CHRONOLOGICAL AGE ESTIMATION FROM DNA METHYLATION DATA

- a case study in regularized linear regression

ELTE EÖTVÖS LORÁND UNIVERSITY

Leiden University, Honours Class Data Science

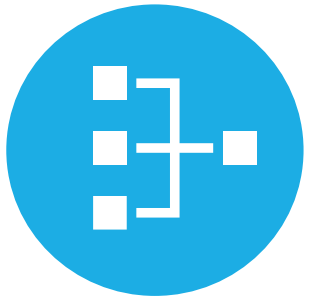
Orsolya Pipek | 4 April 2022

A SHORT OVERVIEW

CHRONOLOGICAL AGE VS. BIOLOGICAL AGE

DNA METHYLATION

LINEAR REGRESSION WITH REGULARIZATION

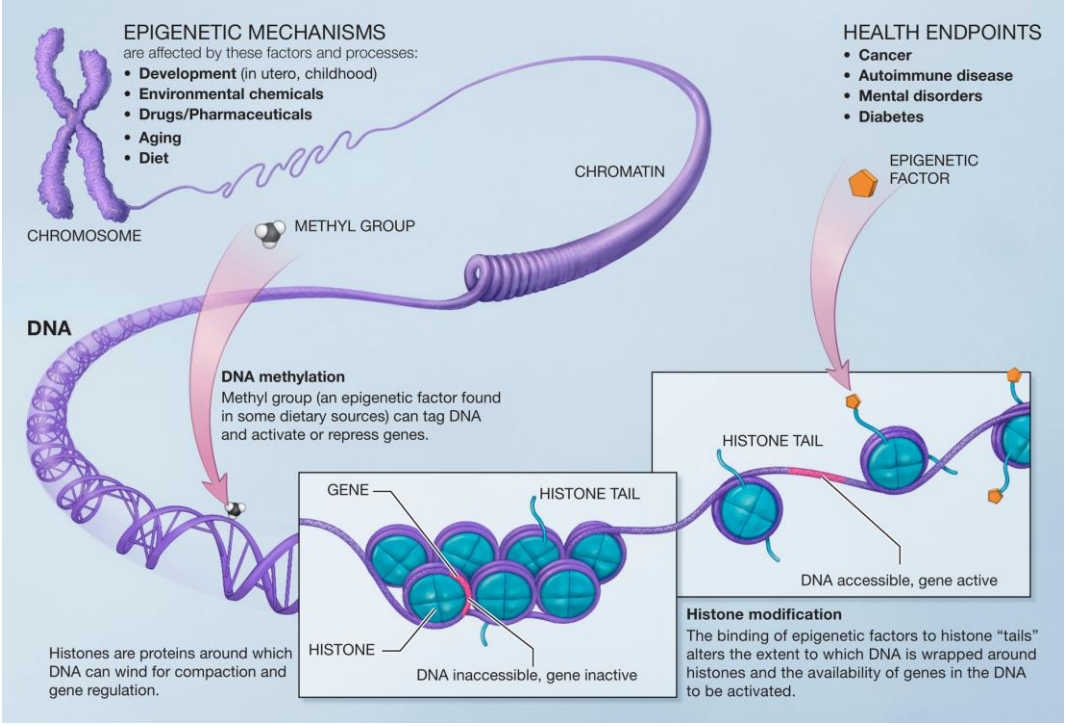


CHRONOLOGICAL AGE PREDICTION

CHRONOLOGICAL/BIOLOGICAL AGE

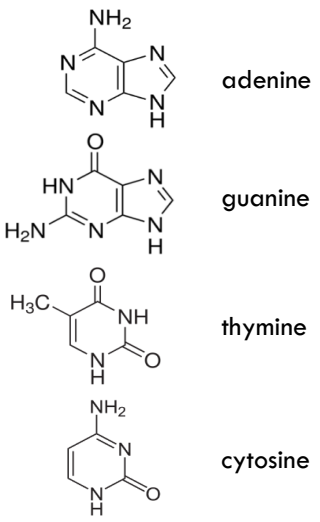
- **Chronological age:** time elapsed since birth
- **Biological age:** captures the different rate of functional deterioration across individuals
  - Aging has many manifestations (greying hair, wrinkles, reduced mobility, etc.)
  - Many manifestations are **epigenetic** (e.g. **DNA methylation**)

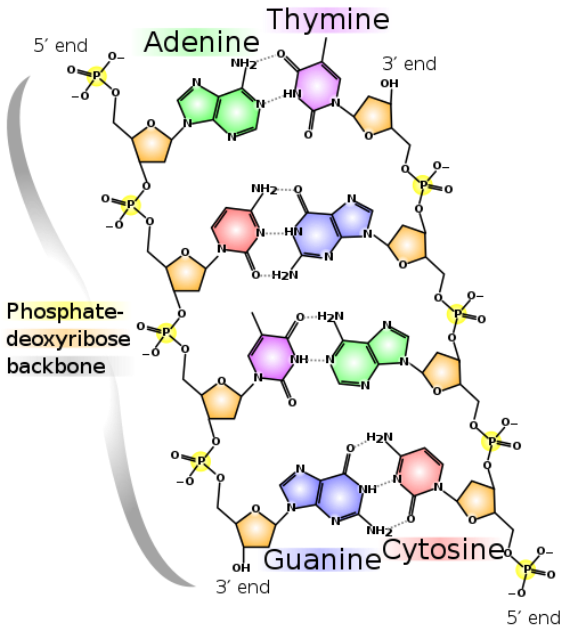
Epigenetic changes: changes **functionally effecting** the genome that **do not involve a change in the nucleotide sequence** of the DNA

