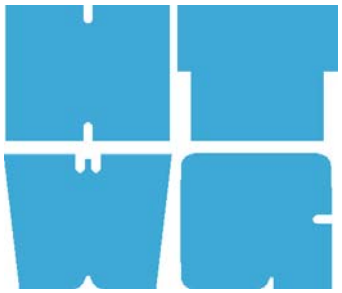


HOCHSCHULE KONSTANZ TECHNIK, WIRTSCHAFT UND GESTALTUNG
UNIVERSITY OF APPLIED SCIENCES

Mustererkennung und Klassifikation

Fakultät Informatik
Technische Informatik
Prof. Dr. Matthias Franz
mfranz@htwg-konstanz.de

Überblick

- **Grundlagen**
  - Einführung in die automatische Mustererkennung
  - Grundlagen der Wahrscheinlichkeitsrechnung
  - Statistische Beschreibung von Daten
- **Klassifikation bei bekannter Wahrscheinlichkeitsverteilung**
  - Entscheidungstheorie
  - Bayes-Klassifikator
  - Entscheidungsfunktionen bei gaußverteilten Daten
- **Überwachtes Lernen bei unbekannter Verteilung der Daten**
  - Nichtparametrische Klassifikation
  - Probleme bei hochdimensionalen Daten
  - Lineare Klassifikation, Perzeptron
  - Nicht linear trennbare Systeme
  - Nichtlineare Klassifikatoren
  - Vergleich von Klassifikatoren
- **Unüberwachtes Lernen**
  - Hauptkomponentenanalyse
  - K-Means-Clustering

Literaturverzeichnis (1)

Standardwerke zu Mustererkennung und Klassifikation

- **R.O.Duda, P.E.Hart & D.G.Stork, *Pattern Classification,***
  Wiley, 654 Seiten, 2001.

  Klassische Einführung in die Mustererkennung.Vorlesungsstoff wird gut erklärt, geht an vielen Stellen über den Vorlesungsstoff hinaus.

- **S. Haykin, *Neural Networks: a comprehensive foundation*,**
  Prentice Hall, 842 Seiten, 1998.

  Ebenfalls ein gut geschriebener Klassiker. Sehr umfangreich, Fokus auf neuronale Netze (ist in diesem Kurs eher Nebensache), aber fast alles aus der Vorlesung wird auch erklärt.

- **T. Hastie, R. Tibshirani & J. Friedman, *The elements of statistical learning*,**
  Springer, 533 Seiten, 2001.

  Standardwerk, geschrieben aus Sicht der Statistik. Schwerer zugänglich, aber lohnend. Geht ebenfalls über den Vorlesungsstoff hinaus.

- **C.M.Bishop, *Pattern recognition and machine learning;***
  Springer, 738 Seiten, 2005.

  Sehr umfangreich, auf dem neuesten Stand in Richtung maschinelles Lernen.

- **R.Rojas, *Theorie der neuronalen Netze;***
  Springer, ca. 300 Seiten, 1996.

  Eines der wenigen deutschsprachigen Bücher zu diesem Thema, deckt nur Teile der Vorlesung ab.

Literaturverzeichnis (2)

Weiterführende Literatur

- **B. Schölkopf & A. Smola, *Learning with kernels,***
  MIT Press, 644 Seiten, 2002

  behandelt im vollen Umfang sogenannte Kernelmethoden (z.B. Supportvektormaschinen), die momentan einige der leistungsfähigsten Algorithmen stellen.

- **D. MacKay, *Information theory, inference and learning algorithms;***
  Cambridge University , 550 Seiten, 2003.

  behandelt maschinelles Lernen mit probabilistischen Methoden im informationstheoretischen Kontext, sehr gut geschrieben.

Überblick

- **Grundlagen**
  - Einführung in die automatische Mustererkennung
  - Grundlagen der Wahrscheinlichkeitsrechnung
- **Klassifikation bei bekannter Wahrscheinlichkeitsverteilung**
  - Entscheidungstheorie
  - Bayes-Klassifikator
  - Entscheidungsfunktionen bei gaußverteilten Daten
- **Unüberwachtes Lernen bei unbekannter Verteilung der Daten**
  - Nichtparametrische Klassifikation
  - Probleme bei hochdimensionalen Daten
  - Lineare Klassifikation, Perzeptron
  - Nicht linear trennbare Systeme
  - Nichtlineare Klassifikatoren
  - Vergleich von Klassifikatoren
- **Unüberwachtes Lernen**
  - Hauptkomponentenanalyse
  - K-Means-Clustering
  - Agglomeratives Clustern

Mustererkennung und Klassifikation: Begriffe

- Klassifizierung oder Klassifikation (von lat. classis, „Klasse“, und facere, „machen“) nennt man das Zusammenfassen von Objekten zu Klassen.

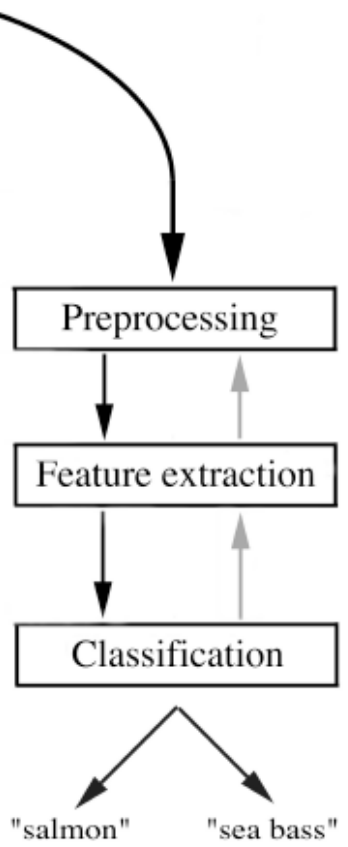
- Kategorisierung ist der Prozess, bei dem unterschiedliche Objekte als gleichwertig betrachtet werden. Sie ist ein fundamentaler kognitiver Vorgang bei Wahrnehmung und Verständnis von Konzepten und Objekten, beim Entscheidungsprozess und bei allen Arten von Interaktion mit der Umwelt.

- Kategorisierung und Klassifizierung bedeuten eigentlich dasselbe, „Klassifizierung“ bezieht sich jedoch auf mathematische oder technische Prozesse, „Kategorisierung“ auf Psychologie und Bedeutung.

- Kategorisierung bzw. Klassifizierung sind Voraussetzung für Abstraktion und Begriffsbildung und damit letztlich für Intelligenz.

