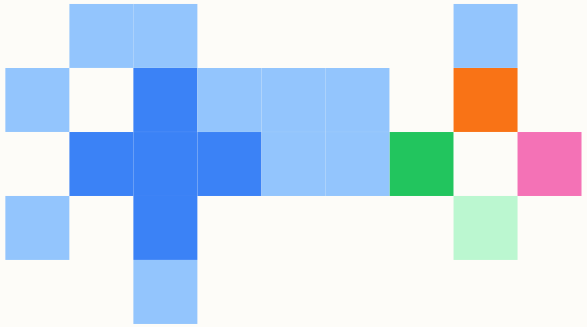


Intro to Machine Learning

Session 1 of 7

What is Machine Learning?

Data-Driven Logic

Programs learn patterns from examples instead of explicit code.

No Hand-Written Rules

Complex problems evade rigid if-else programming structures.

Classic Application

Email spam filters trained on thousands of labelled messages.



Data ✦ Model

A systematic way to measure and improve predictions.

KEY IDEA Instead of writing rules, show examples and let the model find the rules.

Types of Learning

Supervised

Predict target outcomes using labelled historical examples.

Unsupervised

Discover hidden structures in data without explicit labels.

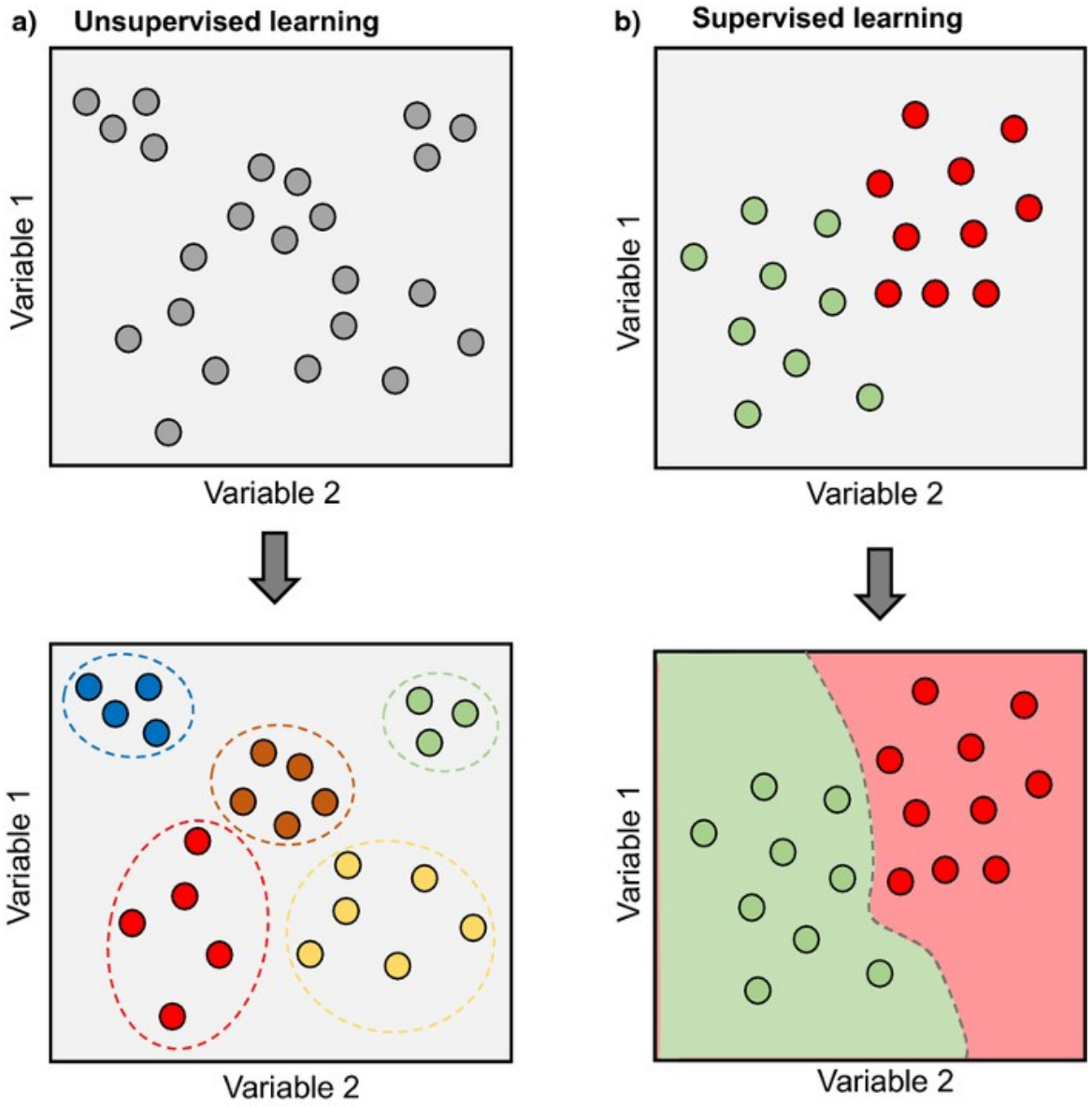
Reinforcement

Using trial, error, and reward to learn

NOTE Most of this tutorial series focuses on supervised learning: predicting y from x.

