Neural networks

Doc. RNDr. Iveta Mrázová, CSc.

Department of Throretical Computer Science and Mathematical Logic

Faculty of Mathematics and Physics

Charles University in Prague

Neural networks

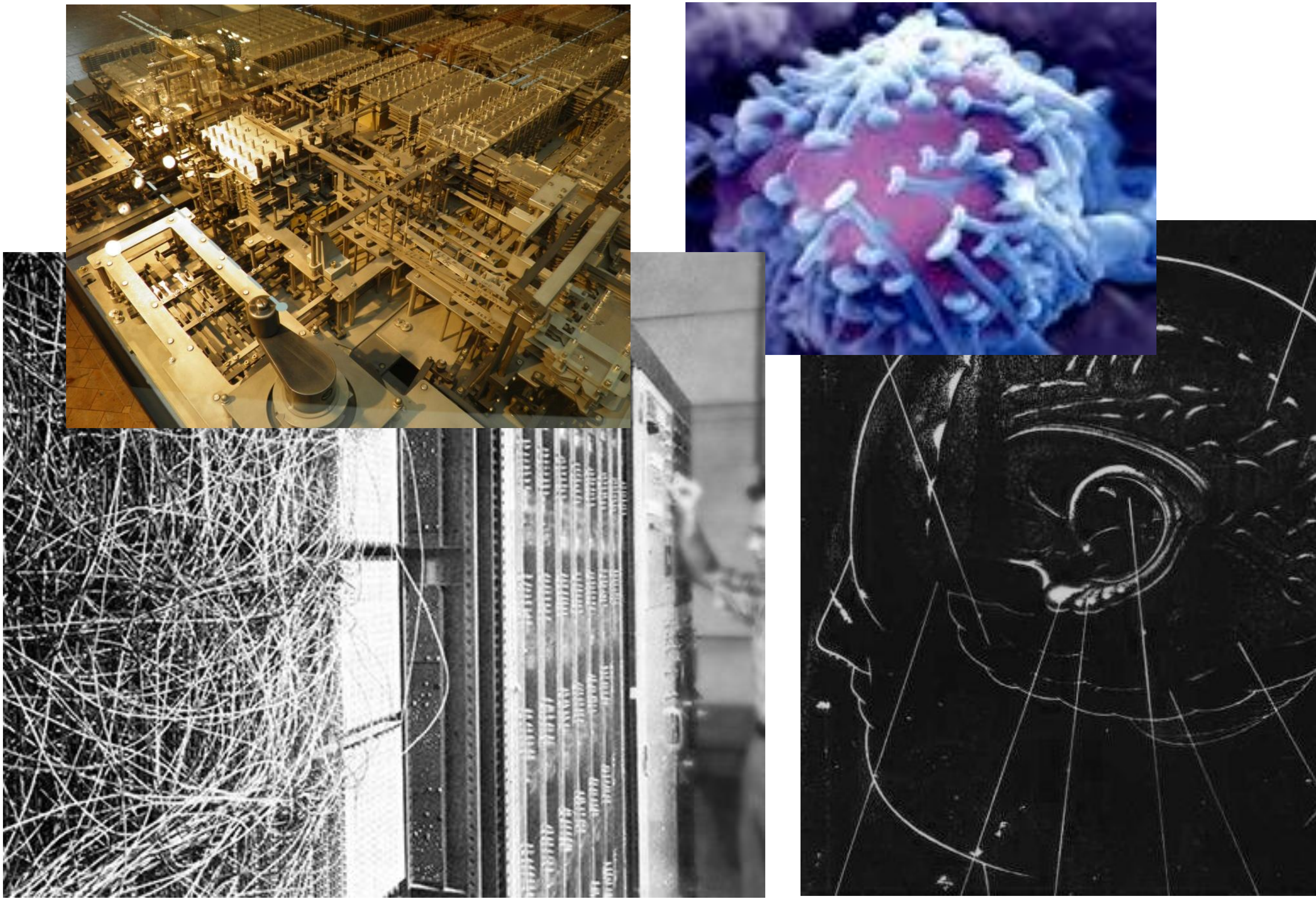
– Introduction into the area –

Doc. RNDr. Iveta Mrázová, CSc.

Katedra teoretické informatiky
Matematicko-fyzikální fakulta
Univerzity Karlovy v Praze

Computer versus brain

- The speed of information processing
- The kind of information processing
  - serial × parallel
- The kind of information storage
- Redundance
- Control



Neural networks – a brief history

- 1943 – formal neuron (W. McCulloch, W. Pitts)
- 1949 – mathematical notion of learning (D. Hebb)
- 1958 – perceptron (F. Rosenblatt)
- 1962 – Adaline and sigmoidal transfer function (B. Widrow, M. Hoff)
- 1969 – The perceptrons (M. Minsky, S. Papert)
- 1980s – a further development

Neural networks – a brief history

- since the eighties – further developments:
  - The back-propagation training algorithm (P. Werbos, D. Rumelhart, G. Hinton, Y. Le Cun)
  - Kohonen maps (T. Kohonen)
  - RBF-networks (Radial Basis Function, J. Moody, C. Darken)
  - GNG-model (Growing Neural Gas, B. Fritzke)
  - Convolutional neural networks (Y. Le Cun)
  - SVM-machines (Support Vector Machines, V. Vapnik)
  - ELM-networks (Extreme Learning Machines, G.-B. Huang)

Neural networks – 21st century

- 2003 - **Allen Brain Atlas** (Allen Institute for Brain Science, USA)
- **HBP – Human Brain Project,** EU (january 2013)

  Goal: mimic the human brain and identify poruchy its function

  Expected costs – 1.2 billions Euro /10 year

  https://www.humanbrainproject.eu/

  http://www.nature.com/news/brain-simulation-and-graphene-projects-win-billion-euro-competition-1.12291
- 2013 – **BigBrain** (Montreal Neurological Institute and German Forschungszentrum Jülich, June 2013)

  https://bigbrain.loris.ca/main.php

Neural networks – 21st century

- **BRAIN Initiative – Brain Research Through Advancing Innovative Neurotechnologies**, USA

„President Obama is calling on the science community to join him in pursuing a GRAND CHALLENGE BRAIN Initiative“, 2. 4. 2013,
http://www.whitehouse.gov/infographics/brain-initiative

