Machine Learning for Biology and Health

CSCI 1851
Spring 2026

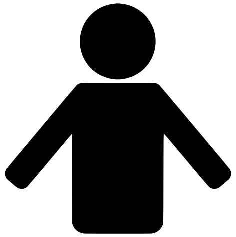
Ritambhara Singh

January 29, 2026
Thursday

Linear Regression

ChatGPT prompt “minimalist landscape painting a cute penguin working on a computer with DNA on its screen”

Recap: Heart disease classification



Input: $\mathbb{X}$

Target: $\mathbb{Y}$

```
age,cholesterol,resting_bp,max_hr
45,180,120,170
48,185,118,172
50,190,125,165
52,200,130,160
55,210,135,158
57,220,140,150
58,235,142,152
60,240,145,148
62,250,150,140
65,260,155,138
```

Function: f

$\boxed{f(\mathbb{X}) \rightarrow \mathbb{Y}}$

```
heart_disease
0
0
0
0
0
1
1
1
1
1
```

Recap: The Perceptron Learning Algorithm

1. set w's to 0.
2. for N iterations, or until the weights do not change:
   a) for each training example $\mathbf{x}^n$ with label y^k
      i. if $y^k - f(\mathbf{x}^k) = 0$ continue
      ii. else for all weights w_i, $\Delta w_i = \left(y^k - f(\mathbf{x}^k)\right) x_i^k$

w/o b = 1/0

- b = bias
- w = weights
- N = maximum number of training iterations
- $\mathbf{x}^k$ = $\mathrm{k^{th}}$ training example
- y^k = label for the $\mathrm{k^{th}}$ example
- w_i = weight for the $\mathrm{i^{th}}$ input where $i \leq n$
- n = number of features
- x_i^k = $\mathrm{i^{th}}$ input of the example where $i \leq n$

Today’s goal - Learn about linear regression

(1) Introducing the task – Predicting age using DNA methylation

(2) Linear regression

(3) Defining the loss function

(3) Optimization – Gradient descent

(4) Class Activity: Linear regression in action

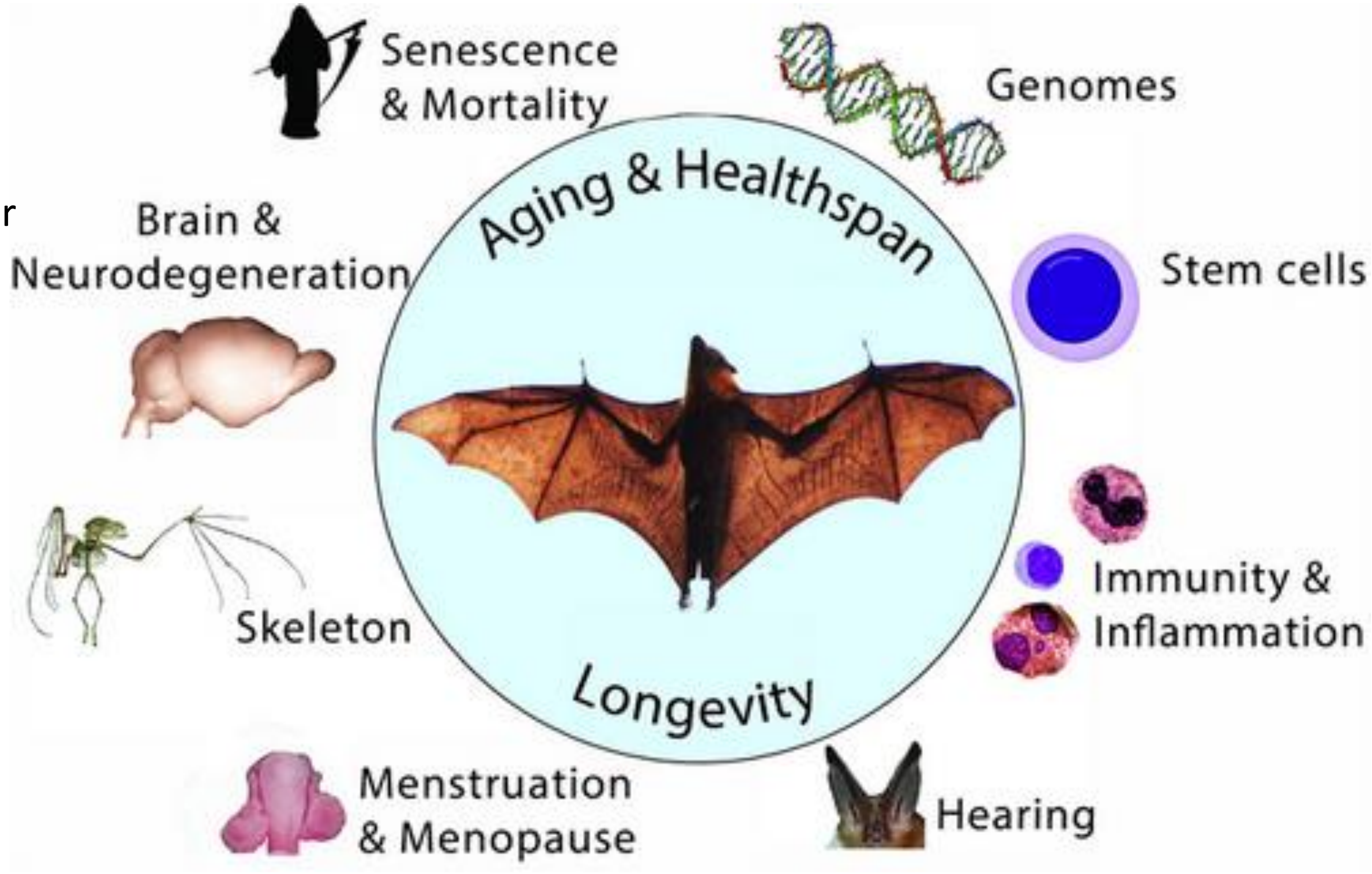
(5) Matrix formulation

Predicting age using DNA methylation

Studying aging in bats!!!

Incredible diversity of physiologies, demonstrate exceptional longevity for their body size

[Gerald Wilkinson | Department of Biology | University of Maryland]()

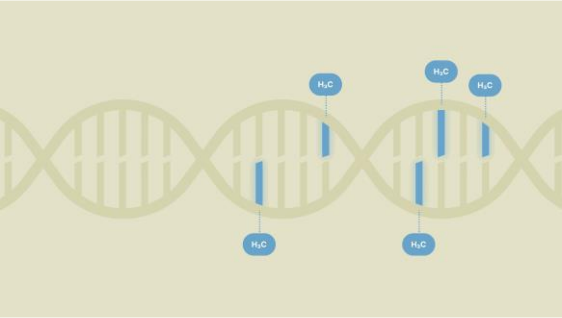


https://nyaspubs.onlinelibrary.wiley.com/doi/full/10.1111/nyas.15233

How can we estimate the age of bats in the wild?

Aging clock (or Epigenetic clock)!!!

DNA sequence



DNA methylation signal

Machine Learning

~ 20

years!

DNA sequence

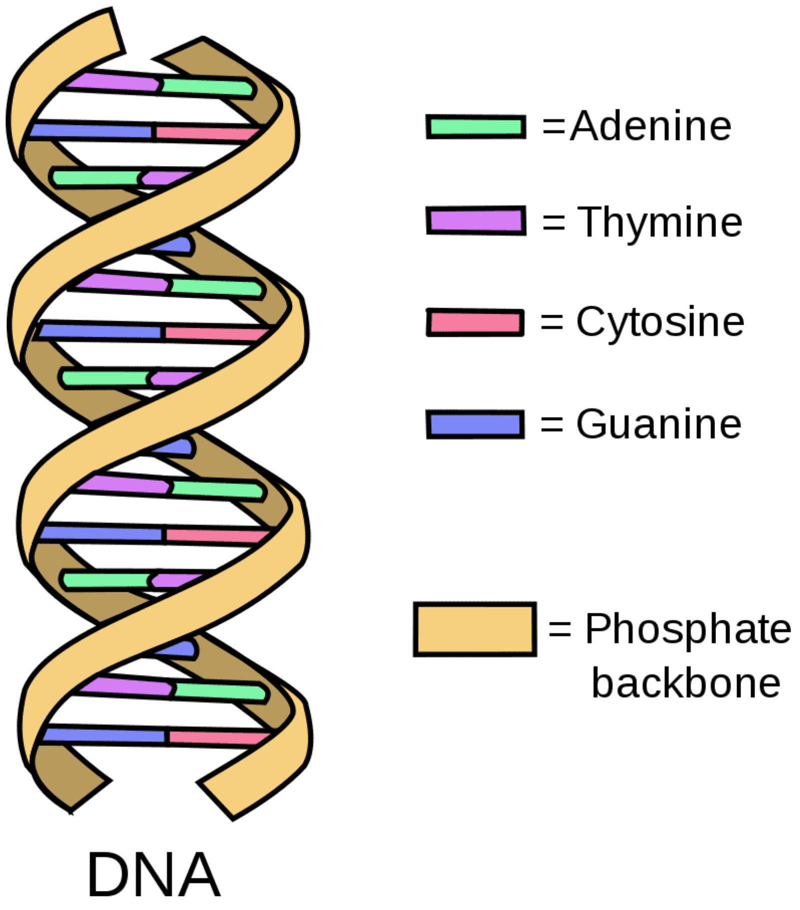


Image courtesy: https://www.ashg.org/discover-genetics/building-blocks/

Disclaimer: I am NOT a biologist!

