Milestone 2

Merck - Healthy Aging

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

What's our topic?

We are studying on one of the most inevitable natural processes — **aging**. In the health industry, this has almost always been associated with diseases.

We are going to study the **signals of aging**, specifically **healthy aging**, by providing

- **an accurate age predictive model for healthy cohort**, as well as
- **comparative model analysis** for differences between healthy aging and unhealthy aging.

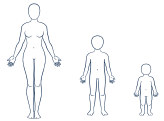
"Unhealthy" definition: neurodegenerative disease

Break our topic up to HEALTHY + AGING

Four step breakdown of our goal

01	Identifying Datasets	✔ What datasets are relevant to age prediction? ✔ Databases: PPMI (Parkinson), EWAS (Methylation) ✔ Different dataset types: MRI, Blood test, Methylation
02	Feature selection	✔ What features are most age related? ✔ How is supported by literature? ✔ Dimension reduction?
03	Age Predictive Modeling	✔ Produce an accurate model for age prediction within the healthy cohort. ✔ Model refinement and model selection.
04	Comparative Model Analysis	✔ Do different age cohorts and unhealthy/unhealthy cohorts behave differently, and if so how? ✔ Comparing results sampled from different tissues.

EWAS Data Hub



Roadmap

Milestone 2:

Milestone 2 has been dedicated to studying Methylation and its relationship with age for different tissues.

- EDA on Methylation data
- Methylation literature reviews.
- Age prediction model with feature selection
- Initial comparison between different tissues.

Milestone 1

- Literature review
- Database Exploration
- DevOps Setup
- Blood and Methylation data preliminary analysis

The explorations and preliminary studies helped us find a direction for healthy aging prediction: **Methylation.**

Final milstone

Goal 1. An accurate age prediction model
Goal 2: Comparative study of age signature between different tissues, age cohorts.
Goal 3: Comparative study on healthy vs unhealthy cohort.

Identifying Datasets

1. Dataset exploration;
2. Literature review.

1

Datasets suggested by Merck

Human Connectome	✔ Lifespan aging dataset, longitudinal study ✔ MRI data for 689 subjects ✔ 5TB total
ADNI Dataset	✔ MRI and clinical data on Alzheimer's patients ✔ 229 Healthy control subjects
PPMI	✔ Variety of data for Parkinson's patients ✔ Including clinical, methylation and MRI ✔ 241 Healthy control subjects
AD Knowledge Portal	✔ Variety of data for Alzheimer's patient ✔ Including genomics, blood and MRI ✔ Healthy control group sizes varying depending on study

CONNECTOME
COORDINATION FACILITY

ADNI
Alzheimer's Disease Neuroimaging Initiative

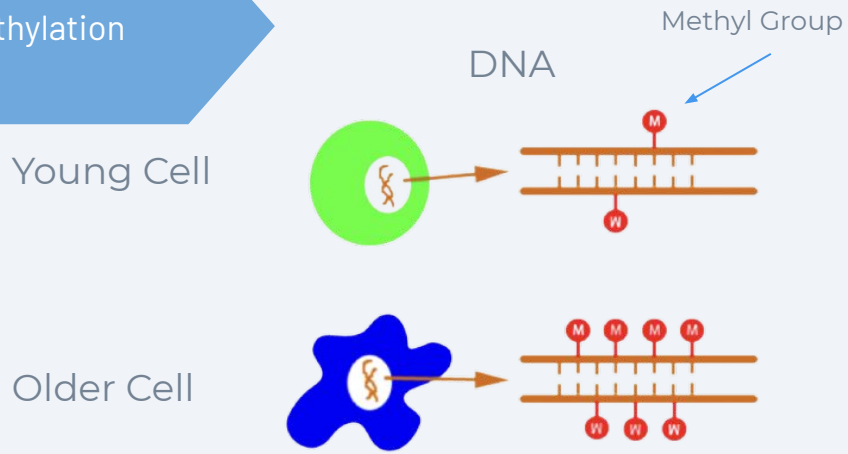
Parkinson's Progression Markers Initiative

AD Knowledge Portal

Literature review

2013

DNA methylation



Literature review

2013	DNA methylation	✔ DNA methylation age of human tissues and cell types ✔ Sites on an individual's DNA become methylated over times ✔ Amount of methylation changes with age

Literature review

2013	DNA methylation	✔ DNA methylation age of human tissues and cell types ✔ Sites on an individual's DNA become methylated over times ✔ Amount of methylation changes with age
2015	Blood chemistry	✔ Application of deep neural networks to biomarker development ✔ Blood chemistry results shown to be good predictors of chronological age ✔ Identified 5 blood markers which proved to be the best predictors of age

Literature review

2013	DNA methylation	✔ DNA methylation age of human tissues and cell types ✔ Sites on an individual's DNA become methylated over times ✔ Amount of methylation changes with age
2015	Blood chemistry	✔ Application of deep neural networks to biomarker development ✔ Blood chemistry results shown to be good predictors of chronological age ✔ Identified 5 blood markers which proved to be the best predictors of age
2020	MRI	✔ Identifying Morphological Indicators of Aging with NN on large-scale whole body MRI ✔ Predicted subject age from whole body MRI images

Other datasets

EWAS Data Hub

A data hub of DNA methylation array data and metadata

95,783
Samples

626
Tissues/cells

431
Diseases

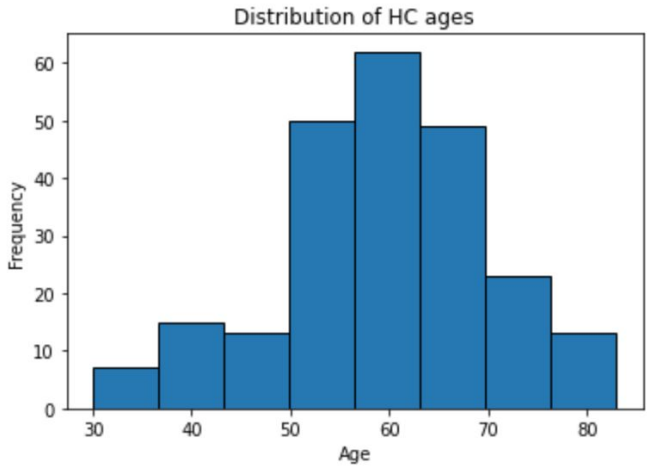
	Healthy Control	Alzheimer's	Parkinson's	Huntington's
Whole Blood	1802	299	400	N/A
Brain	1064	1510	10	406

