How does Machine Learning Work

Data: {x_i, y_i}
i=1, 2, …, n

Goal: learn w
y <—- $f(x; w)$

Minimize an objective function

$$F(w) = \sum_{i=1}^{n} \text{loss}(y_i, f(x_i; w))$$

- Collect Data
- Determine a **function space**
- Pick a **loss function**
- Pick an **optimizer**

Optimizer

1. The minimizer w^* may be in closed form.
2. Try optimization algorithms that can guarantee to converge to the global minimizer.
3. In the worst case, try *gradient descent*.

Loss Function/Metric

1. Regression; Classification
2. https://keras.io/losses/
3. https://keras.io/metrics/

Gradient Descent and Stochastic Gradient Descent

Derivative

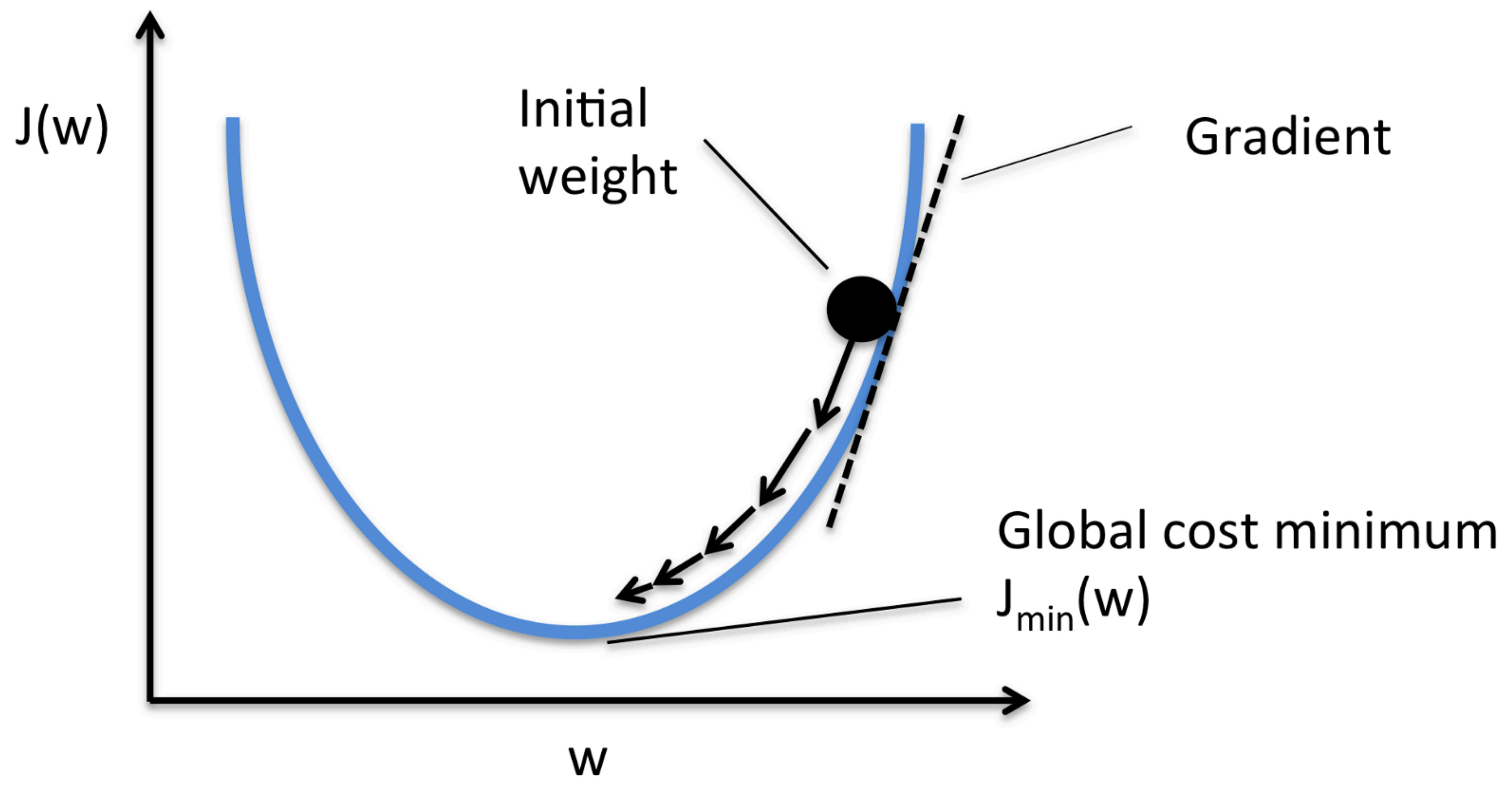
$$f'(x) = \lim_{\delta \to 0} \frac{f(x+\delta) - f(x)}{\delta}$$

