HEALTHY AGING SIGNALS

Partner: Merck
Team members: Dan Cox, Aaron Jacobson, Yaxin Lei, Eleonora Shantsila

General Background - The Topic of Aging

The Top 10 Hot Topics in Aging FREE

John E. Morley

The Journals of Gerontology: Series A, Volume 59, Issue 1, January 2004, Pages M24–M33,
https://doi.org/10.1093/gerona/59.1.M24

Published: 01 January 2004

- The topic of **aging** has almost always been associated with **health issues**.

- A large portion of age-related research has been focused on neurodegenerative disease, like cognitive decline, depression and Parkinson's disease.

- In our project, we are going to study the **signals of aging**, specifically **healthy aging**, potentially through lab imaging data, personal traits and DNA sequences.

Goals

- What are general signals of aging?
- How do we separate healthy aging signals from unhealthy ones?
- Project parts:
  - Finding Relevant Data
  - Building predictive models of aging
  - Comparison between models of healthy and unhealthy cohorts
- We’re not focusing on MRI

Learning goals

- AWS
- Working with high-dimensional datasets
- Learning biological terminology
- Learning advanced ML techniques

Team cooperation/organization structure

- Slack
- Google Drive
- GitHub
- DeepNote / AWS
- Weekly meetings with Merck
- Team meetings twice a week
- Parallel work initially, recently started to merge the work

Literature review

2013 - Horvath - DNA methylation age of human tissues and cell types

2015 - Putin et al - Application of deep neural networks to biomarker development
Levine, M. E., Lu, A. T., Quach, A., Chen, B. H., Assimes, T. L., Bandinelli, S., et al.

2018 - Tozer et al - Textured analysis of T1-weighted and fluid-attenuated inversion recovery images detects abnormalities that correlate with cognitive decline in small vessel disease

2020 - Lagner et al - Identifying Morphological Indicators of Aging with NN on large-scale whole body MRI

Databases

MERCK Suggested candidate datasets and links

Cohort	Data composition	Link
Normal human	MRI and related metadata	https://www.humanconnectome.org/
Alzheimer's disease	Clinical data, MRI, PET images, genetic data, image analysis results, chemical biomarker	http://adni.loni.usc.edu/data-samples/
Parkinson's disease	Clinical, biomarker, imaging and related metaData	http://www.ppmi-info.org/access-data-specimens/
Alzheimer's disease	Gene expression, chromatin activity, proteomics, genomic variants, etc.	https://adknowledgeportal.synapse.org/Explore/Data

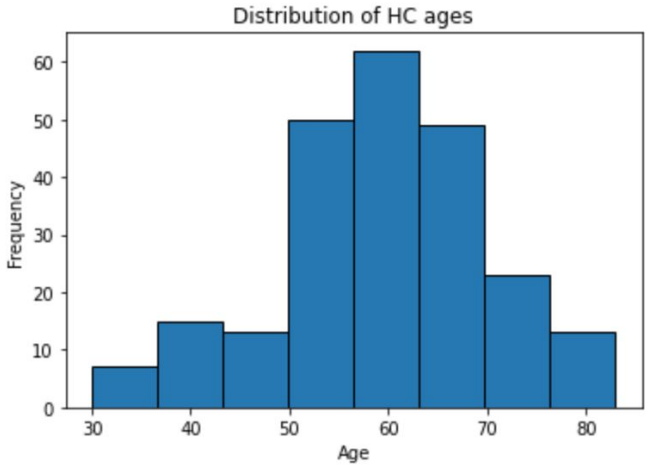
PPMI

- Brain MRI images
- Demographics
- Vital Signs
- Physical Examination
- Blood tests
- Grey matter brain volume
- DaTScan brain cell density
- DNA Methylation

Blood data EDA

	PATNO	YOB	AGE	Prothrombin Time	APTT-QT	Monocytes	Eosinophils	Basophils
0	3404	1954.0	56.0	10.1	21.0	0.32	0.10	0.02
1	3401	1954.0	56.0	10.0	24.8	0.55	0.05	0.02
2	3405	1947.0	63.0	10.0	19.8	0.29	0.10	0.06
3	3100	1942.0	68.0	10.3	21.9	0.64	0.39	0.05
4	3103	1963.0	47.0	10.1	19.5	0.24	0.05	0.03
...	...	...	...	...	...	...	...	...
227	3171	1950.0	60.0	11.2	22.5	0.25	0.08	0.01
228	3157	1946.0	64.0	10.0	22.8	0.50	0.13	0.02
229	3191	1947.0	63.0	10.6	21.8	0.20	0.06	0.05
230	3172	1942.0	68.0	10.8	25.1	0.41	0.07	0.03
231	4105	1946.0	64.0	10.2	27.4	0.38	0.14	0.06

