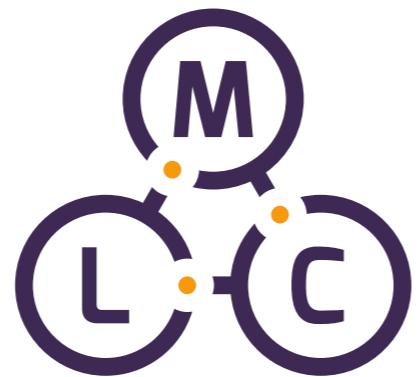


Introduction to Machine Learning and Deep Learning for Phonexia

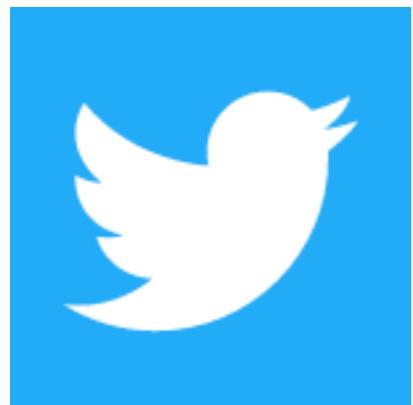
Jiří Materna



Machine
Learning
College



@mlcollegecom



@mlcollegecom



#mlcollege

About me

- Ph.D. in Natural Language Processing and Artificial Intelligence at Masaryk University
- 10 years at seznam.cz (last 8 years as Head Of Research)
- Founder and lecturer at ML College
- Founder and co-organizer of ML Prague
- Mentor at StartupYard
- ML Freelancer and consultant

Outline

Day 1

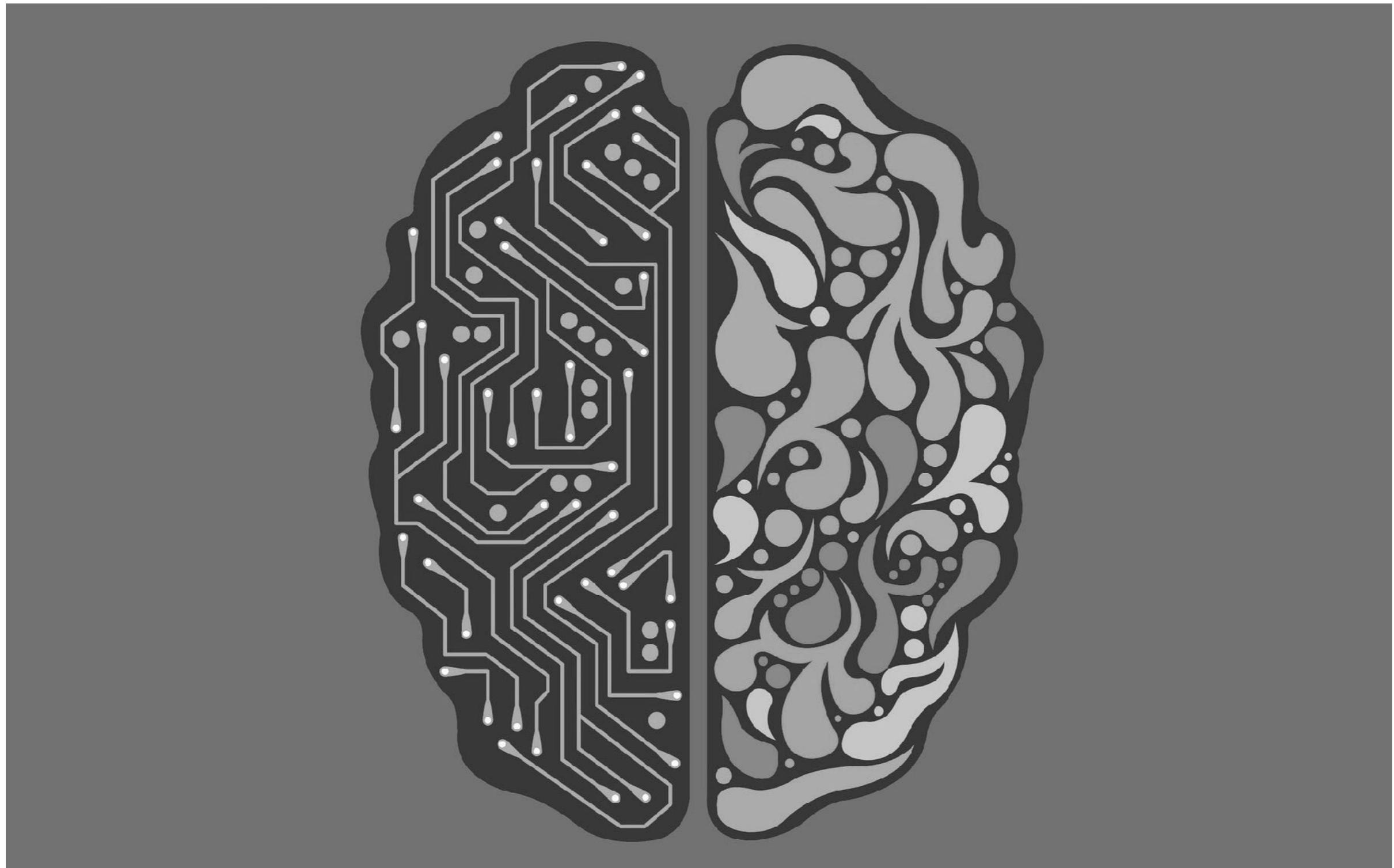
- Introduction to machine learning
- Types of ML tasks
- Data preparation
- Model evaluation
- Basic classification algorithms
- Scikit-learn tutorial
- Practical classification task
- Basic regression algorithms
- Regression model evaluation
- Practical regression task

Outline

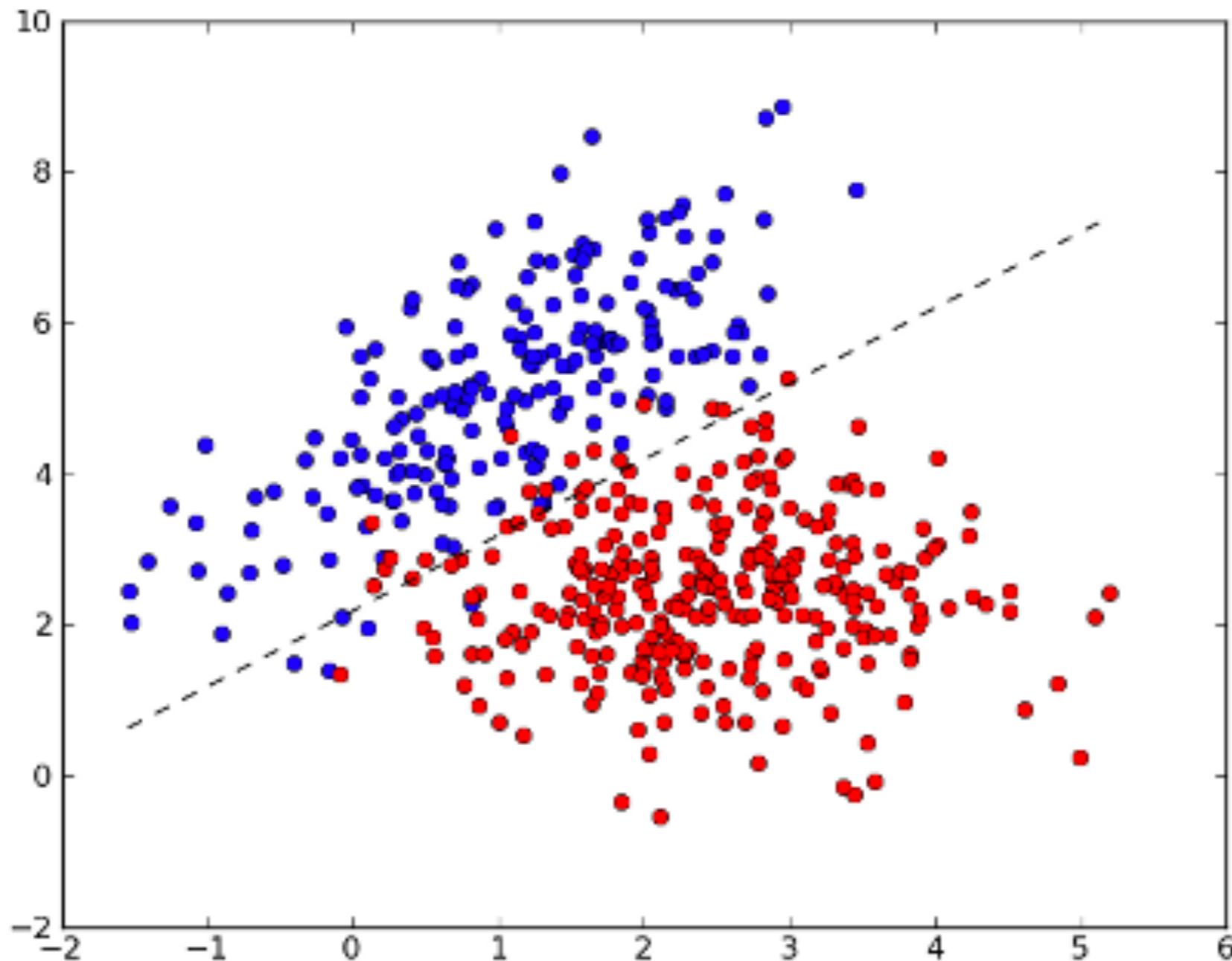
Day 2

- Introduction to neural networks
- Activation functions for neural networks
- Multilayered neural networks
- Methods for training neural networks
- Keras tutorial
- Practical classification and regression tasks solved using neural networks
- Weight initialization
- Neural network regularization and normalization

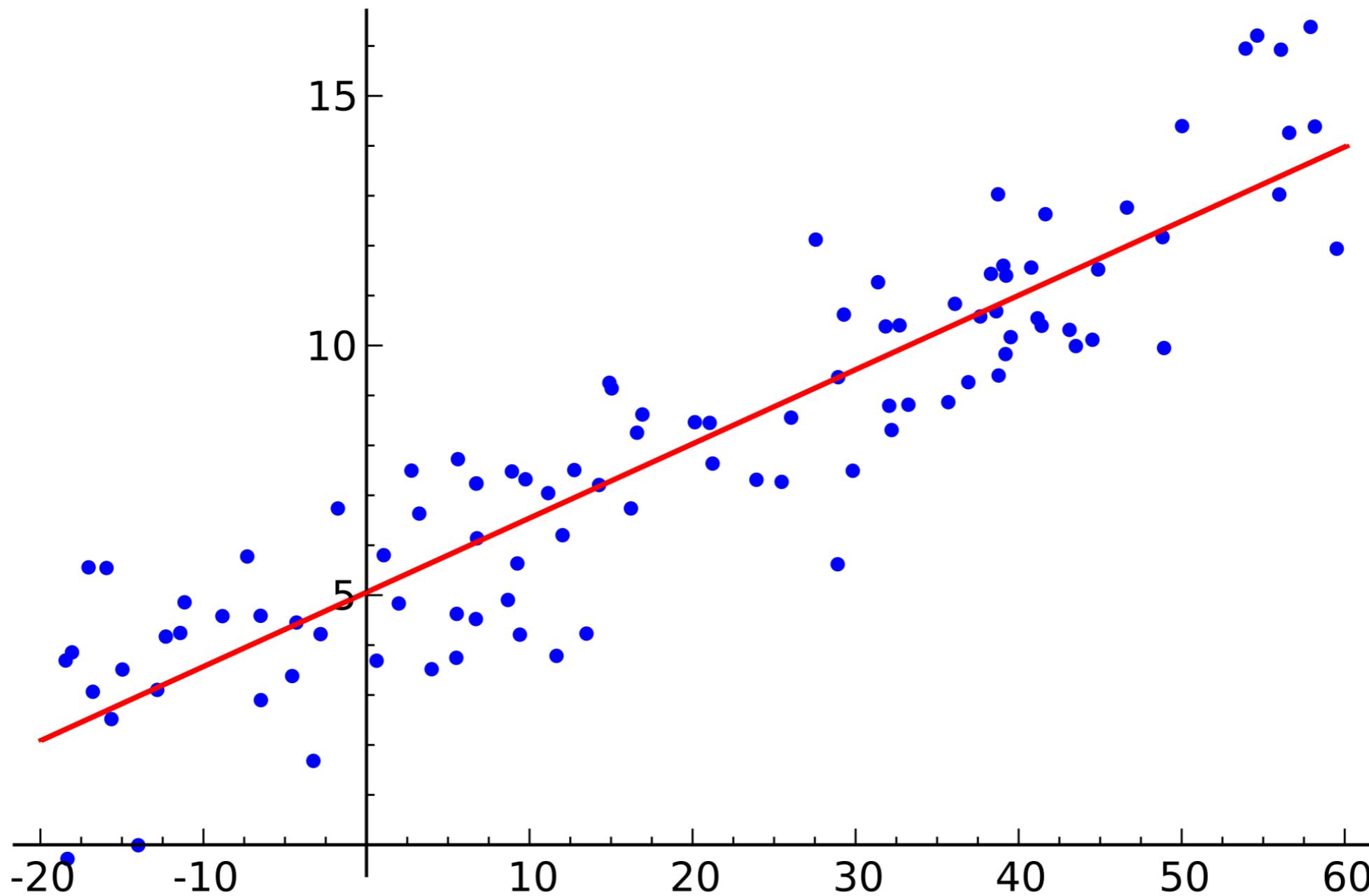
What is (not) machine learning?