$$f(x_0 + \delta) \approx f(x_0) + \delta \cdot f'(x_0)$$

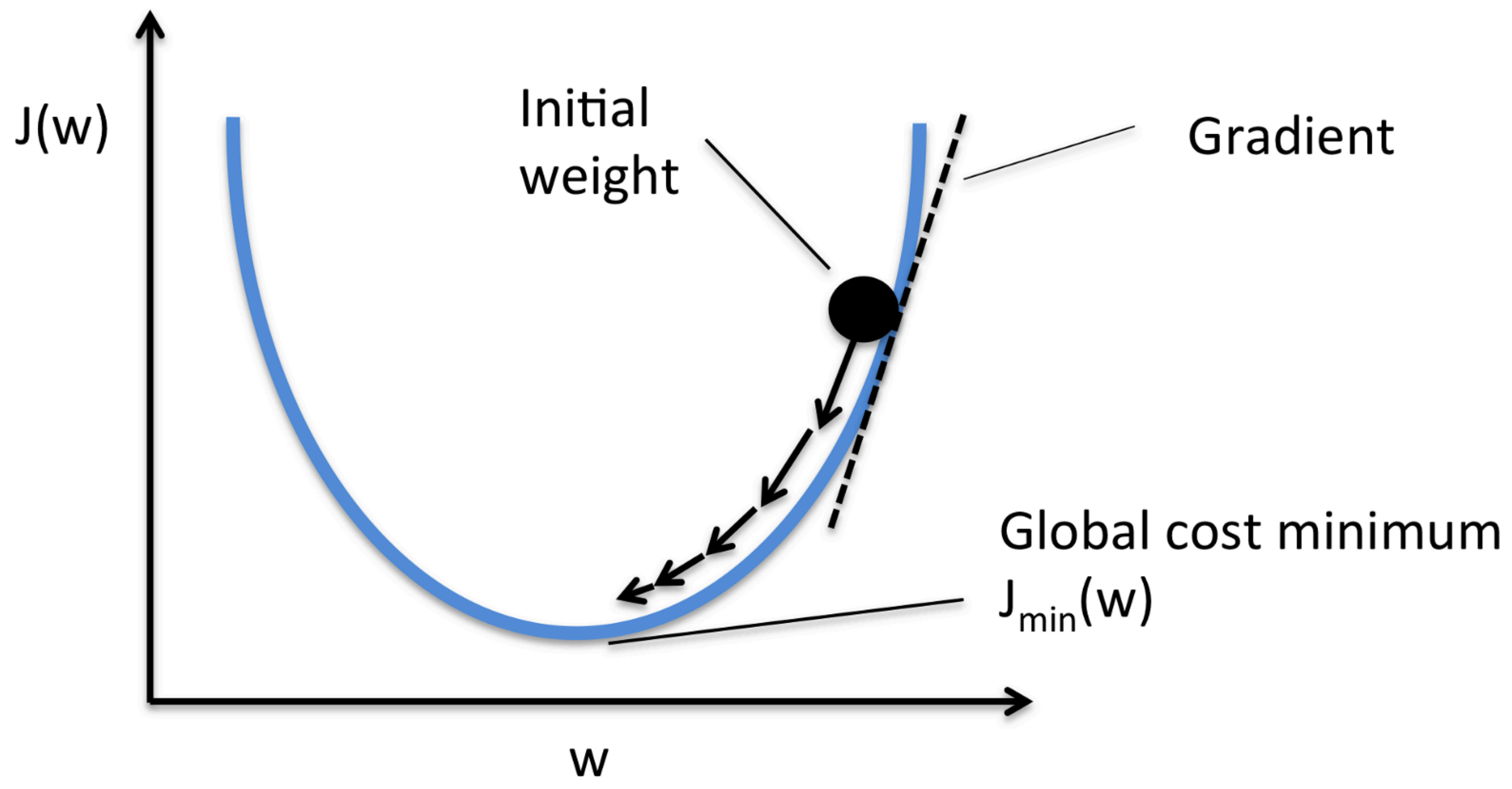
Gradient Descent

$$x_{n+1} = x_n - \gamma_n f'(x_n)$$

Learning Rate



Gradient Descent and Stochastic Gradient Descent



Derivative

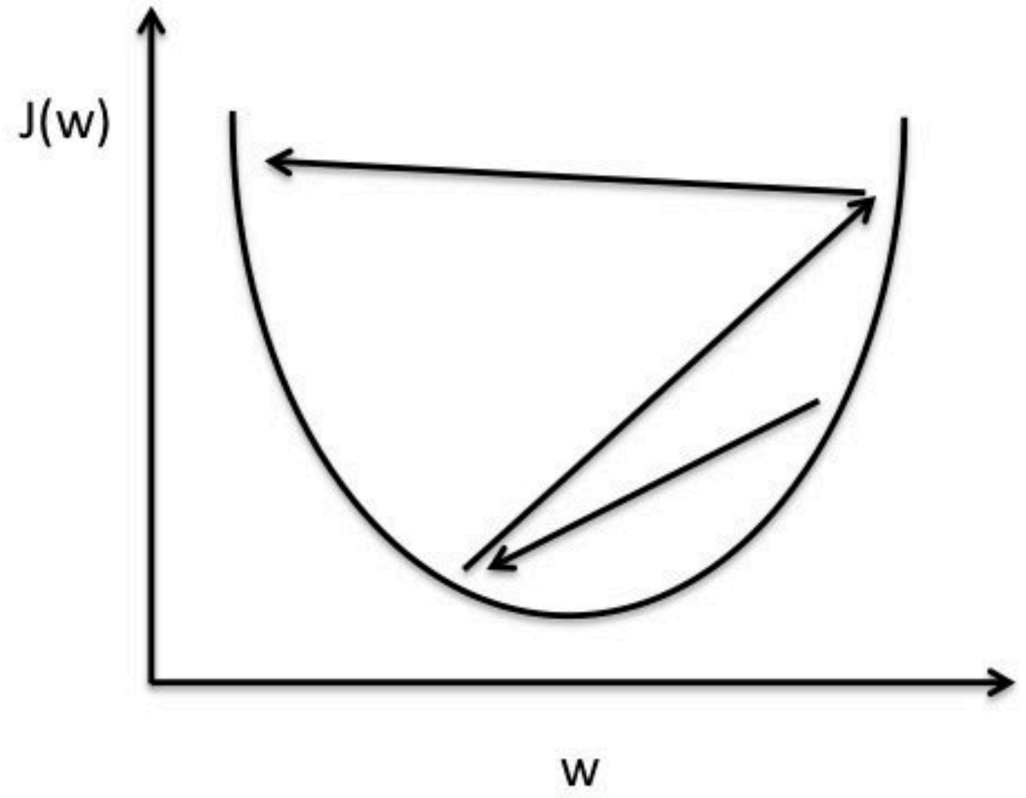
$$f'(x) = \lim_{\delta \to 0} \frac{f(x+\delta) - f(x)}{\delta}$$

$$f(x_0 + \delta) \approx f(x_0) + \delta \cdot f'(x_0)$$

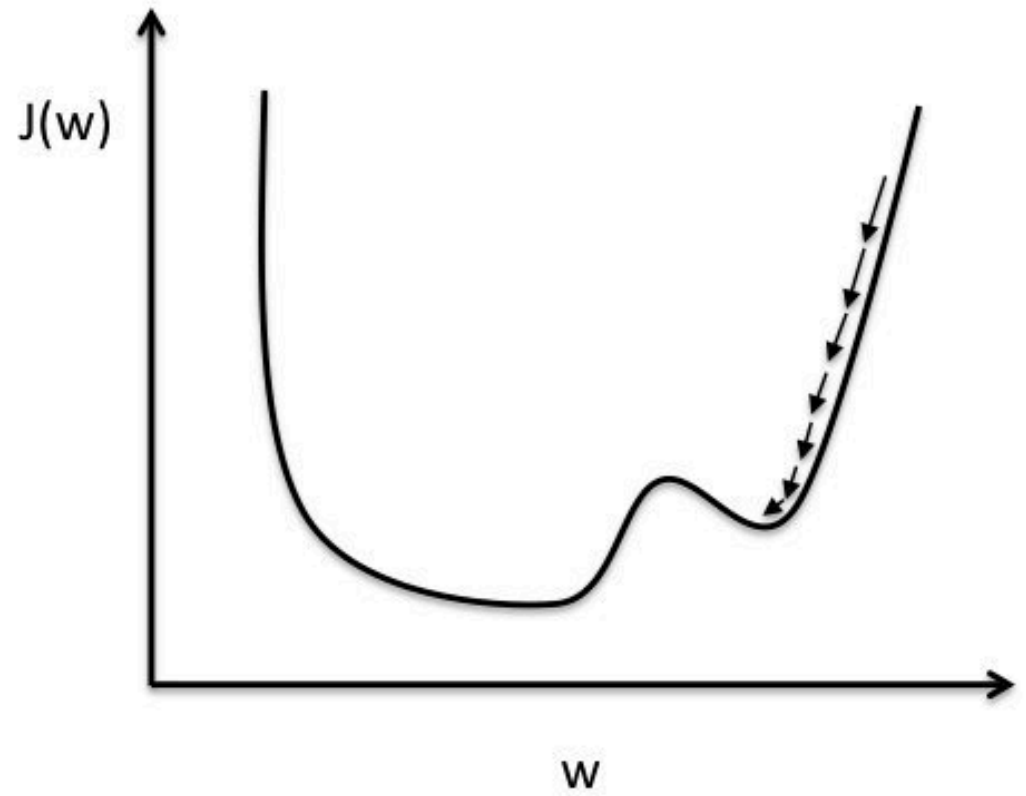
Gradient Descent

$$x_{n+1} = x_n - \gamma_n f'(x_n)$$

Learning Rate



Large learning rate: Overshooting.