Reformatted blood chemistry data with patient ages

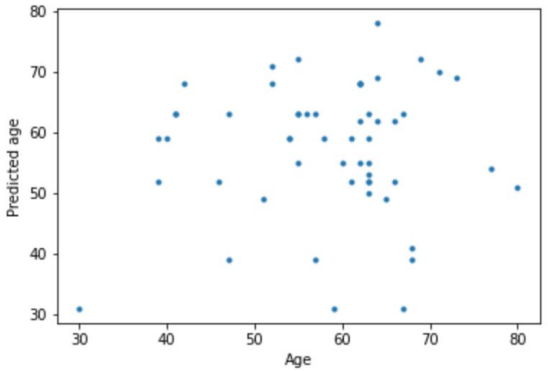


Distribution of healthy control (HC) ages

Correlations

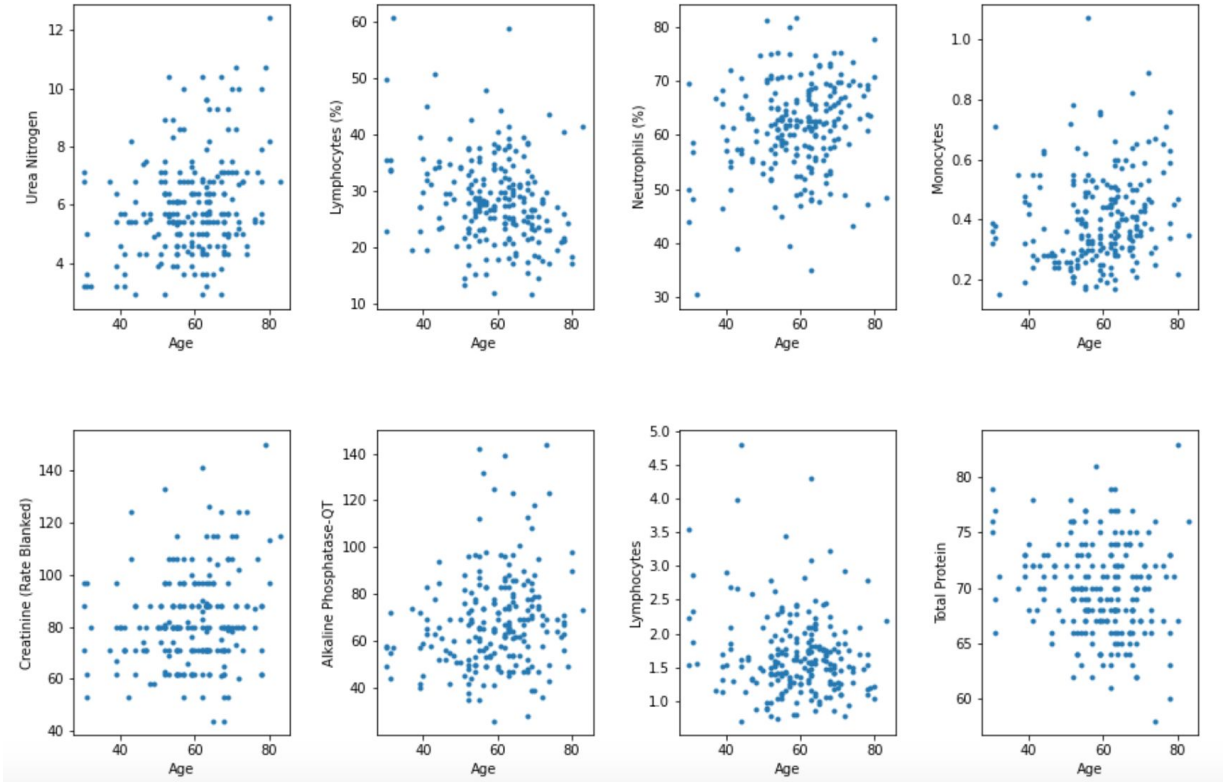
Urea Nitrogen	0.298034
Lymphocytes (%)	-0.255166
Neutrophils (%)	0.197271
Monocytes	0.180632
Creatinine (Rate Blanked)	0.180316
Alkaline Phosphatase-QT	0.178899
Lymphocytes	-0.161696
Total Protein	-0.159974
Monocytes (%)	0.158007
Serum Uric Acid	0.128502
Basophils	0.123712
Albumin-QT	-0.122841
Neutrophils	0.113783
Platelets	-0.099442
AST (SGOT)	0.095477
Serum Glucose	0.089607
Basophils (%)	0.082415
Serum Potassium	0.078685
WBC	0.053332
Serum Chloride	-0.048654
ALT (SGPT)	0.046626
Total Bilirubin	0.040457
Serum Sodium	0.030841
Prothrombin Time	0.030318
Hematocrit	0.025877
Eosinophils (%)	-0.023411
Hemoglobin	0.022398
APTT-QT	0.021057
Eosinophils	-0.013548
RBC	-0.012664
Calcium (EDTA)	0.010038
Serum Bicarbonate	0.009060

Pearson's correlation coefficients of blood chemistry results with age. Pink highlights tests found to be most significant by the Putin study, green highlights tests found to be most significant by the Levine study



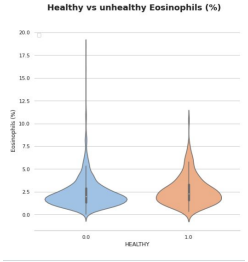
Comparison of the Linear Regression test set predictions vs. the true values

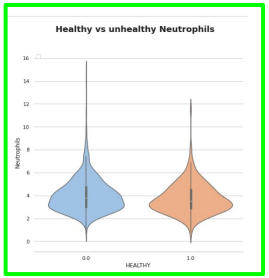
Correlations

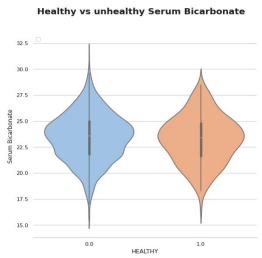


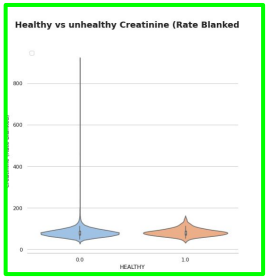
Scatter plots of the 8 most highly correlated blood test results

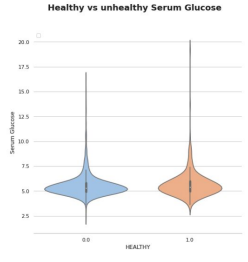
Any true difference between healthy/unhealthy within PPMI?

