# Learning, four substitutive quality criteria

Václav Hlaváč

Czech Technical University in Prague
Faculty of Electrical Engineering, Department of Cybernetics
Center for Machine Perception
http://cmp.felk.cvut.cz/~hlavac, hlavac@fel.cvut.cz

**Outline of the talk:**

- Importance of a fairy tail and a toy stage.
- Learning is treated differently in many disciplines.
- More formal analysis of learning in Pattern Recognition.

# Where does learning stand in the fairy tale parable?

- Some members of artificial intelligence community (including pattern recognition) have ascribed miraculous properties to learning.
- Unaware, 'fairy tale'-like attempts can be noticed even in current days.

---

- The main aim of this lecture is to discuss potential and limits of learning on the background of a 'fairy tale' parable.

# Three stages needed in scientific and technological development

Fairy tales. A miraculous instrument is usually sought that would allow us to perform what has been impossible until now *(e.g., to develop a flying carpet and float in the air)*.

Toys. Various models are created which imitate dreams of the fairy tale stage although they are too far from any practical exploitation, *(e.g., a model glider which is already flying)*.

Prototypes fulfil the practical requirements, a little at the beginning, and more and more later, *(e.g., an airplane)*.

- Thinking in a fairy tale manner is an effort to perceive the result demanded.
- Toys clear up the principles and check whether it is possible to realize this or that dream.

# Learning – a mixture of fairy tales, toys and prototypes

Certain knowledge is needed to create recognition strategies are needed, i.e., functions $q\colon X \to Y$.

Often, there is a dream about a miraculous tool as ‘Lay table, lay!’.

---

*“There is a system (a genetic, evolutionary, neural, or exotic in another way) which works in the following manner:*

- *The system learns first from the* training multi-set, *i.e., couples* $(x_i, y_i)$, $i = 1, \ldots, l$, *where* $x_i$ *is the observation and* $y_i$ *is a label of the class, which is considered correct.*
- *When learning finishes after* $l$ *steps, the normal exploitation stage of the system begins. The system reacts with* $y$ *to each observation* $x$, *and even to one which did not appear in the learning stage.*
- *Thanks to the information about the correct answer not having been provided explicitly, the system is able to solve any pattern recognition task.”*

# Learning: approaches in different disciplines

Engineering – signal processing, system identification, adaptive and optimal control, information theory, robotics, . . .

Computer science – artificial Intelligence, machine learning, computer vision, information retrieval, . . .

Statistics – learning theory, data mining, learning and inference from data, . . .

Cognitive science and psychology – perception, sensorimotor control, reinforcement learning, learning, mathematical psychology, computational linguistics, . . .

Computational neuroscience – neuronal networks, neural information processing, . . .

Economics – decision theory, game theory, operational research, . . .

*Pedagogy – different approach, they do not talk about statistics, . . .*

# Psychology
# the state of learning theories

Start of the 20th century, vast psychological systems were offered as explanations of learning (and of much wider ranges of behavior as well), such as behaviorism and Gestalt psychology.

1940s, comprehensive theories of learning were still believed to be reasonably near at hand. But during the next three decades it grew clear that such theories are tenable only for very limited sets of data.

Late 20th century, learning theory seemed to consist of a set of hypotheses of limited applicability.

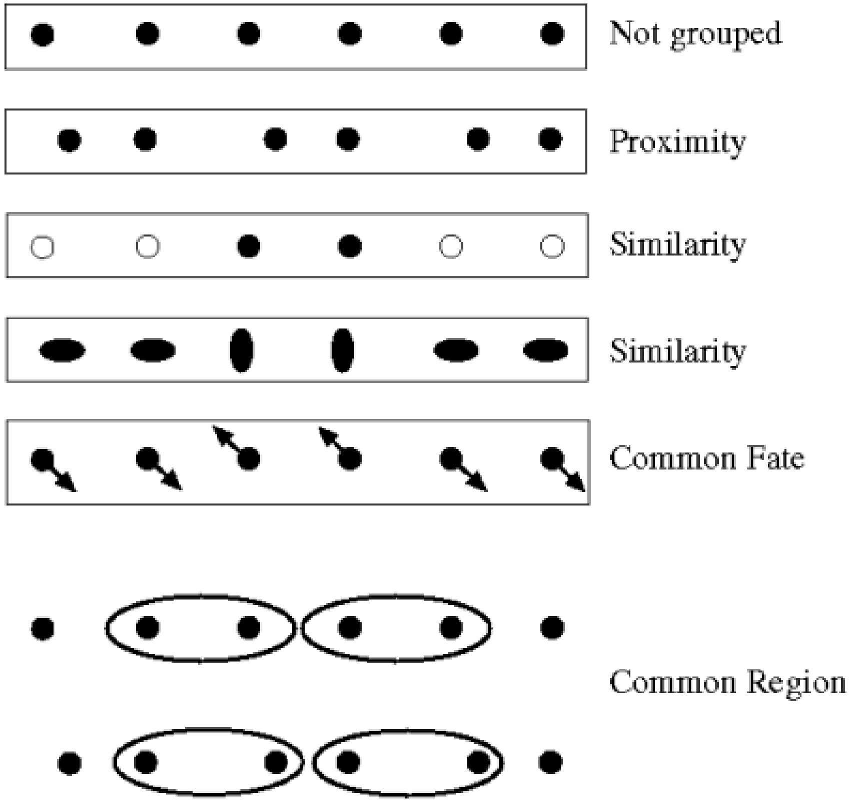
# Behaviorism

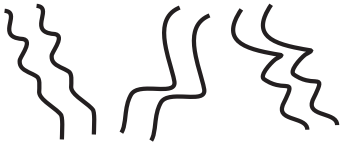
- Concerned itself exclusively with measurable and observable data.
- Ideas, emotions, and the consideration of inner mental experience and activity in general were excluded.
- The organism is seen as “responding” to conditions (stimuli) set by the outer environment and inner biological processes.
- The characteristic method of psychology was thus introspection – observing and reporting upon the working of one’s own mind.

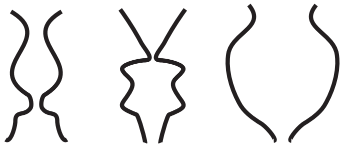
# Gestalt psychology

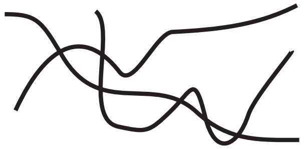
- Provided the foundation for the modern study of perception.
- Founding publication by Max Wertheimer (born in Prague) was issued in 1912 in Frankfurt a.M. It concerns stationary objects shown in a rapid succession appear to move (motion pictures).
- Gestalt theory was meant to have general applicability; its main tenets, however, were induced almost exclusively from observations on visual perception.
- Unlike the atomistic orientation of previous theories, emphasized that the whole of anything is greater than its parts. The attributes of the whole of anything are not deducible from analysis of the parts in isolation.
- "Gestalt" in German means the way a thing has been "placed" or "put together." There is no exact equivalent in English. "Configuration.", "form", "shape", "pattern" are the usual translations.

