XITASO

WIE MAN EINE KI IN DIE IRRE FÜHRT

Anton Winschel

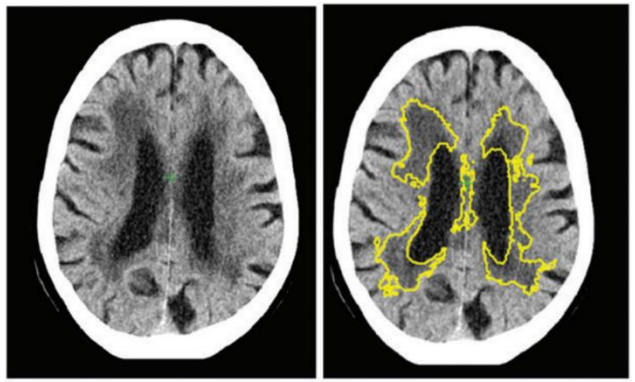
7. November 2019 – IT Kongress Neu-Ulm | Ulm

KÜNSTLICHE INTELLIGENZ – ANWENDUNGEN

XITASO

Autonomes Fahren

Medizinische Diagnosen



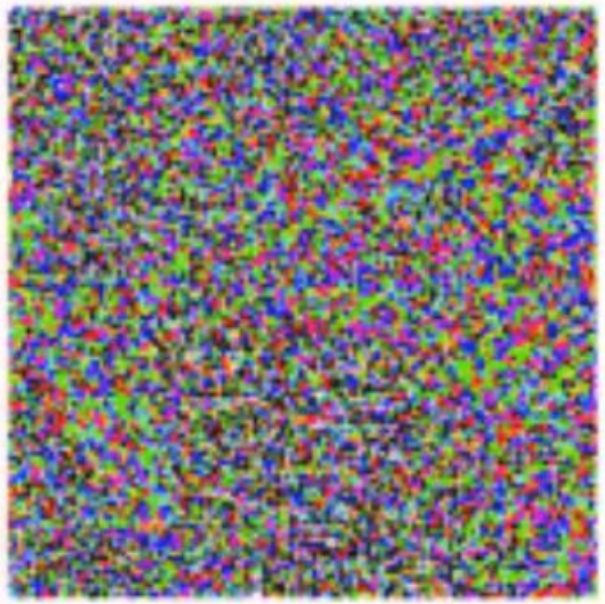
Industrie

- Wie kann man eine KI in die Irre führen?
- Beispiel: Klassifikation
  - Funktionsweise
  - Irreführung
- Wie können KI-Systeme robuster gestaltet werden?

Quelle: medium.com/intro-to-artificial-intelligence
pubs.rsna.org/doi/pdf/10.1148/radiol.2018171567
www.cross-compass.com/article2

IRREFÜHRUNG: OBJEKTERKENNUNG

Ente



× 0.07

Pferd

Quelle: arXiv:1811.07018v1

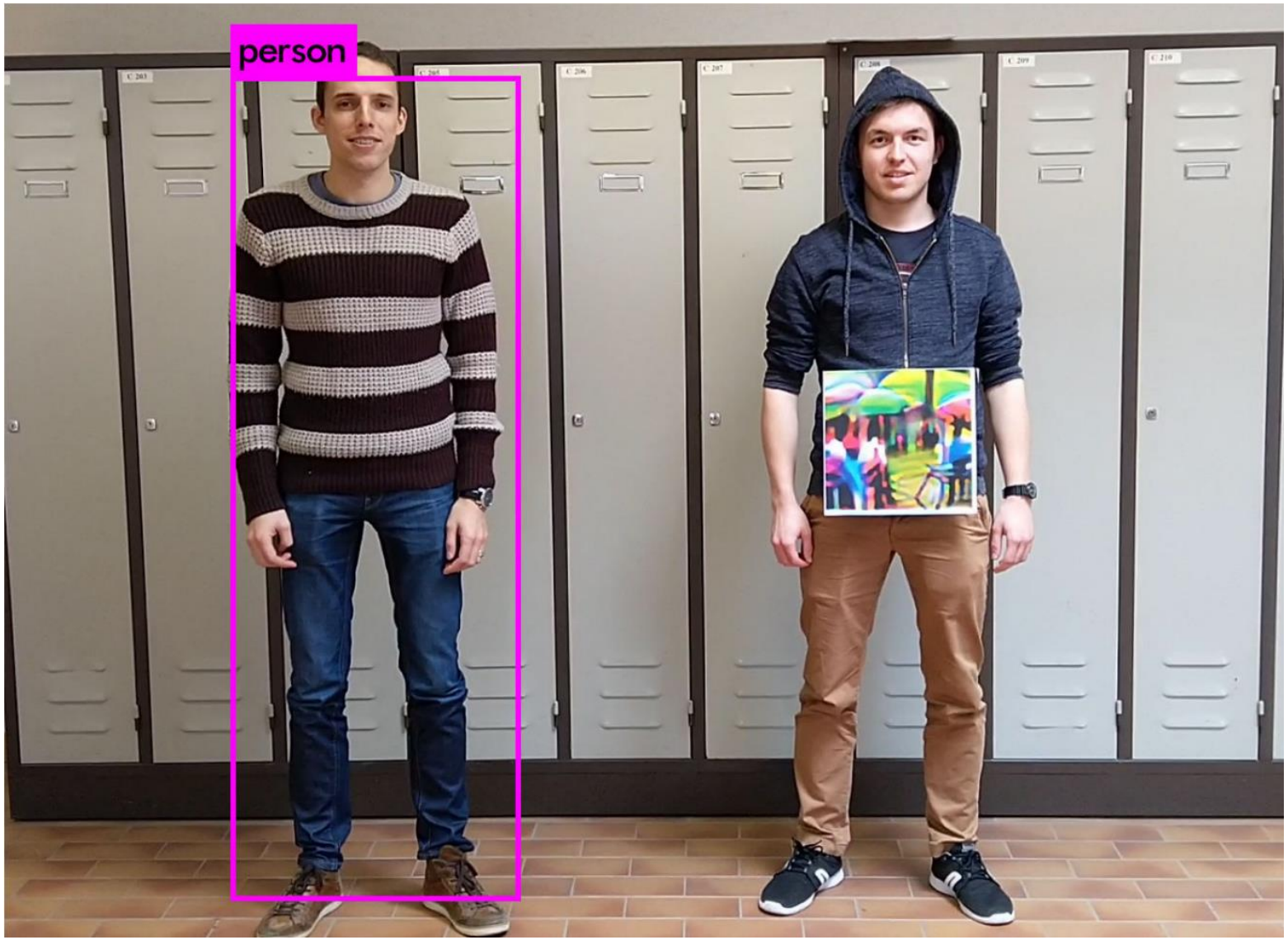
IRREFÜHRUNG: SPRACHERKENNUNG

“Wie geht es dir?“ $\times$ 0.01 “Türe öffnen“

Quelle: arXiv:1811.07018v1

IRREFÜHRUNG: ERKENNUNG VON PERSONEN

XITASO



“Unsichtbarkeits-T-Shirt”