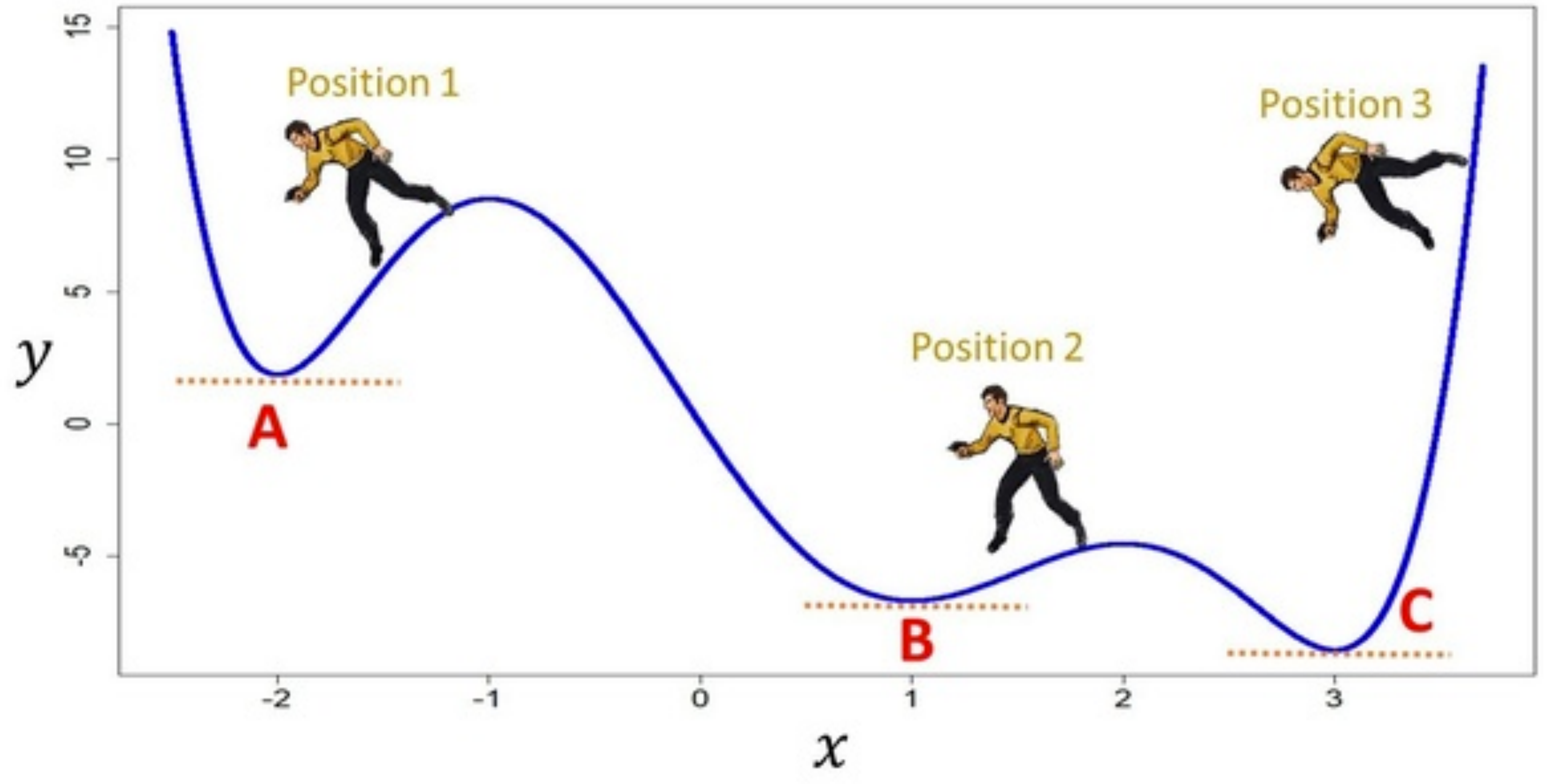
Small learning rate: Many iterations until convergence and trapping in local minima.

SGD

$$F(x) = \sum_{i=1}^{n} f_i(x)$$

$$F'(x) = \sum_{i=1}^{n} f_i'(x)$$

$$F'_{\mathrm{SG}}(x) \propto \sum_{i \in I_t} f_i'(x)$$



Gradient Descent and Stochastic Gradient Descent

Partial Derivative

$$\nabla f(\mathbf{x}) = \begin{pmatrix} \frac{\partial f(\mathbf{x})}{\partial x_1} \\ \vdots \\ \frac{\partial f(\mathbf{x})}{\partial x_d} \end{pmatrix}$$

Gradient Descent

$$\mathbf{x}_{n+1} = \mathbf{x}_n - \gamma_n \cdot \nabla(\mathbf{x}_n)$$

How to Improve Learning Rate?

1. **Reduce Oscillation**
2. **Adaptive to different dims**

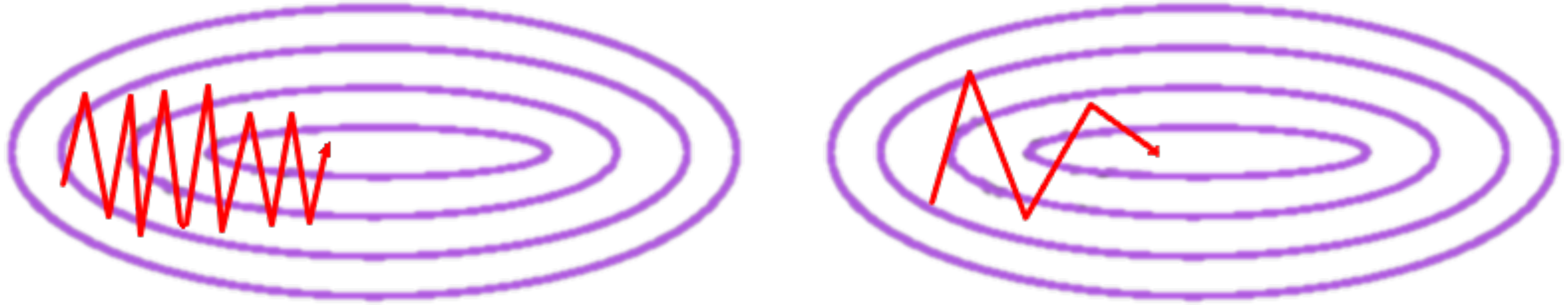
https://keras.io/optimizers/

SGD

$$F(x) = \sum_{i=1}^{n} f_i(x)$$

$$F'(x) = \sum_{i=1}^{n} f_i'(x)$$

$$F'_{\mathrm{SG}}(x) \propto \sum_{i \in I_t} f_i'(x)$$



Chain Rule

Function Space: How to Go Beyond Linear Functions

Suppose $x \in \mathbf{R}^1$ (univariate case).