Feature Selection

1. We investigated as potential features:
    a. Blood chemistry
    b. DNA Methylation

2

Blood data EDA

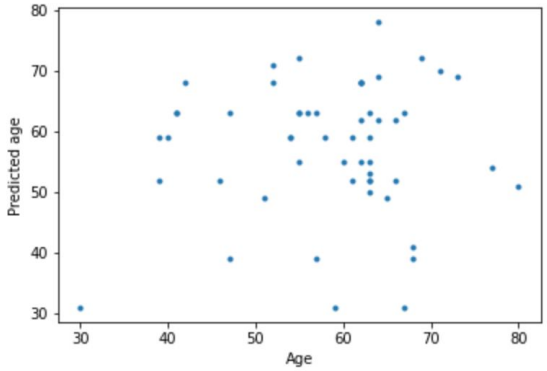


Distribution of healthy control (HC) ages

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

Pink: tests found to be most significant by the Putin study
Navy: tests found to be most significant by the Levine study

Blood data EDA

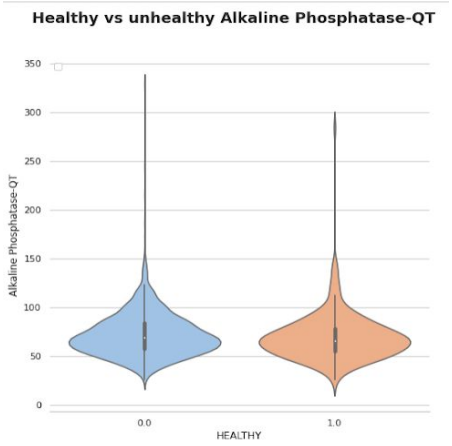


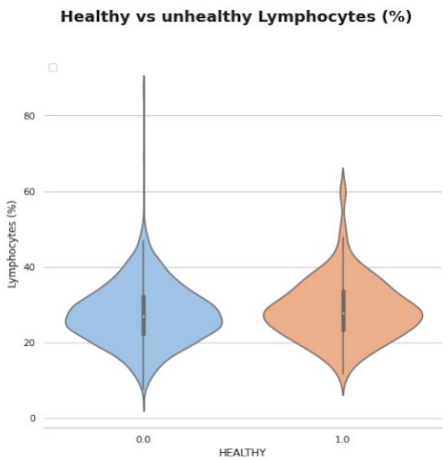
Comparison of the Linear Regression test set predictions vs. the true values

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

Pink: tests found to be most significant by the Putin study
Navy: tests found to be most significant by the Levine study

Blood data EDA





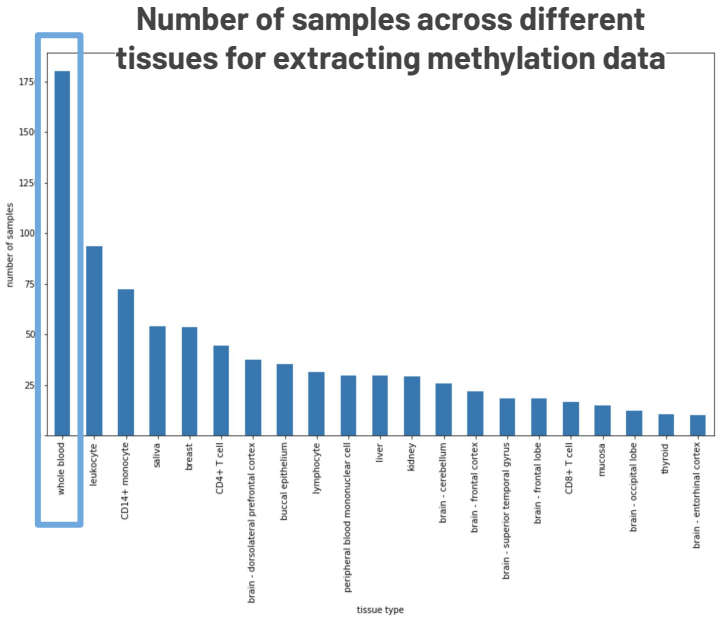
"Q

Is DNA methylation predictive of age?

We answer this question in the following analysis.

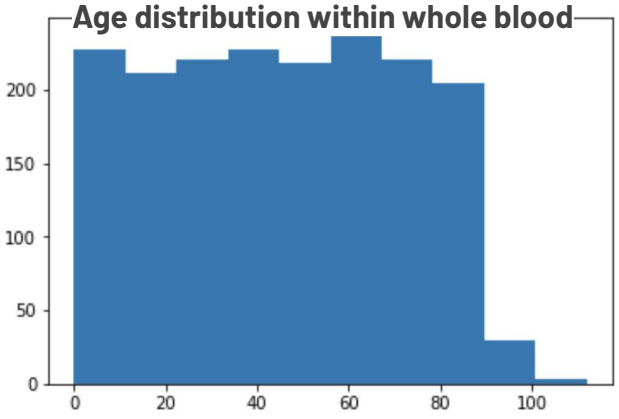
A"

Methylation data from whole blood



whole blood	1802
leukocyte	934
CD14+ monocyte	724
saliva	539
breast	535
CD4+ T cell	443
brain - dorsolateral prefrontal cortex	373
buccal epithelium	351
lymphocyte	313
peripheral blood mononuclear cell	294
liver	294
kidney	290
brain - cerebellum	257
brain - frontal cortex	219
brain - superior temporal gyrus	182
brain - frontal lobe	182
CD8+ T cell	164
mucosa	149
brain - occipital lobe	124
thyroid	105
brain - entorhinal cortex	100