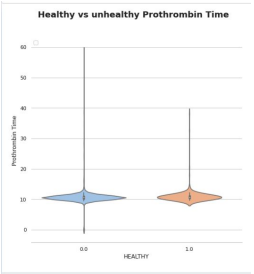


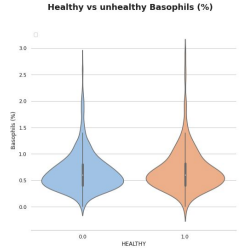


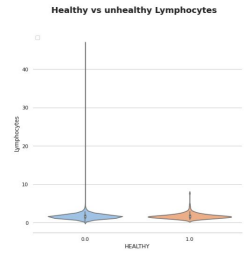


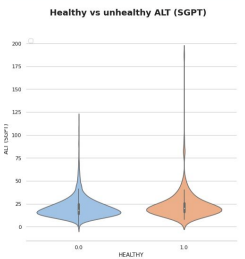


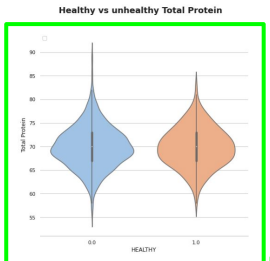


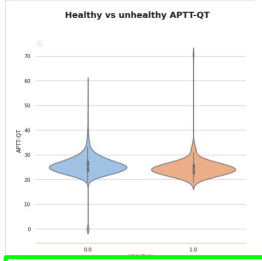


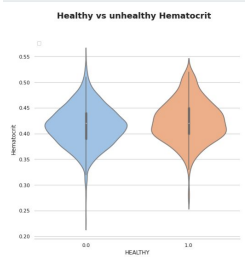


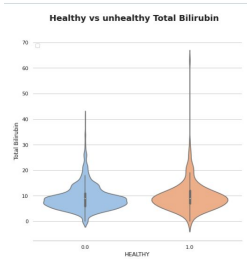


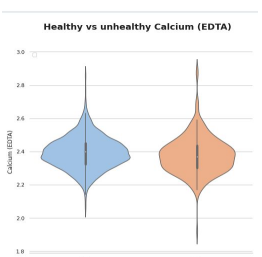


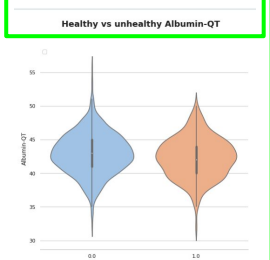


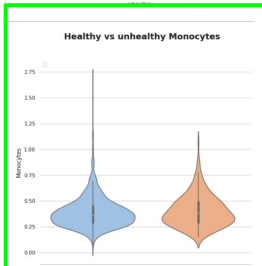




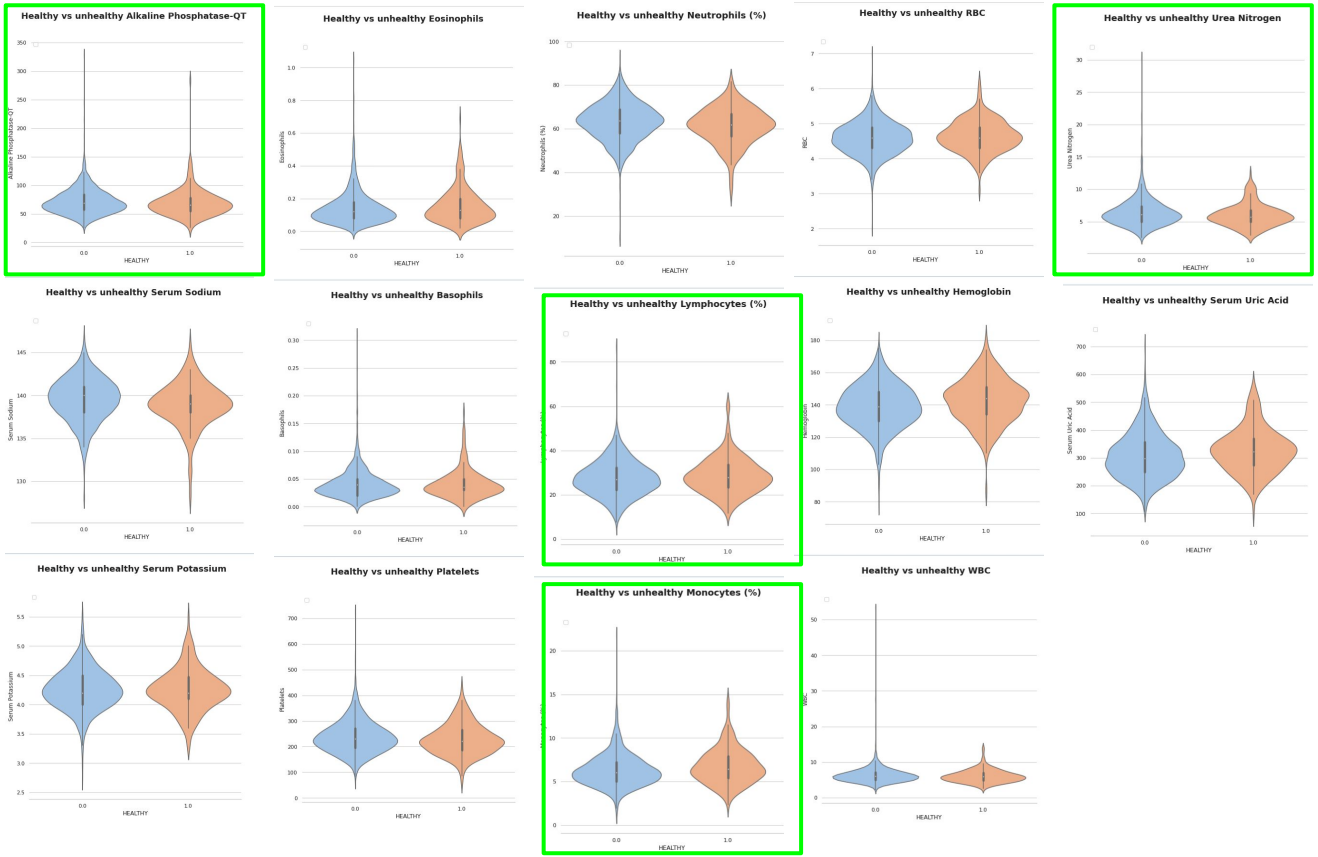








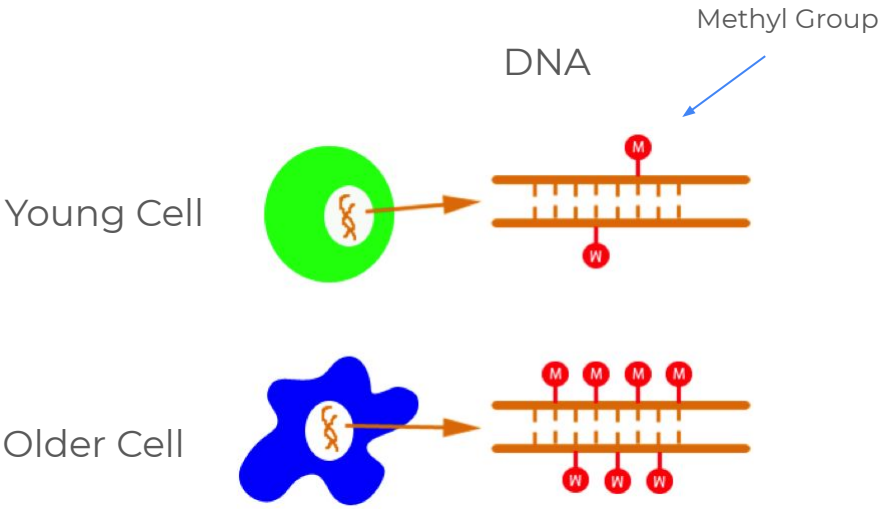
Any true difference between healthy/unhealthy within PPMI?



- Not able to observe clear difference between the healthy, and the unhealthy group blood test metric values.

- Even the minimal differences we observed, can be likely due to the underlying age distribution difference between the two groups.