Types of Learning

Unsupervised



Supervised

Types of Learning:

Supervised Learning

Data: (x, y)
x is an input data, y is a label (e.g. photo with label "cat")

Goal: Learn to map input to output
i.e. $\boldsymbol{x} \rightarrow \boldsymbol{y}$

An example: to classify



This is a cat

Unsupervised Learning

Data: x
x is data, there's no labels!

Goal: Learn an underlying structure of the data.

An example: Comparison



The two things are alike

Supervised Learning:



Data

Collect and clean raw information.

Features

Choose inputs (x) and labels (y).

Model

Pick architecture and loss function.

Train

Evaluate parameters on unseen data.

Test

Compare results against test data, diagnose overfitting

Generalisation:

Train Split

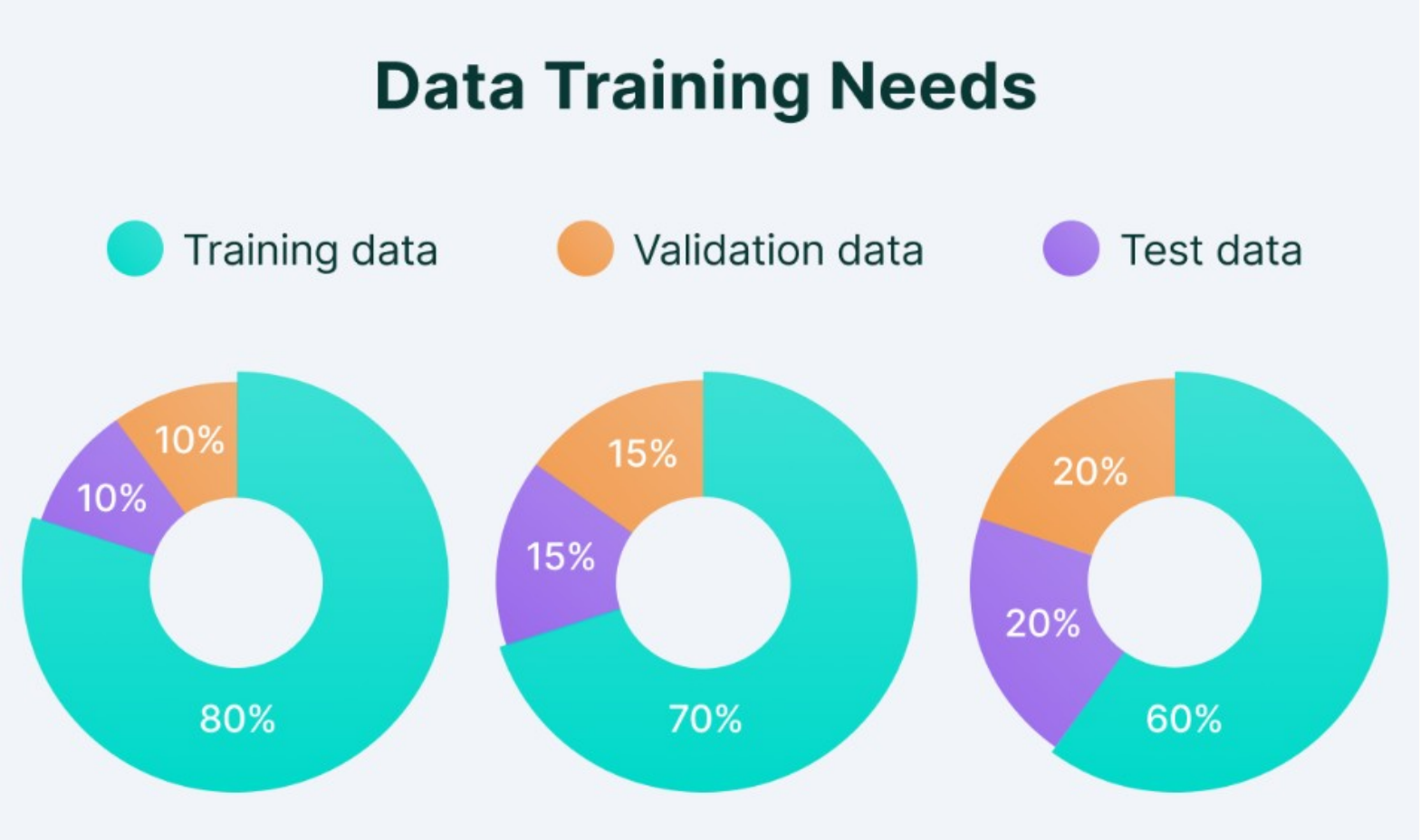
Fit the parameters.