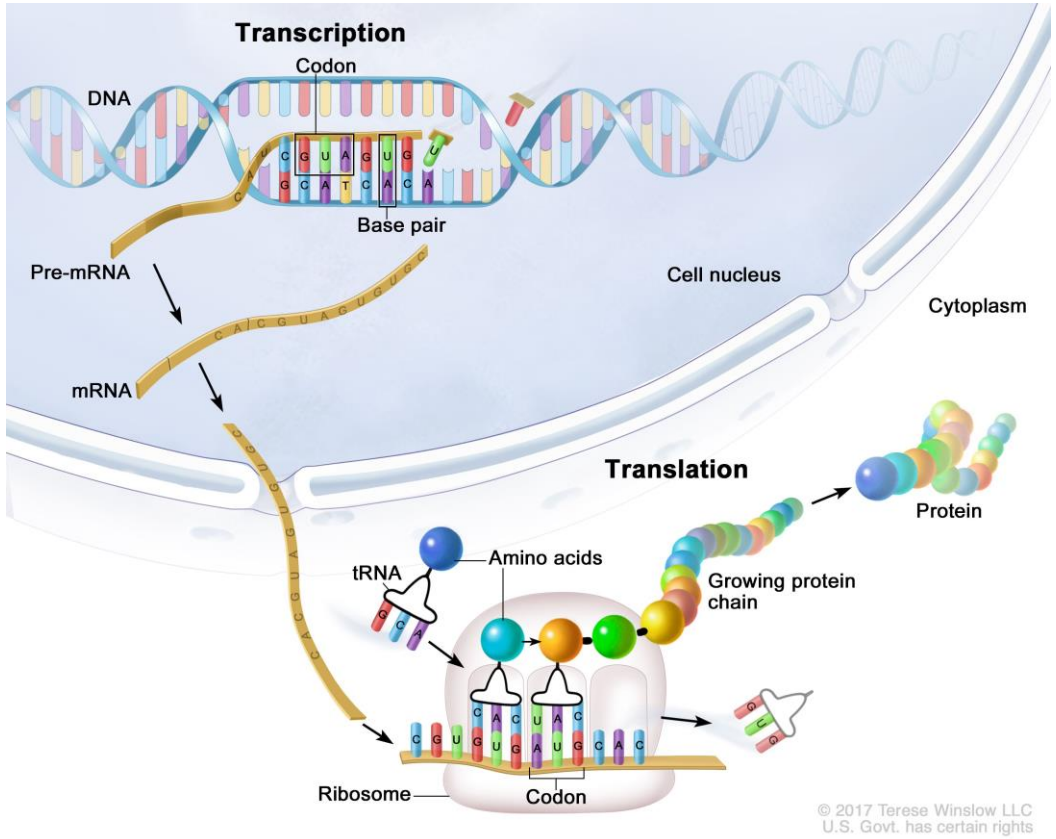


NIH – National Institutes of Health

THE IMPORTANCE OF DNA





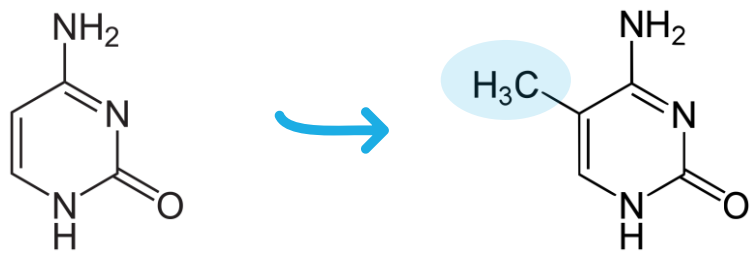


nucleobases

base-pairs,
double-stranded helix structure

transcription,
translation,
proteins

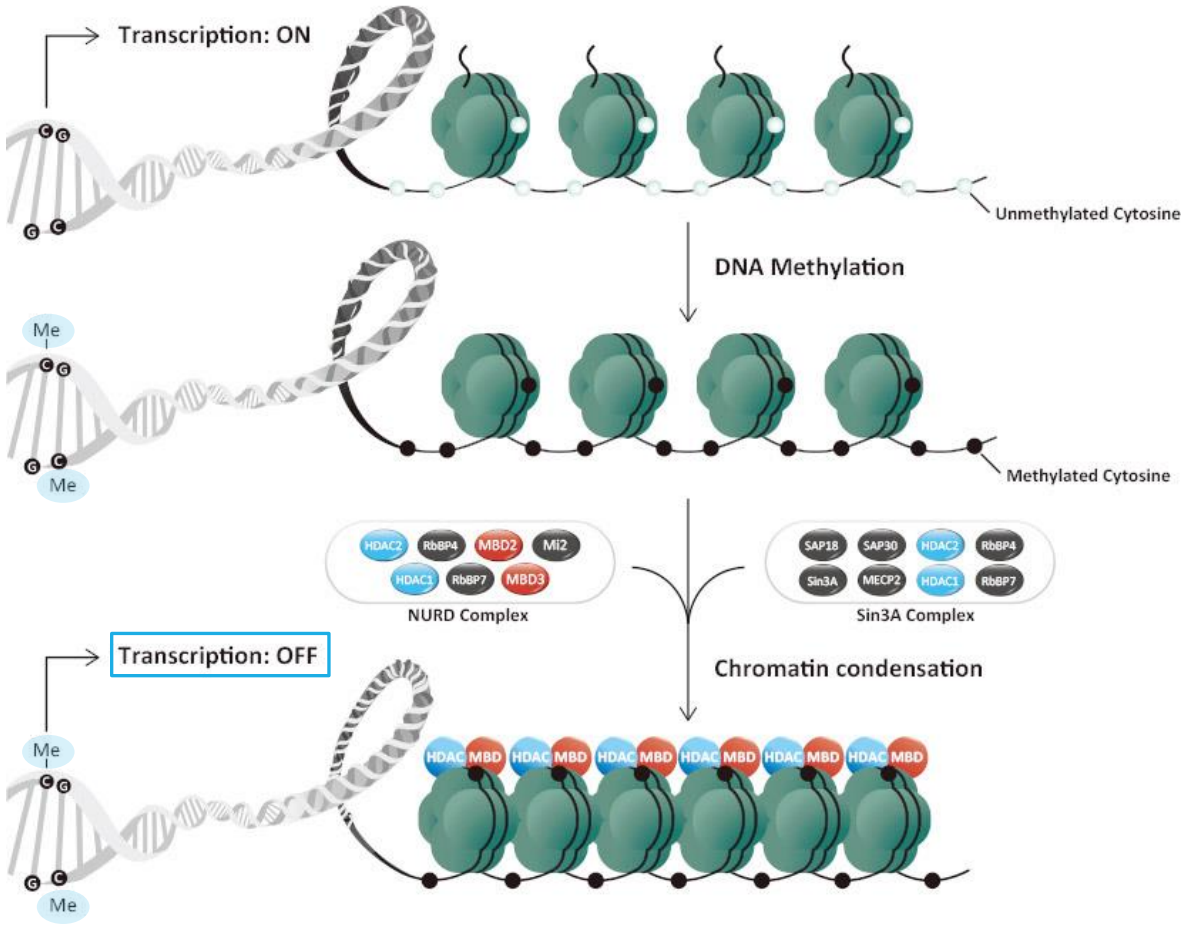
DNA METHYLATION (DNAm)



cytosine 5-methylcytosine

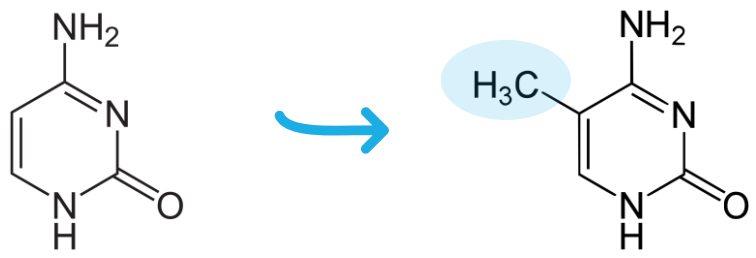
(almost exclusively in CpG dinucleotides)