Methylation EDA



Horvath *Genome Biology* , **14**:R115
http://genomebiology.com//14/10/R115

Genome Biology

RESEARCH Open Access

DNA methylation age of human tissues and cell types

2013

Steve Horvath[1,2,3]

EPIGENETICS

DNA methylation-based biomarkers and the epigenetic clock theory of ageing

2018

Steve Horvath[1,2] and Kenneth Raj[3]*

Abstract | Identifying and validating molecular targets of interventions that extend the human health span and lifespan has been difficult, as most clinical biomarkers are not sufficiently representative of the fundamental mechanisms of ageing to serve as their indicators. In a recent breakthrough, biomarkers of ageing based on DNA methylation data have enabled accurate age estimates for any tissue across the entire life course. These 'epigenetic clocks' link developmental and maintenance processes to biological ageing, giving rise to a unified theory of life course. Epigenetic biomarkers may help to address long-standing questions in many fields, including the central question: why do we age?

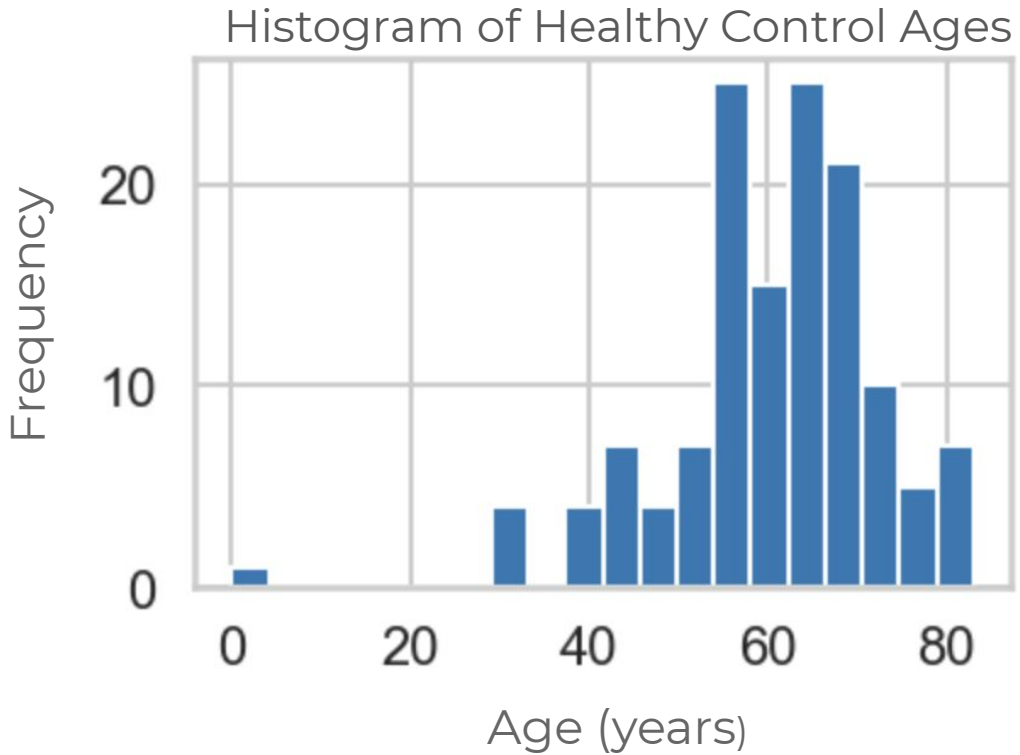
PPMI Methylation Profiling Data

	PATNO	Age	cg16867657	cg07323488	cg22454769	cg06784991	cg11436113	cg19283806	cg13552692	cg15736994	cg13823169	cg177
0	3074	31.160903	0.578927	0.378975	0.457985	0.047398	0.632738	0.352091	0.449920	0.694283	0.598616	0.2
1	3011	31.901370	0.488026	0.293259	0.463545	0.049145	0.726317	0.364094	0.401152	0.689200	0.530503	0.3
2	3619	32.191781	0.565370	0.370947	0.448252	0.084273	0.632496	0.228096	0.393248	0.669290	0.530521	0.3
3	3355	32.331507	0.582600	0.368443	0.421757	0.082836	0.660496	0.413355	0.482633	0.739244	0.543889	0.3
4	3555	39.780822	0.505620	0.324848	0.611662	0.069138	0.647653	0.233859	0.403140	0.664967	0.576215	0.3
...	...	...	...	...	...	...	...	...	...	...	...	
129	4139	80.924231	0.784670	0.192347	0.653453	0.142914	0.525786	0.097309	0.284195	0.492253	0.468315	0.0
130	3274	81.263014	0.764245	0.224250	0.668625	0.134502	0.471430	0.177653	0.299203	0.546103	0.428705	0.1
131	3008	81.890411	0.804439	0.132712	0.563937	0.350124	0.514411	0.091213	0.141609	0.504760	0.432864	0.0
132	3965	82.712329	0.839778	0.226046	0.720982	0.387490	0.536615	0.116127	0.225331	0.597154	0.411576	0.1
133	3009	83.682192	0.791669	0.167237	0.704935	0.259737	0.470238	0.149689	0.260069	0.470776	0.454470	0.1

134 rows x 864067 columns

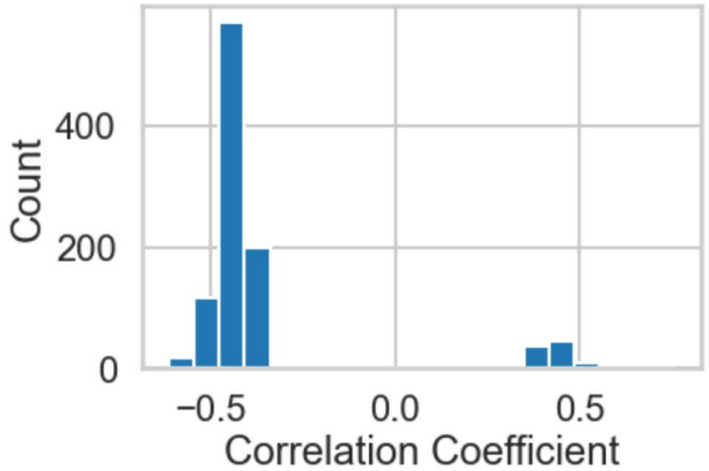
Normalized processed data with 133 samples x 864,067 CpG sites. Beta values range from 0 to 1, indicate probability of being methylated