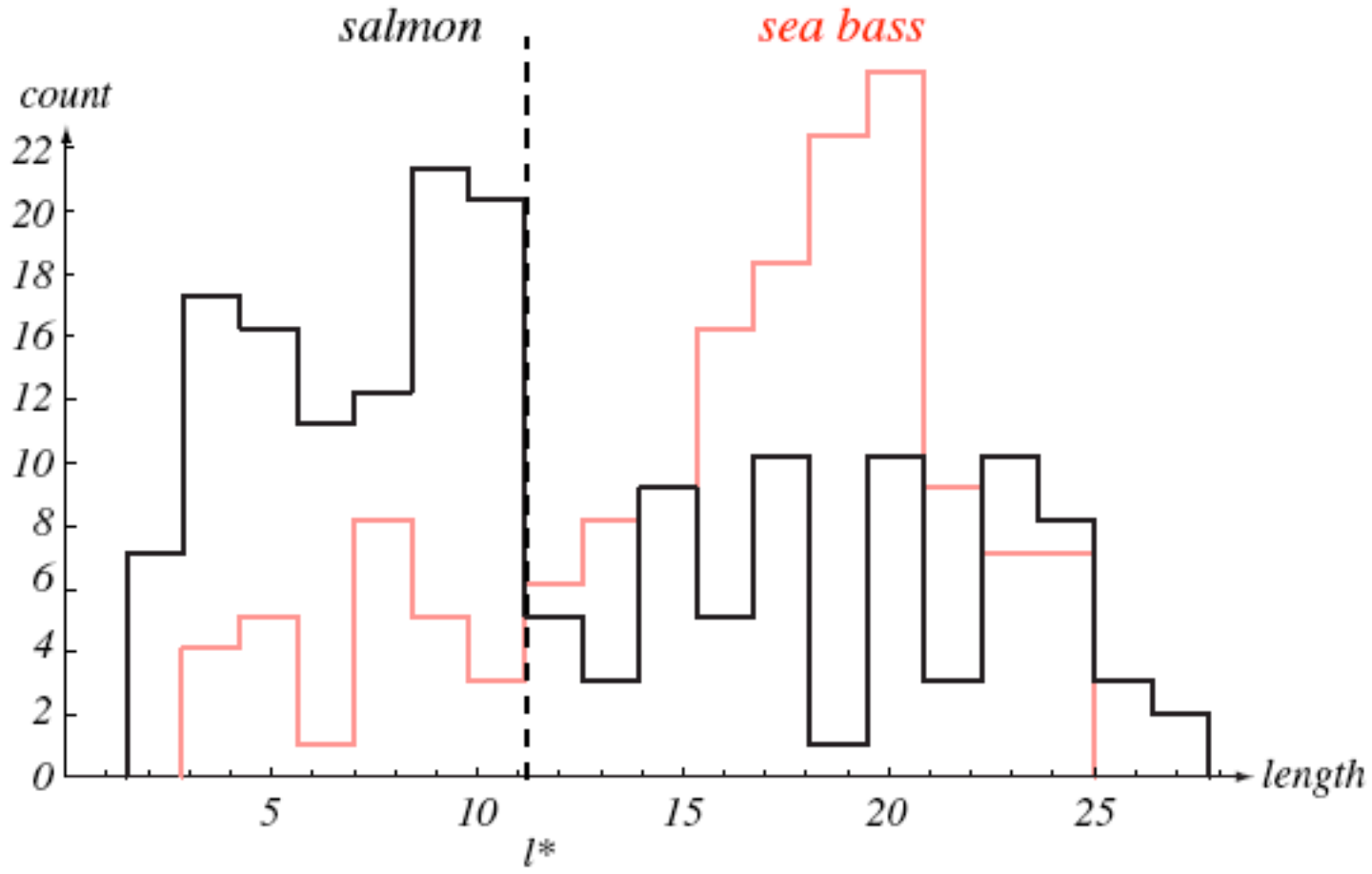
- Mustererkennung bezeichnet ein Verfahren, gemessene Signalen automatisch in Kategorien einzuordnen, d.h sie automatisch zu klassifizieren.

Beispiel für eine Mustererkennungsaufgabe (1)



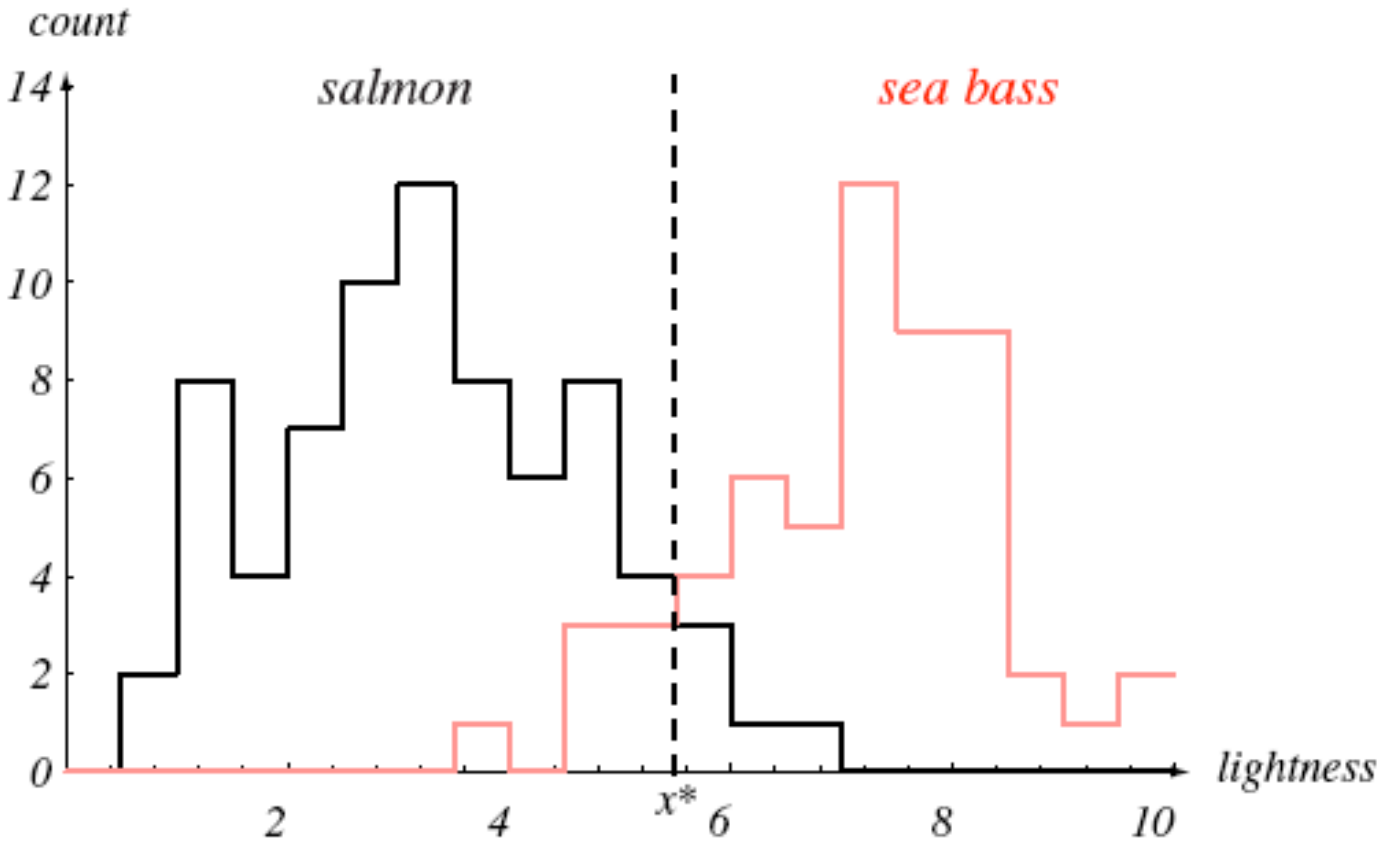
[aus Duda et al., 2001]