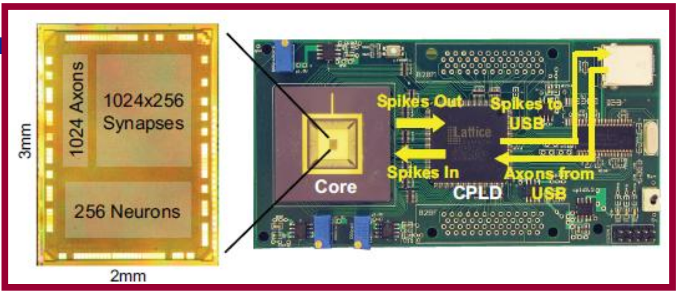
Goal: understand, how we think, how we learn and how works our memory

Expected costs – 3 billions USD / 10 years

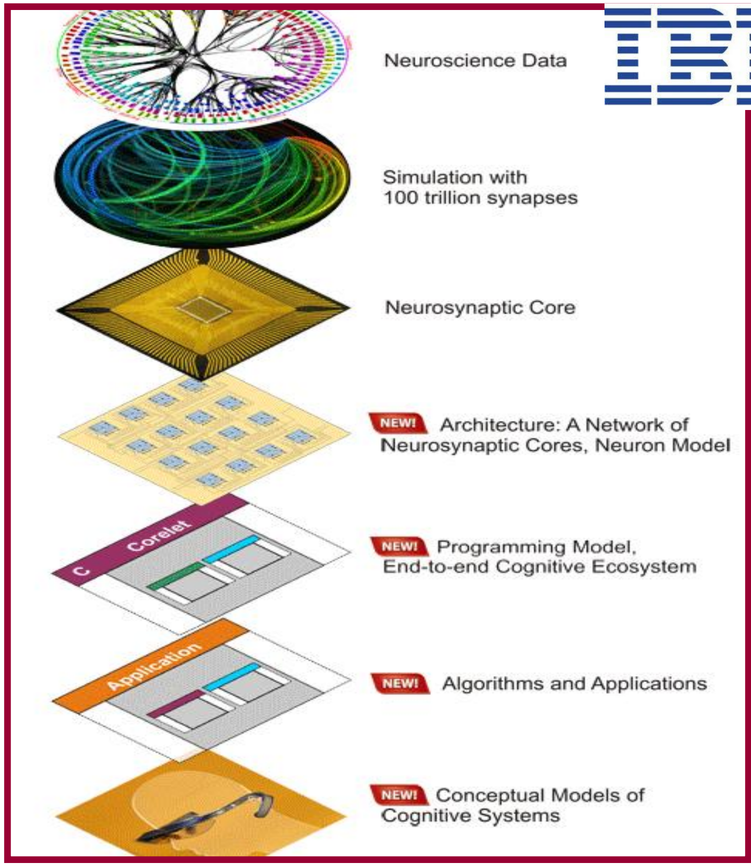
Participants: DARPA ~ Defense Advanced Research Projects Agency
NIH ~ National Institutes of Health
NSF ~ National Science Foudation
private sector

http://www.nih.gov/science/brain/how.htm
http://www.nature.com/news/flashing-fish-brains-filmed-in-action- 1.12621

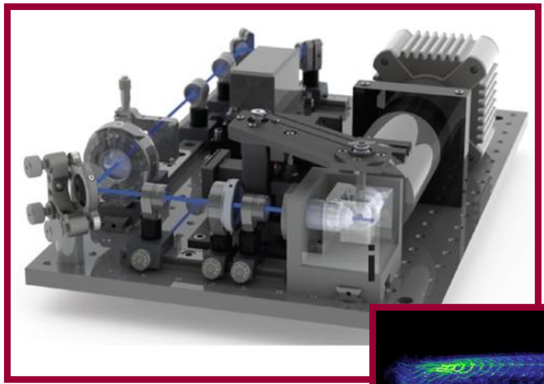
New technologies

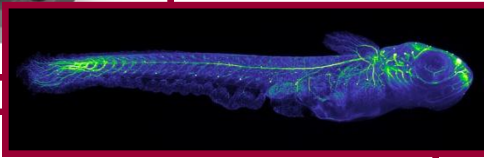


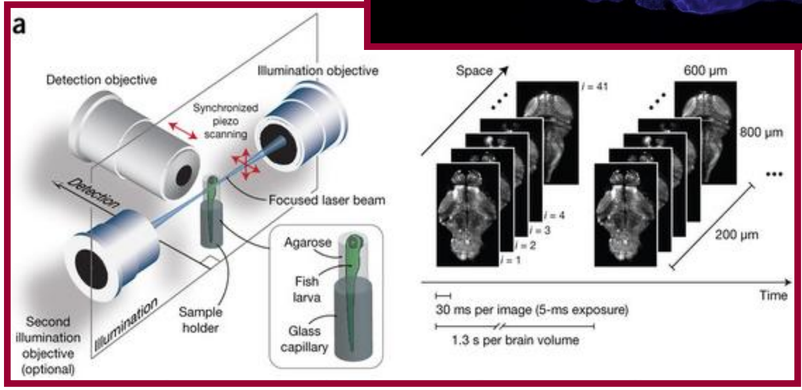
- Neurosynaptic chip



- Lightsheet microscopy







http://www.nature.com/nmeth/journal/v10/n5/fig_tab/nmeth.2434_SV4.html

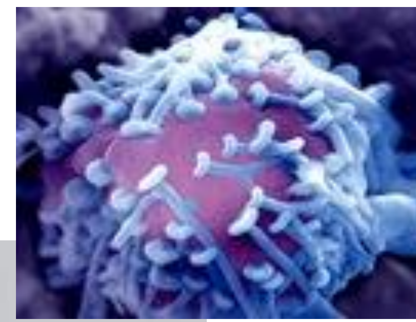
Neural networks – a general introduction