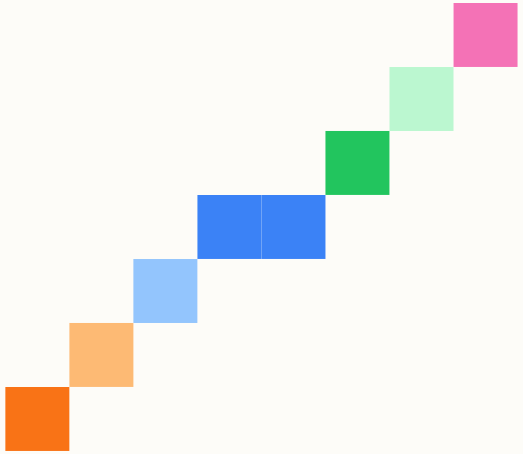
Model learns patterns directly from this specific dataset.

Test Split

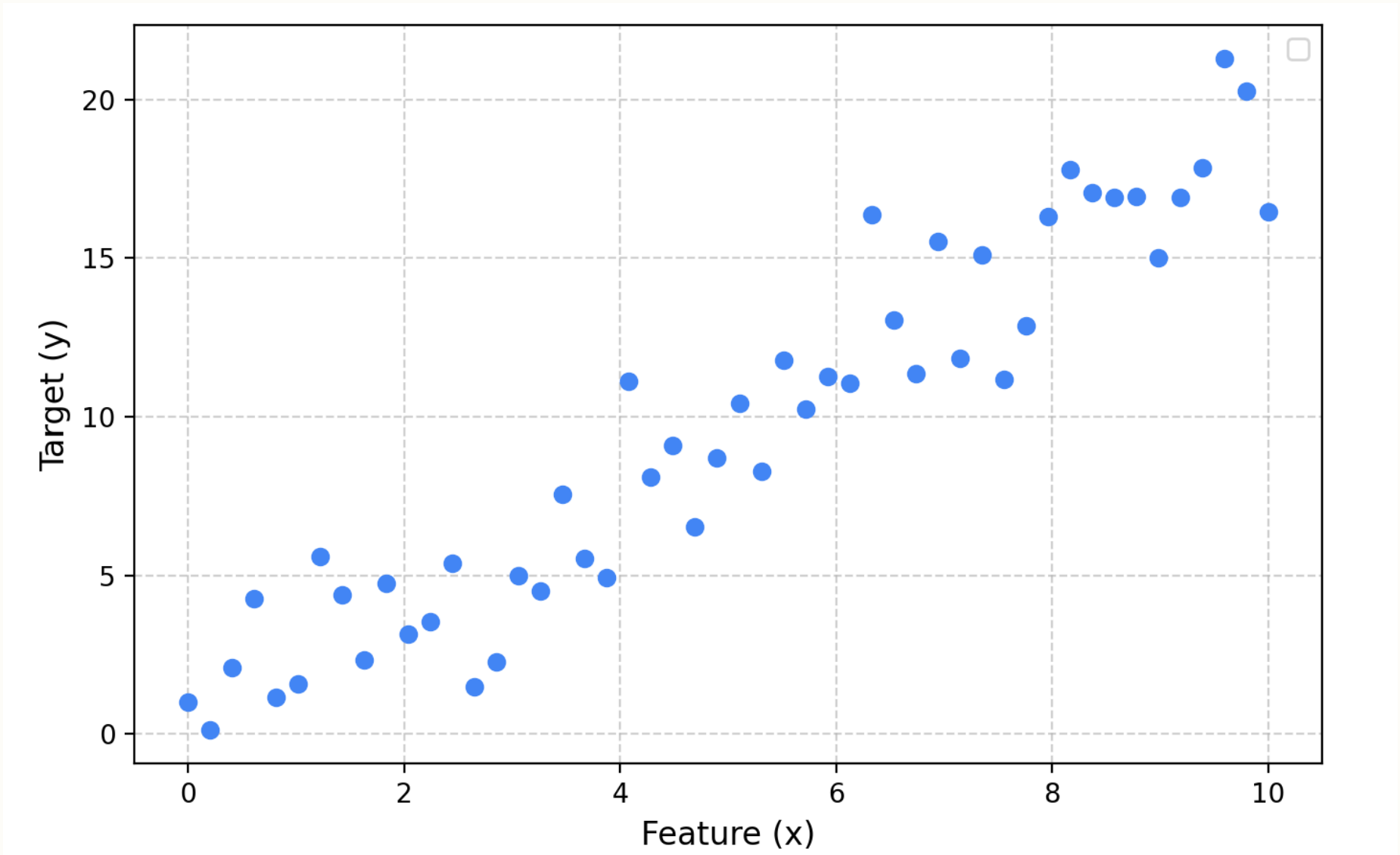
Judge the performance.