Quelle: arXiv:1904.08653v1

cloakwear.co

IRREFÜHRUNG: AUTONOMES FAHREN

Limit 60

Limit 45

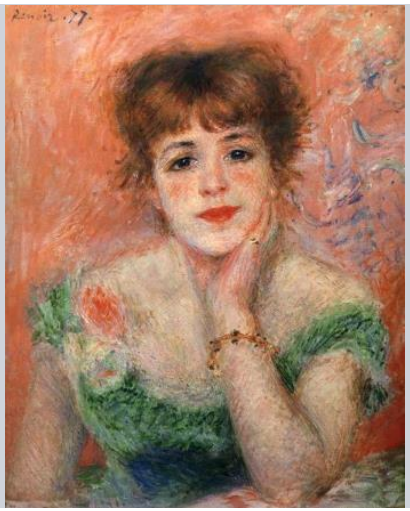
Quelle: arXiv:1707.08945v5

sites.nlsde.buaa.edu.cn/~xlliu/aaai19.pdf

KLASSIFIKATION – KUNSTRICHTUNG ERKENNEN

XITASO



. . .

Training

Klassifikator

Expressionismus?

Impressionismus?

Quelle: wikiart.org

KLASSIFIKATION – GEOMETRISCHE SICHT

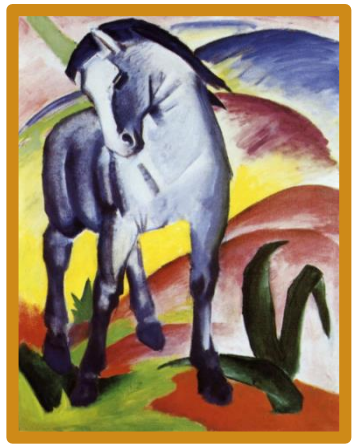
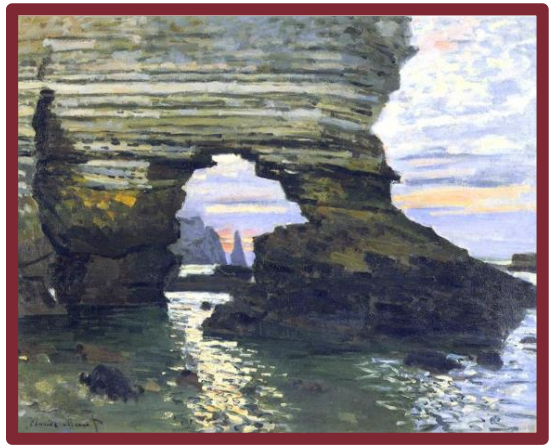
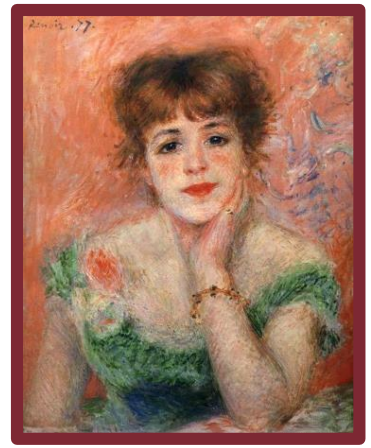
Jedes Bild wird als Punkt in einem hochdimensionalen Raum abgebildet

Z.B.

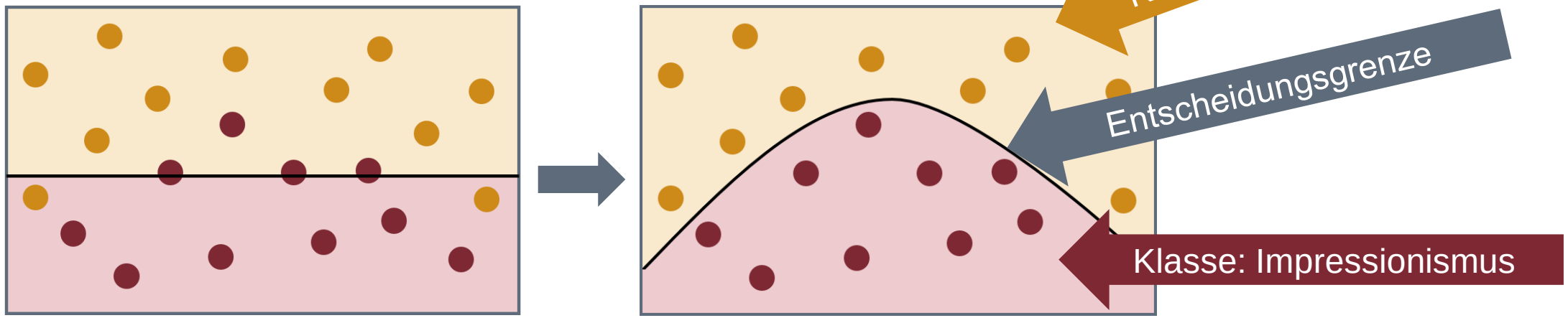
$500px \times 500px \times 3\,(rgb)$
≈ 700.000 dim