Name: tissue, dtype: int64



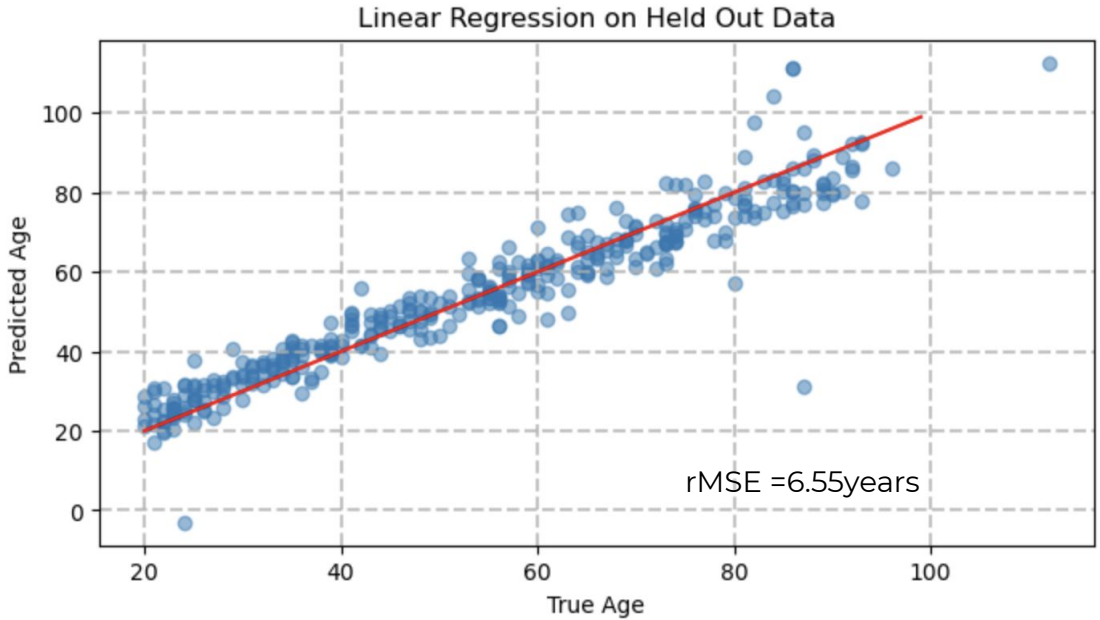
Whole blood had the most samples, it is also the most widely used in industry, and has a very even age distribution.

What does the data look like?

sample_id	tissue	age	cg02494853	cg03706273	cg04023335	cg05213048	cg15295597	cg26520468	cg27539833	cg00008
GSM2334366	whole blood	94	0.078	0.205	0.139	0.904	0.120	0.970	0.912	0.
GSM989863	whole blood	101	0.013	0.008	0.117	0.756	0.033	0.958	0.933	0.
GSM1443696	whole blood	99	0.013	0.017	0.477	0.715	0.017	0.966	0.932	0.
GSM1069241	whole blood	99	0.013	0.017	0.477	0.715	0.017	0.966	0.932	0.
GSM1572442	whole blood	112	0.036	0.255	0.260	0.690	0.065	0.983	0.951	0.
...	...	...	...	...	...	...	...	...	...	
GSM1498536	whole blood	48	0.010	0.048	0.068	0.575	0.034	0.981	0.946	0.
GSM1868331	whole blood	48	0.024	0.019	0.635	0.848	0.035	0.958	0.944	0.
GSM2337452	whole blood	48	0.027	0.032	0.145	0.661	0.068	0.964	0.936	0.
GSM1653326	whole blood	48	0.033	0.023	0.529	0.772	0.064	0.956	0.946	0.
GSM1871289	whole blood	48	0.019	0.024	0.166	0.599	0.048	0.952	0.949	0.

1066 rows × 406628 columns

Modeling with all 406,628 features



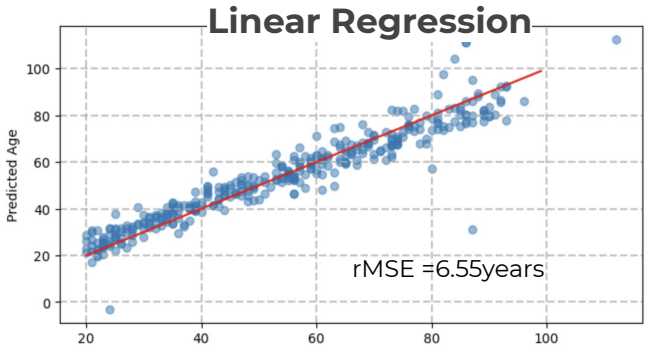
Tissue: Whole blood

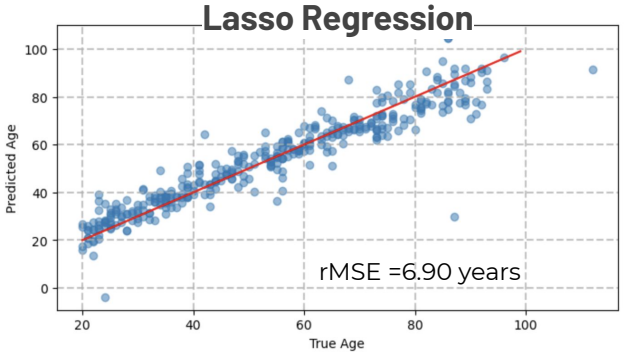
Age: > 20

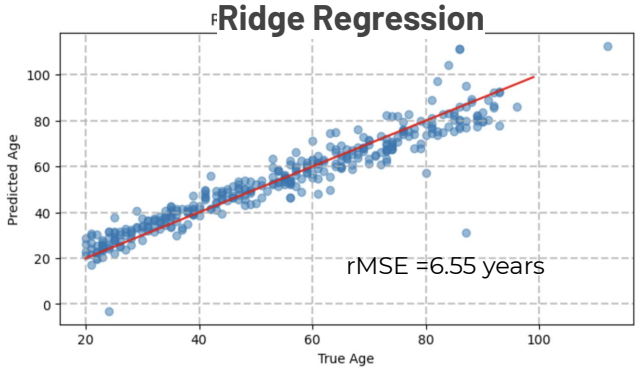
Column filters:
NaNs < 25%

Total number of features:
406,628 features

Modeling with all 406,628 features







Tissue: Whole blood

Age: > 20

Column filters:
NaNs < 25%

Total number of features:
406,628 features

Held out data

"Q Can we use fewer features?