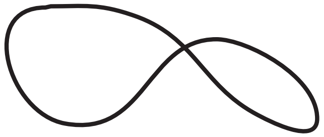
# Gestalt grouping principles



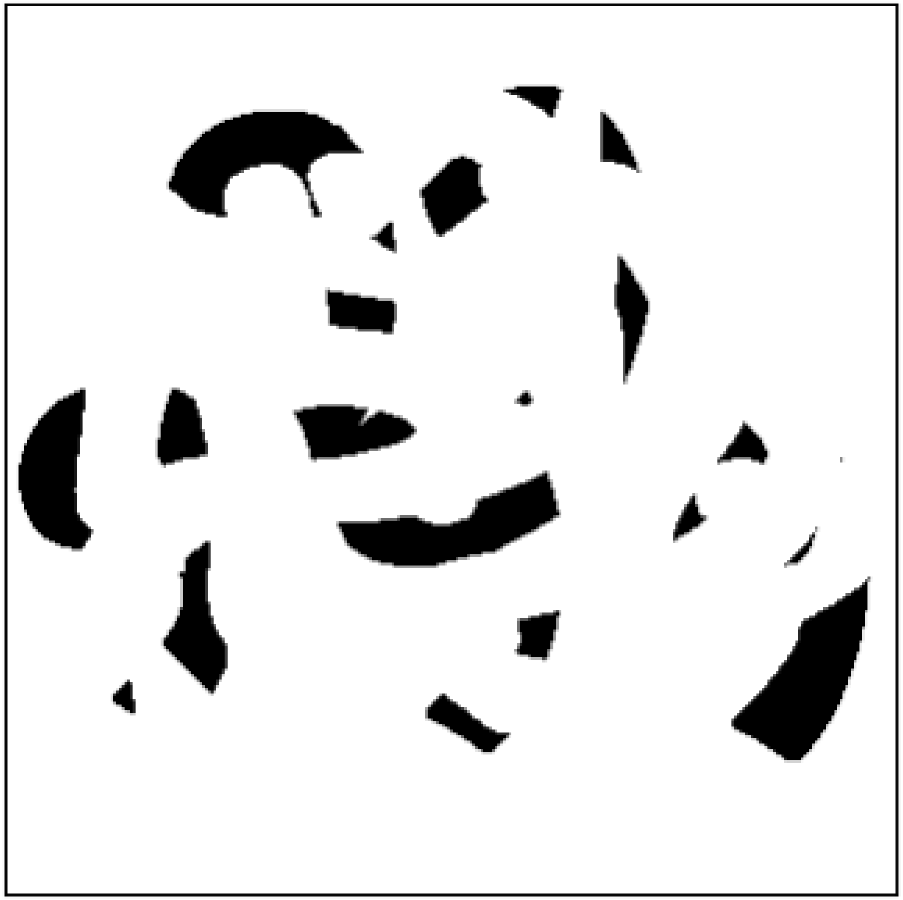

paralellism


symmetry

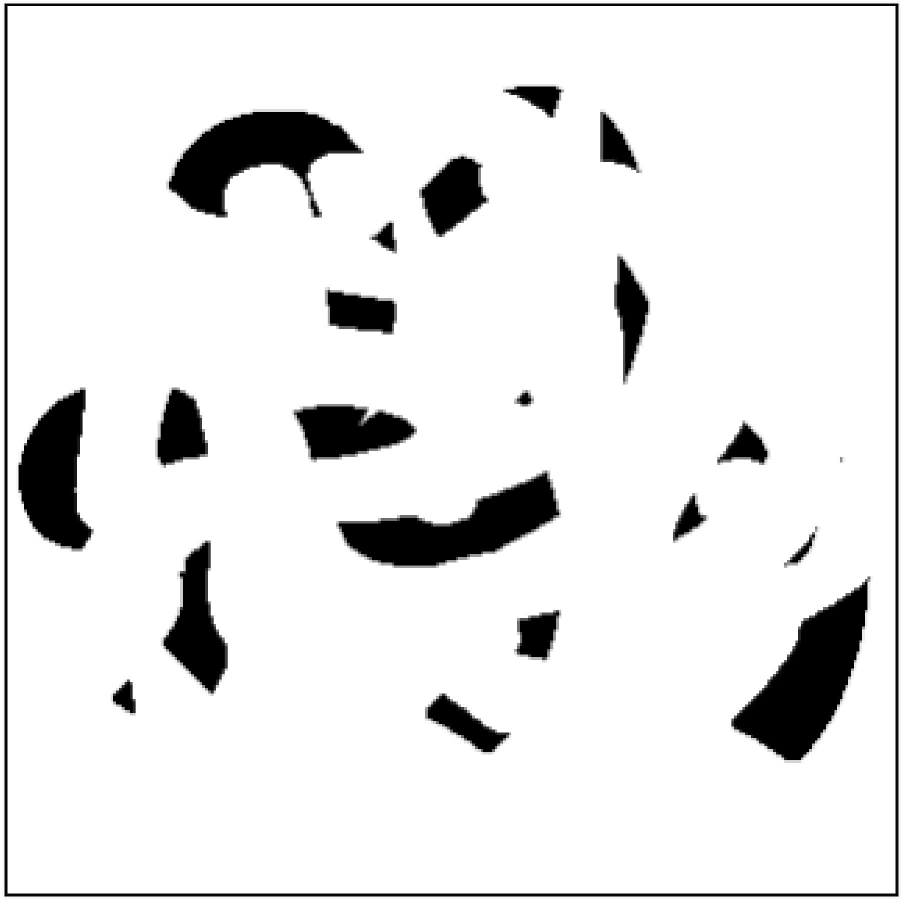

continuity

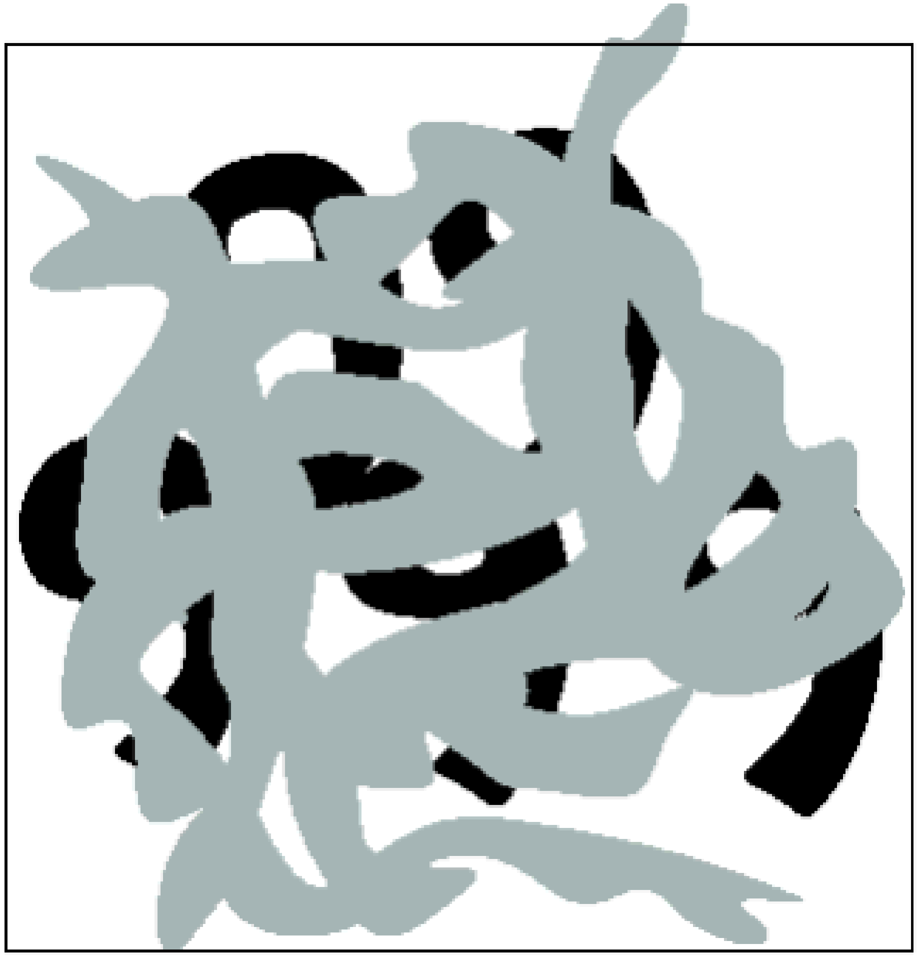

closure

# Humans are good in grouping (1)

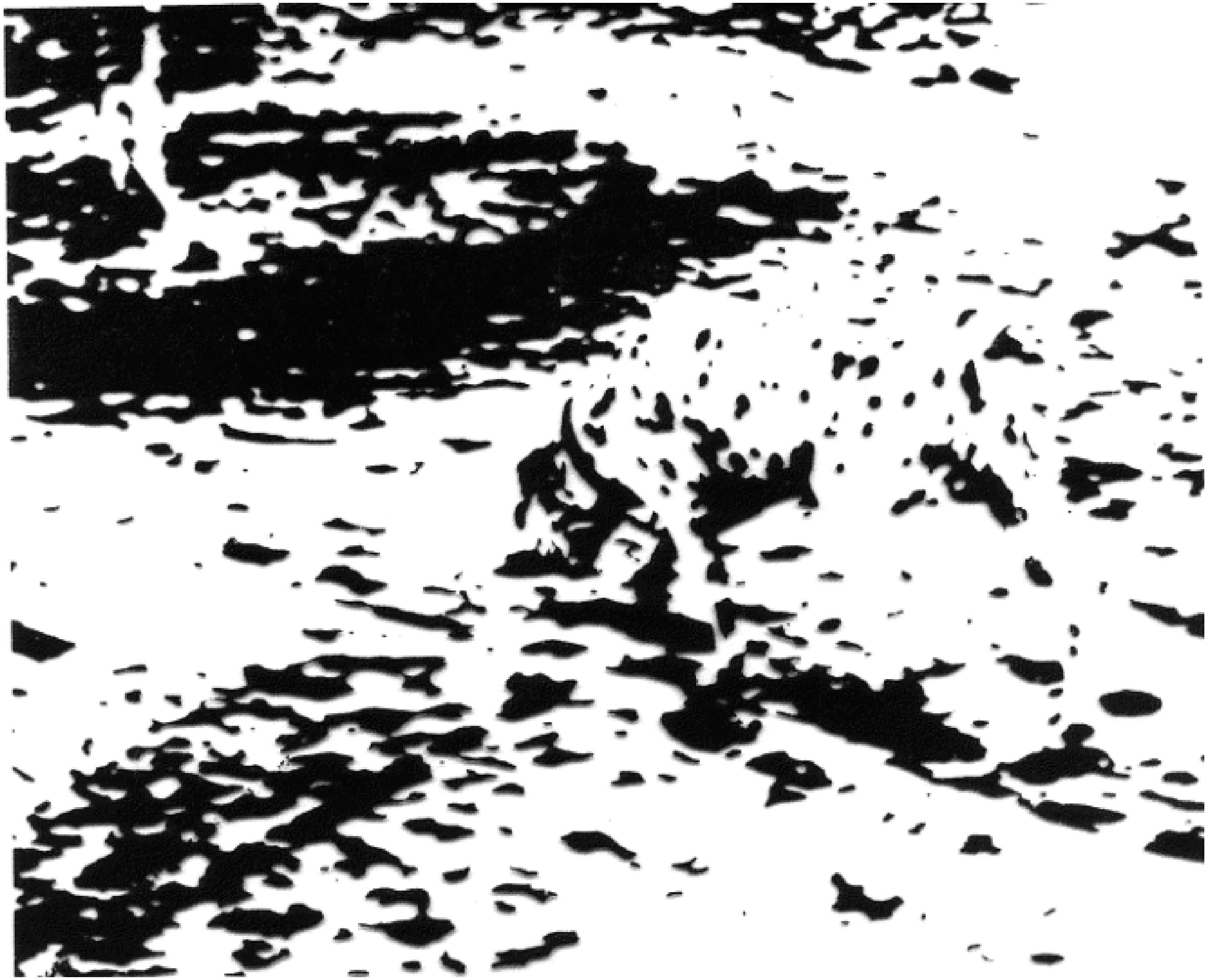


# Humans are good in grouping (2)





# Grouping is not always easy



# Knowledge engineering paradox

In learning by examples:

It is simpler to create good examples than to build general theories or explicit description of pattern or concepts ($\sim$ classes or hidden states in pattern recognition).

---

The aim is to find concepts (classes) description which is

- Complete, i.e., each positive example is satisfied.
- Consistent, i.e., no negative example is satisfied.

---

The training multi-set is finite $\Rightarrow$ concept description is only a hypothesis.

# Inference

Inference is a derivation of conclusions in logic from given information or premises by any acceptable form of reasoning.