Quelle: wikiart.org

KLASSIFIKATION – GEOMETRISCHE SICHT

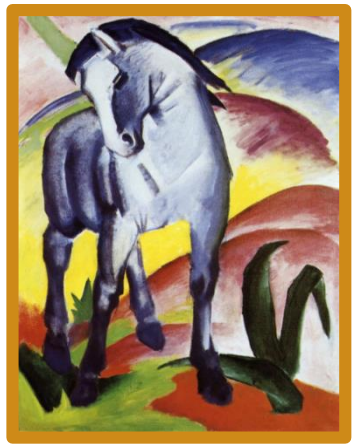
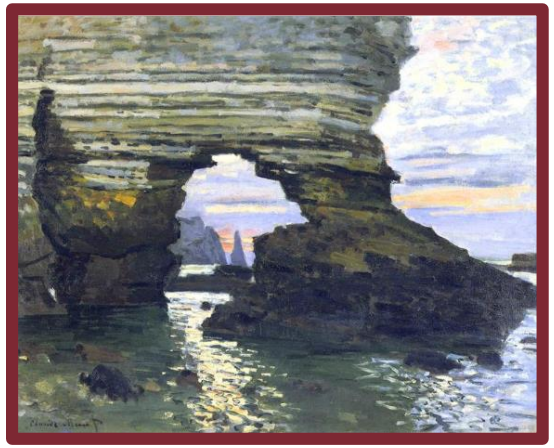
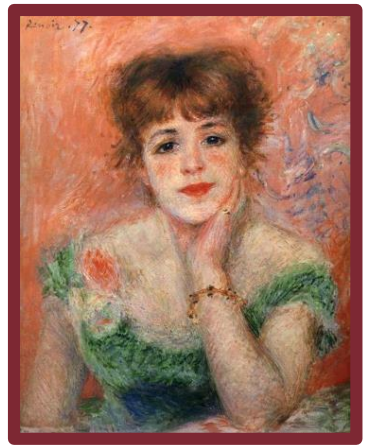
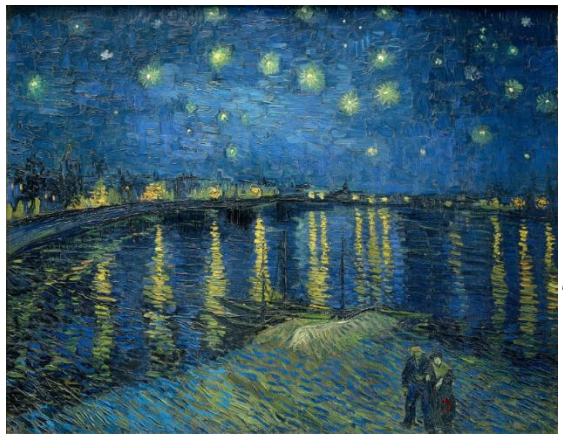
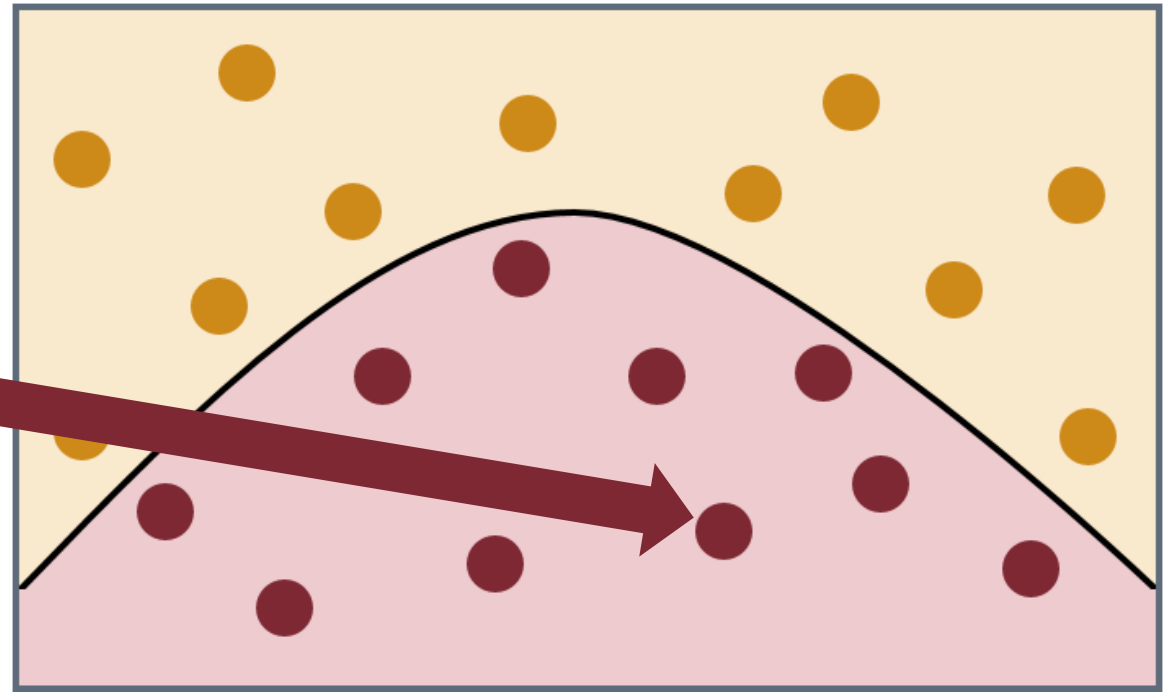





Beim Training wird die Entscheidungsgrenze schrittweise angepasst



Quelle: wikiart.org

KLASSIFIKATION – GEOMETRISCHE SICHT

Mit der Entscheidungsgrenze wird eine Klassifikation durchgeführt

Quelle: wikiart.org

KLASSIFIKATION – FORMALE SICHT

XITASO

$$f(x, o) + \epsilon = y$$

- f: Abbildung
- x: Eingabedaten
- o: Parameter der Abbildung
- ϵ: Fehler der Abbildung
- y: Ausgabedaten

$$f(\ \text{  }\ , o) + \epsilon = \text{Expressionismus}$$

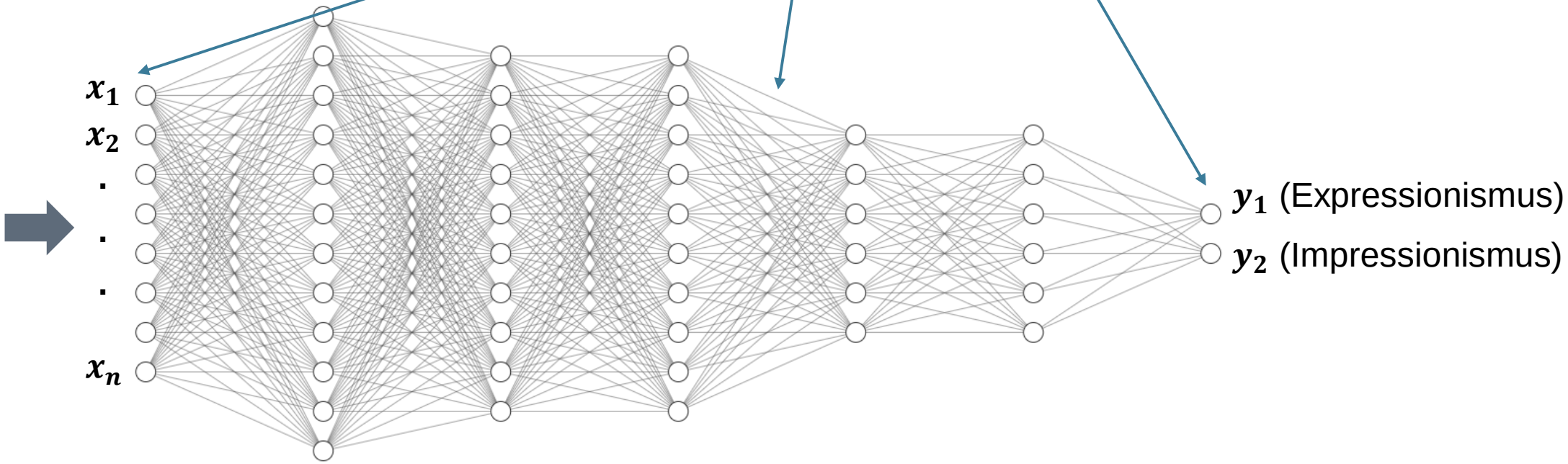
Ziel:

1. Finde o, sodass ϵ minimal ist (Optimierungsproblem)
2. f soll auf unbekannte Daten generalisieren

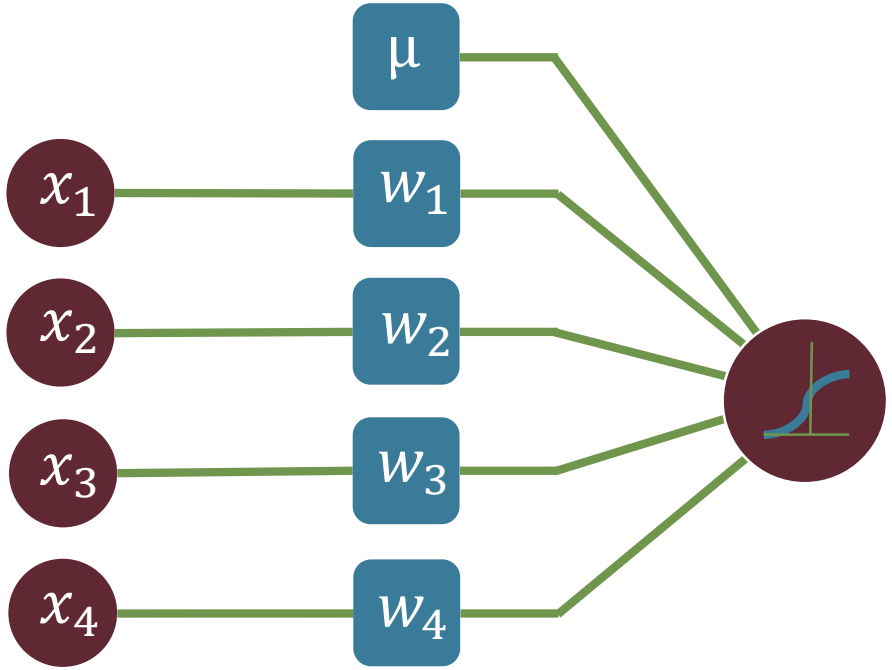
KLASSIFIKATION – ANALYTISCHE SICHT

Beispiel: f = Neuronales Netz

$$f(x, o) + \epsilon = y$$



KLASSIFIKATION – ANALYTISCHE SICHT



$$\sum_i x_i w_i - \mu = y$$

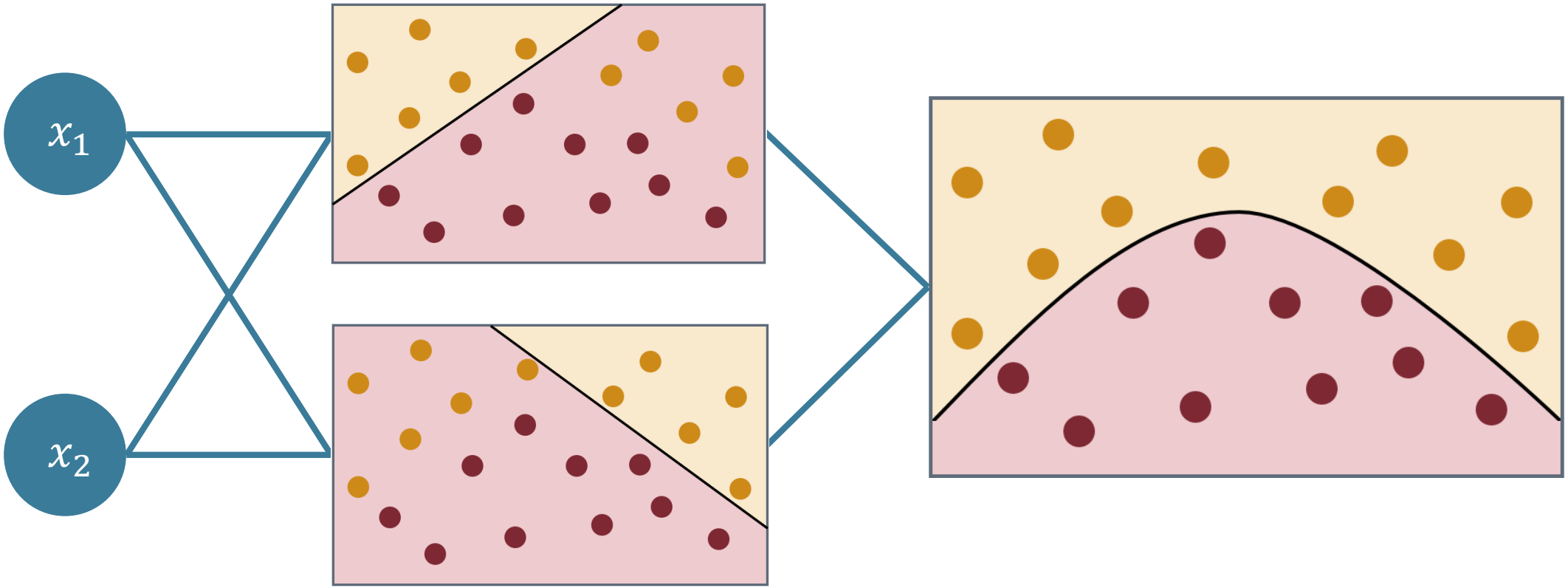
Entscheidungsgrenze

Jedes Neuron berechnet ein Skalarprodukt der Vektoren $\boldsymbol{x}$ und $\boldsymbol{w}$

