Introduction to Machine Learning

Jiří Materna

@mlcollegecom

@mlcollegecom

#mlcollege

About me

- Ph.D. in Natural Language Processing and Artificial Intelligence at Masaryk University
- 10 years at <u>Seznam.cz</u> (last 8 years as Head Of Research)
- Founder and lecturer at ML College
- Founder and co-organizer of ML Prague
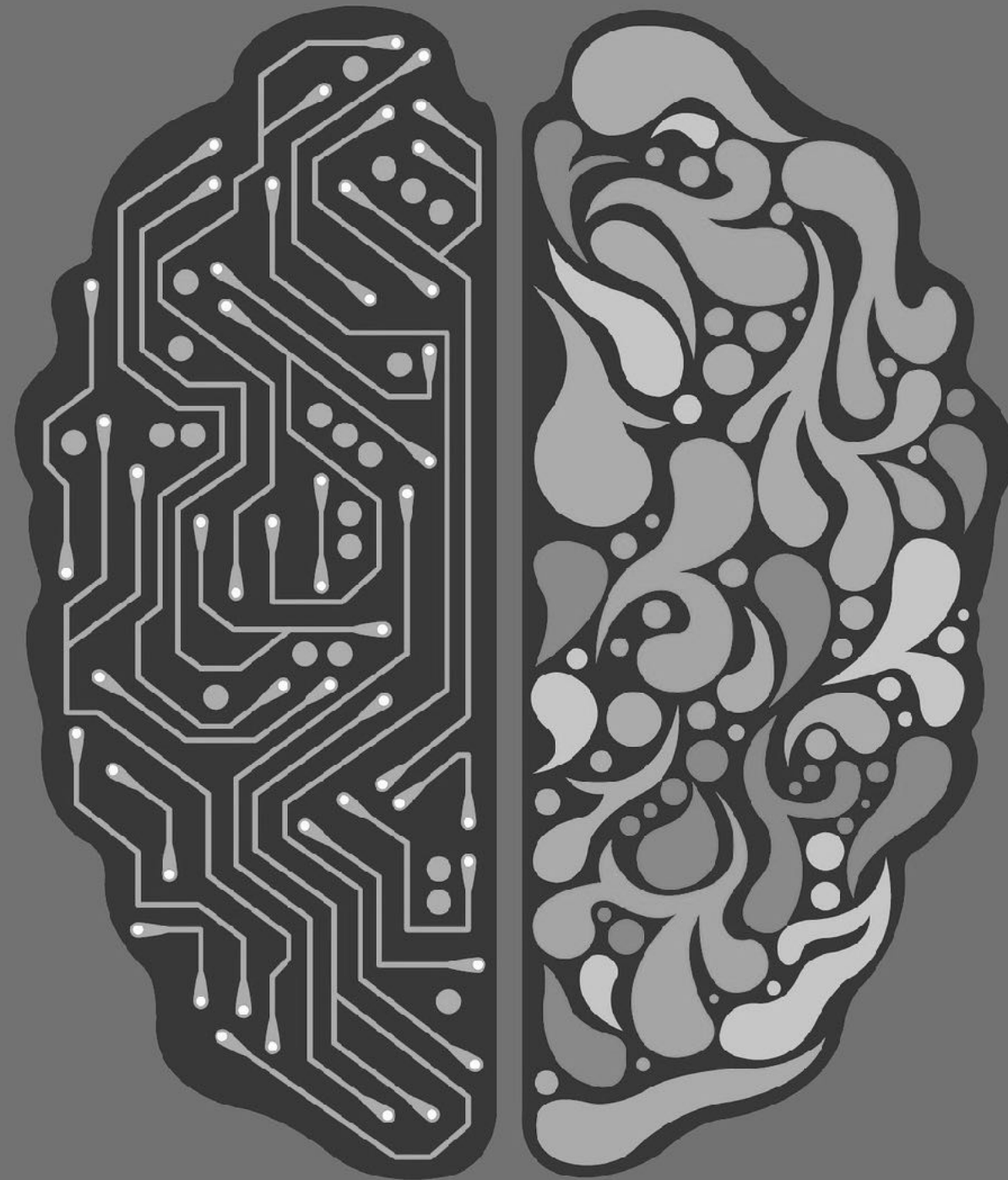- ML Freelance and consultant

<u>www.mlguru.com</u>   <u>www.mlprague.com</u>   <u>www.mlcollege.com</u>

Outline

- 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
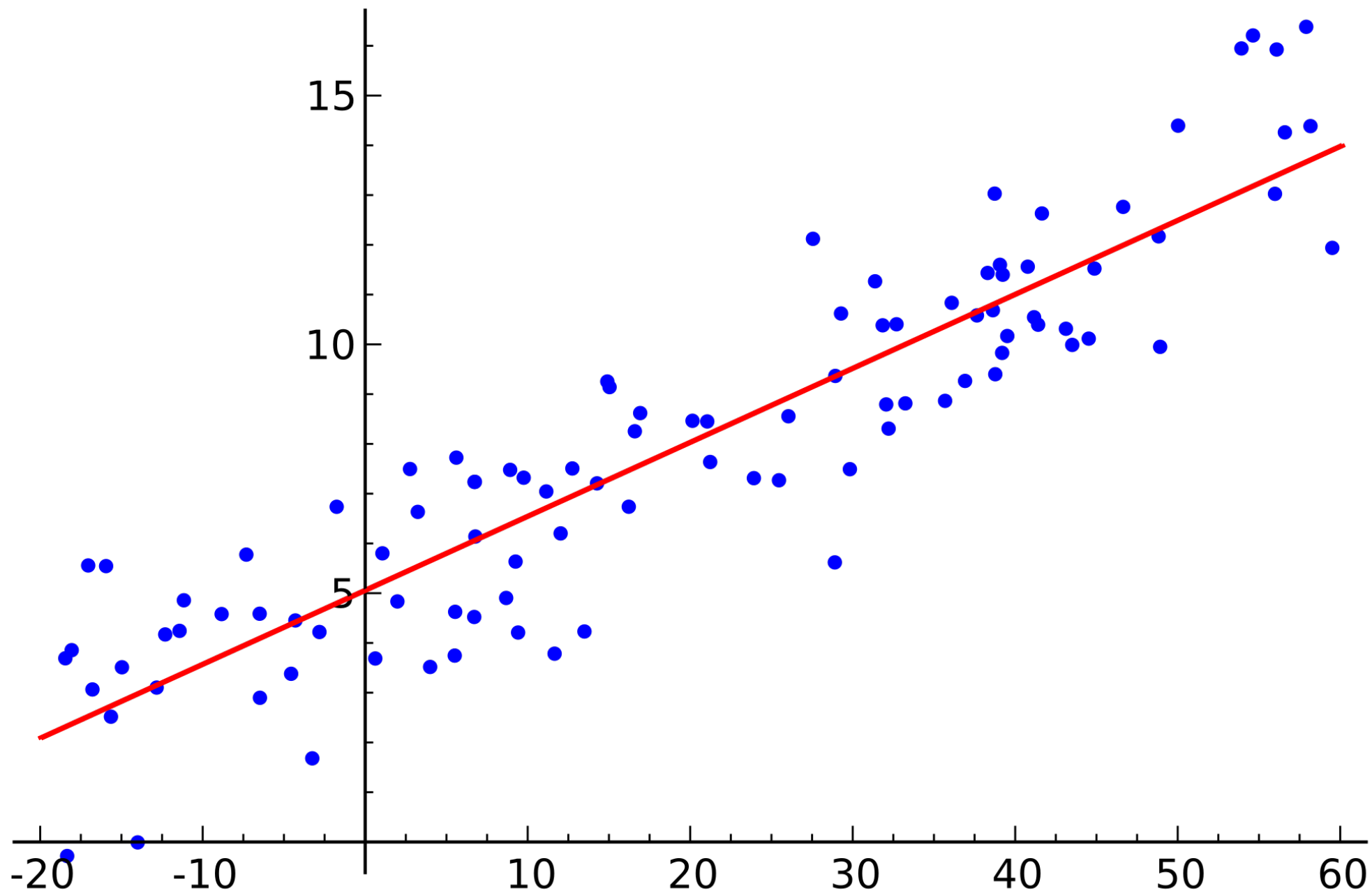- Practical regression task