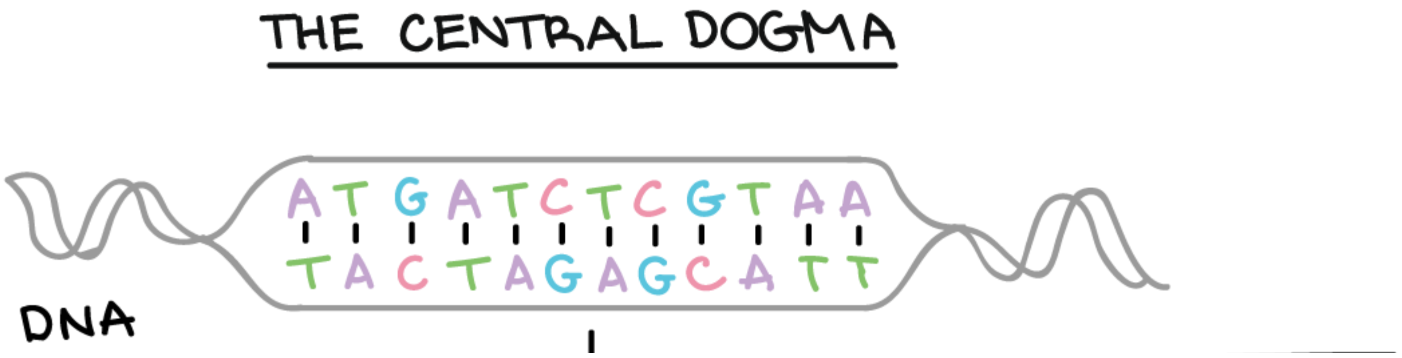


Image courtesy: https://www.khanacademy.org/science/high-school-biology/hs-molecular-genetics/hs-rna-and-protein-synthesis/a/intro-to-gene-expression-central-dogma

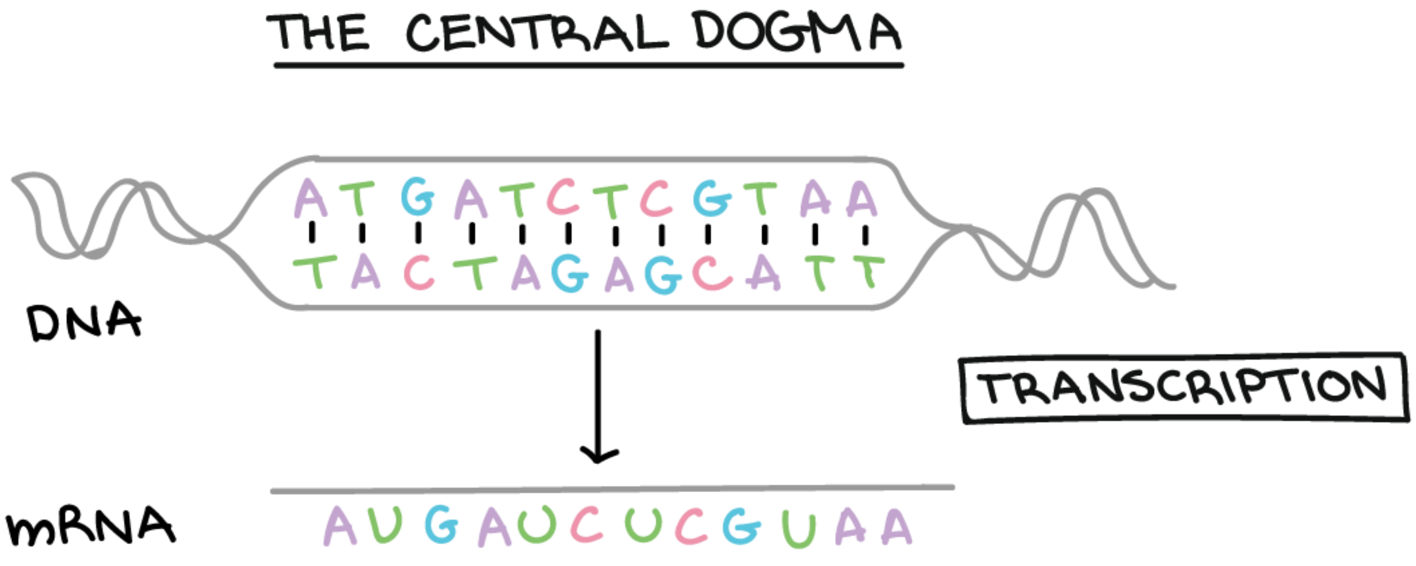


Image courtesy: https://www.khanacademy.org/science/high-school-biology/hs-molecular-genetics/hs-rna-and-protein-synthesis/a/intro-to-gene-expression-central-dogma

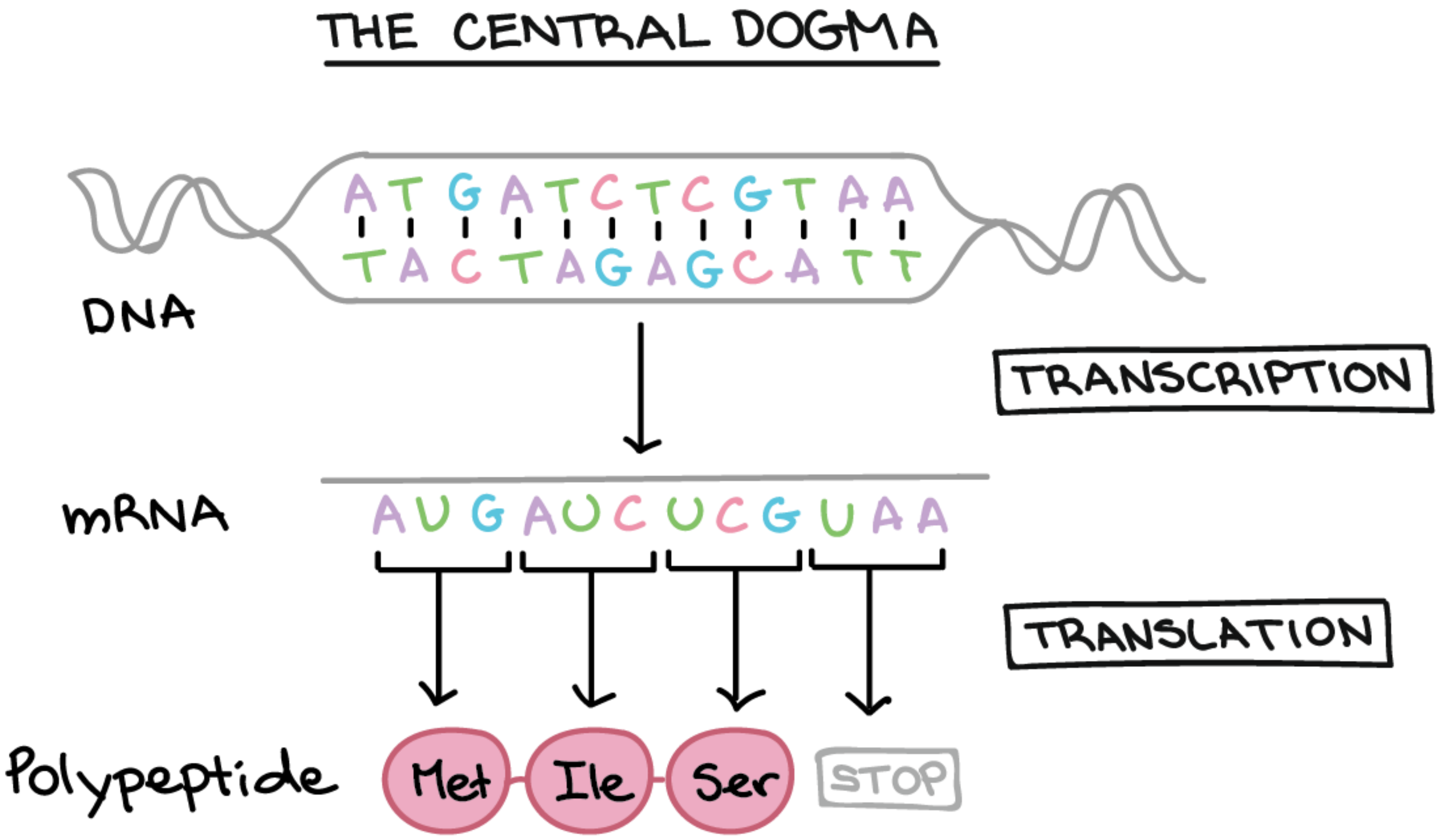
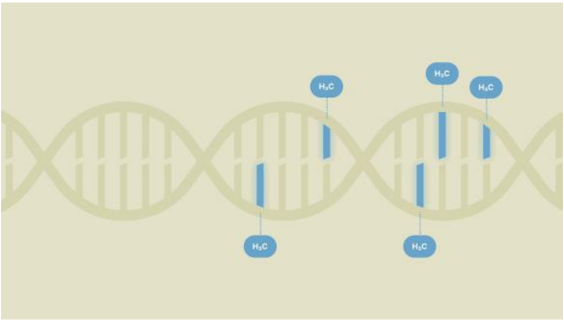


Image courtesy: https://www.khanacademy.org/science/high-school-biology/hs-molecular-genetics/hs-rna-and-protein-synthesis/a/intro-to-gene-expression-central-dogma

Aging clock!!!