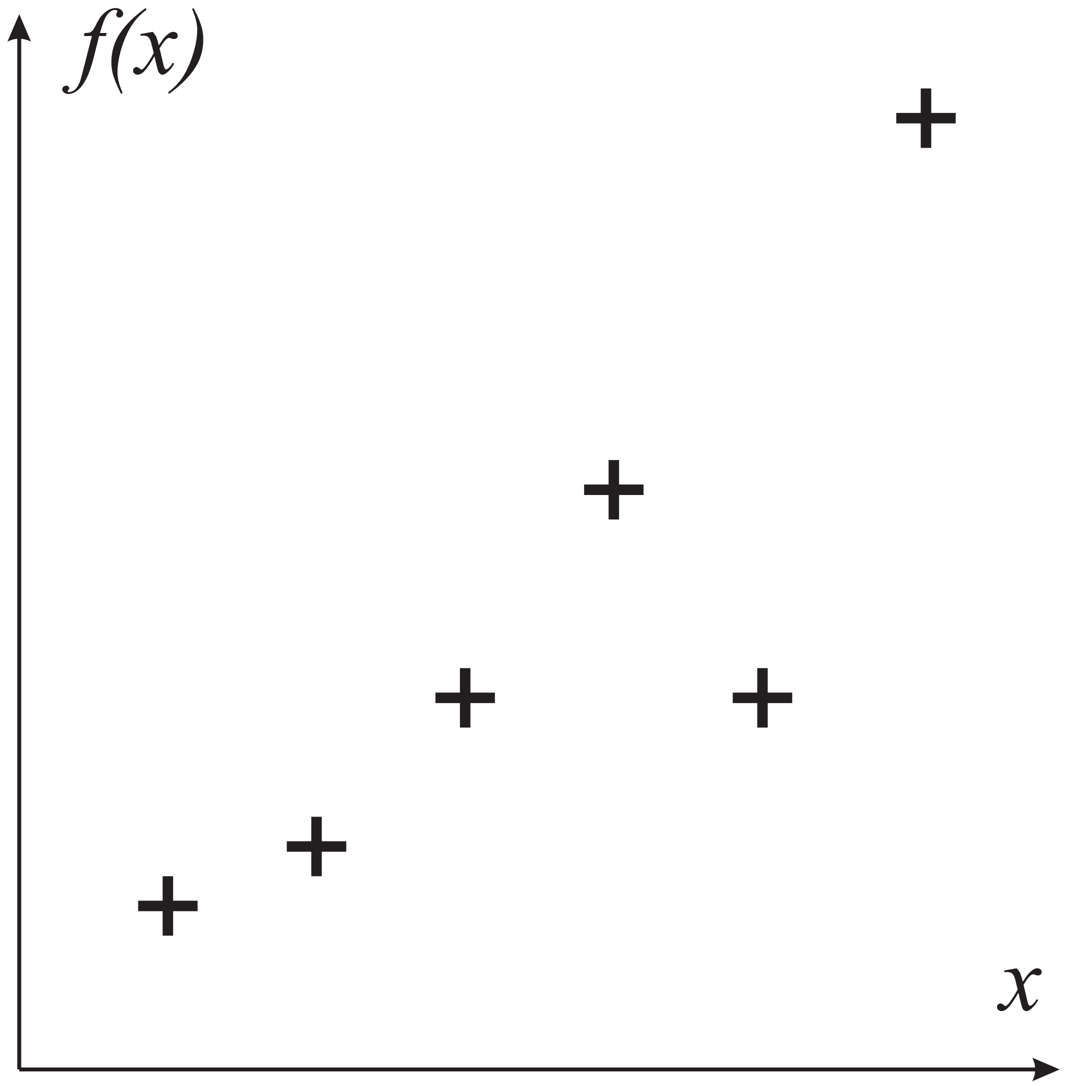
Inferences are commonly drawn

- by deduction, which, by analyzing valid argument forms, draws out the conclusions implicit in their premises (preserves truthfulness),
- by induction, which argues from many instances to a general statement (preserves falsity),
- by probability, which passes from frequencies within a known domain to conclusions of stated likelihood,
- by statistical reasoning, which concludes that, on the average, a certain percentage of a set of entities will satisfy the stated conditions.

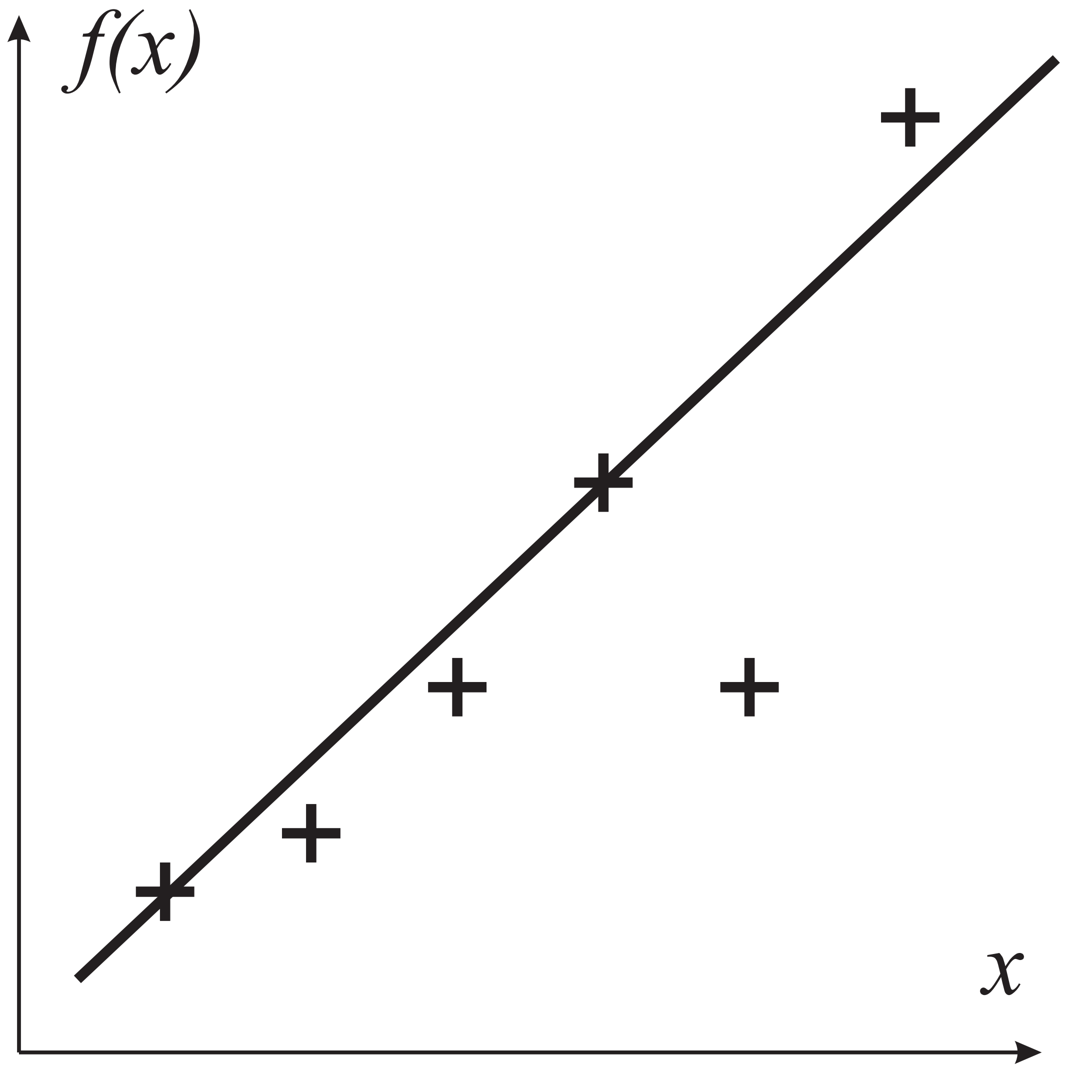
# Motivation: inductive learning

- Simplest form: learn the function $f(x)$ from a training multi-set, i.e., pairs $\{x_i, f(x_i)\}$, $i = 1 \ldots n$.
- Find a hypothesis $h(x) \approx f(x)$.
- This is a highly simplified model of real learning:
  - Ignores prior knowledge.
  - Assumes examples are given.