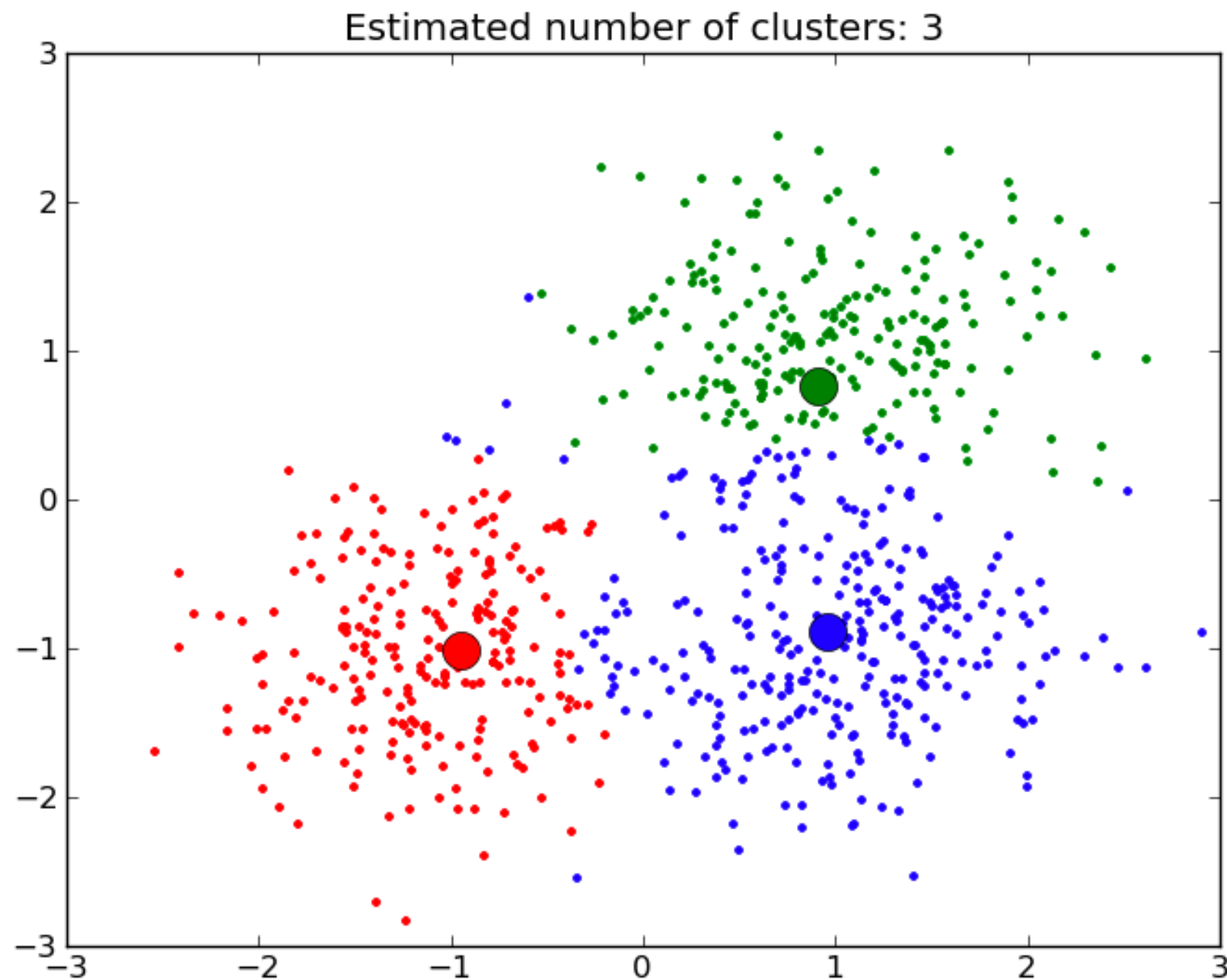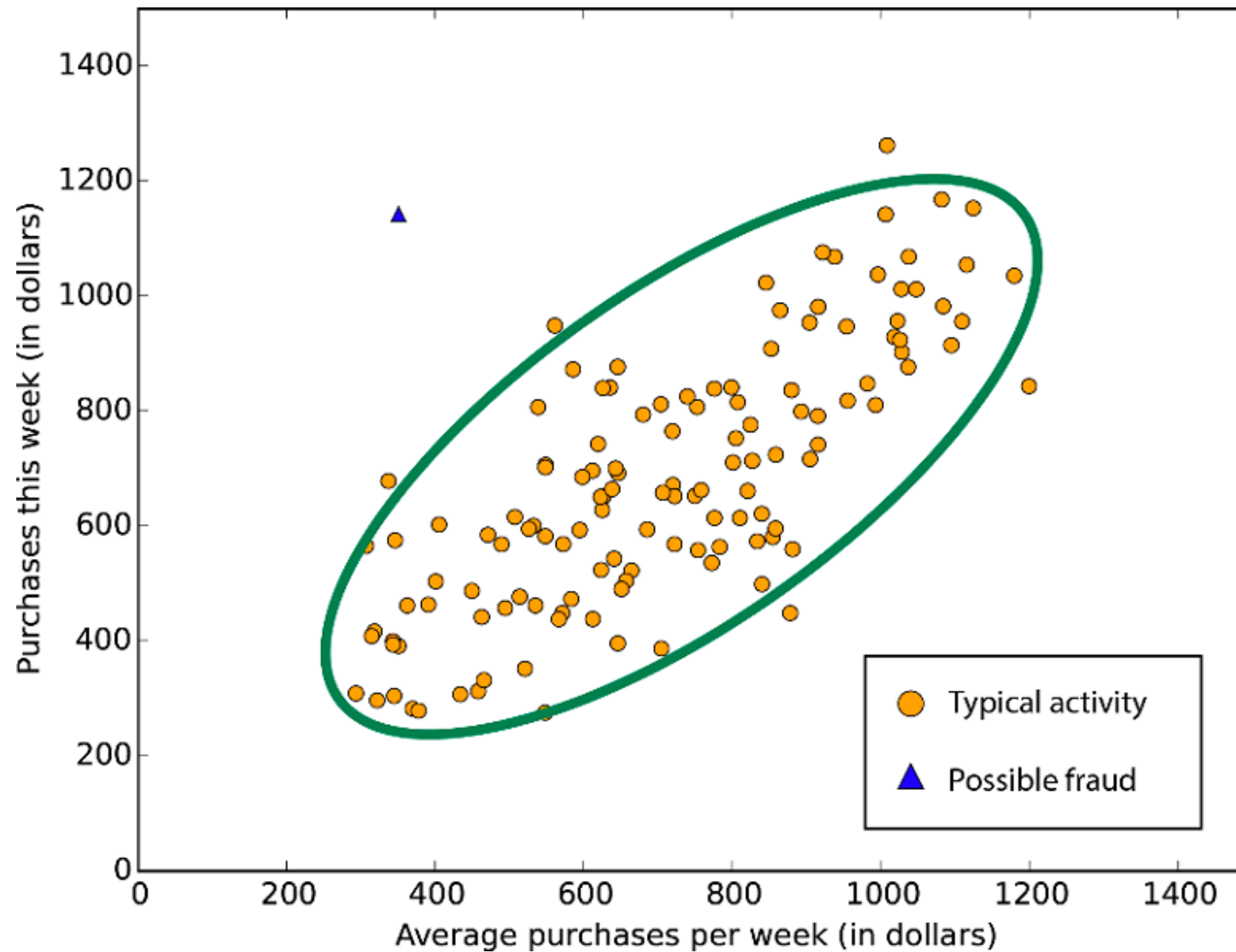# What is (not) machine learning?

Classification

Regression

Clustering

Source: http://scikit-learn.sourceforge.net
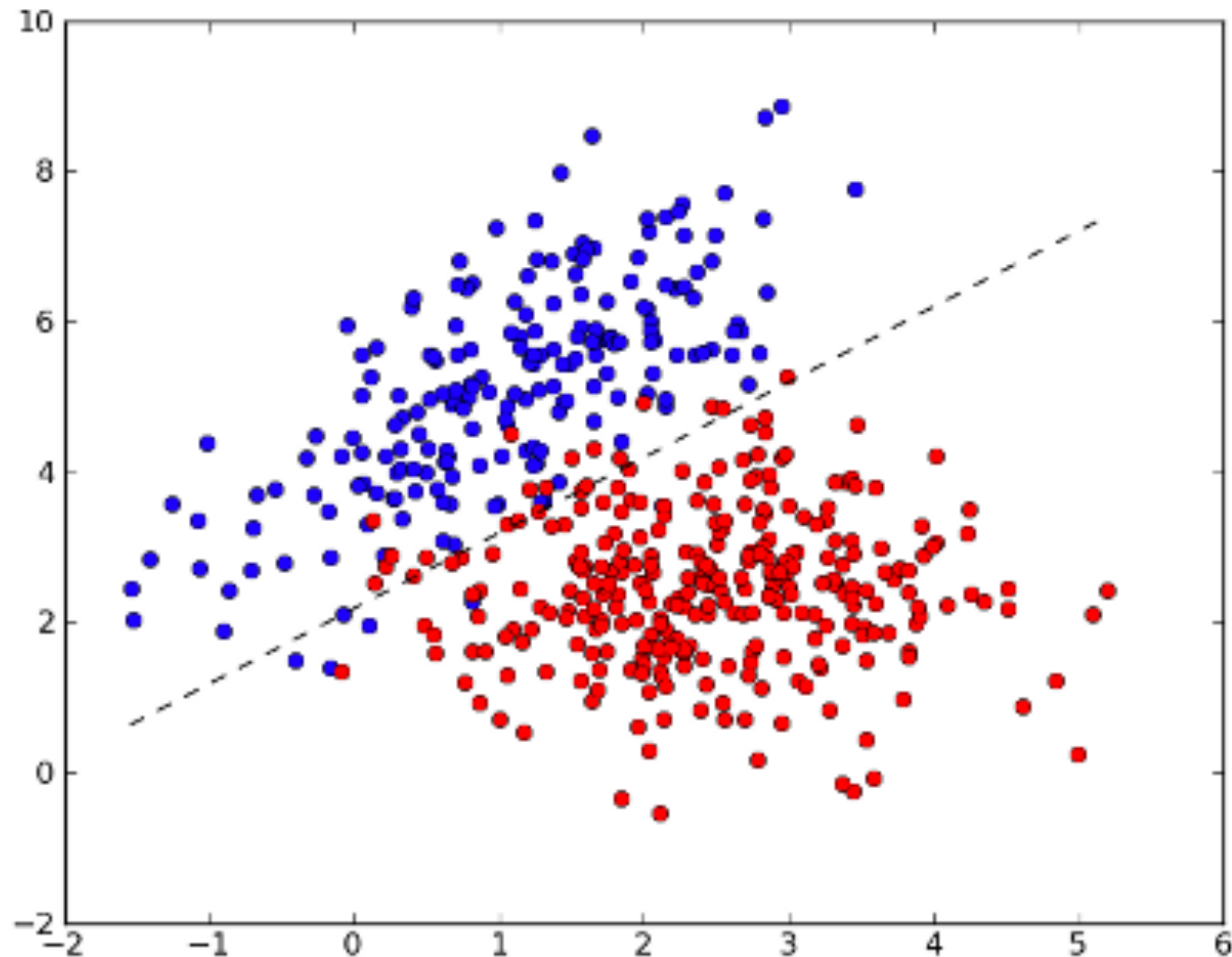
Anomaly detection