KLASSIFIKATION – GEOMETRISCHE SICHT

XITASO

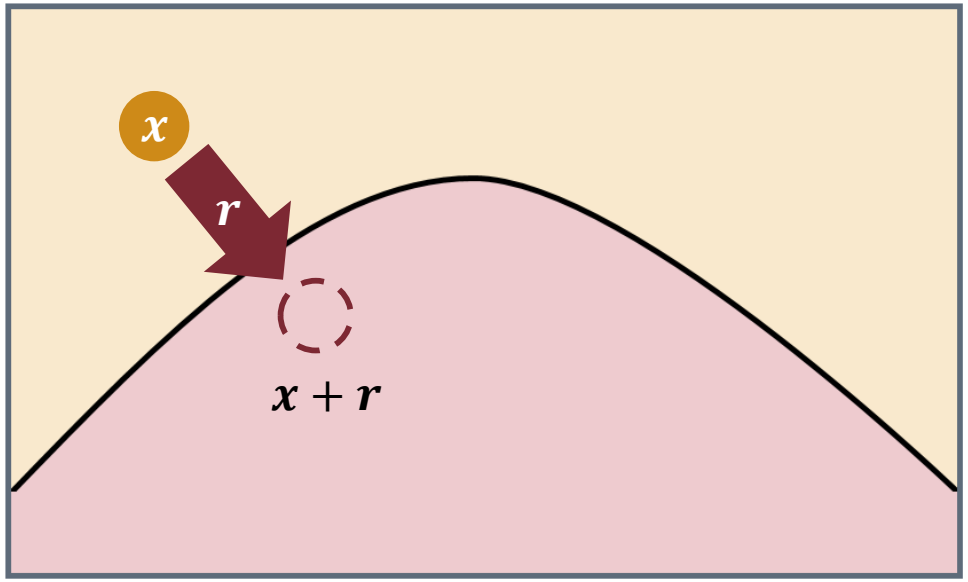
- Jedes Neuron teilt den Eingangsraum in 2 Teile
- Jedes Neuron ist ein binärer linearer Klassifikator
- Ein neuronales Netz mit nur einer Schicht kann jede Funktion beliebig genau approximieren



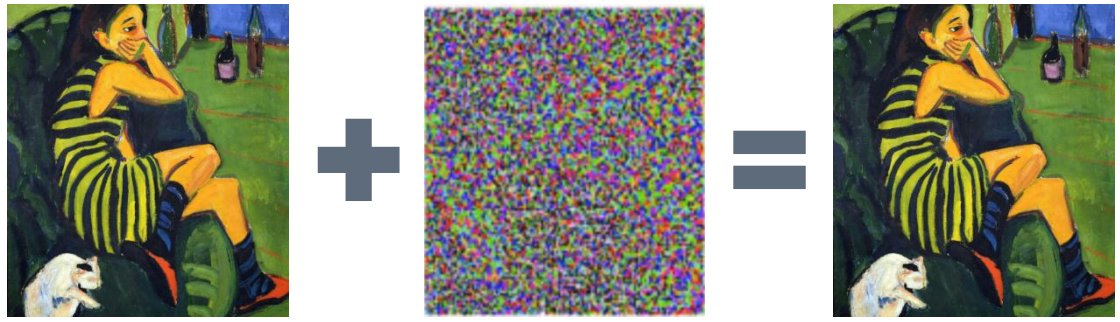
IRREFÜHRUNG – GEOMETRISCHE SICHT

Verschiebung von x auf die andere Seite der Entscheidungsgrenze

- Klassifizierung ändert sich
- Bei ungünstiger Entscheidungsgrenze ist kein visueller Unterschied erkennbar