Beispiel für eine Mustererkennungsaufgabe (2)



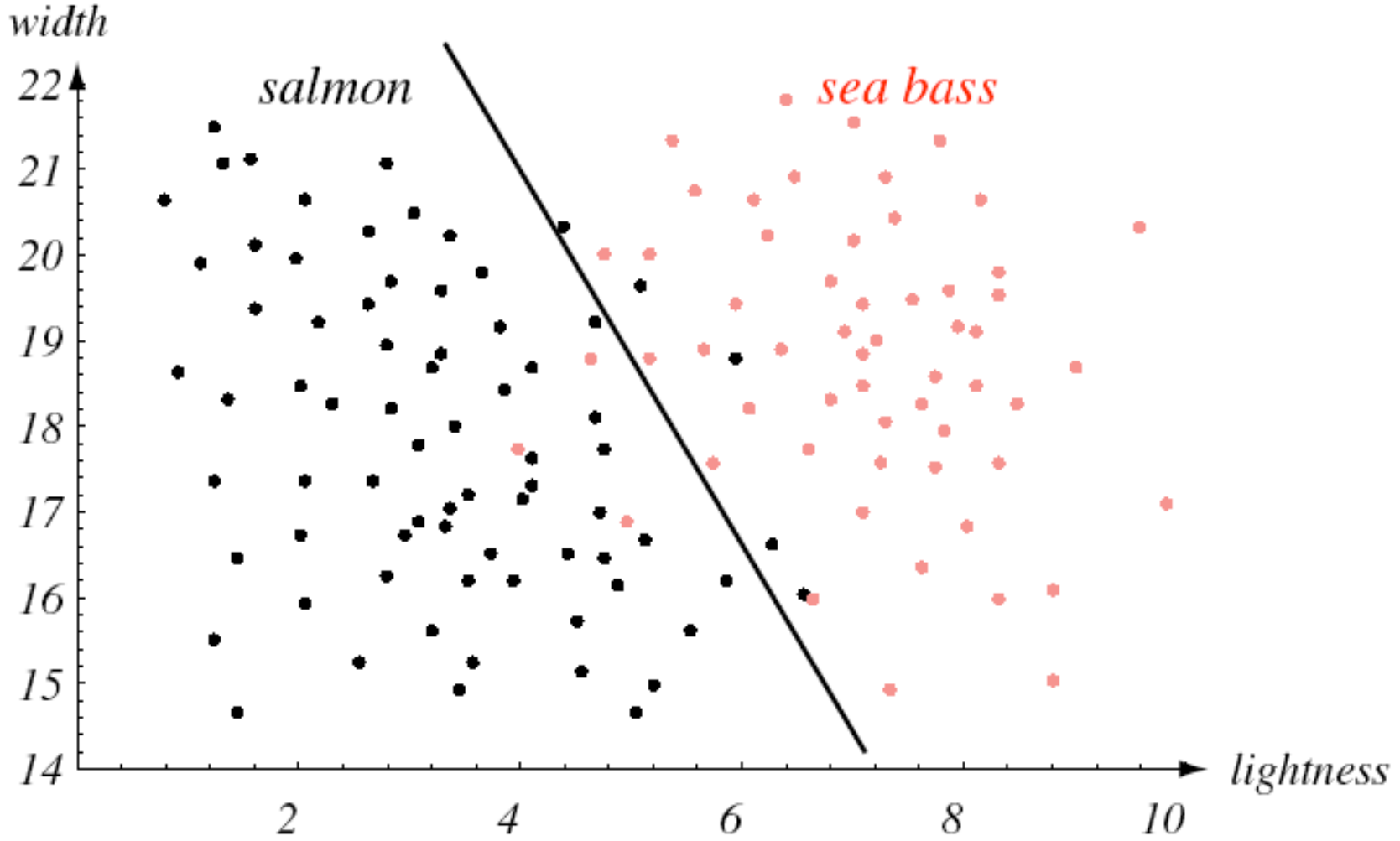
[aus Duda et al., 2001]

Beispiel für eine Mustererkennungsaufgabe (3)



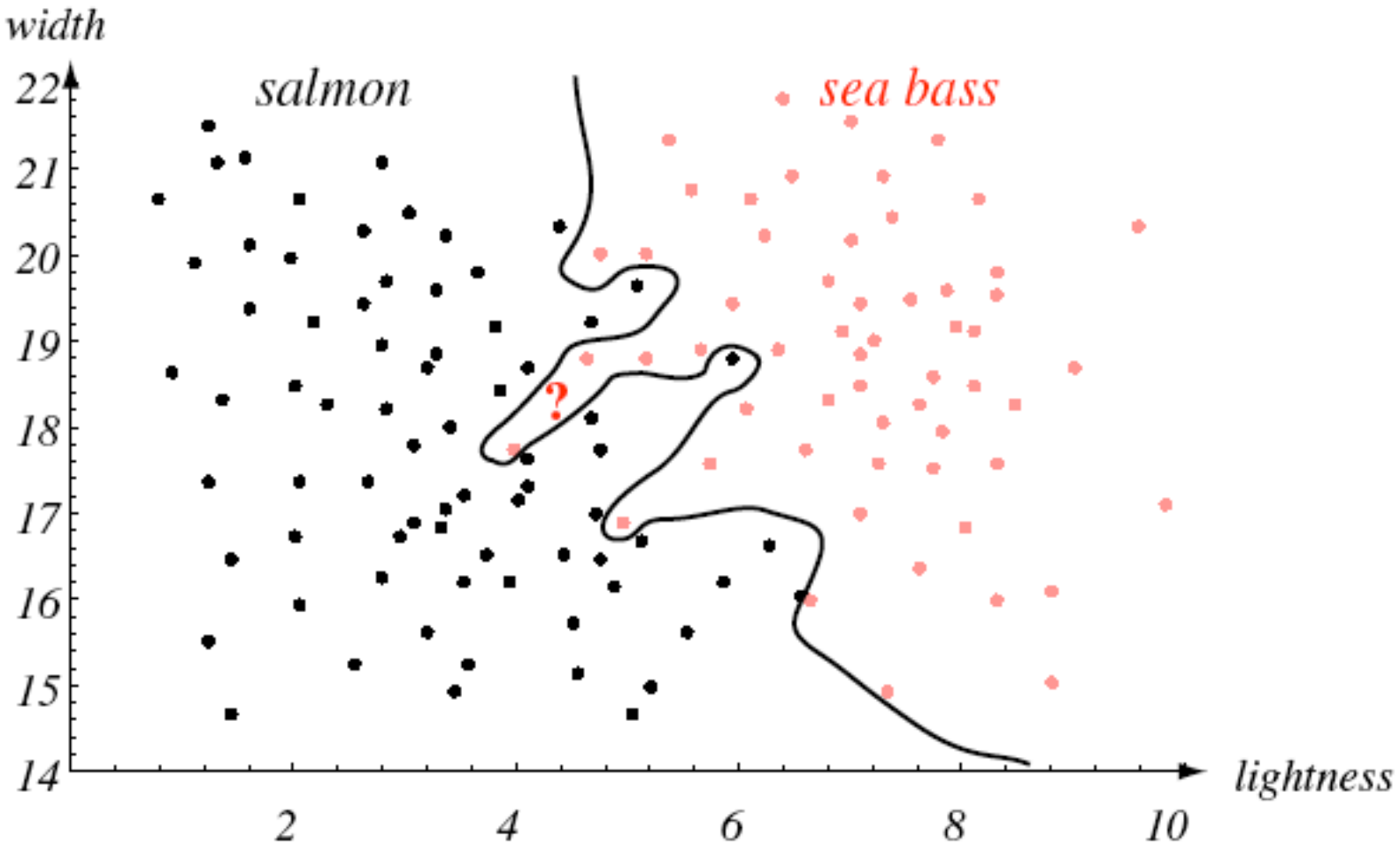
[aus Duda et al., 2001]

