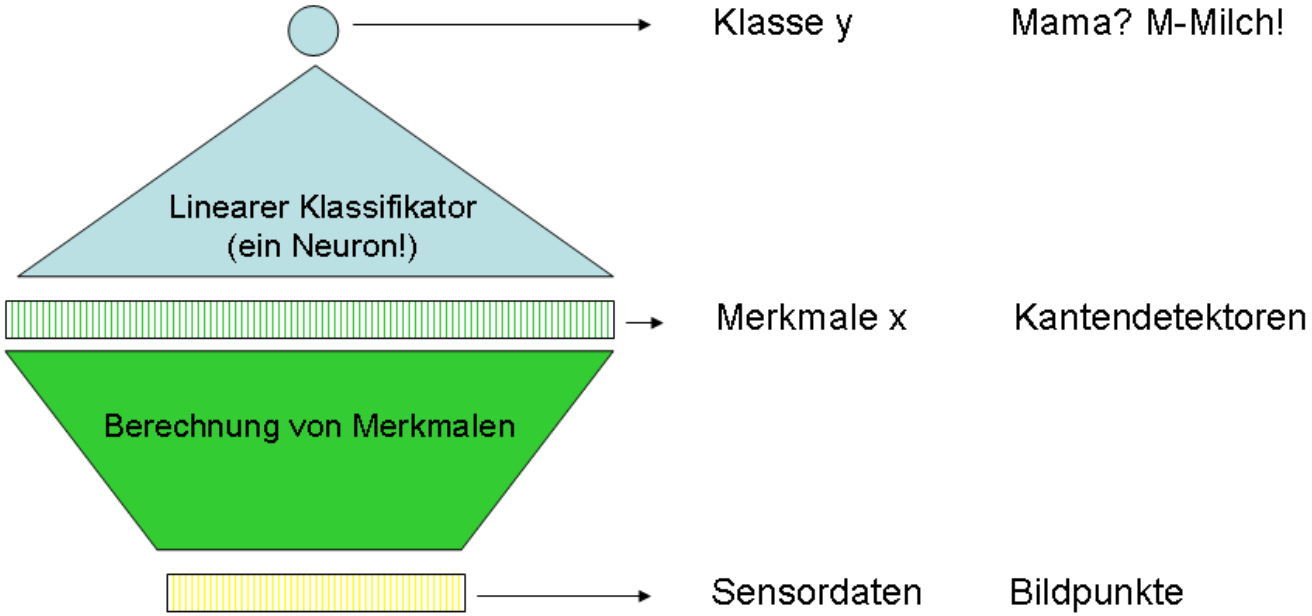
Das Perceptron

Volker Tresp

Ein biologisch motiviertes Lernmodell



Input-Output Modelle

- Ein biologisches System muss basierende auf Sensordaten eine Entscheidung treffen
- Ein OCR System klassifiziert eine handgeschriebene Ziffer
- Ein Prognosesystem sagt den morgigen Energieverbrauch voraus

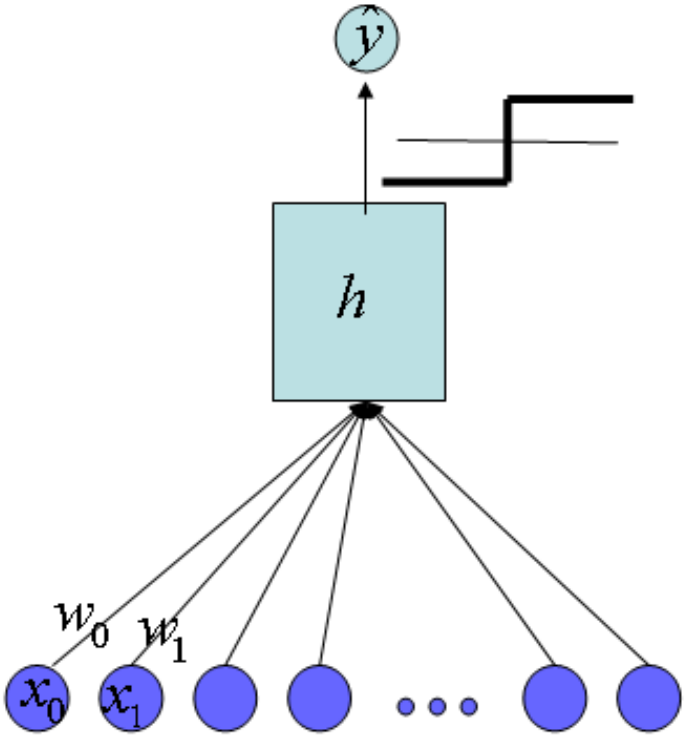
Überwachtes Lernen

- Im überwachten Lernen geht man davon aus, dass in den Trainingsdaten sowohl die Eingangsmuster als auch die Zielgrößen bekannt sind
- Ziel ist die korrekte Klassifikation neuer Muster
- Der vielleicht einfachste aber eigentlich auch schon erstaunlich mächtige Klassifikator ist ein linearer Klassifikator
- Ein linearer Klassifikator kann durch ein Perceptron realisiert werden, also einem einzigen formalisierten Neuron!

Überwachtes Lernen und Lernen von Entscheidungen

- Man kann mit einigem Recht argumentieren, dass Maschinelles Lernen nur interessant ist, inwieweit es zu Entscheidungen führt
- Allerdings lassen sich viele Entscheidungen auf ein überwachtes Lernproblem (Vorhersage/Klassifikationsproblem) zurückführen: Wenn ich die Postleitzahl automatisch lese und erkenne, dann weiß ich wohin ich den Brief schicken muss
- Dies bedeutet: Entscheidung und Aktion kann man häufig auf Klassifikation reduzieren