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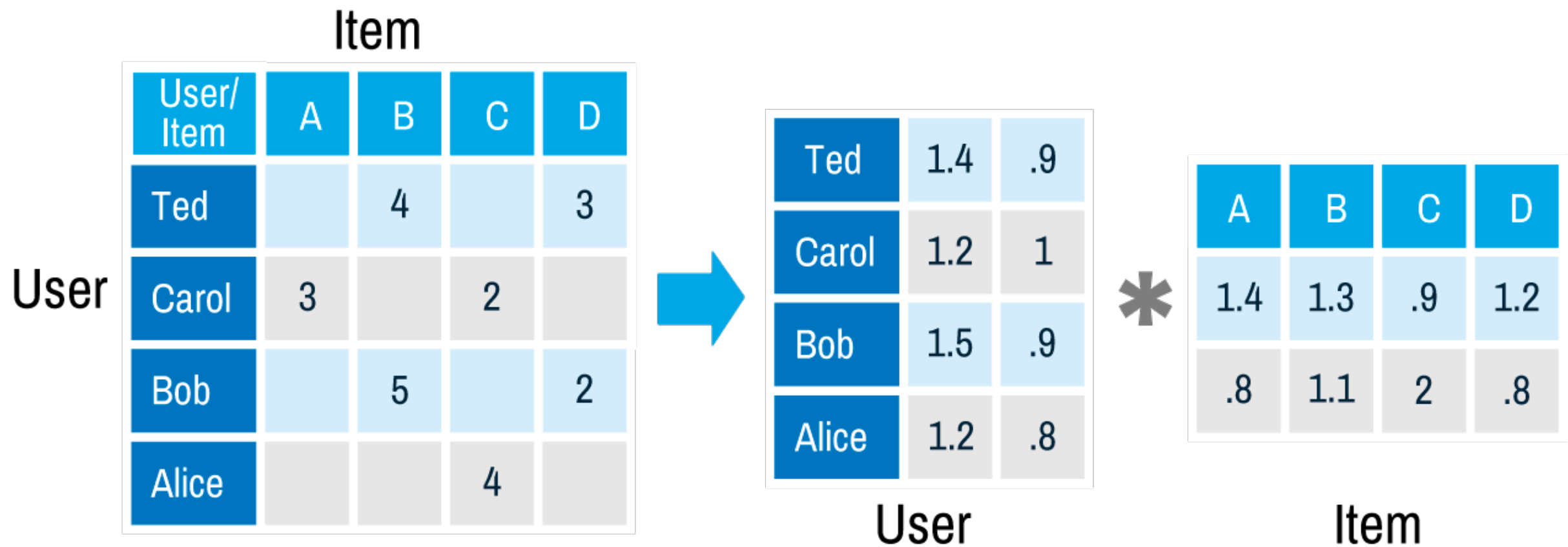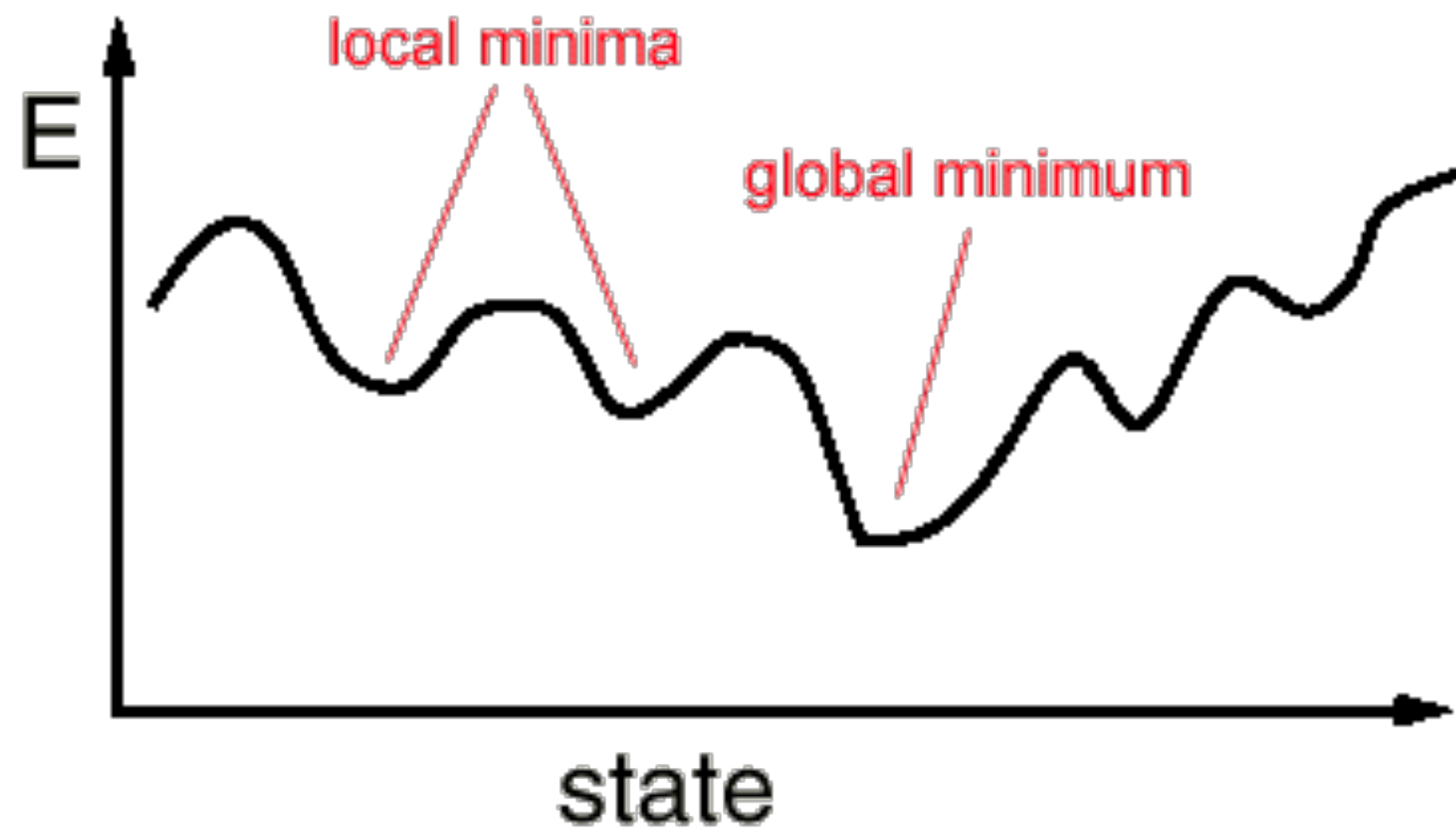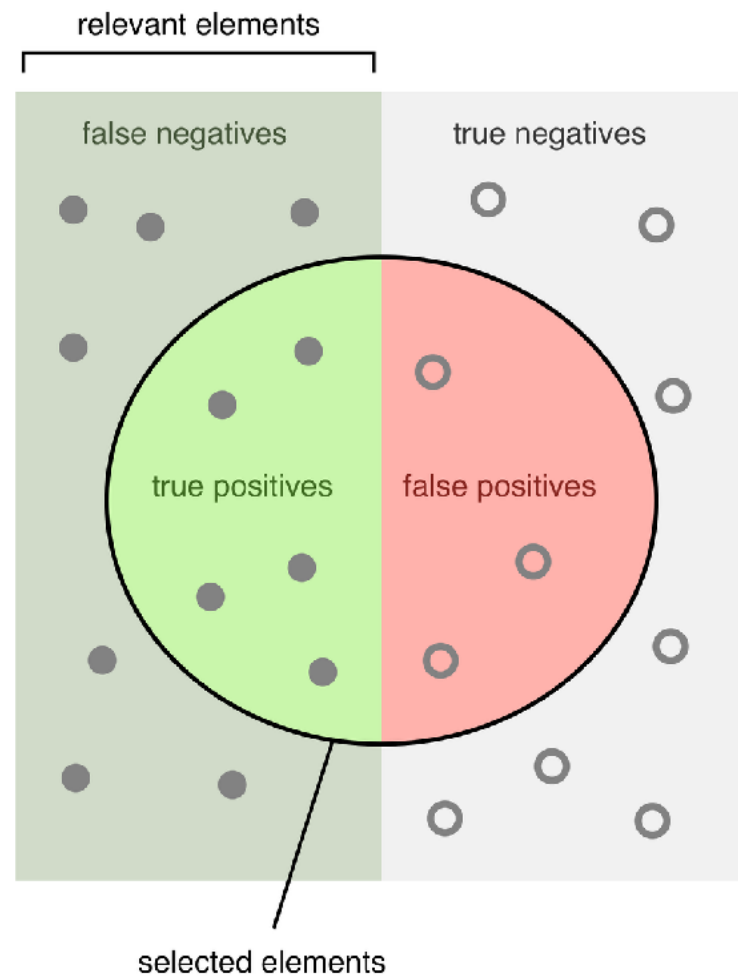
Recommendation

Optimization

Data

- Train, validation, test data sets

- Cross-validation

- Imbalanced data sets
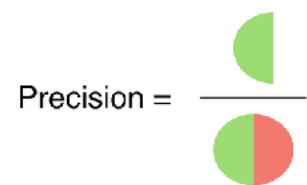
- Baseline models

Model evaluation

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

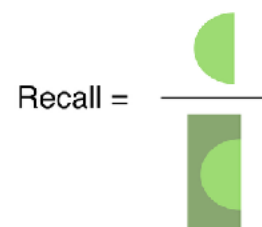$$\text{Precision} = \frac{tp}{tp + fp}$$