Beispiel für eine Mustererkennungsaufgabe (4)



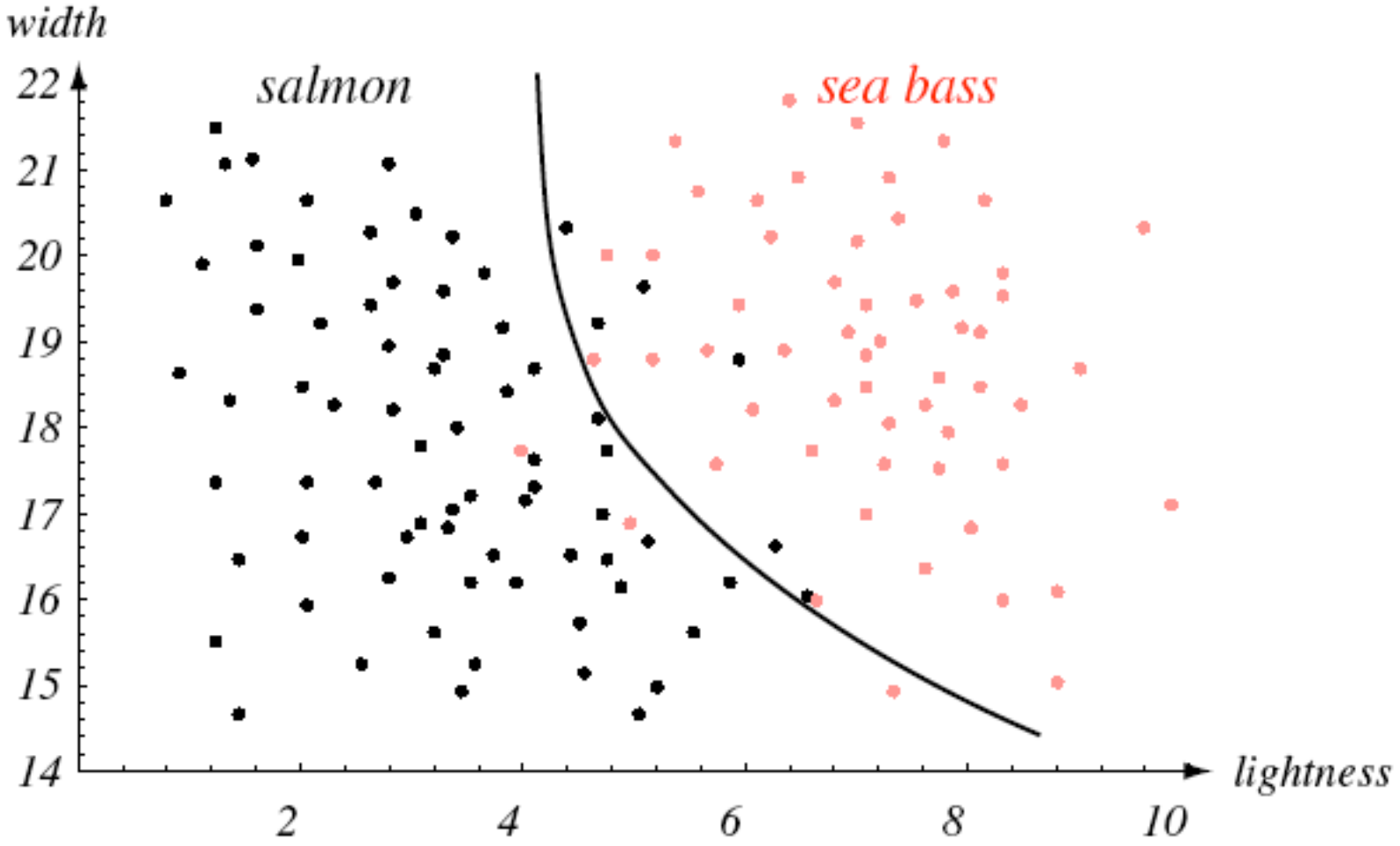
[aus Duda et al., 2001]

Beispiel für eine Mustererkennungsaufgabe (5)



[aus Duda et al., 2001]

Beispiel für eine Mustererkennungsaufgabe (6)



[aus Duda et al., 2001]

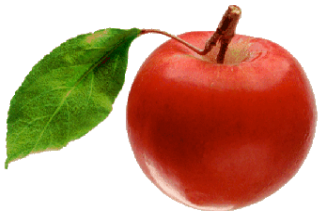
Zentrale Begriffe in der statistischen Mustererkennung

- Merkmal
- Vorverarbeitung
- Segmentierung
- Merkmalsextraktion
- Trainingsbeispiele
- Entscheidungskosten, Kostenfunktion
- Entscheidungstheorie
- Klassengrenze
- Generalisierung

Drei grundlegende Ansätze in der Mustererkennung

1. Syntaktisch: Dinge werden so durch Folgen von Symbolen beschrieben, dass Objekte der gleichen Kategorie die selben Beschreibungen aufweisen. Das Problem der Mustererkennung stellt sich in diesem Fall als Suche nach einer formalen Grammatik dar, also nach einer Menge von Symbolen und Regeln zum Zusammenfügen derselben.
2. Statistisch: Ziel ist es hier, ein Objekt in die Kategorie mit der höchsten Wahrscheinlichkeit einzusortieren. Statt Merkmale nach vorgefertigten Regeln auszuwerten, werden sie hier einfach als Zahlenwerte gemessen und in einem Merkmalsvektor zusammengefasst. Eine mathematische Funktion ordnet dann jedem denkbaren Merkmalsvektor eindeutig eine Kategorie zu.
3. Strukturell: verbindet verschiedene syntaktische und/oder statistische Verfahren zu einem einzigen neuen Verfahren. Die grundlegende Merkmalserkennung wird dabei allgemeinen statistischen Verfahren überlassen, während übergeordnete Inferenzverfahren Spezialwissen über das Sachgebiet einbringen.