Das Perceptron



- Zunächst wird die Aktivierungsfunktion des Perceptrons als gewichtete Summe der Eingangsgrößen x_i berechnet zu

$$h = \sum_{j=0}^{M-1} w_j x_j$$

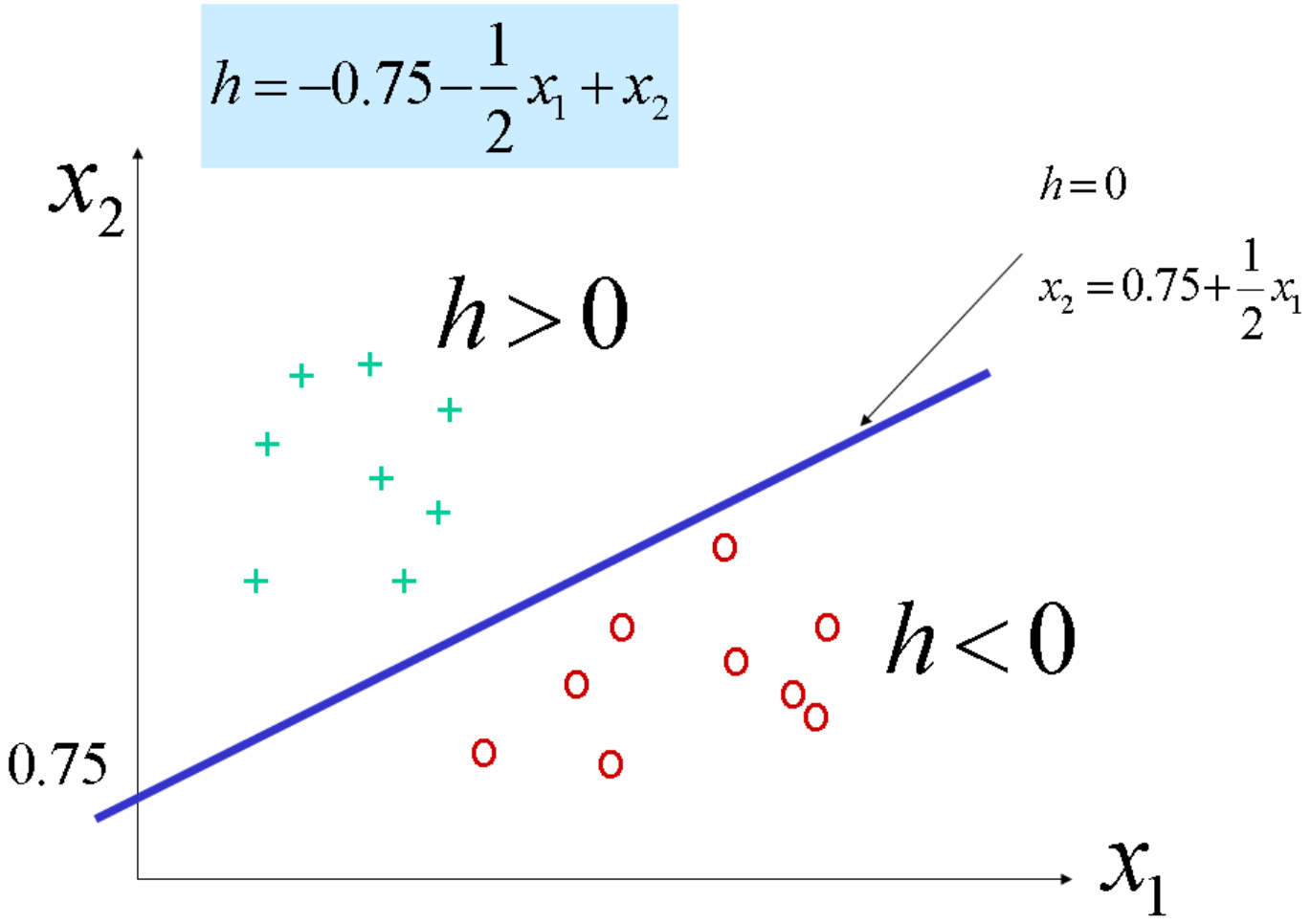
(beachte: $x_0 = 1$ ist ein konstanter Eingang, so dass w_0 dem Bias entspricht)

- Die binäre Klassifikation $y \in \{1, -1\}$ wird berechnet zu

$$y = \mathbf{sign}(h)$$

- Die lineare Klassifikationsgrenze (*separating hyperplane*) ist definiert durch $h = 0$

Zwei linear-separierbare Klassen



Die Perceptron-Lernregel

- Die (synaptischen) Gewichte w_i sollten so eingestellt werden, dass das Perceptron die N Trainingsdaten $\{y_1, \ldots, y_N\}$ richtig klassifiziert
- Dazu definieren wir als Kostenfunktion des Perceptrons

$$\text{cost} = -\sum_{i \in \mathcal{M}} y_i h_i = -\sum_{i \in \mathcal{M}} \left(y_i \sum_{j=0}^{M-1} w_j x_{i,j} \right)$$

  wobei $\mathcal{M} \subseteq \{1, \ldots, N\}$ die Indexmenge der (gegenwärtig) falsch klassifizierten Muster ist und $x_{i,j}$ der Wert des j-ten Eingangs im i-ten Musters ist
- Offensichtlich ist $\text{cost} = 0$ nur dann, wenn alle Muster richtig klassifiziert werden, ansonsten $\text{cost} > 0$
- Die Ableitung der Kostenfunktion nach den Gewichten ist (Beispiel w_j)

$$\frac{\partial \text{cost}}{\partial w_j} = -\frac{1}{2} \sum_{i=1}^{N} (y_i - \widehat{y}_i) x_{i,j} = -\sum_{i \in \mathcal{M}} y_i x_{i,j}$$

Die Perceptron-Lernregel (2)

- Eine sinnvolle Lernregel wird versuchen, die Kostenfunktion zu minimieren, z.B. durch einfachen Gradientenabstieg

$$w_j \longleftarrow w_j + \eta \sum_{i \in \mathcal{M}} y_i \mathbf{x}_{i,j}$$

  hier ist η eine kleine positive Lernrate

- Im tatsächlichen Algorithmus werden in zufälliger Reihenfolge dem Perceptron falsch klassifizierte Muster angeboten (stochastischer Gradientenabstieg). Einmal ist dies biologisch plausibler, und zweitens ist die Konvergenz schneller. Seien $\mathbf{x}(t)$ und $y(t)$ die angebotenen Muster im t-ten Iterationsschritt. Dann wird adaptiert

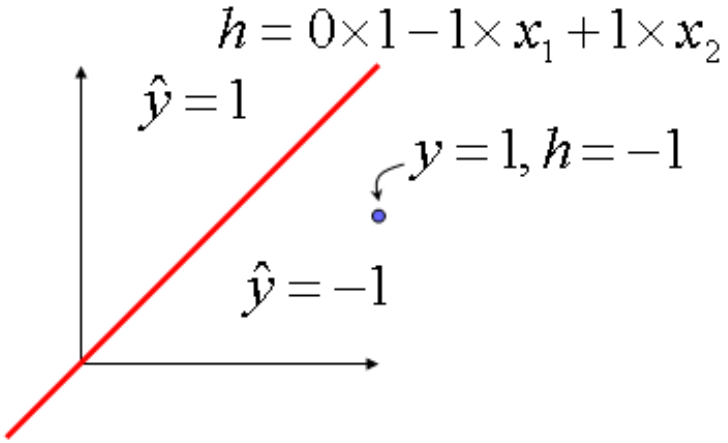
$$w_j \longleftarrow w_j + \eta y(t) x_j(t)$$