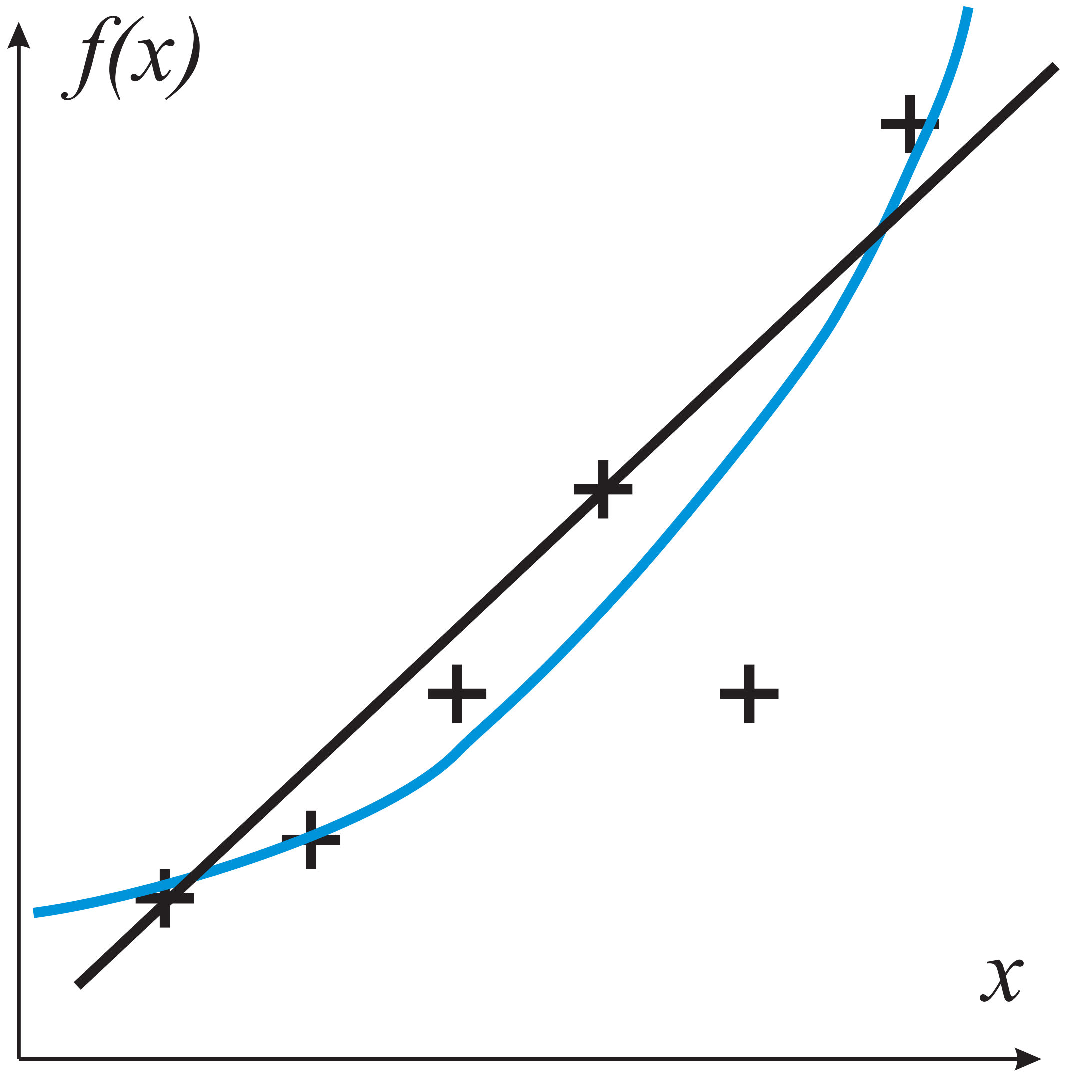
# Example: the function approximation



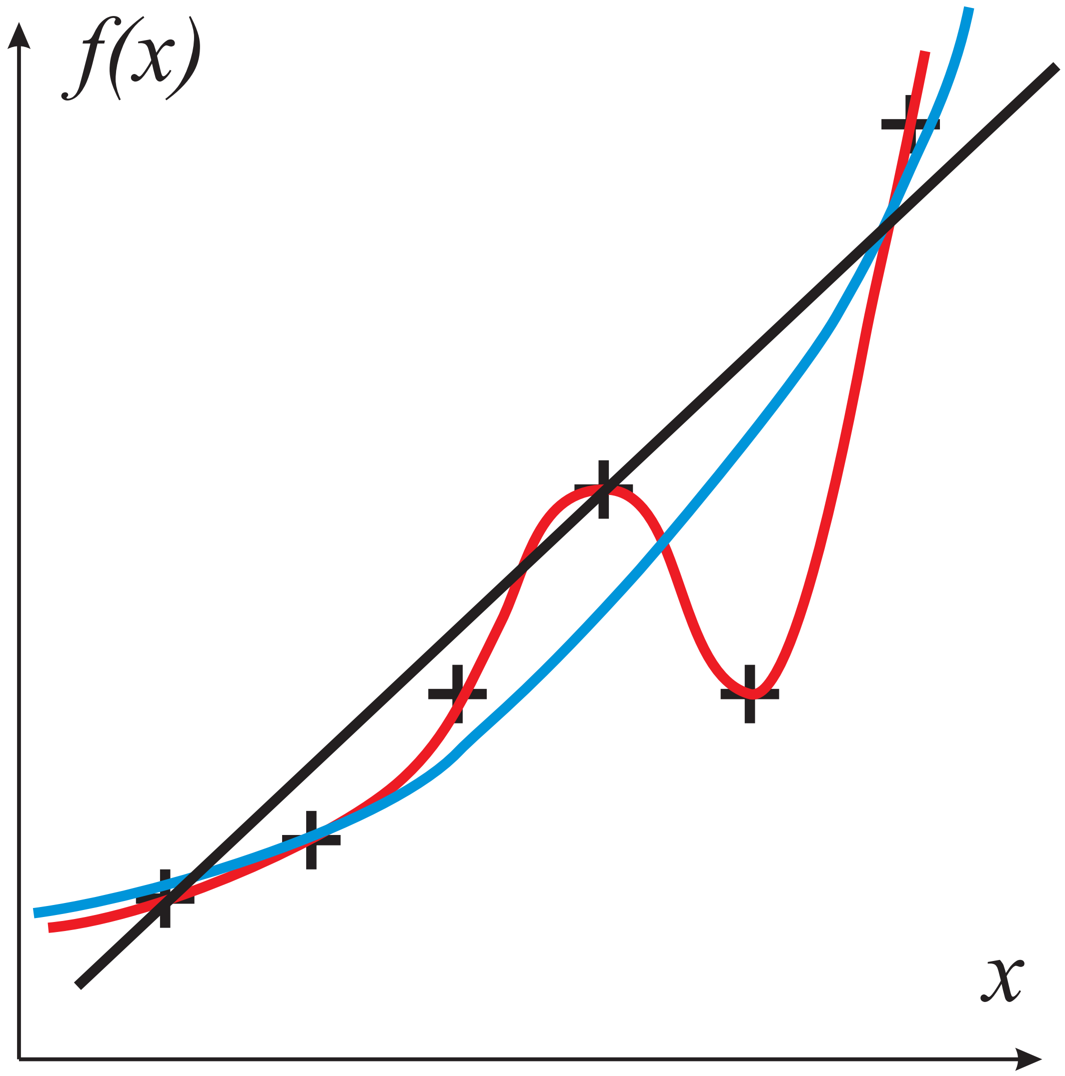
# Example: the function approximation



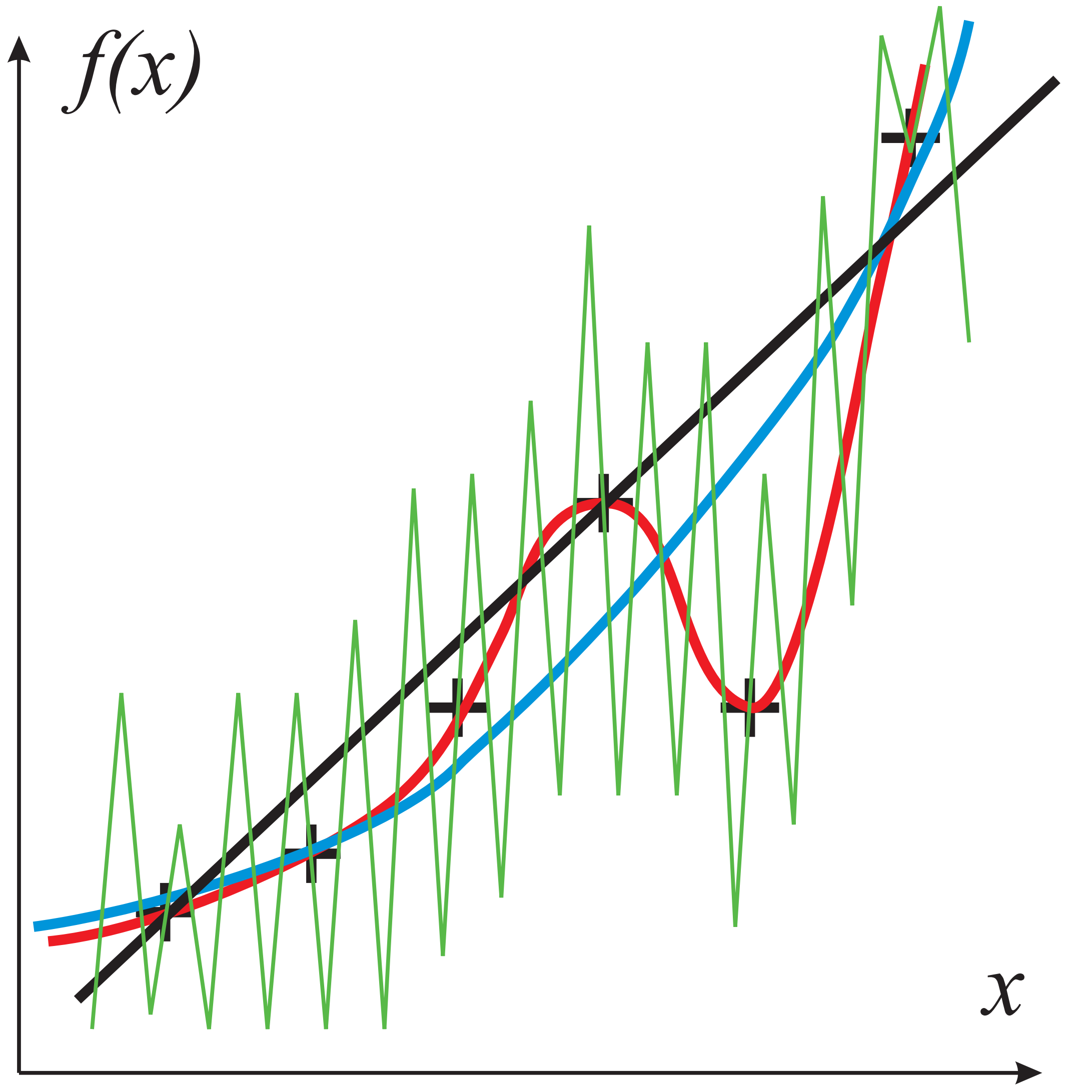
# Example: the function approximation



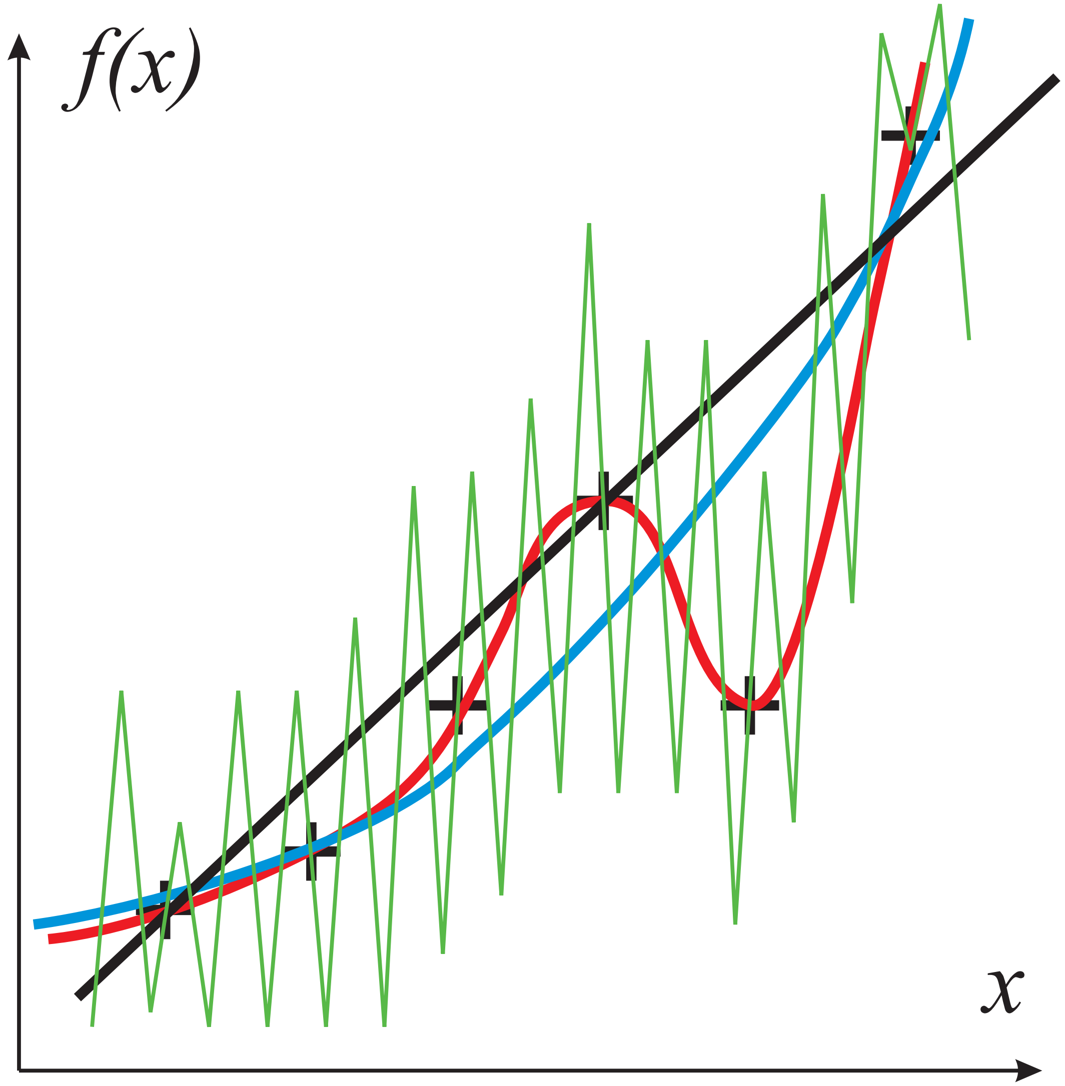
# Example: the function approximation



# Example: the function approximation



# Example: the function approximation



**Occam's razor**, 14th centrury:

*Pluralitas non est ponenda sine necessitate* = Plurality should not be posited without necessity.

Prefer the simplest hypothesis consistent with data.

---

Even ancient Greeks, **Spartans, 5th century BC**:
$\Lambda\alpha\kappa\omega\nu\iota\zeta\iota\nu\ \epsilon\sigma\tau\iota\ \phi\iota\lambda o\sigma o\phi\epsilon\iota\nu.$
*'Lakonidis esti filosofin'*
To talk with as few words as possible is very wise.

# Occam's razor

If two theories explain the facts equally well, then the simpler theory is to be preferred.

- There are fewer short hypotheses than long hypotheses.
- A short hypothesis that fits the data is unlikely to be a coincidence.
- A long hypothesis that fits the data may be a coincidence.

William of Occam
1285-1349, England
Franciscan monk

# Types of feedback in learning

- Supervised learning.

  Correct answers (hidden state, class) are available for each observation. The training multi-set.

- Unsupervised learning.