1. Statistical modeling
2. Cross validation with XGboost.

A"

	feature 1	feature 2	feature 3
sample 1	0.2	0.2	0.2
sample 2	0.7	0.7	0.7
sample 3	0.1	0.1	0.1
sample 4	0.3	0.3	0.3
sample 5	0.3	0.3	0.3
sample 6	0.8	0.8	0.8
sample 7	0.5	0.5	0.5
sample 8	0.4	0.4	0.4
sample 9	0.1	0.1	0.1
sample 10	0.2	0.2	0.2

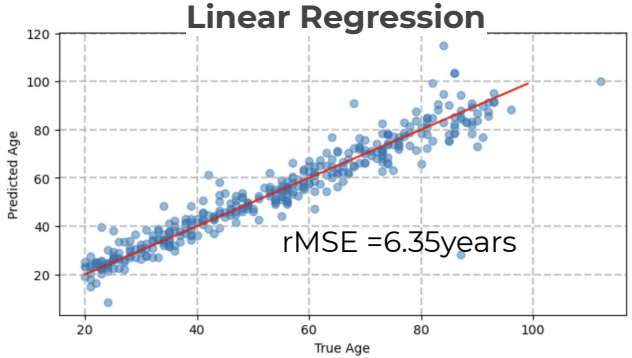
- Optimize XGboost model with all features
- Cycle for 50 cycles
  - Randomly select 50% of the samples
  - Fit data with an XGboost model
  - Record importance scores
- Determine which features most often occur in the top 100 importance scores
- Select the features that appear most often

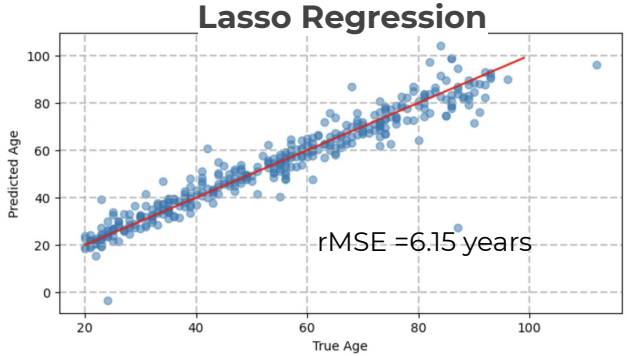
Age Predictive Modeling

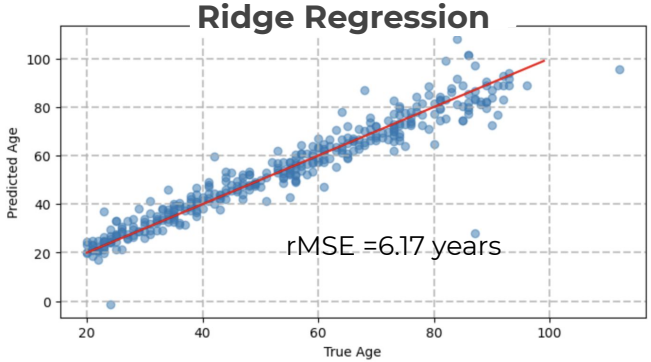
1. Examining simple models with 100 features
2. Examining NN models with ~1300 features

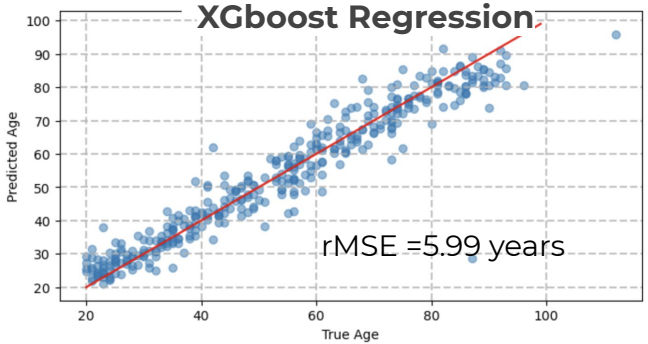
3

Modeling with the top features









Top 100 features.

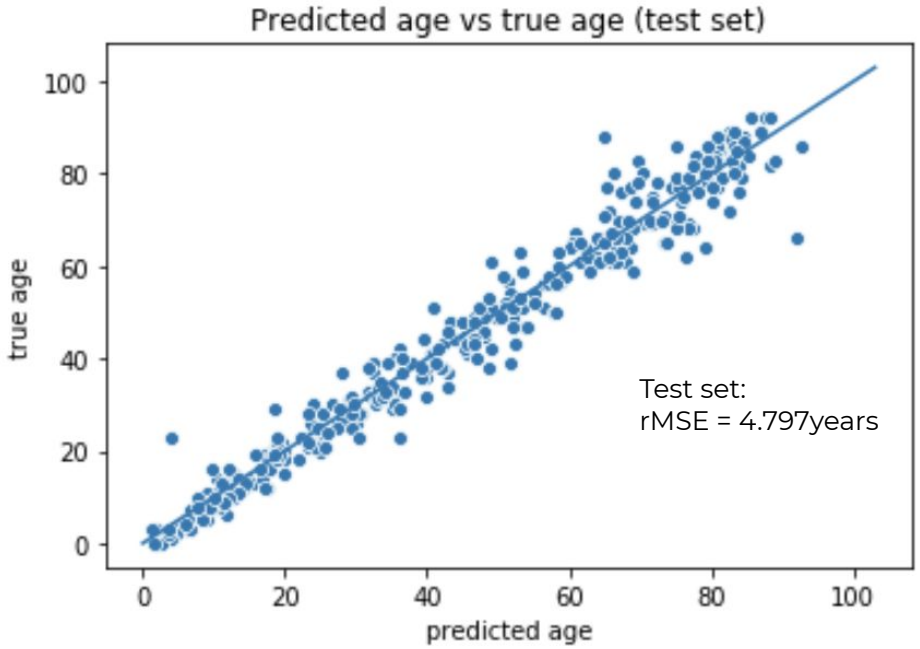
Tissue: Whole blood

Age: > 20

Column filters:
NaNs < 25%

Held out data

Modeling with top features (test set results)



Modeling with top ~1300 generated by Xgboost.