DNA sequence

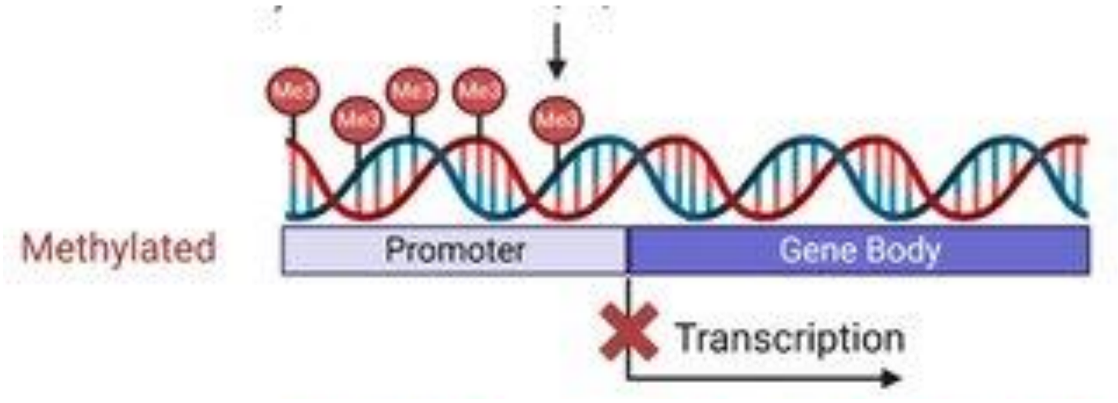


DNA methylation signal

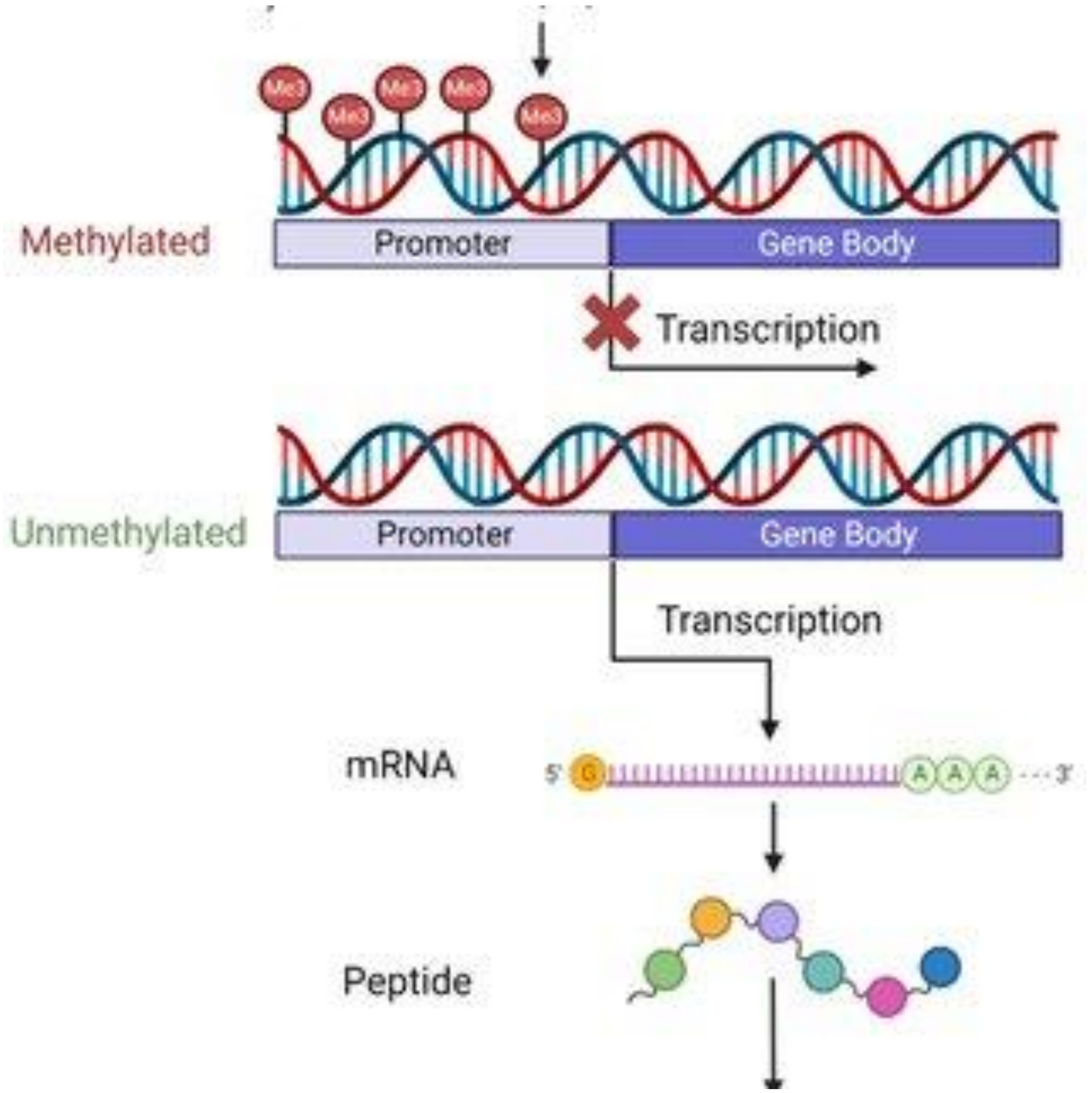
Machine Learning

~ 20 years!