  Correct answers are not available. The answers have to be sought in data itself ⇒ data analysis (data mining, cluster analysis . . . ).

- Semi-supervised. Teacher's classification (hidden state assignments) is available only for a subset of observations.

  *The line of text recognition example; a human is able to label symbols but the position of a particular symbol is difficult to provide. It should be learned from data.*

- Reinforcement learning.

  Occasional awards are provided.

# Statistical learning – two forms how to express conditional probabilities

1. A function of two variables $x, y$.

   $p_{X|Y}(x \,|\, y)\colon X \times Y \to \mathbb{R}$.

2. An ensemble of $|Y|$ functions of a single variable $x$.

   $p_{X|y}(x)\colon Y \to \mathbb{R}$, $y \in Y$.

   - The conditional probability (likelihood) of observation $x$ under the condition of the state $y$ is thus the value of the function $p_{X|y}$ in the point $x$.
   - A specific function from the ensemble is determined by one of the states $y$.

# When is learning necessary?

- Learning is needed in the case, when knowledge about the recognized object is insufficient to solve a pattern recognition task without learning.
- Most often the knowledge about the conditional probabilities (likelihoods) $p_{X|Y}(x \mid y)$ is insufficient, i.e., it is not known exactly enough how the observation $x$ depends on the state $y$.
- *“When designer lacks omniscience”*.

