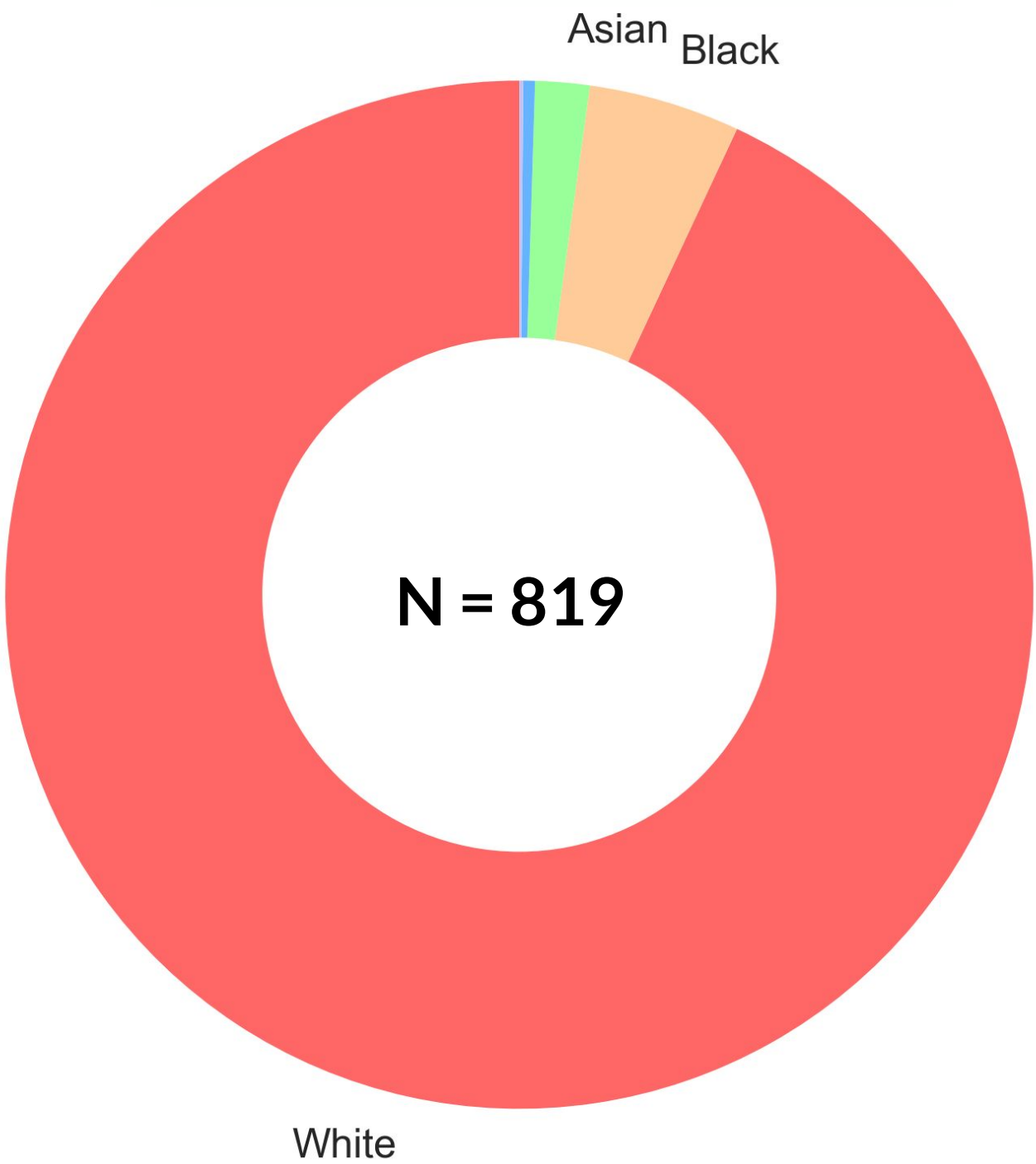
Married Status



Ethnicity



Race





# Exploratory Data Analysis

Healthy



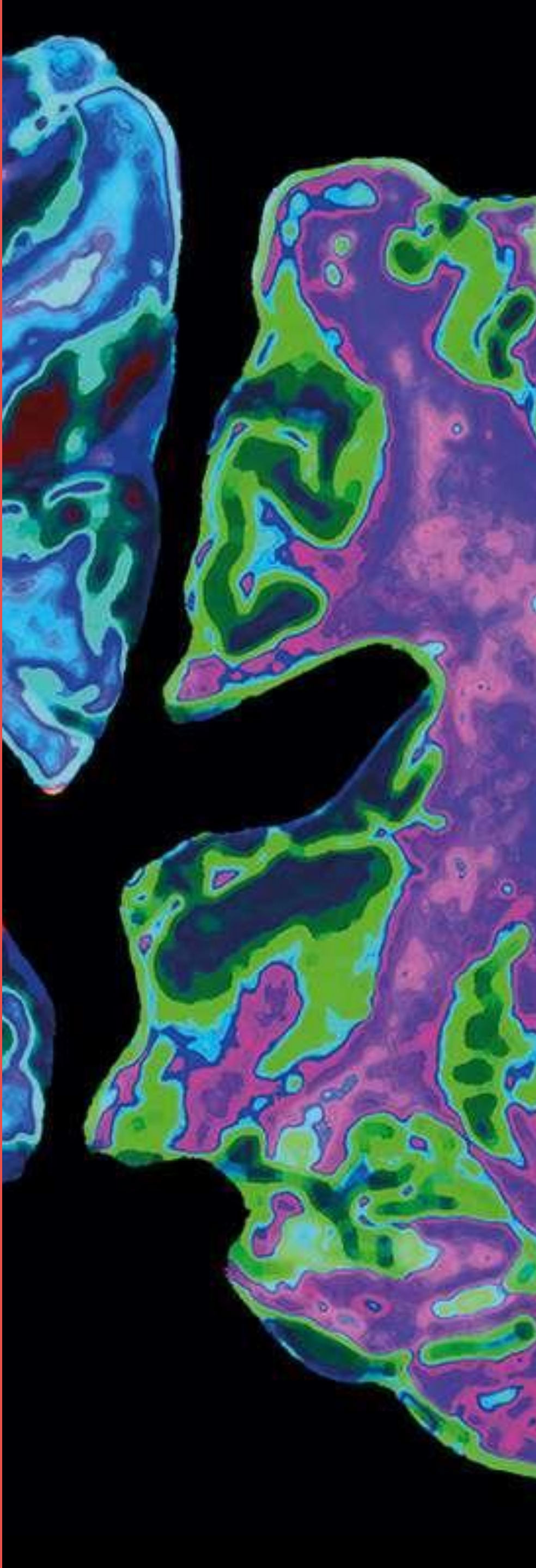
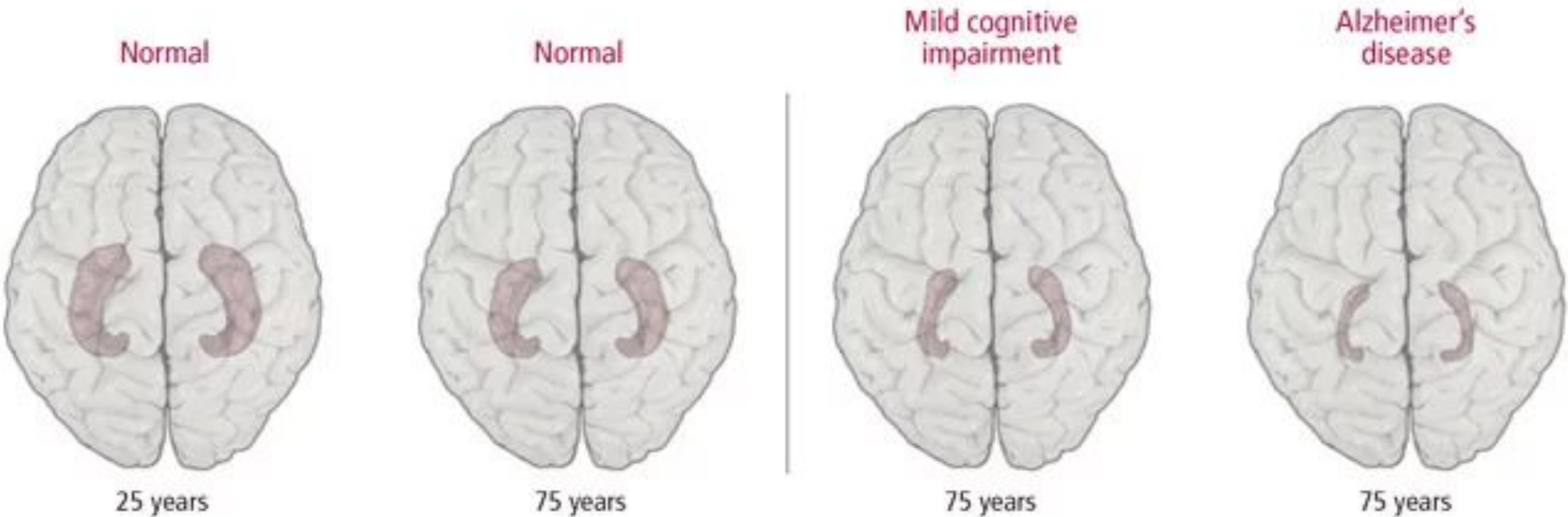
MCI