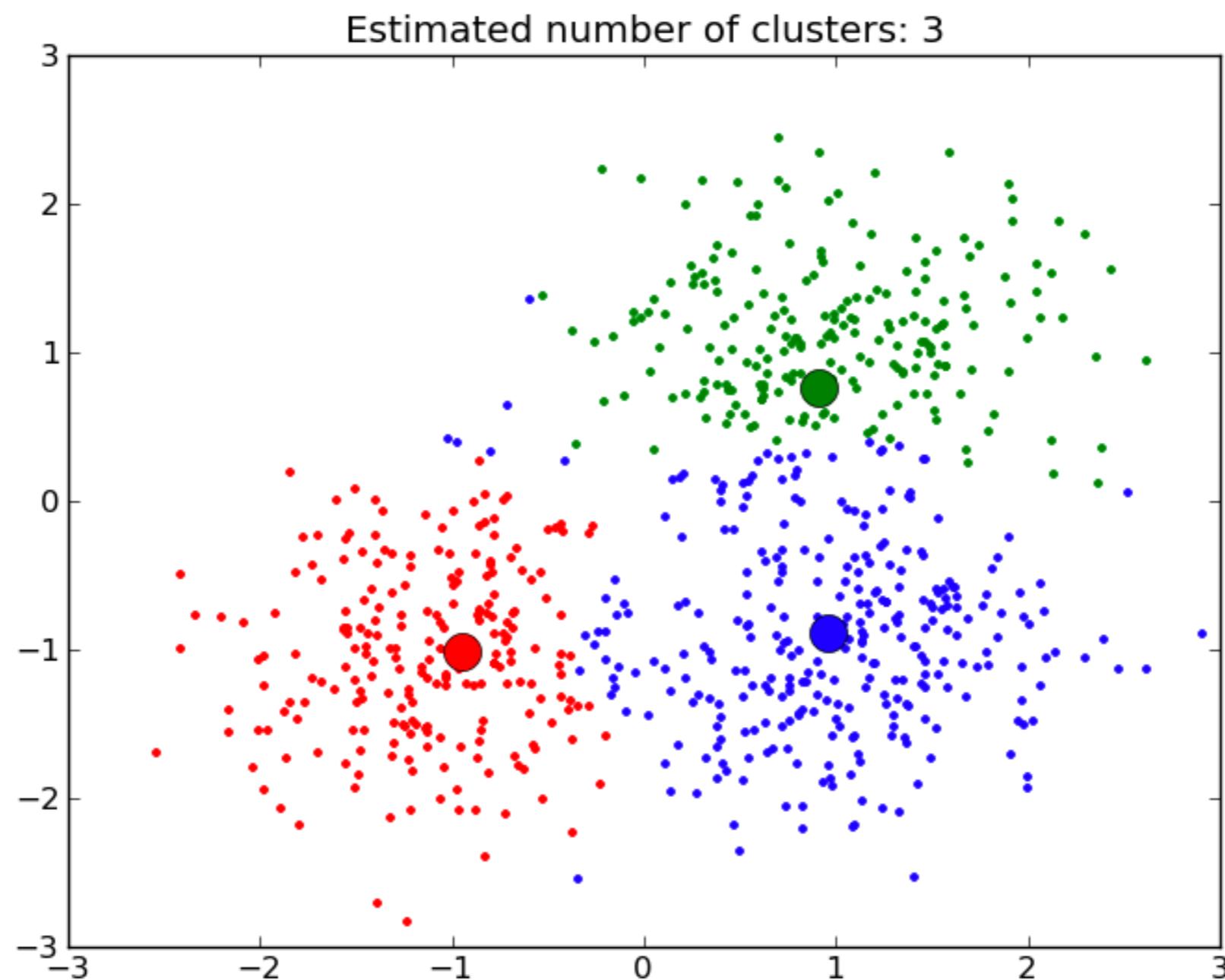
Classification



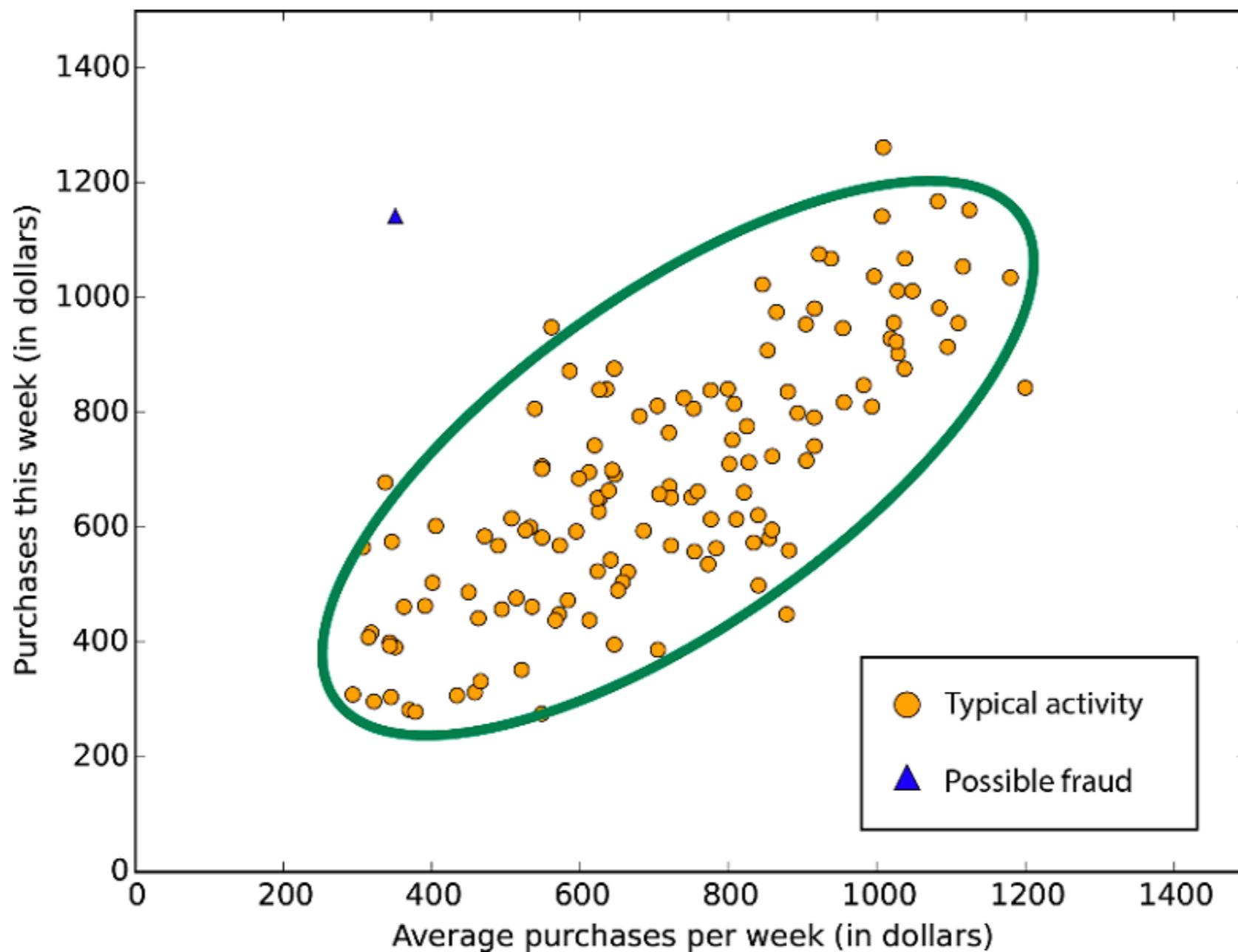
Regression



Clustering

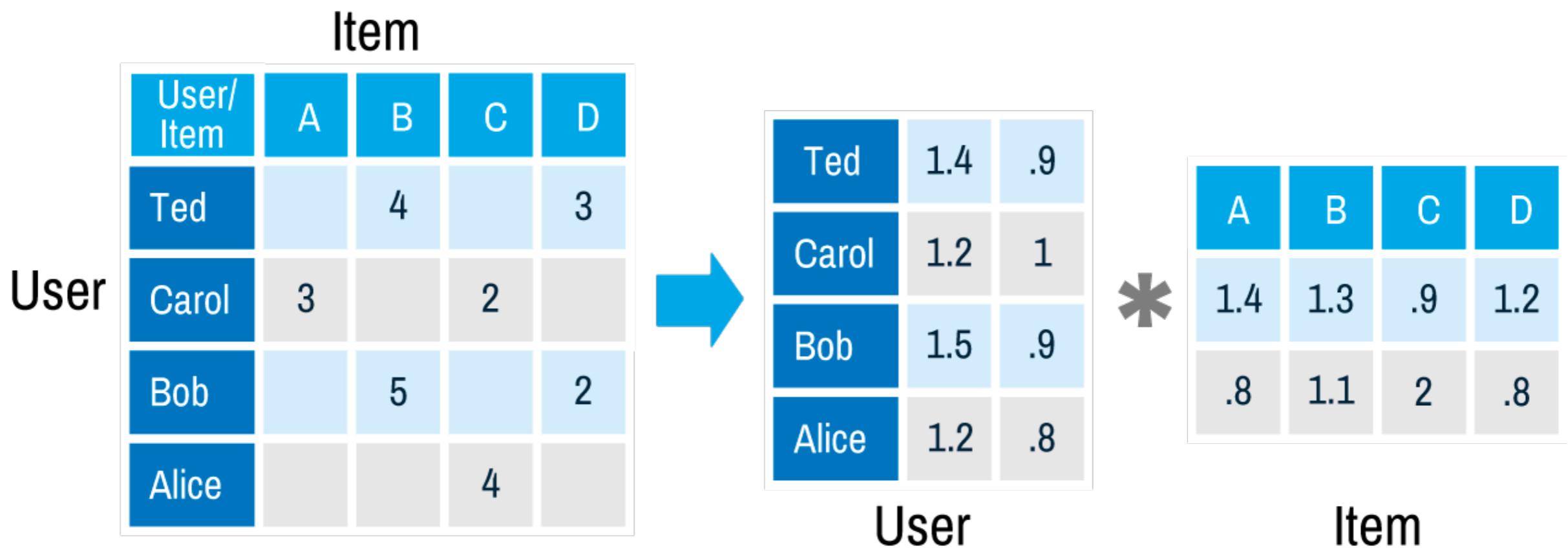


Anomaly detection



Source: <https://docs.microsoft.com/en-us/azure>

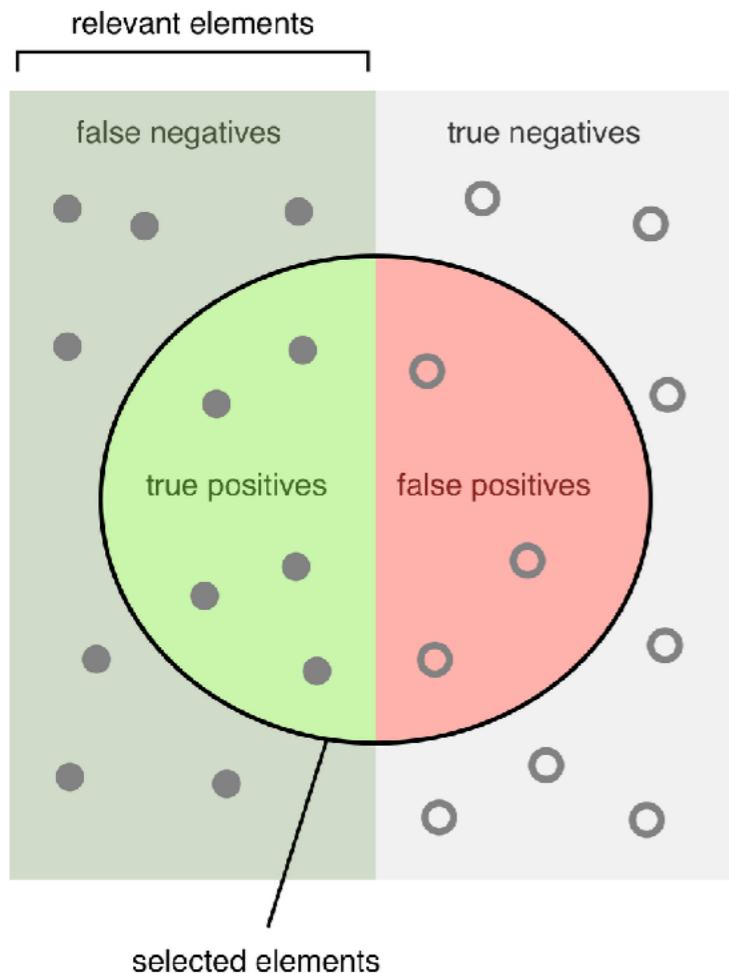
Recommendation



Data

- Train, validation, test data sets
- Cross-validation
- Imbalanced data sets
- Baseline models

Model evaluation



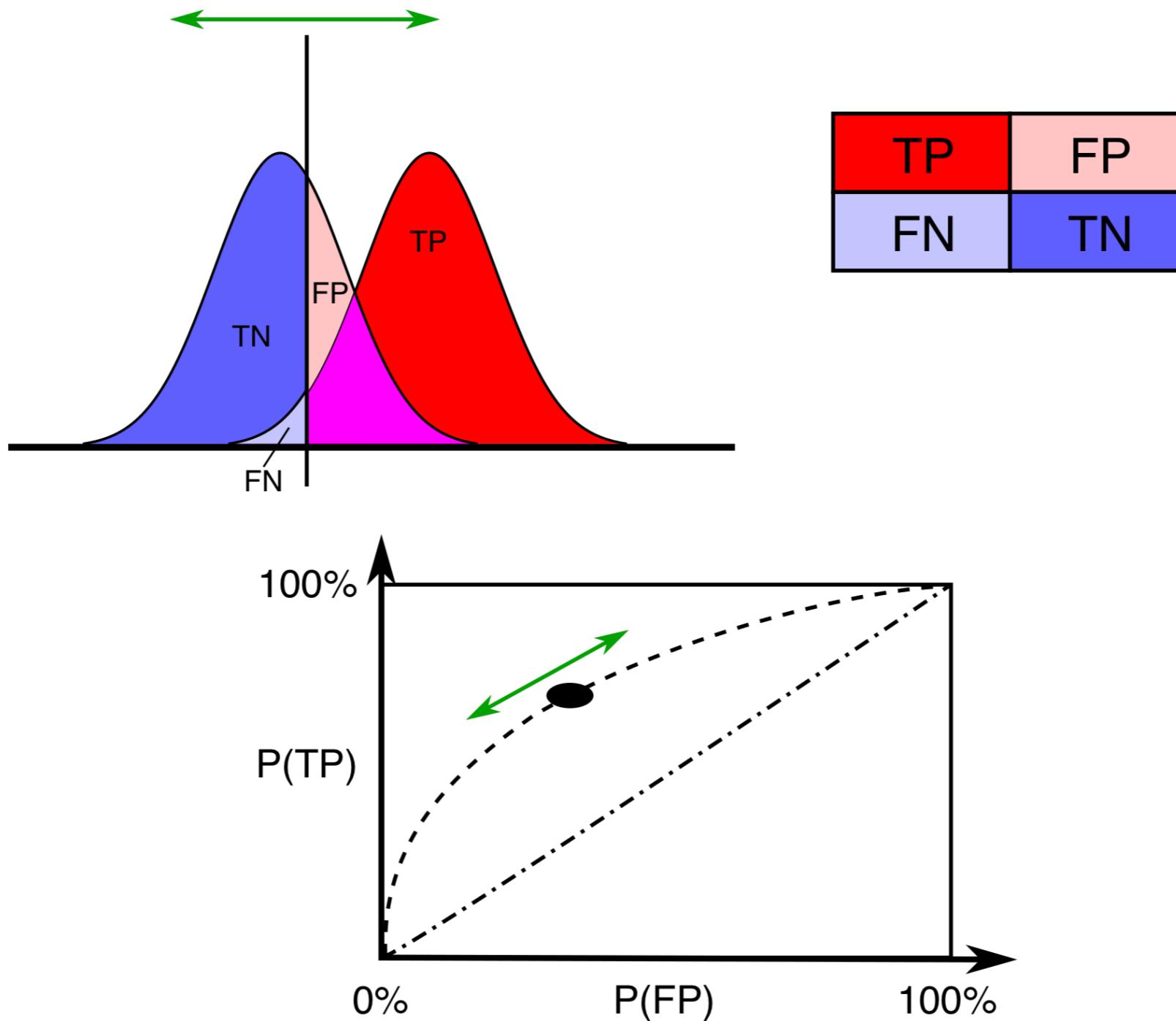
$$\text{Accuracy} = \frac{tp + tn}{tp + tn + fp + fn}$$

$$\text{Precision} = \frac{tp}{tp + fp}$$

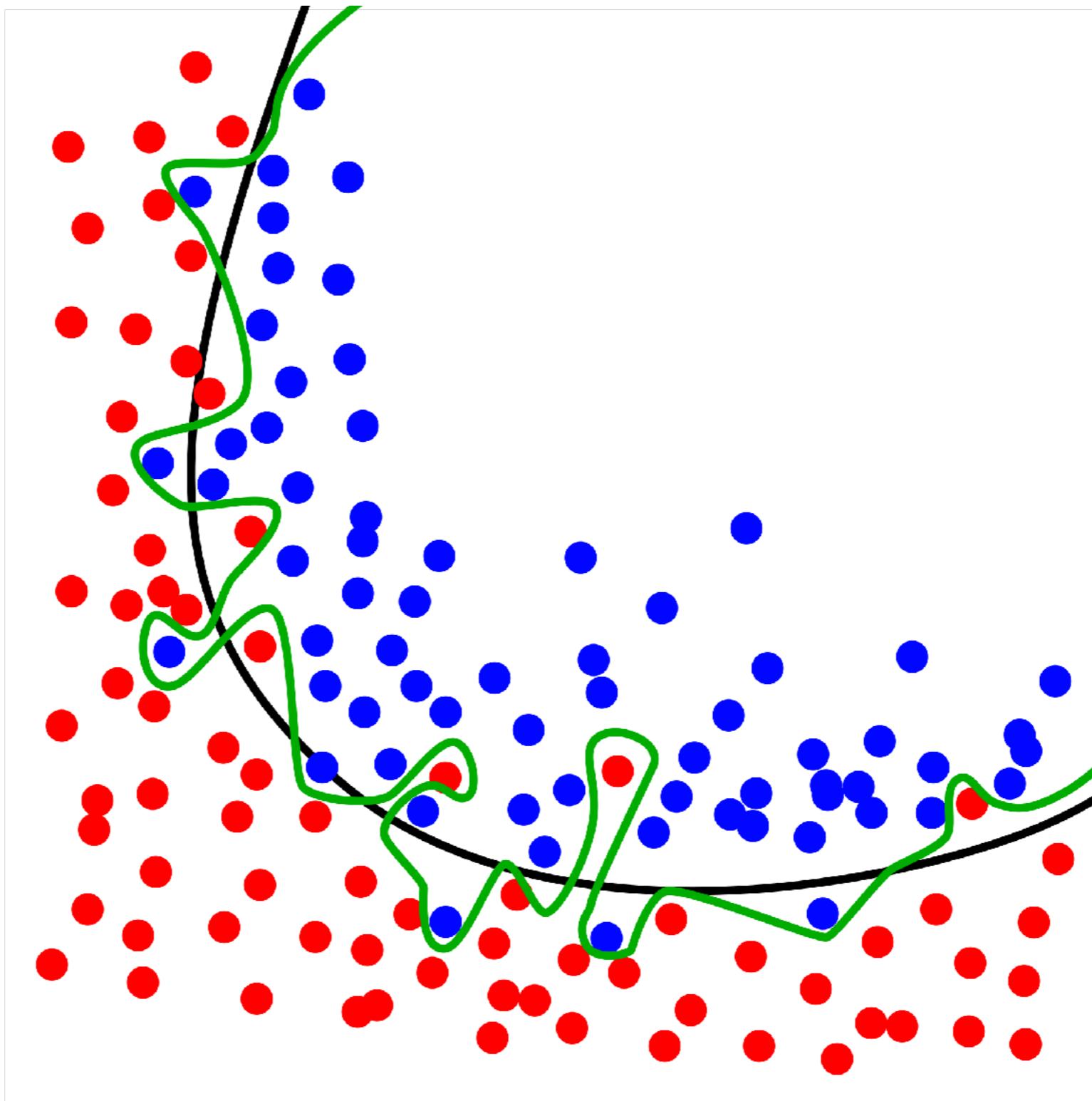
$$\text{Recall} = \frac{tp}{tp + fn}$$

$$F = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

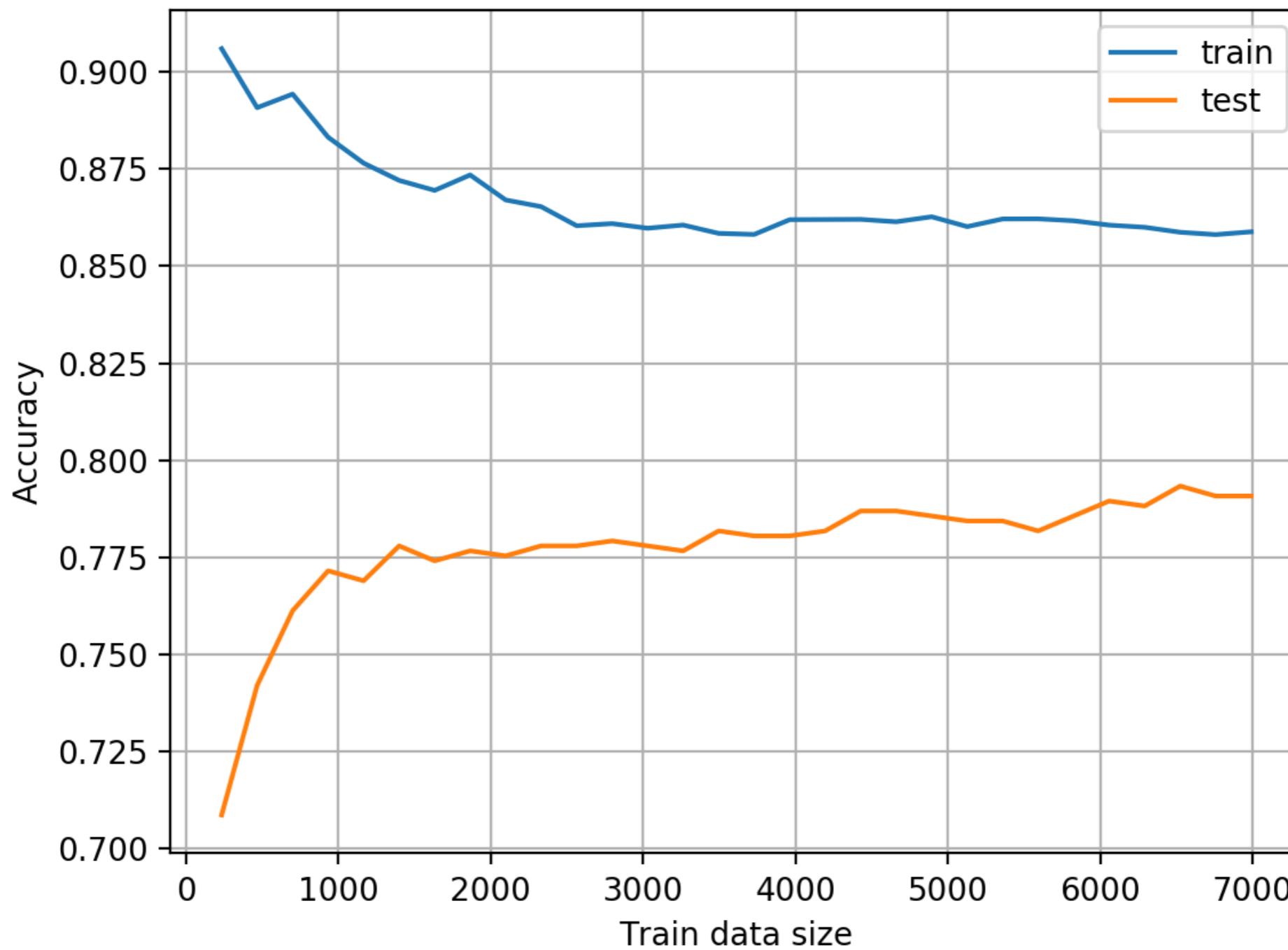
Receiver operating characteristic (ROC Curve), AUC