Tissue: Whole blood

Model used: Two hidden layer fully connected NN.

Testset rMSE/years:
4.797

Comparative Modeling

1. Comparing different tissues,
2. Comparing healthy and unhealthy (future work)

4

Is our blood model transferrable to other tissues?

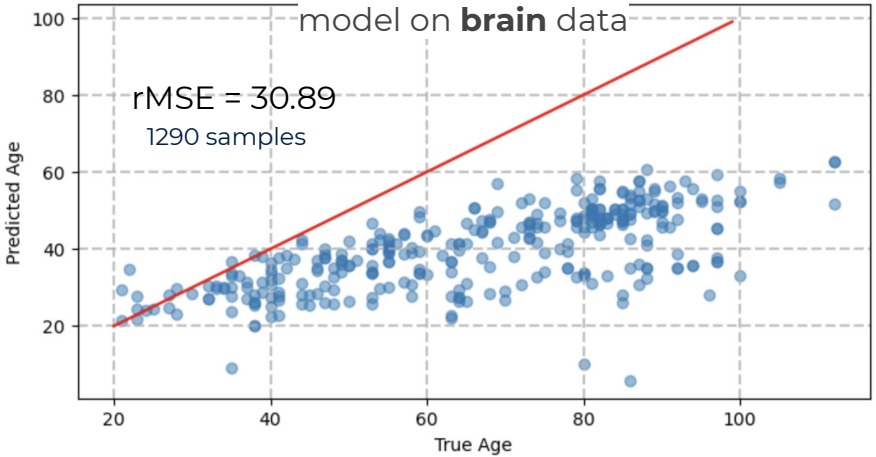
EWAS Tissue types

whole blood	1802
leukocyte	934
CD14+ monocyte	724
saliva	539
breast	535
CD4+ T cell	443
brain - dorsolateral prefrontal cortex	373
buccal epithelium	351
lymphocyte	313
peripheral blood mononuclear cell	294
liver	294
kidney	290
brain - cerebellum	257
brain - frontal cortex	219
brain - superior temporal gyrus	182
brain - frontal lobe	182
CD8+ T cell	164
mucosa	149
brain - occipital lobe	124
thyroid	105
brain - entorhinal cortex	100

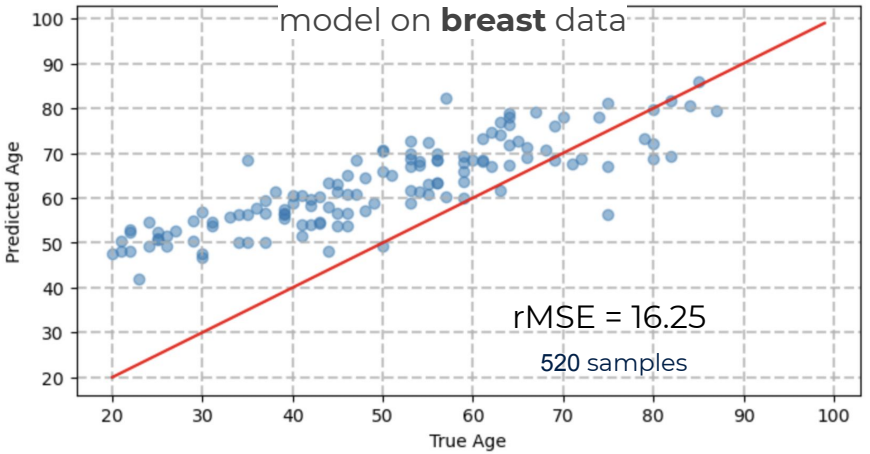
Name: tissue, dtype: int64

"Q Are our blood-based models transferrable to other tissues?

Predicting age using blood trained model on **brain** data



Predicting age using blood trained model on **breast** data



No A"

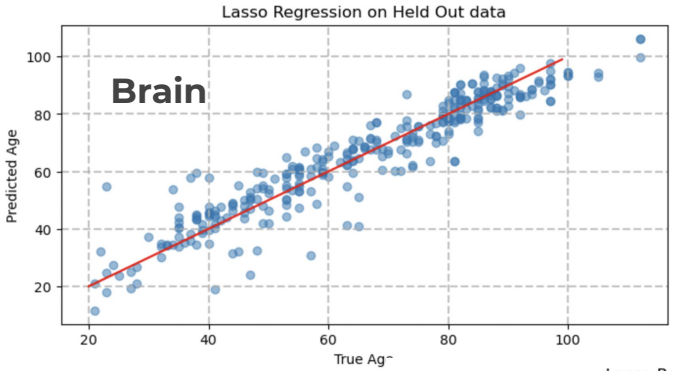
"Q

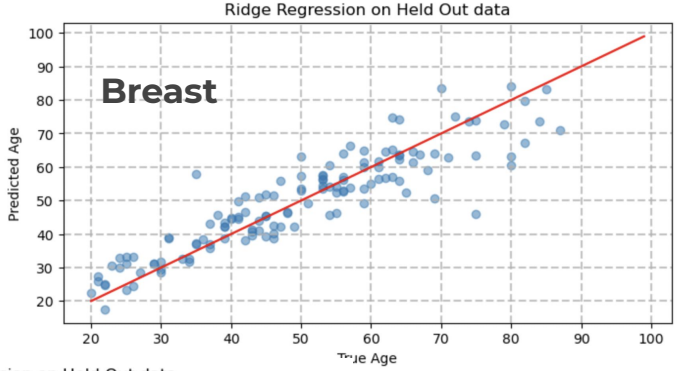
Can methylation data from tissues besides blood be used to predict age?