DNA methylation

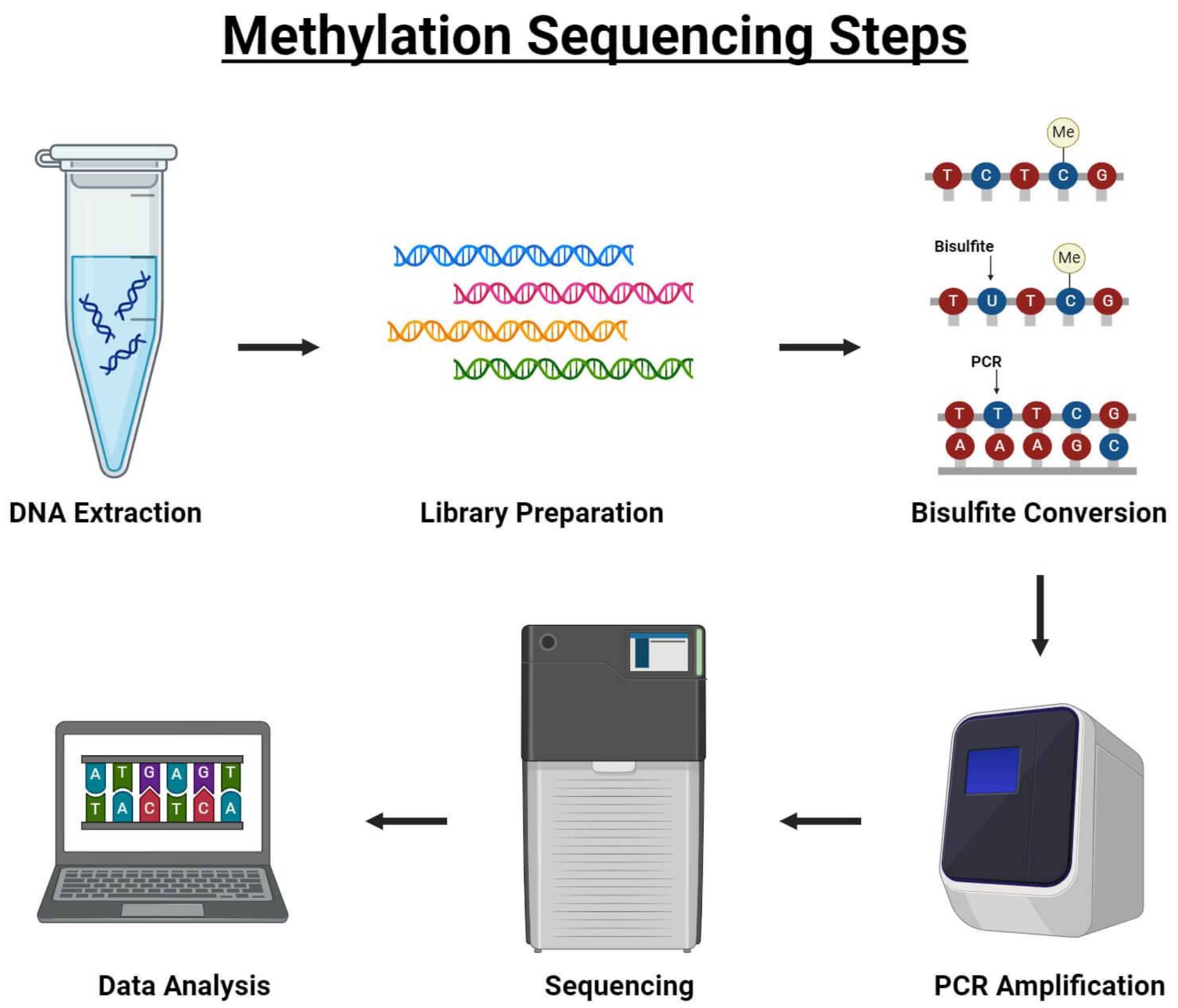


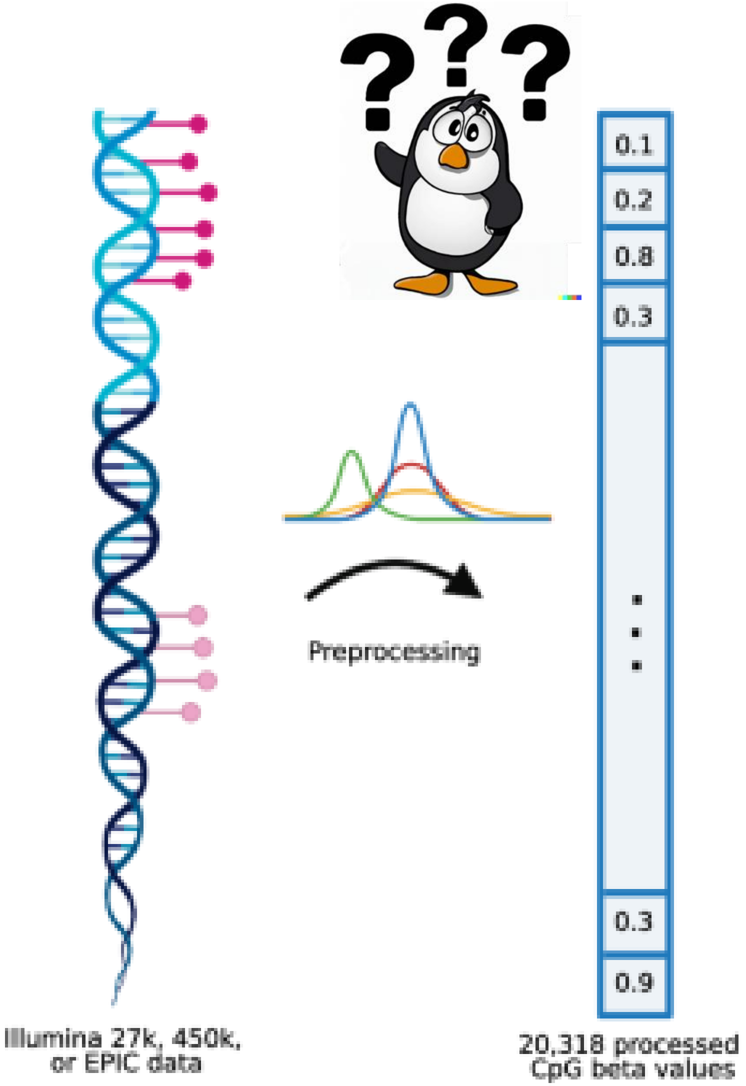
DNA methylation



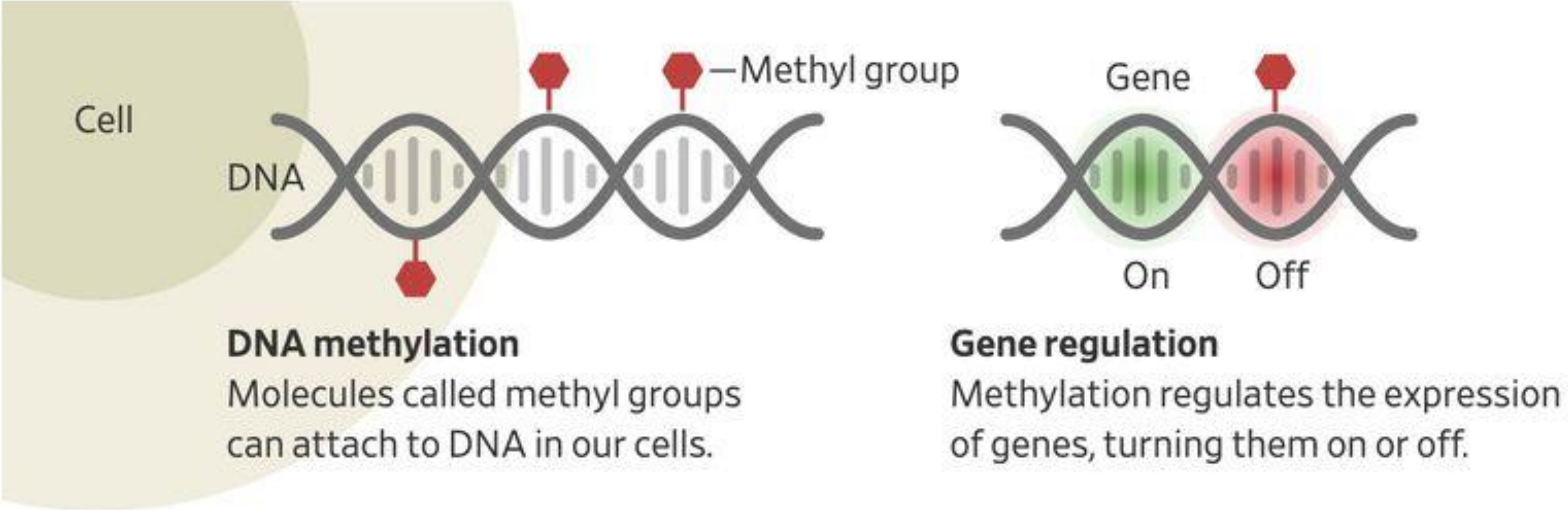
Measuring DNA methylation

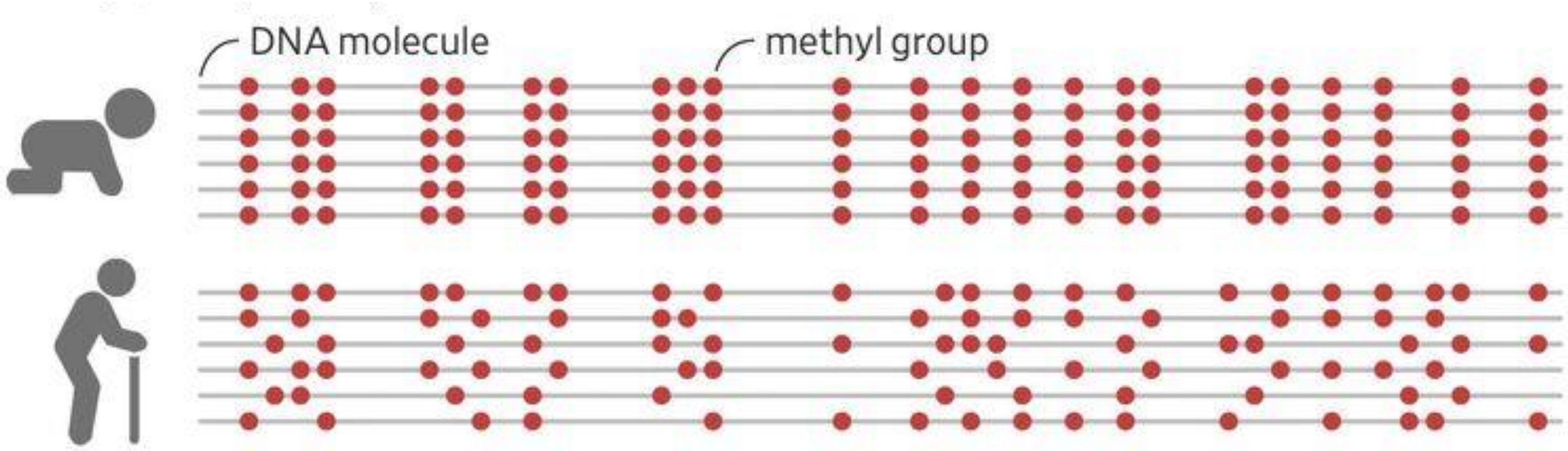
Any questions?





https://microbenotes.com/methylation-sequencing/

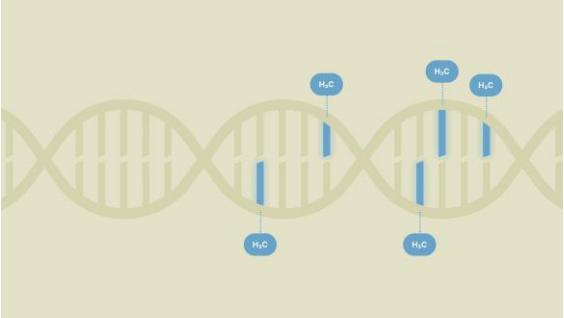




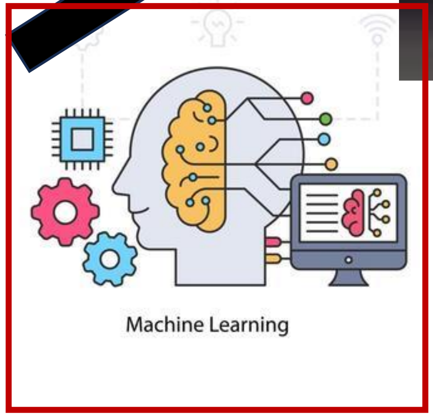
https://www.wsj.com/health/biological-age-test-ccc59a6?mod=hp_lead_pos11

Aging clock (or Epigenetic clock)!!!

DNA sequence



DNA methylation signal



~ 20 years!

Age prediction from DNA methylation

What is our input space?

What is our output space?

What is our prediction task?

Input: $\mathbb{X}$

Target: $\mathbb{Y}$

Sample	CpG_1	CpG_2	CpG_3	Age
1	0.20	0.35	0.60	25
2	0.30	0.45	0.65	35
3	0.40	0.55	0.70	45
4	0.50	0.65	0.75	55

Function: f

$$\boxed{f(\mathbb{X}) \rightarrow \mathbb{Y}}$$

Linear regression problem

Simpler example: How do we represent input/output?

Input: $\mathbb{X}$

"CpG site 1"

$x^{(1)}$ 0.20

$x^{(2)}$ 0.30

$x^{(3)}$ 0.40

$\mathbb{X} \in \mathbb{R}$

Regression

Function: f

f(X)→Y

Do you see a trend here?

What is different about the output here?

Target: $\mathbb{Y}$

"Age"

$y^{(1)}$ 25

$y^{(2)}$ 35

$y^{(3)}$ 45

$\mathbb{Y} \in \mathbb{Z}$

(Numerical output)

Learning function f

Input: $\mathbb{X}$

"CpG site 1"

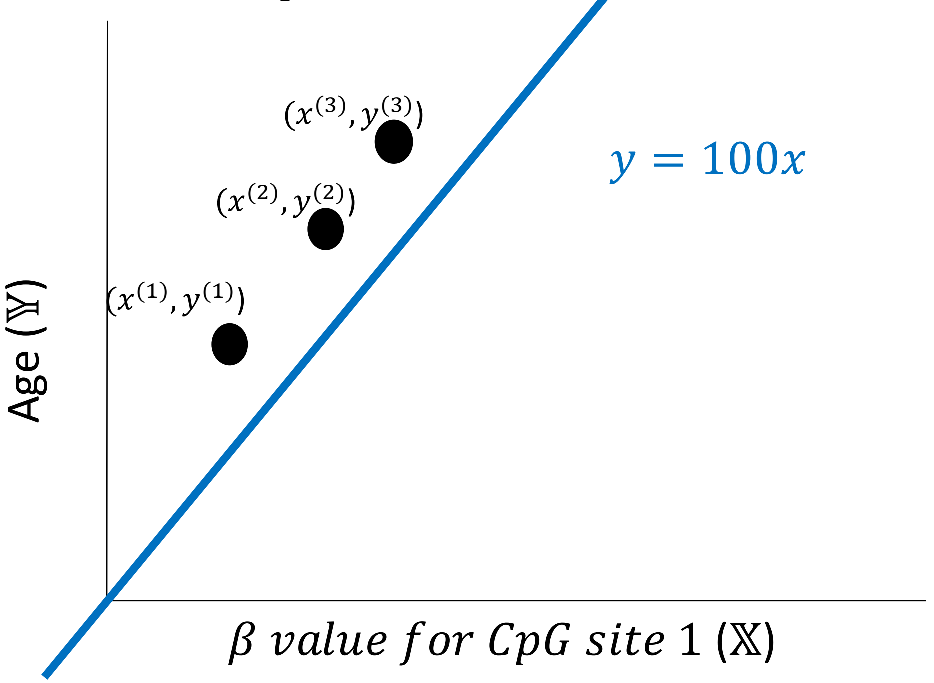
$\mathbb{X} \in \mathbb{R}$

$x^{(1)}$	0.20
$x^{(2)}$	0.30
$x^{(3)}$	0.40

Regression

Linear function

$$y = wx + b$$



Target: $\mathbb{Y}$

"Age"

$y^{(1)}$	25
$y^{(2)}$	35
$y^{(3)}$	45

$\mathbb{Y} \in \mathbb{Z}$

(Numerical output)

Learning function f

Input: $\mathbb{X}$

"CpG site 1"

Regression

Target: $\mathbb{Y}$

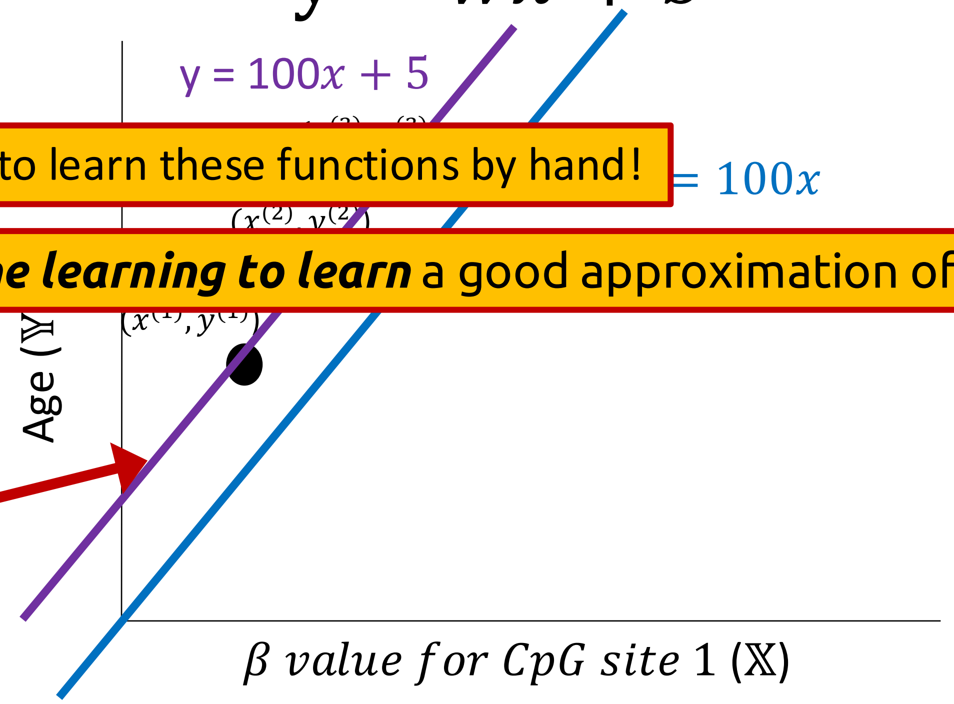
"Age"

Linear function

$$y = wx + b$$

$x^{(1)}$	0.20	$y^{(1)}$	25
$x^{(2)}$	0.30	$y^{(2)}$	35
$x^{(3)}$	0.40	$y^{(3)}$	45

$\mathbb{X} \in \mathbb{R}$



Very hard to learn these functions by hand!