DNA methylation has a regulatory role by affecting the accessibility of different genomic regions, thus influencing transcription (~ activity)



https://www.genetex.com/Research/Overview/epigenetics/dna_methylation

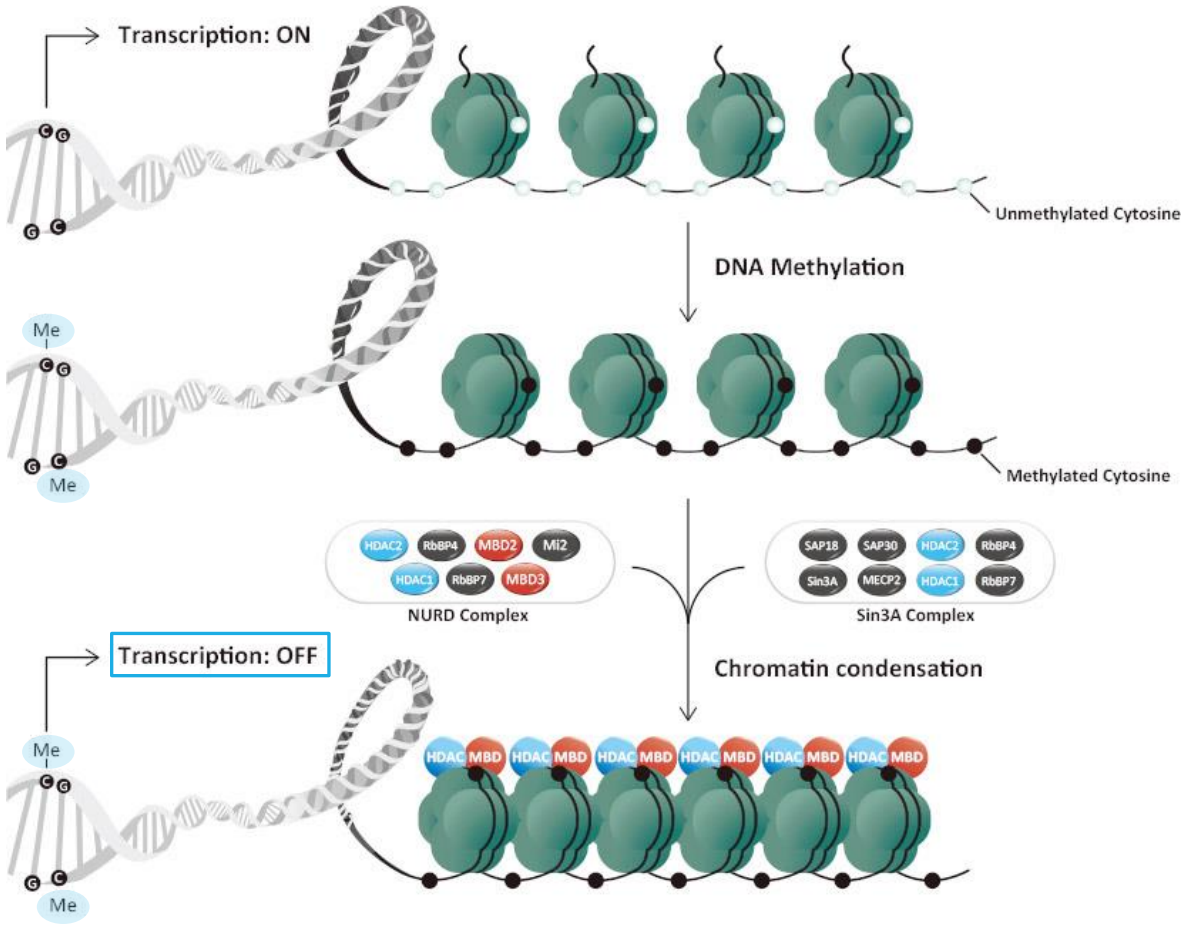
DNA METHYLATION (DNAm)



cytosine 5-methylcytosine

(almost exclusively in CpG dinucleotides)

DNA methylation has a regulatory role by affecting the accessibility of different genomic regions, thus influencing transcription (~ activity)



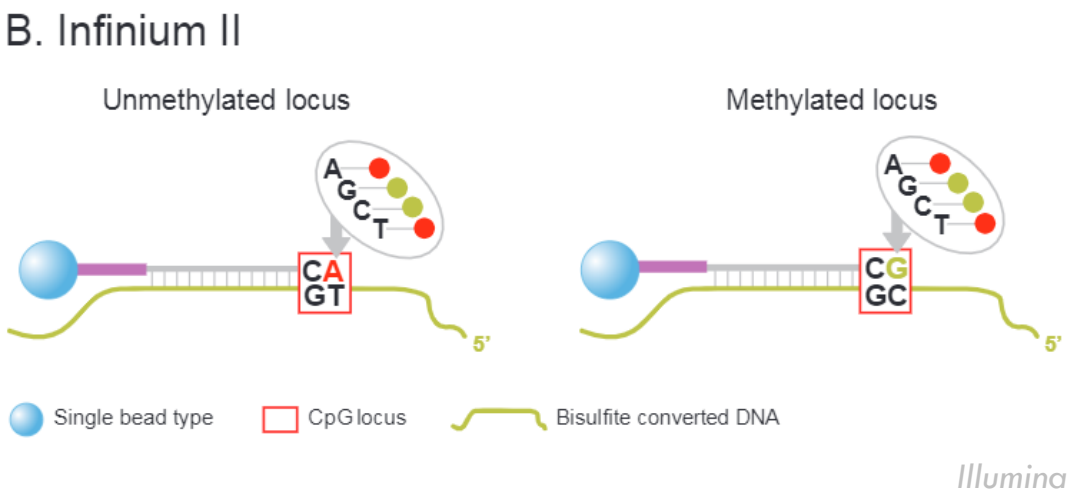
https://www.genetex.com/Research/Overview/epigenetics/dna_methylation

… but generally, it's much more complex

DNA METHYLATION (DNAm)