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

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

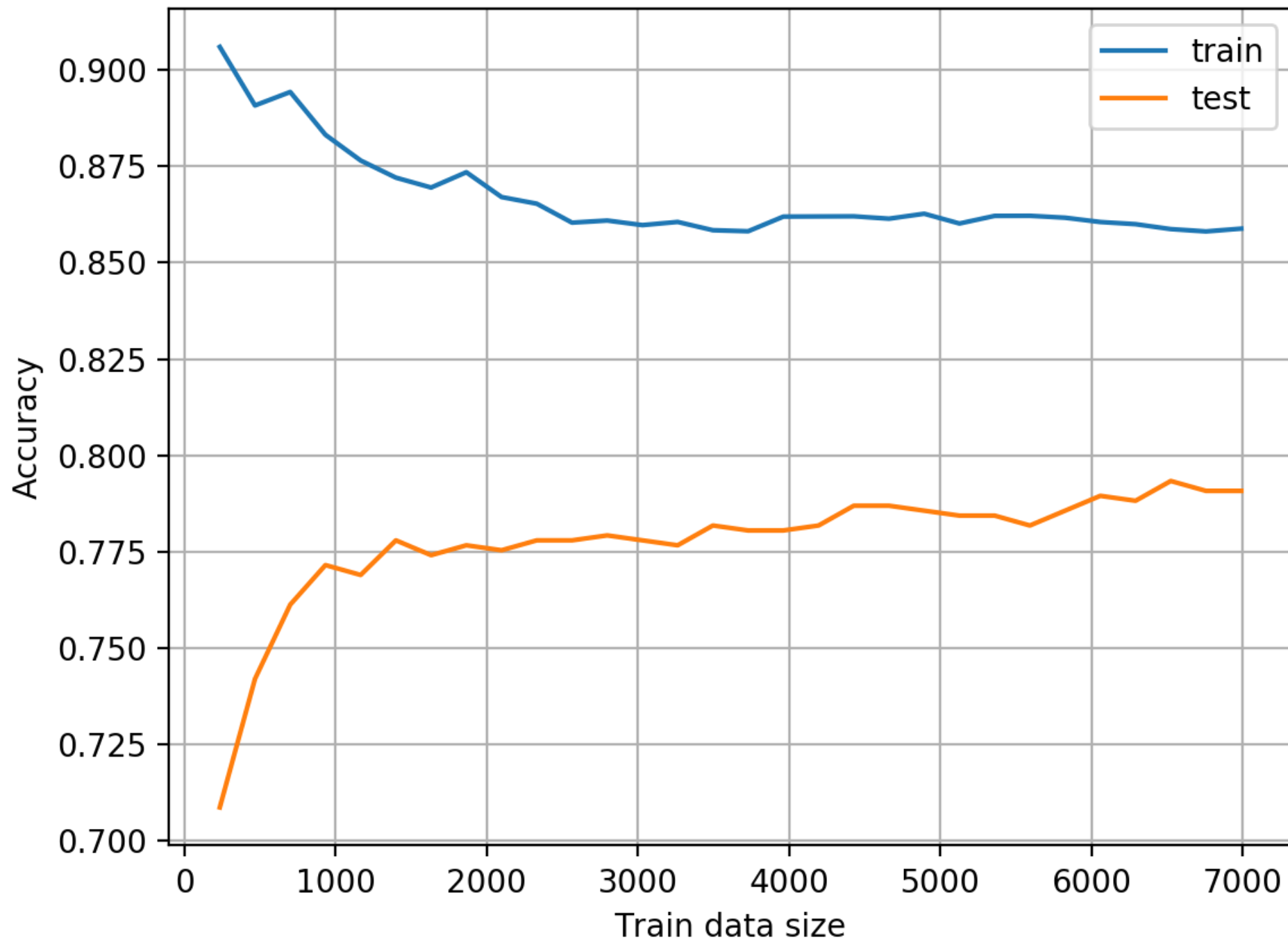Source: https://www.wikipedia.org

Receiver operating characteristic (ROC Curve), AUC
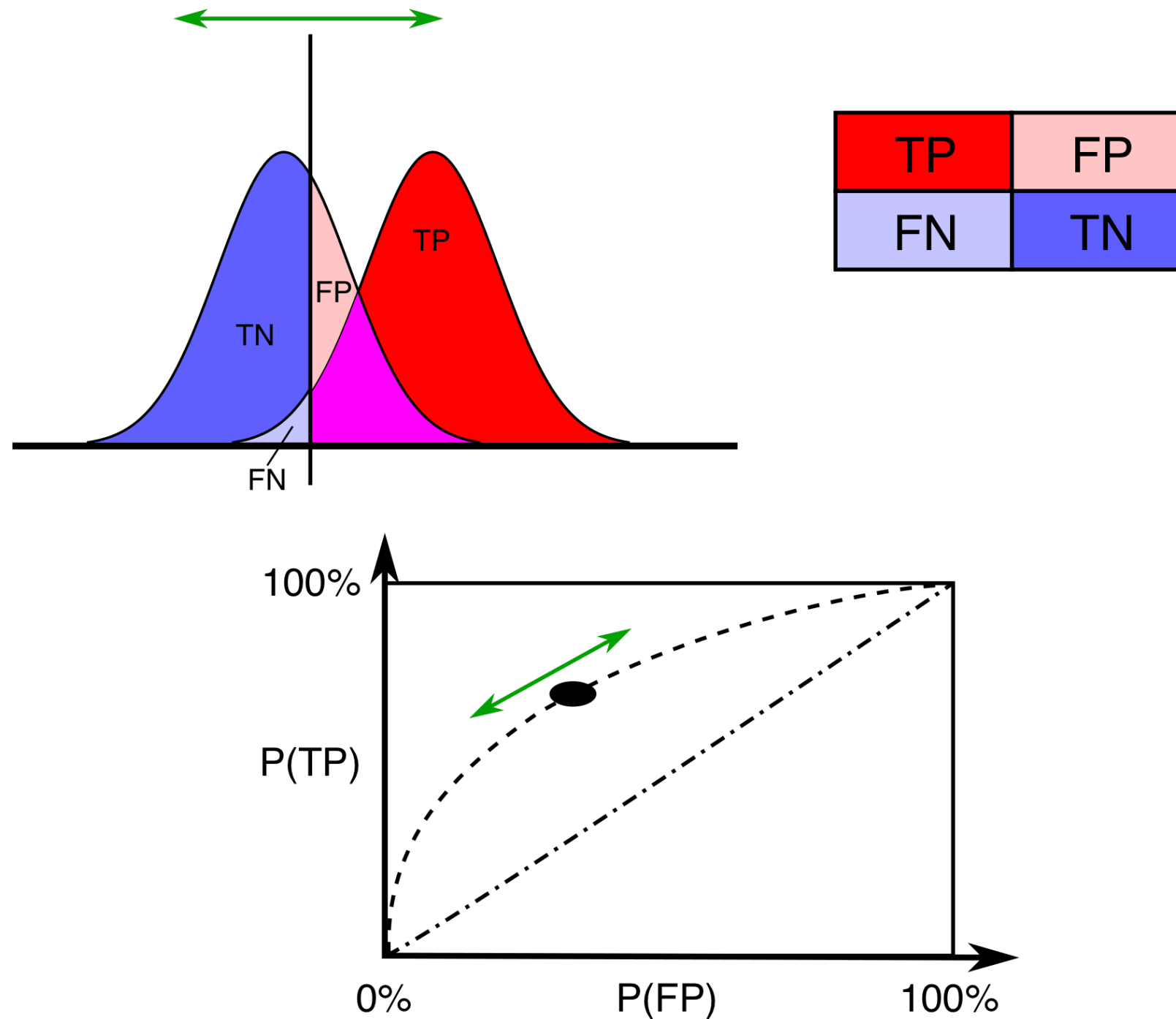
Overfitting
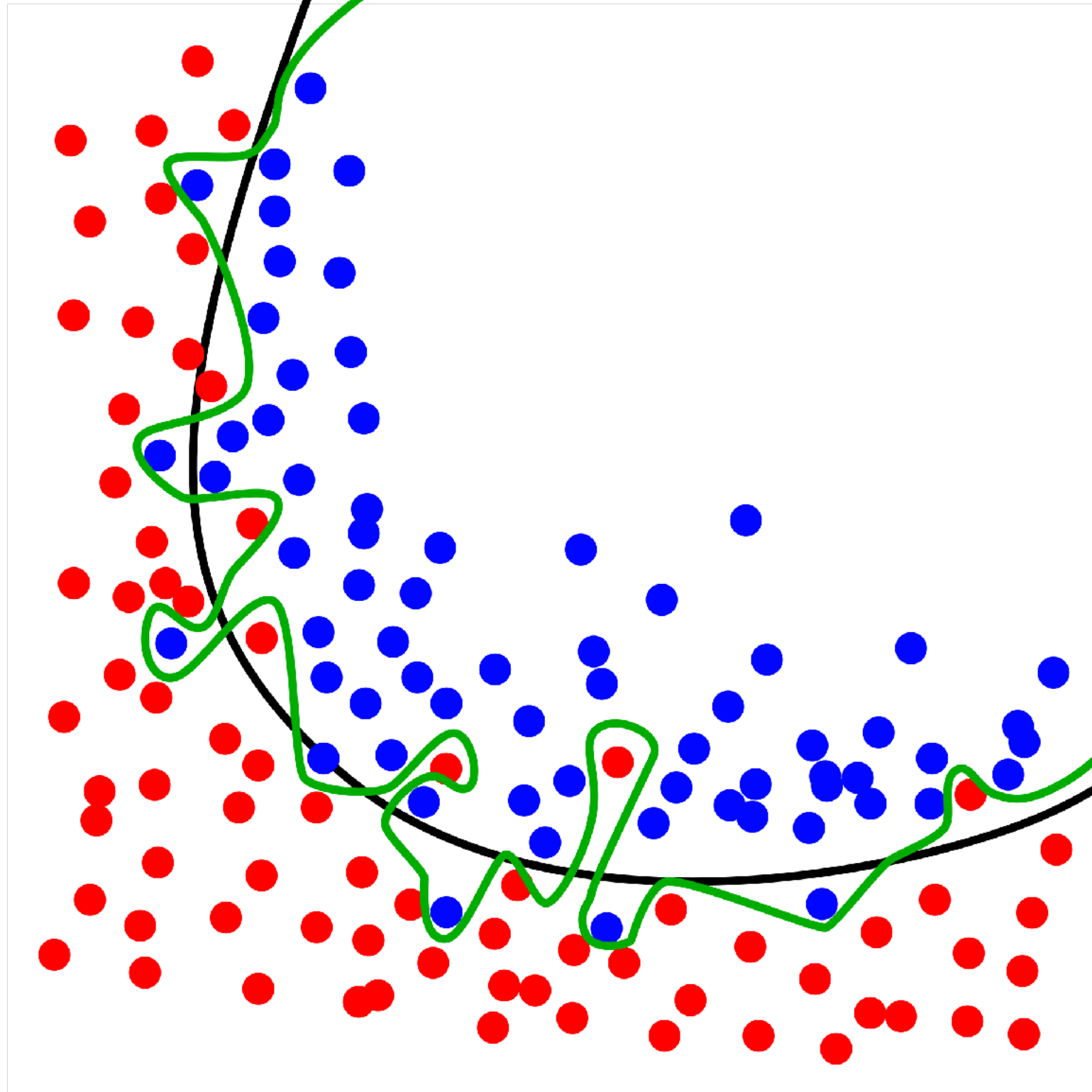
Overfitting detection