Use machine learning to learn a good approximation of the function ***from data***

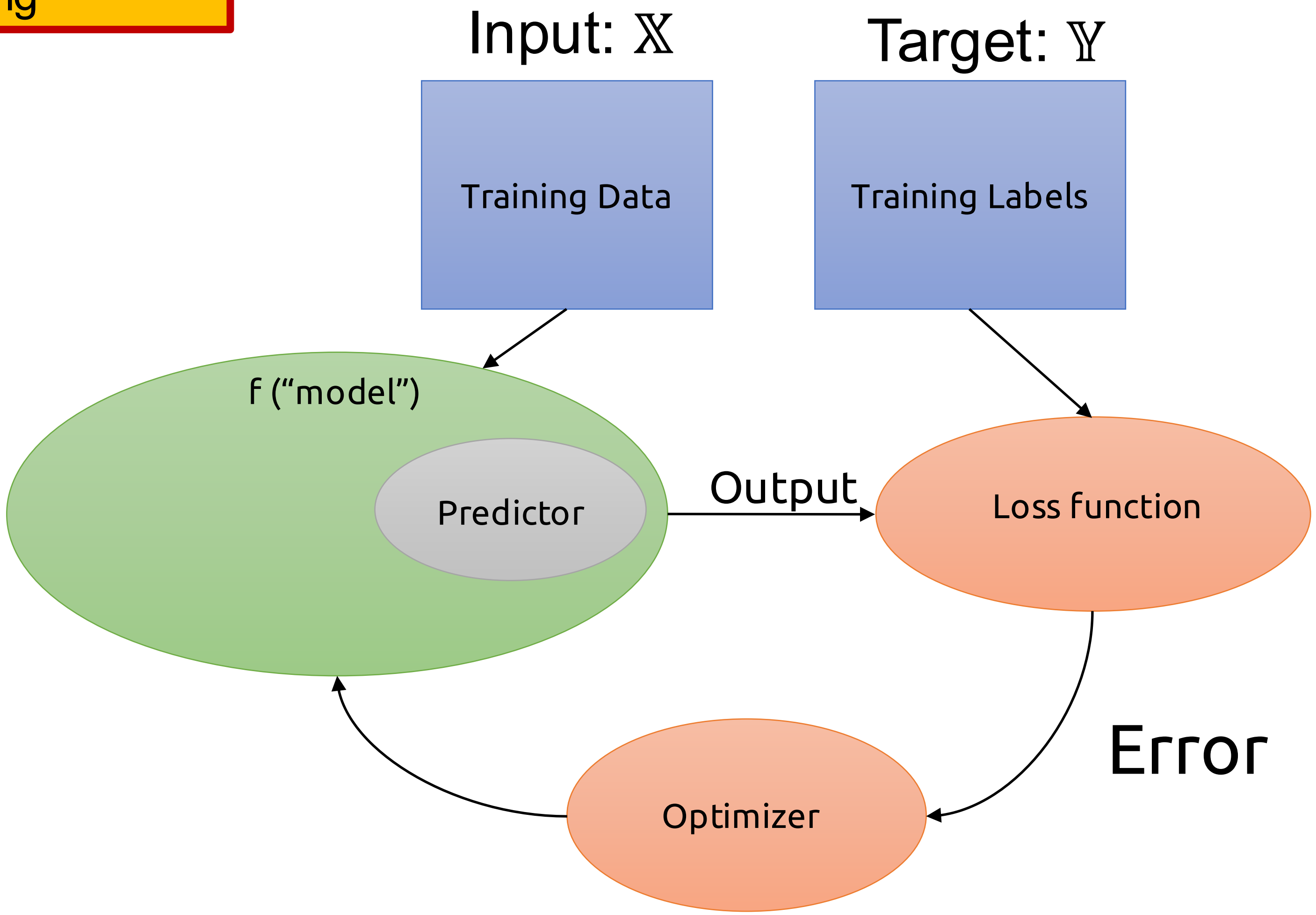
Only the line with bias can fit the data

$\mathbb{Y} \in \mathbb{Z}$

(Numerical output)

“Classic” Supervised Learning in Machine Learning

Training



Any questions?

Loss function for regression

Mean Squared Error (MSE)

Average squared residual (residual: difference between predicted and true value)

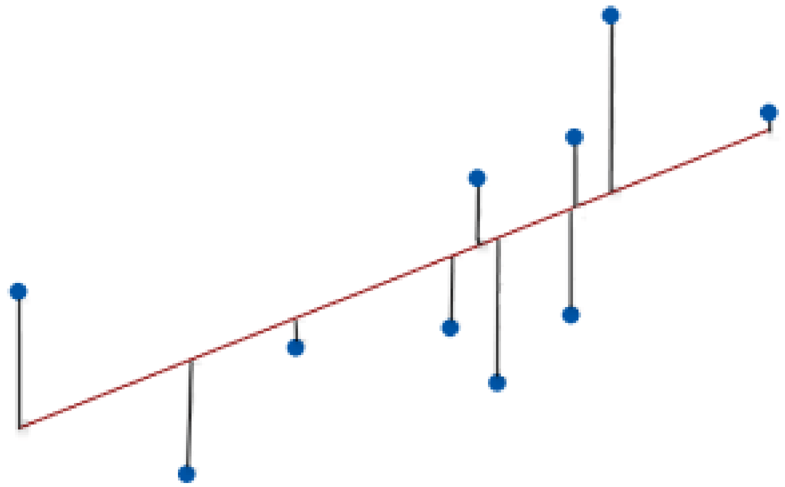
Decreasing the MSE = the model has less error = data points fall closer to the regression line

$$MSE = \frac{\sum_{k=1}^{n}(y^k - \hat{y}^k)^2}{n}$$

y^k : *true output value*

$\hat{y}^k$: *predicted output value*

n: *number of samples*



MSE is the average squared distance between the observed and predicted values

What could be the purpose of squaring the distance?

Courtesy: https://statisticsbyjim.com/regression/mean-squared-error-mse/

Mean Squared Error (MSE)

Average squared residual (residual: difference between predicted and true value)

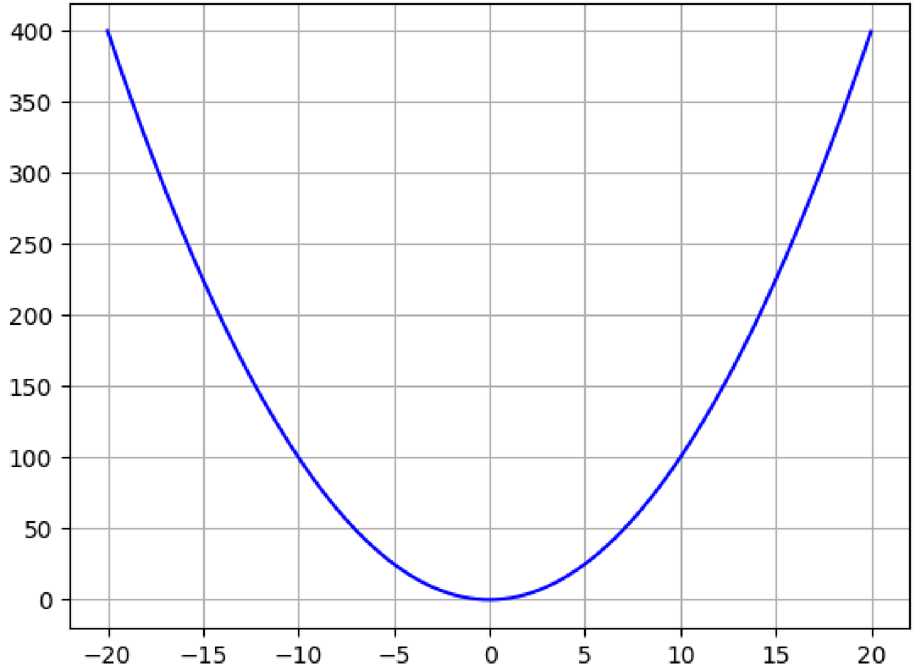
Decreasing the MSE = the model has less error = data points fall closer to the regression line

$$MSE = \frac{\sum_{k=1}^{n}(y^k - \hat{y}^k)^2}{n}$$

y^k : *true output value*

$\hat{y}^k$: *predicted output value*

n: *number of samples*



What could be the purpose of squaring the distance?

Courtesy: https://statisticsbyjim.com/regression/mean-squared-error-mse/

Convex functions

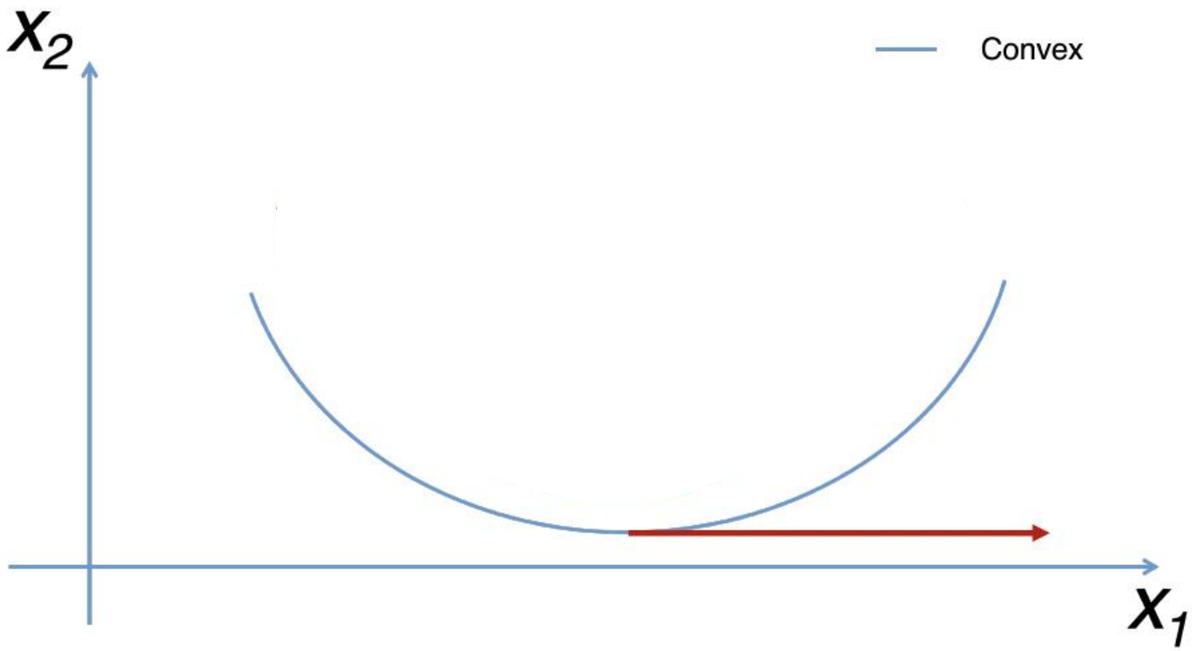


Figure: https://fmin.xyz/docs/theory/Convex_function/

Any questions?