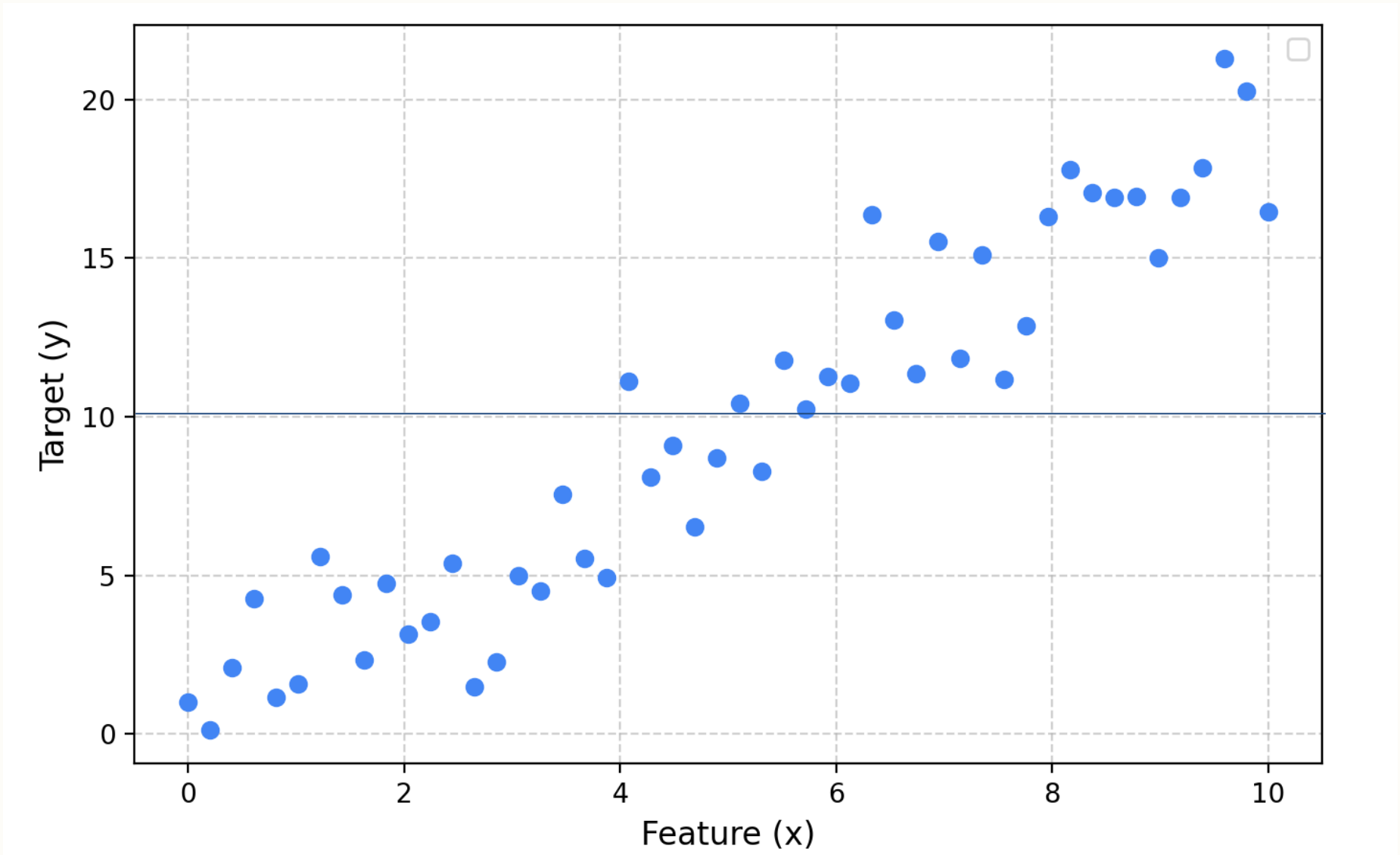
Unseen data used strictly to verify real-world accuracy.