How is it measured? – a simplified picture

1. Bisulfite conversion
   - cytosine ⟶ uracil ⟶ thymine
   - methyl-cytosine ⟶ cytosine
2. Amplification
3. Attachment of single DNA strands to probe sequences terminating right before the target site
4. Trying to elongate probe sequence with either A/T or G/C nucleotides labelled with different colors
5. Measuring resulting light intensities



$$\beta = \frac{I_{meth}}{(I_{meth} + I_{unmeth} + \alpha)}$$

⟶ a single β-value for each genomic position of interest

LINEAR REGRESSION

~ ”**fit a line** to the data (sometimes in many-many dimensions) that describes it well enough”

LINEAR REGRESSION

~ ”**fit a line** to the data (sometimes in many-many dimensions) that describes it well enough”

… or more precisely:

- Given a $\left\{x_{i_j}\right\}_{i\in\{1,2,\ldots s\}}^{j\in\{1,2,\ldots k\}}$ set of $\boldsymbol{k}$ **inputs/independent variables/features/predictors**
- for $\boldsymbol{s}$ **samples/observations,**
- find the b_j **coefficients** of the (linear) function $f(\boldsymbol{x_i}) = b_0 + \sum_{j=1}^{k} b_j x_{i_j} = \hat{y}_i$
- that **minimizes the „prediction error”** defined as $\sum_{i=1}^{s}(y_i - \hat{y}_i)^2$ for the **output/dependent variable/response** of $\boldsymbol{y} = (y_1, y_2, \ldots y_s)$.

LINEAR REGRESSION

~ ”**fit a line** to the data (sometimes in many-many dimensions) that describes it well enough”

… or more precisely:

- Given a $\left\{x_{i_j}\right\}_{i\in\{1,2,\ldots s\}}^{j\in\{1,2,\ldots k\}}$ set of $\boldsymbol{k}$ **inputs/independent variables/features/predictors**
- for $\boldsymbol{s}$ **samples/observations,**
- find the b_j **coefficients** of the (linear) function $f(\boldsymbol{x_i}) = b_0 + \sum_{j=1}^{k} b_j x_{i_j} = \hat{y}_i$
- that **minimizes the „prediction error”** defined as $\sum_{i=1}^{s}(y_i - \hat{y}_i)^2$ for the **output/dependent variable/response** of $\boldsymbol{y} = (y_1, y_2, \ldots y_s)$.

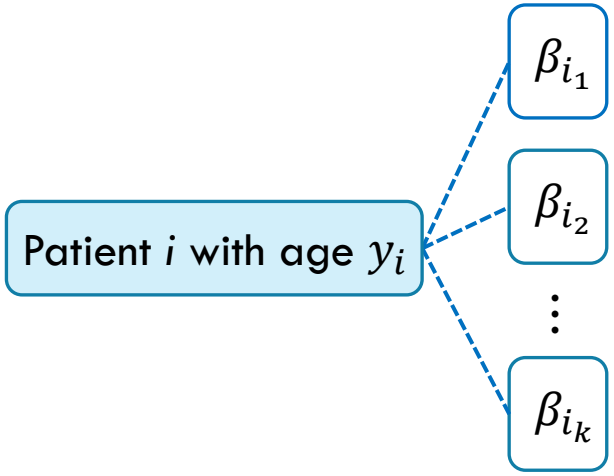
~ ”method of ordinary least squares”

LINEAR REGRESSION – IN OUR CASE

- **inputs/independent variables/features/predictors:** methylation β-values in targeted genomic positions
- **samples/observations:** patients
- **output/dependent variable/response:** chronological age

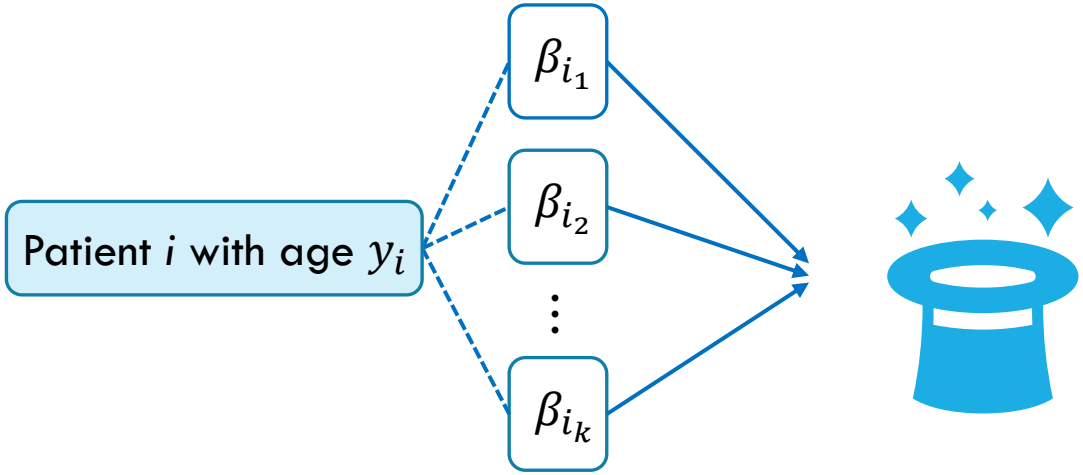
LINEAR REGRESSION – IN OUR CASE

- **inputs/independent variables/features/predictors:** methylation β-values in targeted genomic positions
- **samples/observations:** patients
- **output/dependent variable/response:** chronological age



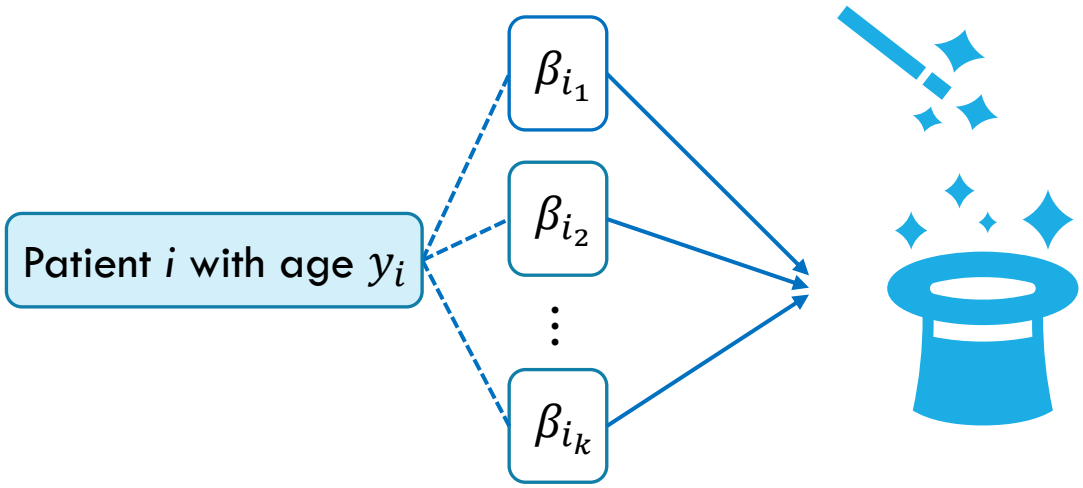
LINEAR REGRESSION – IN OUR CASE

- **inputs/independent variables/features/predictors:** methylation β-values in targeted genomic positions
- **samples/observations:** patients
- **output/dependent variable/response:** chronological age