# Expressing the lack of knowledge (1)

- The likelihood function $p_{X|Y}$ is known to belong to a class $\mathcal{P}$ of functions.
- It is not known which specific function from the class $\mathcal{P}$ actually describes the object.
- *Expressed equivalently:*
    - Knowledge can be determined by the ensemble of sets $\mathcal{P}(y)$, $y \in Y$.
    - Each of the sets comprises the actual function $p_{X|y}$ .
    - It is not known, which one of the sets it is.

# Expressing the lack of knowledge (2)

- The set $\mathcal{P}$ or, what is the same, the ensemble of sets $\mathcal{P}(y)$, $y \in Y$, can be parameterized almost always in such a way that the function $f(x, a)$ of two variables $x$, $a$ is known. It determines the function $f(x)\colon X \to \mathbb{R}$ of one single variable for each fixed value of the parameter $a$.

- The set $\mathcal{P}(y)$ is thus $\{f(a) \mid a \in A\}$, where $A$ is the set of values of the unknown parameter $a$.

- Our knowledge about the probabilities $p_{X|Y}(x \mid y)$ which is given by the relation $p_{X|Y} \in \mathcal{P}$ means that the value $a^*$ of the parameter $a$ is known to exist for which $p_{X|y} = f(a^*)$.

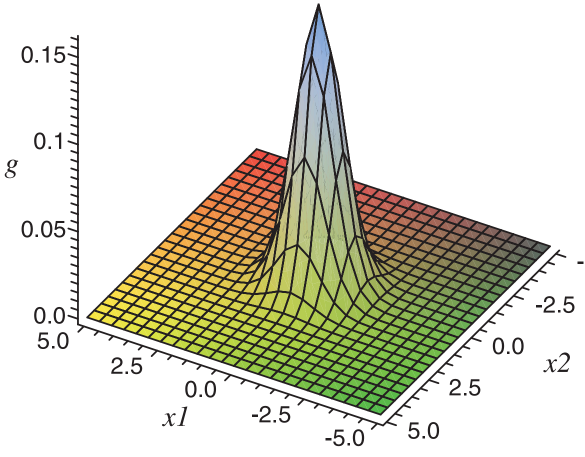
# Example, parametrization of $\mathcal{P}(y)$ $n$-dimensional Gaussian

- Let $\mathcal{P}$ be a set consisting of a probability distributions of $n$-dimensional Gaussian random variables with mutually independent components and unit variances.

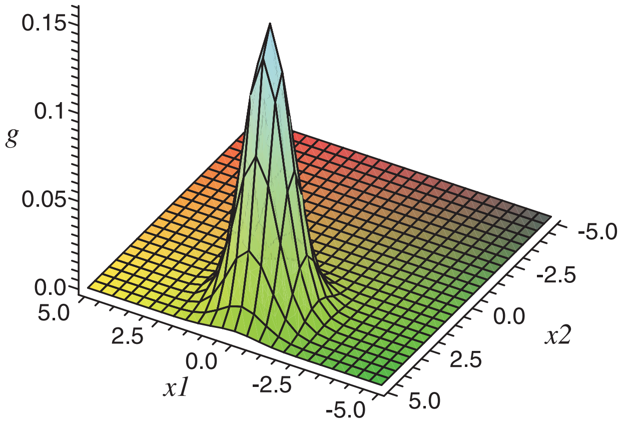
- The set $\mathcal{P}(y)$ in a parameterized form is the set $\{f(\mu) \mid \mu \in \mathbb{R}^n\}$ of the functions $f(\mu)\colon X \to \mathbb{R}$ of the form

$$f(\mu)(x) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-(x_i - \mu_i)^2}{2}\right) \; .$$

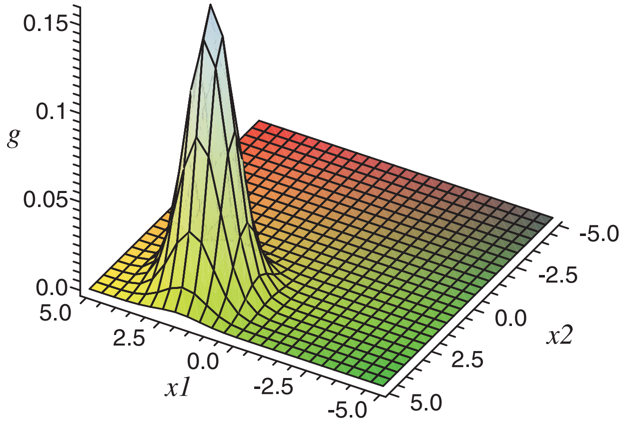
$\mu_1 = 0,\ \mu_2 = 0$



$\mu_1 = 0,\ \mu_2 = 3$



$\mu_1 = 2,\ \mu_2 = 3$



# Parametric set of correct strategies

- Let have $p_{X|y}$, $y \in Y$, defined up to values of the unknown parameters $a_1, a_2, \dots, a_n$, $a_i \in A$.
- The strategy is provided up to the values of unknown parameters can be created
$$q(x, a_1, a_2, \dots, a_n) \ .$$
- It illustrates how the observation $x$ would be assessed if the parameters $a_y$, $y = 1, 2, \dots, n$, determining the distribution $p_{X|y}$, were known.

---

- The parametric set of strategies can be created

$$Q = \{q(a_1, a_2, \dots, a_n) \mid a_1 \in A, a_2 \in A, \dots, a_n \in A \ \} \ ,$$

into which the sought strategy surely belongs.

# Formulation as a non-Bayesian task with non-random interventions

Non-random interventions = unknown parameters.

- It can happen that in such an approach the guaranteed level of risk will be insufficient. It happens when an *a priori* known set of models is too extensive.
- In such situations, it is necessary to narrow the set of models or the possible strategies set by using additional information.
- This additional piece of information is obtained from a teacher in a process of learning. The information has the form of a multi-set $T = \big((x_1, y_1),$ $(x_2, y_2), \ldots, (x_l, y_l)\big)$ in which $x_i \in X$ and $y_i \in Y$.

Note: Training set × training multi-set.

# Learning = selecting a single strategy

from the set of *a priori* known strategies using information provided in a learning process.

- Natural selection criterion for the strategy selection is the risk

$$\sum_{x\in X}\sum_{y\in Y} p_{XY}(x,y)\, W\big(y, q(x)\big)\ .$$

- Wrong decisions of which are quantified by the penalty $W$.
- Unfortunately, the criterion cannot be computed because the function $p_{XY}(x,y)$ is not known.
- The lack of knowledge about the function $p_{XY}(x,y)$ is substituted to a certain degree by the training set or multi-set.

# Substitutive optimality selection criteria

- Substitutive optimality criterion can be calculated based on the information obtained during the empirical learning.
- Nevertheless, a gap always remains between the criterion that should, but cannot, be calculated, and the substitute criterion which can be computed.
- This gap can be based on conscientiousness (intuition or experience) of the learning algorithm's designer or it can be estimated in some way.

# Four substitutive optimization criteria

1. Learning according to the maximal likelihood.

   Originates in the statistical literature, can be anchored to F. Gauß (end of 18th century), R. Fisher (1936).

2. Learning according to a non-random training set.

   M.I. Schlesinger (1989).

3. Learning by minimization of the empirical risk.

   F. Rosenblatt (1962), M. Ajzerman, E. Braverman, L. Rozoner (1970).

4. Learning by minimizing of the structural risk.

   V. Vapnik, A. Chervonenkis (1974, 1998).

   *(will be discussed in a separate lecture.)*

# Learning according to the maximal likelihood (1)

Given:

Conditional probability $p_{X|Y}(x \,|\, y, a_y)$ (likelihood) known up to an unknown value of the parameter $a_k$.