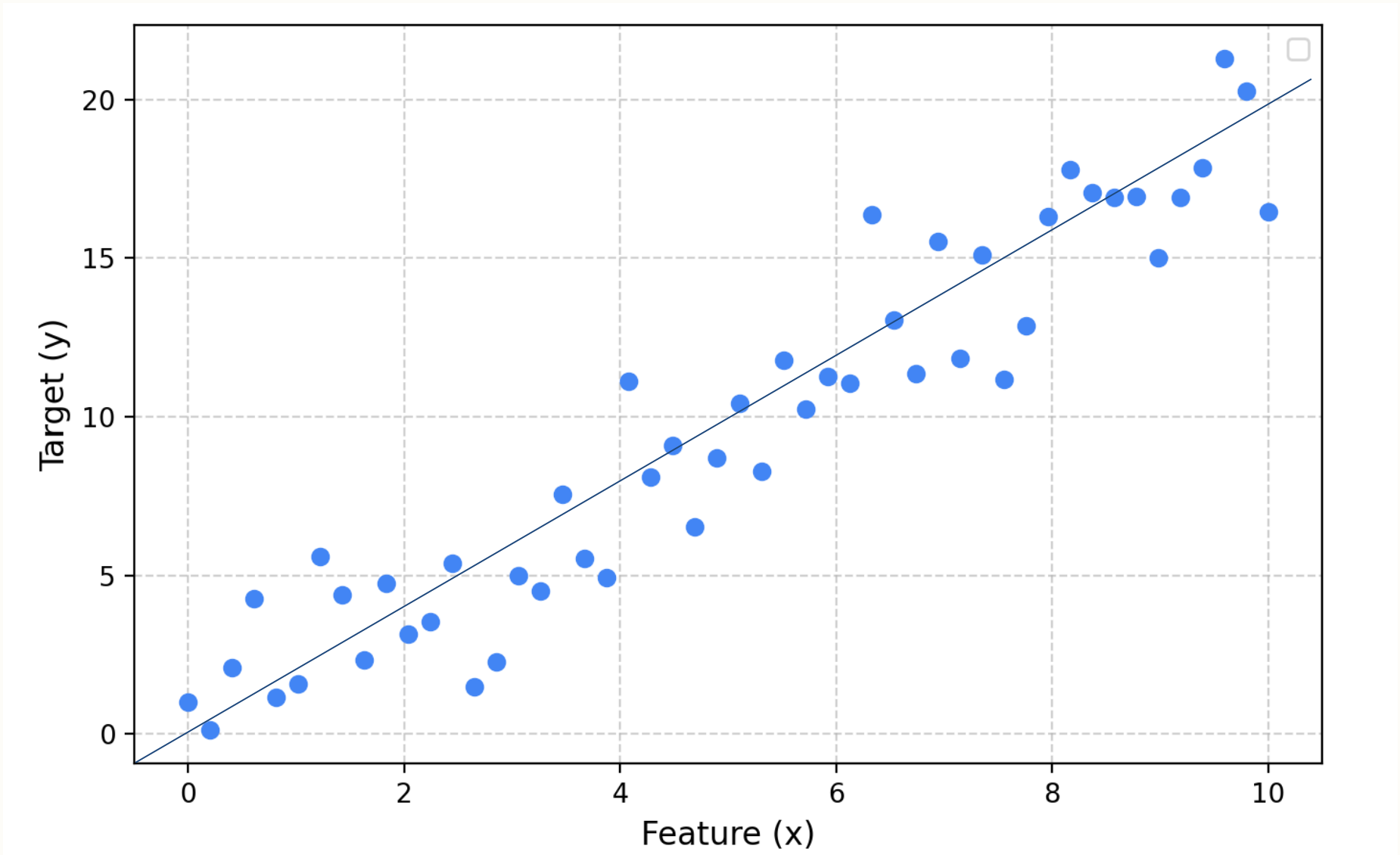




Linear Regression


Target (y)
Feature (x)
0
5
10
15
20
0
2
4
6
8
10


Target (y)
Feature (x)
0
5
10
15
20
0
2
4
6
8
10


Target (y)
Feature (x)
0
5
10
15
20
0
2
4
6
8
10

Basic Idea:

How can a machine create a line of best fit for any given graph?