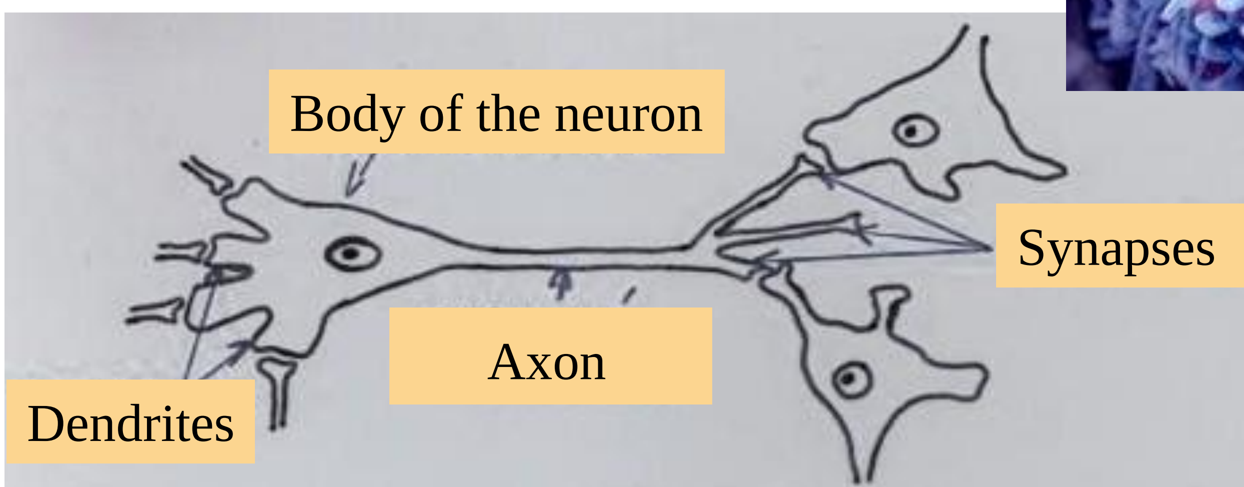
- **Recent problems:**
  - Training strategies – paralellization and efficiency
    - „This Will Revolutionize Education“ (http://youtu.be/GEmuEWjHr5c)
  - Architecture – generalization and robustness
  - Convergence and over-training
  - Prediction
- **Applications:**
  - Data mining – „black-box“, „white-box“
  - Clustering and classification
  - Information processing – speech, vision, olfactory, tactile, motoric
  - Data compression
  - Solutions of optimization tasks
  - and many others ….

Biological background (1)

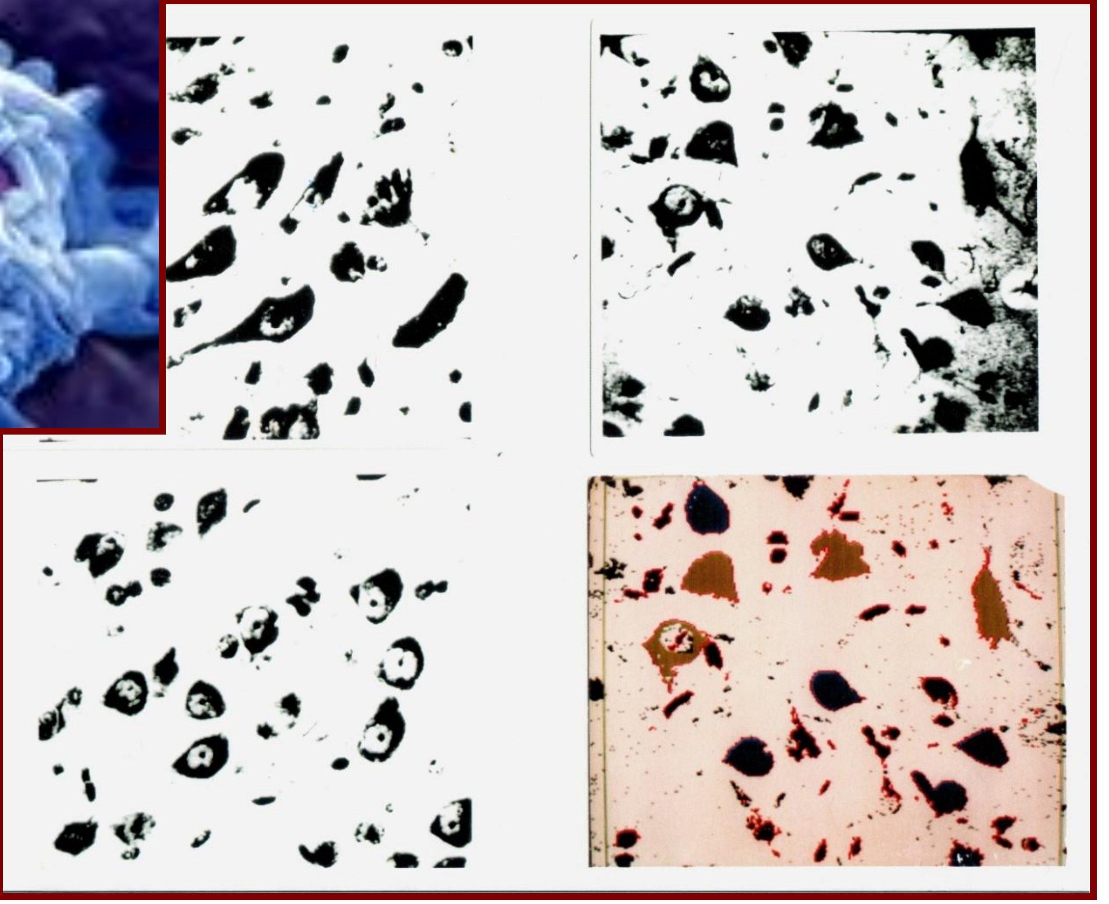
- **Model of a neuron**
  - basic „computational unit“ of a more complex system – neural network (contains cca 80 billion neurons)
  - biological neurons consist of:

    body (soma), dendrits, axon and synapses





Biological background (1)
biological neuron



Biological background (2)

- **Body (soma):**
  - summarizes signals transmitted by surrounding neurons $\rightarrow$ **potential**
  - inner potential leads to the **excitation** of the neuron
  - the size of neuron body varies from several μm to several tens of μm
- **Dendrites:**
  - represent signal input to neuron body
  - their length varies around 2-3 mm

Biological background (3)

- **Axon:**
  - the only output of a neuron, branched out widely, however, at it end
  - transmits the signal given by the level of excitation to the synapses
  - its length can reach over 1 m
- **Synapse:**
  - represent the „output device“ of the neurons, can the signal amplify or diminish and transmit it to other neurons
  - for each neuron, there are up to 10^6 connections to other neurons
- **Neuron output:**
  - Depends on neuron inputs and their processing inside neuron body

Biological background (4)

Biological neural networks:

- **neurons are mutually interconnected into networks**
  - by means of axons, that are connected to dendrites of other neurons via synapses
- **density of the neurons:**
  - reaches cca $70 - 80 \cdot 10^3$ / mm^3 in the human brain
  - cca $10 \cdot 10^3$ neurons die every day without replacement
  - synapses are formed on the dendrites during the whole life

  → **new synapses are formed**, resp. **non-functioning synapses can be revived**

=> **L E A R N I N G**

Biological background (8)