Overfitting



Overfitting detection



Conditional probability

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A)P(B|A)}{P(B)}$$

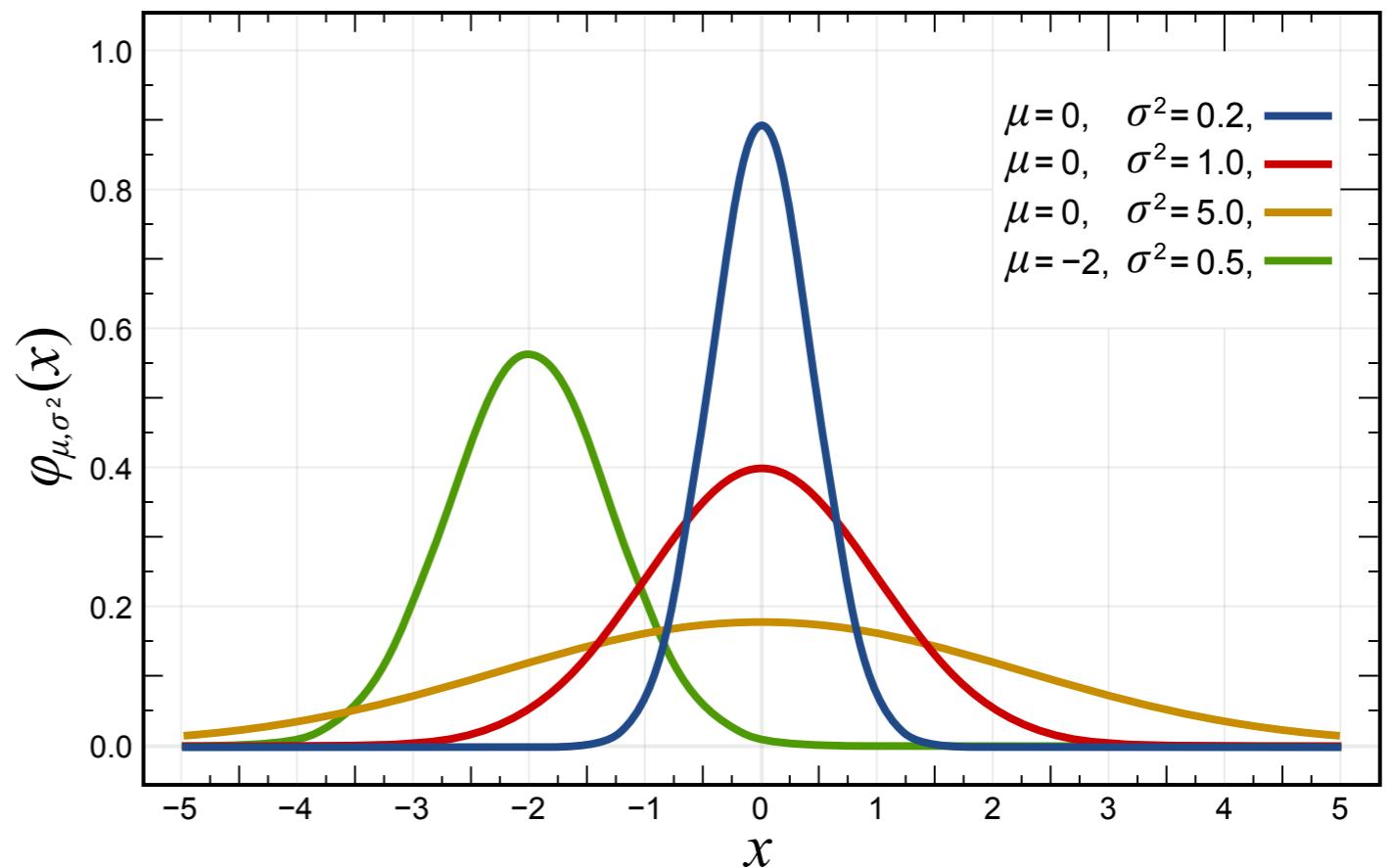
$$A \perp B \iff P(A \cap B) = P(A)P(B)$$

Naive Bayes Classifier

client	balance	income	sex	unemployed	loan
1	H	H	F	F	T
2	L	H	M	F	T
3	L	L	M	T	F
4	H	L	F	T	T
5	L	L	F	T	F
6	H	L	M	F	?

Gaussian Naive Bayes Classifier

$$p(x = v|C_k) = \frac{1}{\sqrt{2\pi\sigma_k^2}} e^{-\frac{(v - \mu_k)^2}{2\sigma_k^2}}$$



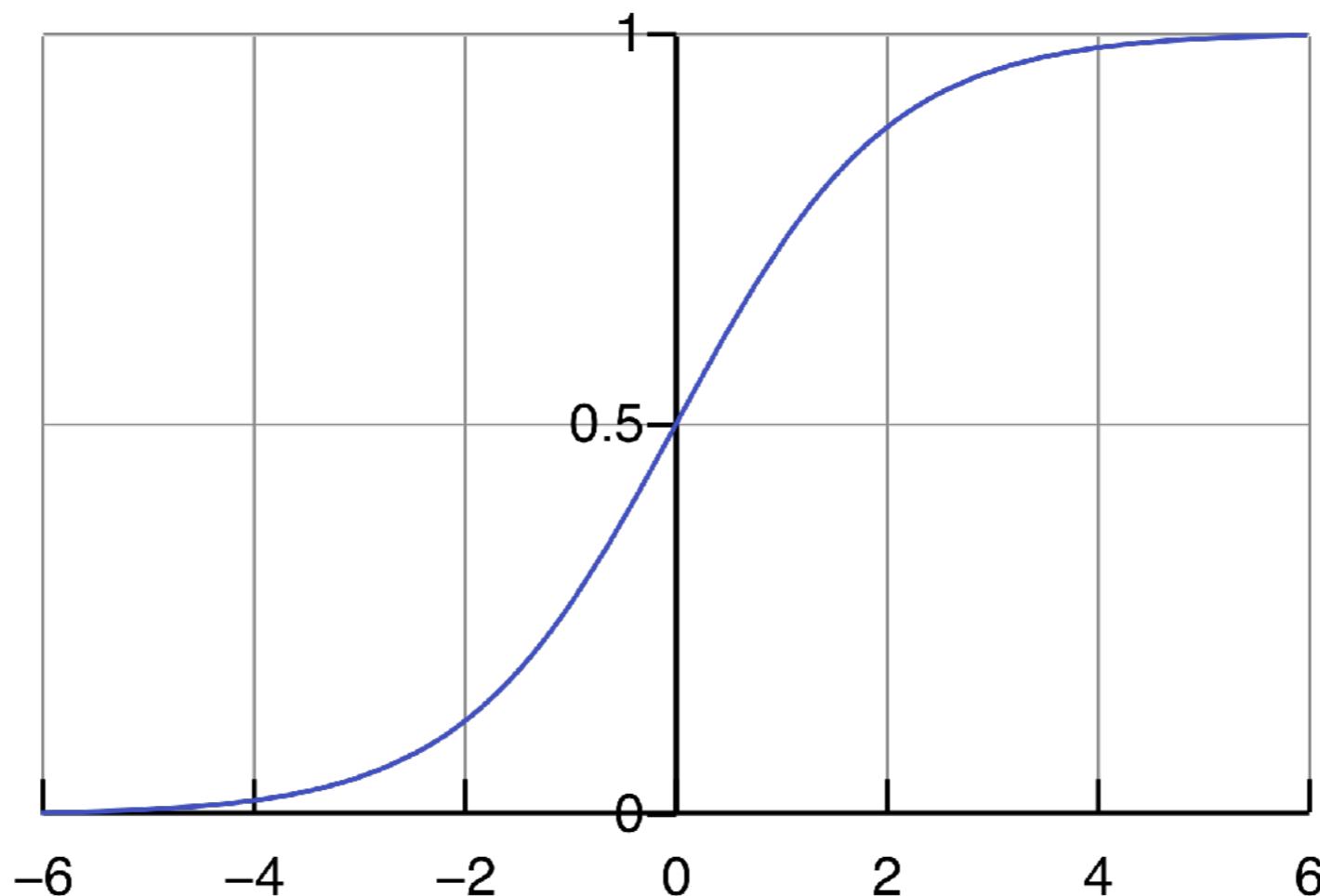
Scikit-learn tutorial

<http://scikit-learn.org/stable/>

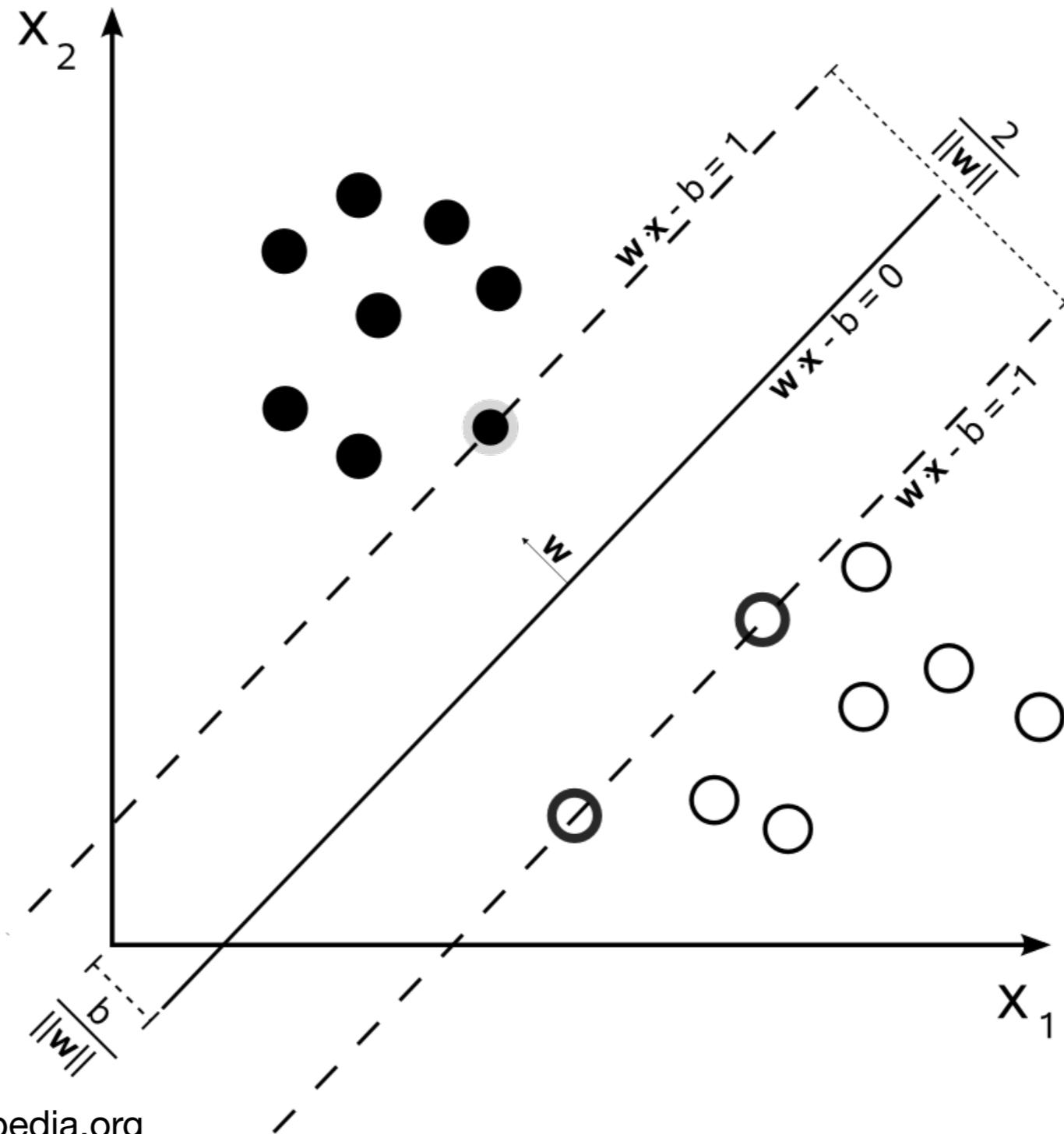
[01-Scikit-introduction.ipynb](#)

Logistic regression

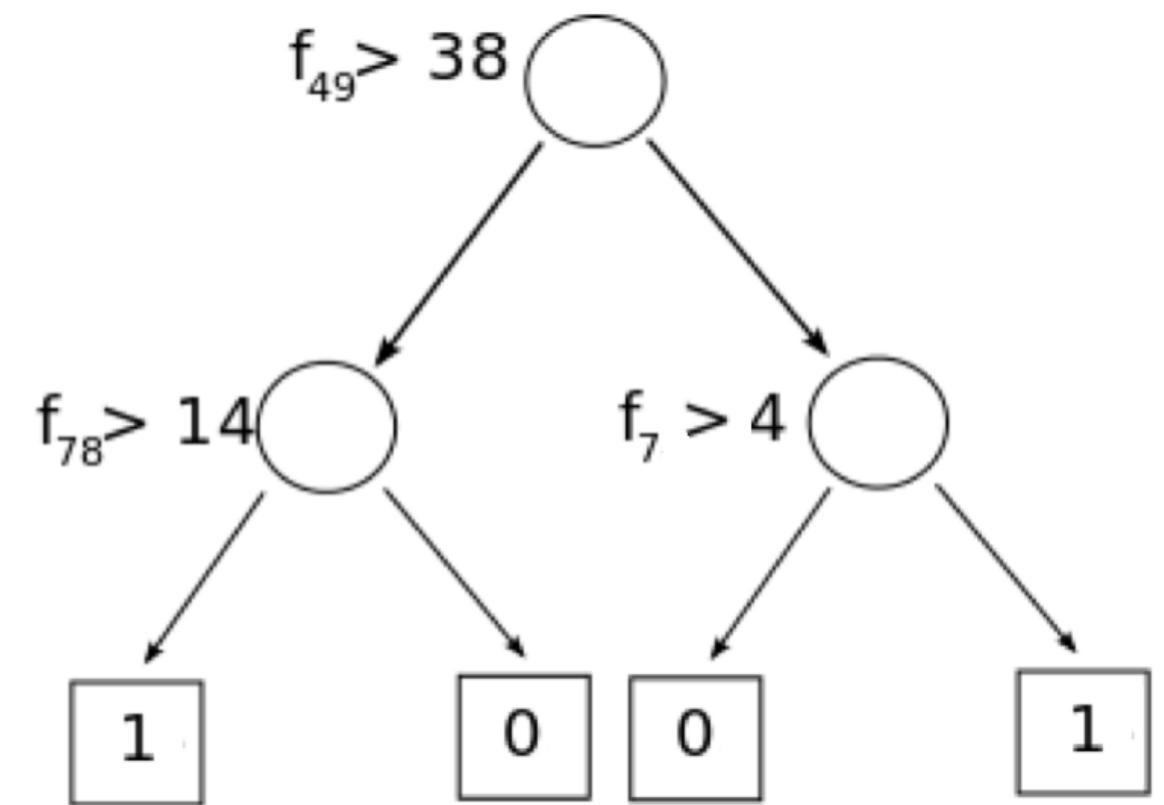
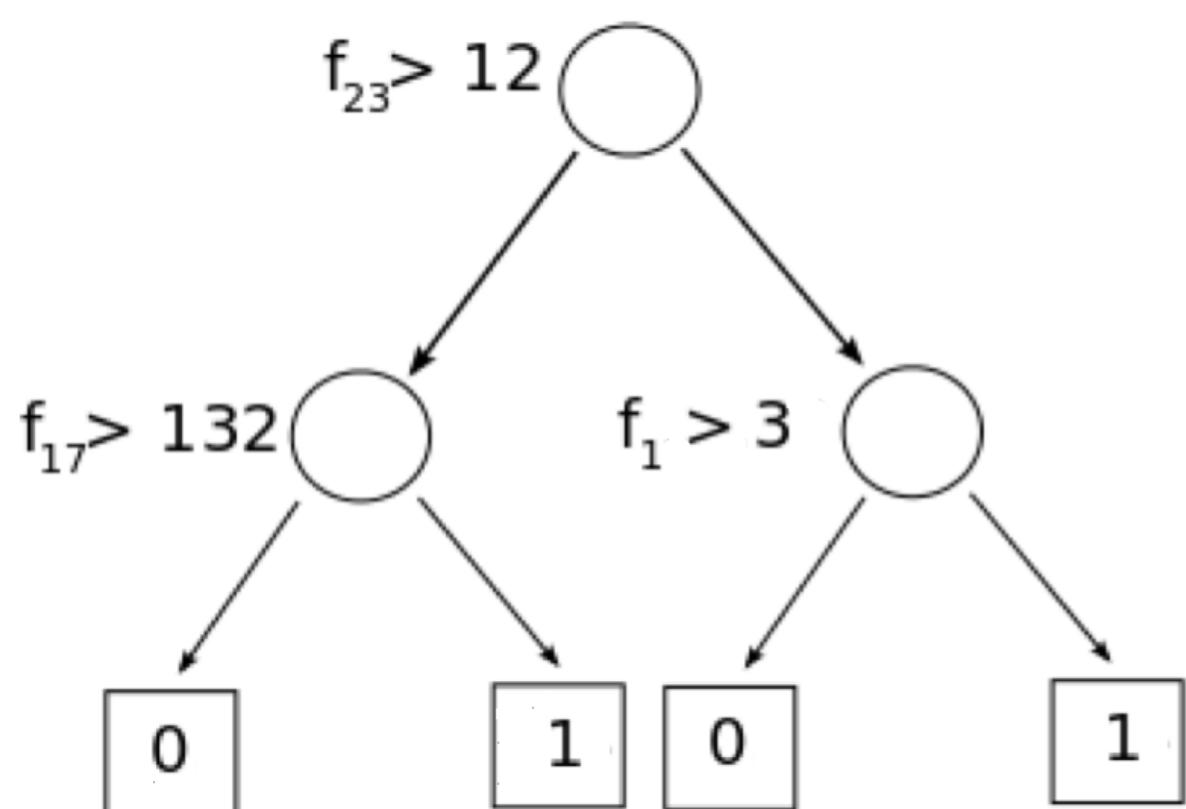
$$P(y|\vec{x}) = \frac{1}{1 + e^{-(\vec{x}\vec{w} + w_0)}}$$