Predict a Number

Output is continuous (e.g. house price).

Weights and Intercept

Learn the slope (w) and baseline (b).

Multiple Features

Expand to many input variables.

ONE FEATURE

$$\hat{y} = wx + b$$

MANY FEATURES

$$\hat{y} = w_1 x_1 + w_2 x_2 + \cdots + w_d x_d + b = w^T x + b$$

Least Squares

$$L(w,b) = \frac{1}{2n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Residuals

The difference between the actual value and the prediction.

$$e_i = y_i - \hat{y}_i$$

Squared Error

Squaring stops negative/positive signs from cancelling out and amplifies large error.

$$e_i^2 = (y_i - \hat{y}_i)^2$$

Solving It

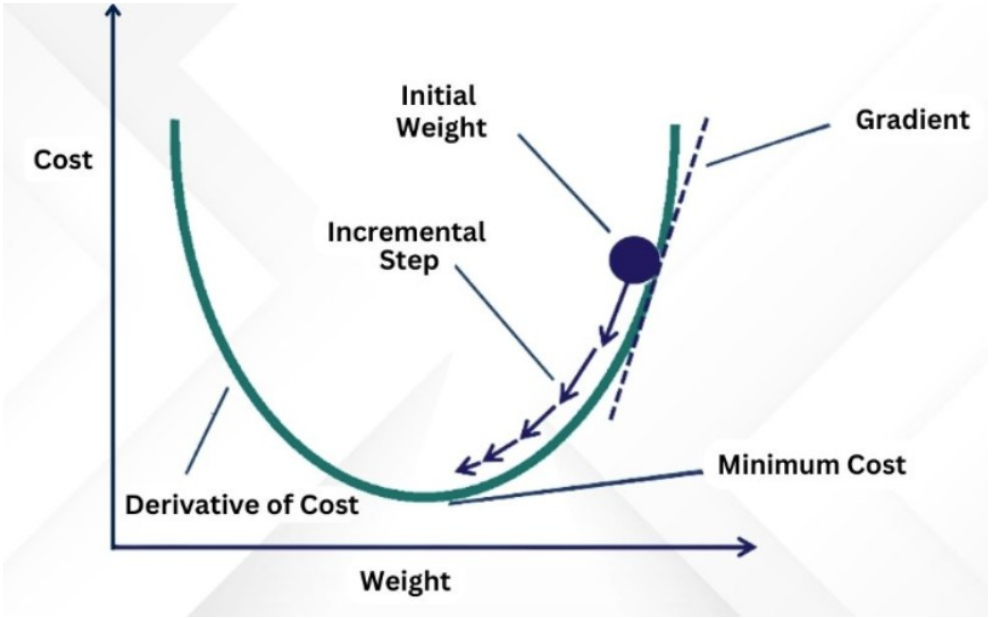
Method 1: Finding the minimum directly

Solving It

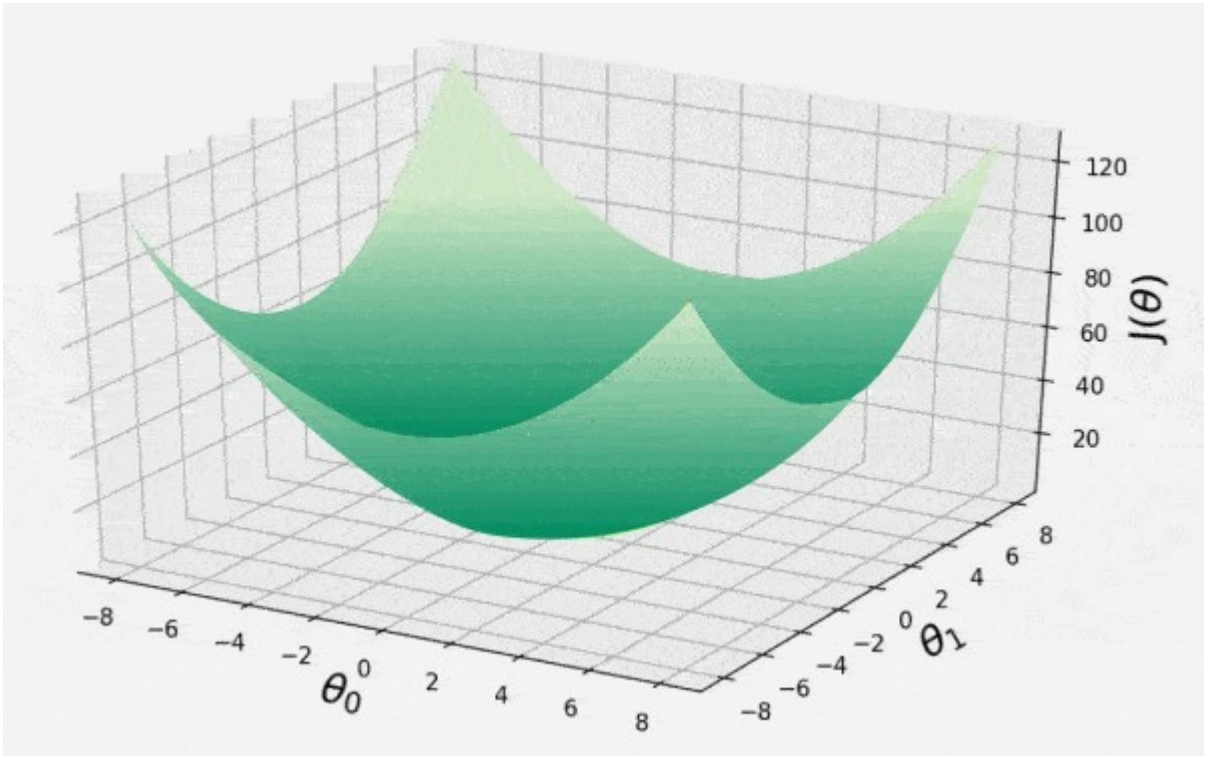
Method 2: Using Gradient Descent

KEY IDEA The loss function is a bowl. Gradient descent rolls to the bottom.

Solving It



From https://medium.com/@datasciencewizards/a-simple-guide-to-gradient-descent-algorithm-60cbb66a0df9



From https://www.geeksforgeeks.org/machine-learning/what-is-gradient-descent/

Solving It

Normal Equations

Exact closed-form mathematical solution in a single step.

$$w = (X^T X)^{-1} X^T y$$

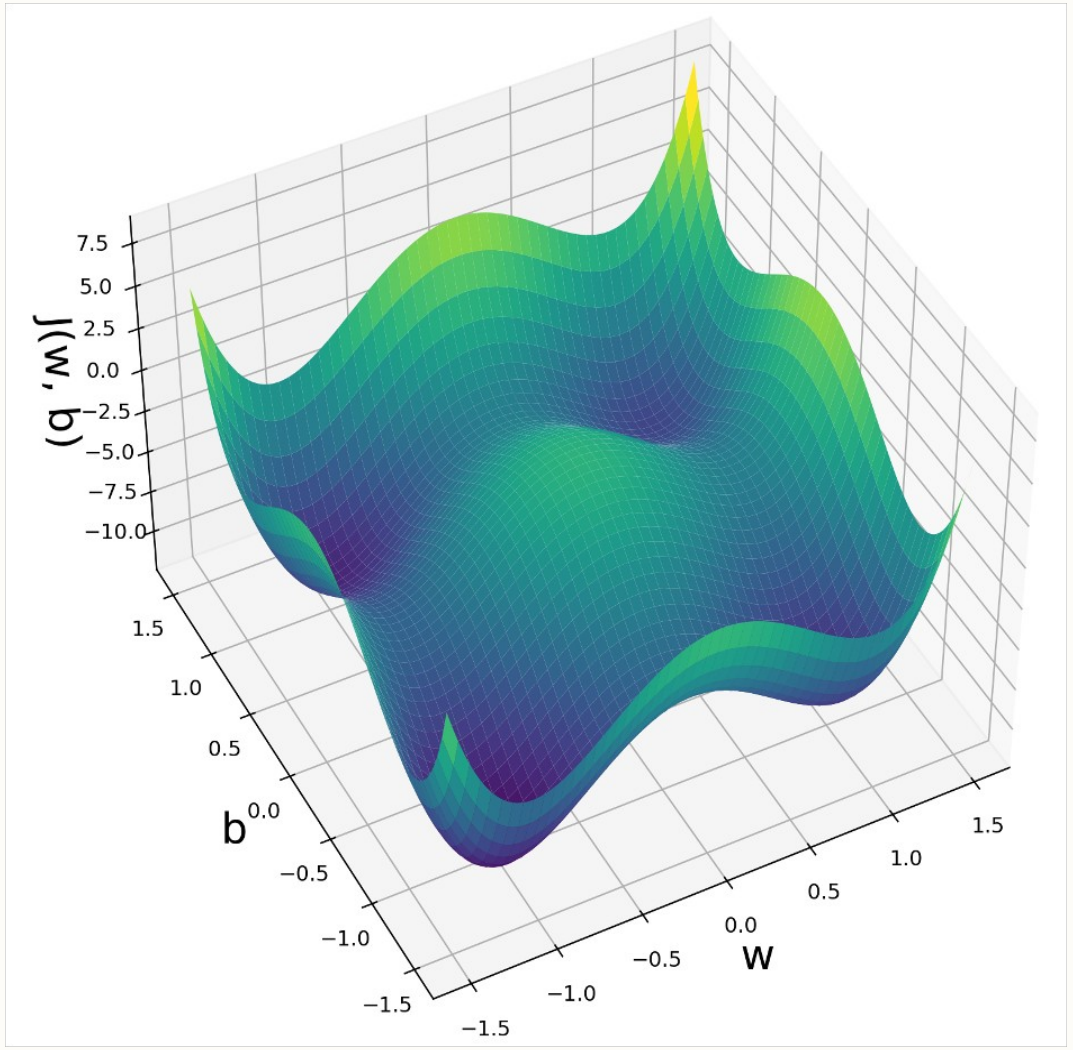
Gets very slow with millions of features.