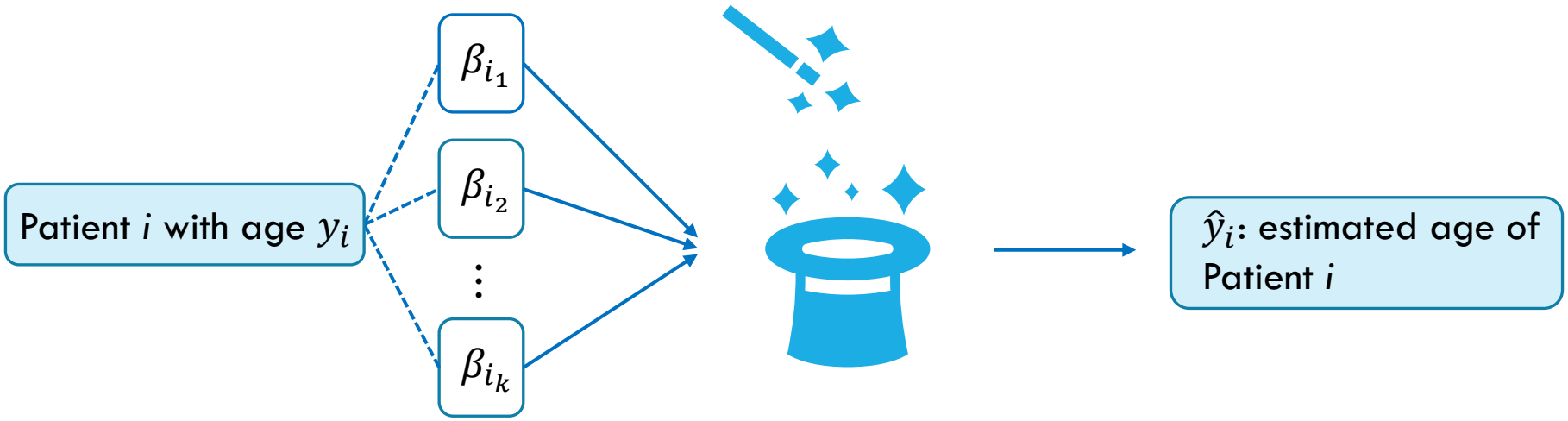
LINEAR REGRESSION – IN OUR CASE

- **inputs/independent variables/features/predictors:** methylation β-values in targeted genomic positions
- **samples/observations:** patients
- **output/dependent variable/response:** chronological age



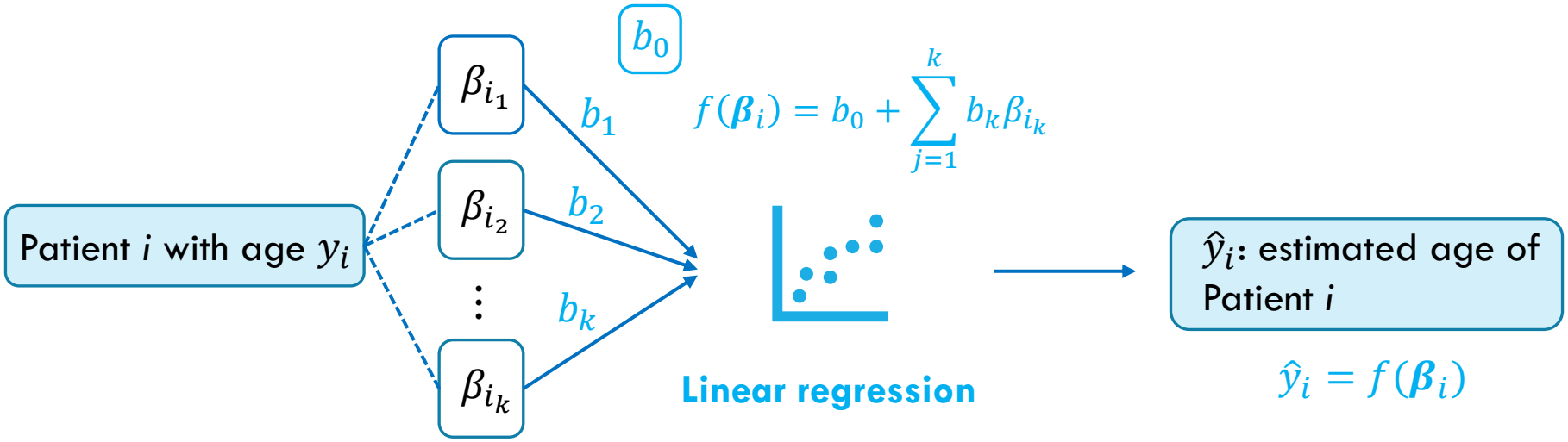
LINEAR REGRESSION – IN OUR CASE

- **inputs/independent variables/features/predictors:** methylation β-values in targeted genomic positions
- **samples/observations:** patients
- **output/dependent variable/response:** chronological age



LINEAR REGRESSION – IN OUR CASE

- **inputs/independent variables/features/predictors:** methylation β-values in targeted genomic positions
- **samples/observations:** patients
- **output/dependent variable/response:** chronological age