Syntaktische Mustererkennung



rot, rund

gelb, länglich

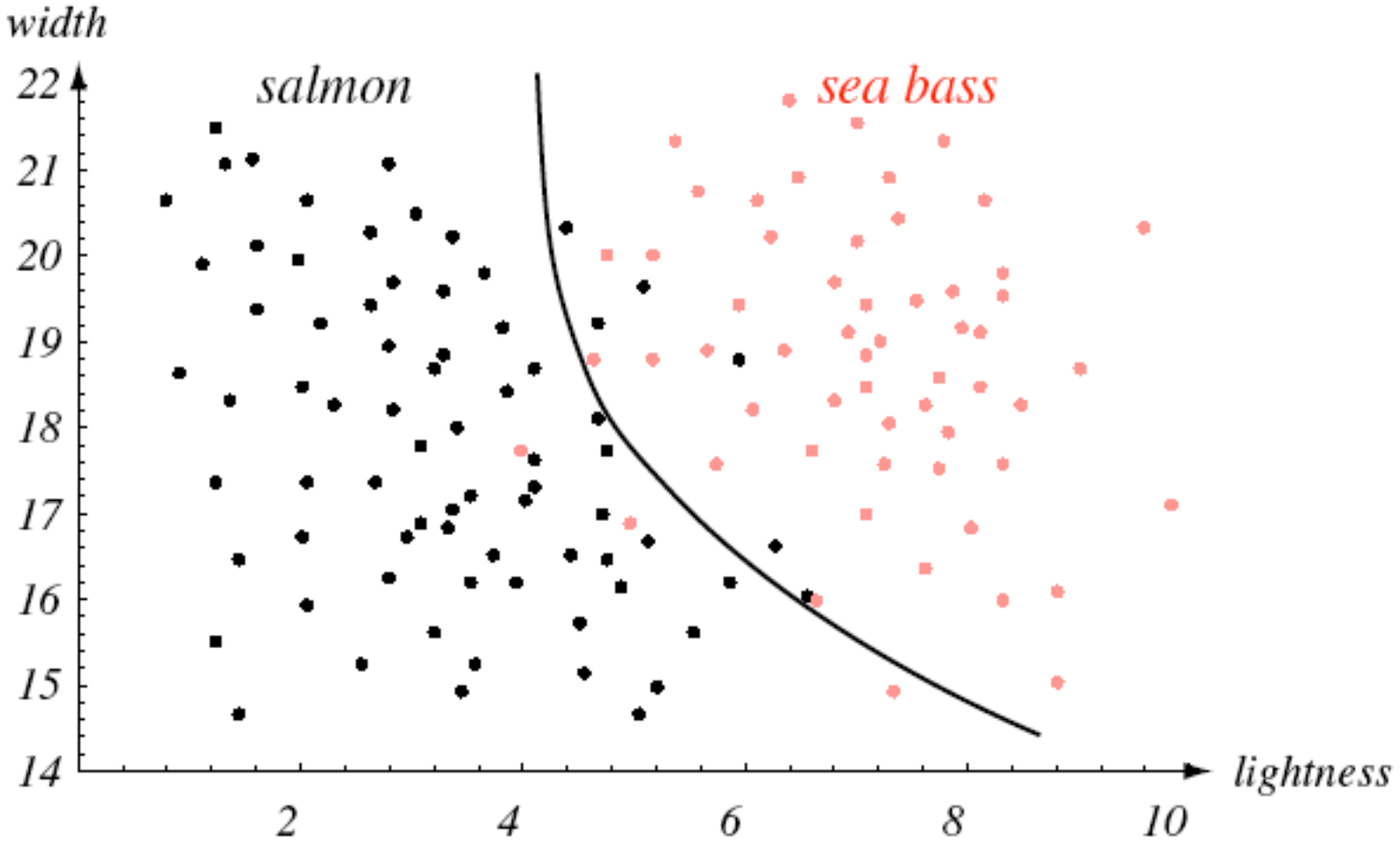
Syntaktischer Mustererkenner:

Regel 1: „Wenn gelb und länglich, dann Banane“

Regel 2: „Wenn rot und rund, dann Apfel“

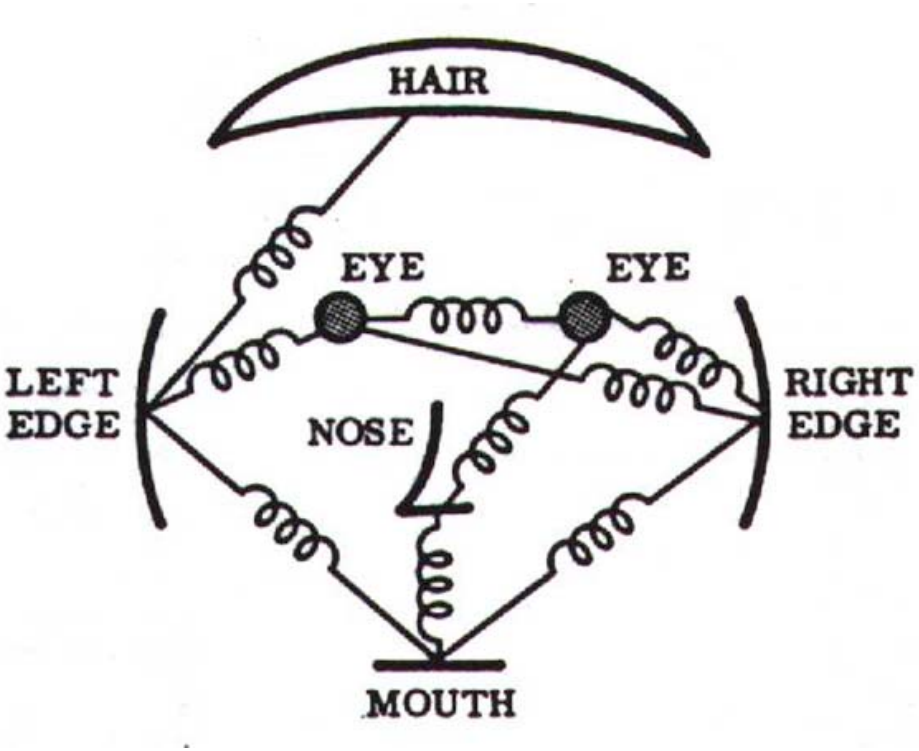
Erfordert eindeutige Attribute

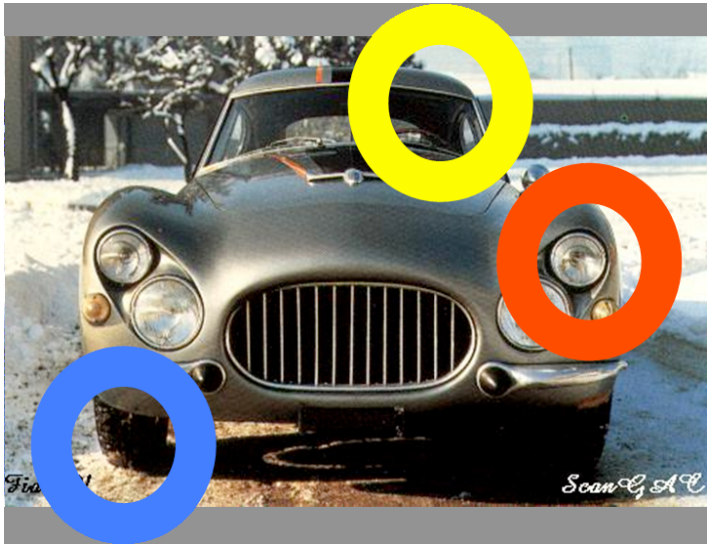
Statistische Mustererkennung



[aus Duda et al., 2001]