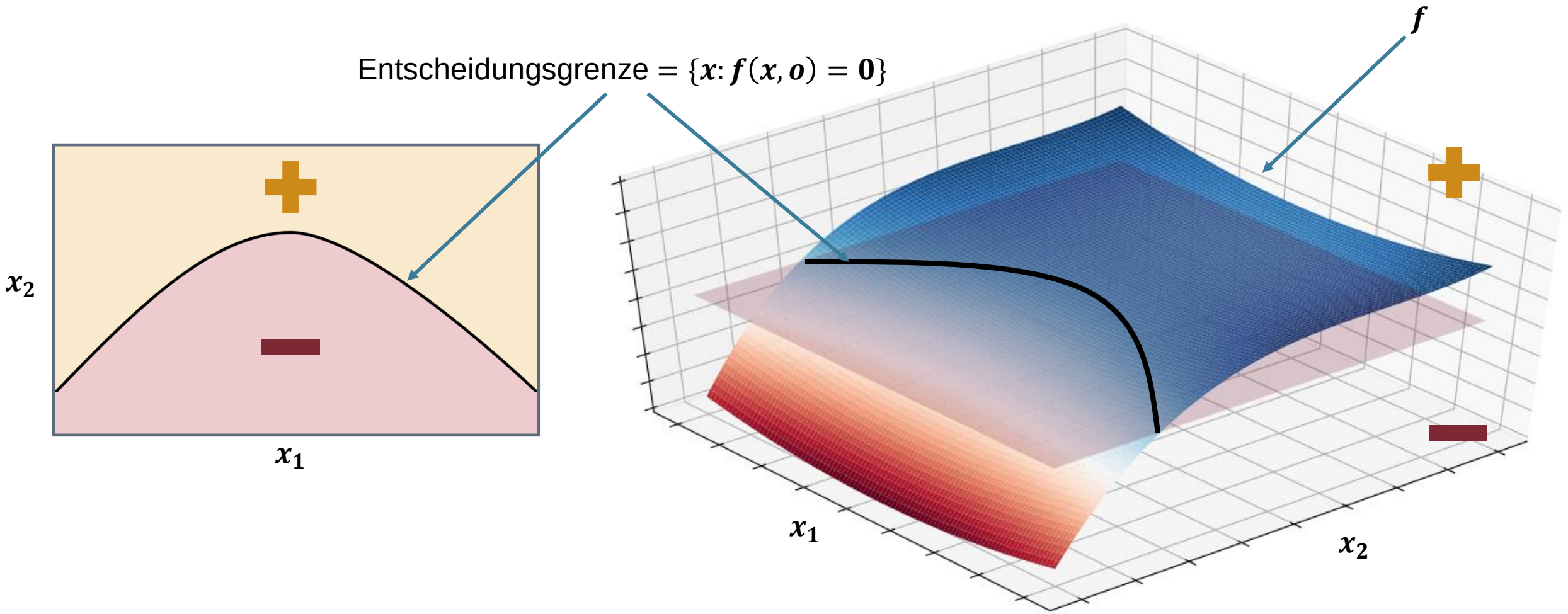


x r $x + r$

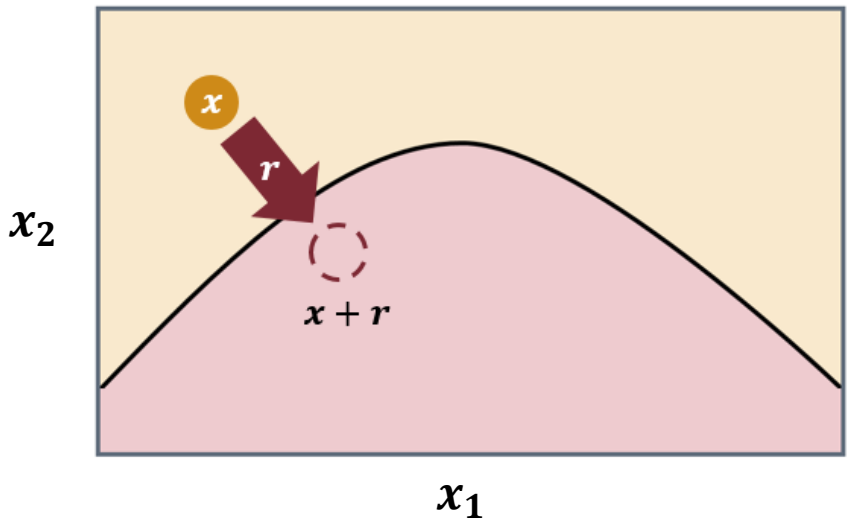


Expressionismus Impressionismus

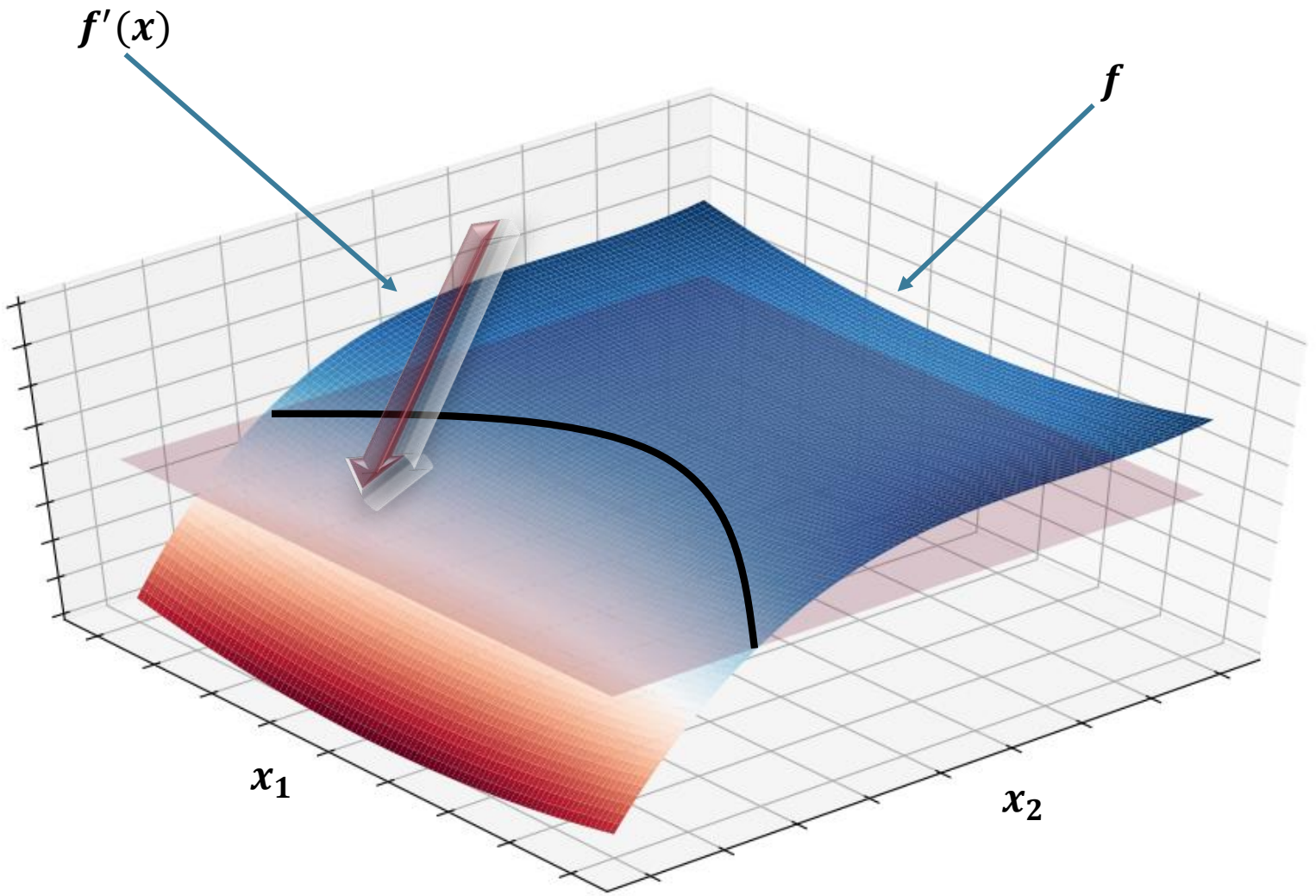
IRREFÜHRUNG – ANALYTISCHE SICHT



IRREFÜHRUNG – ANALYTISCHE SICHT



r kann durch die Ableitung von f berechnet werden



IRREFÜHRUNG – BEISPIEL

- Klassifikator für 1000 Klassen
- Wir möchten dass ein Wal als Hai klassifiziert wird

Wal

Hai?