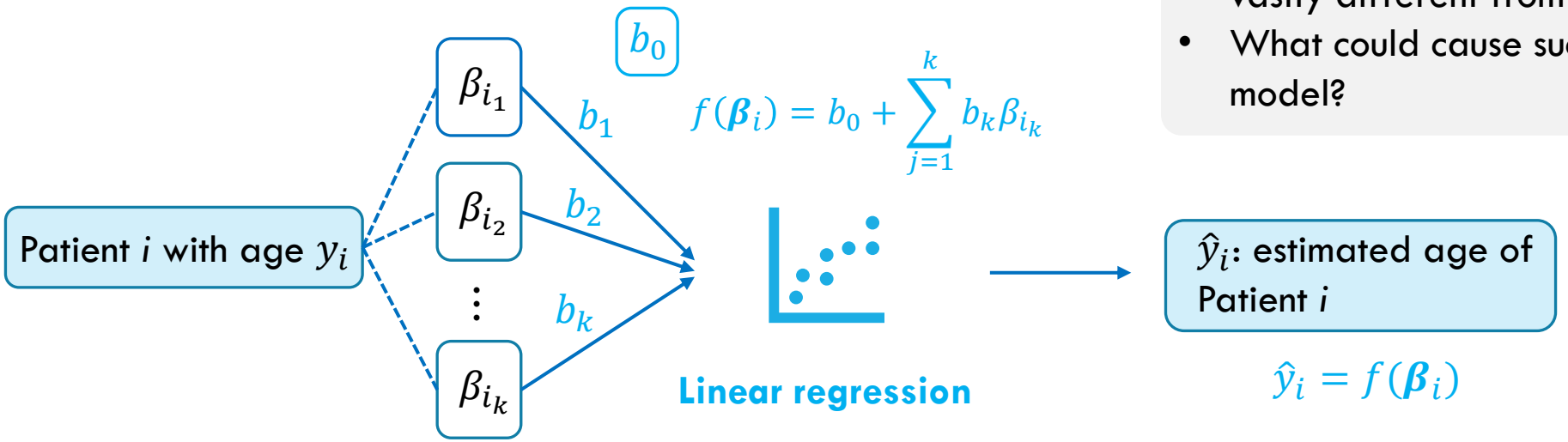


Hopefully, $\hat{y}_i \approx y_i$

LINEAR REGRESSION – IN OUR CASE

- **inputs/independent variables/features/predictors:** methylation β-values in targeted genomic positions
- **samples/observations:** patients
- **output/dependent variable/response:** chronological age

- Is the model bad if a patient's predicted age is vastly different from his/her chronological age?
- What could cause such an effect besides a bad model?



$$\hat{y}_i = f(\boldsymbol{\beta}_i)$$

Hopefully, $\hat{y}_i \approx y_i$

LINEAR REGRESSION – AVOIDING OVERFITTING

- **Number of features: 27K / 450K / 850K** (depending on measurement platform)
- **Number of patients: 10K at best** (depending on data availability and research funds)
- **There is an enormous chance of overfitting the model if we use all β-values for prediction!**

LINEAR REGRESSION – AVOIDING OVERFITTING

- **Number of features: 27K / 450K / 850K** (depending on measurement platform)
- **Number of patients: 10K at best** (depending on data availability and research funds)
- **There is an enormous chance of overfitting the model if we use all β-values for prediction!**

  → **REGULARIZATION!**

LINEAR REGRESSION – AVOIDING OVERFITTING

- **Number of features: 27K / 450K / 850K** (depending on measurement platform)
- **Number of patients: 10K at best** (depending on data availability and research funds)
- **There is an enormous chance of overfitting the model if we use all β-values for prediction!**

→ **REGULARIZATION!** ~ punishing the use of many large coefficients

Instead of simply minimizing $\sum_{i=1}^{s}(y_i - \hat{y}_i)^2$, let's minimize

$$\sum_{i=1}^{s}(y_i - \hat{y}_i)^2 + \alpha \sum_{j=1}^{k} b_j^2 + \lambda \sum_{j=1}^{k} |b_j|$$

LINEAR REGRESSION – AVOIDING OVERFITTING

- **Number of features: 27K / 450K / 850K** (depending on measurement platform)
- **Number of patients: 10K at best** (depending on data availability and research funds)
- **There is an enormous chance of overfitting the model if we use all β-values for prediction!**

→ **REGULARIZATION!** ~ punishing the use of many large coefficients

Instead of simply minimizing $\sum_{i=1}^{s}(y_i - \hat{y}_i)^2$, let's minimize

$$\sum_{i=1}^{s}(y_i - \hat{y}_i)^2 + \alpha \sum_{j=1}^{k} b_j^2 + \lambda \sum_{j=1}^{k} |b_j|$$

- $\alpha = 0, \lambda = 0 \rightarrow$ ordinary least squares
- $\alpha = 0, \lambda \neq 0 \rightarrow$ "Lasso regression" (L1-regularization)
- $\alpha \neq 0, \lambda = 0 \rightarrow$ "Ridge regression" (L2-regularization)
- $\alpha \neq 0, \lambda \neq 0 \rightarrow$ "ElasticNet"

LINEAR REGRESSION – AVOIDING OVERFITTING

- **Number of features: 27K / 450K / 850K** (depending on measurement platform)
- **Number of patients: 10K at best** (depending on data availability and research funds)
- **There is an enormous chance of overfitting the model if we use all β-values for prediction!**

→ **REGULARIZATION!** ~ punishing the use of many large coefficients

Instead of simply minimizing $\sum_{i=1}^{s}(y_i - \hat{y}_i)^2$, let's minimize

$$\sum_{i=1}^{s}(y_i - \hat{y}_i)^2 + \alpha \sum_{j=1}^{k} b_j^2 + \lambda \sum_{j=1}^{k} |b_j|$$

- $\alpha = 0, \lambda = 0 \rightarrow$ ordinary least squares
- $\alpha = 0, \lambda \neq 0 \rightarrow$ "Lasso regression" (L1-regularization)
- $\alpha \neq 0, \lambda = 0 \rightarrow$ "Ridge regression" (L2-regularization)
- $\alpha \neq 0, \lambda \neq 0 \rightarrow$ "ElasticNet"

- What do you think would happen if either $\alpha \rightarrow \infty$ or $\lambda \rightarrow \infty$?
- How would you choose the optimal α and λ ?