Yes, but they are not as good.

A"

Age prediction models in other tissues





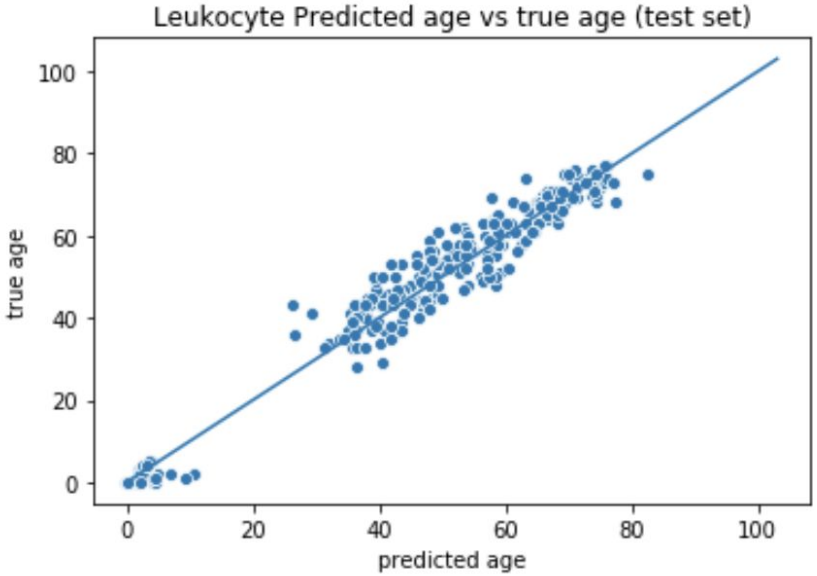
Modeling with top 100 cpg features from their perspective tissues.

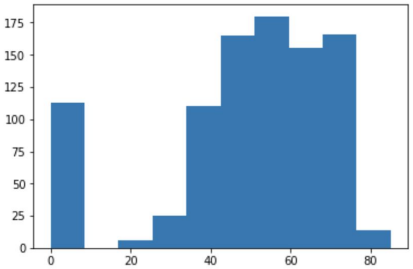
Brain:
Lasso Regression
rMSE: 6.9

Breast:
Ridge Regression
rMSE: 7.03

Blood:
Lasso Regression
rMSE: 6.15

Age prediction models in other tissues





Modeling with cpg features produced by the blood xgboost CV.

Tissue: Leukocyte

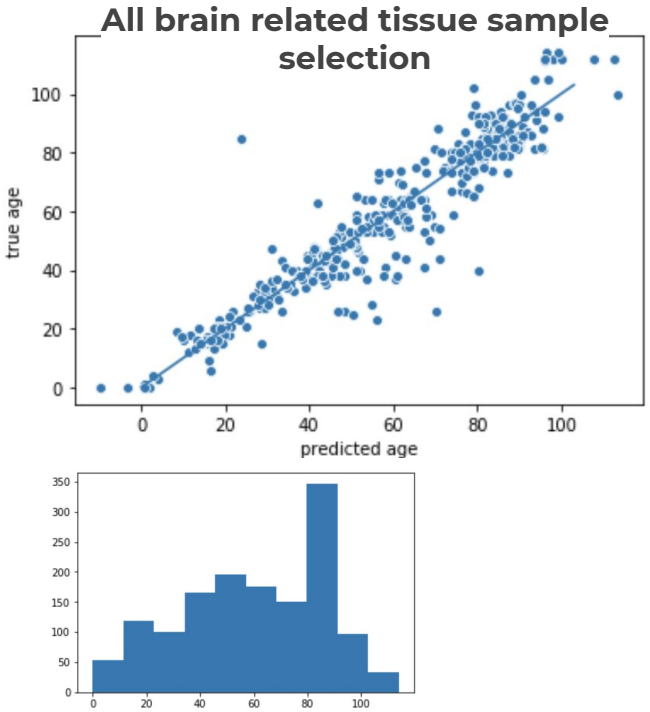
2 hidden layer, fully connected NN

rMSE: 4.5762

Age prediction models in other tissues

whole blood	1802
leukocyte	934
CD14+ monocyte	724
saliva	539
breast	535
CD4+ T cell	443
brain - dorsolateral prefrontal cortex	373
buccal epithelium	351
lymphocyte	313
peripheral blood mononuclear cell	294
liver	294
kidney	290
brain - cerebellum	257
brain - frontal cortex	219
brain - superior temporal gyrus	182
brain - frontal lobe	182
CD8+ T cell	164
mucosa	149
brain - occipital lobe	124
thyroid	105
brain - entorhinal cortex	100

Name: tissue, dtype: int64



Modeling with top ~1300 cpg features produced by the xgboost CV.

Tissue: All brain

Sample number: 1437

2 hidden layer, fully connected NN

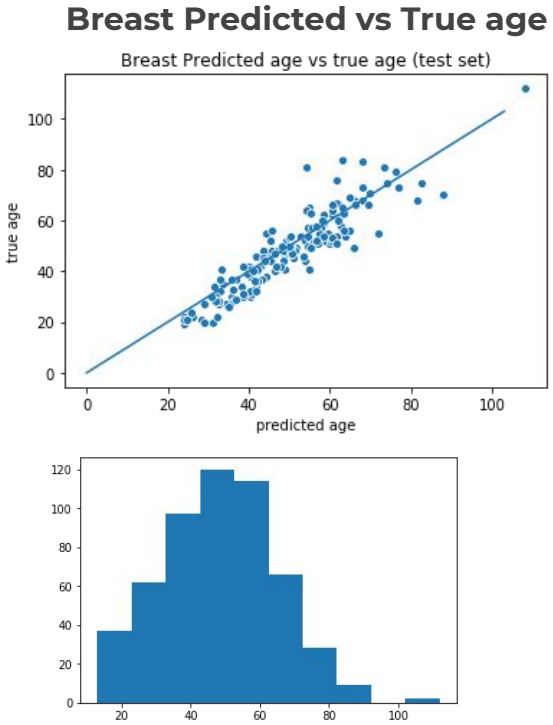
rMSE: 8.525

Different parts of brain tissues have different behavior and cpg sites.