Support Vector Machines



Boosted Decision Trees

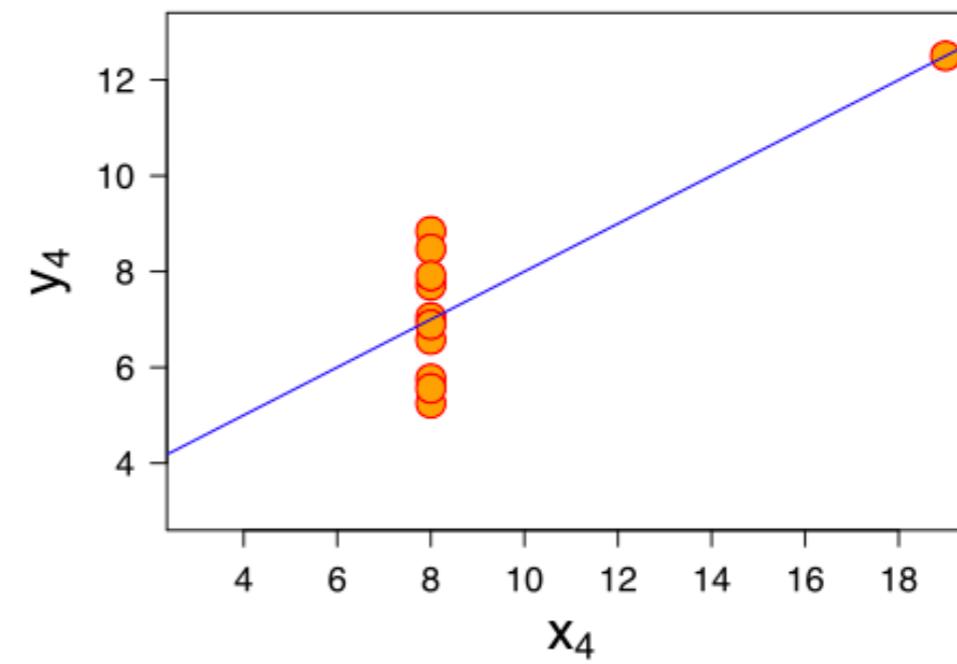
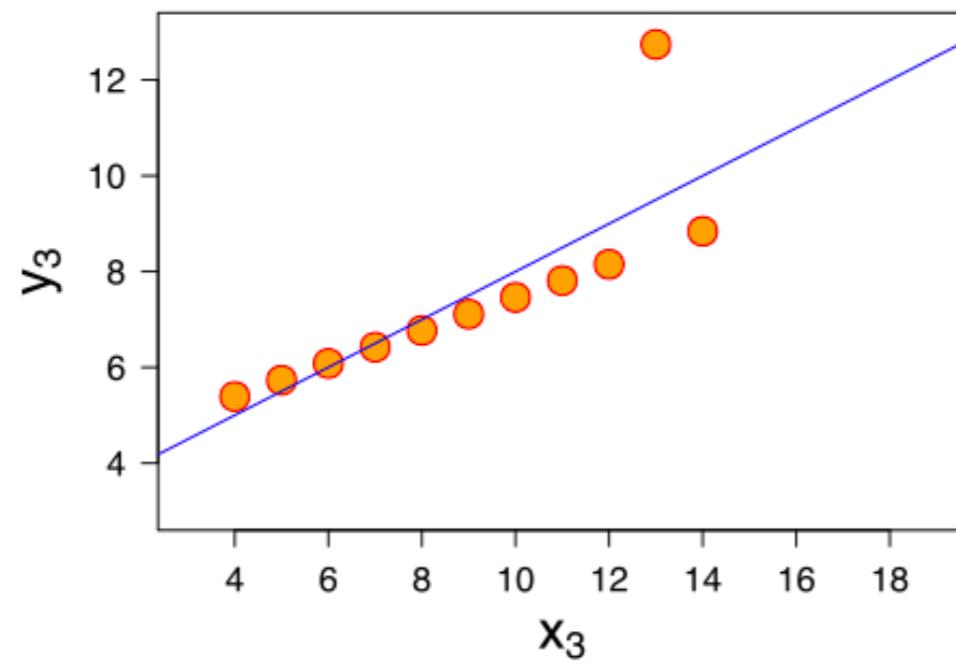
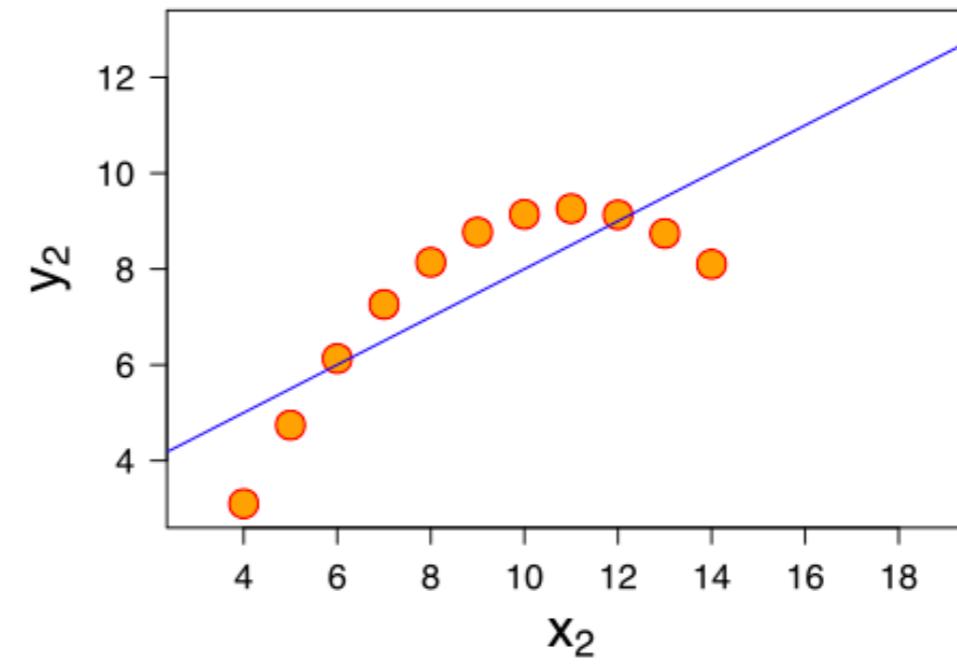
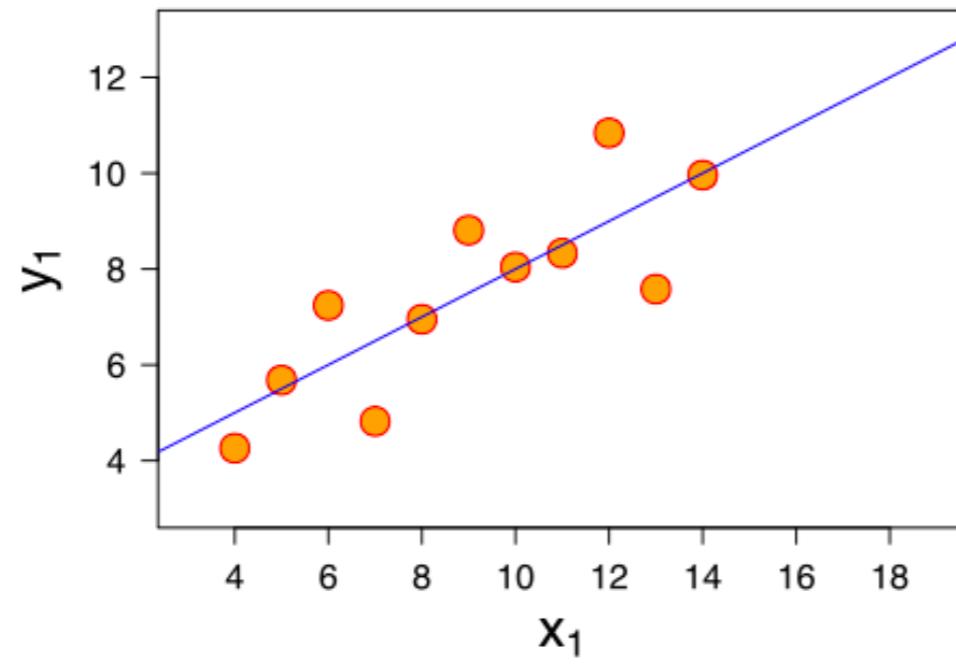


Classification task

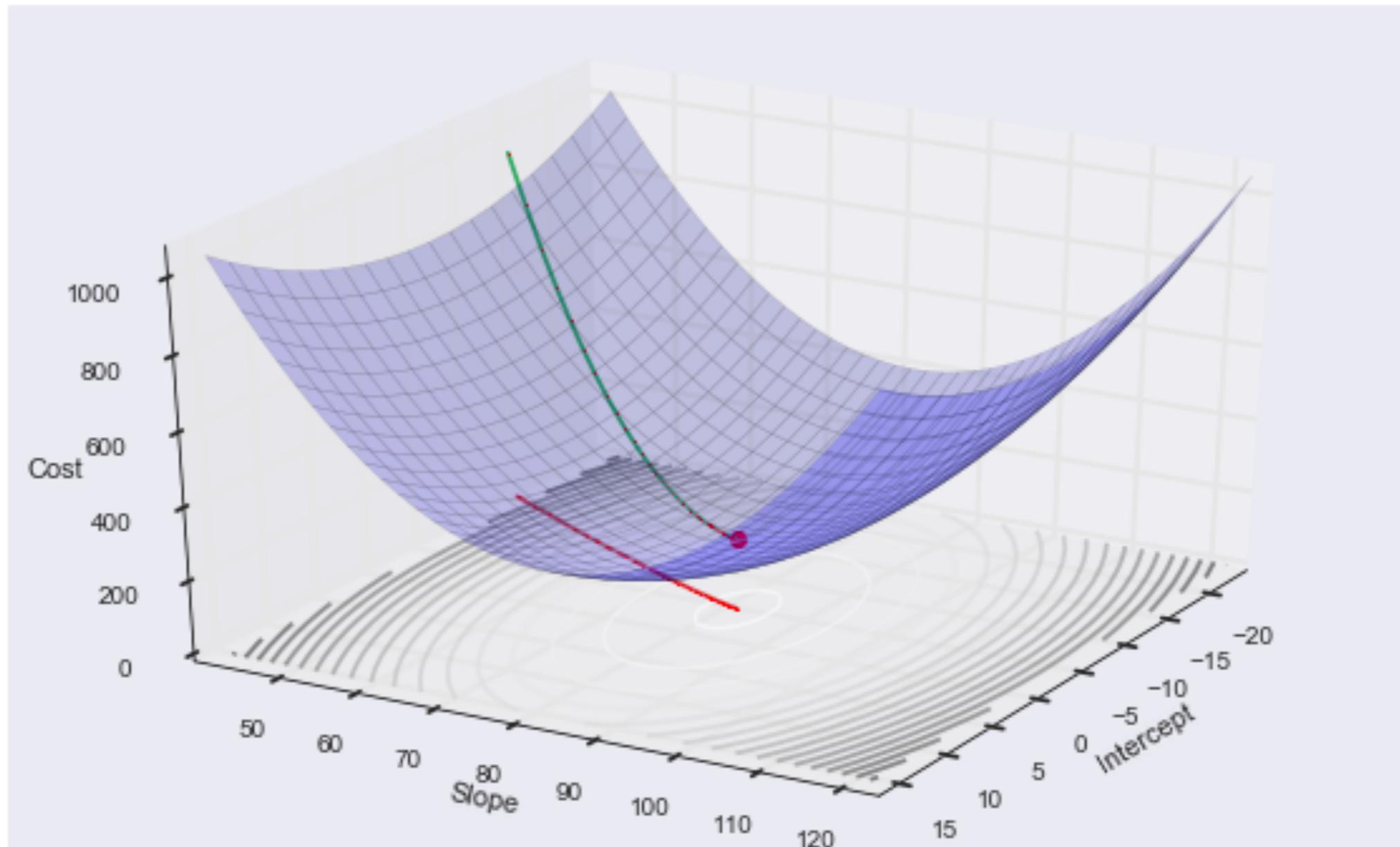
02-Classification1-assignment.ipynb

03-Classification2-assignment.ipynb

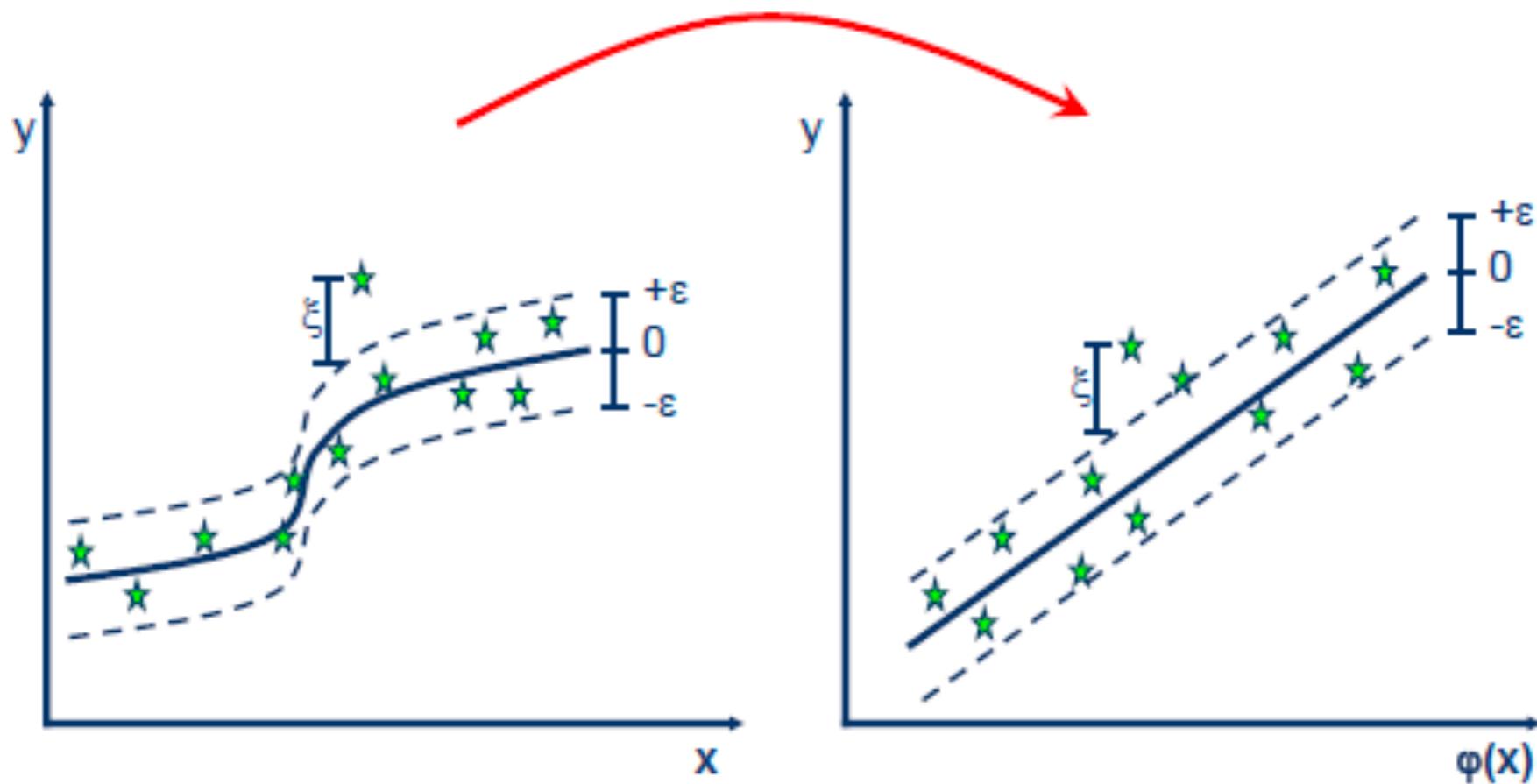
Regression



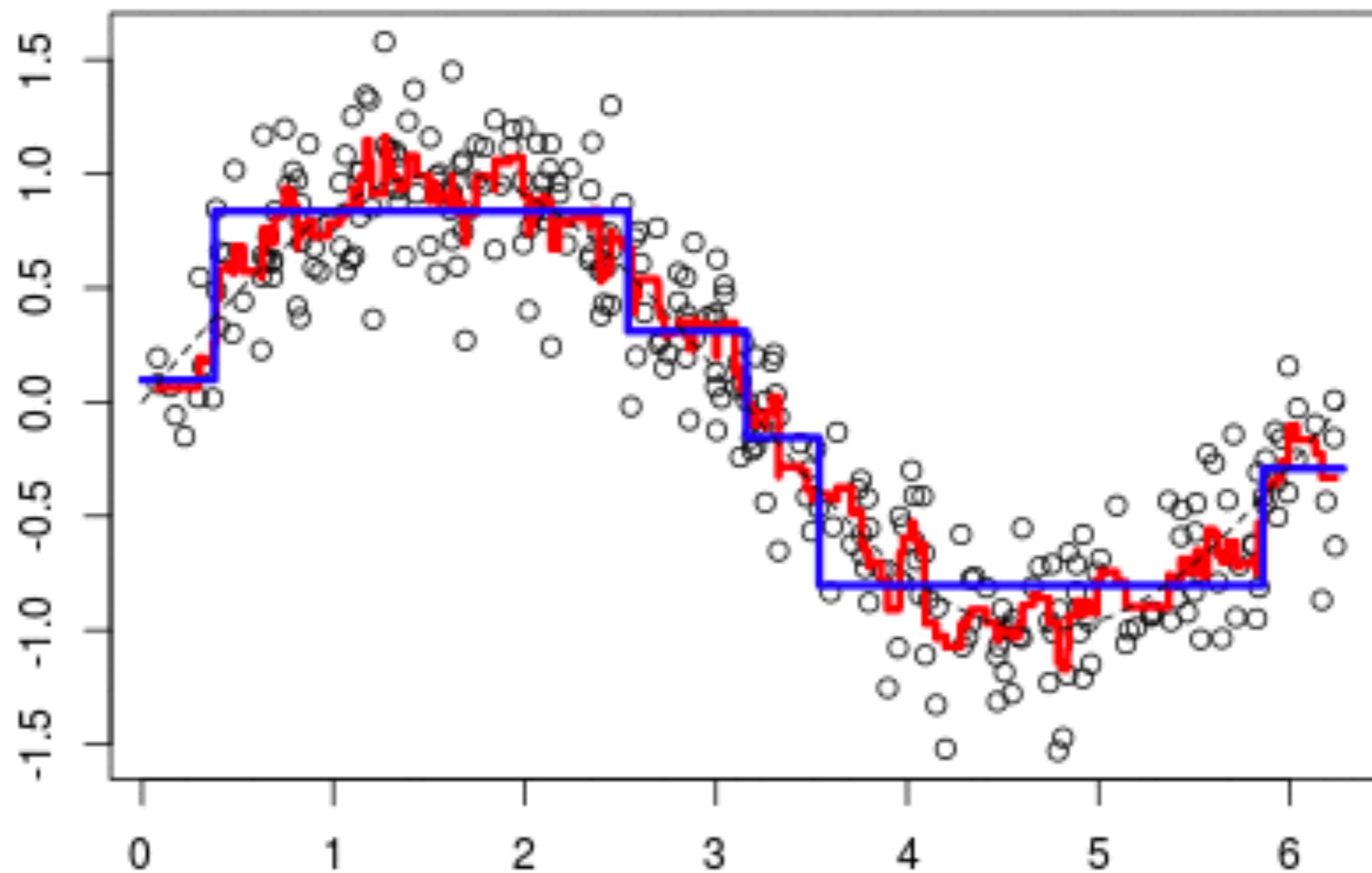
Linear regression with SGD



Support vector machines for Regression



Boosted regression trees



Evaluation of regression models

Root mean squared error

$$\text{RMSE} = \sqrt{\frac{\sum_i (y_i - \hat{y}_i)^2}{n}}$$

Mean absolute error

$$\text{MAE} = \frac{\sum_i |y_i - \hat{y}_i|}{n}$$

R Squared

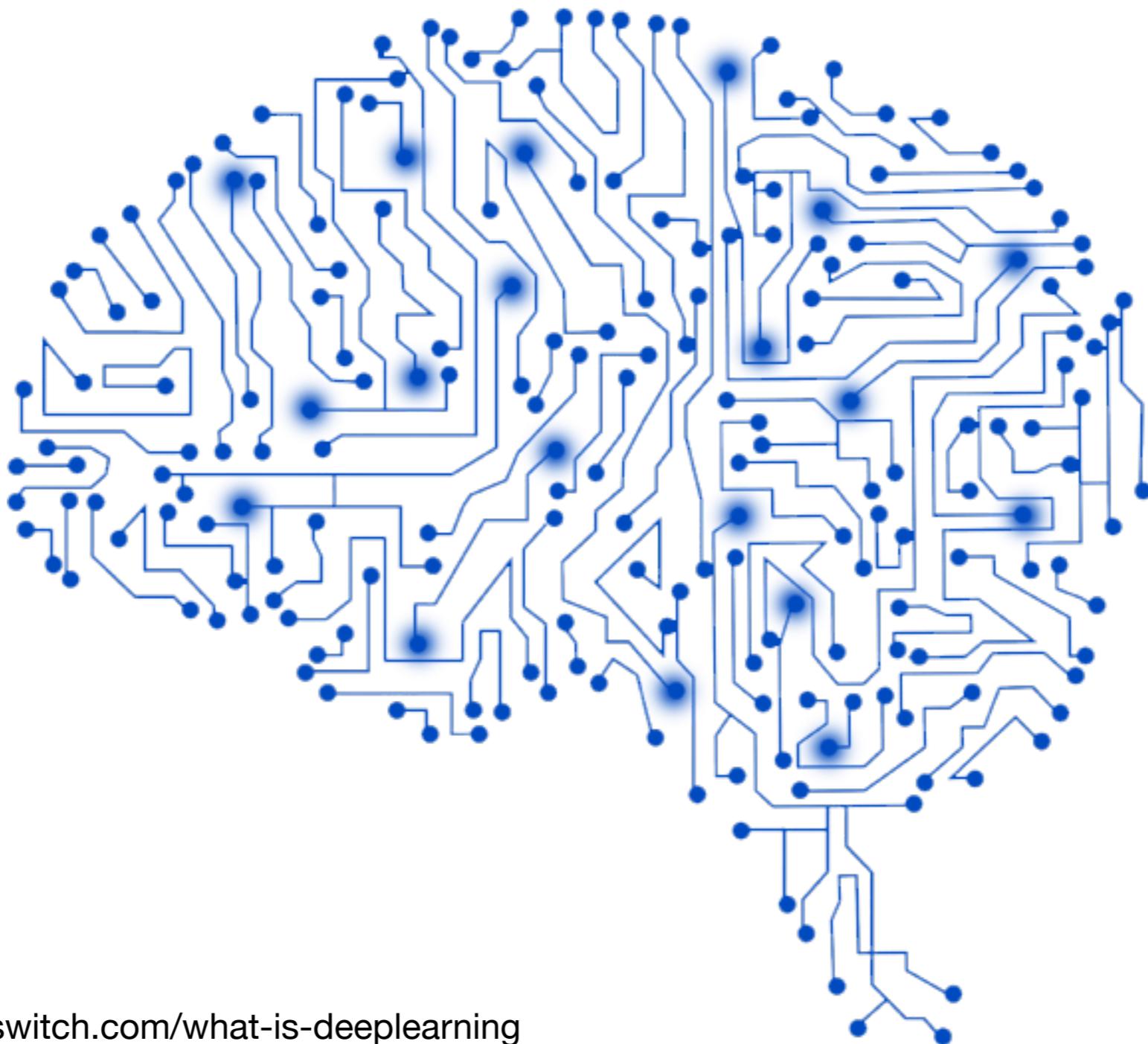
$$\hat{R}^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