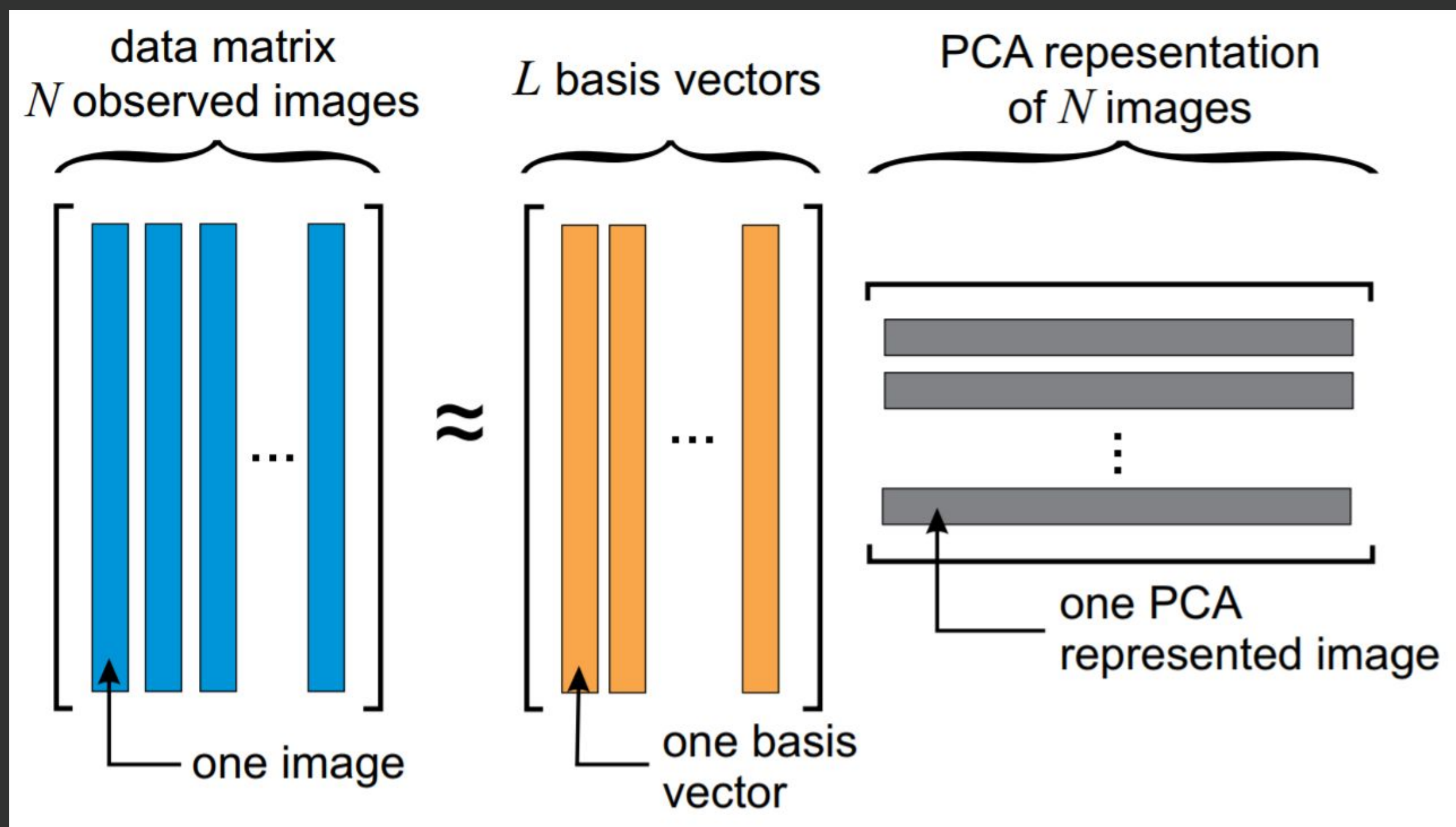
AD



Figure 6 The shrinking hippocampus







# Principal Component Analysis



# Classification Methods (Supervised)

## Decision Tree

- Required little of work for data preparation
- Easy to interpret
- Base learner for RF and GBM