Age prediction models in other tissues

whole blood	1802
leukocyte	934
CD14+ monocyte	724
saliva	539
breast	535
CD4+ T cell	443
brain - dorsolateral prefrontal cortex	373
buccal epithelium	351
lymphocyte	313
peripheral blood mononuclear cell	294
liver	294
kidney	290
brain - cerebellum	257
brain - frontal cortex	219
brain - superior temporal gyrus	182
brain - frontal lobe	182
CD8+ T cell	164
mucosa	149
brain - occipital lobe	124
thyroid	105
brain - entorhinal cortex	100

Name: tissue, dtype: int64



Modeling with top ~1300 cpg features produced by the xgboost CV.

Tissue: breast

Sample number: 535

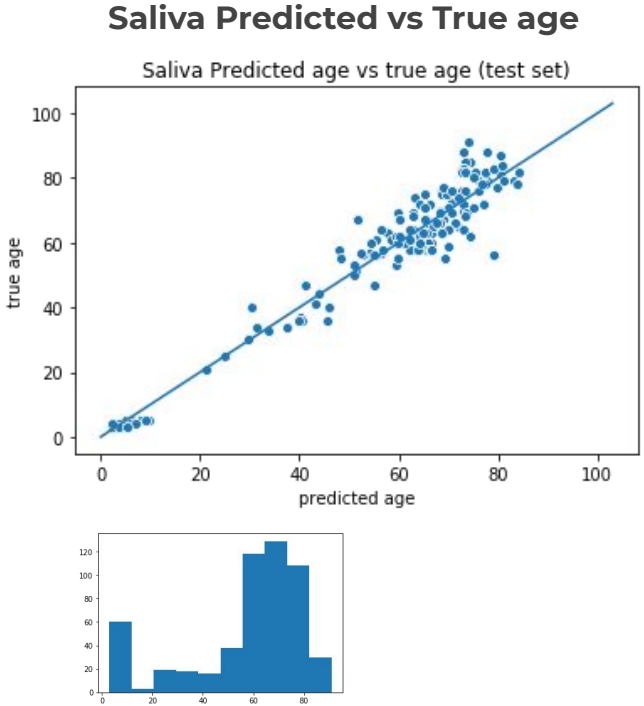
2 hidden layer, fully connected NN

rMSE: 6.605131863

Age prediction models in other tissues

whole blood	1802
leukocyte	934
CD14+ monocyte	724
saliva	539
breast	535
CD4+ T cell	443
brain - dorsolateral prefrontal cortex	373
buccal epithelium	351
lymphocyte	313
peripheral blood mononuclear cell	294
liver	294
kidney	290
brain - cerebellum	257
brain - frontal cortex	219
brain - superior temporal gyrus	182
brain - frontal lobe	182
CD8+ T cell	164
mucosa	149
brain - occipital lobe	124
thyroid	105
brain - entorhinal cortex	100

Name: tissue, dtype: int64



Modeling with top ~1300 cpg features produced by the xgboost CV.

Tissue: breast

Sample number: 539

2 hidden layer, fully connected NN

rMSE: 6.0209

"Q Do tissues share common important cpg sites?

Top 100 cpgs

- Common between blood and brain 4
- Common between blood and breast 7
- Common between brain and breast 10
- Common to all 3 tissues 2

Do different tissues within the brain share common methylation sites?

After filtering 40,000+ features by the 2462 features ones XGboost initially gave back.

Cerebellum is left with 1052 non NA columns;
Prefrontal cortex is left with 560
Overlap with no NAs: 438

Very few. Difference sites are most predictive of age in different tissues A"

Past Challenges and Failures

Database Identification

We spent considerable amount of time finding high quality datasets, and testing which types of data are good age predictors.

Blood test data not working out

We weren't able to replicate literature review results on blood test within the PPMI dataset.

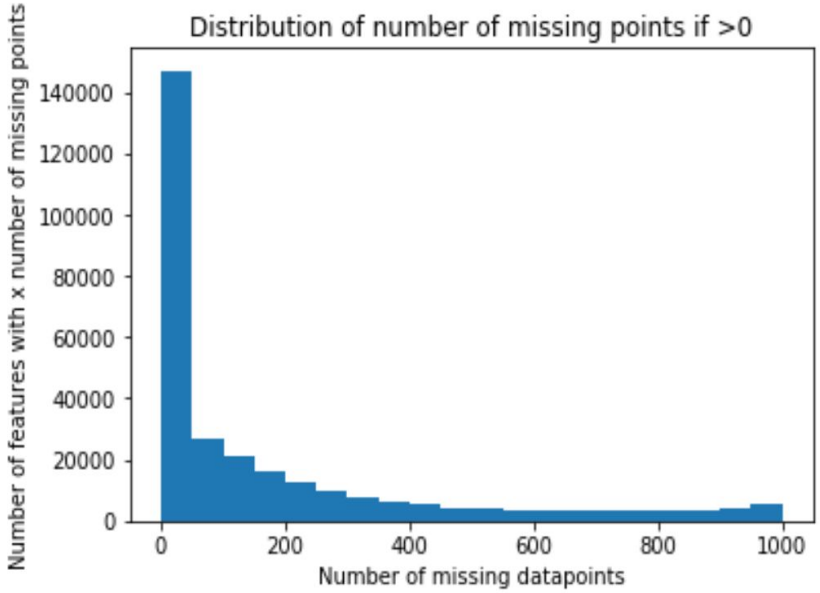
Methylation data size

~400,000 columns to load and process

NA data

Many columns with NA in methylation data, and different tissue types had different NA columns.