PPMI Methylation Profiling Data

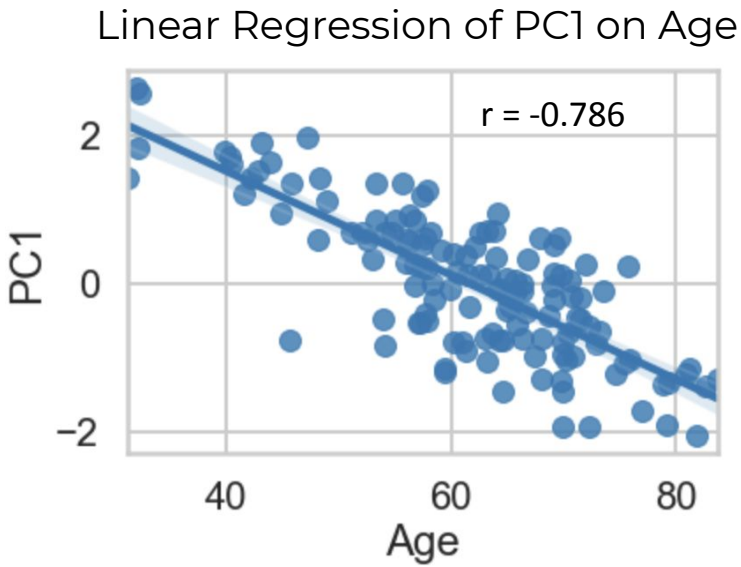


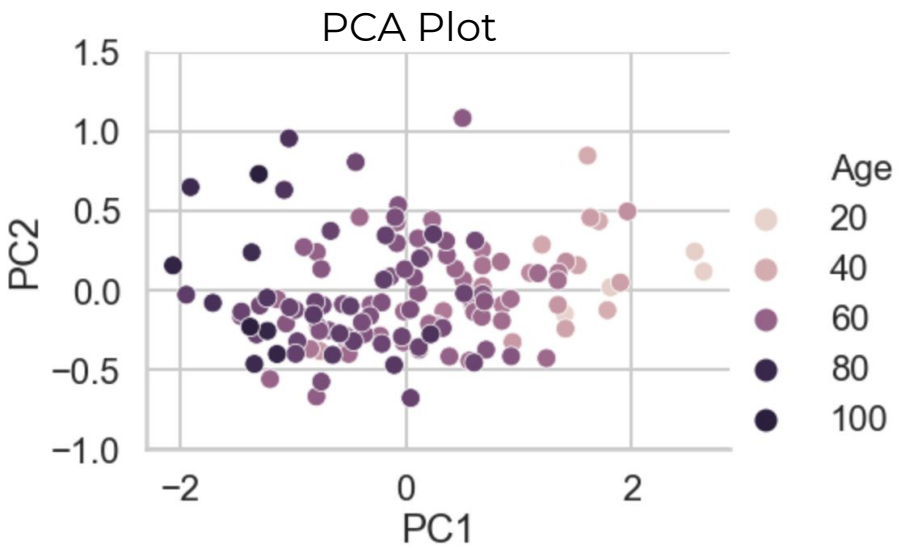
EDA of the top 1000 CpGs correlated with Age

- Correlations calculated for each CpG and Age
- The top 1000 saved
- PCA done to look for segregation by age
- Regression model created using 10 Pcs



PCA Plot of 134 Healthy Control Samples with top 1000 CpGs





Linear Regression of the first 10 PCs on Age

OLS Regression Results

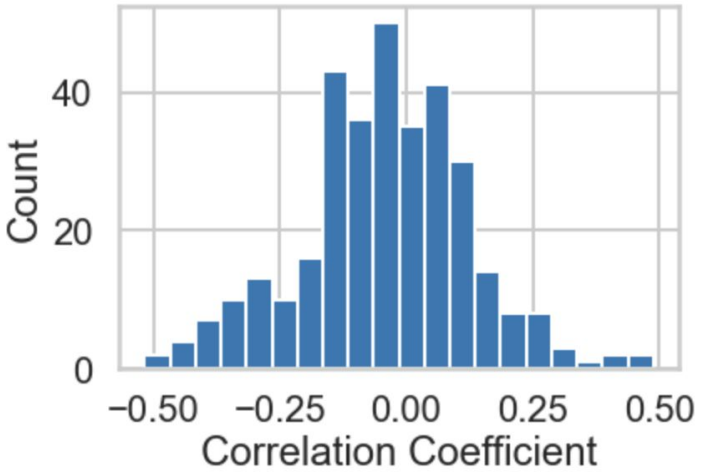
Dep. Variable:	Age	R-squared:	0.849
Model:	OLS	Adj. R-squared:	0.837
Method:	Least Squares	F-statistic:	69.27
Date:	Wed, 24 Feb 2021	Prob (F-statistic):	1.09e-45
Time:	17:26:59	Log-Likelihood:	-384.12
No. Observations:	134	AIC:	790.2
Df Residuals:	123	BIC:	822.1
Df Model:	10		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	61.5384	0.383	160.473	0.000	60.779	62.297
PC1	-8.7932	0.392	-22.441	0.000	-9.569	-8.018
PC2	-0.3287	0.866	-0.380	0.705	-2.042	1.385
PC3	1.0302	1.183	0.871	0.386	-1.312	3.373
PC4	-11.8697	1.479	-8.026	0.000	-14.797	-8.942
PC5	8.6396	1.490	5.798	0.000	5.690	11.589
PC6	-4.9106	1.532	-3.206	0.002	-7.943	-1.879
PC7	-11.6706	1.680	-6.948	0.000	-14.996	-8.346
PC8	8.7994	1.758	5.006	0.000	5.320	12.279
PC9	0.4007	1.994	0.201	0.841	-3.547	4.349
PC10	-5.1522	2.025	-2.545	0.012	-9.160	-1.144

Omnibus:	4.867	Durbin-Watson:	2.060
Prob(Omnibus):	0.088	Jarque-Bera (JB):	5.939
Skew:	0.172	Prob(JB):	0.0513
Kurtosis:	3.972	Cond. No.	5.28

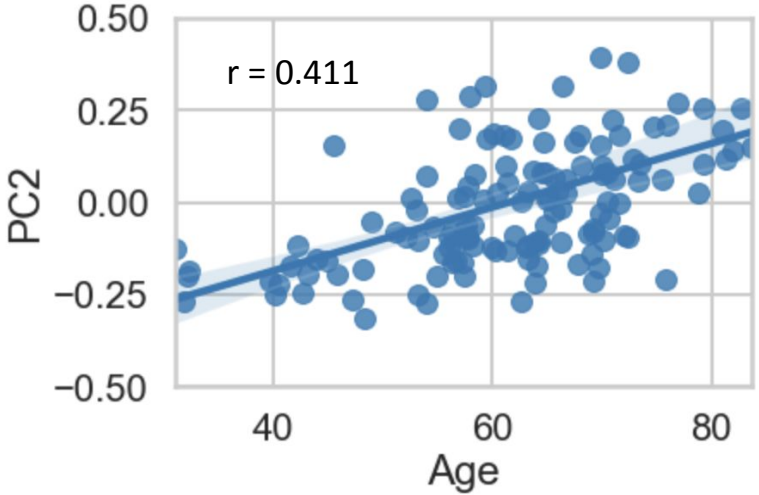
EDA of the 335 CpGs common with those of Horvath 2013