Optimization

What does it mean to optimize?

"Optimization" comes from the same root as "optimal", which means *best*. When you optimize something, you are "making it best".

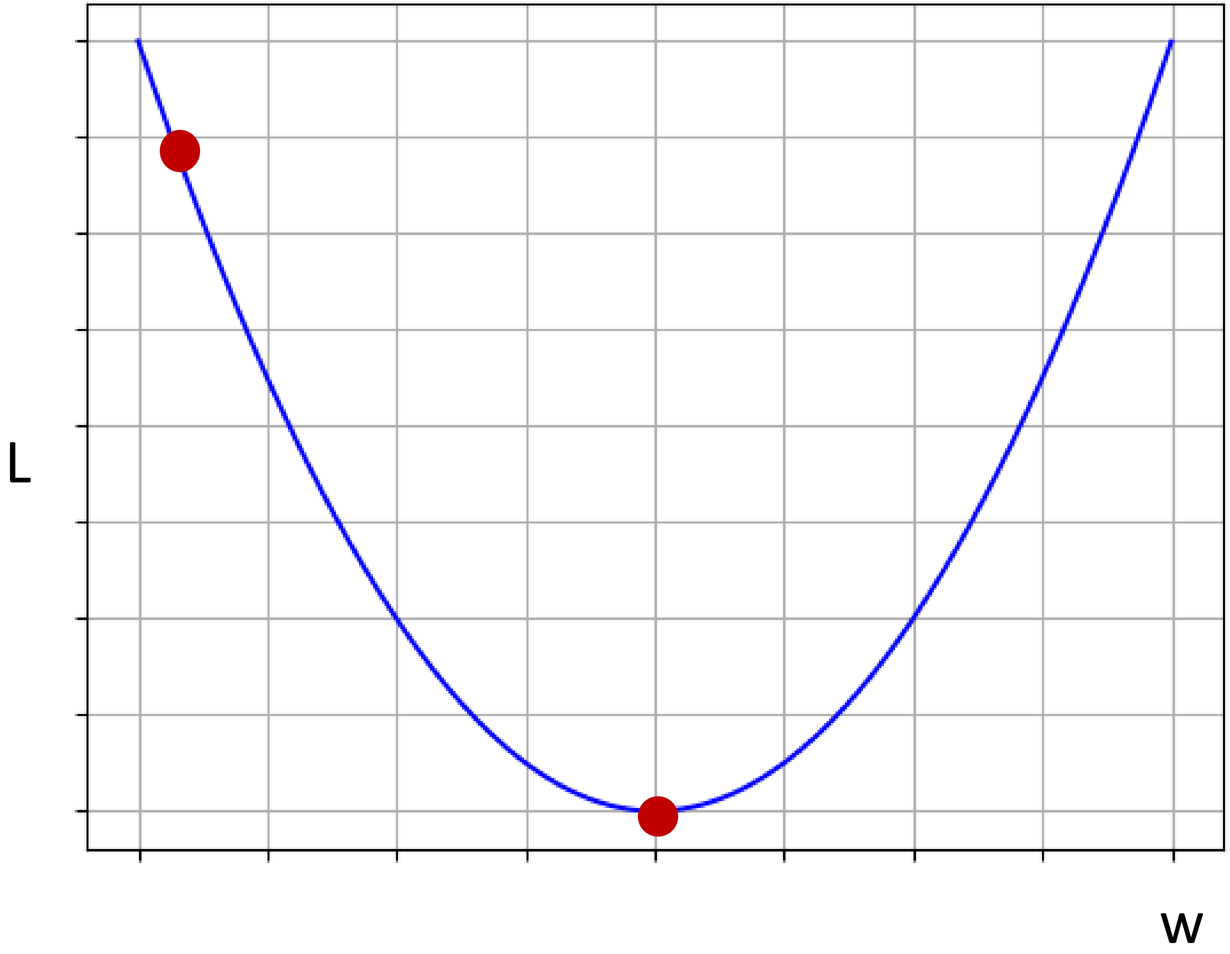
For our case, we want to minimize the loss function to get the "best" model!

https://web.stanford.edu/group/sisl/k12/optimization/MO-unit1-pdfs/1.1optimization.pdf

What does it mean to optimize?

1. Calculate the parameter update values

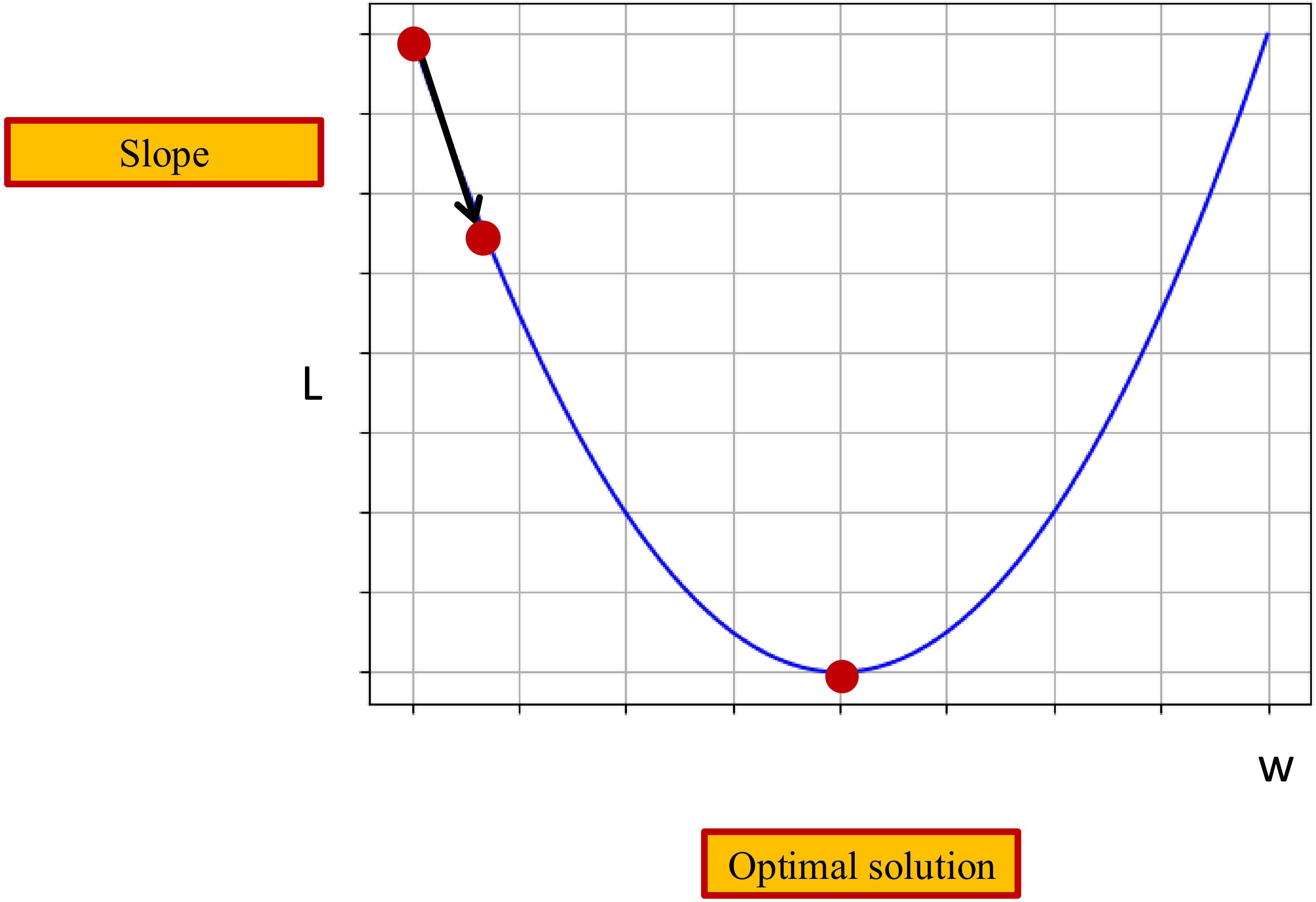
2. Update the parameters



Optimal solution

Gradient (measuring the change)

Calculating partial derivative of the Loss
with respect to the weights/parameters



Courtesy: https://statisticsbyjim.com/regression/mean-squared-error-mse/

Vector Calculus Recap

- Partial derivative: the derivative of a multivariable function with respect to one of its variables

Vector Calculus Recap

- Partial derivative: the derivative of a multivariable function with respect to one of its variables
- Example: $f(x, w, b) = wx + b$
- The partial derivative of f with respect to w is $\frac{\partial f}{\partial w}$