- Polynomial regression

$$f(x) = \alpha_0 + \alpha_1 x + \cdots + \alpha_d x^d.$$

- Spline models: regression splines or smoothing splines.

- Local polynomial regression: Loess

What else ?

SVM,
Tree Models,
Model Ensemble,

Suppose $\mathbf{x} \in \mathbf{R}^p$ (multivariate)

- For a polynomial model with degree 3, how many terms in the model?

$$\{X_j\}_{j=1}^p, \{X_j^2\}, \{X_j^3\}, \{X_j X_l\}, \{X_j^2 X_l\}, \{X_j X_k X_l\}.$$

  COD (Curse Of Dimensionality): num of parameters explodes.

- A Compromise: Additive model

$$f(\mathbf{X}) = \mu + f_1(X_1) + f_2(X_2) + \cdots + f_p(X_p).$$

  Drawback: totally ignore interactions among the p covariates.

Function Space: How to Go Beyond Linear Functions

Suppose $x \in \mathbf{R}^1$ (univariate case).

- Polynomial regression

$$f(x) = \alpha_0 + \alpha_1 x + \cdots + \alpha_d x^d.$$

- Spline models: regression splines or smoothing splines.
- Local polynomial regression: Loess

Suppose $\mathbf{x} \in \mathbf{R}^p$ (multivariate)

- For a polynomial model with degree 3, how many terms in the model?

$$\{X_j\}_{j=1}^p, \{X_j^2\}, \{X_j^3\}, \{X_j X_l\}, \{X_j^2 X_l\}, \{X_j X_k X_l\}.$$

COD (Curse Of Dimensionality): num of parameters explodes.