Regression task

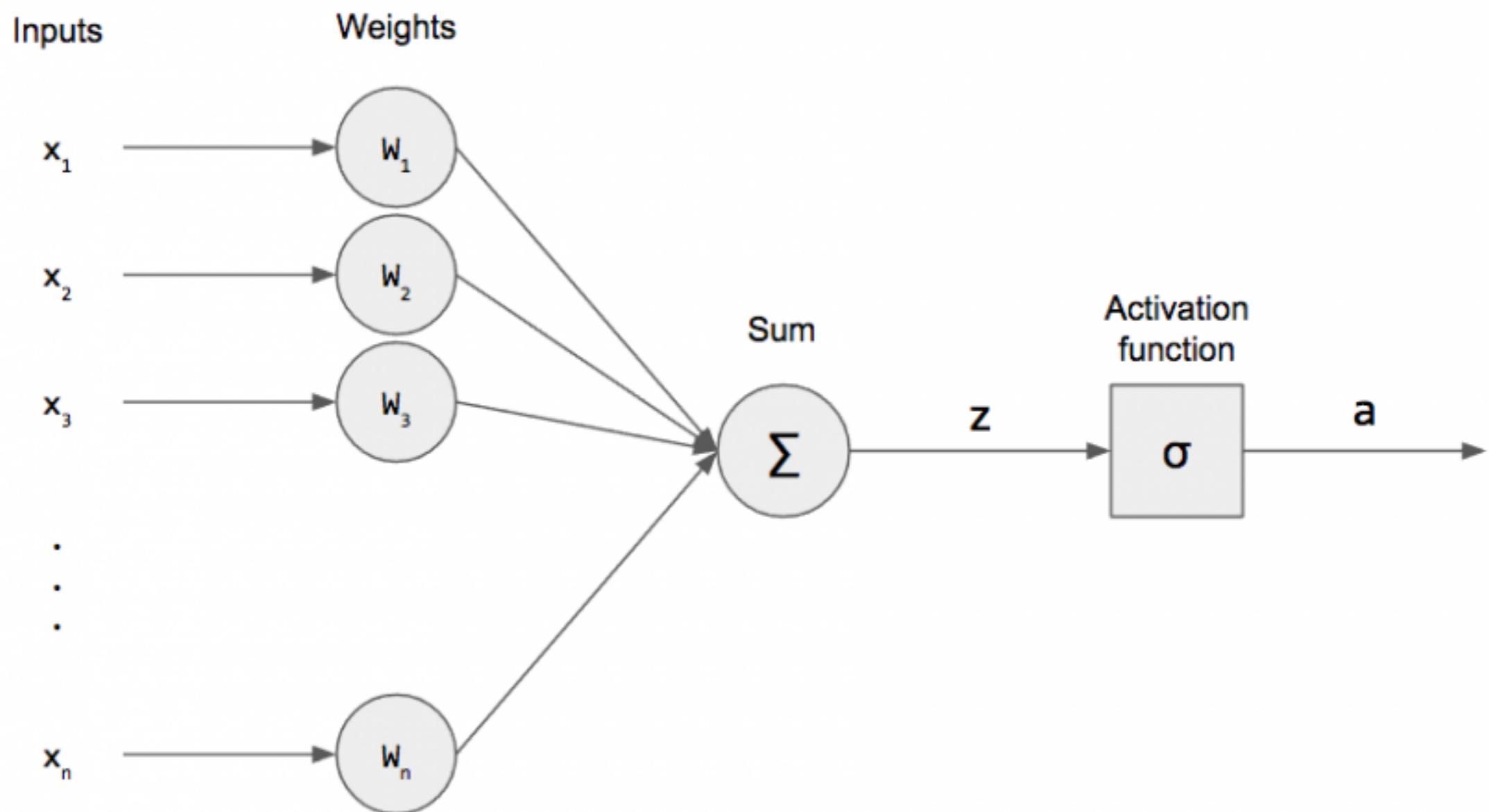
[04-Regression1-assignment.ipynb](#)

[05-Regression2-assignment.ipynb](#)

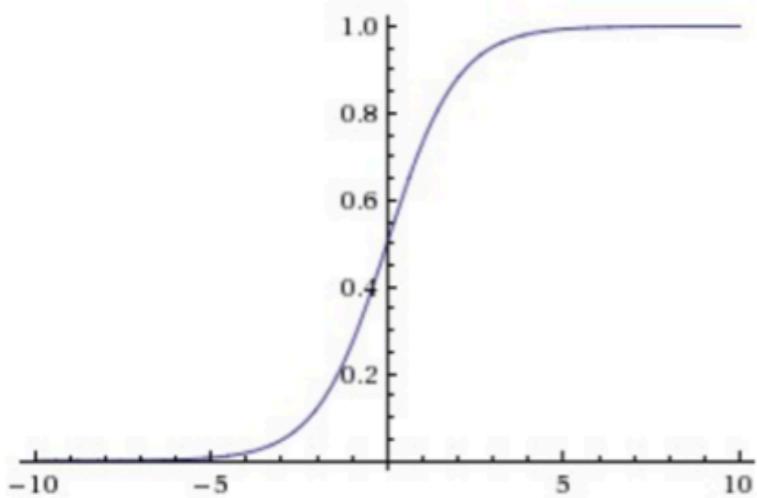
Neural networks and deep learning



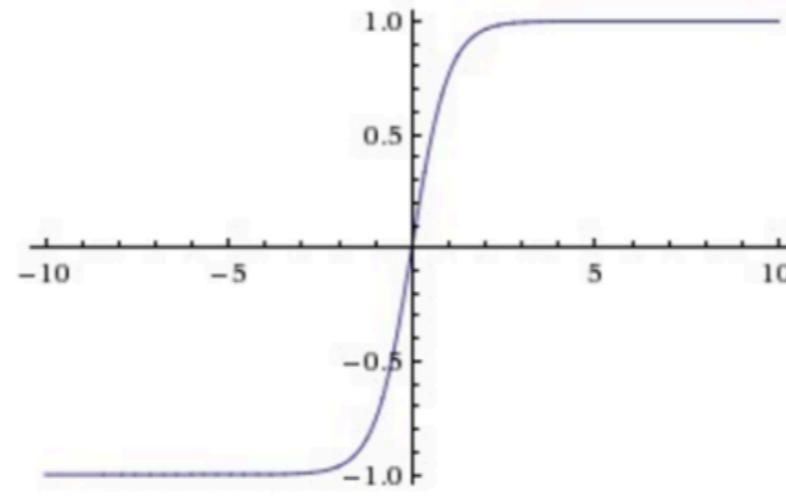
Perceptron



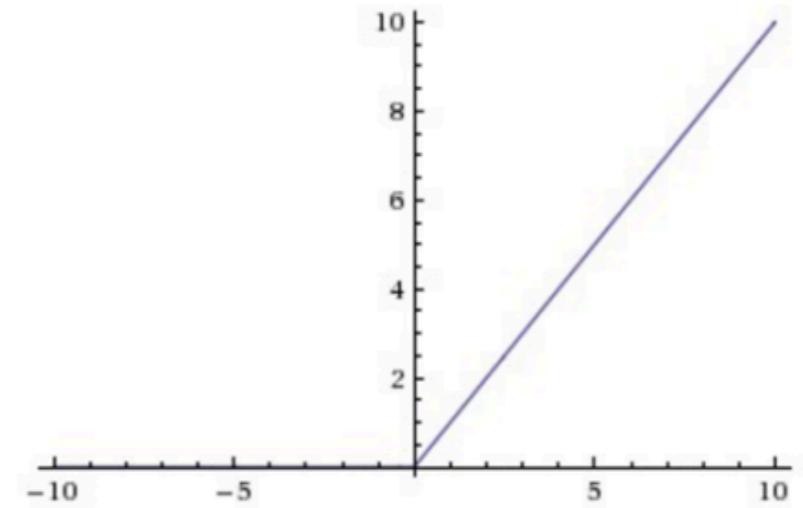
Activation functions



Sigmoid



tanh

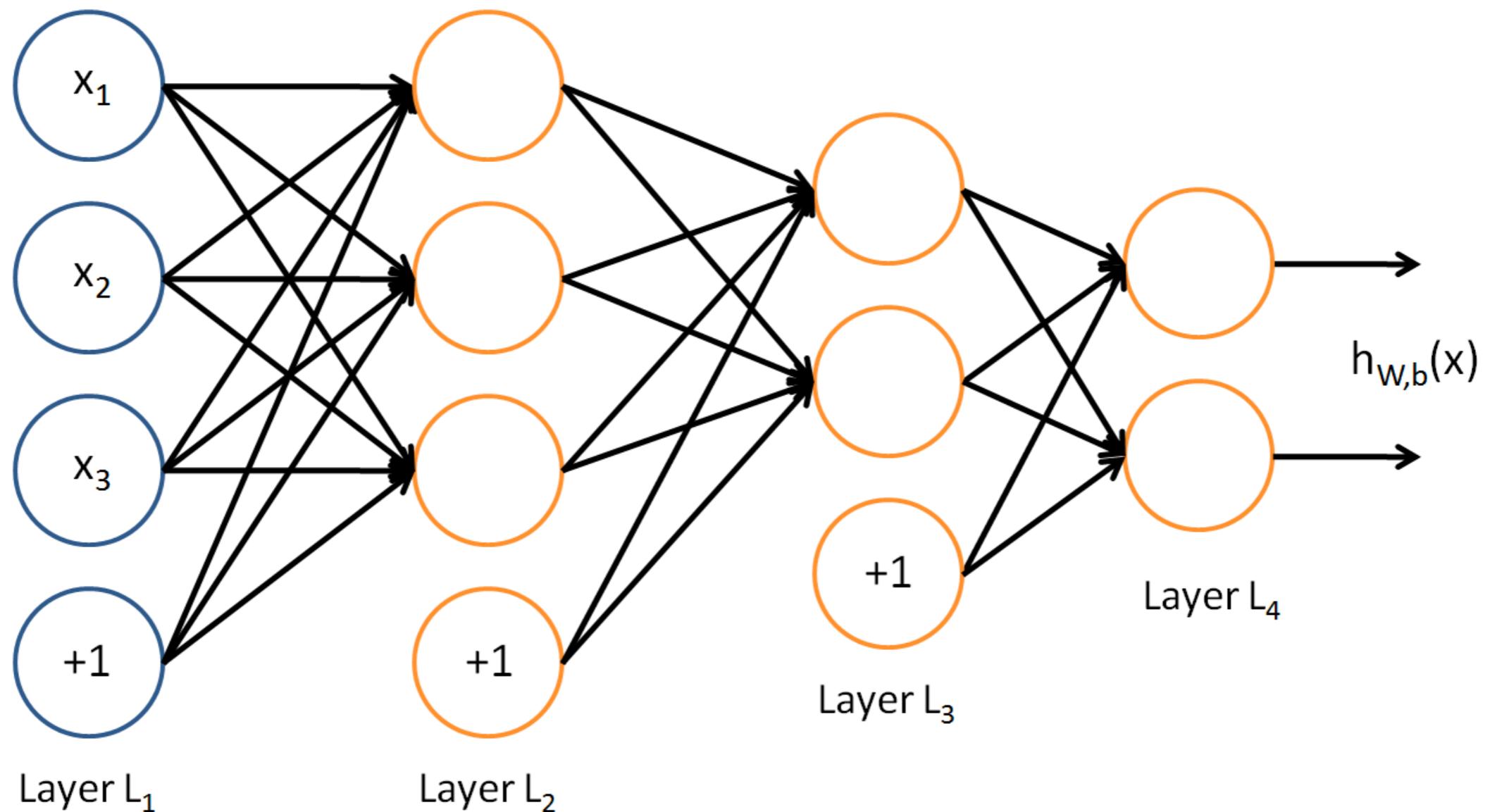


ReLU

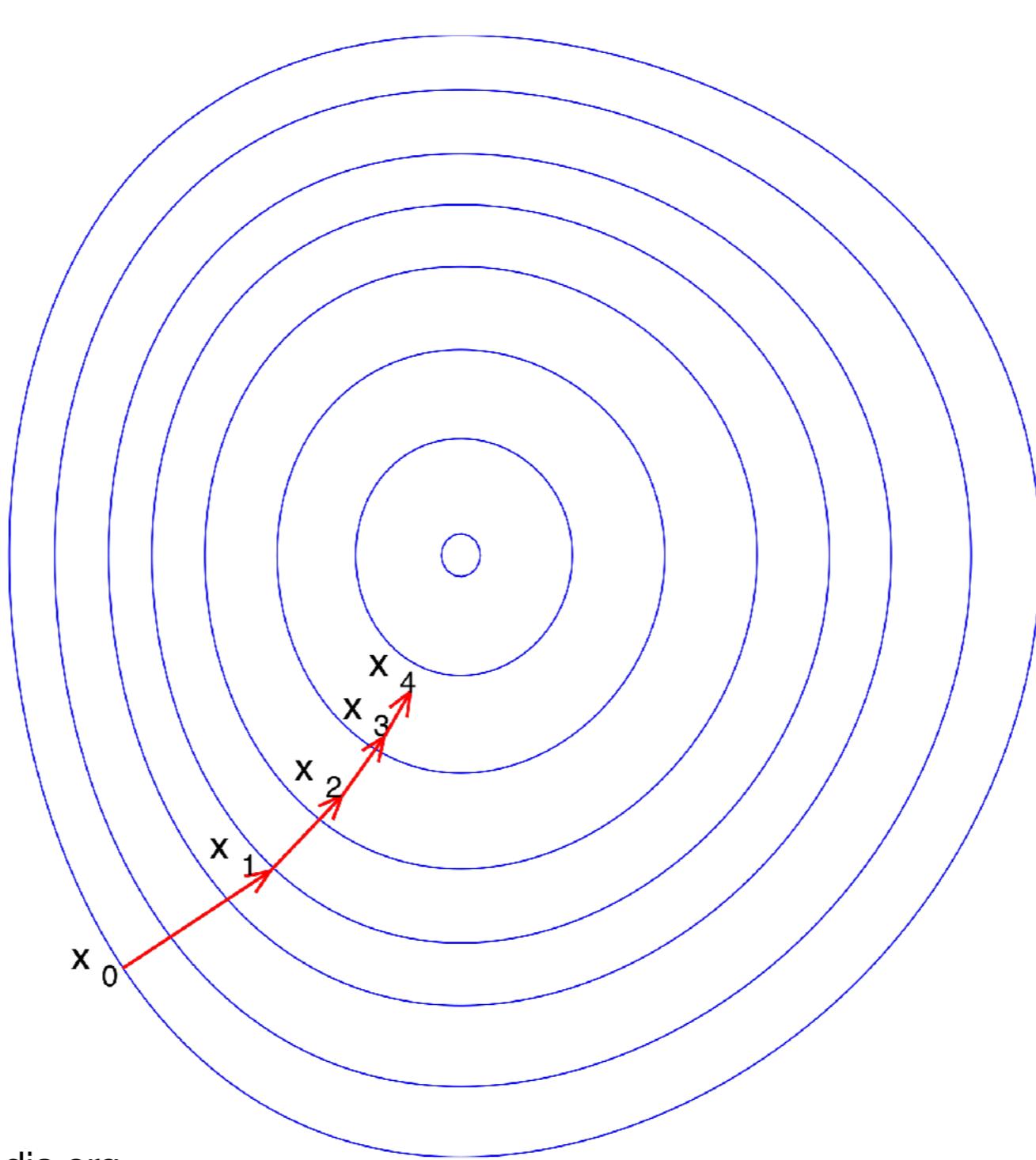
Softmax:

$$\sigma(\mathbf{z})_j = \frac{e^{z_j}}{\sum_{k=1}^K e^{z_k}}$$

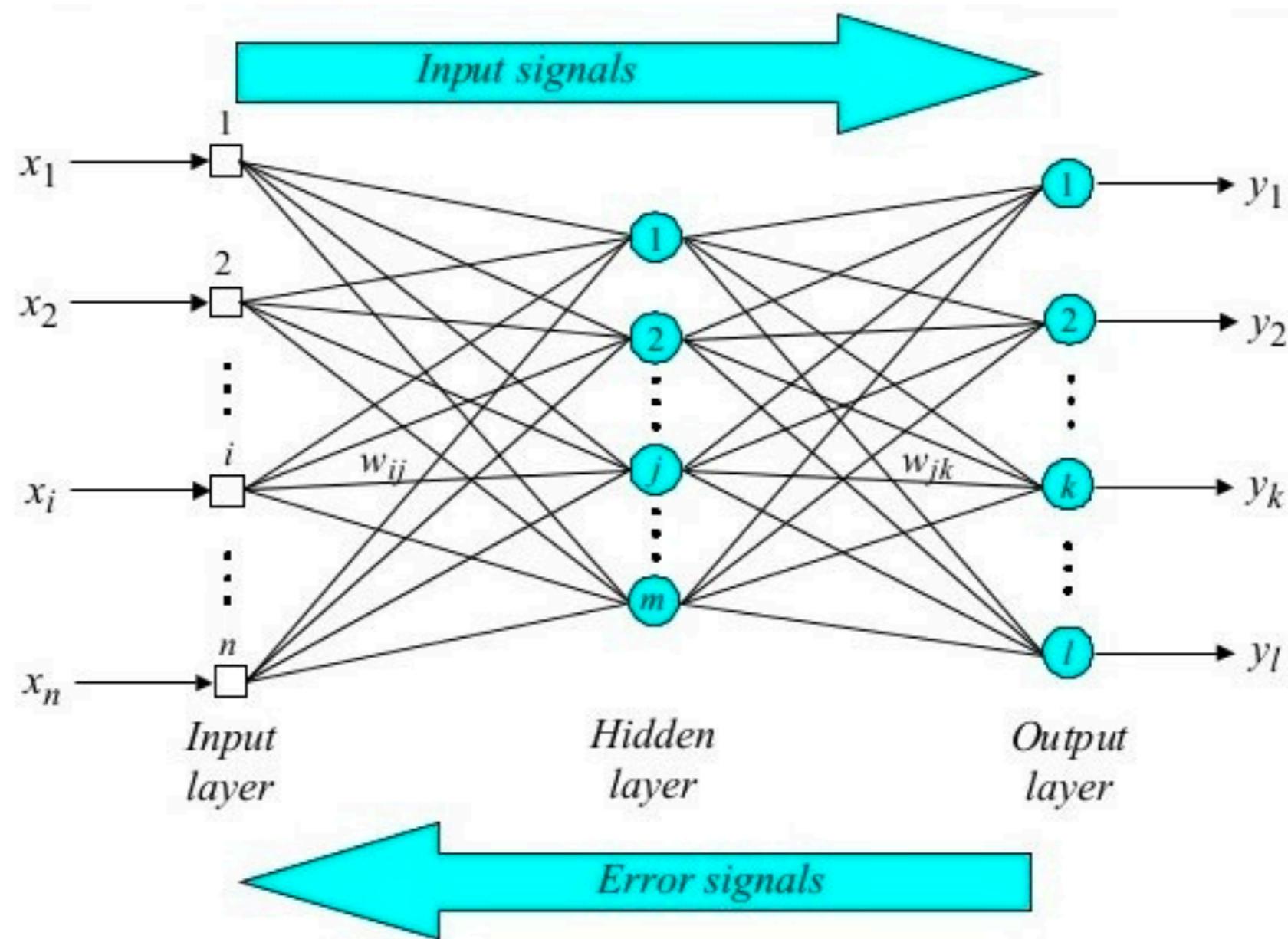
Multilayer Neural Networks



Steepest gradient descent



Back propagation



Gradient Descent Variants

(Batch) Gradient Descent

$$w_{t+1} = w_t - \lambda \frac{\partial e(X, y)}{\partial w_t}$$

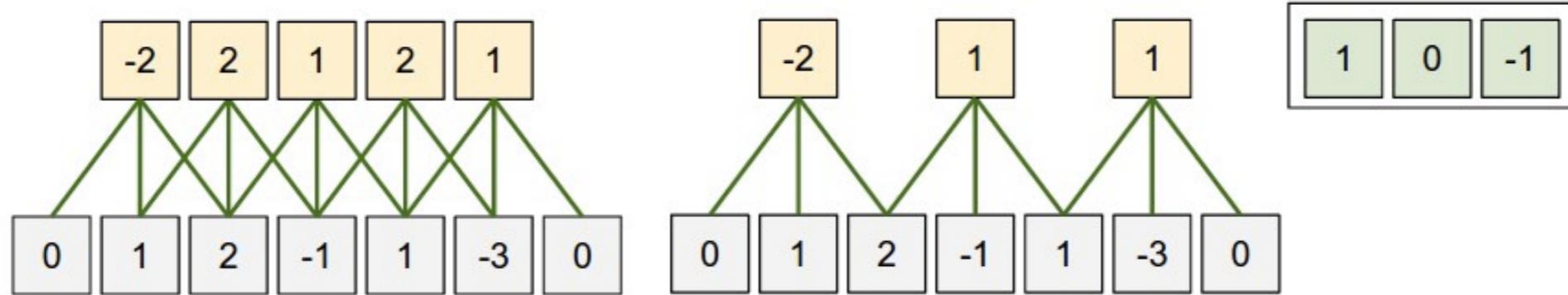
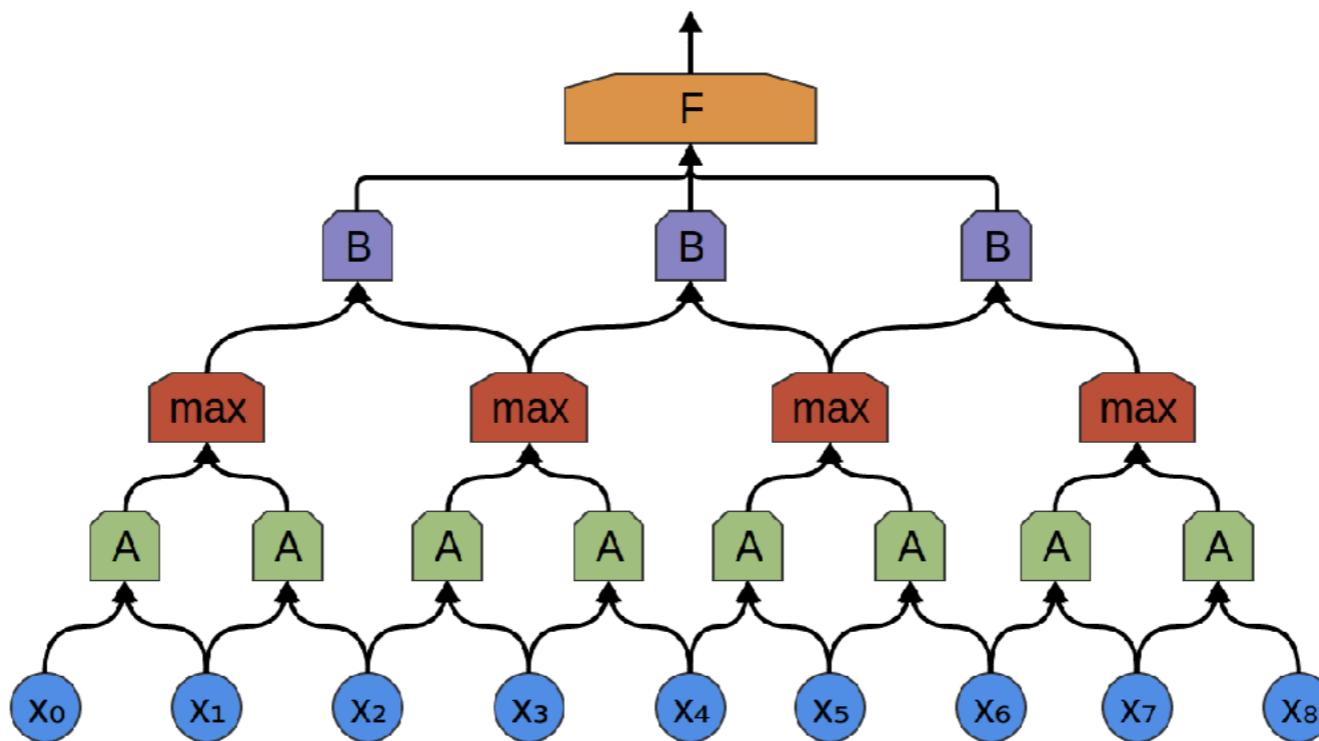
Stochastic Gradient Descent

$$w_{t+1} = w_t - \lambda \frac{\partial e(X^i, y^i)}{\partial w_t}$$

Mini-Batch Gradient Descent

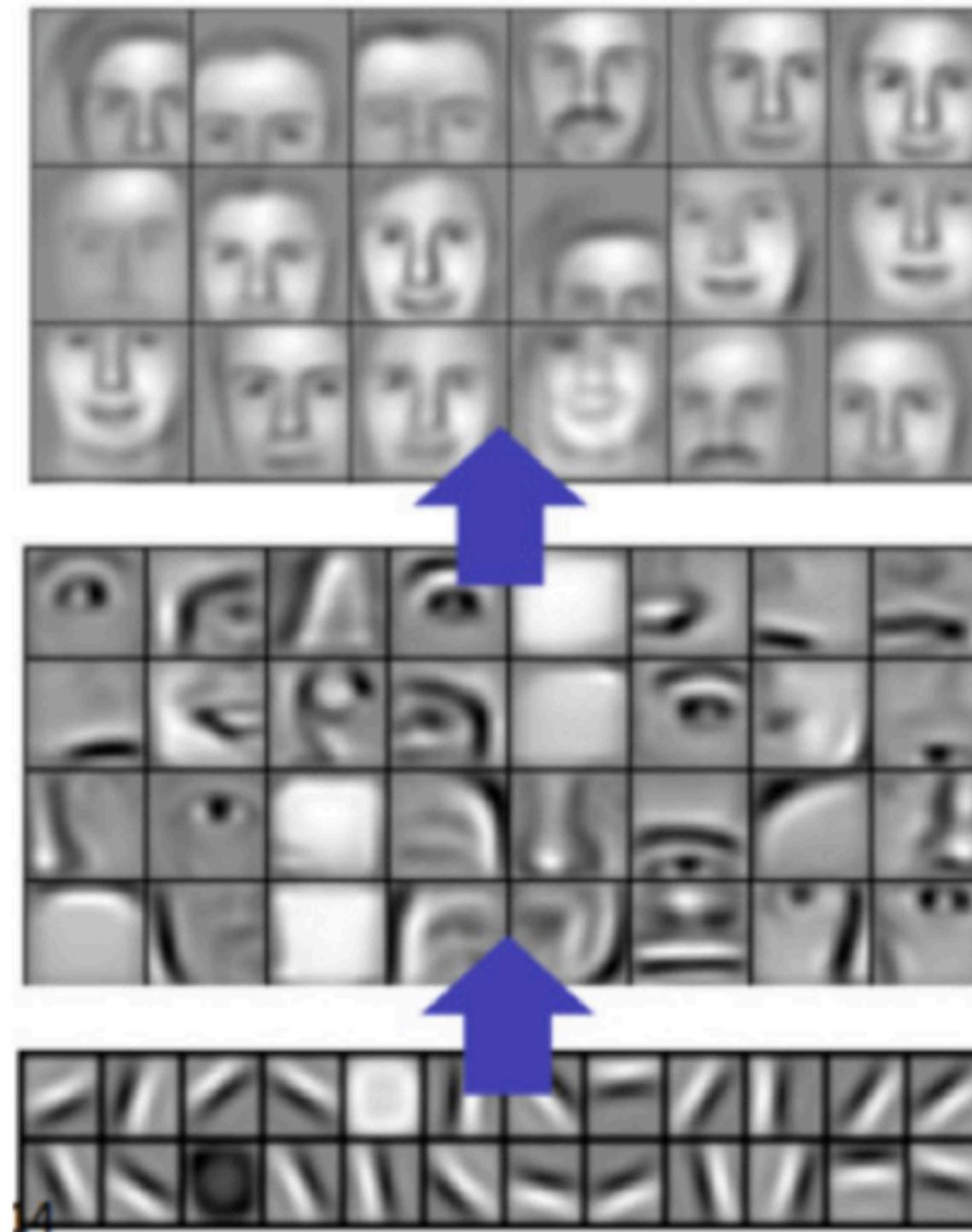
$$w_{t+1} = w_t - \lambda \frac{\partial e(X^{(i,i+n)}, y^{(i,i+n)})}{\partial w_t}$$

Convolution



Source: <https://www.tensorflow.org>

Weights visualization



Layer 3

Layer 2

Layer 1

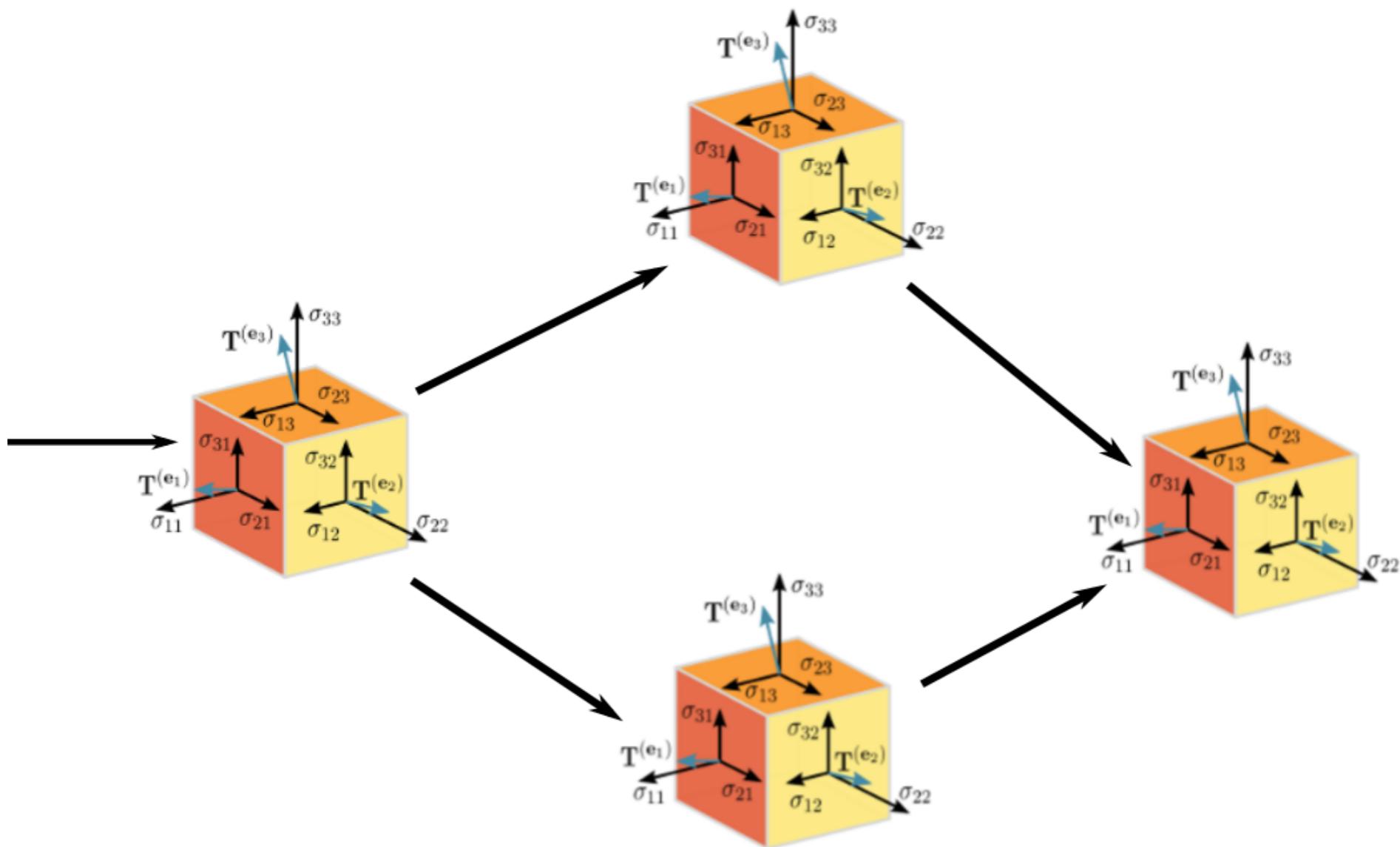
Important terms

- deep learning
- epoch
- dropout

What is not TensorFlow



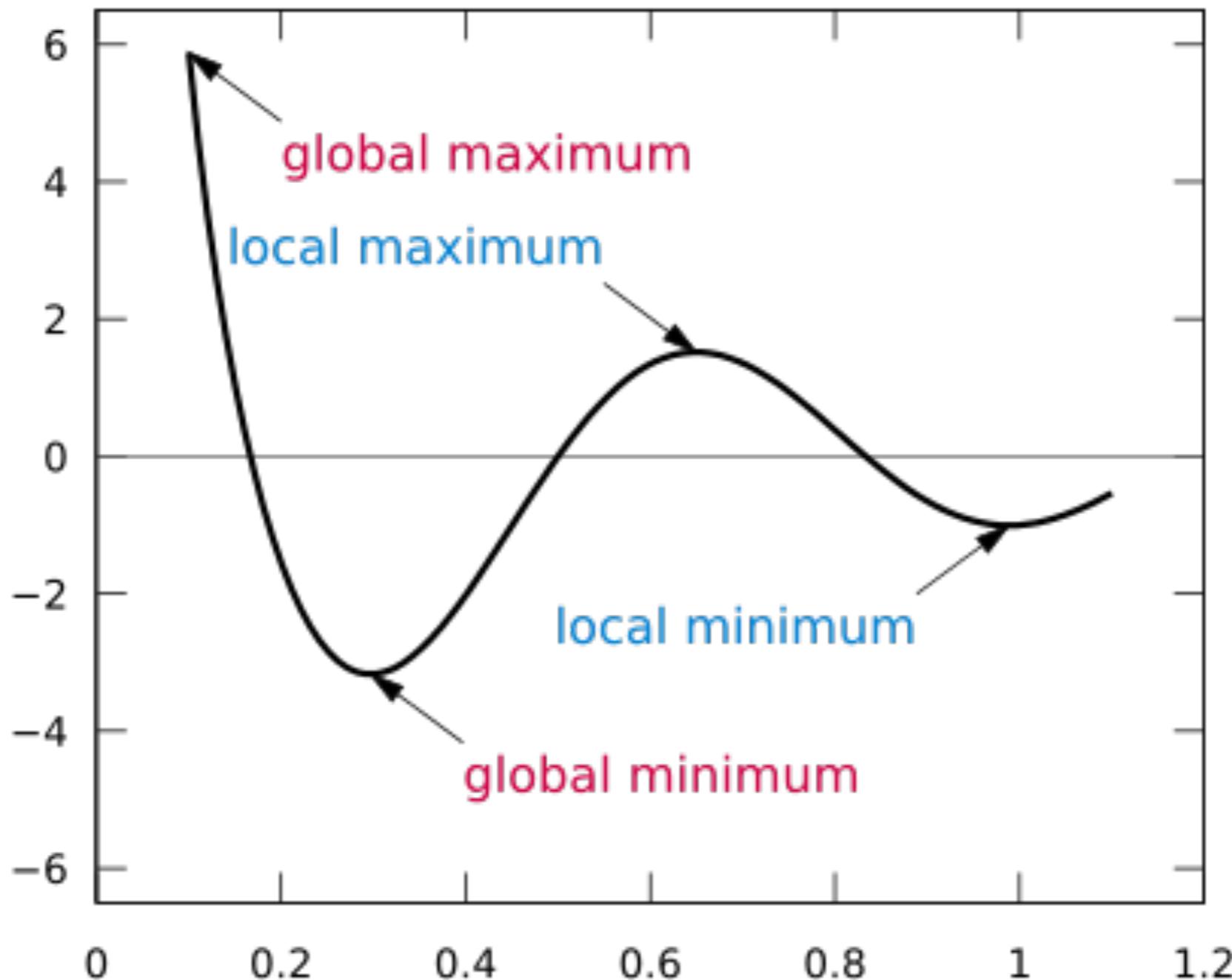
What is TensorFlow?