- Eine Gewicht wächst an, wenn (postsynaptisches) $y(t)$ und (präsynaptisches) $x_j(t)$ gleiches Vorzeichen besitzen; bei unterschiedlichen Vorzeichen verringert sich das Gewicht (vergleiche: **Hebb'sches Lernen**)

Die Perceptron-Lernregel (3)

- Konvergenzbeweis: Bei trennbaren Problemen konvergiert der Algorithmus nach endlich vielen Schritten

Beispiel: Perceptron Lernen, $\eta = 0.1$



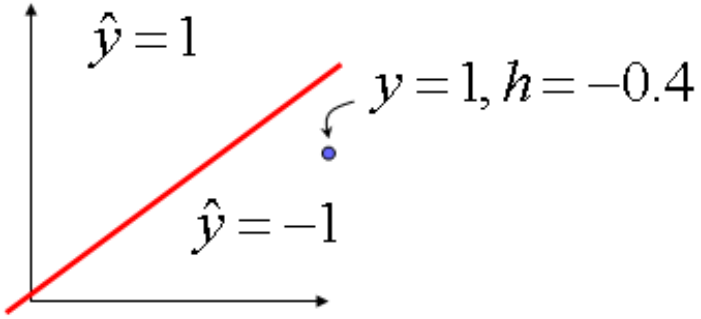
Adaption:

$w_0 \leftarrow 0 + 0.1 \times 1 = 0.1$

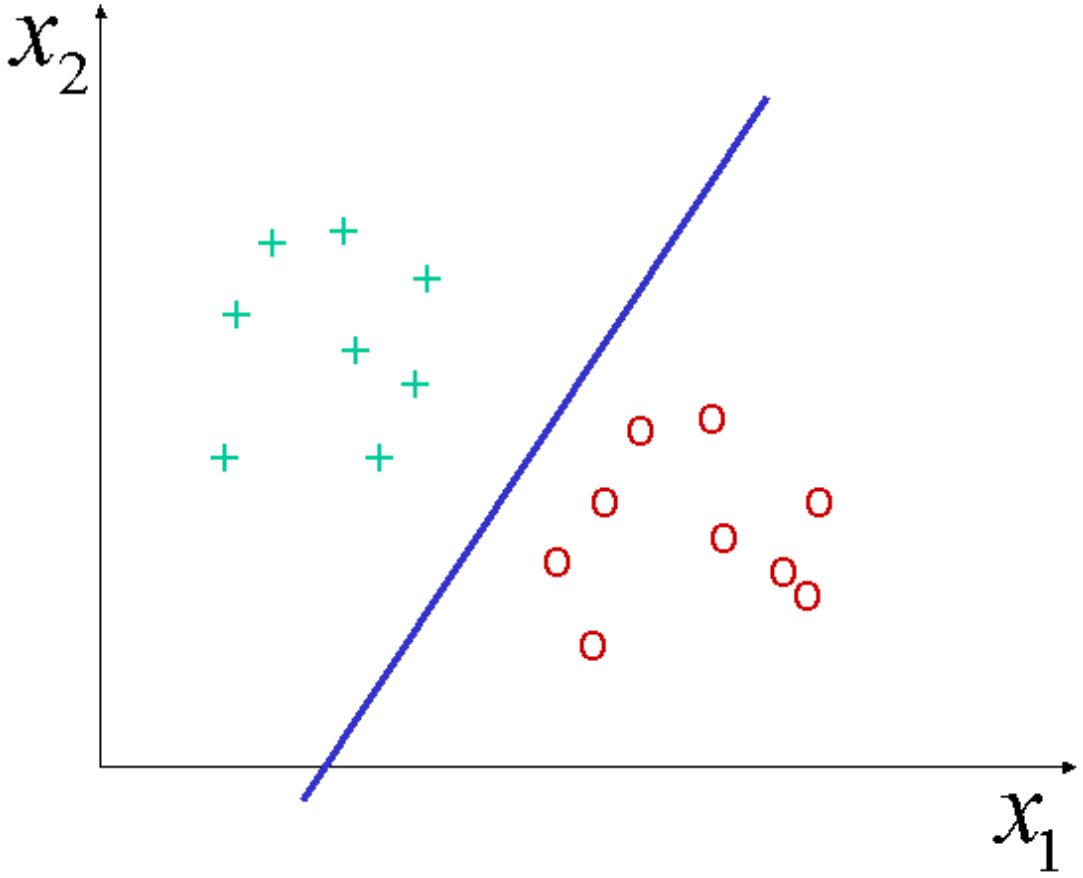
$w_1 \leftarrow -1 + 0.1 \times 2 = -0.8$