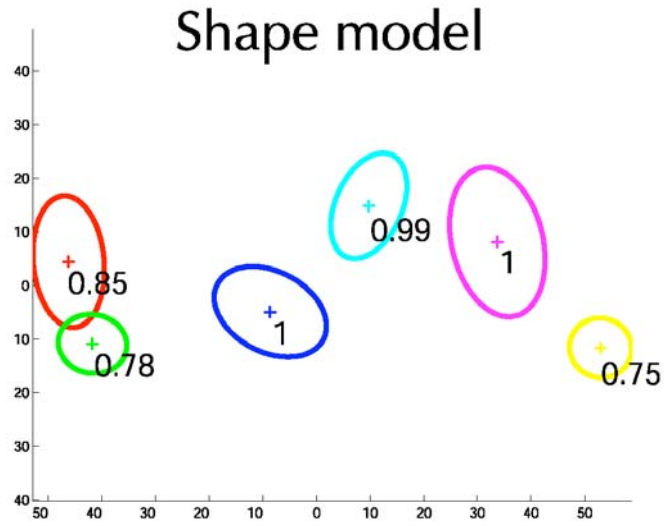
Strukturelle Mustererkennung





-Teile des Objekts werden mit statistischen Techniken detektiert.
-Struktur, d.h. Konstellation der Teile wird zusätzlich gelernt.

Strukturelle Mustererkennung (2)



Abgrenzung zu anderen Gebieten

- Bei der Klassifikation werden den Inputs diskrete Outputs zugewiesen (d.h. die jeweilige Bezeichnung der Klasse). Sind die Outputs dagegen kontinuierliche Größen, spricht man von Regression, bei ordinalen Outputs von Rangfolgenbestimmung (Ranking).

- Künstliche Intelligenz: Teilgebiet der Informatik, das mit der Automatisierung intelligenten Verhaltens befasst ist. Mustererkennung als grundlegende Voraussetzung für intelligentes Verhalten ist ein zentraler Bestandteil der KI.

- Data mining: explorative Datenanalyse zur Entdeckung von Mustern in (meist großen) Datenbeständen. Methoden der Mustererkennung werden als zentraler Bestandteil im Data mining eingesetzt.

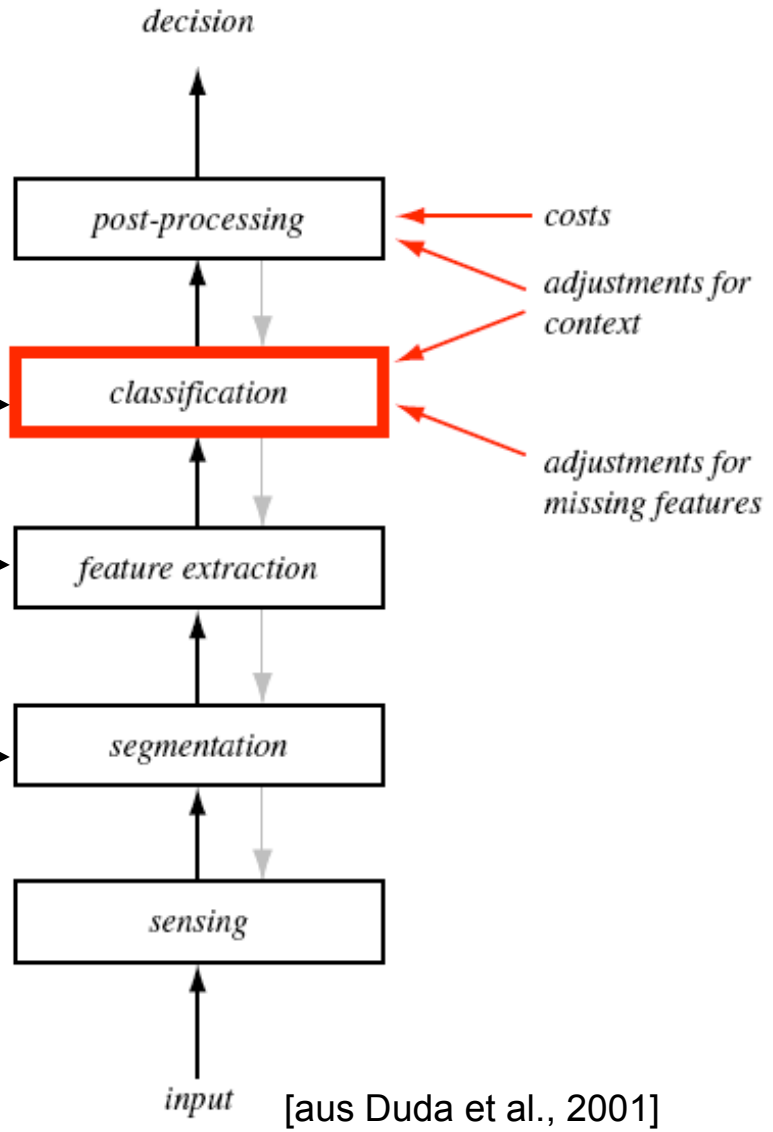
- Maschinelles Lernen: Oberbegriff für die „künstliche“ Generierung von Wissen aus Erfahrung. Damit beinhaltet maschinelles Lernen nicht nur die Mustererkennung, sondern auch Techniken wie Regression, Interpolation und Dichteschätzung

Schema eines Mustererkennungsystems

Zentrales Thema der Vorlesung

Merkmale sollten invariant gegenüber unwichtigenTransformationen sein (z.B. Verschiebung, Drehung, Skalierung)