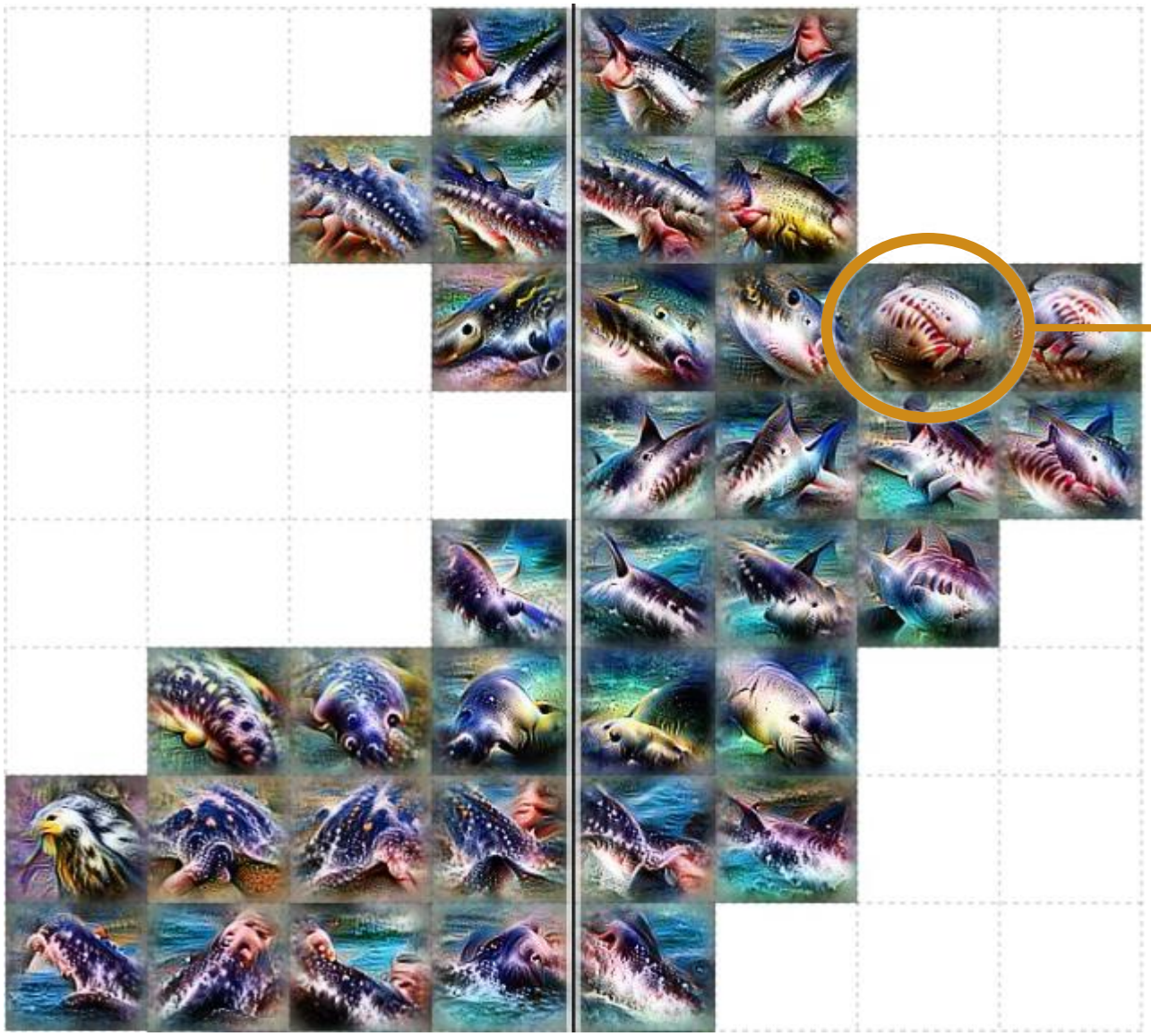
Quelle: distill.pub/2019/activation-atlas

IRREFÜHRUNG – BEISPIEL

Entscheidungsgrenze

Wal

Hai



Baseball?

“Debugging“ von neuronalen Netzen

Quelle: distill.pub/2019/activation-atlas

IRREFÜHRUNG – BEISPIEL

XITASO

x r $x + r$

+

=

Wal Baseball Hai

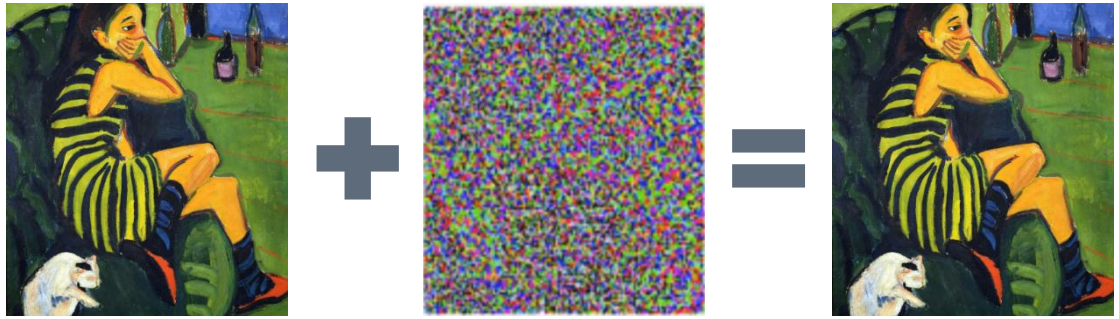
Quelle: distill.pub/2019/activation-atlas