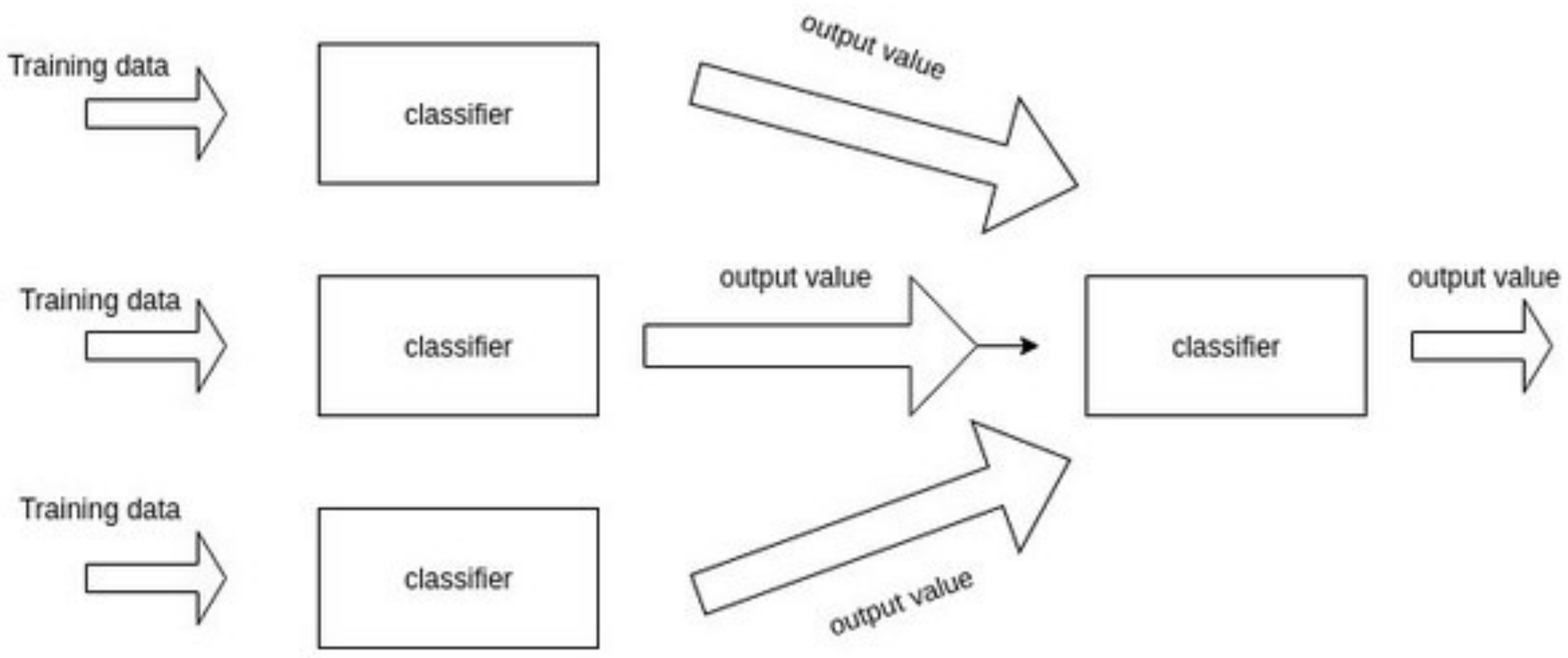
- A Compromise: Additive model

$$f(\mathbf{X}) = \mu + f_1(X_1) + f_2(X_2) + \cdots + f_p(X_p).$$

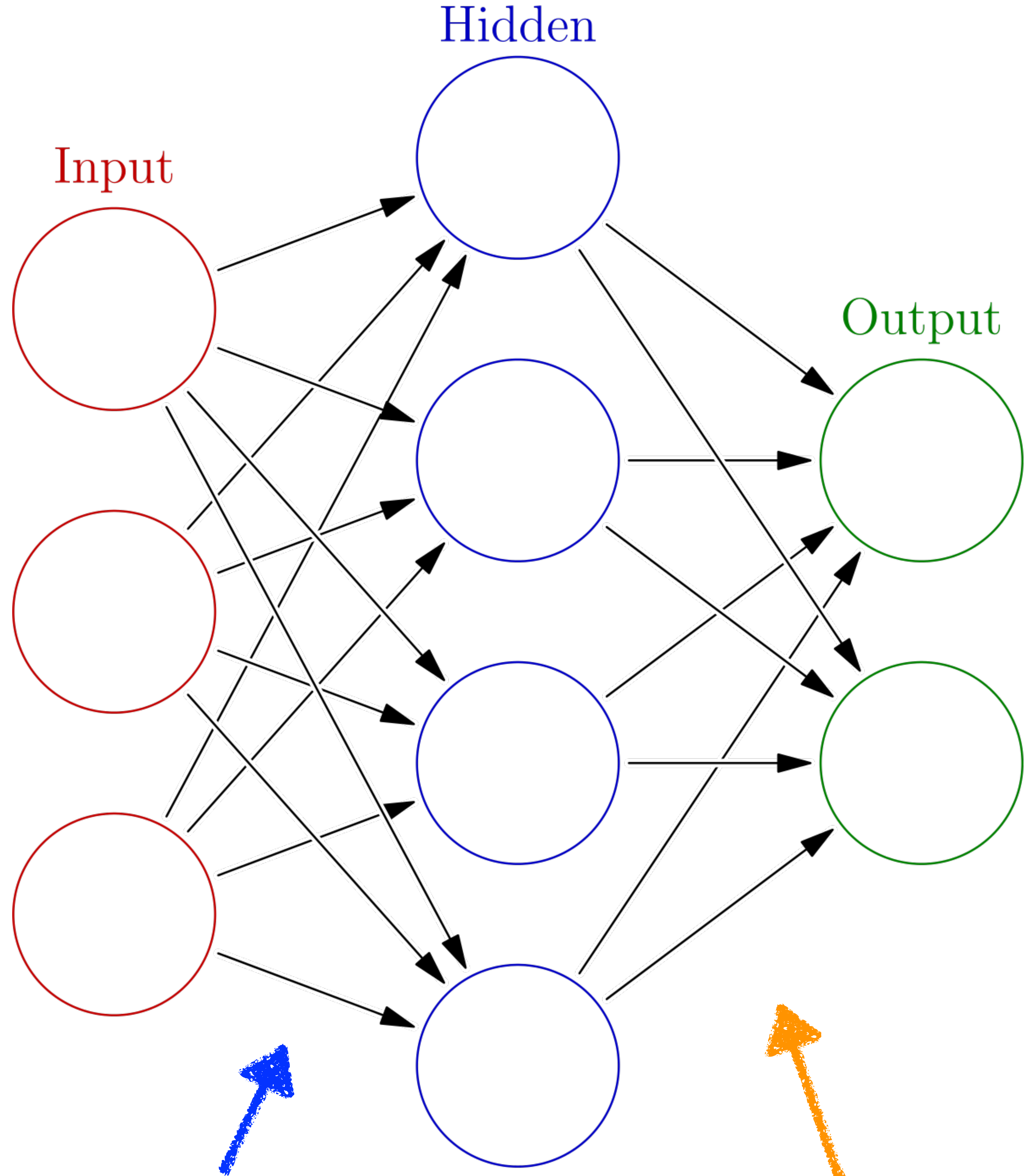
Drawback: totally ignore interactions among the p covariates.

What else ?

SVM,
Tree Models,
Model Ensemble,
Model Stacking



Neutral Networks



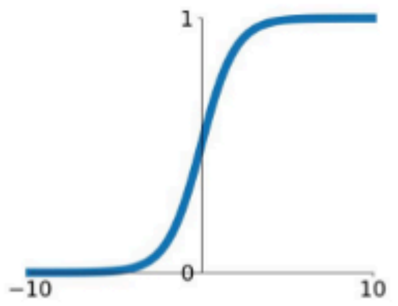
(W_4x3, c_4x1)

(w_2x4, b_2x1)

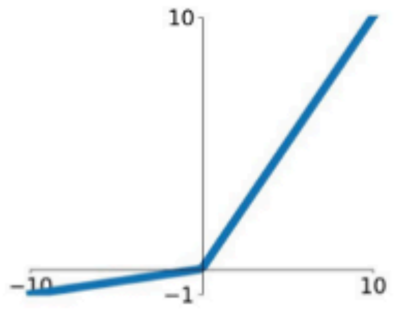
$$h_1 = \sigma(x_1 \cdot W_{1,1} + x_2 \cdot W_{1,2} + x_3 \cdot W_{1,3} + b_1)$$
$$h_2 = \sigma(x_1 \cdot W_{2,1} + x_2 \cdot W_{2,2} + x_3 \cdot W_{2,3} + b_2)$$
$$h_3 = \sigma(x_1 \cdot W_{3,1} + x_2 \cdot W_{3,2} + x_3 \cdot W_{3,3} + b_3)$$
$$h_4 = \sigma(x_1 \cdot W_{4,1} + x_2 \cdot W_{4,2} + x_3 \cdot W_{4,3} + b_4)$$
$$\mathbf{h} = \sigma(\mathbf{W}\mathbf{x} + \mathbf{c}) \quad \mathbf{y} = \mathbf{w}\mathbf{h} + \mathbf{b}$$

Activation Functions

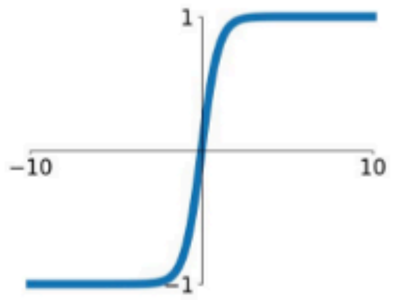
Sigmoid
$\sigma(x) = \frac{1}{1+e^{-x}}$