- Correlations calculated for each CpG and Age
- PCA done to look for segregation by age
- Regression model created using 10 Pcs

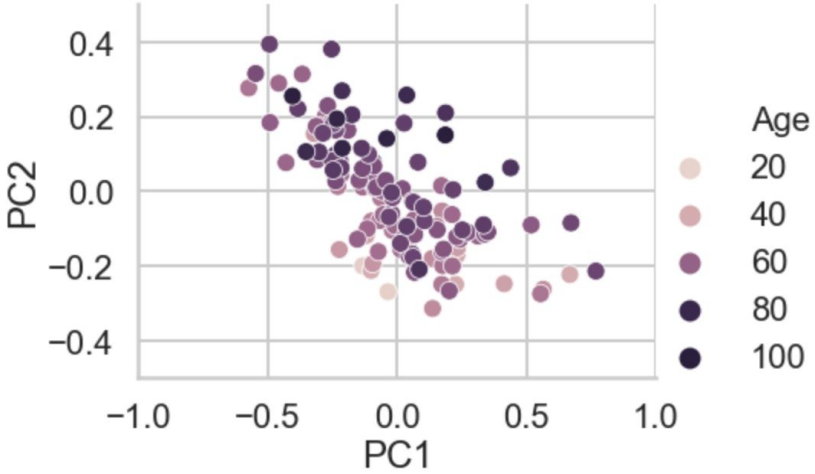


PCA Plot of 134 Healthy Control Samples with the 335 CpGs

Linear Regression of PC2 on



PCA Plot



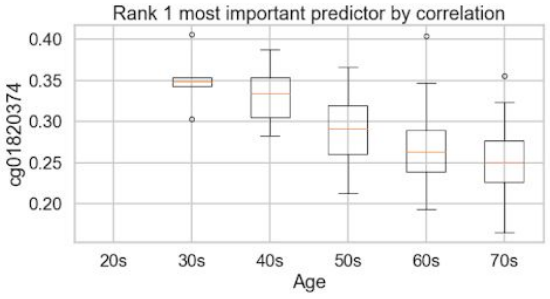
Linear Regression of the first 10 PCs on Age

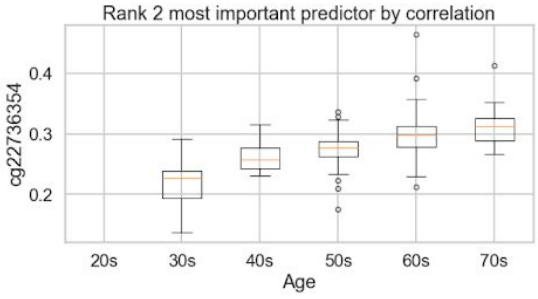
OLS Regression Results

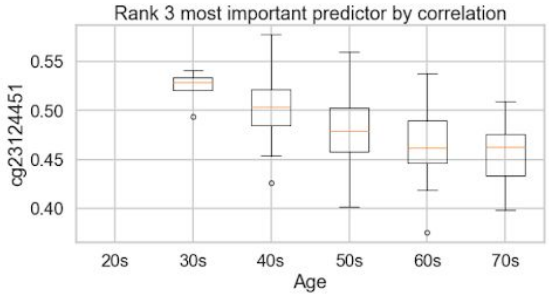
Dep. Variable:	Age	R-squared:	0.729
Model:	OLS	Adj. R-squared:	0.707
Method:	Least Squares	F-statistic:	33.10
Date:	Thu, 25 Feb 2021	Prob (F-statistic):	2.65e-30
Time:	09:47:05	Log-Likelihood:	-423.37
No. Observations:	134	AIC:	868.7
Df Residuals:	123	BIC:	900.6
Df Model:	10		
Covariance Type:	nonrobust		

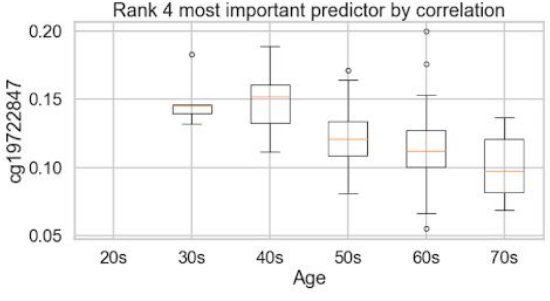
	coef	std err	t	P>|t|	[0.025	0.975]
const	61.5384	0.514	119.727	0.000	60.521	62.556
PC1	-2.6703	1.742	-1.533	0.128	-6.119	0.778
PC2	19.5339	2.232	8.753	0.000	15.117	23.951
PC3	-16.0569	3.094	-5.189	0.000	-22.182	-9.932
PC4	44.3673	3.280	13.527	0.000	37.875	50.860
PC5	-3.9829	3.575	-1.114	0.267	-11.059	3.093
PC6	-2.6676	3.630	-0.735	0.464	-9.854	4.519
PC7	5.6708	4.121	1.376	0.171	-2.486	13.828
PC8	5.0989	4.338	1.175	0.242	-3.488	13.685
PC9	15.2962	4.603	3.323	0.001	6.184	24.408
PC10	24.8927	4.875	5.106	0.000	15.243	34.543

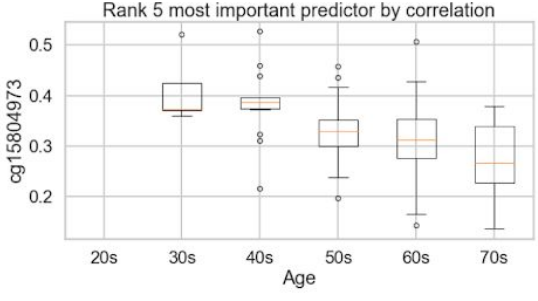
Omnibus:	0.273	Durbin-Watson:	1.589
Prob(Omnibus):	0.873	Jarque-Bera (JB):	0.073
Skew:	0.040	Prob(JB):	0.964
Kurtosis:	3.082	Cond. No.	9.49