Keras tutorial

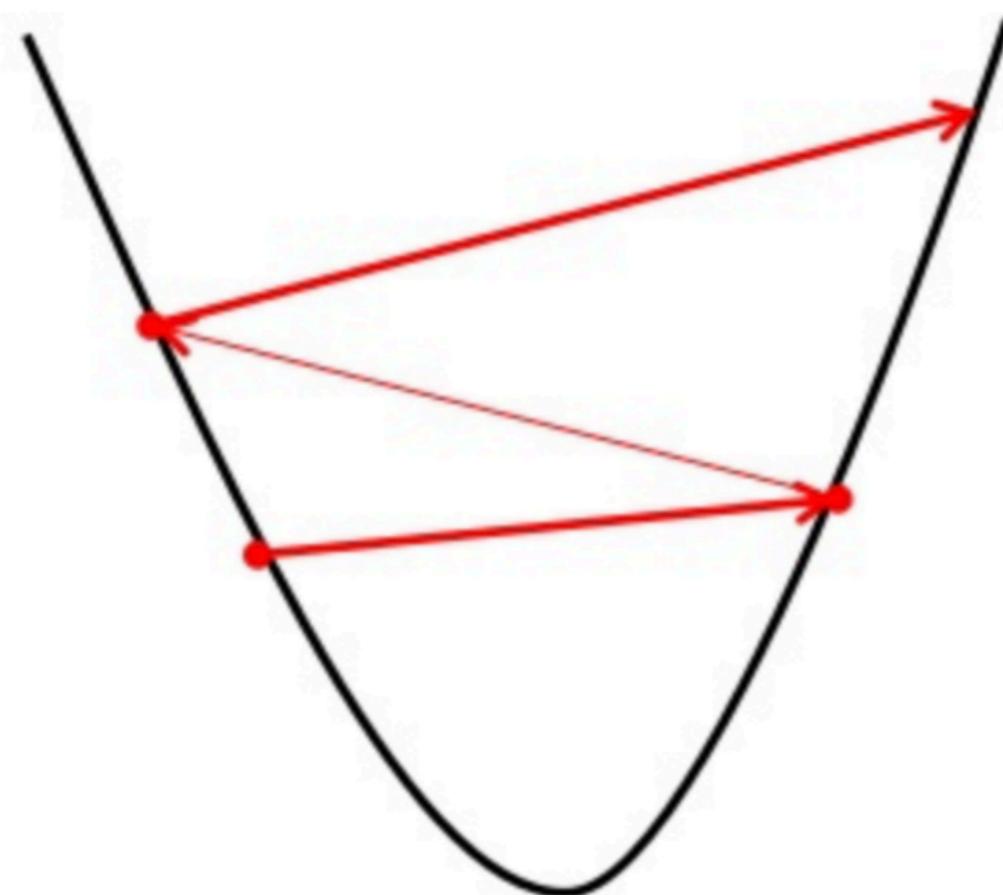
[08-Keras-introduction.ipynb](#)

Parameter optimization strategies

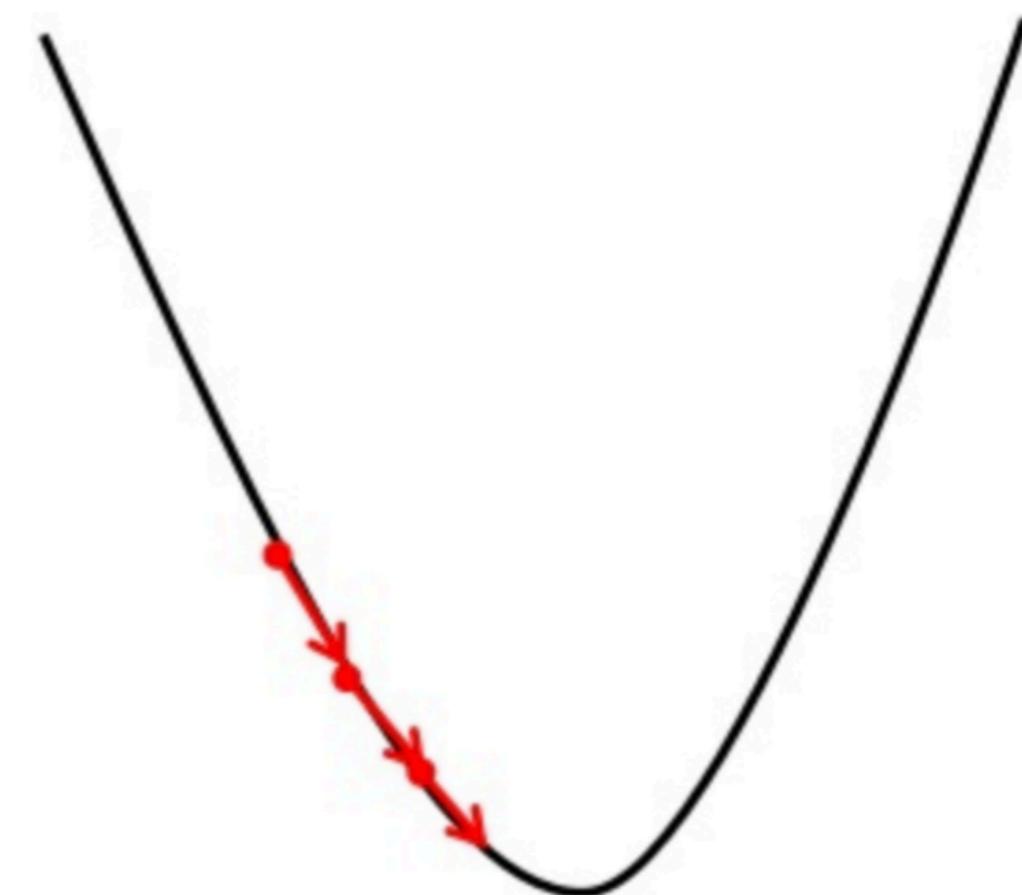


Learning rate tuning

Big learning rate



Small learning rate



Momentum and Nesterov Accelerated Gradient

$$v_t = \gamma v_{t-1} + \lambda \frac{\partial e(w_t)}{\partial w_t}$$

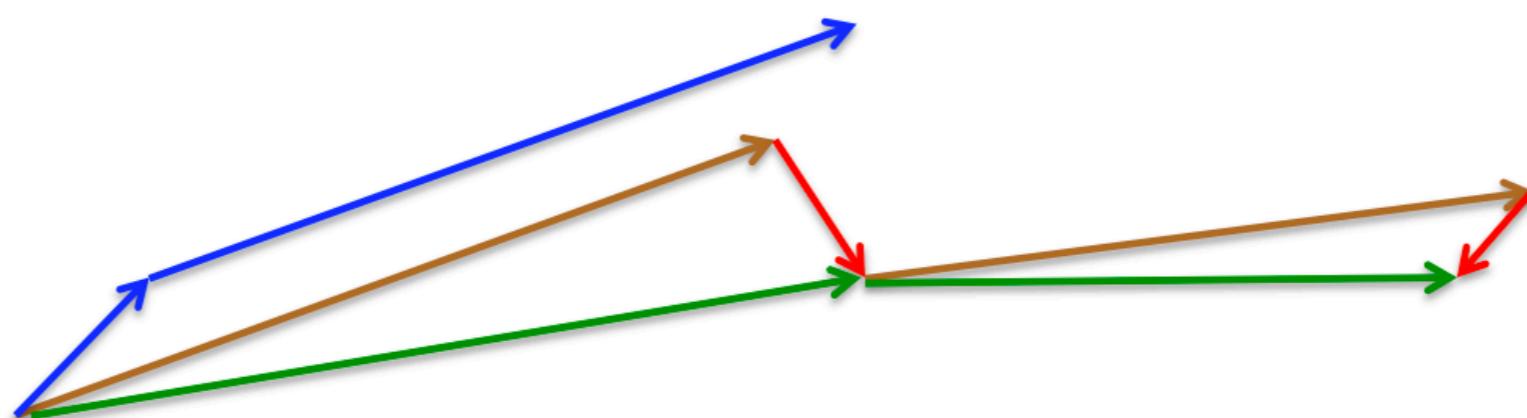
Naive momentum

$$w_{t+1} = w_t - v_t$$

Nesterov Accelerated Gradient

$$v_t = \gamma v_{t-1} + \lambda \frac{\partial e(w_t - \gamma v_{t-1})}{\partial w_t}$$

$$w_{t+1} = w_t - v_t$$



Adaptive Gradient Algorithms

$$w_{t+1} = w_t - \frac{\lambda}{\sqrt{\sum_{i=1}^t g_i^2 + \epsilon}} g_t$$

Adagrad

$$g_i = \frac{\partial e(w_i)}{\partial w_i}$$

$$w_{t+1} = w_t - \frac{\lambda}{\sqrt{\mathbb{E}[g^2]_t - \epsilon}} g_t^2$$

RMSProp

$$\mathbb{E}[g^2]_t = \gamma \mathbb{E}[g^2]_{t-1} + (1 - \gamma) g_t^2$$

Adam and Nadam

Adam

Combination of RMSProp with momentum

Nadam

Combination of RMSProp with Nesterov momentum

Loss functions for deep learning

Mean Squared Error

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

Mean Absolute Error

$$MSE = \frac{\sum_{i=1}^n |y_i - \hat{y}_i|}{n}$$

Cross Entropy (Negative Log Likelihood)

Categorical Cross Entropy
$$CCE = -\frac{\sum_{i=1}^n \sum_{j=1}^c y_{i,j} \log(\hat{y}_{i,j})}{n}$$

Binary Cross Entropy

$$BCE = -\frac{\sum_{i=1}^n [y_{i,j} \log(\hat{y}_{i,j}) + (1 - y_{i,j}) \log(1 - \hat{y}_{i,j})]}{n}$$

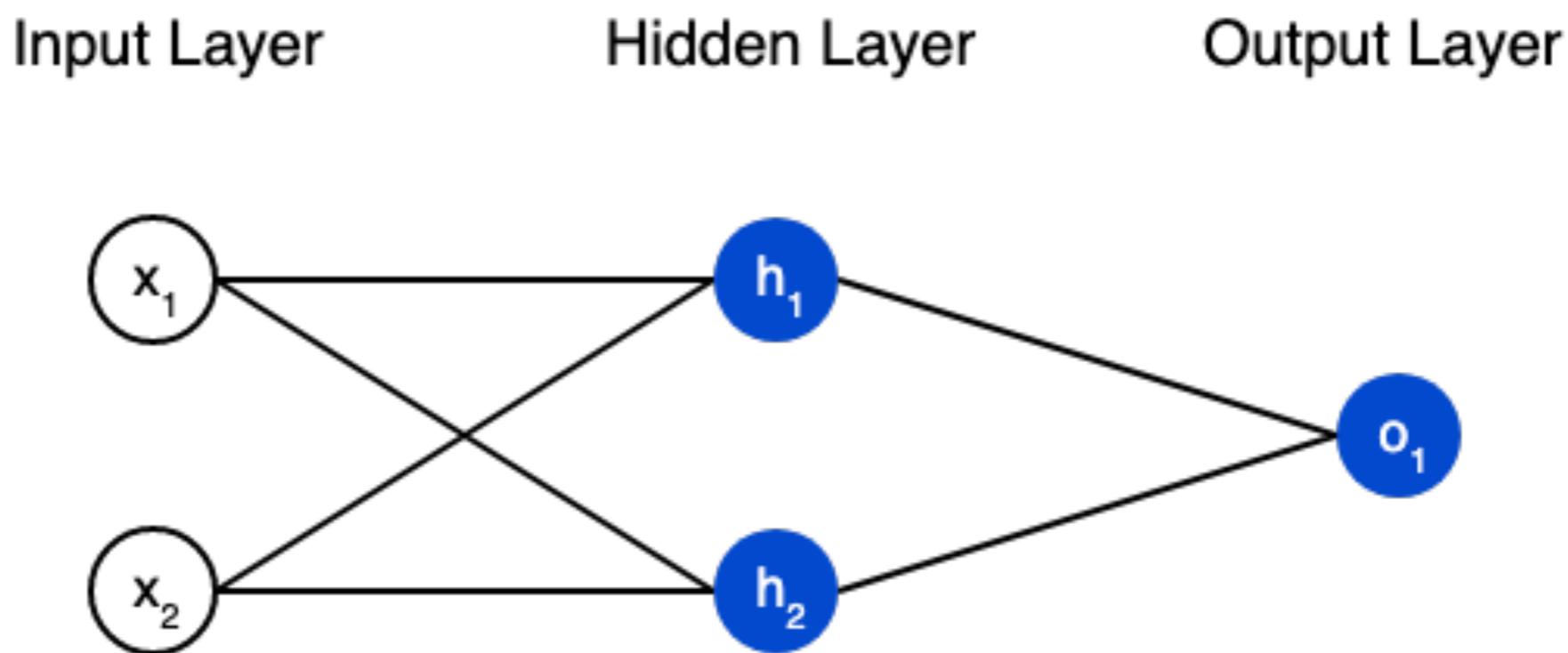
Reimplementation of the classification and regression task using NN

[**09-Classification-nn-assignment.ipynb**](#)

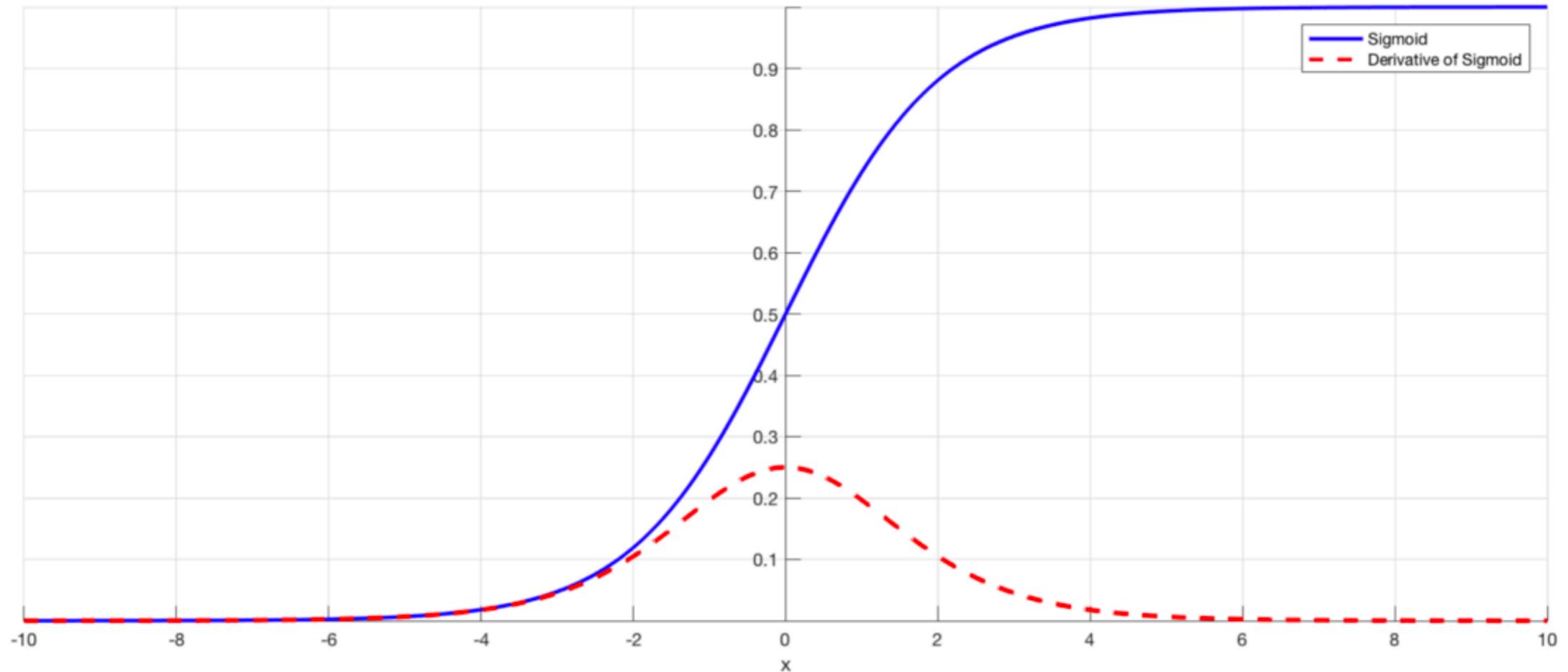
[**10-Regression-nn-assignment.ipynb**](#)

Weight initialization

Zero or constant initialization



Too low or too high initialization



Xavier and He initializers

1. The mean of the activations should be zero
2. The variance of the activations should stay the same across every layer

**Xavier (Glorot) initialization
for tanh**

$$\mathbf{W}^l \sim \mathcal{N}(\mu = 0, \sigma^2 = \frac{1}{n^{l-1}})$$

$$b^l = 0$$

**He (Kaiming) initialization
for relu**

$$\mathbf{W}^l \sim \mathcal{N}(\mu = 0, \sigma^2 = \frac{2}{n^{l-1}})$$

$$b^l = 0$$

Experiment with various initializations for a deep network

[08-Regression-nn-assignment.ipynb](#)

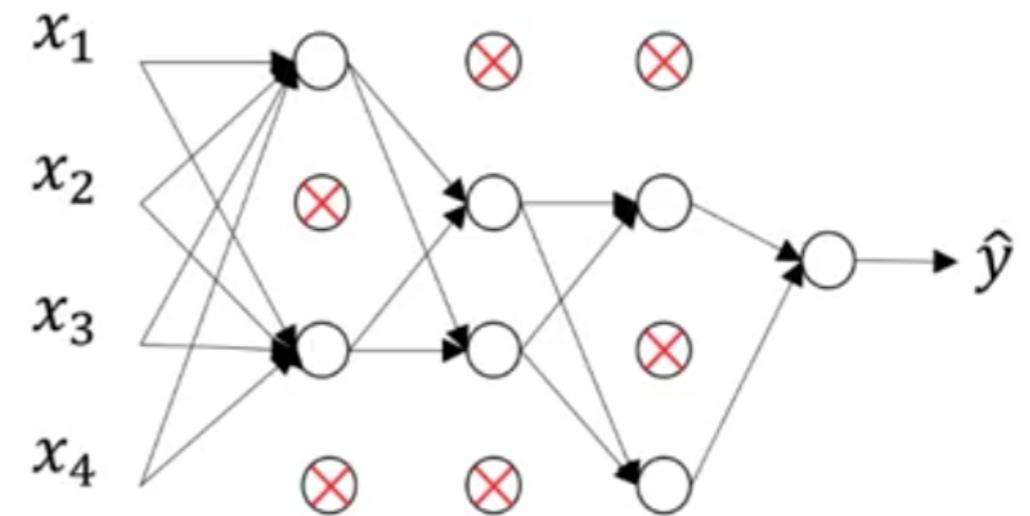
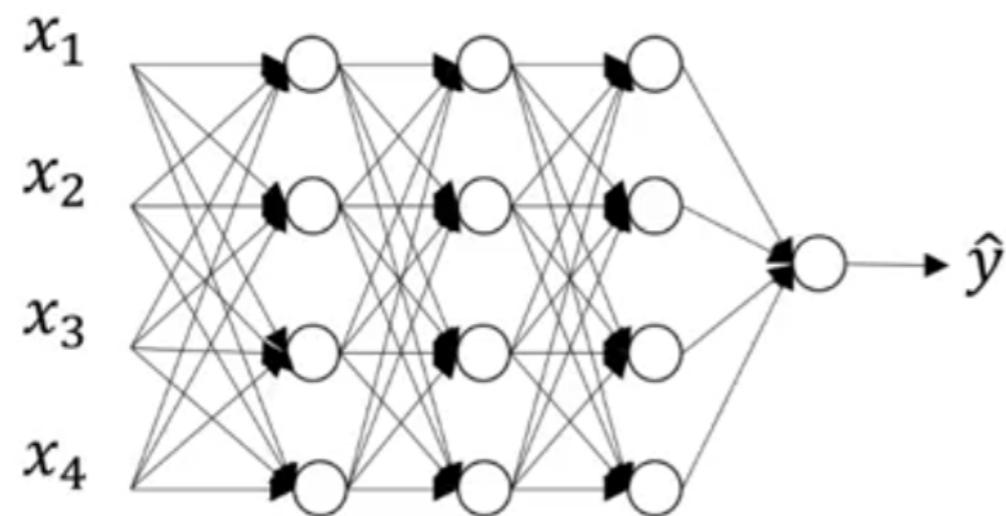
L2 Regularization in deep learning

$$cost(w^1, b^1, \dots, w^L, b^L) = \frac{1}{n} \sum_{i=1}^n Loss(y_i, \hat{y}_i) + \frac{\lambda}{2n} \sum_{l=1}^L \|w^l\|_F^2$$