Memory types

- **Short-term memory mechanism**
  - based on cyclical circulation of signals in neural networks
  - after cca 300 ciculations, fixation of the information starts in mid-term memory – this takes cca 30 s
- **Mid-term memory mechanism**
  - based on the changes of „neural weights“
  - the change of synaptic weight coefficients is caused by multiple actions of the same signal on the respective synapse

Biological background (9)

Memory types

- **Mid-term memory mechanism**
  - some information stored in mid-term memory moves to long-term memory while sleeping
  - information stays in mid-term memoryfor several hours or days

- **Long-term memory mechanism**
  - consists in copying the informations from mid-term memory to proteins inside the neurons – in particular in their nuclei
  - information stored in this way can remain in the organism for its entire life

Adaptation and learning

Adaptation:

- ability to accommodate to the changes of the environment

Adaptive process: the process of the adjustment

- every adaptation represents for the system some costs (material, energy, …)
- living organisms are capable of reducing these costs during multiply repeated adaptations to environment changes

L E A R N I N G:

- minimalization of costs spent for adaptation
- result of a multiply repeated adaptation

Adaptation and learning: the formalism (1)

- **Manifestation of the environment:** $\boldsymbol{x}$
- **Feature description of the objects:**
  - selection of n basic characteristics – features $x_1, \ldots, x_n$
  - $\boldsymbol{x} = (x_1, \ldots, x_n)$
- **Information about the desired system reaction to the manifested environment:** Ω
- The system reacts to any manifestation of the environment $\boldsymbol{x}$ and information Ω by yielding one of the symbols $\omega_r \,;\, r = 1, \ldots, R$ at its output

Adaptation and learning: the formalism (2)

- Every assignment $[\mathbf{x}, \Omega] \rightarrow \omega_r$ is accompanied by some costs given by the function $Q(\mathbf{x}, \Omega, \omega_r)$ for each time unit
- **The goal of the system:**
  - find for any $\mathbf{x}$ and Ω such an assignment

    $$[\mathbf{x}, \Omega] \rightarrow \omega_r\,,$$

    for which the **cost is minimal:**

    $$Q(x, \Omega, \omega_r) = \min_{\omega} Q(x, \Omega, \omega)$$

Adaptive systems (1)

Adaptive system

~ a system with two inputs and one output determined by:

1) a set X **of manifestations of the environment** $\boldsymbol{x}$
2) a set O_1 **of informations about the desired system reaction** Ω
3) a set O_2 **of output symbols** ω
4) a set D **of decision rules** $\omega = d(\mathbf{x}, q)$
5) **the cost** $Q(\mathbf{x}, \Omega, q)$

For any pair $[\mathbf{x}, \Omega]$ we seek for such a parametr q^* ,
for which it holds: $\boldsymbol{Q(x, \Omega, q^*) = \min_q Q(x, \Omega, q)}$

Adaptive systems (2)

- Initial assignment $[\mathbf{x}, \Omega] \rightarrow \omega_s$
- If the system stays for time T in its initial assignment, this will be associated with total costs corresponding to $T\,Q(\mathbf{x}, \Omega, \omega_s)$
- If the system is able to change its behavior based on an ongoing cost assessment, it finds **after a certain time τ necessary for evaluation** ω_r, for which the cost is minimal

Adaptive systems (3)

Total costs after time *T* :

$$\tau\, Q(\mathbf{x}, \Omega, \omega_s) + (T - \tau)\, Q(\mathbf{x}, \Omega, \omega_r)$$

- bigger than the least possible total costs $T\, Q(\mathbf{x}, \Omega, \omega_r)$
- smaller than the total costs of a system, that cannot change its decision, $T\, Q(\mathbf{x}, \Omega, \omega_s)$

$$T\, Q(x, \Omega, \omega_r) < \tau\, Q(x, \Omega, \omega_s) + (T - \tau)\, Q(x, \Omega, \omega_r) < $$
$$< T\, Q(x, \Omega, \omega_s)$$

Learning systems (1)

The result of adaptation is stored in the memory:

- Save the time τ necessary to find minimum costs for repeated manifestations of the environment
- Further, it is not necessary to evaluate the costs

  → after training, the information Ω about the desired system reaction is not necessary anymore

Total costs of a learning system after training:

$$T\,Q(\mathbf{x}, \Omega, \omega_r)$$

- smaller than total costs of an adaptive system

Learning systems (2)

Learning system

~ a system with two inputs and one output determined by:

1) a set X **of manifestations of the environment** $\boldsymbol{x}$
2) a set O_1 **of informations about the desired system reaction** Ω
3) a set O_2 **of output symbols** ω
4) a set D **of decision rules** $\omega = d\,(\mathbf{x}, q)$
5) The **desired behavior** $\Omega = T(x)$
6) **Mean costs** $J(q)$ evaluated over $\mathrm{X} \times O_1$

Learning systems (3)

Learning system

- Finds after presenting the pair elements from the sequence

  $\{ [x_k, \Omega_k] \}; 1 \leq k \leq \infty$, where $\Omega = T_k (x_k)$,

  such a parametr q^*, for which it holds:

  $$J(q^*) = \min_q J(q)$$

- **Sequential** ~ sequential presentation of the pairs $[x_k, \Omega_k]$
- **Inductive** ~ find after the evaluation of countably many pairs $[x_k, \Omega_k]$ the parametr q^*, that minimizes the mean costs over the entire set X

Efficiency of adaptation and learning

Efficiency of an adaptive system is the higher, the shorter is the time τ necessary for its adaptation and the longer are the time intervals T when the environment does not change:

- $\tau / T \rightarrow 0$:

  Efficiency of the AS is comparable with the efficiency of a learning system after training

- $\tau / T \rightarrow 1 \quad (\tau / T < 1)$:

  AS has about the same efficiency like a non-adaptive system

- $\tau / T \geq 1$: no adaptation takes place

Efficiency of the (trained) learning system is the highest possible

Selection and order of features

Probability of a wrong decision

$\times$

Information contained in the input patterns

- **Too many features:**
  - technical feasibility
  - speed of processing
  - danger of over-training
    - the number of variables $\times$ the number of training patterns
  - correlated features

Selection of informative features

- **Selection of the minimum number of features** from the considered set of features
  - the chosen set is not guaranteed to contain really informative features
  - the choice depends on the actual task solved
- **The order of features** from the considered set of features
  - according to the amount of information contained
  - can be used, e.g., in the case of sequential classifiers

Karhunen-Loeve transform (1)

Properties of the Karhunen-Loeve transform:

1. For the given number of expansion members it yields the **least mean squared error** between the original and the transformed patterns
2. After the application of the covariance matrix the approximated patterns are decorrelated

   → **decorrelation of features**

Karhunen-Loeve transform (2)

3. Expansion members **do not contribute equally to the approximation**
   - The influence of each respective expansion member on the approximation accuracy falls with its index

     → The impact of members with high indexes will be small and we can thus omit them
4. **The magnitude of the approximation error does not influence the structure of the expansion**
   - Changed demands on the approximation error do not require the recomputation of the entire expansion

     → It is sufficient to add or remove a few of the last members

Of advantage especially for **sequential classification metods**

Karhunen-Loeve transform (3)

- The choice of a suitable mapping $V: X^m \rightarrow X^p$ such that the patterns from X^p will represent the best approximation of the original patterns from X^m in the sense of the mean squared error

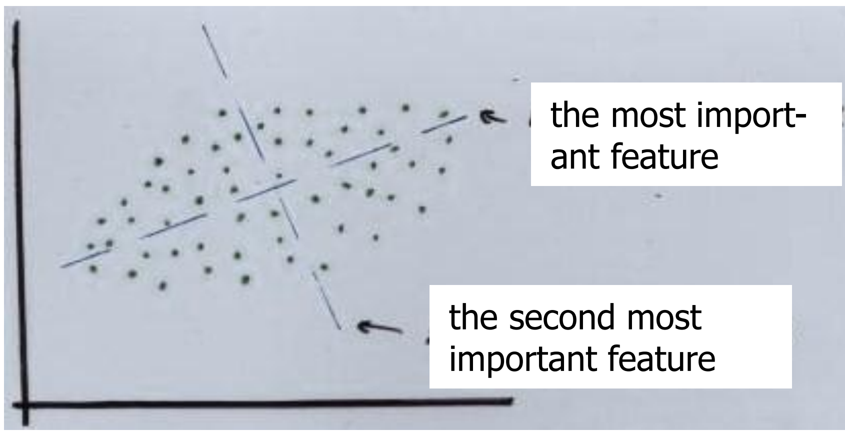
  K patterns from the same class
  m features
  p orthonormal vectors $\mathbf{e}_i$ $(1 \le i \le p)$ in X^m $(p \le m)$

$\rightarrow$ Approximate the vectors $\mathbf{x}_k$ from X^m $(1 \le k \le K)$ by a linear combination of $\mathbf{e}_i$:

$$\mathbf{y}_k = \sum_{i=1}^{p} c_{ki}\mathbf{e}_i$$

such that the squared error between $\mathbf{x}_k$ and $\mathbf{y}_k$, $\varepsilon_k^2 = \left\|\mathbf{x}_k - \mathbf{y}_k\right\|^2$, will be minimal

Karhunen-Loeve transform (4)



From m measured features, we want to get the p most important features$(1 \leq p << m)$

Matrix $\mathbf{V}$: $m \times p$

$$\mathbf{V} = \begin{pmatrix} v_{11} & \cdots & v_{1p} \\ \vdots & \ddots & \vdots \\ v_{m1} & \cdots & v_{mp} \end{pmatrix}$$

$$\mathbf{v} = (v_1, v_2, \ldots)^T ,$$

$$\mathbf{x} = (x_1, x_2, \ldots)^T$$

$$\mathbf{y} = \mathbf{v}^T \mathbf{x} = v_1 x_1 + v_2 x_2 + \ldots$$

Compute the vector *p* of the most important features:

$$\mathbf{y} = \mathbf{V}^T \mathbf{x}$$

Karhunen-Loeve transform (5)

Computation of the matrix V:

- Center the data:
$$\mu_j = \frac{1}{K}\sum_{k=1}^{K} x_{kj}$$
- covariance matrix for the training set:
$$w_{ij} = w_{ji} = \frac{1}{K}\sum_{k=1}^{K}\left(x_{ki}-\mu_i\right)\left(x_{kj}-\mu_j\right)$$
- The vectors defining the most important features correspond to the eigenvectors of the covariance matrix

Karhunen-Loeve transform (6)

- The eigenvalues correspond to the variance of the most important features
  - the first column of the matrix $\boldsymbol{V}$ will be the eigenvector corresponding to the biggest eigenvalue,…
  - further columns of $\boldsymbol{V}$ will be added until the following eigenvalues are too small and can be omitted

Problem:

- The choice of an adequate number of eigenvalues (p)
- An optimal choice of p cannot be guaranteed as the expansion does not reflect the true importance of each respective feature

Karhunen-Loeve transform (7)

Modifications:

1. **Centered most important features**

$$\mathbf{y} = \mathbf{V}^T (\mathbf{x} - \boldsymbol{\mu}),$$

where $\boldsymbol{\mu} = (\mu_1, \ldots)$ is the vector of mean values

2. **Normalized most important features**

$$\mathbf{y} = \mathbf{L}^{-1/2} \mathbf{V}^T (\mathbf{x} - \boldsymbol{\mu}),$$

where $\mathbf{L}$ is the matrix $p \times p$, diagonal elements are the eigenvalues corresponding to the columns of $\mathbf{V}$, the other elements are zero

Probability – basic notions (1)

Probability (of an event *A* from the space *S*):

- $P(A) \geq 0$ ($P(\{\}) = 0$)
- $P(S) = 1$
- For a finite number of mutually exclusive events $A_1, A_2, \ldots, A_n$ the probability

$$P(A_1 \cup A_2 \cup \ldots \cup A_n) = \sum_{i=1}^{n} P(A_i)$$

- For an infinite number of mutually exclusive events $A_1, A_2, \ldots, A_n$ the probability

$$P(A_1 \cup A_2 \cup \ldots) = \sum_{i=1}^{\infty} P(A_i)$$

Probability – basic notions (2)

- **Conditional probability** of the event ***B*** given that the event ***A*** has occurred (***P(A) > 0***):

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

- **Mutual independence** of the events ***A*** and ***B*:**

$$\boldsymbol{P(A \cap B) = P(A) \cdot P(B)}$$

- **Formula for the probability** of ***A***:

$$P(A) = \sum_i P(A \mid B_i) P(B_i)$$

Probability – basic notions (3)

Bayesian formula for the conditional probability:

$$P(B \mid A) = \frac{P(A \mid B)\, P(B)}{P(A)} \; ; \qquad P(A), P(B) > 0$$

- **Random variable:**
  - ´the name of an experiment with a probabilistic outcome´
  - its value is the outcome of the experiment
- **Probability distribution (for the random variable Y):**
  - Probability $P\,(Y = y_i)$, that Y will také on the value y_i
- **Expected value (~mean) of a random variable Y :**

$$\mu_Y = E\,(Y) = \sum_i y_i\, P\,(Y = y_i)$$

Probability – basic notions (4)