Conditional probability

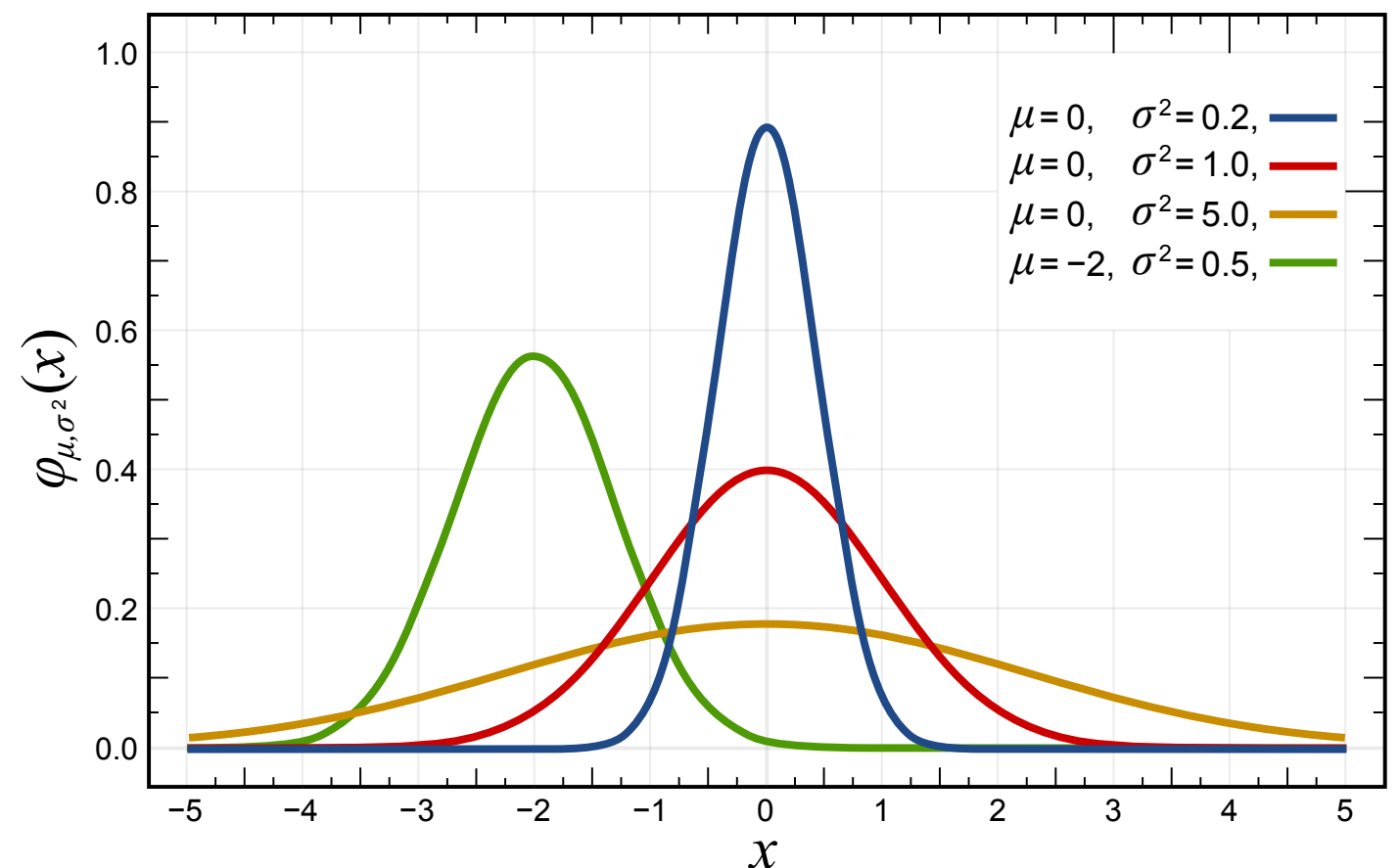$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A)P(B|A)}{P(B)}$$

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

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

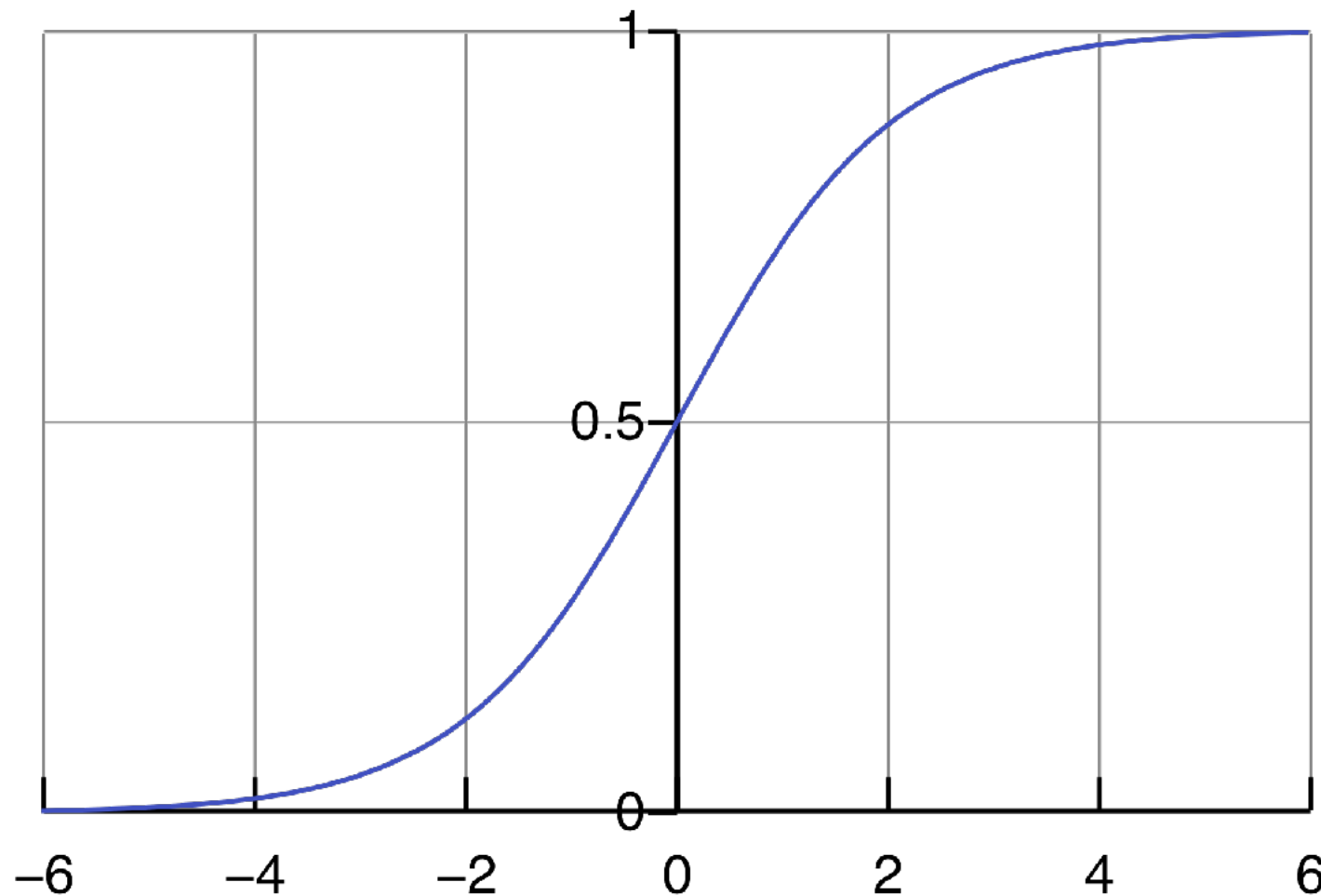$$p(x = v | C_k) = \frac{1}{\sqrt{2\pi\sigma_k^2}} e^{-\frac{(v - \mu_k)^2}{2\sigma_k^2}}$$