Leaky ReLU
$\max(0.1x, x)$

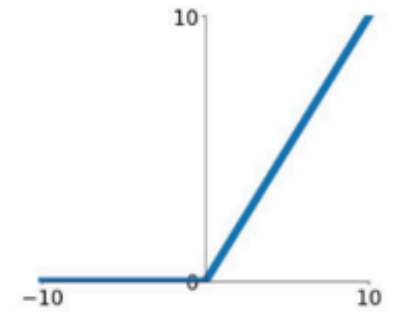


tanh
$\tanh(x)$

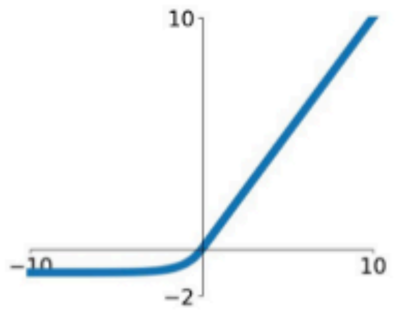


Maxout
$\max(w_1^T x + b_1, w_2^T x + b_2)$

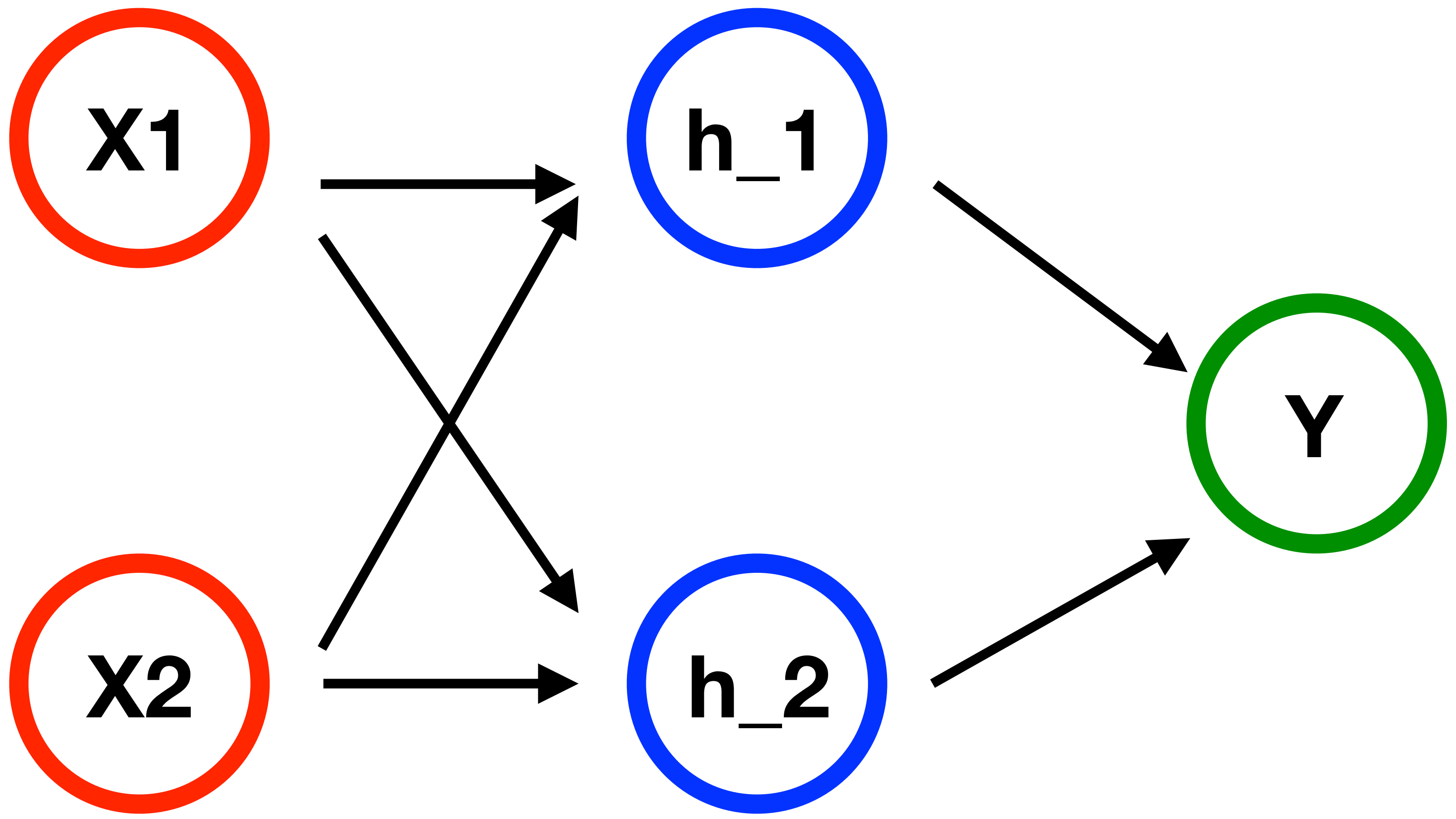
ReLU
$\max(0, x)$



ELU
$$\begin{cases} x & x \geq 0 \\ \alpha(e^x - 1) & x < 0 \end{cases}$$



A Simple Example of NN



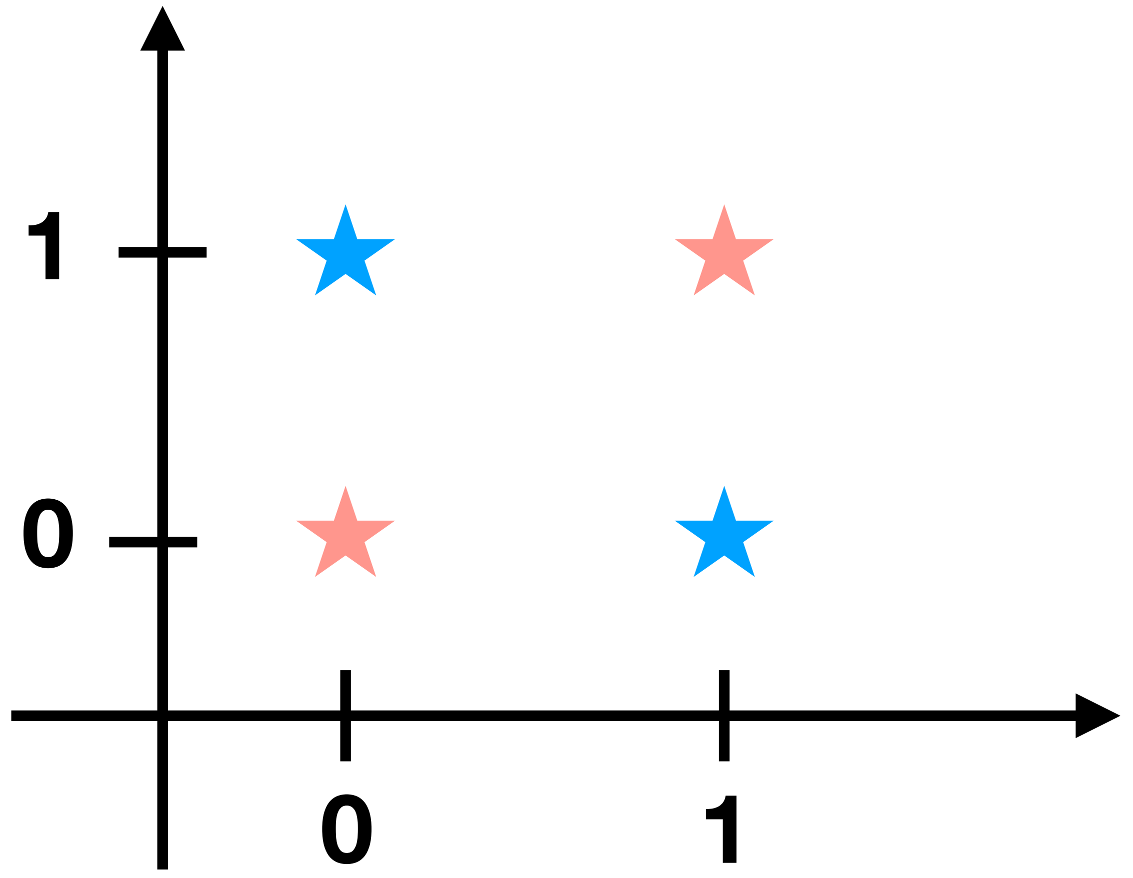
$$\mathbf{W} = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$$