$w_2 \leftarrow 1 + 0.1 \times 1 = 1.1$

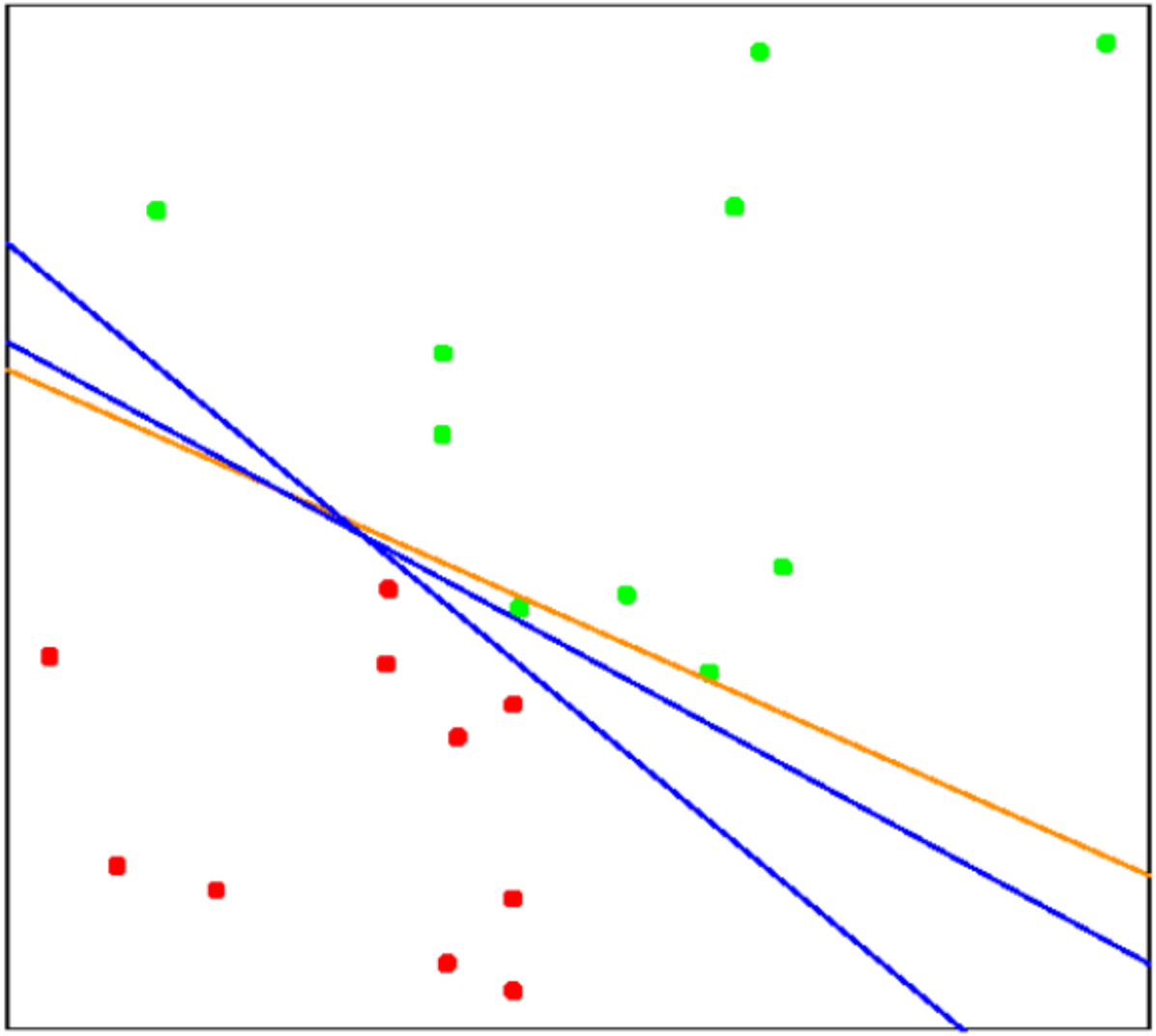
$h = 0.1 \times 1 - 0.8 \times x_1 + 1.1 \times x_2$