Scikit-learn tutorial
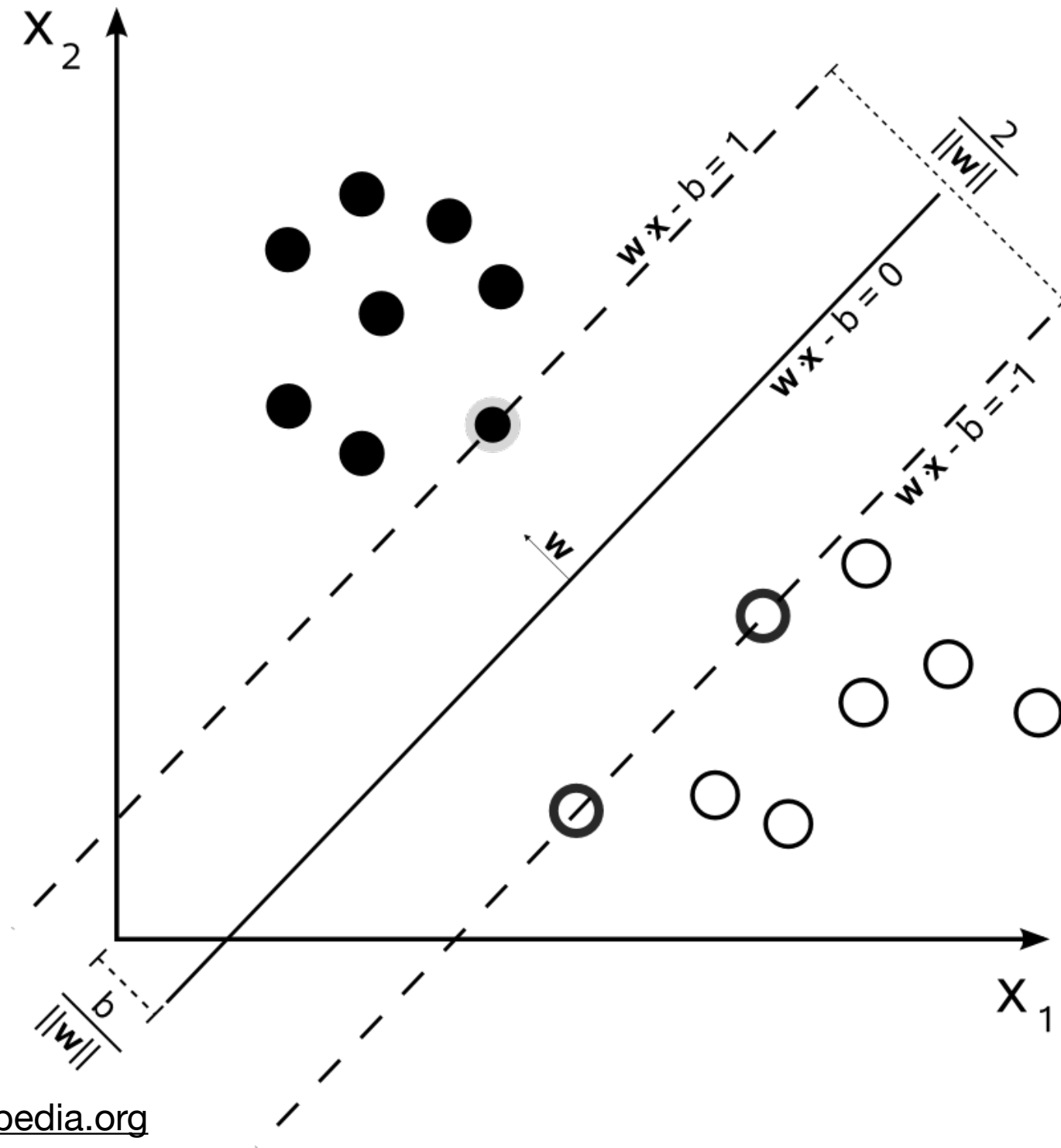
http://scikit-learn.org/stable/

01-Scikit-introduction.ipynb
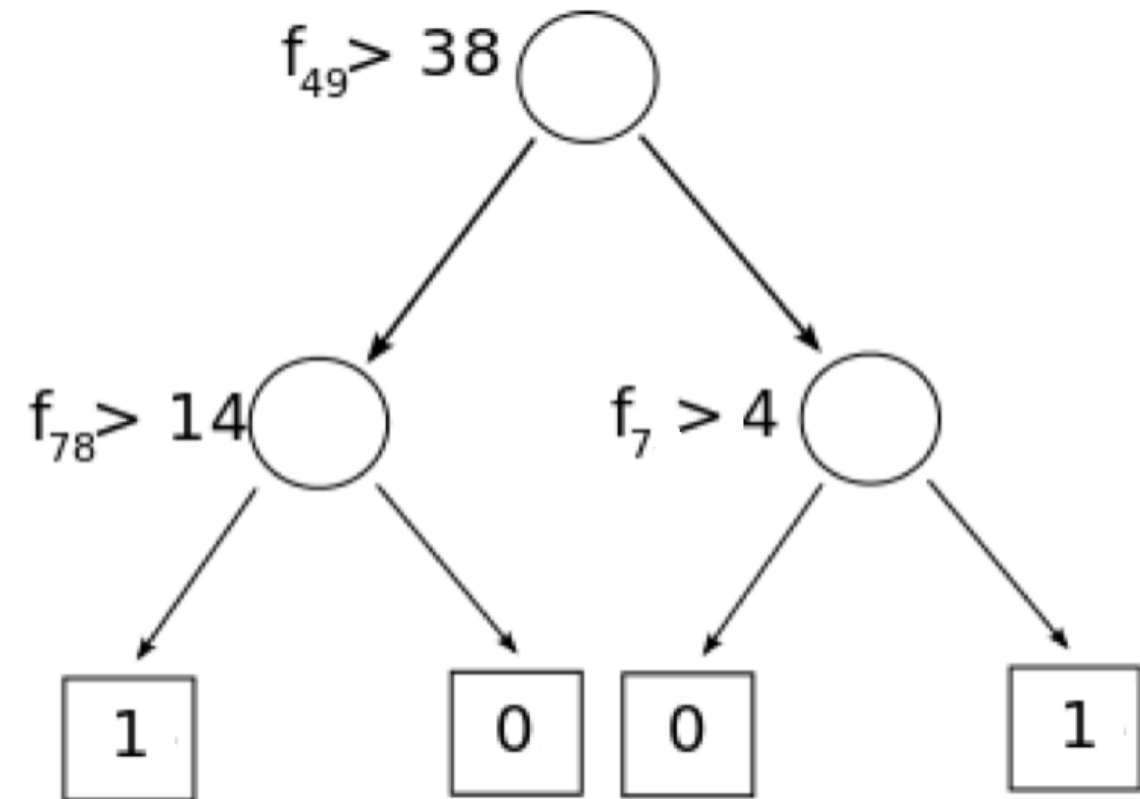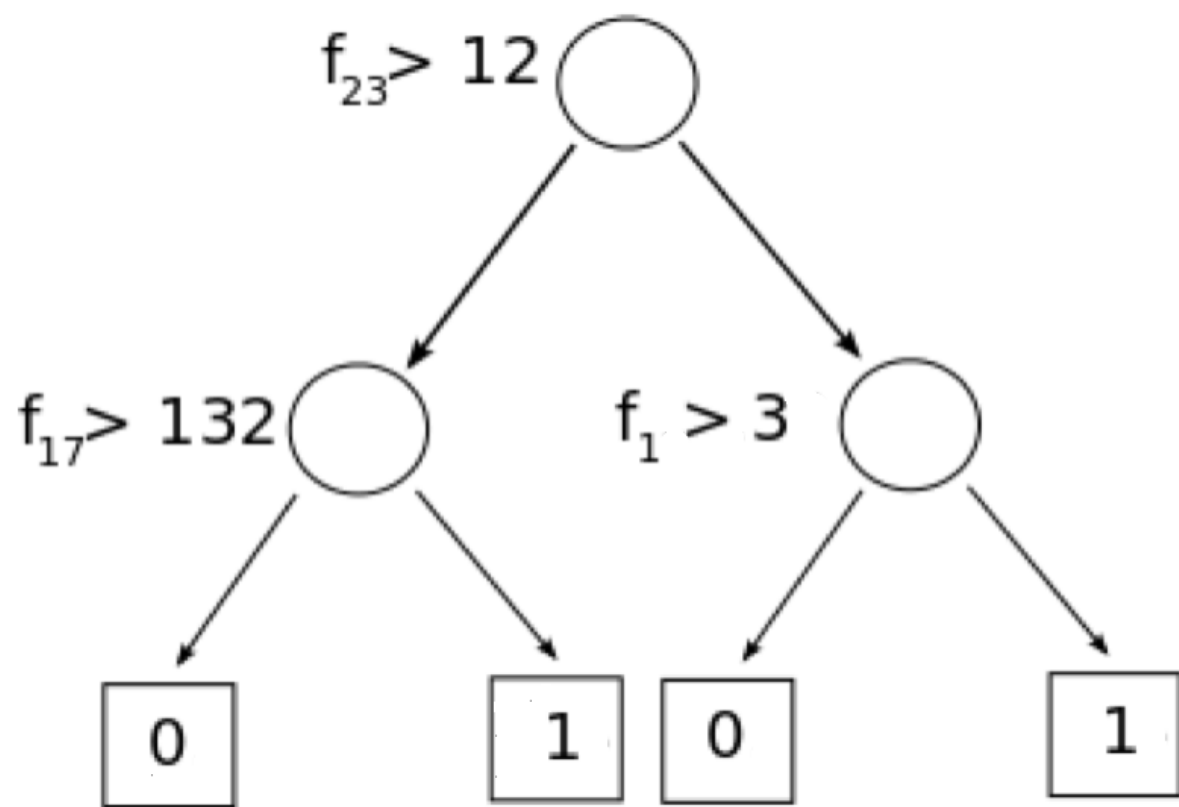
Logistic regression