## Neural Network

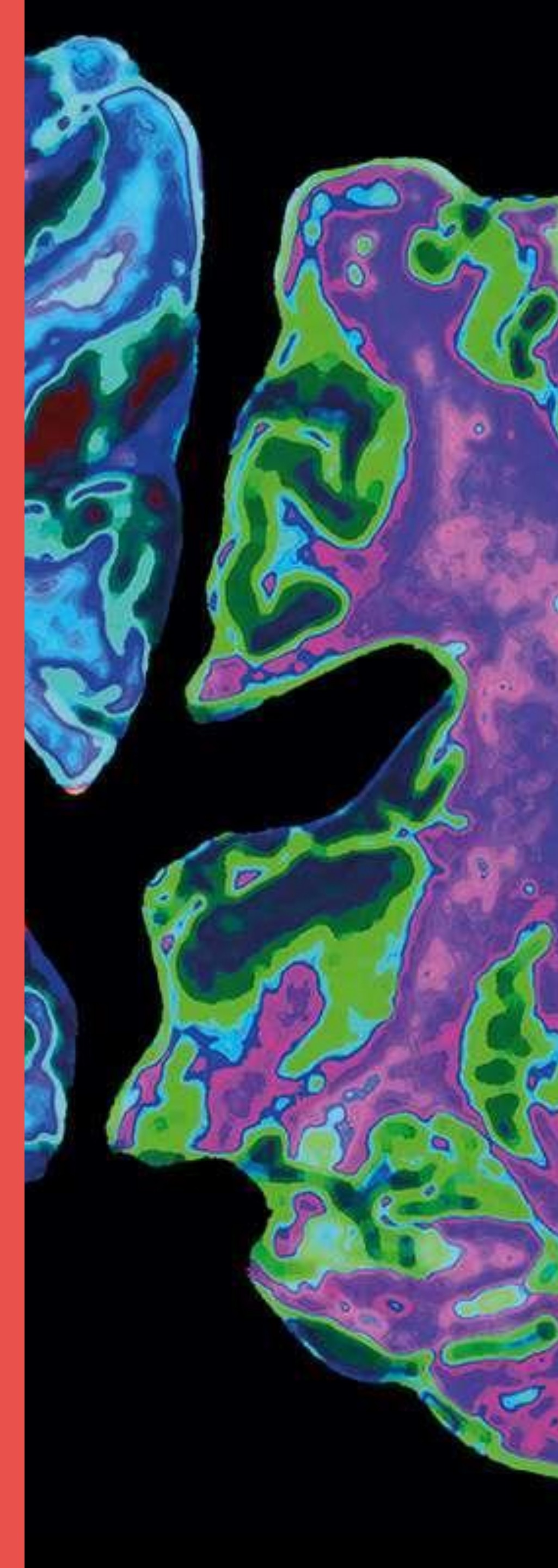
- Neural Network can used for classification problem
- Simulate the parameter and choose the best model base on MSE and accuracy

## Random Forest

- Random Forest can used for classification problem
- Base learner as Decision Tree
- Low bias and high variance

## Gradient Boosting Machine

- Gradient Boosting Machine can used for classification problem
- Base learner as Decision Tree
- High bias and low variance



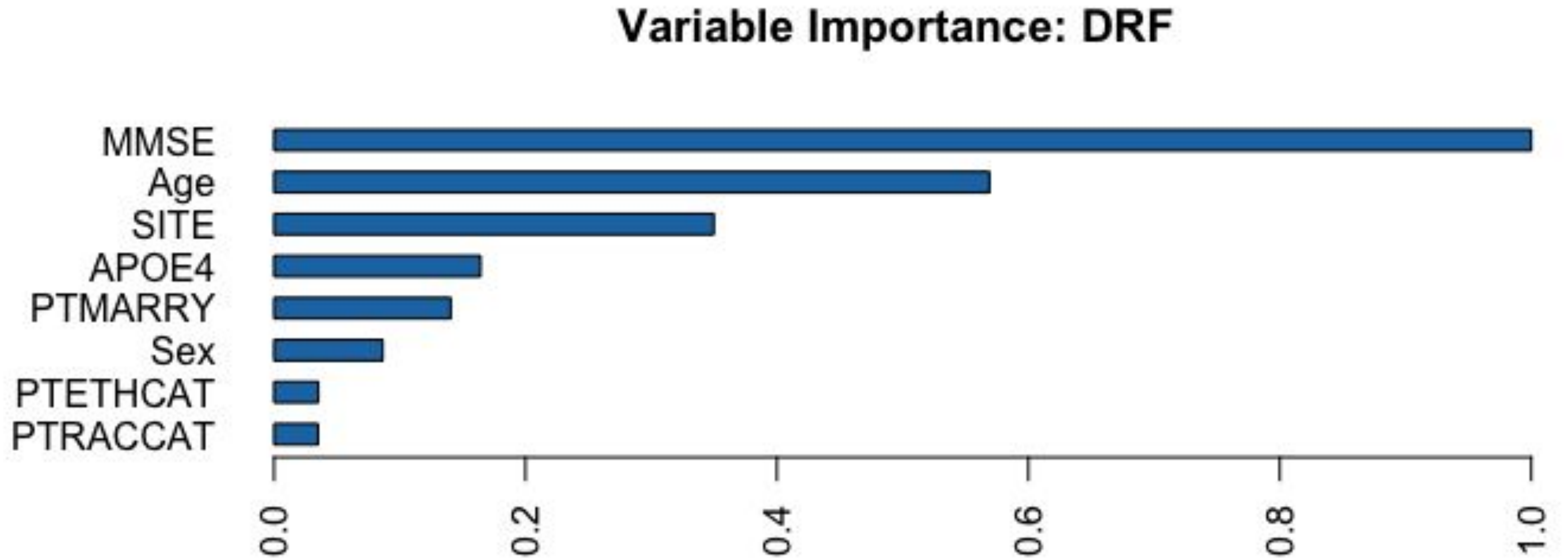


# Neural Network Classification Result

Actual	Predicted					
	AD	CN	LMCI	Error	Rate	
	AD	14	0	10	0.4167	= 10 / 24
	CN	0	15	5	0.2500	= 5 / 20
	LMCI	7	4	29	0.2750	= 11 / 40
Totals	21	19	44	0.3095	= 26	/ 84