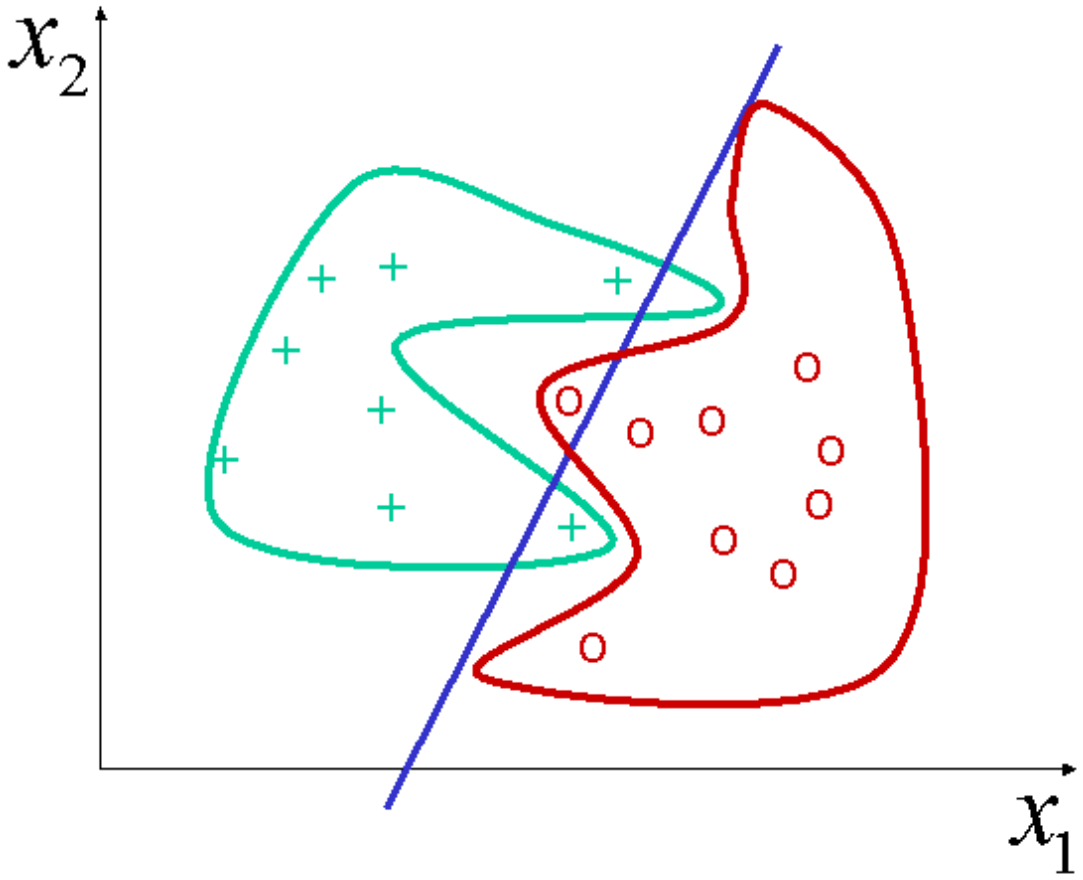
Zwei linear-separierbare Klassen



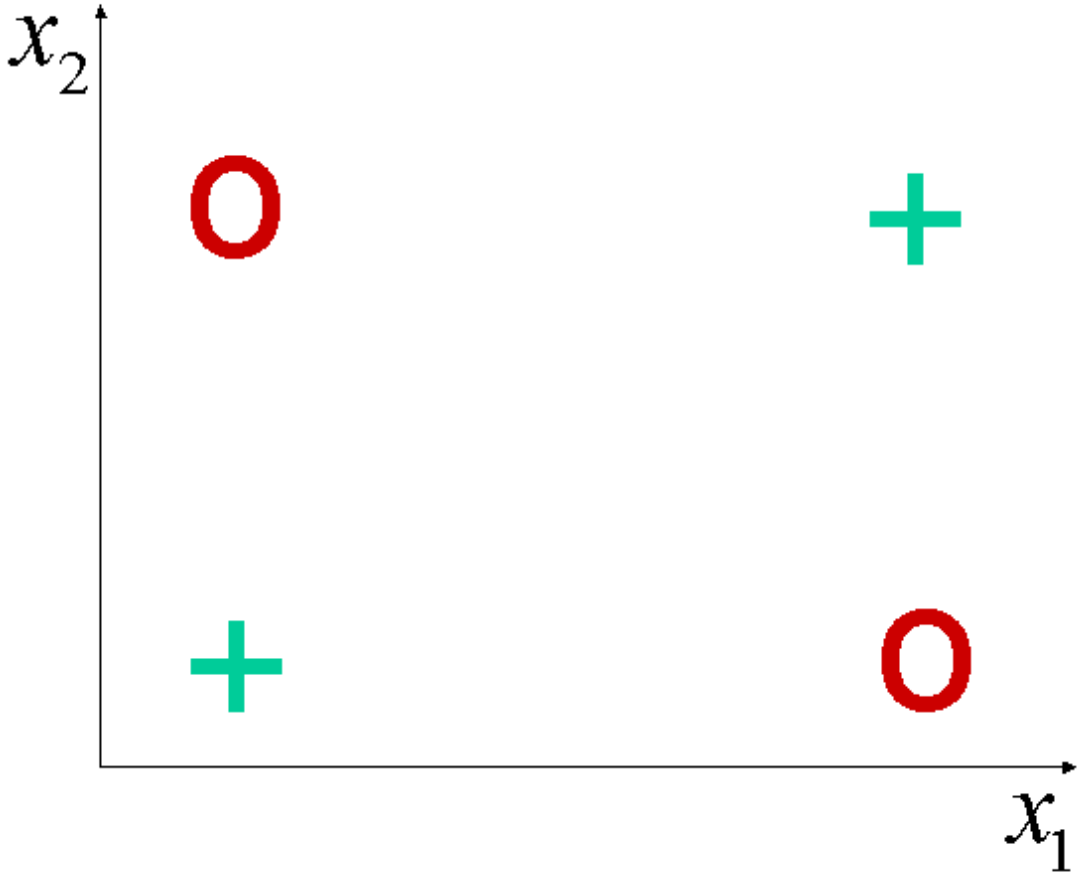
Konvergenz des Perceptrons und Mehrdeutigkeit der Lösung



Zwei Klassen, die nicht linear separierbar sind



Das klassische Beispiel nicht-separierbarer Klassen: XOR



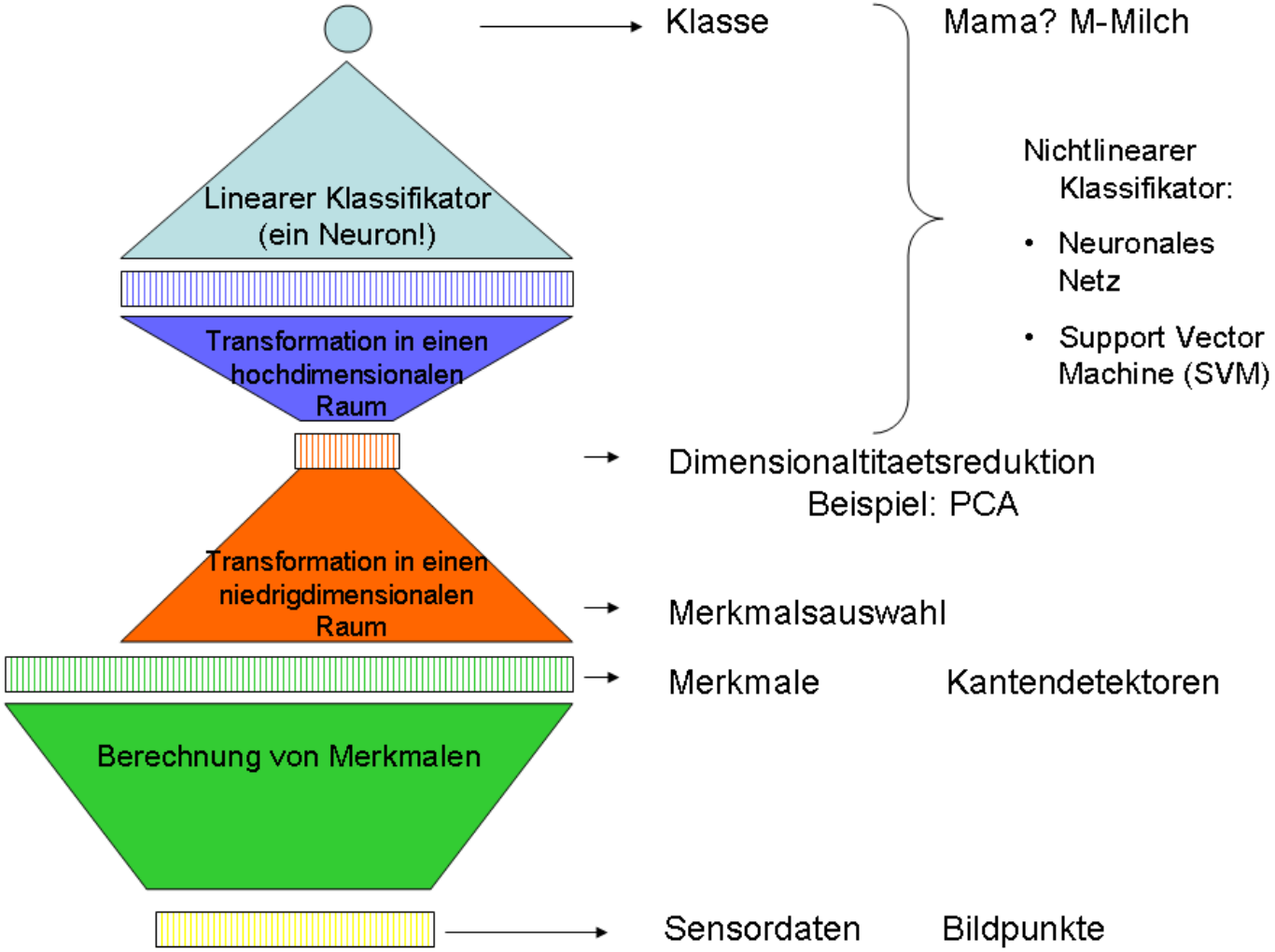
Kommentare zum Perceptron

- Konvergenz kann sehr schnell sein
- Gibt es solche "Grossmutterzellen" (*grandmother cells*)? Grossmutterareale?
- Lineare Klassifikatoren sind auch heute von zentraler Bedeutung: mit $M \longrightarrow \infty$ werden alle Probleme linear separierbar!
- In manchen Fällen sind die Rohdaten schon hochdimensional: Texte haben Merkmalsvektoren > 10000 (Anzahl der möglichen Worte)
- In anderen Fällen transformiert man zunächst die Eingangsdaten in einen hoch-dimensionalen (∞-dimensional) Raum, und wendet dann einen linearen Klassifikator an (Kernel-trick in der SVM, Neuronale Netze)
- Betrachtet man die Mächtigkeit eines einzigen Neurons, wieviel Rechenleistung kann man von 100 Milliarden Neuronen erwarten?