- **Variance (of a random variable):**

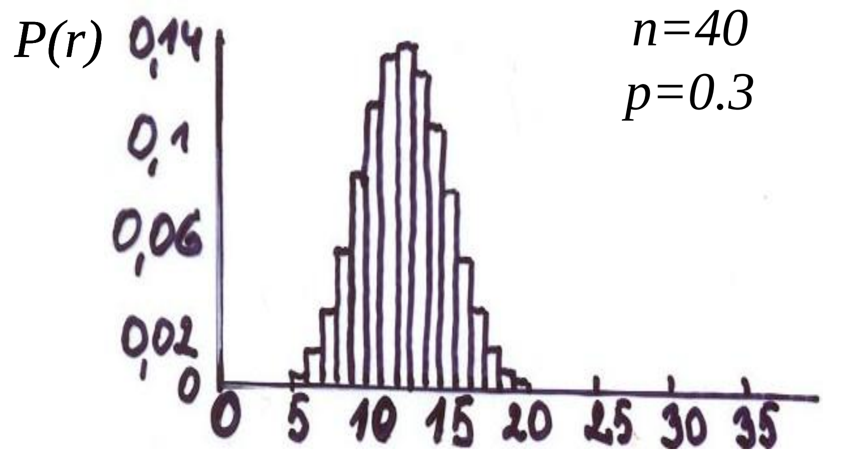
$$VAR(Y) = E\left\lfloor (Y - \mu_Y)^2 \right\rfloor$$

  - Characterizes the width (dispersion) of the distribution about its mean
- **Standard deviation of *Y*:** $\sigma_Y = \sqrt{VAR(Y)}$
- **Binomial distribution**
  - The probability of observing ***r*** ´heads´ in a series of ***n*** independent coin tosses
  - The probability of ´heads´ in a single toss is ***p***

Probability – basic notions (5)

Binomial distribution

- The probability of observing ***r*** ´heads´ in a series of ***n*** independent coin tosses
- The probability of ´heads´ in a single toss is ***p***



- **Probability function** (probability that ***X*** will take on the value ***r***)**:**

$$P(r) = \frac{n!}{r!(n-r)!} p^r (1-p)^{n-r}$$

Probability – basic notions (6)

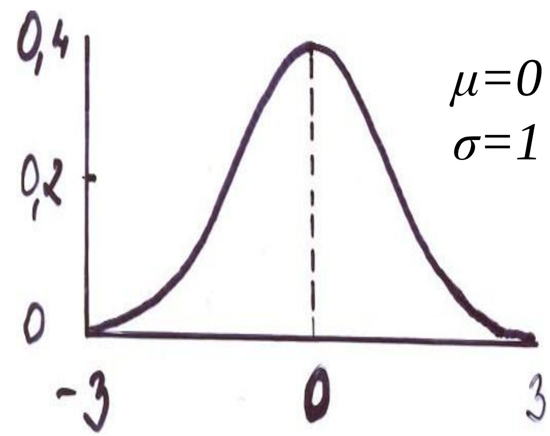
- Expected (mean) value of X: $E[X] = np$
- Variance: $VAR(X) = np(1-p)$
- Standard deviation: $\sigma_X = \sqrt{np(1-p)}$
- For sufficiently large values of n the binomial distribution is closely approximated by a normal distribution with the same mean and variance
- **Recommendation:** use the normal approximation only when: $np(1-p) \geq 5$

Probability – basic notions (7)

Normal distribution

- also called **Gaussian distribution**
- **Normal probability density function**

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$



$\mu=0$
$\sigma=1$

- Probability that the value of the random variable ***X*** will fall into the interval ***(a, b)*** :

$$\int_a^b p(x)\,dx$$

Probability – basic notions (8)

Normal distribution

- Suitable for a large number of natural phenomena
- Expected (mean) value of X: $E[X] = \mu$
- Variance: $VAR(X) = \sigma^2$
- Standard deviation: $\sigma_X = \sigma$

- **Central limit theorem:**

  ´The distribution of the mean of a large number of independent random variables of the same distribution approximates the normal distribution´

Probability – basic notions (9)

- **Estimator** ~ random variable Y
  - Used to estimate the parameter p from the tested population
- **Estimation bias** of Y for p : $E[Y] - p$
  - ´unbiased´ estimator for p : $E[Y] = p$
- $N\%$ **confidence interval** for the parameter p
  - Interval that contains p with probability $N\%$
- **Test** ~ procedure deciding on the correctness of a statistical hypothesis H
  - **Significance level** α corresponds to the probability of rejecting the true hypothesis $\longrightarrow$ usually set as $\alpha = 0.05$

Hypotheses testing (1)

1. Given the observed accuracy of a hypothesis over a limited sample of data → how well does this estimate its accuracy over additional examples?
2. Given that one hypothesis outperforms another over some sample of data → how probable is it that this hypothesis is more accurate in general?
3. When data is limited → what is the best way to use this data to both learn a hypothesis and estimate its accuracy as well as to compare the performance of two learning algorithms?

→ limit the difference between the accuracy observed on the given data and the actual accuracy of the whole data distribution

Hypotheses testing (2)

Aim: 1) Understand whether to use the hypothesis or not

2) Evaluating hypotheses represents an integral component of many learning methods (e.g., when post-pruning decision trees to avoid overfitting)

Estimate future accuracy of a hypothesis given only a limited set of data:

- **Bias in the estimate**: over-training × unbiased estimate of future accuracy (mutually independent training and test sets)
- **Variance in the estimate**: the measured accuracy can vary from the true accuracy; bigger variance for fewer test examples

Hypotheses testing (3)

Estimating hypothesis accuracy

- Space of possible instances ***X***, e.g., the set of all people
- Various target functions may be defined over ***X*** **,**
  $f: X \rightarrow \{0,1\}$, e.g., people who plan to purchase new skis this year
- Different instances $x \in X$ may be encountered with different frequencies, e.g., probability that ***x*** arrives at the ski resort
  - ***D*** … probability of encountering the instances in ***X***

Hypotheses testing (4)

Task: learn the target function f from the space H of possible hypotheses

- provided are training examples $\boldsymbol{x}$, along with their correct target value $\boldsymbol{f(x)}$, drawn randomly from $\boldsymbol{X}$ according to the distribution $\boldsymbol{D}$

Questions:

- Given a hypothesis $\boldsymbol{h}$ and a data sample containing $\boldsymbol{n}$ examples drawn at random according to the distribution $\boldsymbol{D}$:

1. **What is the best estimate of the accuracy of $\boldsymbol{h}$ over future instances drawn from the same distribution?**
2. **What is the probable error in this accuracy estimate?**

Hypotheses testing (5)

The sample error on the training set $S \subset X$

~ the fraction of S , misclassified by h

$$ERROR_S(h) \equiv \frac{1}{n} \sum_{x \in S} \delta(f(x), h(x))$$

- n … the number of examples in S
- $\delta(f(x), h(x)) = 1$ pro $f(x) \neq h(x)$
- $\delta(f(x), h(x)) = 0$ pro $f(x) = h(x)$
- Binomial distribution $ERROR_S(h)$: $ERROR_S(h) = r / n$
  - r … the number of examples from S, that were misclassified by h

Hypotheses testing (6)

The true error of hypothesis *h*

~ probability of misclassification for an instance $x \in X$ drawn at random according to D

$$ERROR_D(h) \equiv \Pr_{x \in D}[f(x) \neq h(x)]$$

- Binomial distribution: $ERROR_D(h) = p$ (= r/n … estimate for p)
  - p … probability of misclassifying a single instance drawn from D
  - unbiased estimator $ERROR_D(h)$ (~ $p = r/n$)
    - The hypothesis h and the sample set S must be chosen independently.
    - The training set S contains n (≥ 30) examples drawn at random from X according to the probability distribution D

Hypotheses testing (7)

Estimator variance

- Unbiased estimator with the least variance would yield the smallest expected squared error between the estimate and the true value of the parameter
- Given no other information, the most probable value of ***$ERROR_D(h)$*** is ***$ERROR_S(h)$***
- With approximately ***95%*** probability, the true error ***$ERROR_D(h)$*** lies in the interval

$$ERROR_S(h) \pm 1.96\sqrt{\frac{ERROR_S(h)\left(1 - ERROR_S(h)\right)}{n}}$$

→ for approximately 95% of experiments, the calculated interval will contain the true error value

Hypotheses testing (8)

Expression for general (***N%***) confidence intervals – constant z_N:

$$ERROR_S(h) \pm z_N \sqrt{\frac{ERROR_S(h)\left(1 - ERROR_S(h)\right)}{n}}$$

	The values of z_N for two-sided $N\%$ confidence intervals						
N%	50%	68%	80%	90%	95%	98%	99%
z_N	0.67	1.00	1.28	1.64	1.96	2.33	2.58

- Wider intervals for a higher probability
- Good approximation for $n \geq 30$, resp.

$$n \cdot ERROR_S(h)\,(1 - ERROR_S(h)) \geq 5$$

Hypotheses testing (9)

General approach to derive the confidence intervals:

1. **Identify the underlying population parameter** p to be estimated, e.g., ($ERROR_D\,(h)$)
2. **Define the estimator** Y (e.g., $ERROR_S\,(h)$)

   – choose a minimum-variance, unbiased estimator
3. **Determine the probability distribution** D_Y that governs the estimator Y including its mean and variance
4. **Determine the** $N\%$ **confidence interval**

   – find the thresholds L and U such that $N\%$ of the mass in the probability distribution D_Y falls between L a U

Hypotheses testing (10)

Difference in error of two hypotheses:

- Discrete-valued target function
- Hypothesis h_1 has been tested on a sample S_1 containing n_1 randomly drawn examples
- Hypothesis h_2 has been tested on an independent sample S_2 containing n_2 examples drawn from the same distribution
- **We want to estimate the difference d between the true errors of these two hypotheses:**

$$d = ERROR_D(h_1) - ERROR_D(h_2)$$

Hypotheses testing (11)

→ Estimator $\hat{d}$ ~ difference between sample errors:

$$\hat{d} \equiv ERROR_{S_1}(h_1) - ERROR_{S_2}(h_2)$$

$\hat{d}$ yields an unbiased estimate of d

- Normal distribution with the mean $E\left[\hat{d}\right] = d$ and variance $\sigma_{\hat{d}}^2$

$$\sigma_{\hat{d}}^2 \approx \frac{ERROR_{S_1}(h_1)\,(1 - ERROR_{S_1}(h_1))}{n_1} + \frac{ERROR_{S_2}(h_2)\,(1 - ERROR_{S_2}(h_2))}{n_2}$$

- ***N%*** confidence interval:

$$\hat{d} \pm z_N \sqrt{\frac{ERROR_{S_1}(h_1)\,(1 - ERROR_{S_1}(h_1))}{n_1} + \frac{ERROR_{S_2}(h_2)\,(1 - ERROR_{S_2}(h_2))}{n_2}}$$

Hypotheses testing (12)

Comparing the learning algoritms:

- test for comparing the learning algoritms L_A a L_B
- statistical significance of the observed difference between the algorithms

$\rightarrow$ determine which of the learning methods, L_A and L_B, is better for learning the target function f

- Consider the relative performance of the two algorithms averaged over all the training sets of size n that might be drawn from the distribution D

Hypotheses testing (13)

Comparing learning algorithms:

→ **estimate the expected value of the difference in the errors**

$$\underset{S \subset D}{E}\left[ERROR_D\left(L_A(S)\right) - ERROR_D\left(L_B(S)\right)\right]$$

$L\,(\,S\,)$ … hypothesis obtained by the learning algorithm $\boldsymbol{L}$ on the training set $\boldsymbol{S}$

$S \subset D$… the expected value is taken over the samples S drawn according to the underlying instance distribution $\boldsymbol{D}$

→ **in practice, just a limited number of training data $\boldsymbol{D_0}$ is available to compare the considered learning algorithms**

Hypotheses testing (14)

- Divide the set D_0 into a training set S_0 and a disjoint test set množinu T_0
  - Training data are used to train both L_A and L_B
  - Test data are used to compare the accuracy of the two learned hypotheses:

$$ERROR_{T_0}\left(L_A\left(S_0\right)\right) - ERROR_{T_0}\left(L_B\left(S_0\right)\right)$$

  - $ERROR_{T_0}(h)$ approximates the true error $ERROR_D(h)$
  - The difference in errors is measured only for the training set S_0 (rather than taking the expected value of this difference over all samples S that might be drawn from the distribution D)

k-fold cross validation (1)