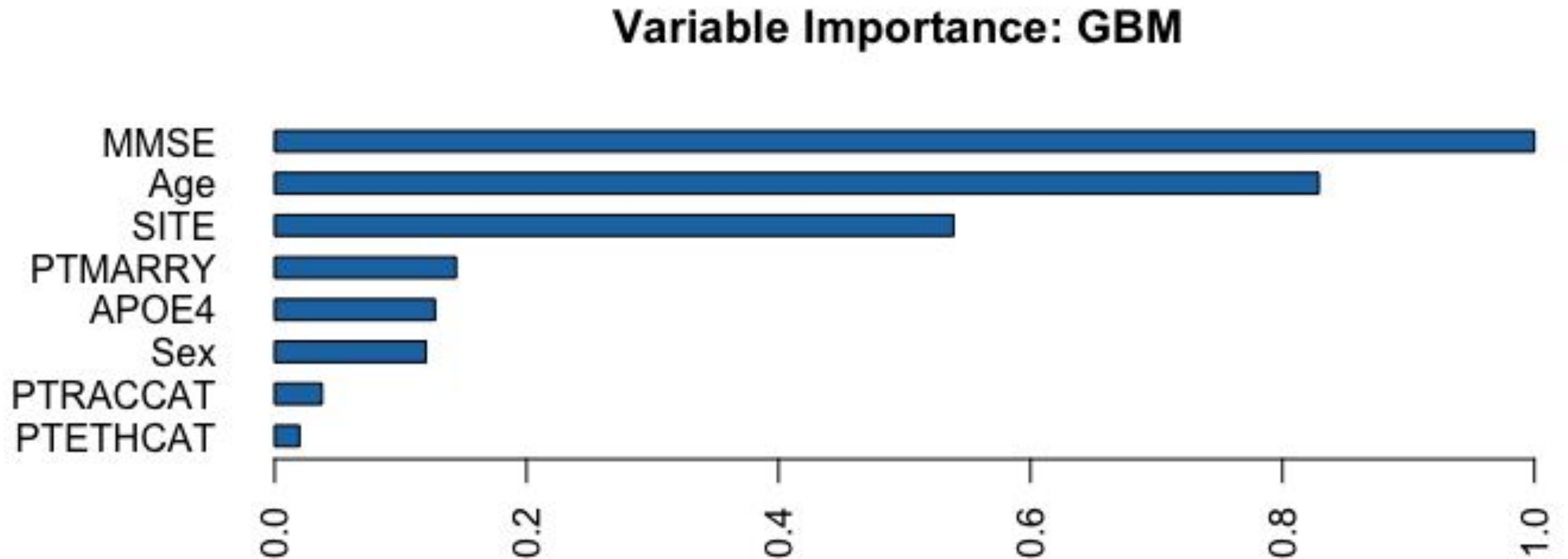


# Variable Importance for Random Forest





# Variable Importance for Gradient Boosting Machine





# Classification Results: Accuracy

	AD	Healthy	MCI	Total
DT	75%	62.5%	73.2%	71.6%
NN	75%	85%	80%	79.8%
RF	41.7%	50%	80%	61.9%
GBM	70.8%	90%	87.5%	83.3%

Baseline: 33%



# Characterizing PET Scans: Regularized Tensor-on-scalar Regression <sup>[3]</sup>

$$Y = X\Gamma + \varepsilon$$

$Y$  ( $n \times M$ ): vectorized images

$X$  ( $n \times p$ ): patient's data

$\Gamma$  ( $p \times M$ ): coefficient to estimate

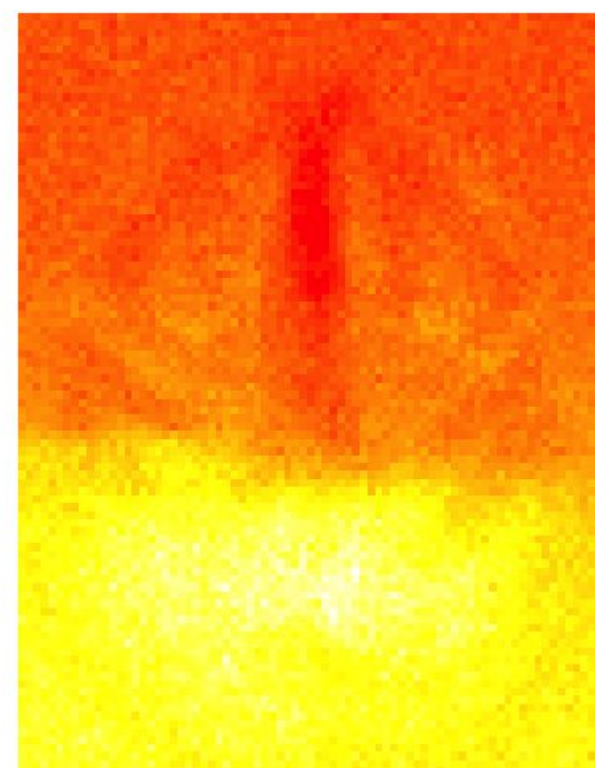
$M$ : # picture voxels, e.g.  $79 \times 95 = 7505$

$n$ : # images/subjects, e.g. 402

$p$ : # parameters in scalar data, e.g. 6

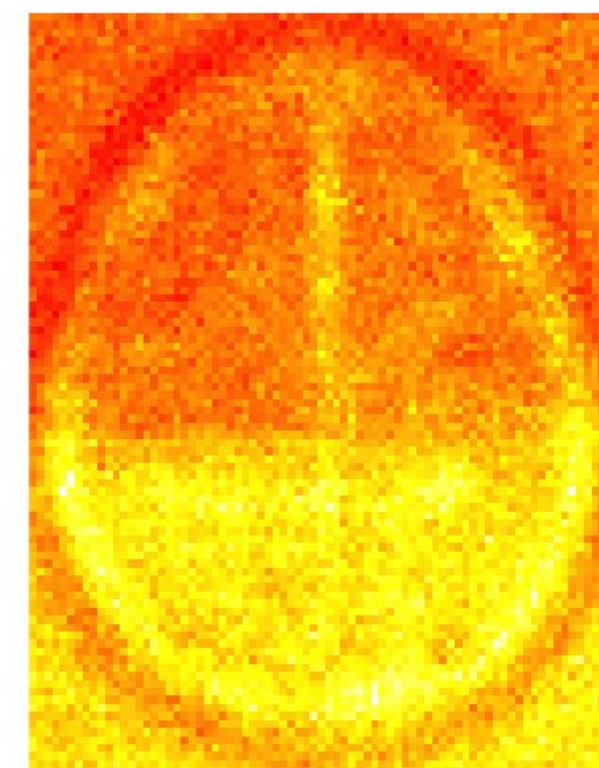
$$\min \frac{1}{2} \|Y - X\Gamma\|_F^2 + \lambda \|X\Gamma D\|_{l_1}$$