Kommentare zum Perceptron (2)

- Die Perceptron Lernregel wird heutzutage nicht mehr viel benutzt
  - keine Konvergenz bei nicht trennbaren Klassen
  - Trennebene nicht eindeutig
- Alternative lineare Klassifikatoren:
  - linear Support Vector Maschine
  - Fisher linear Discriminant
  - Logistic Regression
- Die folgende Graphik zeigt ein mächtigeres Lernmodell; gleichzeitig das Programm eines Großteils der Vorlesung!

Ein mächtigeres Lernmodell

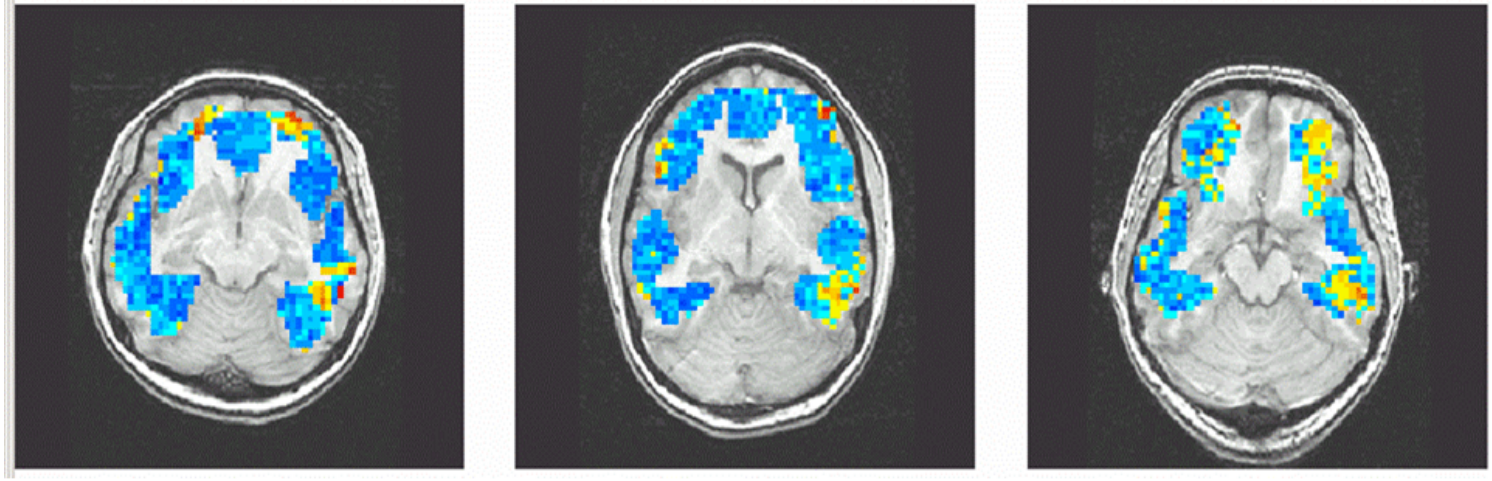


Brain Image Analysis Research Group (Tom Mitchel et al., CMU)

Internal (restricted) New instruments for imaging human brain activity, such as fMRI, offer a wonderful opportunity to study mechanisms in the brain. Our group develops statistical machine learning algorithms to analyze fMRI data. We are specifically interested in algorithms that can learn to identify and track the cognitive processes that give rise to observed fMRI data.

Example: In one fMRI study we trained our algorithms to decode whether the words being read by a human subject were about tools, buildings, food, or several other semantic categories. The trained classifier is 90% accurate, for example, discriminating whether the subject is reading words about tools or buildings.

The following figure shows, for each of three different subjects, the degree to which different brain locations can help predict the word's semantic category. Red and yellow voxels are most predictive. Note the most predictive regions in different subjects are in similar locations.



Paradigma der Mustererkennung

- von Neumann: *... the brain uses a peculiar statistical language unlike that employed in the operation of man-made computers...*
- Eine Klassifikationsentscheidung wird parallel anhand des gesamten Musters getroffen und nicht als logische Verknüpfung einer kleinen Anzahl von Größen oder als mehrschichtiges logisches Programm
- Die linear gewichtete Summe entspricht mehr einem Abstimmen als einer logischen Entscheidungsfunktion; jeder Eingang hat entweder immer einen positiven oder immer einen negativen (gewichteten) Einfluss
- Robustheit: in hohen Dimensionen kommt es nicht auf einen einzigen Eingangswert an; jeder Eingang macht seinen "kleinen" Beitrag

Nachbemerkungen

Warum Mustererkennung?

- Eines der großen Rätsel des Maschinellen Lernens ist der mangelnde Erfolg im Versuch, einfache deterministische logische Regeln zu erlernen
- Probleme: Die erlernten Regeln sind oft trivial, bekannt, extrem komplex oder unmöglich zu interpretieren; die Performanz eines Klassifikators basierend auf einfachen deterministischen logischen Regeln ist in der Regel nicht sehr gut!
- Dies steht scheinbar im Gegensatz zum eigenen Eindruck, dass die Welt wohldefinierten einfachen Regeln gehorcht
- Ebenso folgen (fast) alle Computer Programme, Maschinen (Autos), deterministischen Regeln