$$\mathbf{c} = \begin{pmatrix} 0 \\ -1 \end{pmatrix}$$

$$\mathbf{w} = \begin{pmatrix} 1 \\ -2 \end{pmatrix}$$

$$b = 0$$

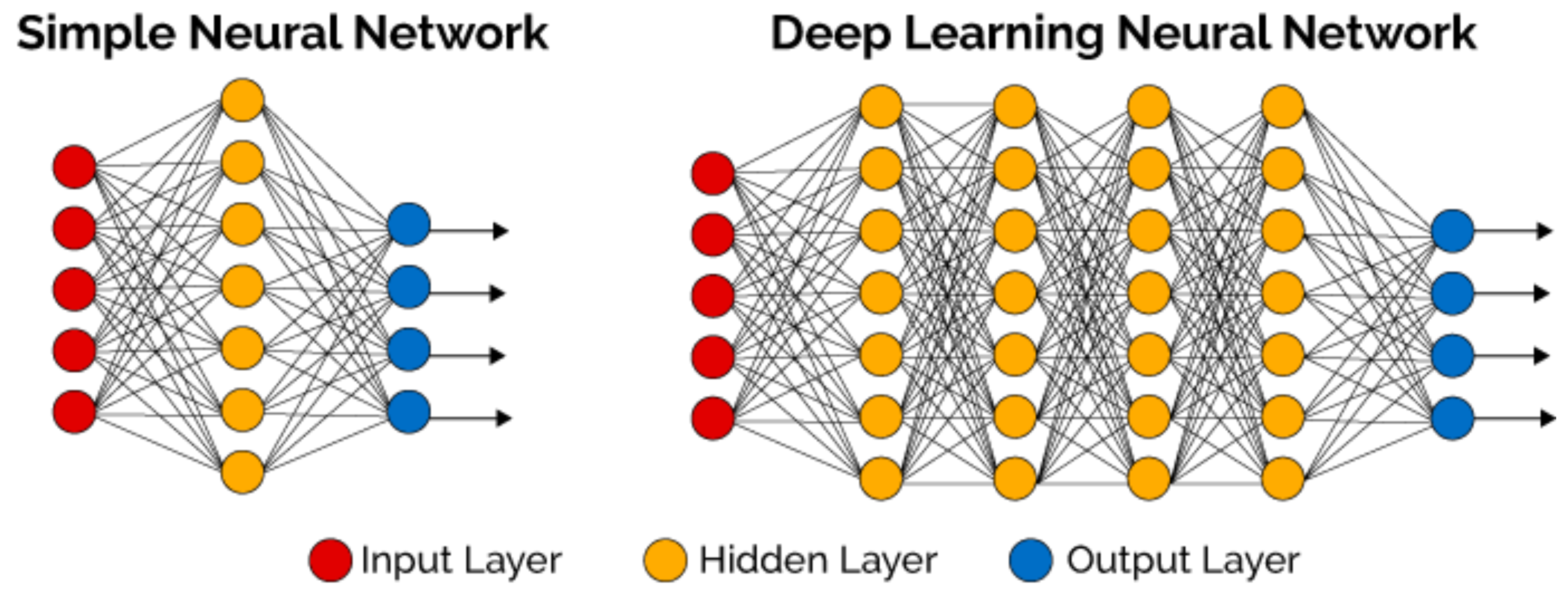
$$y = \mathbf{w}\sigma(\mathbf{W}\mathbf{x} + \mathbf{c}) + b$$