**Age**



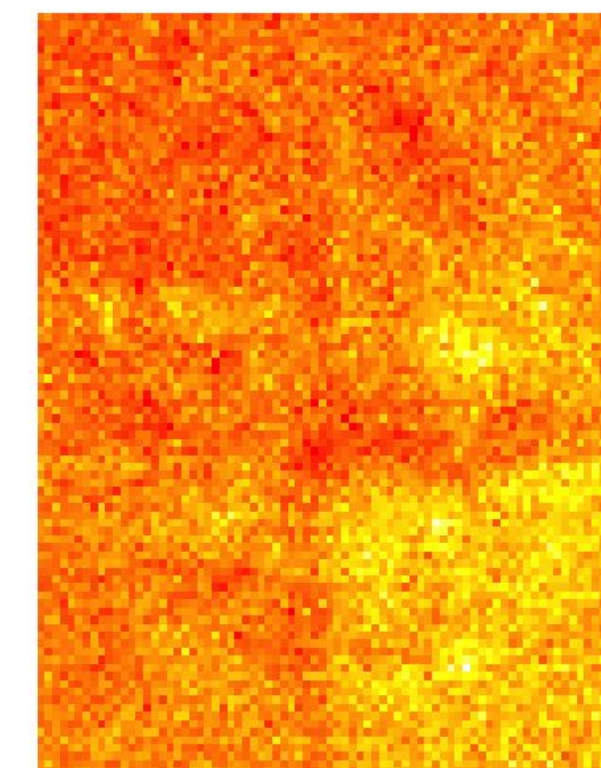
mean:1.92e-4

**Female**



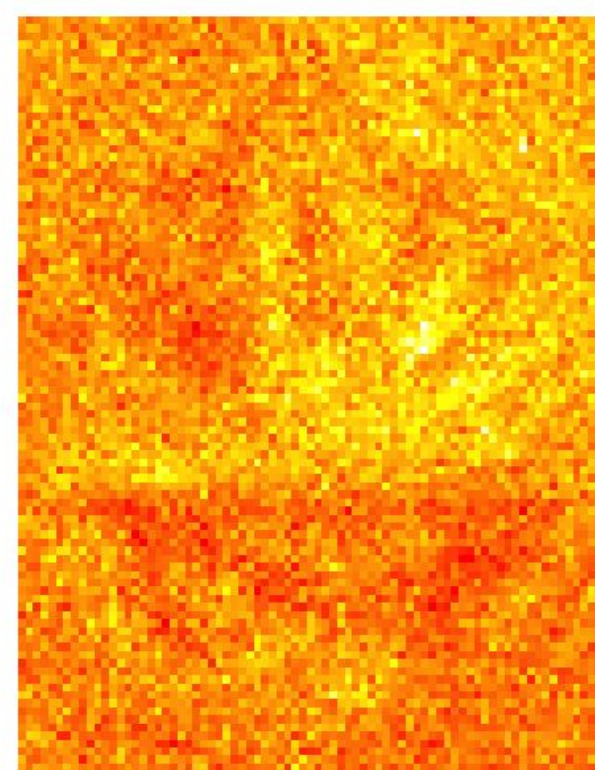
mean:-24e-4

**APOE4**



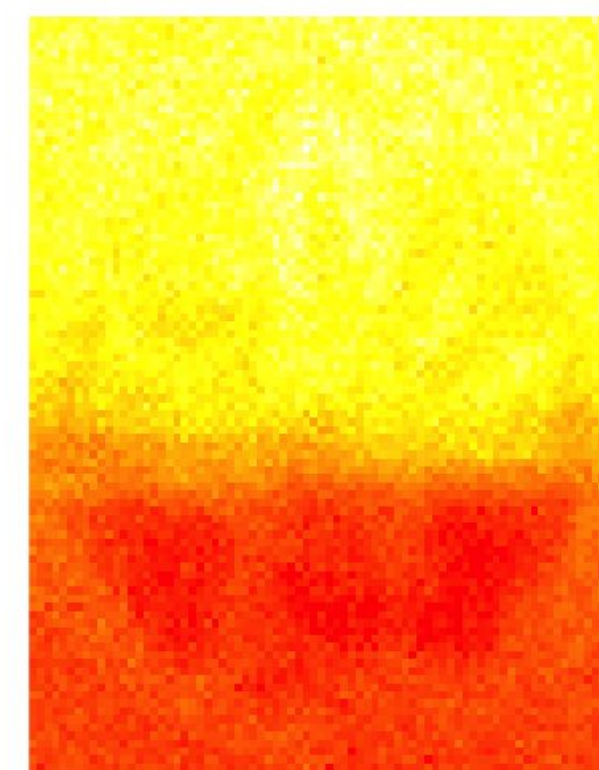
mean:6.93e-4

**D2**



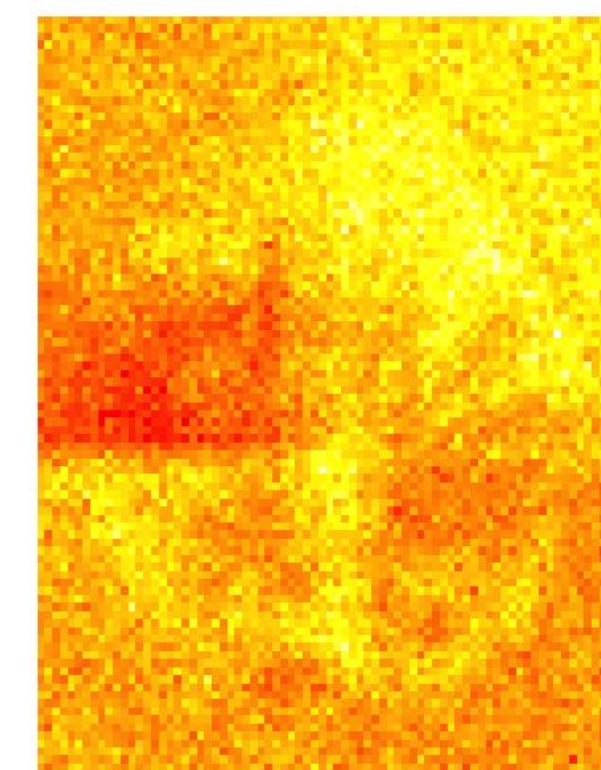
mean:-7.4e-4

**D3**



mean:-34e-4

**MMSE**



mean:8e-4



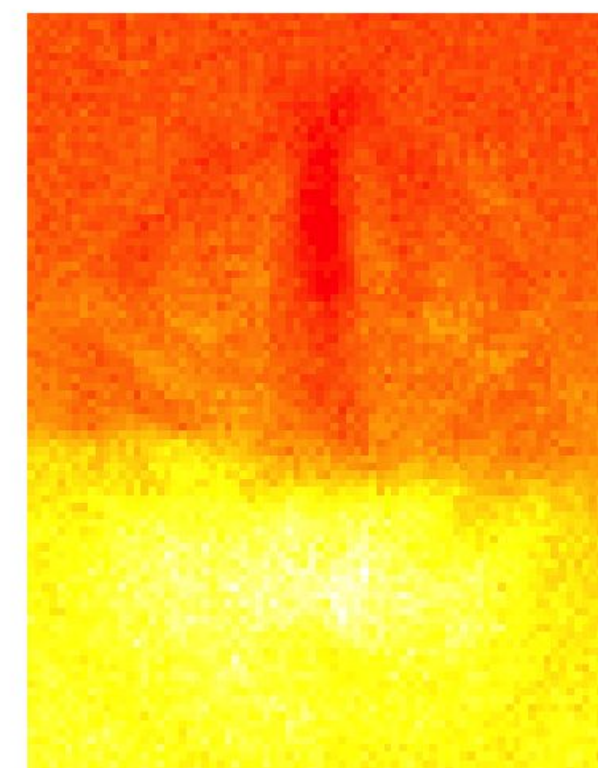
# Characterizing PET Scans: Regularized Tensor-on-scalar Regression <sup>[3]</sup>

Diagnosis stage 2 and Diagnosis stage 3 results in a decrease of brain activity compared with Diagnosis stage 1.

Lower MMSE score correlates with a lower brain activity.

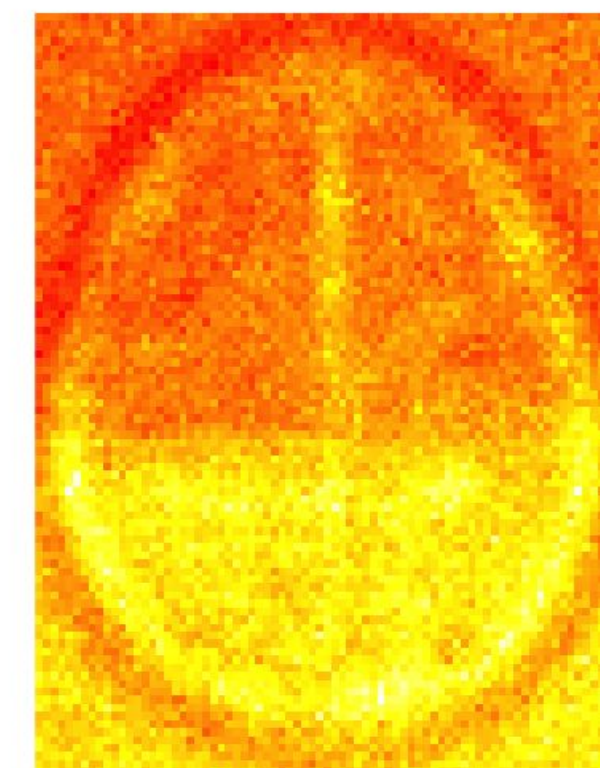
Higher age and non-zero APOE4 type correlates with a higher brain activity. Thus, collinearity issues still exist.