Dataset with 1000 samples and all sites



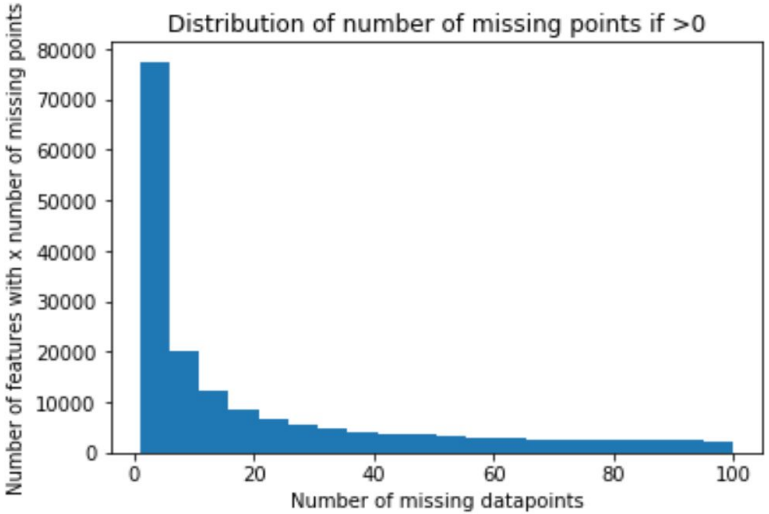
- 485,242 columns
- 60.45% of columns have missing data
- 175.5 missing values per column, on average
- 275 columns with no data

Impute mean on all data

- Linear regression on top 100 features selected by XGBoost
- Baseline model- columns with any NaN removed (293,218)
- Noticeably better performance on test set

	Train rmse	Test rmse	Saved rmse	Train r^2	Test r^2	Saved r^2
Baseline model	3.940	5.878	5.674	0.924	0.815	0.833
Mean on all data	3.848	5.211	5.640	0.927	0.854	0.835

Dropped columns with >10% missing



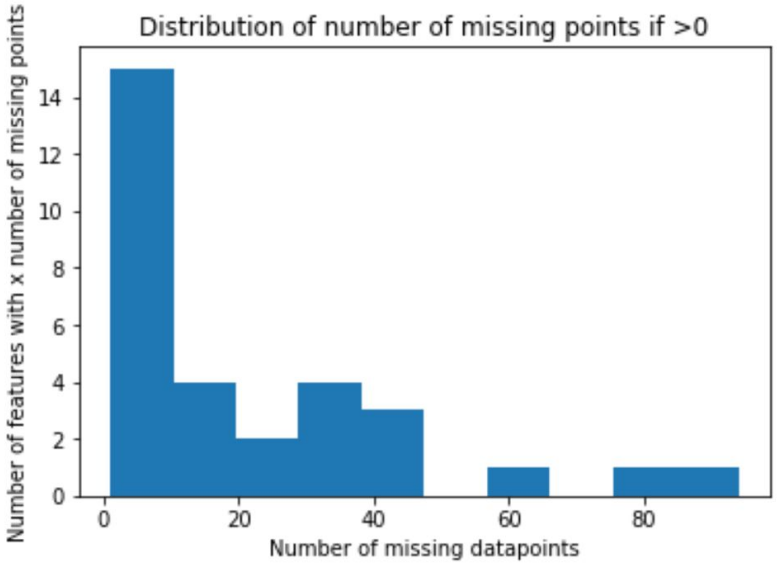
- Dropped 119,610 columns
- 47.5% of columns have missing data
- 20.5 missing values per column, on average

Impute mean on subset of data

- Same model type
- Improvement on saved data
- Only 13 sites in common with baseline method

	Train rmse	Test rmse	Saved rmse	Train r^2	Test r^2	Saved r^2
Baseline model	3.940	5.878	5.674	0.924	0.815	0.833
Mean on all data	3.848	5.211	5.640	0.927	0.854	0.835
Mean on >25% missing removed	3.885	5.562	5.529	0.926	0.834	0.841
Mean on >10% missing removed	4.145	5.160	5.310	0.916	0.857	0.853

Iterative imputer



- Dataset with 100 features chosen from last model
- 31% of columns have missing data
- 20.2 missing values per column, on average

	Train rmse	Test rmse	Saved rmse	Train r^2	Test r^2	Saved r^2
Baseline model	3.940	5.878	5.674	0.924	0.815	0.833
Mean on all data	3.848	5.211	5.640	0.927	0.854	0.835
Mean on >25% missing removed	3.885	5.562	5.529	0.926	0.834	0.841
Mean on >10% missing removed	4.145	5.160	5.310	0.916	0.857	0.853
Ridge Regression	4.155	5.169	5.293	0.915	0.857	0.854
Decision Tree	4.151	5.113	5.308	0.915	0.860	0.853
Extra Trees	4.148	5.154	5.326	0.915	0.857	0.852
KNN Regression n=15	4.149	5.165	5.301	0.915	0.857	0.854

Where are we now?

		COMPLETED	ONGOING	FUTURE PLANS
01	Identifying Datasets	✔ EDA on PPMI (Parkinson), EWAS (Methylation) ✔ Literature review		
02	Feature selection	✔ Narrowed our features to Methylation data ✔ Completed cpg feature selection	Computation and memory issue for large datasets.	
03	Age Predictive Modeling	✔ Completed modeling with Xgboost, linear regression and NN	Refine models; Impute NAs.	Refine models; Biological interpretation of models
04	Comparative Model Analysis	✔ Compared model transferability and feature transferability ✔ Trained separate models for subset of different tissues	Compare different tissues.	Compare Healthy and Unhealthy cohorts; Interpretation of results.

Thank you

Roadmap

Milestone 2:

Milestone 2 has been dedicated to studying Methylation and its relationship with age for different tissues.

- EDA on Methylation data
- Methylation literature reviews.
- Age prediction model with feature selection
- Initial comparison between different tissues.

Milestone 1

- Literature review
- Database Exploration
- DevOps Setup
- Blood and Methylation data preliminary analysis

The explorations and preliminary studies helped us find a direction for healthy aging prediction: **Methylation.**

Final milstone

Goal 1. An accurate age prediction model
Goal 2: Comparative study of age signature between different issues, age cohorts.
Goal 3: Comparative study on healthy vs unhealthy cohort.