Vector Calculus Recap

- Partial derivative: the derivative of a multivariable function with respect to one of its variables
- Example: $f(x, w, b) = wx + b$
- The partial derivative of f with respect to w is $\frac{\partial f}{\partial w}$
- How to compute? -- treat all other variables as constants and differentiate

$$\frac{\partial f}{\partial w} =$$

Vector Calculus Recap

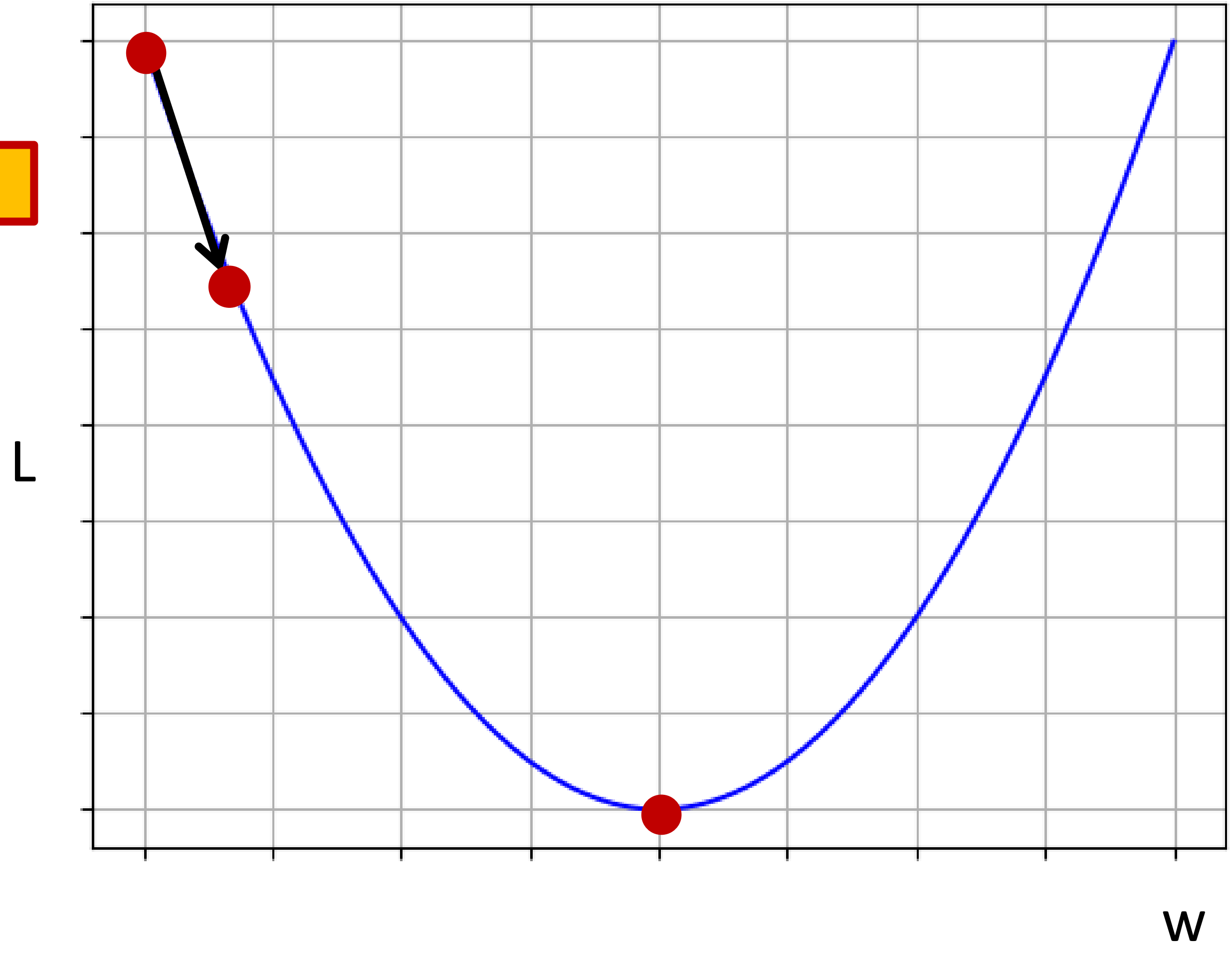
- Partial derivative: the derivative of a multivariable function with respect to one of its variables
- Example: $f(x, w, b) = wx + b$
- The partial derivative of f with respect to w is $\frac{\partial f}{\partial w}$
- How to compute? -- treat all other variables as constants and differentiate

$$\frac{\partial f}{\partial w} = \frac{\partial}{\partial w}(wx + \mathrm{b}\,) = \frac{\partial}{\partial w}(wx) + \frac{\partial}{\partial w}(b) = x + 0 = x$$

Gradient Descent

$$\Delta w = -\alpha \cdot \frac{\partial L}{\partial w}$$

Learning rate



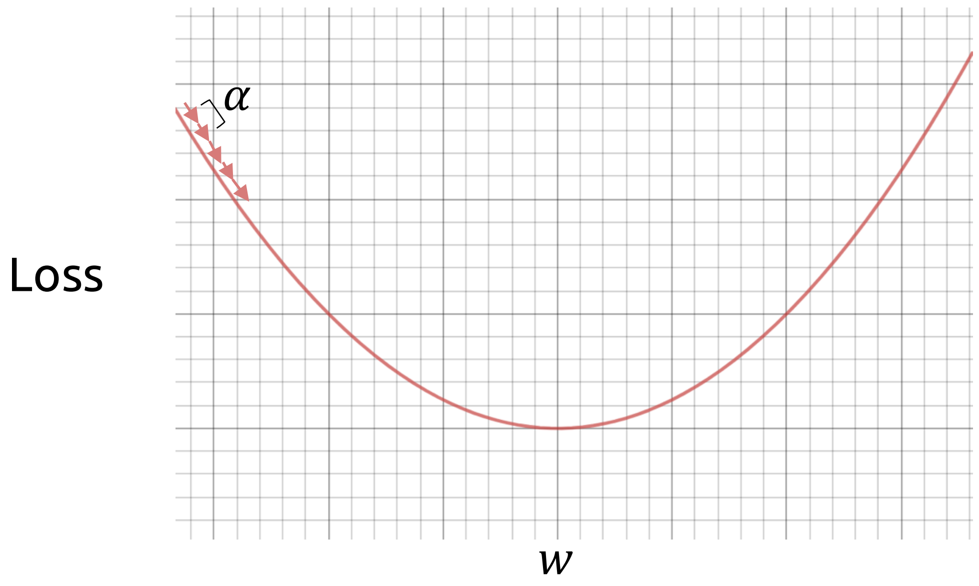
Courtesy: https://statisticsbyjim.com/regression/mean-squared-error-mse/

Impact of Learning Rate

$$\Delta w = -\alpha \cdot \frac{\partial L}{\partial w}$$

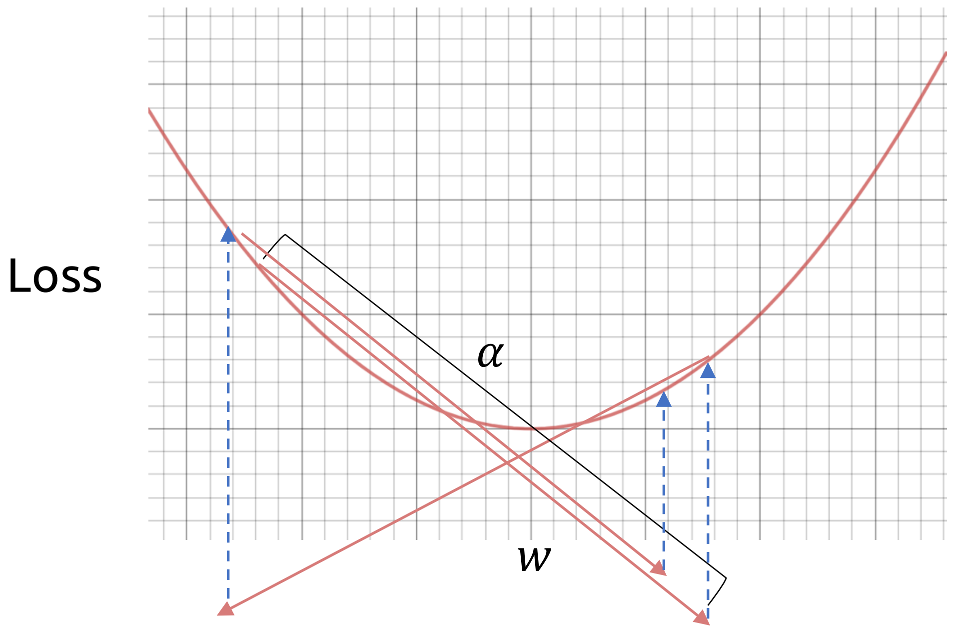
Learning rate too small?
Slow Convergence

$\alpha = 10^{-8}$



Learning rate too big?
Instability (“overshooting”)

$\alpha = 10^{-1}$

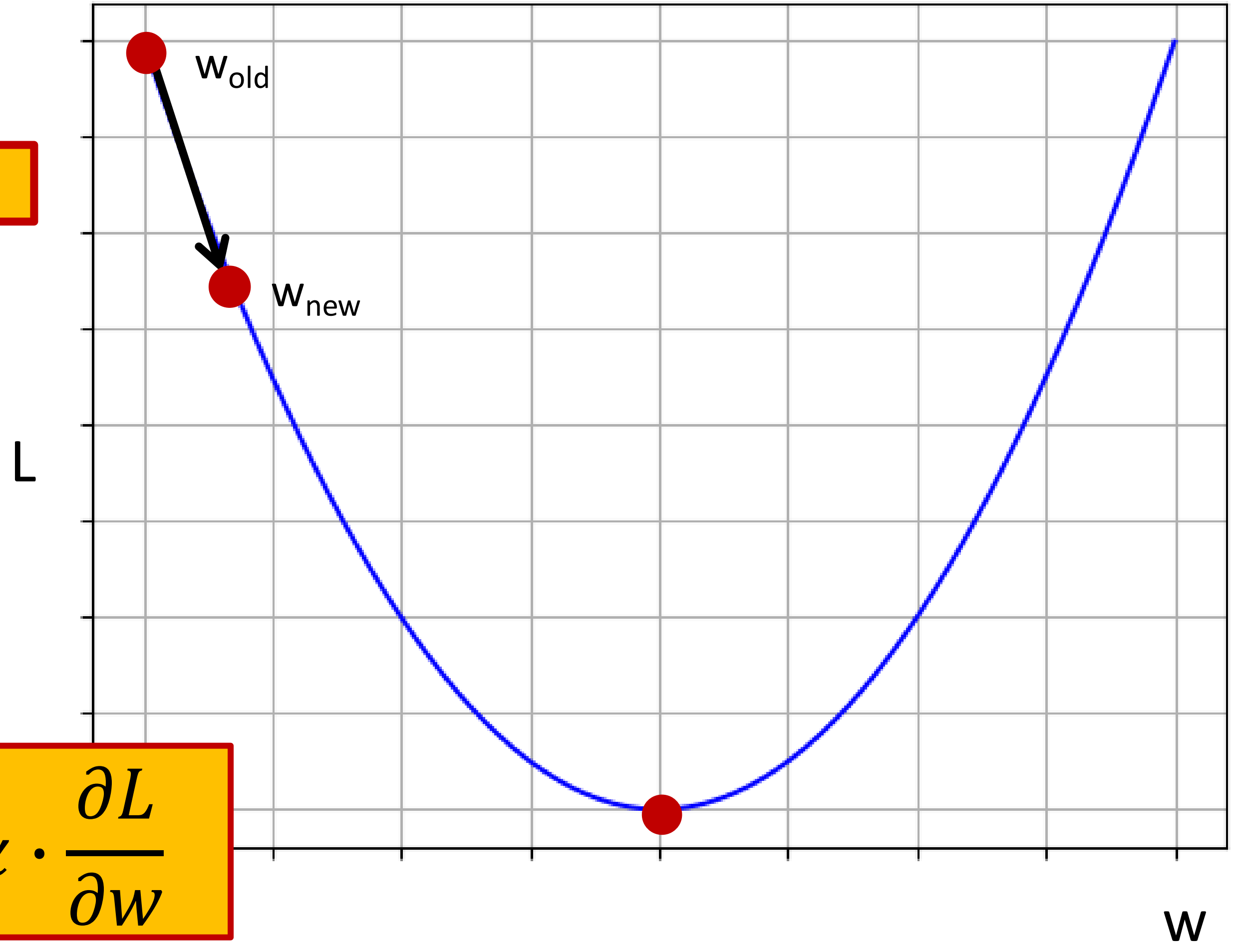


Gradient Descent (updating parameters)