**Age**



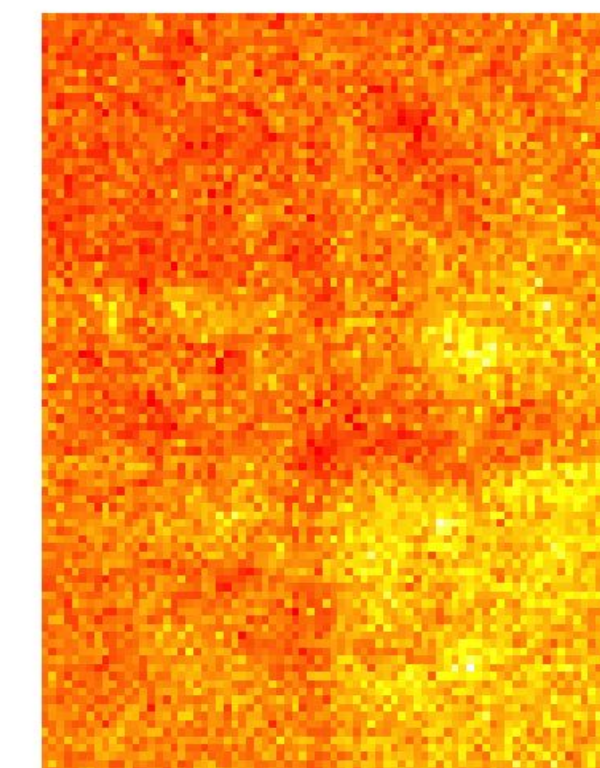
mean:1.92e-4

**Female**



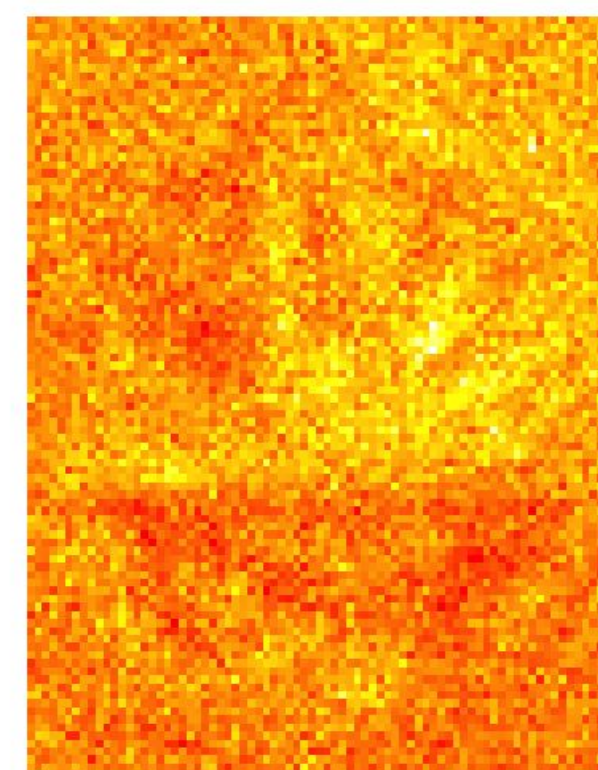
mean:-24e-4

**APOE4**



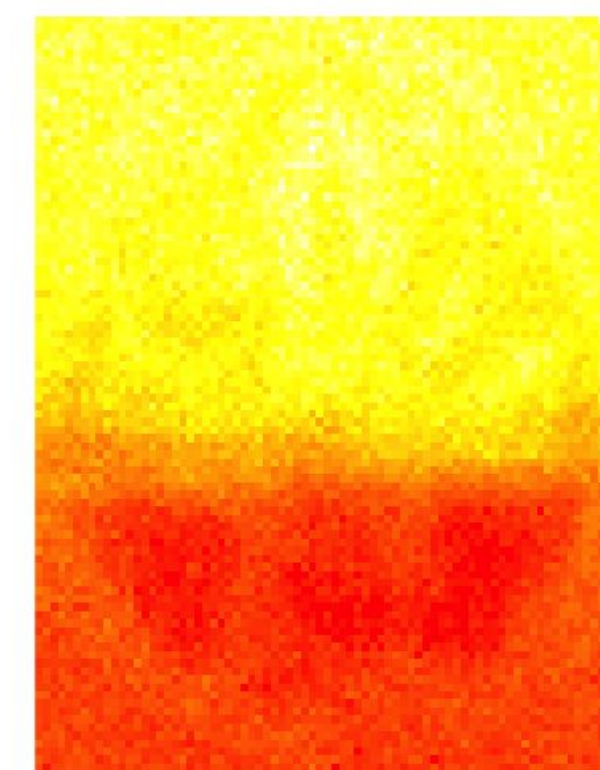
mean:6.93e-4

**D2**



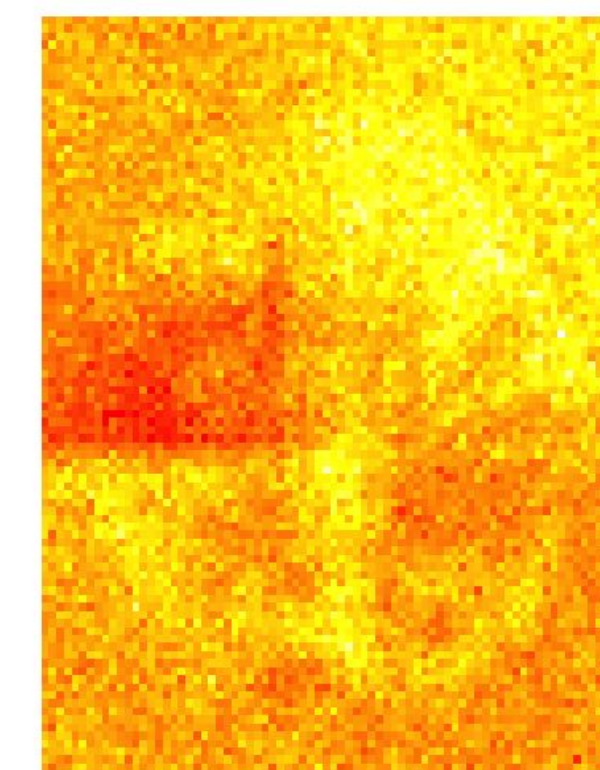
mean:-7.4e-4

**D3**



mean:-34e-4

**MMSE**



mean:8e-4



# Conclusions

## Important Features

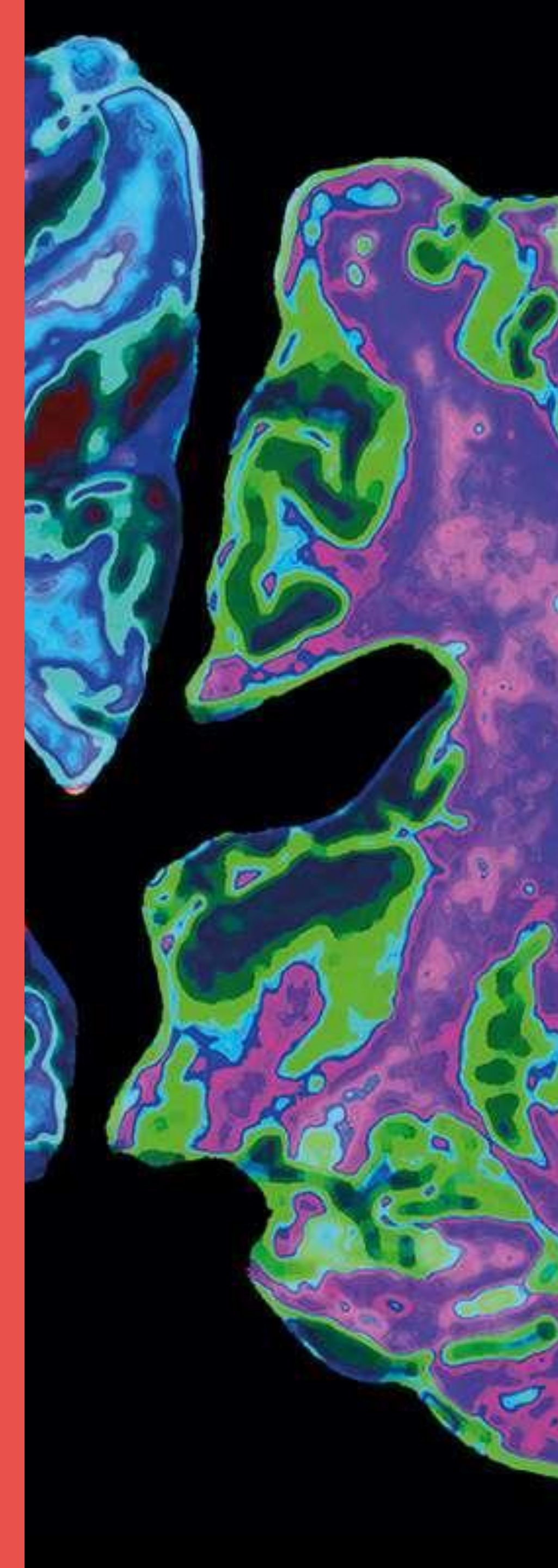
- Age
- MMSE
- Site

## Classification

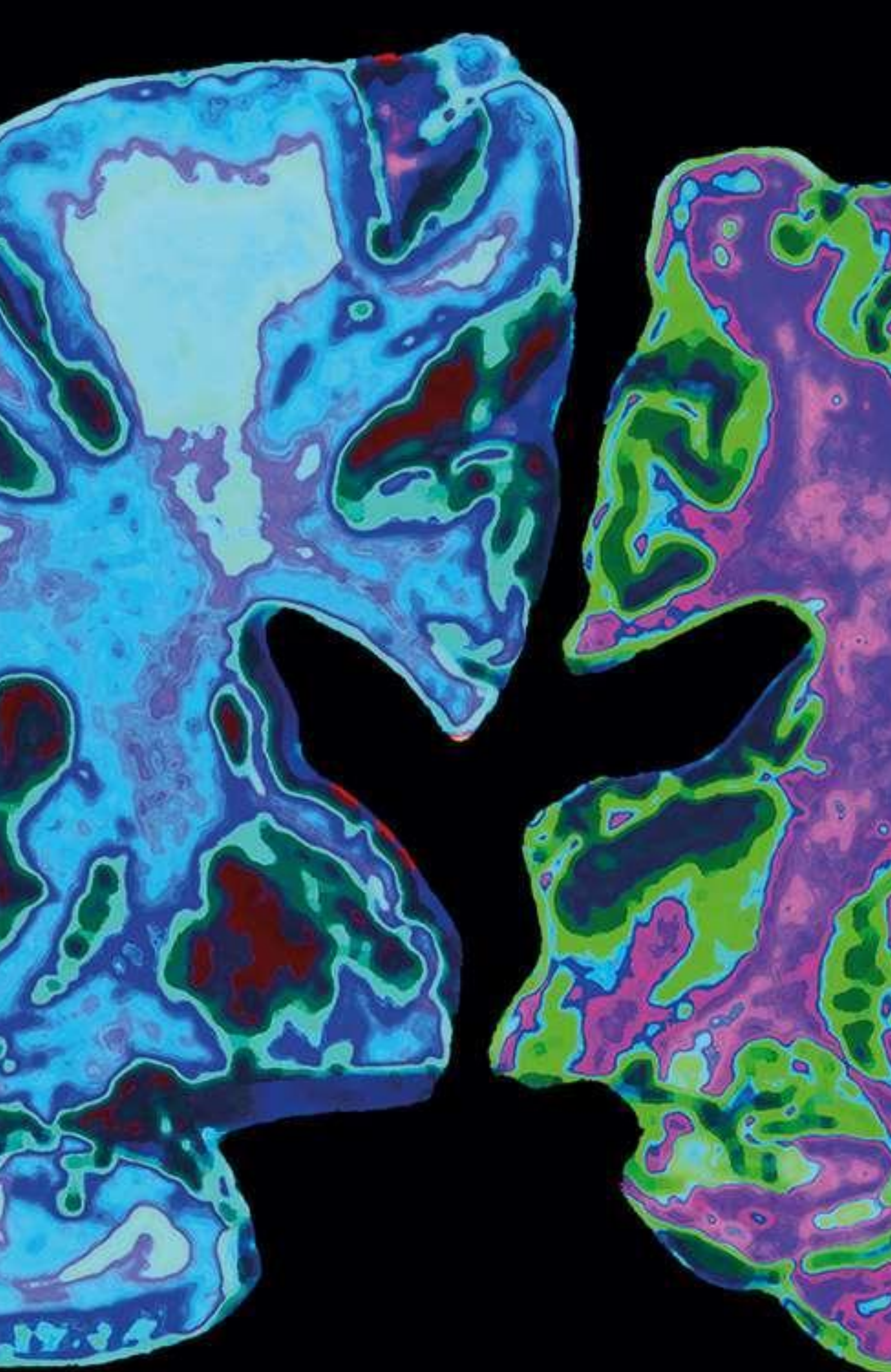
Significant accuracy above baseline can be achieved - could be used to help doctors make diagnoses

## Future Works

- Improve the efficiency of tensor-on-scalar regression
- Compare the coefficient map with result of other algorithms







SAMSI

NC State University

Dr. Xinyi Li

Dr. Mansoor Haider

Thomas Gehrmann

**THANK YOU!**



# Reference

- [1] Alzheimer's and Dementia. (n.d.). Retrieved from [https://www.alz.org/alzheimer\\_s\\_dementia](https://www.alz.org/alzheimer_s_dementia)
- [2] Powles, R. (2016, August 18). Why Does ApoE4 Increase Alzheimer's Risk? Retrieved from <http://longevityreporter.org/blog/2016/8/18/why-does-apoe4-increase-alzheimers-risk>
- [3] Total Variation Regularized Tensor-on-scalar Regression. (n.d.). Retrieved from <https://arxiv.org/pdf/1703.05264.pdf>
- [4] Overview about h2o. Retrieved from <http://docs.h2o.ai/h2o/latest-stable/h2o-docs/index.html>