1. Partition the available data D_0 into k disjoint subsets T_1, $T_2, \dots, T_k$ of equal size (≥ 30).

2. **FOR i:=1 TO k DO**

   use T_i for the test set, and the remaining data to build the training set S_i

   $S_i \leftarrow \{ D_0 \setminus T_i \}$

   $h_A \leftarrow L_A(S_i)$

   $h_B \leftarrow L_B(S_i)$

   $\delta_i \leftarrow ERROR_{T_i}(h_A) - ERROR_{T_i}(h_B)$

3. Return the value $\bar{\delta}$, where: $\bar{\delta} \equiv \frac{1}{k} \sum_{i=1}^{k} \delta_i$

k-fold cross validation (2)

***N* %** - confidence interval: $\bar{\delta} \pm \boldsymbol{t}_{N,k=1}\, \boldsymbol{s}_{\bar{\delta}}$

$\boldsymbol{s}_{\bar{\delta}}$ … estimate of the standard deviation:

$$s_{\bar{\delta}} \equiv \sqrt{\frac{1}{k(k-1)} \sum_{i=1}^{k} \left(\delta_i - \bar{\delta}\right)^2}$$

$\boldsymbol{t}_{N,k\text{-}1}$ … constant (values of $\boldsymbol{t}_{N,\nu}$ for two-sided confidence intervals approach the values of $\boldsymbol{z}_N$ with $\boldsymbol{\nu} \rightarrow \infty$)

N …… the desired confidence level

$\boldsymbol{\nu}$ ……. Nr. of degrees of freedom (nr. of independent random events that influence the value of $\bar{\delta}$)

k-fold cross validation (3)

	Confidence level N			
	90%	95%	98%	99%
$\nu = 2$	2.92	4.30	6.96	9.92
$\nu = 5$	2.02	2.57	3.36	4.03
$\nu = 10$	1.81	2.23	2.76	3.17
$\nu = 20$	1.72	2.09	2.53	2.84
$\nu = 30$	1.70	2.04	2.46	2.75
$\nu = 120$	1.66	1.98	2.36	2.62
$\nu = \infty$	1.64	1.96	2.33	2.58

N … the desired confidence level

ν … Nr. of degrees of freedom

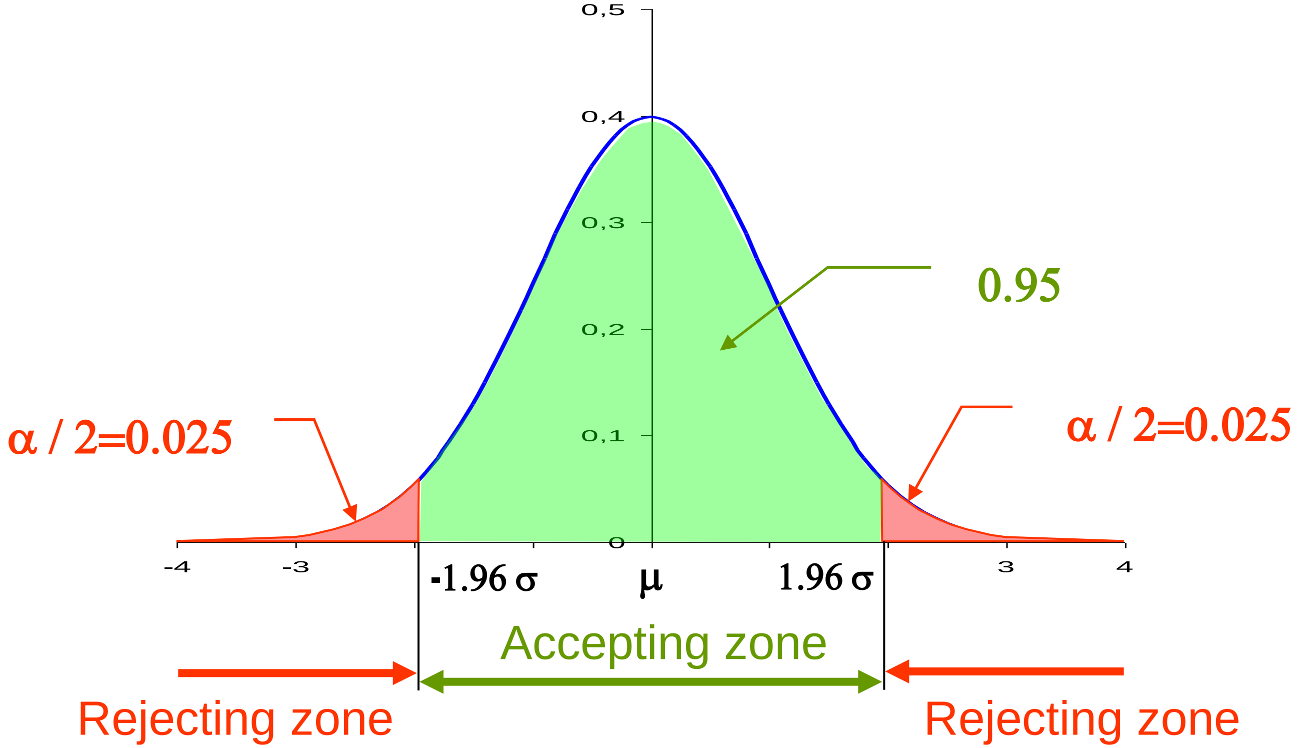
k-fold cross validation (4)

- **Testing has to be done on identical test sets!**
  - in contrast to comparing hypotheses that requires independent test sets

→ **Paired tests**

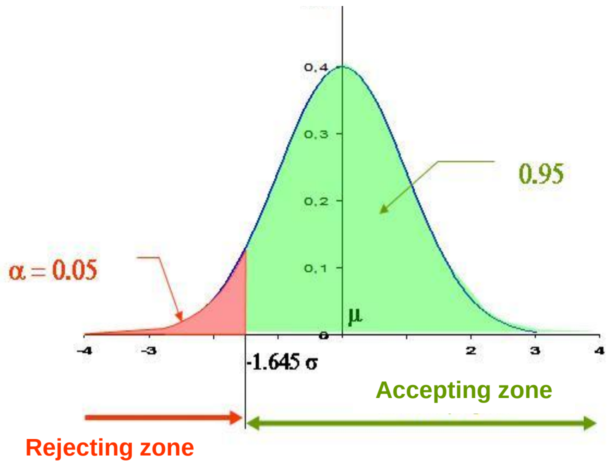
- typically produce tighter confidence intervals because any differences in observed errors are due to differences between the hypotheses and not due to differences in the makeup of the sampled data

A two-sided test



A one-sided × a two-sided test

A one-sided test



A two-sided test