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

Support Vector Machines
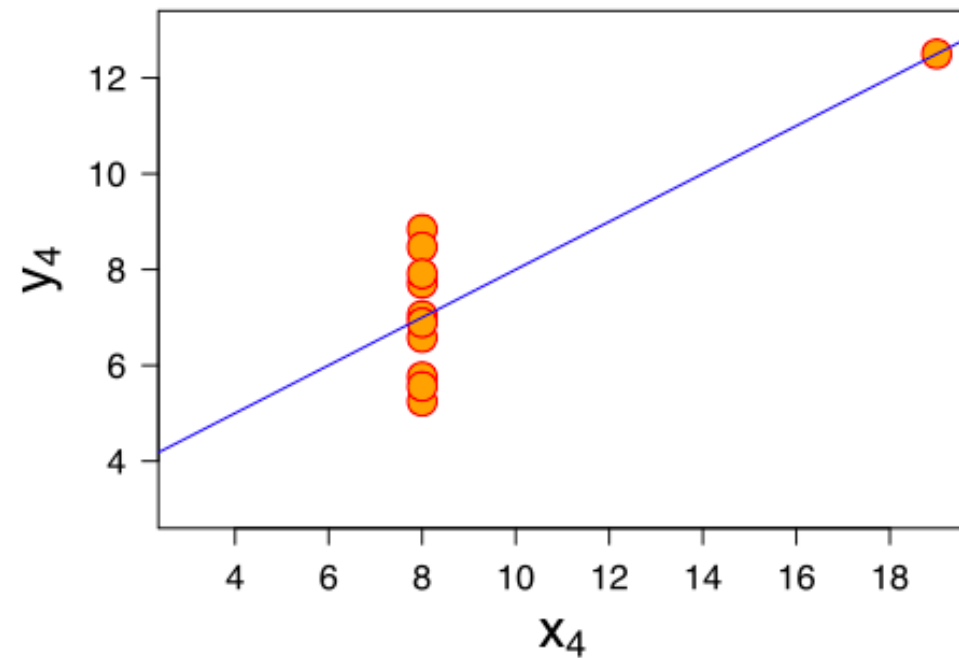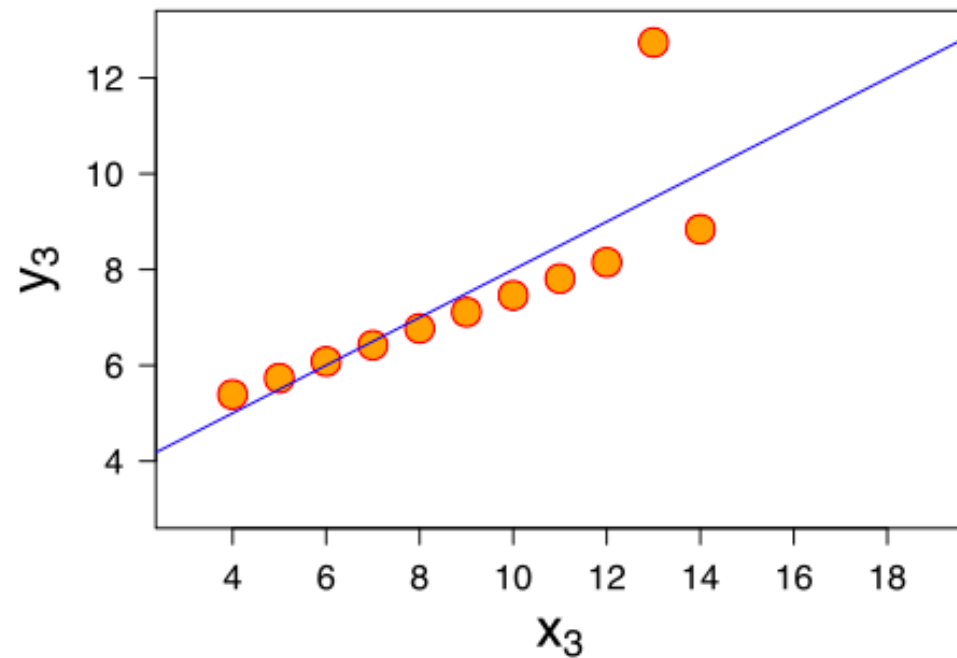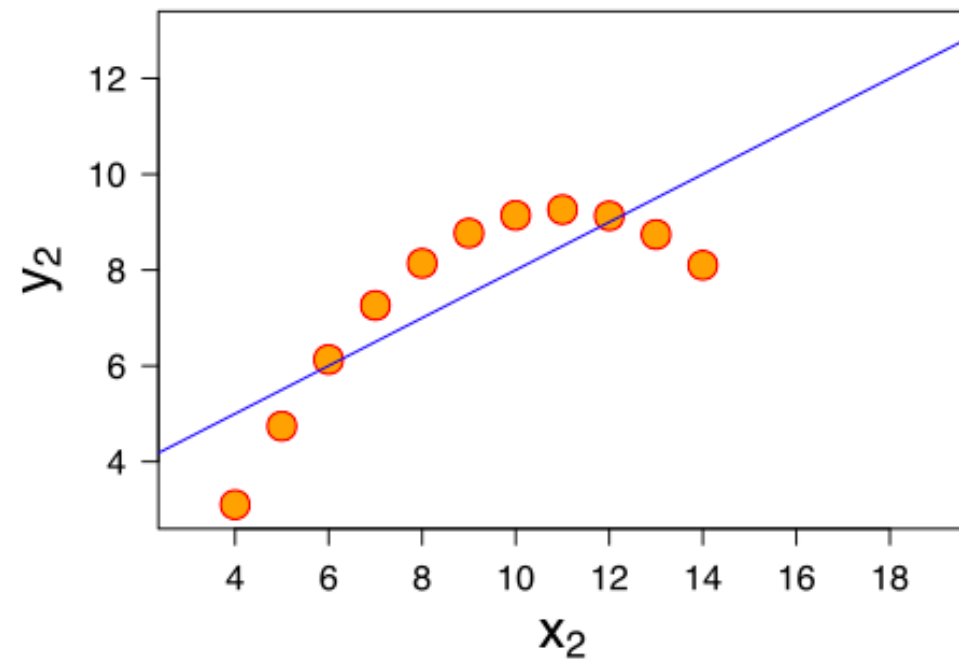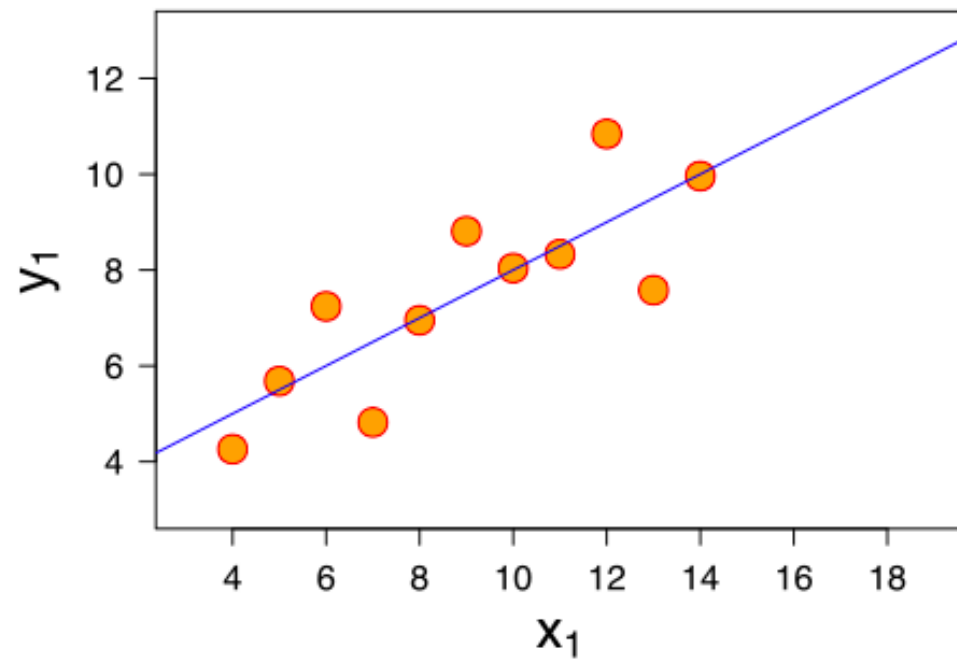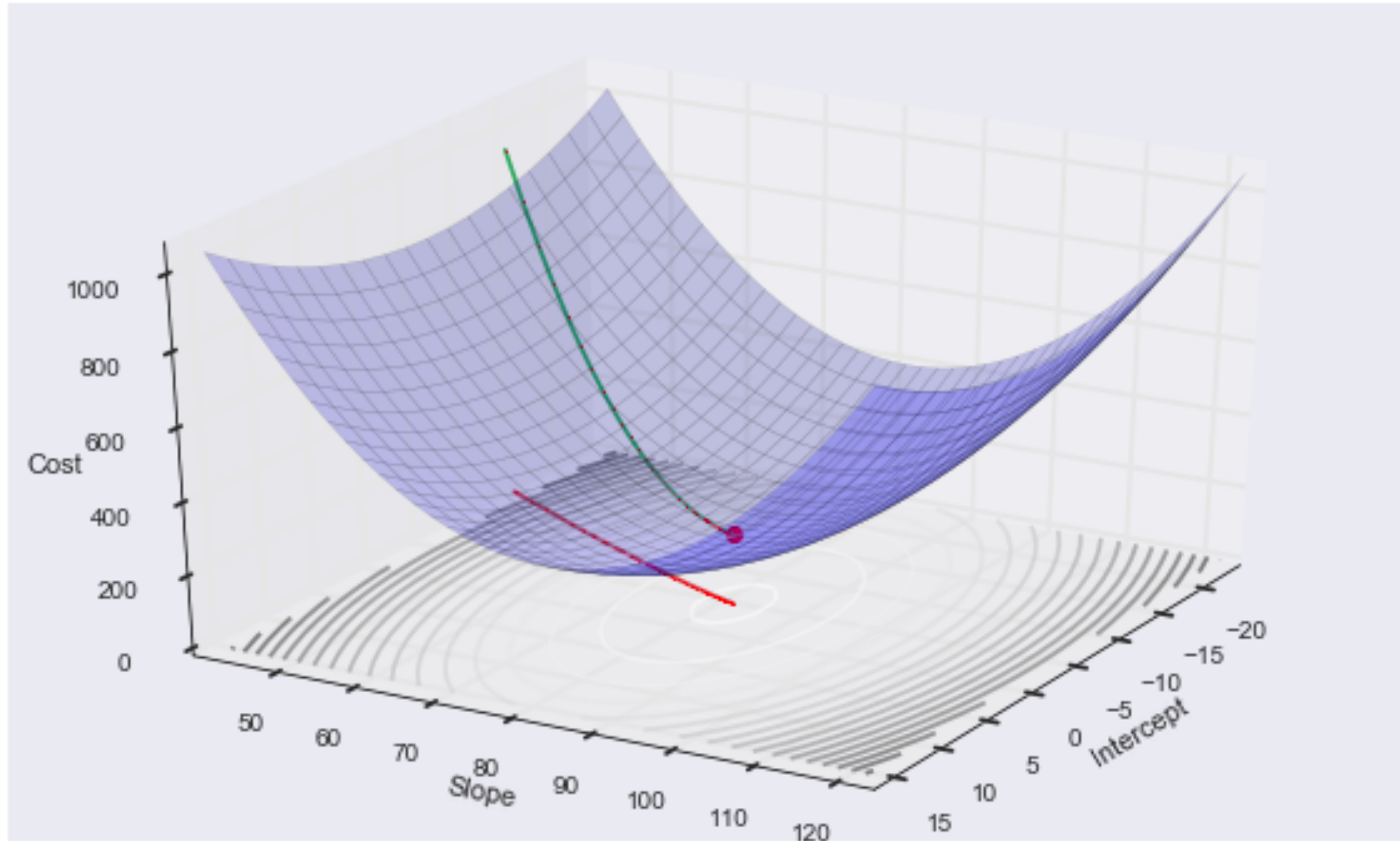
Boosted Decision Trees