Training multi-set

$$T = \big((x_1, y_1), (x_2, y_2), \ldots, (x_l, y_l)\big) \,, \quad x_i \in X \,, \quad y_i \in Y \,,$$

which is treated as in statistics, i.e., by assuming that the elements are mutually independent random variables with the probability distribution

$$p_{XY}(x, y) = p_Y(y) \; p_{X|Y}(x \,|\, y, a_y) \;.$$

# Learning according to the maximal likelihood (2)

In this case, the probability of the training multi-set $T$ can be computed for each ensemble of unknown parameters $a = (a_y, y \in Y)$ as the likelihood function

$$L(T, a) = \prod_{i=1}^{l} p_Y(y_i)\, p_{X|Y}(x_i \,|\, y_i, a_{y_i}) \; .$$

Learning according to the maximal likelihood seeks such values $a_y^*$, $y \in Y$ maximizing the probability (likelihood function), i.e.,

$$a^* = (a_y^*, y \in Y) = \operatorname*{argmax}_{(a_y, y \in Y)} \prod_{i=1}^{l} p_Y(y_i)\, p_{X|Y}(x_i \,|\, y_i, a_{y_i}) \; .$$

The ensemble $a^*$ of values $(a_y^*,\ y \in Y)$ is treated as if the values were real. The ensemble $(a_y^*, y \in Y)$ is substituted into the general expression $q(x, a_1, a_2, \ldots, a_n)$ and the recognition is performed according to the strategy $q(x, a_1^*, a_2^*, \ldots, a_n^*)$.

# Learning according to the maximal likelihood (3)

There is an equivalent formulation.

Let $\alpha(x, y)$ gives the frequency of the pair $(x, y)$ in the training multi-set. We can write under the condition of non-zero likelihoods $p_{X|Y}(x \,|\, y, a_y)$

$$\begin{aligned} a^* &= \operatorname*{argmax}_{(a_y, y \in Y)} \prod_{x \in X} \prod_{y \in Y} \left( p_Y(y)\, p_{X|Y}(x \,|\, y, a_y) \right)^{\alpha(x,y)} \\ &= \operatorname*{argmax}_{(a_y, y \in Y)} \sum_{y \in Y} \sum_{x \in X} \alpha(x, y) \log p_Y(y)\, p_{X|Y}(x \,|\, y, a_y) \\ &= \operatorname*{argmax}_{(a_y, y \in Y)} \sum_{y \in Y} \sum_{x \in X} \alpha(x, y) \log p_{X|Y}(x \,|\, y, a_y) \,. \end{aligned}$$

Observation:

In the sum, each term of addition depends only on one single element of this set.

# Learning according to the maximal likelihood (4)

- The maximization task decomposes into $|Y|$ independent maximization tasks that seek $a_y^*$ according to the requirement

$$a_y^* = \operatorname*{argmax}_{a_y} \sum_{x \in X} \alpha(x,y) \log p_{X|Y}(x \,|\, y, a_y) \; .$$

- Notice that it is not needed to know *a priori* probabilities $p_Y(y)$ when determining $a_y^*$.

# Learning according to a non-random training set (1)

- The approach is common in the recognition of/in images.
- The random examples are not easy to be obtained.
- Instead, a carefully selected patterns are used for learning (i.e., tuning the recognition algorithm).
- Designers requires that the selected patterns:

  1. represent well the whole set of images which are to be recognized, and
  2. any of the images chosen for learning is good enough, of a satisfying quality, not damaged, so the recognition algorithm should evaluate it as a very probable representative of its class.

# Learning according to a non-random training set (2)

*Formally:*

- Let $X(y)$, $y \in Y$, be the ensemble of quite probable examples reliably selected by the teacher.
- The parameter $a_y^*$ which determines the probability distribution $p_{X|Y}$ (likelihood) is to be chosen in such a way that

$$a_y^* = \operatorname*{argmax}_{a_y \in A} \; \min_{x \in X(y)} \; p_{X|Y}(x \,|\, y, a_y) \,.$$

---

- Notice that a training set is used and not a training multi-set.
- The solution of the task does not depend any longer on how many times this or that observation has occurred. It is significant that it has occurred at least once.

# Example comparing two learning approaches for multi-dimensional Gaussians

If $\mathcal{P}(y)$ is a set of functions of the form

$$p(x|y,\mu_y) = \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi}} \exp\left( \frac{-(x_i - \mu_{iy})^2}{2} \right)$$

then in the case of

- Learning according to the maximal likelihood: the $\mu_y^*$ is estimated as the mean value $(1/l)\sum_{i=1}^{l} x_i$ of observations of the object in the $y$-th state.
- Based on the non-random training set: the $\mu_y^*$ is estimated as the center of the smallest circle containing all vectors which were selected by the teacher as rather good representatives of objects in the $y$-th state.

# Learning by minimization of the empirical risk (1)

- $W(y,d)$ is a penalty function.
- $Q = q(\Theta)$ is a parameterized set of strategies expressed as the strategy $q(\Theta)\colon X \to D$ defined up to unknown values of certain parameters $\Theta$.
- The quality of each strategy $q(\Theta)$ is measured by the risk $R(\Theta)$

$$R(\Theta) = \sum_{y \in Y} \sum_{x \in X} p_{XY}(x,y)\, W\big(y, q(\Theta)(x)\big)\ .$$

- The risk $R(\Theta)$ should be minimized by an appropriate selection of the value $\Theta$.

# Learning by minimization of the empirical risk (2)

- The risk cannot be measured because the statistical model $p_{XY}(x,y)$ is not known.

- Fortunately, based on the training multi-set $T = \big((x_1,y_1),\ (x_2,y_2),\ \ldots,(x_l,y_l)\big)$ the empirical risk can be defined,

$$R_{\mathrm{emp}}(\Theta) = \frac{1}{l}\sum_{i=1}^{l} W\big(y_i, q(\Theta)(x_i)\big)\ ,$$

  which can be measured and can substitute the actual risk.

- This approach to learning creates a set of strategies on the basis of partial knowledge about the statistical model of the object.

- From this parametric set, such a strategy is next chosen which secures the minimal empirical risk on the submitted training multi-set.

# Example, $n$-dim Gaussian, learning by minimization of the empirical risk

- Consider a special case as in previous example, a multidimensional Gaussian with unit variance.
- If the number of states and number of decisions is equal to two and the observation is a multi-dimensional Gaussian random variable with mutually independent components and unit variance

  then

  the set of strategies contains strategies separating classes by the hyperplane.
- Learning aims at finding the hyperplane which secures the minimal value of the empirical risk (or the minimal number of errors in this particular case) on the training multi-set.

# Learning does not have a magic ability

- Variety of approaches to learning exist based on how unavailable risk is approximated. Four of them were mentioned in this talk
  - Maximal likelihood learning.
  - Learning according to a non-random training set.
  - Learning minimizing empirical risk.
  - *Structural risk minimization (to be explained later).*

---

- Learning lost a hope to have miraculous properties.
- Recognition (without learning) is used to solve a single particular problem.
- Recognition with learning solves an unambiguously defined class of problems.

# What is learning in pattern recognition ?

- Learning consists of delimiting a task to be recognized (decided on) and finding a good algorithm to make such decisions.
- A designer of a learning algorithm has to understand the variety of all possible task that can occur in the delimited task.
- Said in another words, a designer has to find a general solution to the whole delimited class of problems.
- The solution is expressed as a set of parametric strategies.
- Parameters are learned from the training set (or multi-set).