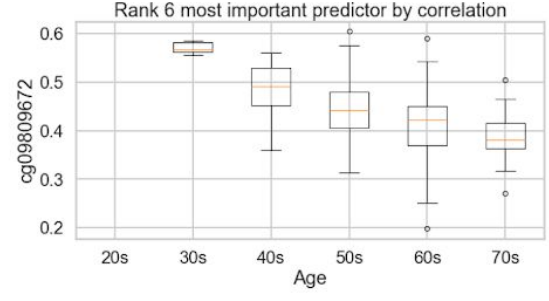

Rank 1 most important predictor by correlation
cg01820374
0.40
0.35
0.30
0.25
0.20
20s
30s
40s
50s
60s
70s
Age

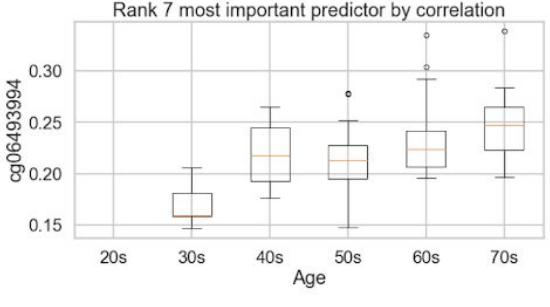

Rank 2 most important predictor by correlation
cg22736354
0.4
0.3
0.2
20s
30s
40s
50s
60s
70s
Age

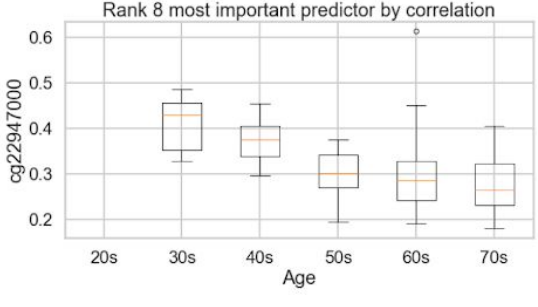

Rank 3 most important predictor by correlation
cg23124451
0.55
0.50
0.45
0.40
20s
30s
40s
50s
60s
70s
Age

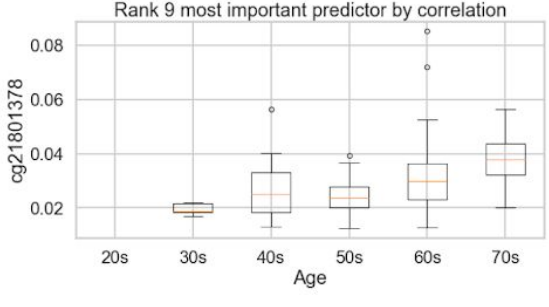

Rank 4 most important predictor by correlation
cg19722847
0.20
0.15
0.10
0.05
20s
30s
40s
50s
60s
70s
Age


Rank 5 most important predictor by correlation
cg15804973
0.5
0.4
0.3
0.2
20s
30s
40s
50s
60s
70s
Age


Rank 6 most important predictor by correlation
cg09809672
0.6
0.5
0.4
0.3
0.2
20s
30s
40s
50s
60s
70s
Age


Rank 7 most important predictor by correlation
cg06493994
0.30
0.25
0.20
0.15
20s
30s
40s
50s
60s
70s
Age


Rank 8 most important predictor by correlation
cg22947000
0.6
0.5
0.4
0.3
0.2
20s
30s
40s
50s
60s
70s
Age

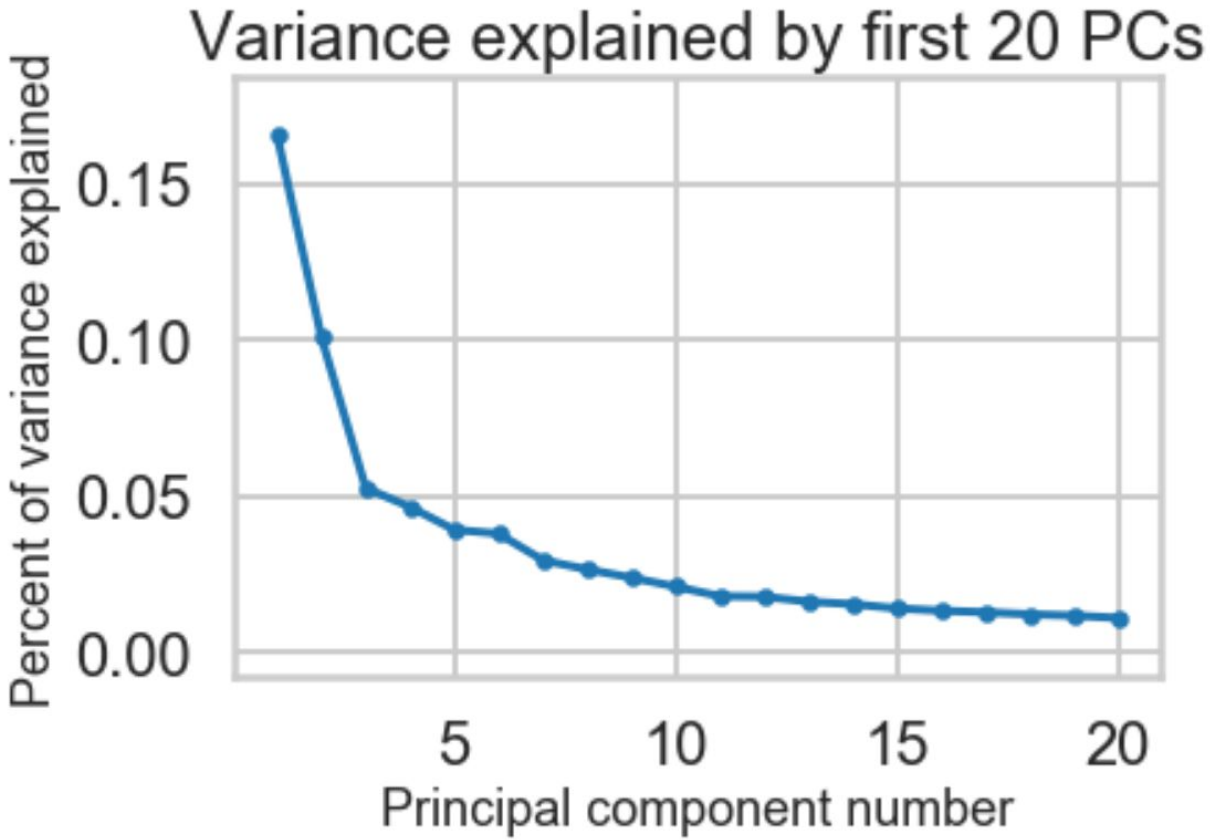

Rank 9 most important predictor by correlation
cg21801378
0.08
0.06
0.04
0.02
20s
30s
40s
50s
60s
70s
Age