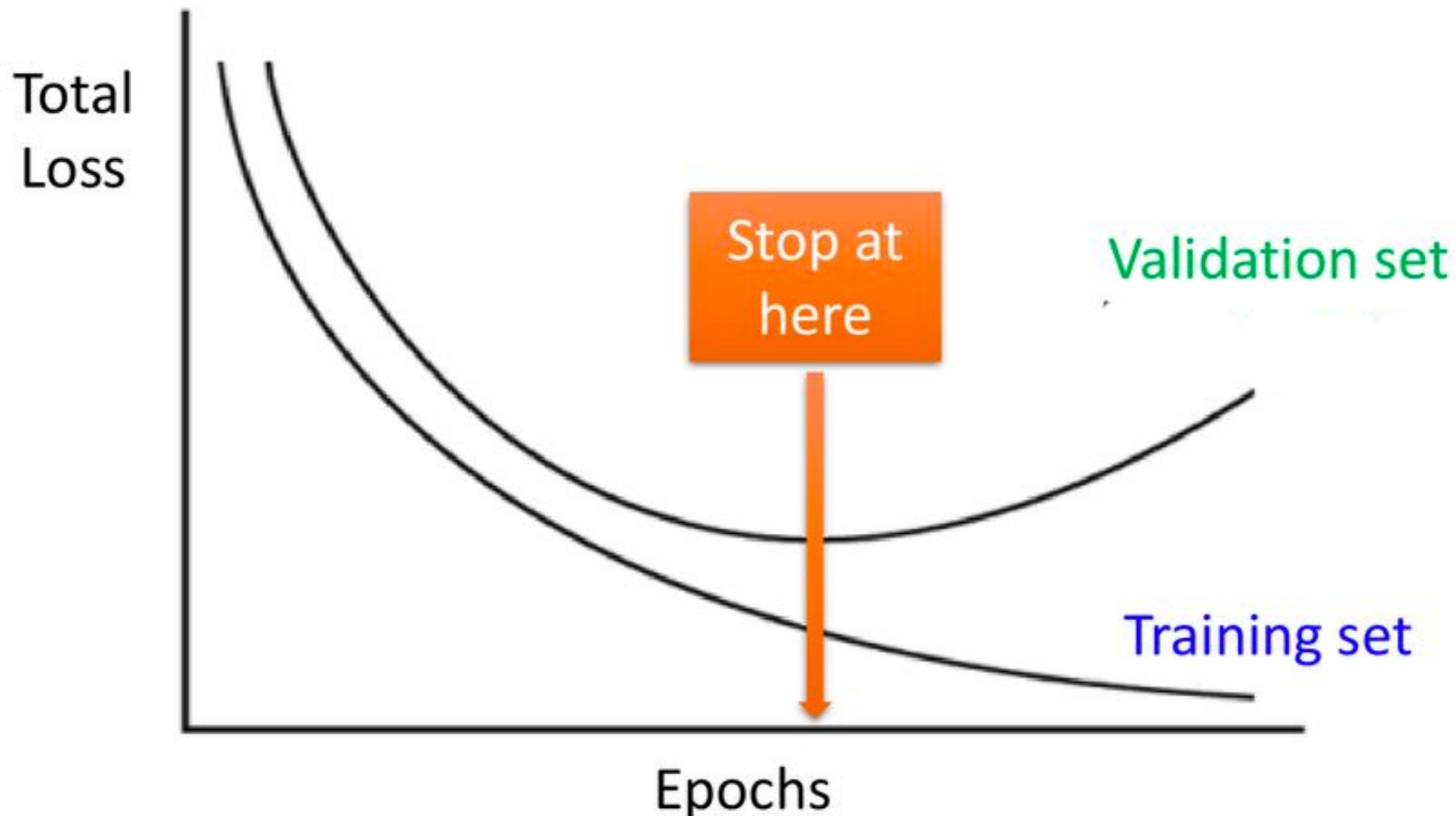
Frobenius norm

$$\|w\|_F = \sqrt{\sum_{i=1}^n \sum_{j=1}^m |w_{i,j}|^2}$$

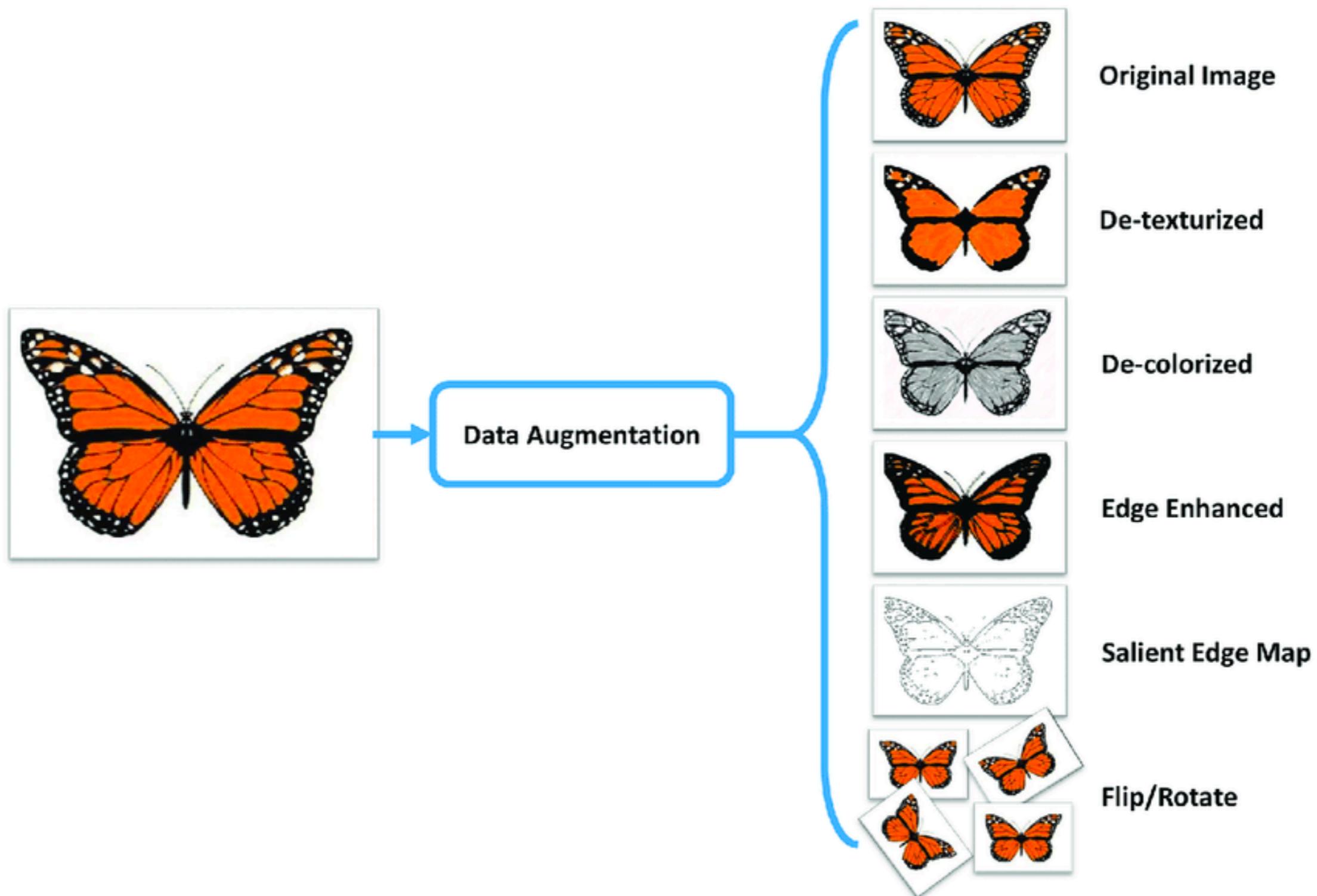
Dropout



Early stopping



Data augmentation



Batch normalization

Input: Values of x over a mini-batch: $\mathcal{B} = \{x_1 \dots m\}$;

Parameters to be learned: γ, β

Output: $\{y_i = \text{BN}_{\gamma, \beta}(x_i)\}$

$$\mu_{\mathcal{B}} \leftarrow \frac{1}{m} \sum_{i=1}^m x_i \quad // \text{mini-batch mean}$$

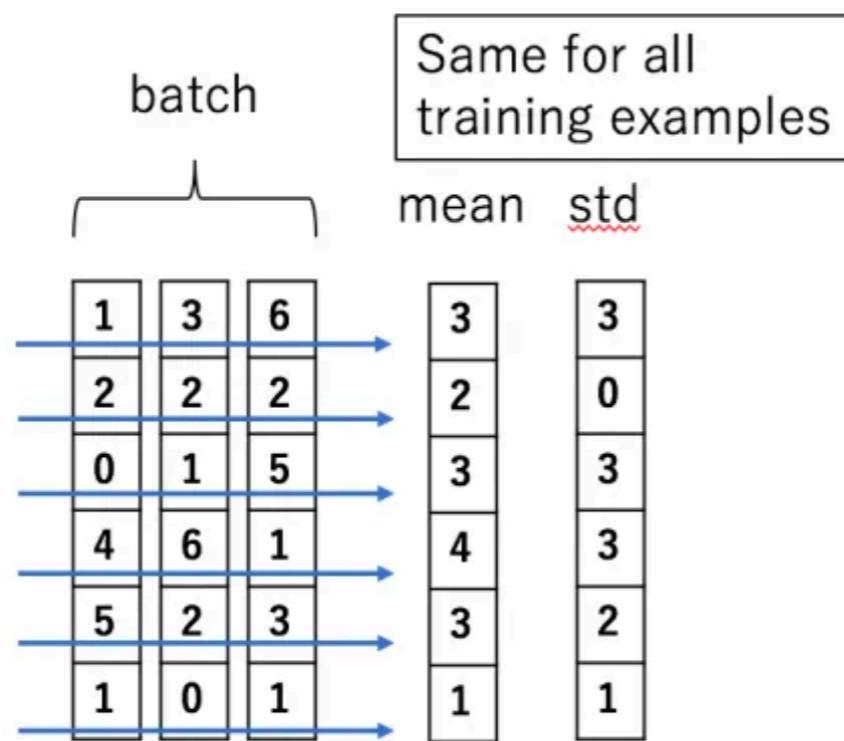
$$\sigma_{\mathcal{B}}^2 \leftarrow \frac{1}{m} \sum_{i=1}^m (x_i - \mu_{\mathcal{B}})^2 \quad // \text{mini-batch variance}$$

$$\hat{x}_i \leftarrow \frac{x_i - \mu_{\mathcal{B}}}{\sqrt{\sigma_{\mathcal{B}}^2 + \epsilon}} \quad // \text{normalize}$$

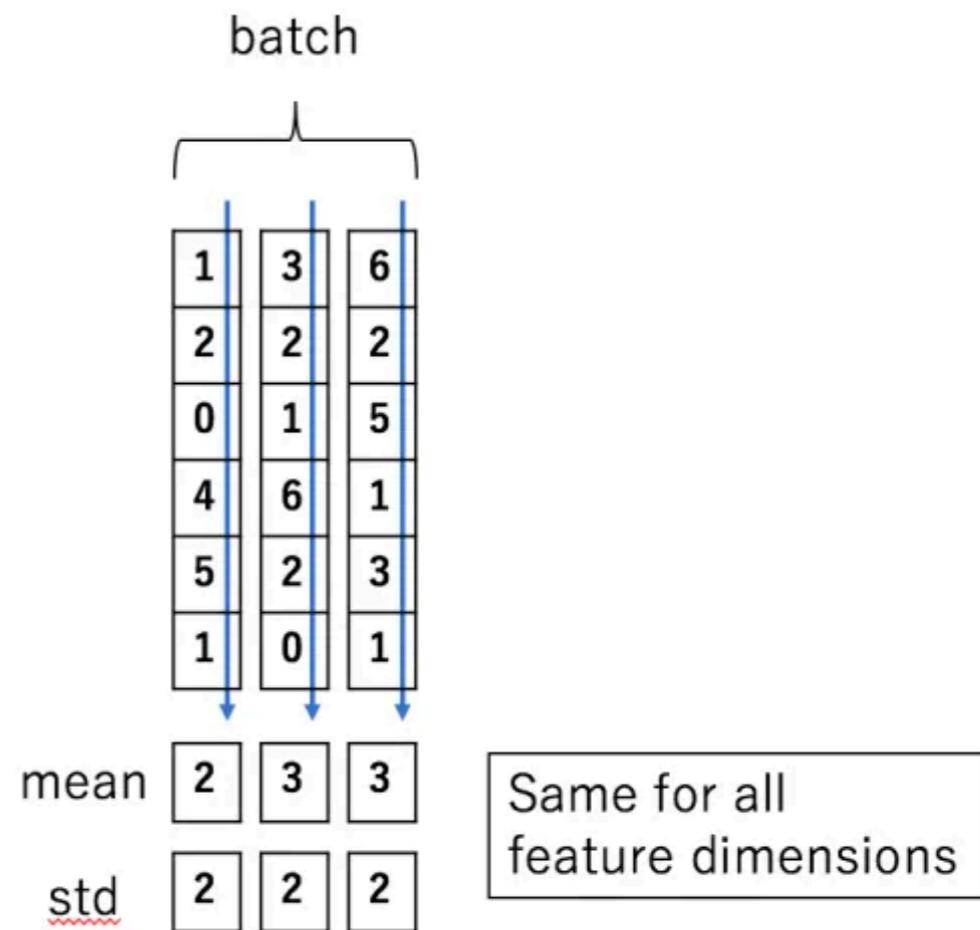
$$y_i \leftarrow \gamma \hat{x}_i + \beta \equiv \text{BN}_{\gamma, \beta}(x_i) \quad // \text{scale and shift}$$

Layer normalization

Batch Normalization



Layer Normalization



Functional API in Keras

```
1 # Sequential model
2 from keras.models import Sequential
3 from keras.layers import Dense
4
5 model = Sequential()
6 model.add(10, input_shape=(10,), activation='relu')
7 model.add(Dense(20, activation='relu'))
8 model.add(Dense(10, activation='relu'))
9 model.add(Dense(1, activation='sigmoid'))
```

```
1 # Functional model
2 from keras.models import Model
3 from keras.layers import Input, Dense
4
5 visible = Input(shape=(10,))
6 hidden1 = Dense(10, activation='relu')(visible)
7 hidden2 = Dense(20, activation='relu')(hidden1)
8 hidden3 = Dense(10, activation='relu')(hidden2)
9 output = Dense(1, activation='sigmoid')(hidden3)
10 model = Model(inputs=visible, outputs=output)
```

Shared Input

```
1 # Shared Input Layer
2 from keras.utils import plot_model
3 from keras.models import Model
4 from keras.layers import Input, Dense, Flatten
5 from keras.layers.convolutional import Conv2D
6 from keras.layers.pooling import MaxPooling2D
7 from keras.layers.merge import concatenate
8 # input layer
9 visible = Input(shape=(64,64,1))
10 # first feature extractor
11 conv1 = Conv2D(32, kernel_size=4, activation='relu')(visible)
12 pool1 = MaxPooling2D(pool_size=(2, 2))(conv1)
13 flat1 = Flatten()(pool1)
14 # second feature extractor
15 conv2 = Conv2D(16, kernel_size=8, activation='relu')(visible)
16 pool2 = MaxPooling2D(pool_size=(2, 2))(conv2)
17 flat2 = Flatten()(pool2)
18 # merge feature extractors
19 merge = concatenate([flat1, flat2])
20 # interpretation layer
21 hidden1 = Dense(10, activation='relu')(merge)
22 # prediction output
23 output = Dense(1, activation='sigmoid')(hidden1)
24 model = Model(inputs=visible, outputs=output)
```

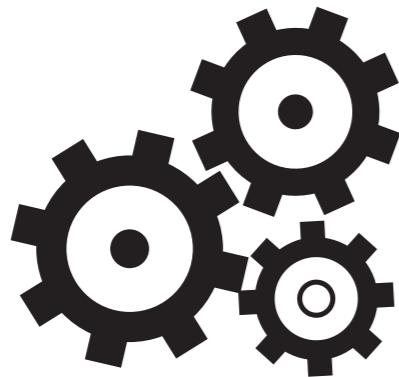
Multiple inputs (outputs)

```
1 # Multiple Inputs
2 from keras.utils import plot_model
3 from keras.models import Model
4 from keras.layers import Input
5 from keras.layers import Dense
6 from keras.layers import Flatten
7 from keras.layers.convolutional import Conv2D
8 from keras.layers.pooling import MaxPooling2D
9 from keras.layers.merge import concatenate
10 # first input model
11 visible1 = Input(shape=(64,64,1))
12 conv11 = Conv2D(32, kernel_size=4, activation='relu')(visible1)
13 pool11 = MaxPooling2D(pool_size=(2, 2))(conv11)
14 conv12 = Conv2D(16, kernel_size=4, activation='relu')(pool11)
15 pool12 = MaxPooling2D(pool_size=(2, 2))(conv12)
16 flat1 = Flatten()(pool12)
17 # second input model
18 visible2 = Input(shape=(32,32,3))
19 conv21 = Conv2D(32, kernel_size=4, activation='relu')(visible2)
20 pool21 = MaxPooling2D(pool_size=(2, 2))(conv21)
21 conv22 = Conv2D(16, kernel_size=4, activation='relu')(pool21)
22 pool22 = MaxPooling2D(pool_size=(2, 2))(conv22)
23 flat2 = Flatten()(pool22)
24 # merge input models
25 merge = concatenate([flat1, flat2])
26 # interpretation model
27 hidden1 = Dense(10, activation='relu')(merge)
28 hidden2 = Dense(10, activation='relu')(hidden1)
29 output = Dense(1, activation='sigmoid')(hidden2)
30 model = Model(inputs=[visible1, visible2], outputs=output)
```

Practical example on regularization and normalization

[**09-Normalization-and-regularization-assignment.ipynb**](#)

What next?



Machine Learning Prague

ML MACHINE LEARNING
MU meetups

Thank you for your attention

e-mail: jiri@mlcollege.com

Web: www.mlcollege.com

Twitter: @JiriMaterna

Facebook: <https://www.facebook.com/maternajiri>

LinkedIn: <https://www.linkedin.com/in/jirimaterna/>