$$\Delta w = -\alpha \cdot \frac{\partial L}{\partial w}$$

Learning rate

Slope



$$w_{new} = w_{old} - \alpha \cdot \frac{\partial L}{\partial w}$$

Optimal solution

Courtesy: https://statisticsbyjim.com/regression/mean-squared-error-mse/

Gradient Descent of MSE (1 sample)

$$\Delta w = -\alpha \cdot \frac{\partial L}{\partial w}$$

$$L = (y - \hat{y})^2$$

$$= (y - f(x))^2$$

$$= y^2 + f(x)^2 - 2yf(x)$$

$$= y^2 + (wx + b)^2 - 2y(wx + b)$$

$$= y^2 + w^2x^2 + b^2 + 2wxb - 2ywx - 2yb$$

$$\frac{\partial L}{\partial w} = ?$$

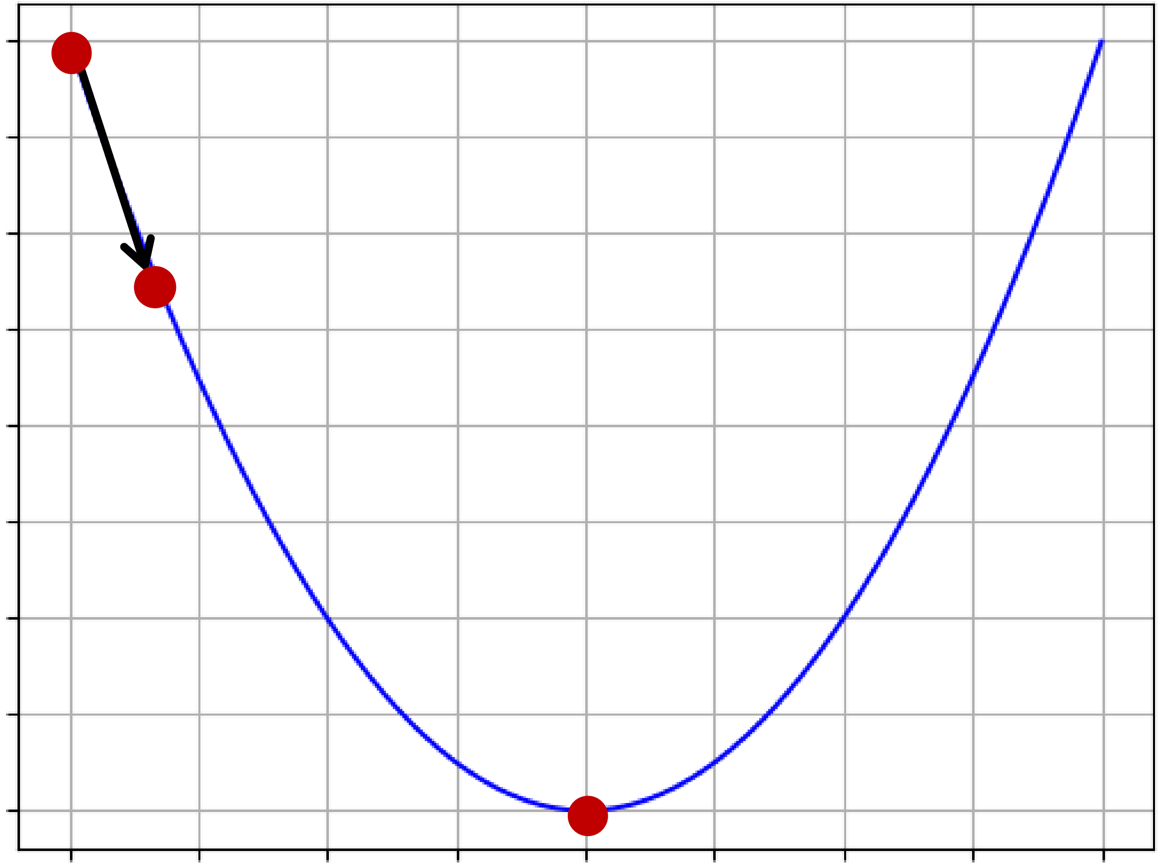
$$\frac{\partial L}{\partial w} = 2wx^2 + 2xb - 2yx$$

$$\frac{\partial L}{\partial w} = 2x(wx + b - y)$$

$$\frac{\partial L}{\partial w} = 2x(error)$$

Any questions?

L



w

$$\boxed{w_{new} = w_{old} - \alpha \cdot \frac{\partial L}{\partial w}}$$

Class activity

[Linear Regression Demo]

Matrix Formulation

Linear regression using Matrix Operations

$$X = \begin{bmatrix} x_{11} & x_{12} & x_{13} \\ x_{21} & x_{22} & x_{23} \\ x_{31} & x_{32} & x_{33} \\ x_{41} & x_{42} & x_{43} \end{bmatrix}$$

y_pred = X @ w + b

$$Xw = \begin{bmatrix} x_{11}w_1 + x_{12}w_2 + x_{13}w_3 \\ x_{21}w_1 + x_{22}w_2 + x_{23}w_3 \\ x_{31}w_1 + x_{32}w_2 + x_{33}w_3 \\ x_{41}w_1 + x_{42}w_2 + x_{43}w_3 \end{bmatrix}$$

Broadcasting

- Actually not a problem because of broadcasting!
- Broadcasting: implicitly replicating a matrix along some dimension to make math operations possible.
- NumPy will broadcast for you.

Linear regression using Matrix Operations

```
def compute_gradients(X, y, y_pred):
n = len(y)

error = y_pred - y

dw = (2 / n) * X.T @ error
db = (2 / n) * np.sum(error)

return dw, db
```

$$\text{error} = \begin{bmatrix} \hat{y}_1 - y_1 \\ \hat{y}_2 - y_2 \\ \hat{y}_3 - y_3 \\ \hat{y}_4 - y_4 \end{bmatrix}$$

$$X^T = \begin{bmatrix} x_{11} & x_{21} & x_{31} & x_{41} \\ x_{12} & x_{22} & x_{32} & x_{42} \\ x_{13} & x_{23} & x_{33} & x_{43} \end{bmatrix}$$

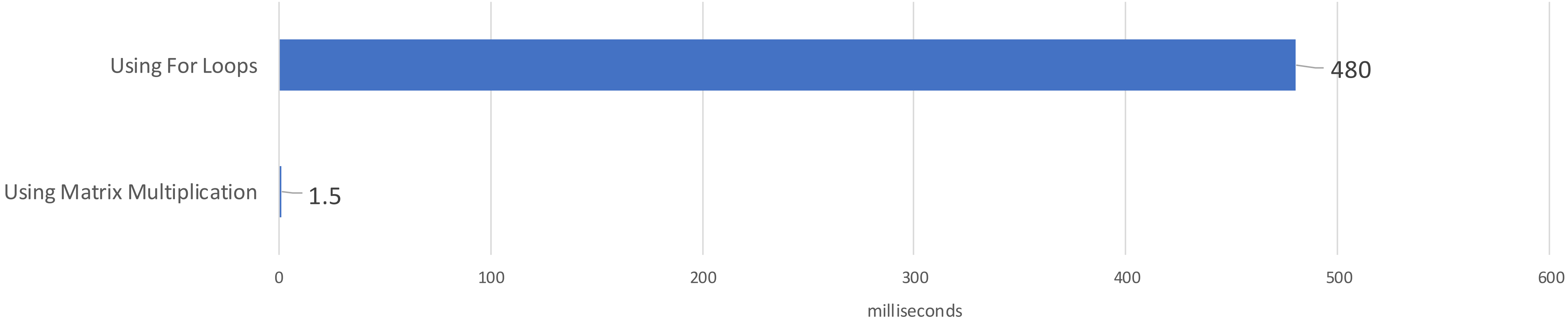
$$y^2 + w^2x^2 + b^2 + 2wxb \; - 2ywx - 2yb$$

$$\begin{bmatrix} \sum_i x_{i1}(\hat{y}_i - y_i) \\ \sum_i x_{i2}(\hat{y}_i - y_i) \\ \sum_i x_{i3}(\hat{y}_i - y_i) \end{bmatrix}$$

Why is matrix formulation useful?

Existing linear algebra optimizations

- Matrix multiplication can be ***way*** faster than *for* loops
- Example: time required to compute dot product of $a, b \in \mathbb{R}^{1,000,000}$



From: https://www.coursera.org/lecture/neural-networks-deep-learning/vectorization-NYnog

- Lots of existing effort to build fast linear algebra code (e.g. NumPy)
- Leads to order of magnitude speedup!

Horvath’s clock

DNA methylation age of human tissues and cell types

Research | Open access | Published: 10 December 2013

Volume 14, article number 3156, (2013) Cite this article

You have full access to this open access article

Download PDF

Save article

Genome Biology

Aims and scope →

Submit manuscript →

Steve Horvath

Part of a collection:
Twenty years with Genome Biology

Sections | Figures | References

Today's goal - Learn about linear regression

(1) Introducing the task – Predicting age using DNA methylation

(2) Linear regression

(3) Defining the loss function

(3) Optimization – Gradient descent

(4) Class Activity: Linear regression in action

(5) Matrix formulation

Homework reading:
[Section 7.6 in CIML book]

If interested, further reading:
[DNA methylation preprocessing]

Wrap up

What was the clearest point today?

What was the muddiest point today?