CONSTRUCTING THE ”EPIGENETIC CLOCK”

- For details, check out **Horvath (2013): DNA methylation age of human tissues and cell types**

CONSTRUCTING THE "EPIGENETIC CLOCK"

- For details, check out **Horvath (2013): DNA methylation age of human tissues and cell types**
- Training & test data:
  - Publicly available
  - From healthy donors
  - The data from a specific study was never split between train & test sets
  - Two different measurement platforms (27K & 450K Illumina beadchip)
  - Many different tissues

CONSTRUCTING THE "EPIGENETIC CLOCK"

- For details, check out **Horvath (2013): DNA methylation age of human tissues and cell types**
- Training & test data:
  - Publicly available
  - From healthy donors → Why?
  - The data from a specific study was never split between train & test sets
  - Two different measurement platforms (27K & 450K Illumina beadchip)
  - Many different tissues

CONSTRUCTING THE "EPIGENETIC CLOCK"

- For details, check out **Horvath (2013): DNA methylation age of human tissues and cell types**
- Training & test data:
  - Publicly available
  - From healthy donors
  - The data from a specific study was never split between train & test sets → Why?
  - Two different measurement platforms (27K & 450K Illumina beadchip)
  - Many different tissues

CONSTRUCTING THE "EPIGENETIC CLOCK"

- For details, check out **Horvath (2013): DNA methylation age of human tissues and cell types**
- Training & test data:
  - Publicly available
  - From healthy donors
  - The data from a specific study was never split between train & test sets
  - Two different measurement platforms (27K & 450K Illumina beadchip)
  - Many different tissues
- ElasticNet model with 10-fold cross-validation → **353 non-zero coefficients + intercept (b_0)**

CONSTRUCTING THE "EPIGENETIC CLOCK"

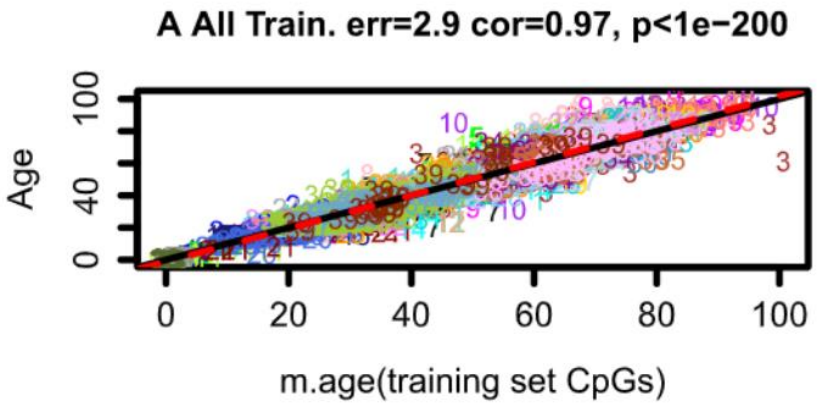
- For details, check out **Horvath (2013): DNA methylation age of human tissues and cell types**
- Training & test data:
  - Publicly available
  - From healthy donors
  - The data from a specific study was never split between train & test sets
  - Two different measurement platforms (27K & 450K Illumina beadchip)
  - Many different tissues
- ElasticNet model with 10-fold cross-validation → **353 non-zero coefficients + intercept (b_0)**

What's that?

CONSTRUCTING THE "EPIGENETIC CLOCK"

Results:

Fig 1A.

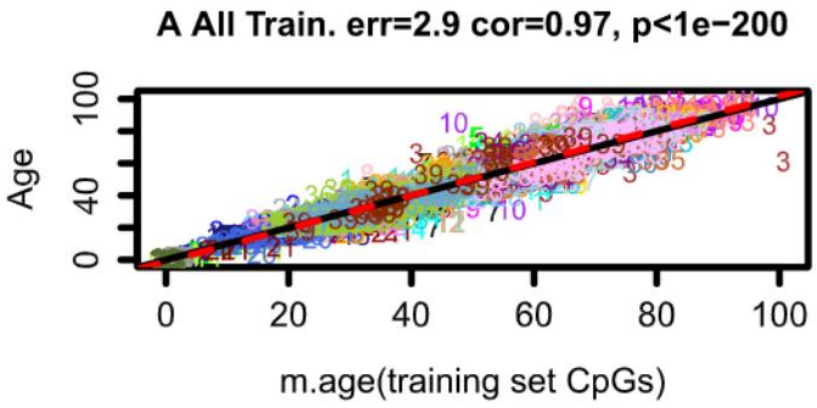


CONSTRUCTING THE "EPIGENETIC CLOCK"

Results:

on all training data

Fig 1A.

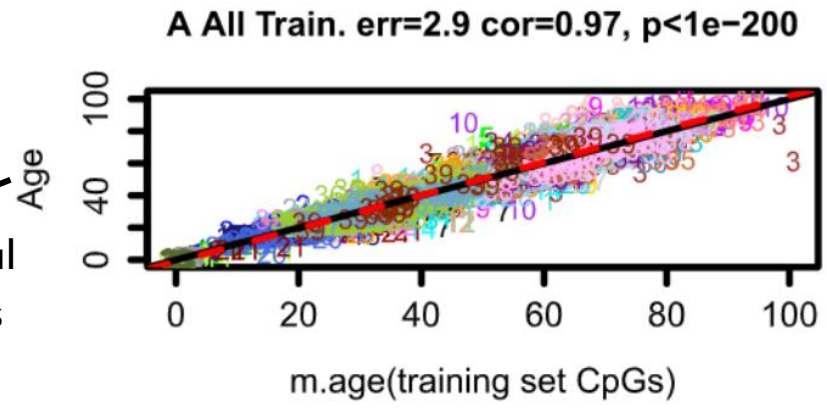


CONSTRUCTING THE "EPIGENETIC CLOCK"

Results:

on all training data

Fig 1A.



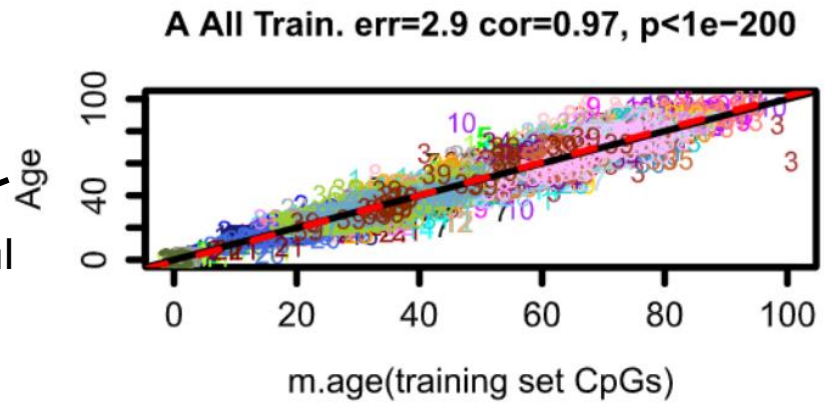
chronological age in years

CONSTRUCTING THE "EPIGENETIC CLOCK"

Results:

on all training data

Fig 1A.