Segmentierung ist eines der schwierigsten Probleme in der Mustererkennung

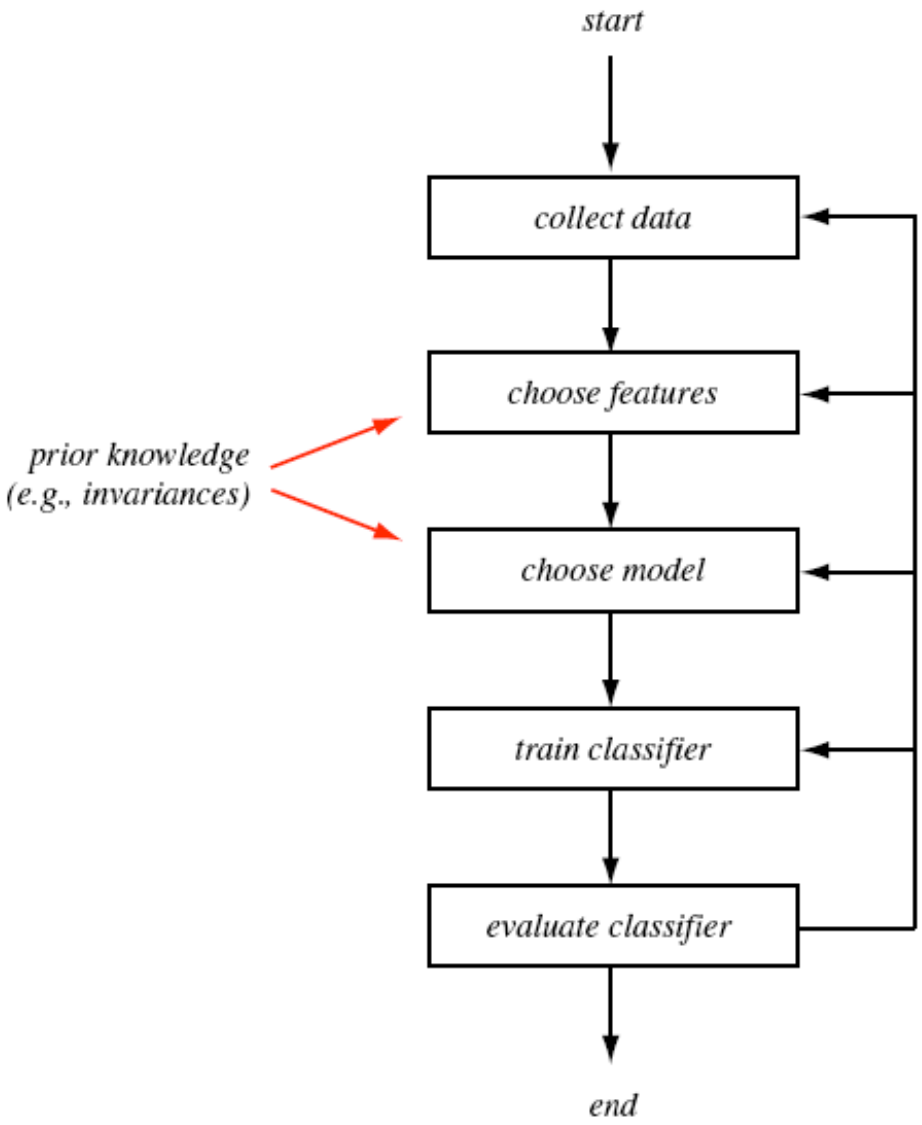


[aus Duda et al., 2001]

Invariante Merkmale?

...

Designzyklus eines Mustererkennungssystems



Hauptprobleme:

1. Überanpassung An Trainingsbeispiele (Overfitting)

2. Begrenzte Rechenkapazität

[aus Duda et al., 2001]

Lernen anhand von Beispielen

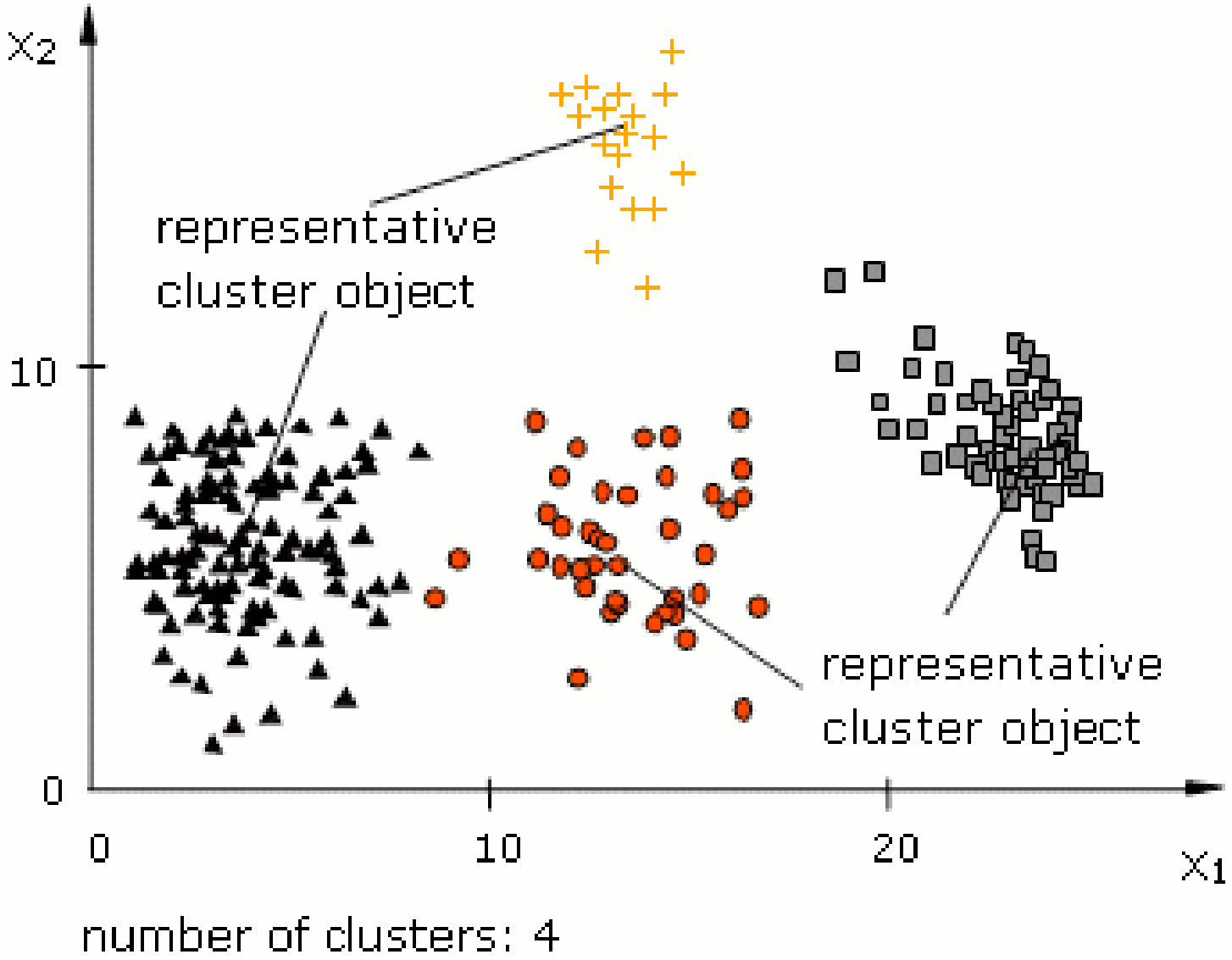
• Überwachtes Lernen: Zu jedem Trainingsbeispiel ist die Klassenzugehörigkeit bekannt

• Unüberwachtes Lernen: Klassen müssen über „natürliche Gruppen“ gefunden werden (Clustering)