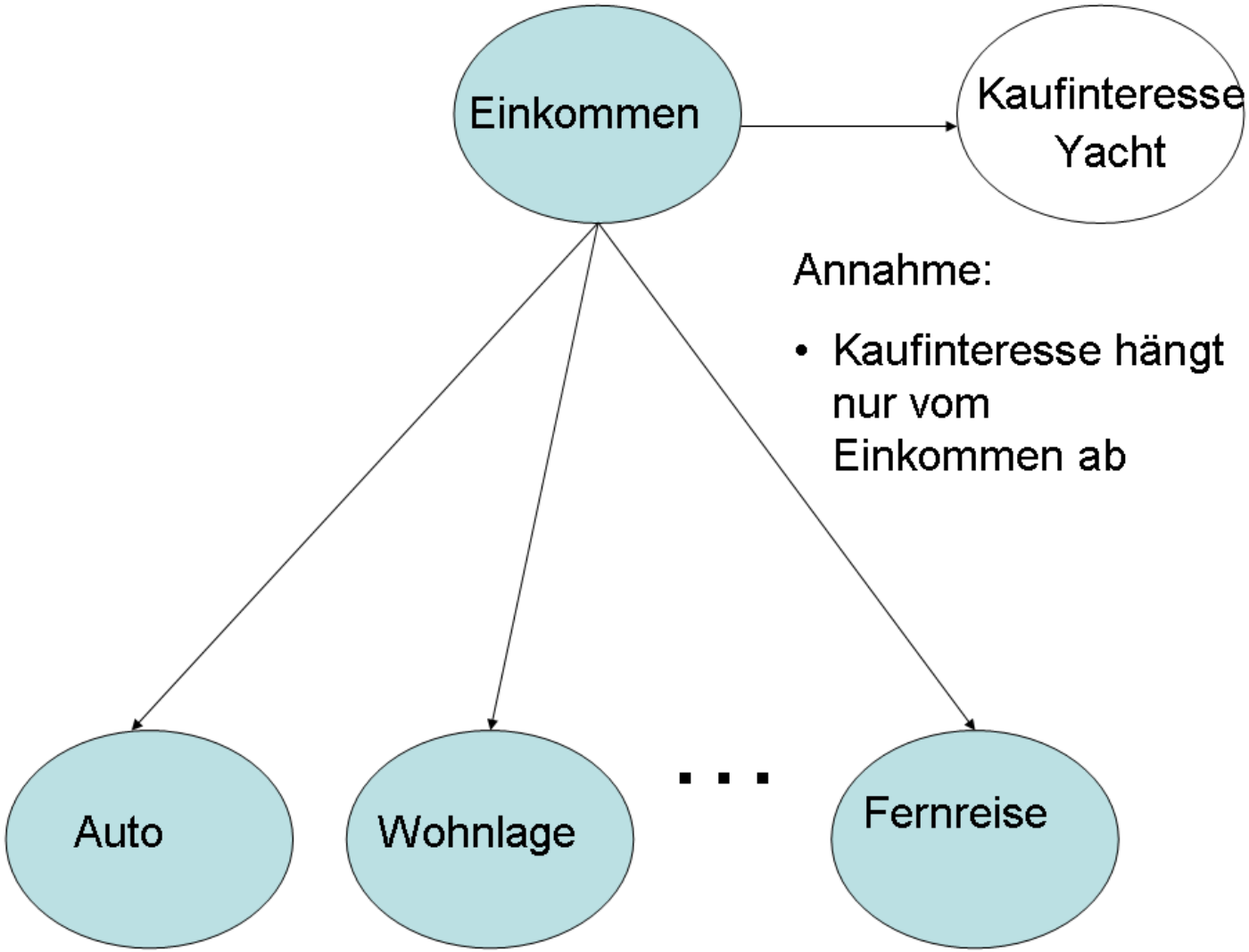
Beispiel: Alle Vögel fliegen hoch

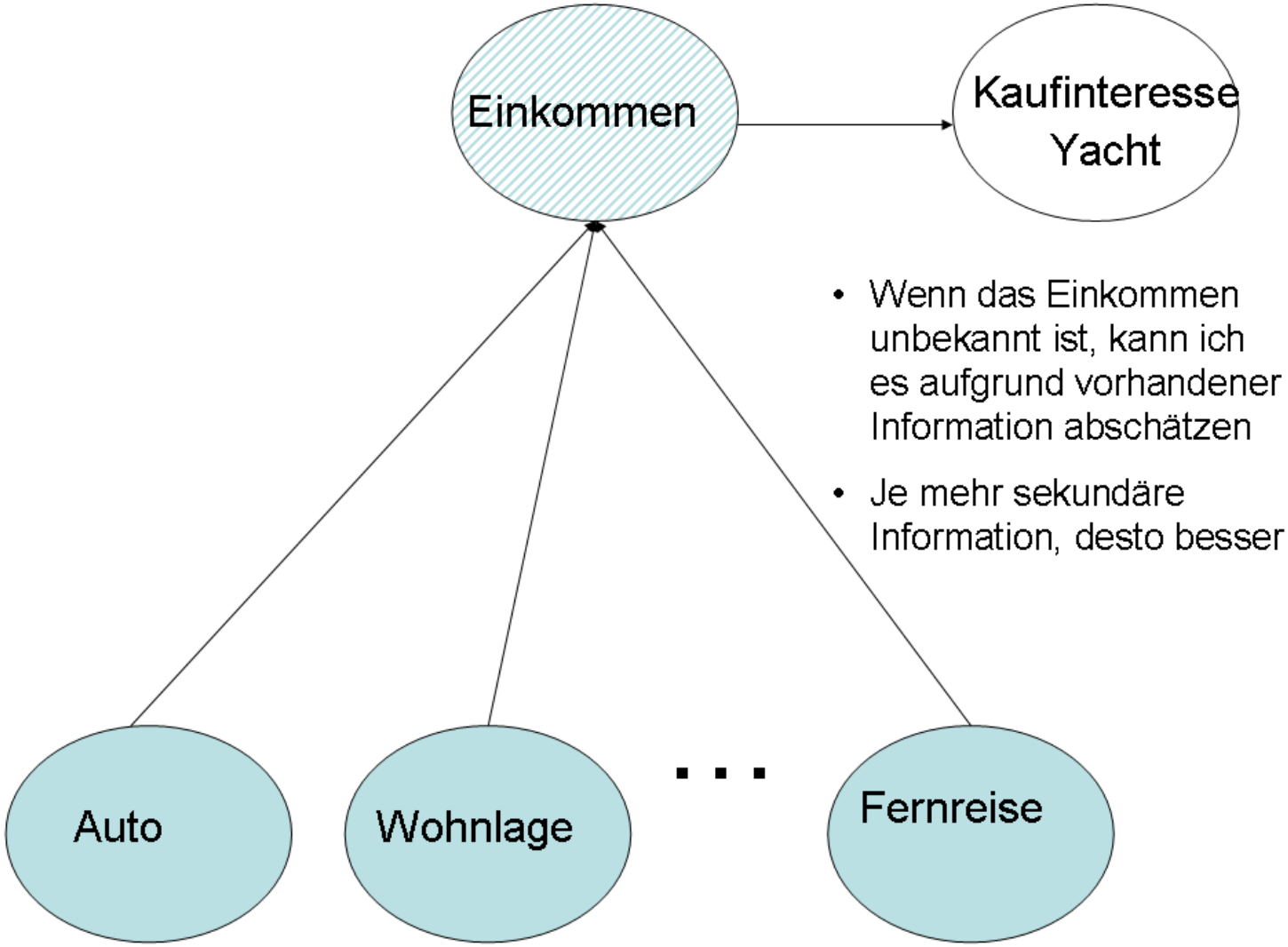
- Definiere fliegen: aus eigener Kraft, mindestens 20 m weit, mindestents 1 m hoch, mindestens einmal am Tag seines erwachsenen Lebens, ...
- Ein Objekt ist als Vogel klassifiziert; der Vogel kann fliegen,
  - Außer er ist ein Pinguin, oder ein, ...
  - Außer er ist schwer verletzt oder tot
  - Außer er ist zu alt
  - Außer er ist gestutzt worden
  - Außer er hat eine Reihe von Krankheiten
  - Außer er lebt ausschließlich in einem Stall
  - Außer mat hat ihm ein schweres Gewicht an seine Körperteile befestigt
  - Außer er lebt als Papagei in einem zu kleinen Käfig, ...