# Gradient Boosting Machine with Parameter

Confusion Matrix: Row labels: Actual class; Column labels: Predicted class

	AD	CN	LMCI	Error	Rate
AD	17	1	6	0.2917 = 7 / 24	
CN	1	18	1	0.1000 = 2 / 20	
LMCI	2	3	35	0.1250 = 5 / 40	
Totals	20	22	42	0.1667 = 14 / 84	

Confusion Matrix: Row labels: Actual class; Column labels: Predicted class

	AD	CN	LMCI	Error	Rate
AD	15	0	9	0.3750 = 9 / 24	
CN	0	16	4	0.2000 = 4 / 20	
LMCI	4	4	32	0.2000 = 8 / 40	
Totals	19	20	45	0.2500 = 21 / 84	

Confusion Matrix: Row labels: Actual class; Column labels: Predicted class

	AD	CN	LMCI	Error	Rate
AD	12	0	10	0.4545 = 10 / 22	
CN	0	12	3	0.2000 = 3 / 15	
LMCI	6	3	36	0.2000 = 9 / 45	
Totals	18	15	49	0.2683 = 22 / 82	

# Best Neural Network model base on simulate parameter

> nnl\_confusion\_best\_model\_mse

Confusion Matrix: Row labels: Actual class; Column labels: Predicted class

	AD	CN	LMCI	Error	Rate
AD	13	0	11	0.4583 = 11 / 24	
CN	0	15	5	0.2500 = 5 / 20	
LMCI	5	8	27	0.3250 = 13 / 40	
Totals	18	23	43	0.3452 = 29 / 84	

> nnl\_confusion\_best\_model\_accuracy

Confusion Matrix: Row labels: Actual class; Column labels: Predicted class

	AD	CN	LMCI	Error	Rate
AD	18	1	5	0.2500 = 6 / 24	
CN	0	17	3	0.1500 = 3 / 20	
LMCI	2	6	32	0.2000 = 8 / 40	
Totals	20	24	40	0.2024 = 17 / 84	



# Random Forest Classification

Confusion Matrix: Row labels: Actual class; Column labels: Predicted class

	AD	CN	LMCI	Error	Rate
AD	10	1	13	0.5833 = 14 / 24	
CN	1	10	9	0.5000 = 10 / 20	
LMCI	4	4	32	0.2000 = 8 / 40	
Totals	15	15	54	0.3810 = 32 / 84	

# Decision Tree Classification Result

## Confusion Matrix and Statistics

	Reference		
Prediction	AD	CN	LMCI
AD	18	1	5
CN	0	10	6
LMCI	1	10	30

## Overall Statistics

Accuracy : 0.716  
95% CI : (0.605, 0.8107)  
No Information Rate : 0.5062  
P-Value [Acc > NIR] : 9.833e-05

Kappa : 0.5443

McNemar's Test P-Value : 0.1979

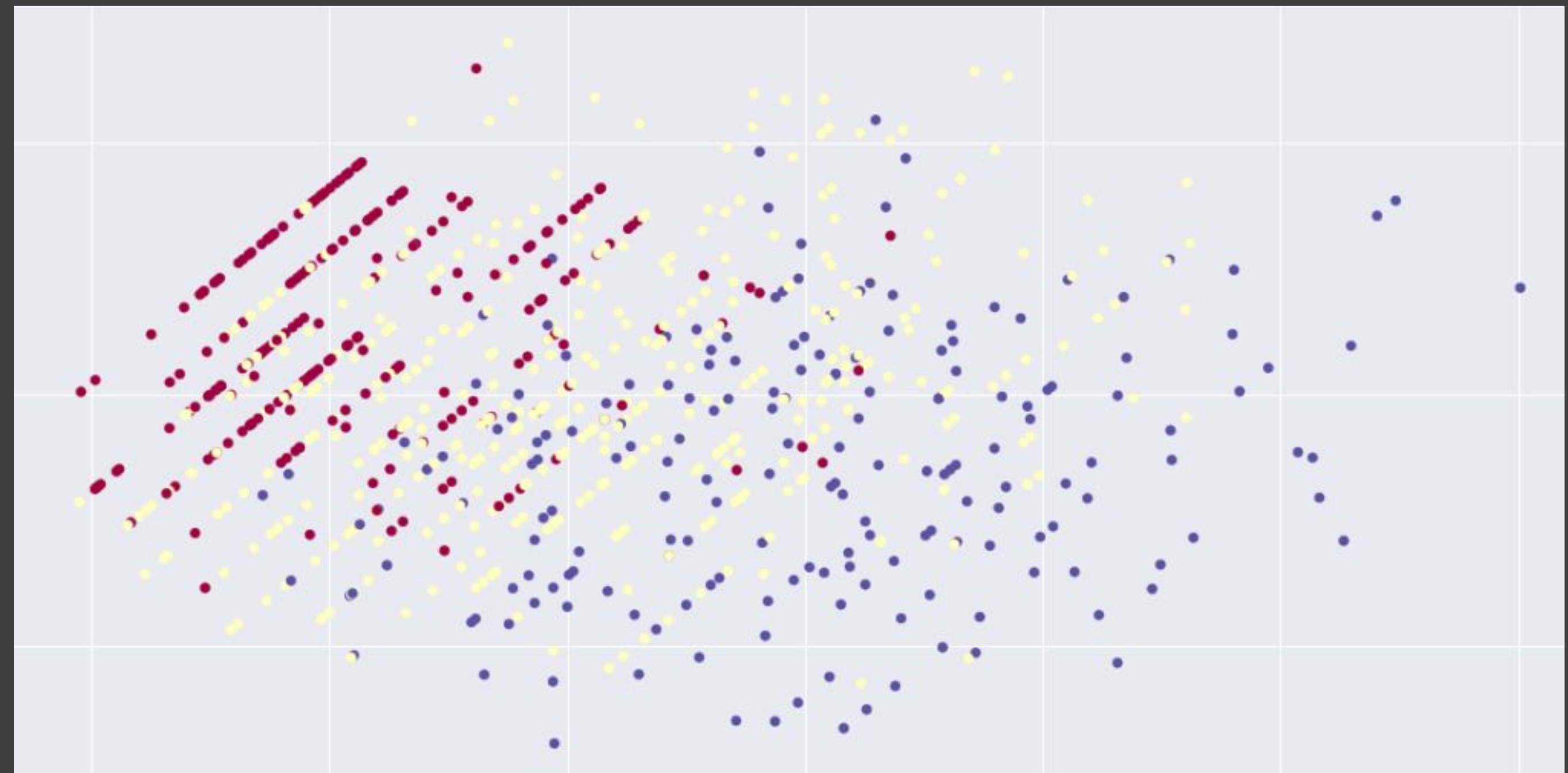
## Statistics by Class:

	Class: AD	Class: CN	Class: LMCI
Sensitivity	0.9474	0.4762	0.7317
Specificity	0.9032	0.9000	0.7250
Pos Pred Value	0.7500	0.6250	0.7317
Neg Pred Value	0.9825	0.8308	0.7250
Prevalence	0.2346	0.2593	0.5062
Detection Rate	0.2222	0.1235	0.3704
Detection Prevalence	0.2963	0.1975	0.5062
Balanced Accuracy	0.9253	0.6881	0.7284



# Principal Component Analysis

PC 2 (32%)



PC 1 (58%)