Classification task

02-Classification1-assignment.ipynb

03-Classification2-assignment.ipynb

Regression

Linear regression with SGD

Support vector machines for Regression

Source: http://www.saedsayad.com/support_vector_machine_reg.htm
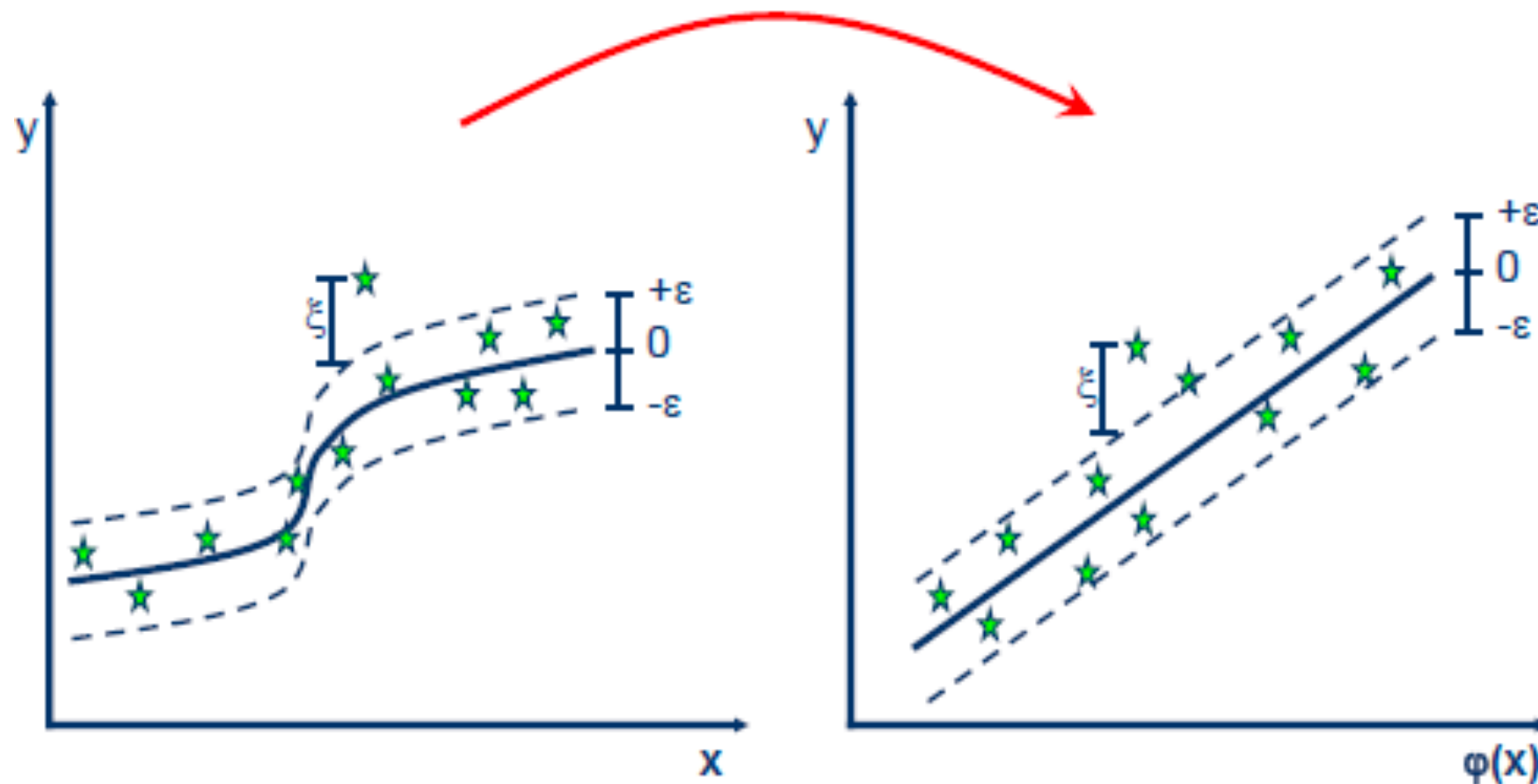
Boosted regression trees
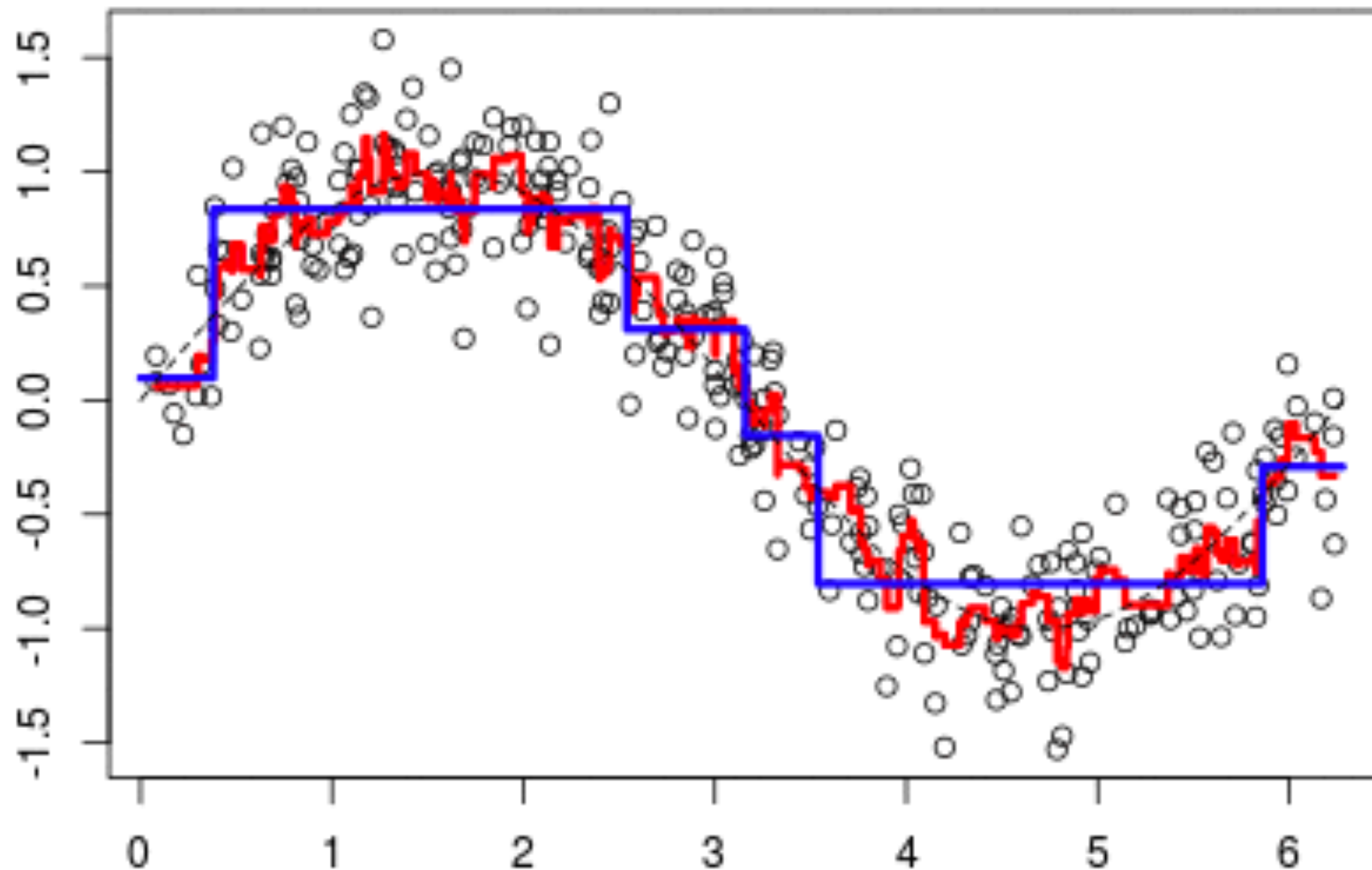
Evaluation of regression models

Root mean squared error

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

Mean absolute error

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

R Squared

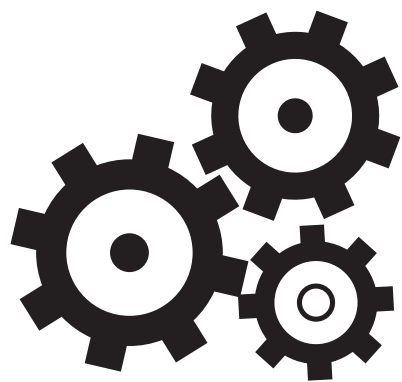$$\hat{R}^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \overline{y})^2}$$

Regression task

04-Regression1-assignment.ipynb

05-Regression2-assignment.ipynb

What next?

https://www.mlcollege.com/en/#courses

Machine Learning Prague

ML MU MACHINE LEARNING 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/