Mustererkennung

- Von allen Vögeln auf der Welt fliegen 90%
- Von allen Vögeln auf der Welt, die nicht zur Klasse flugunfähiger Arten gehören, fliegen 94%
- Von allen Vögeln auf der Welt, die nicht zur Klasse flugunfähiger Arten gehören, und die nicht von Menschen gehalten werden, fliegen 96% ...
- Grundsätzliches Probleme:
  - Komplexität des unterliegenden (deterministischen) Systems
  - Unvollständige Information
- Hieraus begründet sich der Erfolg statistisch-basierter Ansätze

Beispiel: Kaufentscheidung



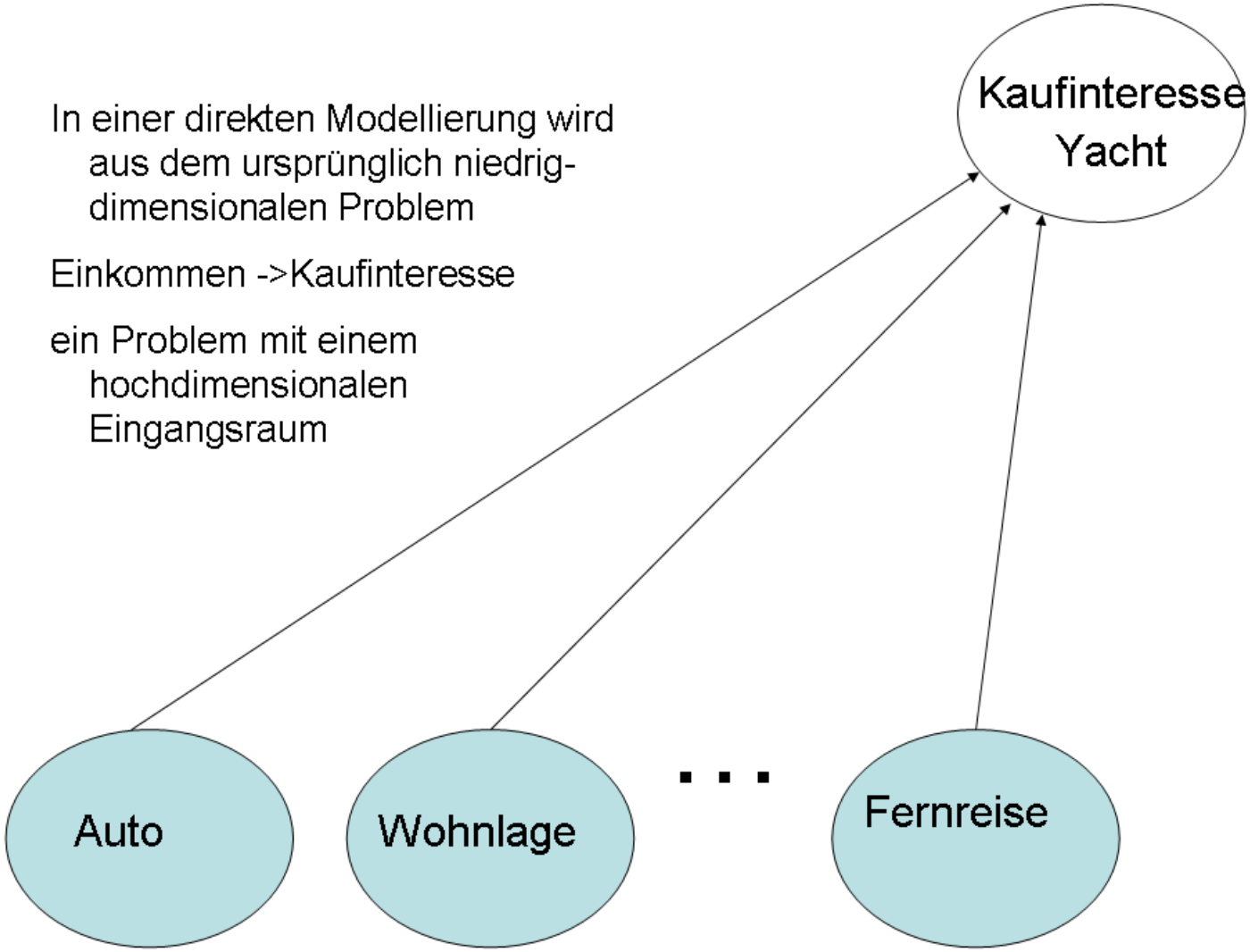


- Wenn das Einkommen unbekannt ist, kann ich es aufgrund vorhandener Information abschätzen
- Je mehr sekundäre Information, desto besser

In einer direkten Modellierung wird aus dem ursprünglich niedrig-dimensionalen Problem

Einkommen ->Kaufinteresse

ein Problem mit einem hochdimensionalen Eingangsraum



Ontologien

- Es gibt weiterhin Anstrengungen das menschliche Wissen zu formalisieren; die gegenwärtigen Anstrengungen in Richtung auf ein Ontologie-basiertes Semantisches Netz können auch in diesem Zusammenhang gesehen werden
- Cyc (vom englischen encyclopedia) ist eine Wissensdatenbank des Alltagswissen. Sie wird seit 1984 weiterentwickelt, um Anwendungen der Künstlichen Intelligenz das logische Schlußfolgern über Sachverhalte des "Gesunden Menschenverstandes" zu ermöglichen. Dabei werden alle Inhalte als logische Aussagen in der Ontologiesprache CycL formuliert, die auf der Prädikatenlogik aufbaut. Zusätzlich enthält CyC eine Inferenzmaschine zum Schlussfolgern über die gespeicherten Zusammenhänge und Plausibilitätskontrollen.
- Cyc besteht aus einer Menge von einfachen Regeln (zum Beispiel dass Wasser nass macht). Beispielsweise kann ein Programm mit Hilfe der Cyc-Ontologie aus der Aussage, dass Peter im Meer schwimmt und dass das Meer größtenteils aus Wasser besteht, schlussfolgern, dass die betreffende Person nass ist.

Vision einer Deduktiven und Induktiven Wissensbasis