Linear Regression with all 335 predictors

- 85-15 train-test split (114 and 21 patients)

```
Train MSE: 4.258630997673122e-27
Test MSE: 66.12140699483992
Train r^2: 1.0
Test r^2: 0.17572533681805735
```

- 47.3% from first 7 PCs



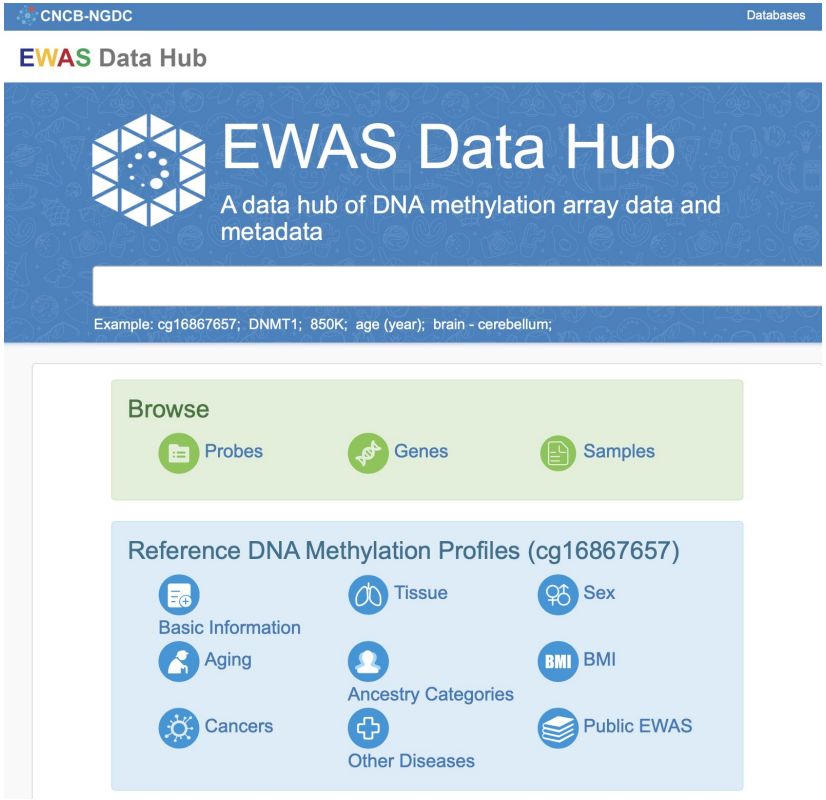
```
Train MSE: 73.96400487585797
Test MSE: 45.57224182366759
Train r^2: 0.5299063528882135
Test r^2: 0.4318928470082145
```

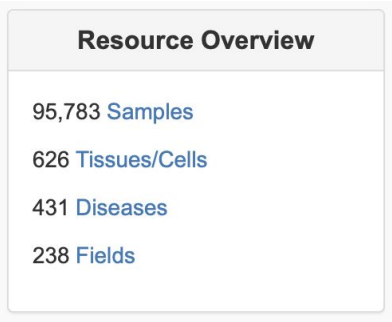
Linear regression with first 7 PCs as features

```
Train MSEs: [72.19648502 40.71750456 73.90331455 78.14491506 74.57141582]
Test MSEs: [295.79891641 185.31292021  50.7801157   34.04216107  48.56972427]
Train r^2s: [0.48491266 0.69011935 0.50097329 0.48347877 0.52460435]
Test r^2s: [-0.75812388  0.08069901  0.6035098   0.72008465  0.47127988]
```

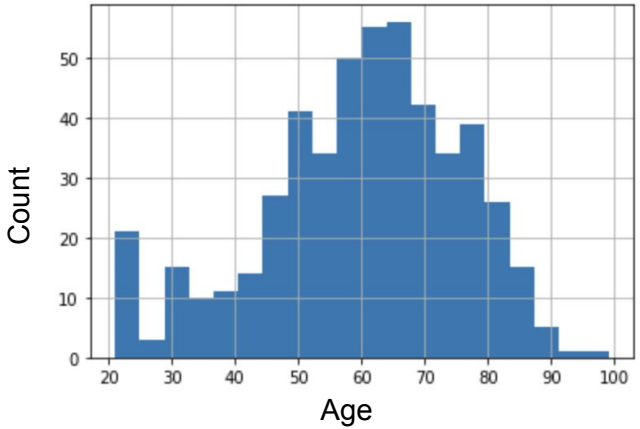
Results of 5-fold cross validation

EWAS Database for DNA Methylation





- Healthy
- Whole Blood
- Ages 20 and over
- ~ 9000 samples



Milestone 1: Overview of current progress

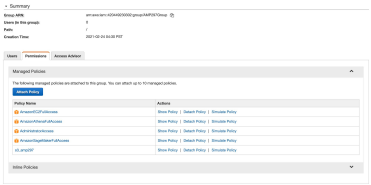
Database Exploration

Databases:
PPMI Database, EWAS, Kaggle, Alzheimer, National Cancer Institute, Etc.

Data Types:
Imaging, Methylation, blood, clinical

Existing codebase/projects for processing/modeling imaging data.

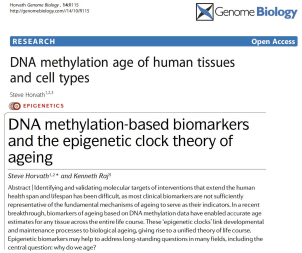
Devops Setup



Github (external)
AWS (internal data and modeling)

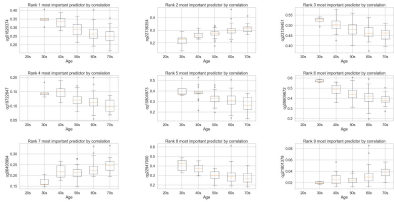
Team cooperation channel.

Literature review



Existing research on Methylation, blood test and aging provided direction on what are potential signatures we can look at.

EDA



We've completed EDA on the blood test data and methylation data within the PPMI dataset, following the inspiration from our literature reviews.

Current/Future project ideas

- **Data:**
  - Blood test data is not a promising avenue
  - Current focus methylation
  - EWAS datahub methylation data contains over 9000 HC samples
- **Analysis:**
  - Different tissue comparison
  - Healthy vs. unhealthy comparison
    - *Here we define unhealthy narrowly as individuals with neurodegenerative disease e.g. Huntington's, Parkinson's, Alzheimer's*
- **Modeling:** Explore more sophisticated ML models suggested by Merck and improve current literature
  - Multi-level perceptron, XGBoost, higher dimensionality in training data