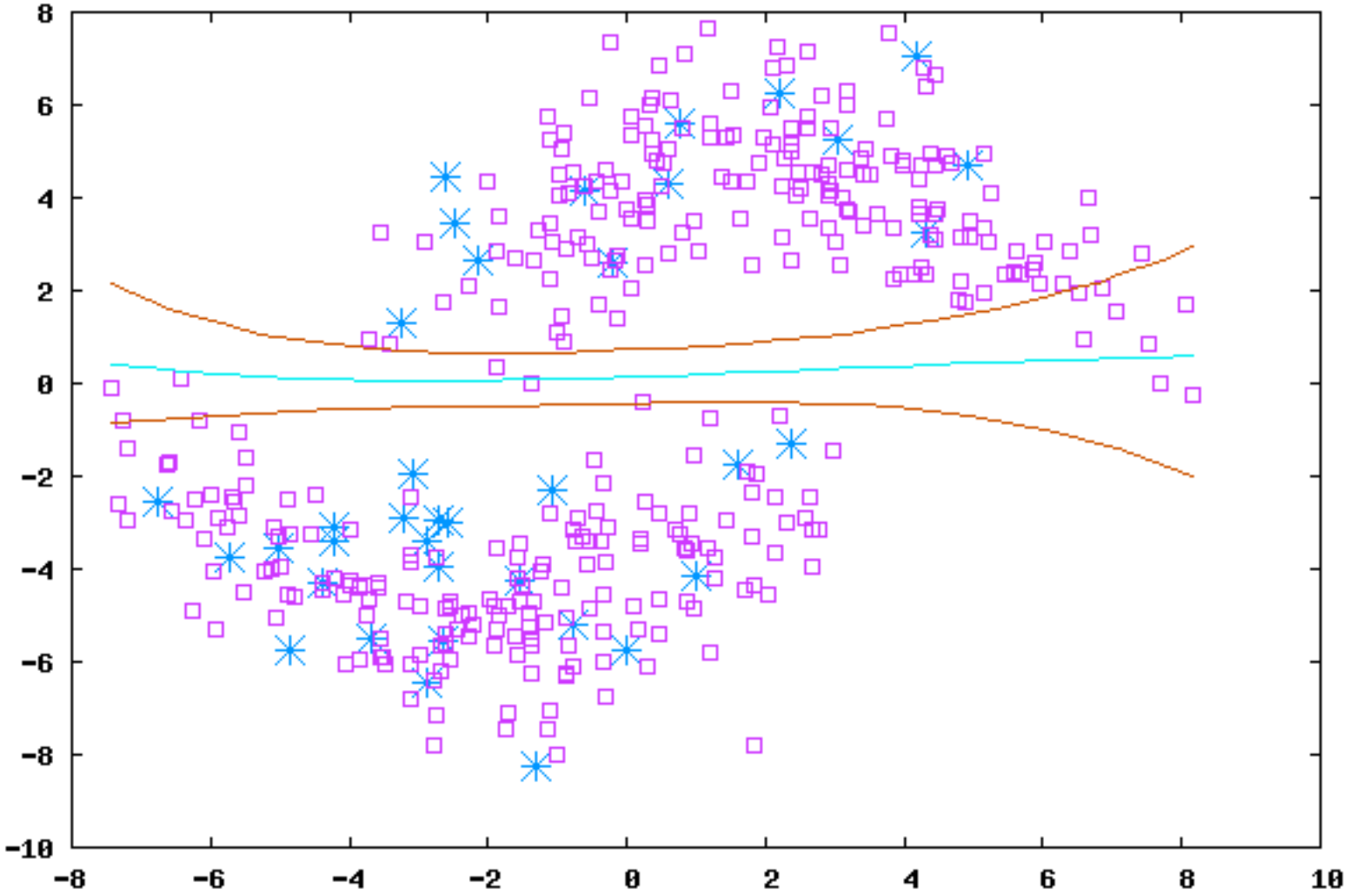
• Halbüberwachtes Lernen: die Klassenzugehörigkeit ist nur für ein paar Trainingsbeispiele bekannt, die anderen müssen sinnvoll eingeordnet werden

• Bekräftigungslernen (Reinforcement Learning): Aktionen des Mustererkenners werden nur mit falsch oder richtig bewertet, die Klassenzugehörigkeit wird nicht mitgeteilt.

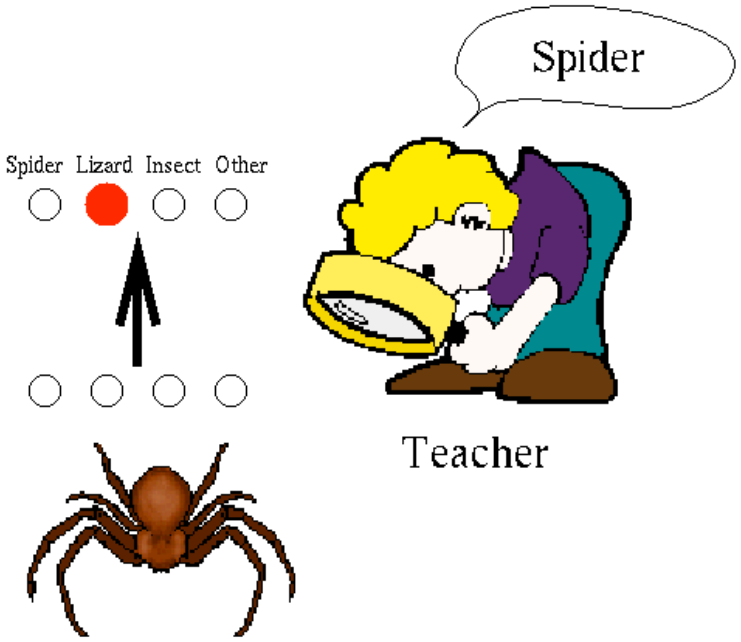
Unüberwachtes Lernen



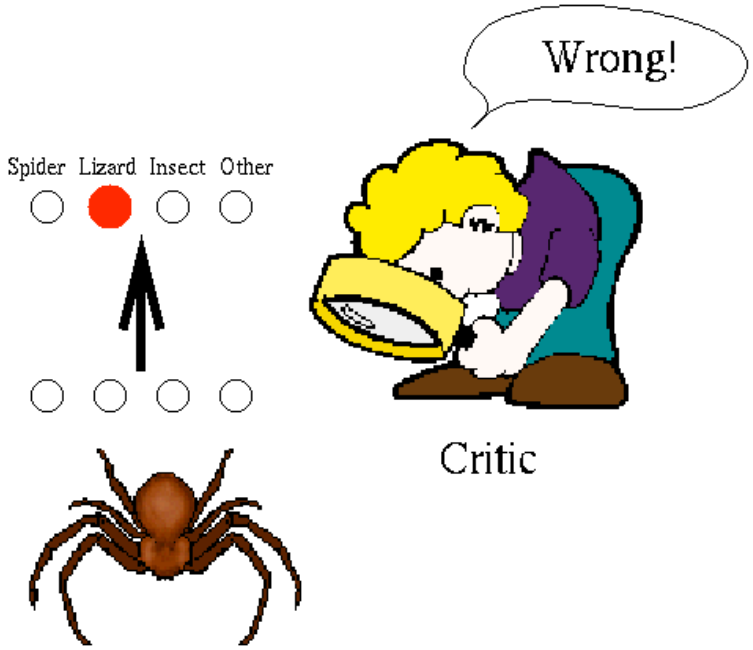
Halbüberwachtes Lernen



Überwachtes und Bekräftigungslernen



Überwachtes Lernen



Reinforcement-Lernen

[S. Dennis, 1997]