LÖSUNGSANSÄTZE FÜR ROBUSTERE SYSTEME

XITASO

Trainingsdatensatz

x r $x + r$

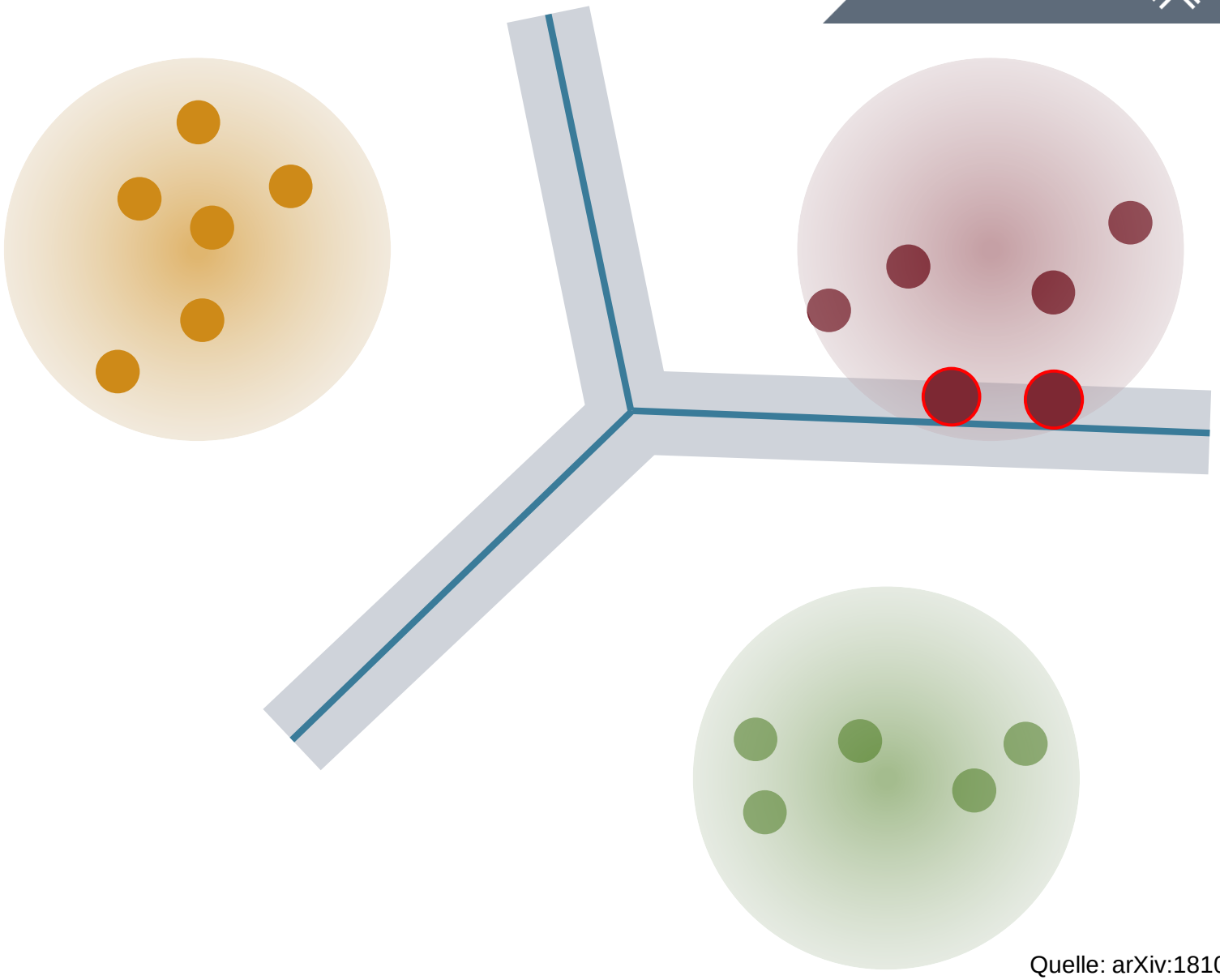


Irreführende Daten zum Training hinzunehmen

Aber: Kein Schutz vor anderen Arten von Angriffen

LÖSUNGSANSÄTZE FÜR ROBUSTERE SYSTEME

XITASO



Sonderbehandlung für Daten in der Nähe von Entscheidungsgrenzen

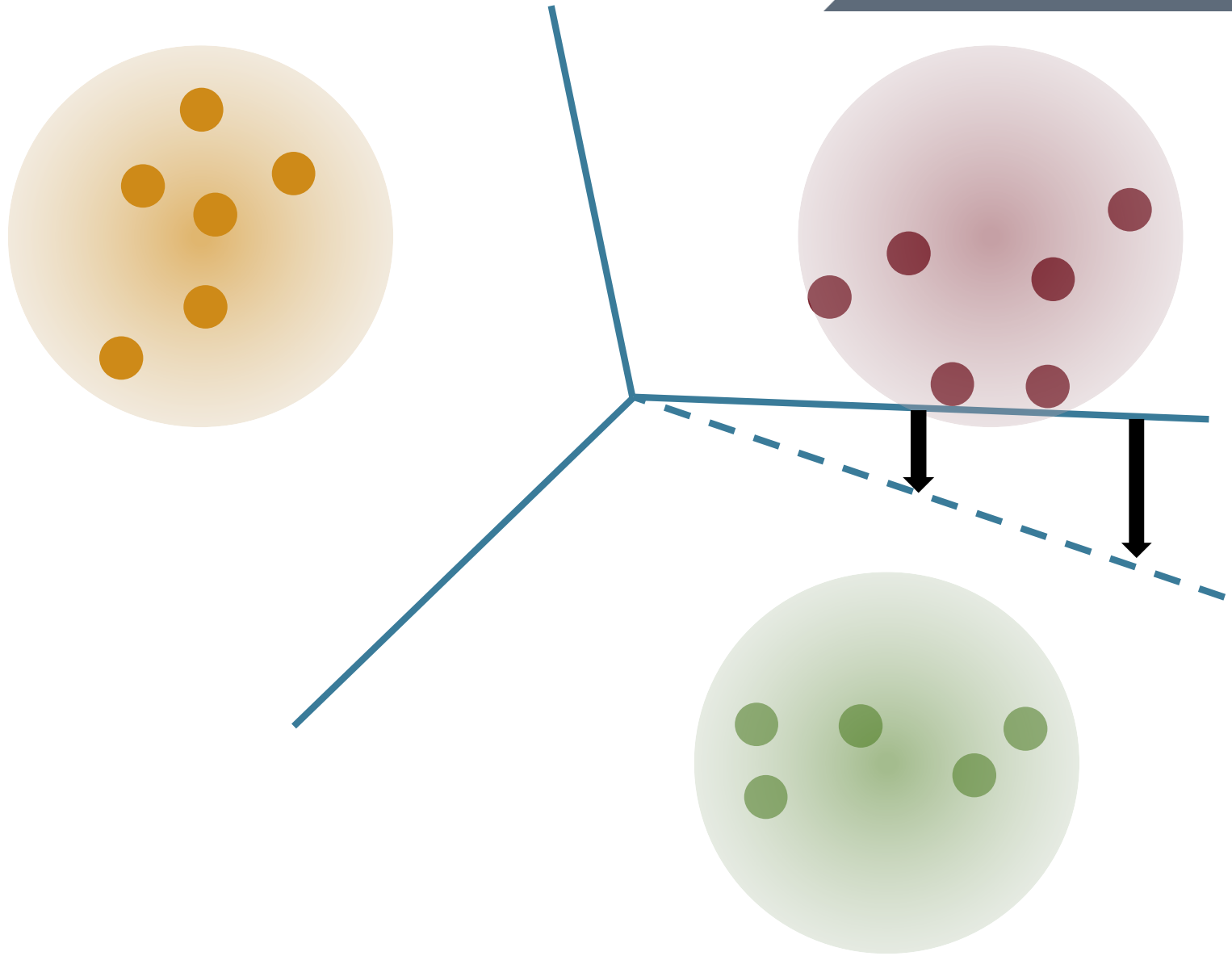
Quelle: arXiv:1810.12715v4

LÖSUNGSANSÄTZE FÜR ROBUSTERE SYSTEME

XITASO

- Entscheidungsgrenzen verschieben
- Maximalen Abstand zu den Klassen gewährleisten
- Aufgrund hoher Dimension und Nichtlinearität schwierig

- Erste Lösungsansätze aus der Forschung: arXiv:1810.12715v4



FAZIT

XITASO

- KI wird bereits in vielen Szenarien sinnvoll eingesetzt
- KI birgt aber auch Schwächen und Risiken (Robustheit, Angriffssicherheit, …)

- Bei einer nicht-robusten KI reicht zufälliges Sensorrauschen für eine Irreführung aus
- Ein “normales“ Training erzeugt kein robustes System
- Für den robusten Einsatz in sicherheitskritischen Szenarien sind zusätzliche Maßnahmen nötig

VIELEN DANK

XITASO
in Augsburg

Austraße 35
86153 Augsburg

Tel. +49 (0)821 885 882 0
E-Mail info@xitaso.com
Web www.xitaso.com

XITASO
in Magdeburg

Werner-Heisenberg-Straße 1
39106 Magdeburg

Tel. +49 (0)391 / 792 930 00
E-Mail info@xitaso.com
Web www.xitaso.com