Input	Hidden	Output
(0, 0)		0
(1, 1)		0
(0, 1)		-1
(1, 0)		-1

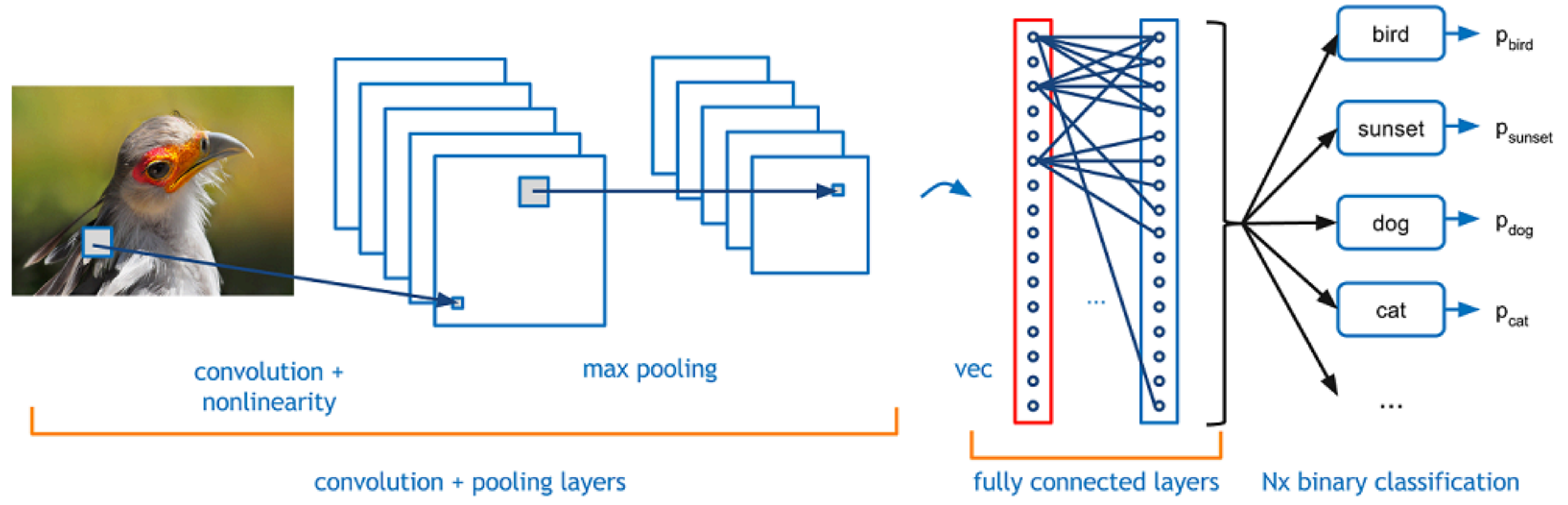
Deep Neutral Networks

Fully Connected

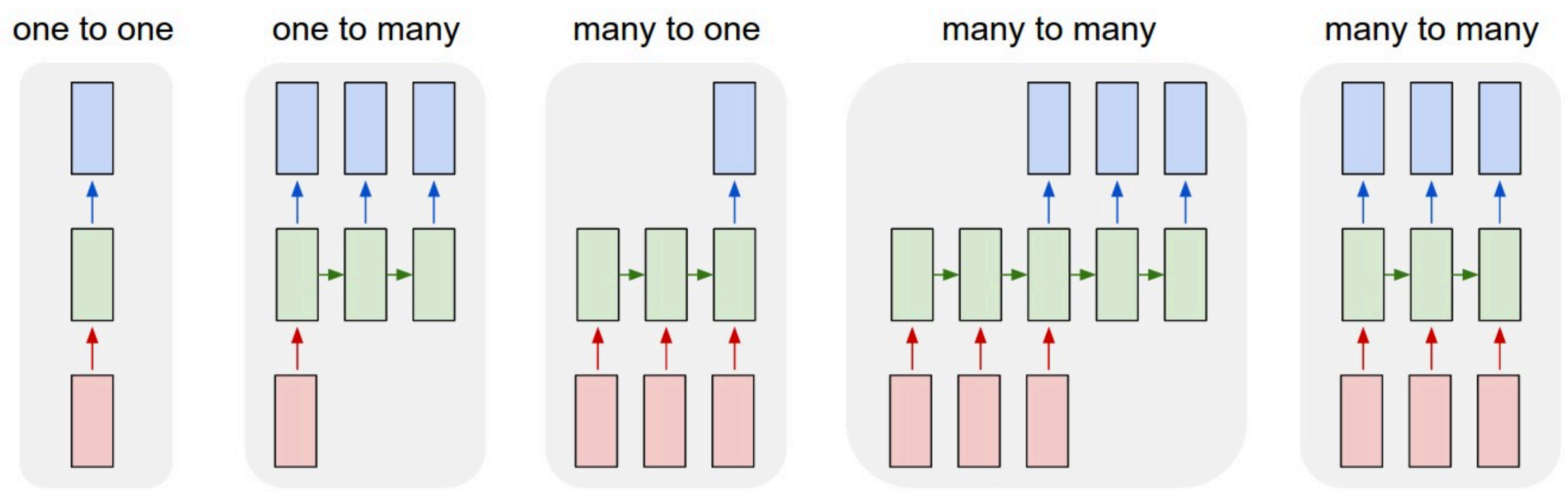


- Collect Data
- Pick a Network Architecture
- Pick a **loss function**
- Pick an **optimizer**

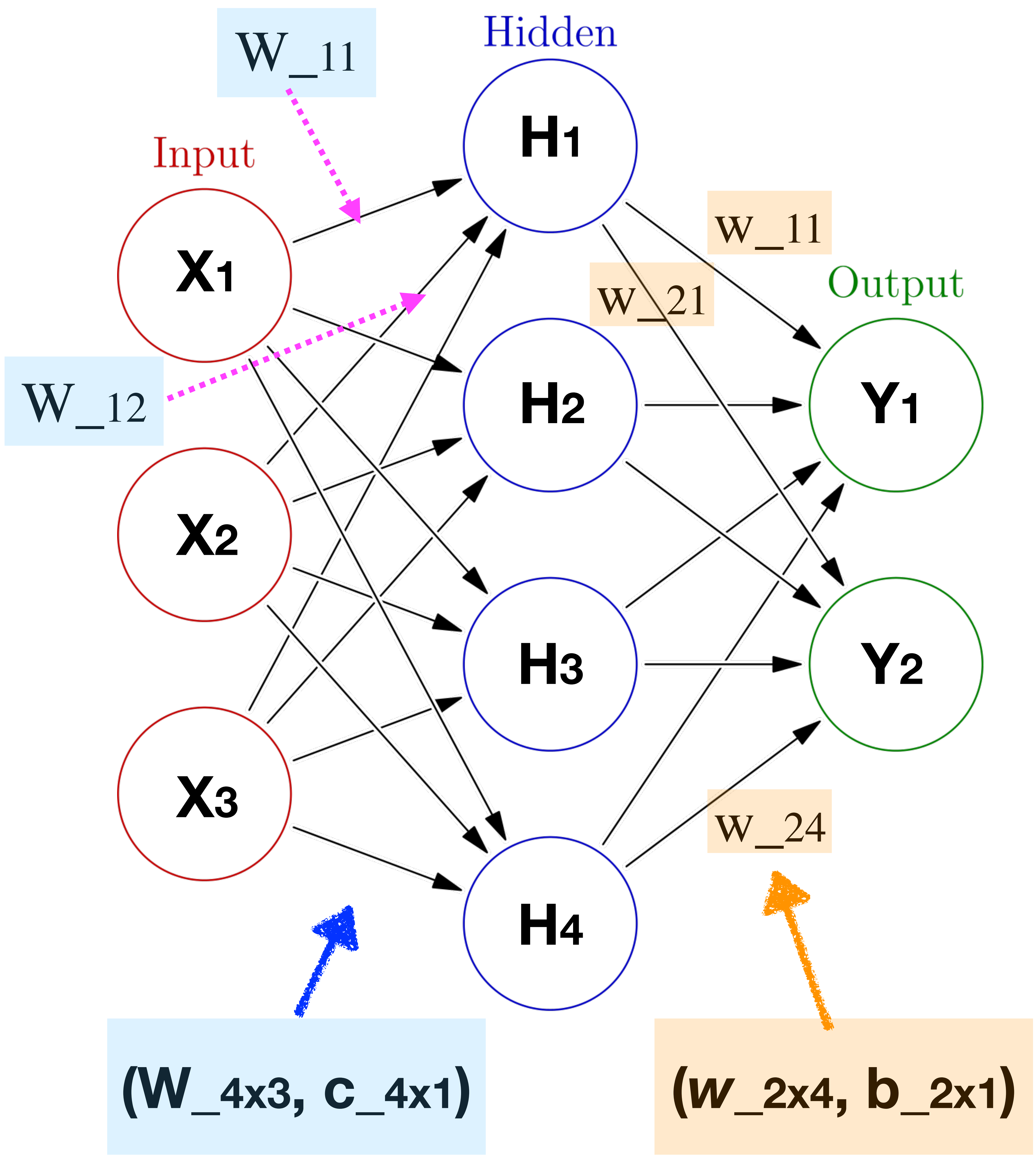
CNN



RNN



Chain Rule and Back-Propagation



$$h_1 = \sigma(x_1 \cdot W_{1,1} + x_2 \cdot W_{1,2} + x_3 \cdot W_{1,3} + b_1)$$
$$h_2 = \sigma(x_1 \cdot W_{2,1} + x_2 \cdot W_{2,2} + x_3 \cdot W_{2,3} + b_2)$$
$$h_3 = \sigma(x_1 \cdot W_{3,1} + x_2 \cdot W_{3,2} + x_3 \cdot W_{3,3} + b_3)$$
$$h_4 = \sigma(x_1 \cdot W_{4,1} + x_2 \cdot W_{4,2} + x_3 \cdot W_{4,3} + b_4)$$
$$\mathbf{h} = \sigma(\mathbf{W}\mathbf{x} + \mathbf{c}) \quad \mathbf{y} = \mathbf{w}\mathbf{h} + \mathbf{b}$$

$$\frac{\partial y_1}{\partial w_{11}} = h_1$$
$$\frac{\partial y_2}{\partial w_{21}} = h_1$$
$$\frac{\partial y_i}{\partial w_{il}} = h_l$$

$$\frac{\partial y_1}{\partial W_{11}} = \frac{\partial y_1}{\partial h_1}\frac{\partial h_1}{\partial z_1}\frac{\partial z_1}{\partial W_{11}}$$
$$\frac{\partial y_1}{\partial W_{12}} = \frac{\partial y_1}{\partial h_1}\frac{\partial h_1}{\partial z_1}\frac{\partial z_1}{\partial W_{12}}$$
$$\frac{\partial y_i}{\partial W_{jl}} = \frac{\partial y_i}{\partial h_j}\frac{\partial h_j}{\partial z_j}\frac{\partial z_j}{\partial W_{jl}}$$

What about Multiple Hidden Layers?