chronological age in years

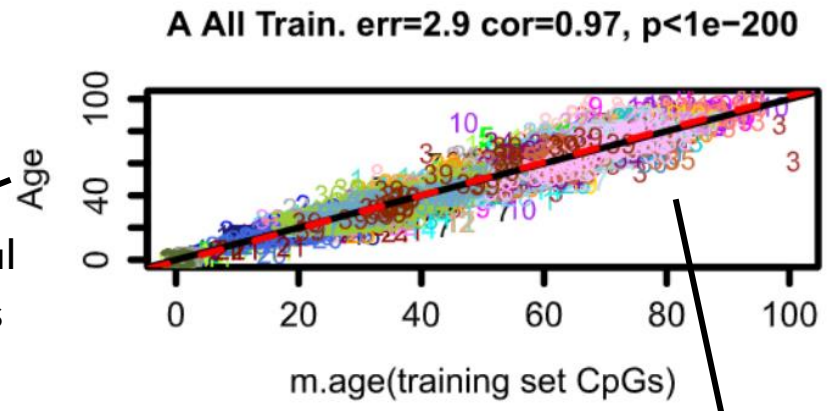
estimated "DNAm age" in years

CONSTRUCTING THE "EPIGENETIC CLOCK"

Results:

on all training data

Fig 1A.



chronological age in years

estimated "DNAm age" in years

different colours indicate different data sources

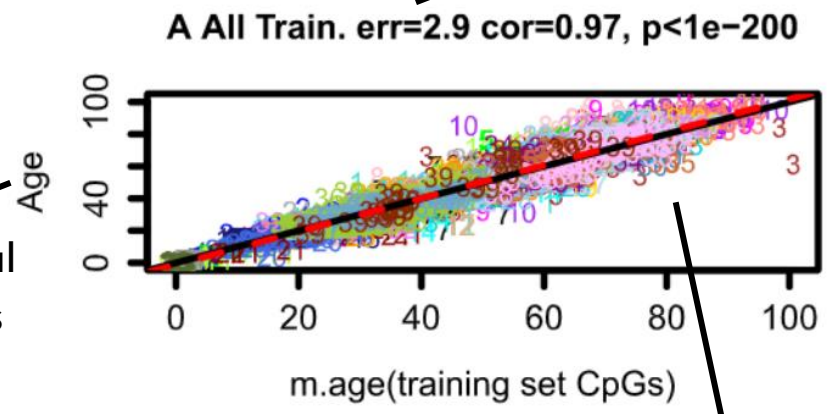
CONSTRUCTING THE "EPIGENETIC CLOCK"

Results:

on all training data

Fig 1A.

median absolute error in years: *for half of the patients in the training set, the difference between the estimated and the actual age was smaller than 2.9 years*



chronological age in years

estimated "DNAm age" in years

different colours indicate different data sources

CONSTRUCTING THE "EPIGENETIC CLOCK"

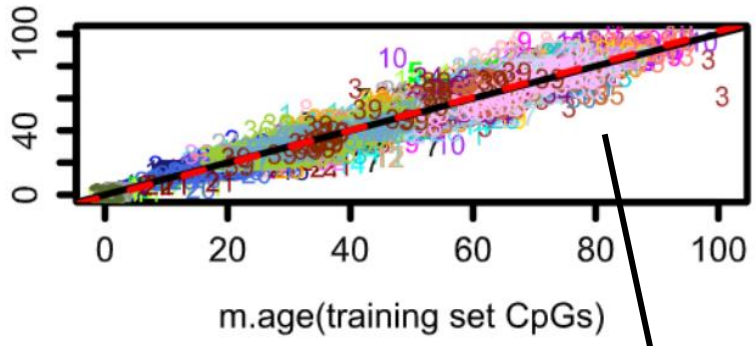
Results:

Fig 1A.

on all training data

median absolute error in years

Pearson-correlation between actual and predicted age



chronological age in years

estimated "DNAm age" in years

different colours indicate different data sources

CONSTRUCTING THE "EPIGENETIC CLOCK"

Results:

Fig 1A. on all training data

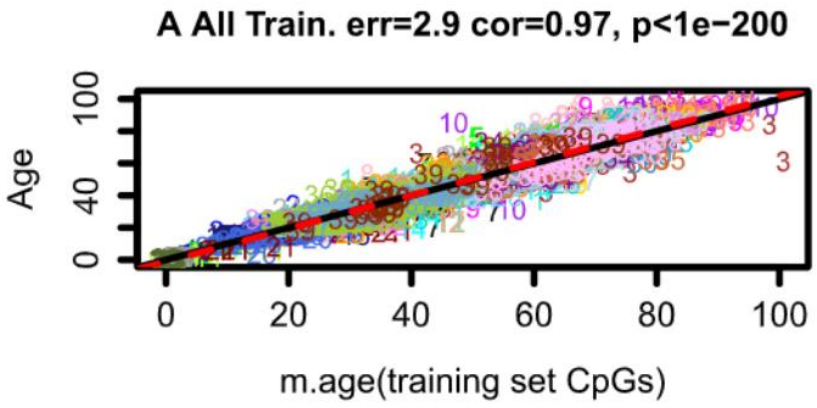
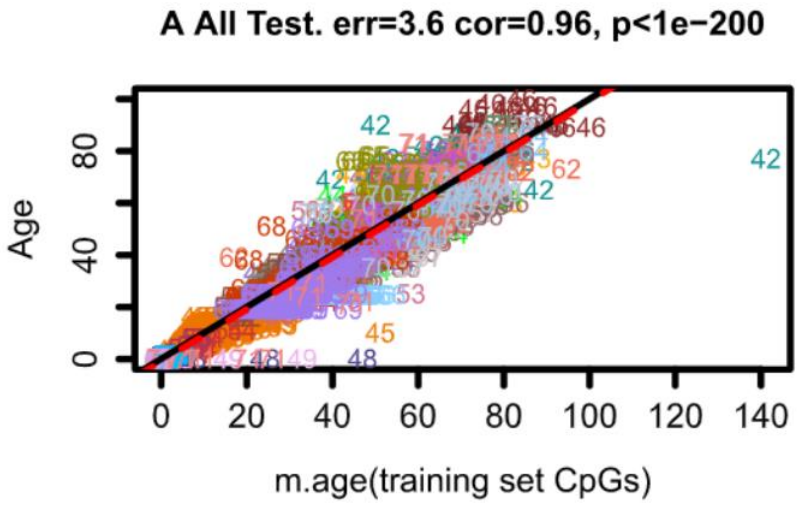


Fig 2A. on all test data



- On which dataset would you expect higher correlation? Why?
- On which dataset would you expect higher median absolute error? Why?

TRYING IT OUT FOR REAL…

1. Downloading an appropriate dataset
2. Building a model that predicts age
3. Testing the Horvath multi-tissue epigenetic clock
4. Comparing the results