Gradient Descent

Repeatedly step downhill opposite to the gradient.

$$w = w - lr \frac{\partial}{\partial w} J(w, b) \qquad b = b - lr \frac{\partial}{\partial b} J(w, b)$$

Scales efficiently to massive datasets.



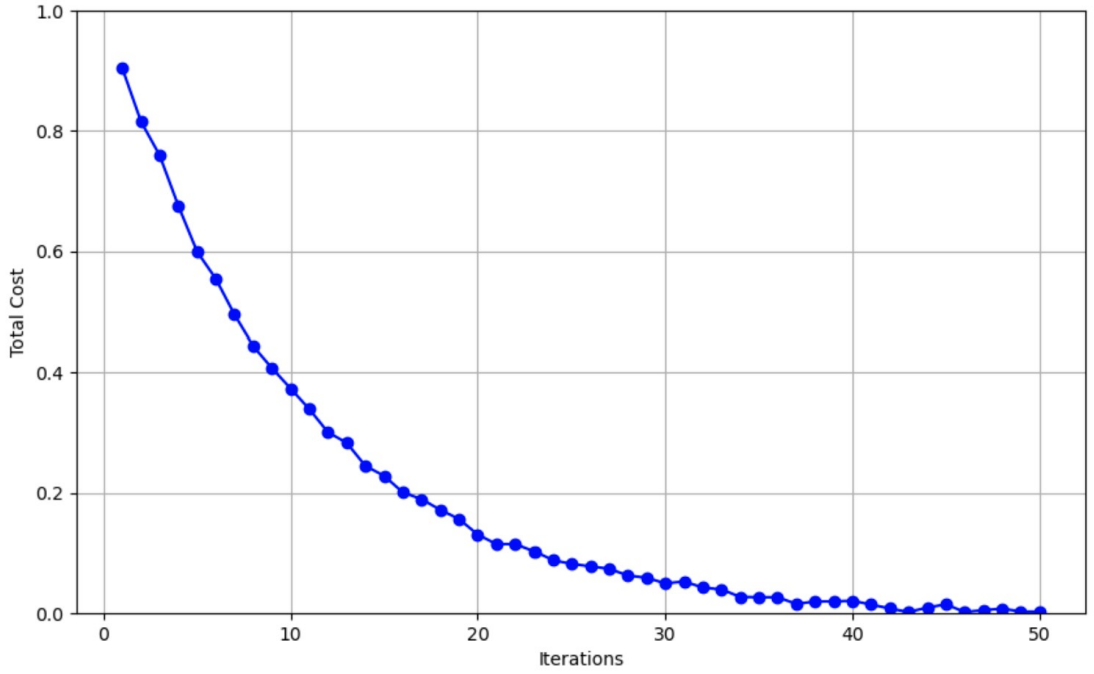
Solving It

Gradient Descent

Repeatedly step downhill opposite to the gradient.

$$w = w - lr\frac{\partial}{\partial w}J(w,b) \qquad b = b - lr\frac{\partial}{\partial b}J(w,b)$$

Scales efficiently to massive datasets.



Where do we stop?