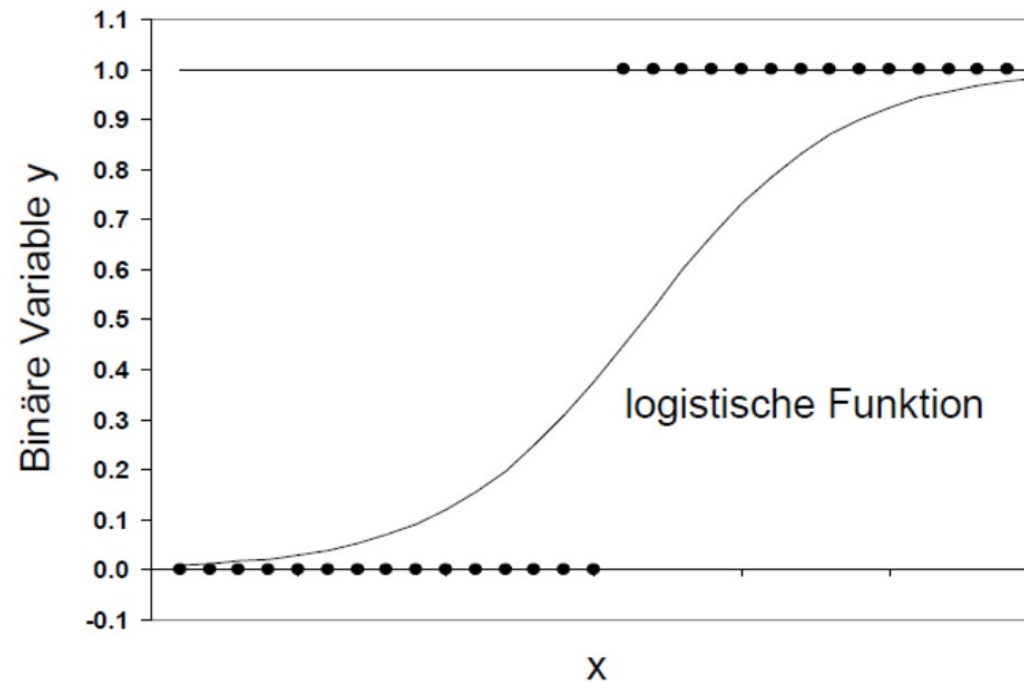


Datenanalyse auf Basis von KI-Methoden

Logistische Regression

Die (binär) logistische Regressionsanalyse testet, ob ein Zusammenhang zwischen mehreren unabhängigen und einer binären abhängigen Variable besteht.



Logistische Regression - Modelgüte

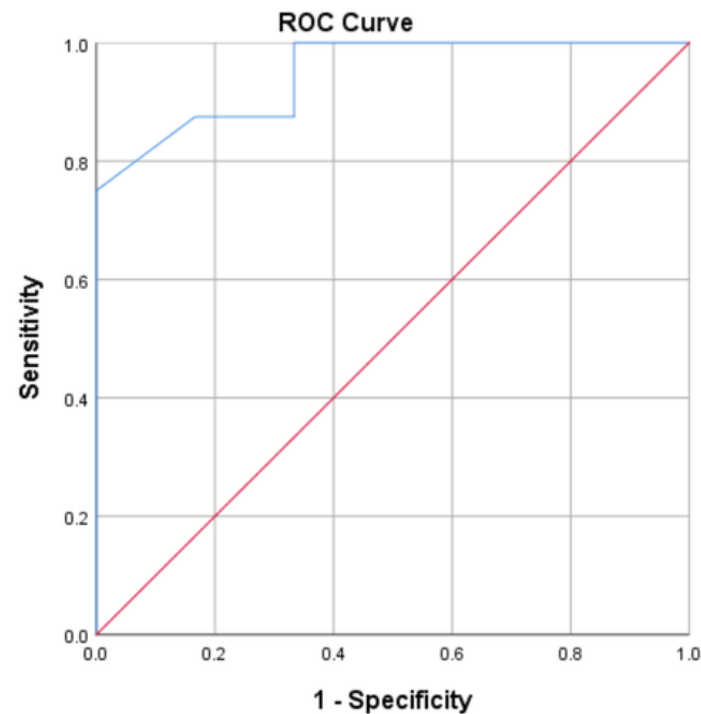
Um zu beurteilen, wie gut ein logistisches Regressionsmodell zu einem Datensatz passt, können wir die folgenden zwei Metriken betrachten:

- **Sensitivität:** Die Wahrscheinlichkeit, dass das Modell ein positives Ergebnis für eine Beobachtung vorhersagt, wenn das Ergebnis tatsächlich positiv ist.
- **Spezifität:** Die Wahrscheinlichkeit, dass das Modell ein negatives Ergebnis für eine Beobachtung vorhersagt, wenn das Ergebnis tatsächlich negativ ist.

Eine einfache Möglichkeit, diese beiden Metriken zu visualisieren, besteht darin, eine **ROC-Kurve** zu erstellen.

Logistische Regression – ROC-Kurve

ROC-Kurve ist ein Diagramm, das die Sensitivität und Spezifität eines logistischen Regressionsmodells anzeigt.



Diagonal segments are produced by ties.

Support Vector Machines

Die Support Vector Machine ist eine mathematische Methode und einen zur Klassifizierung von Objekten.

Sie lässt sich im maschinellen Lernen als Algorithmus einsetzen, um Daten zu analysieren und die Objekte bestimmten Klassen zuzuordnen.

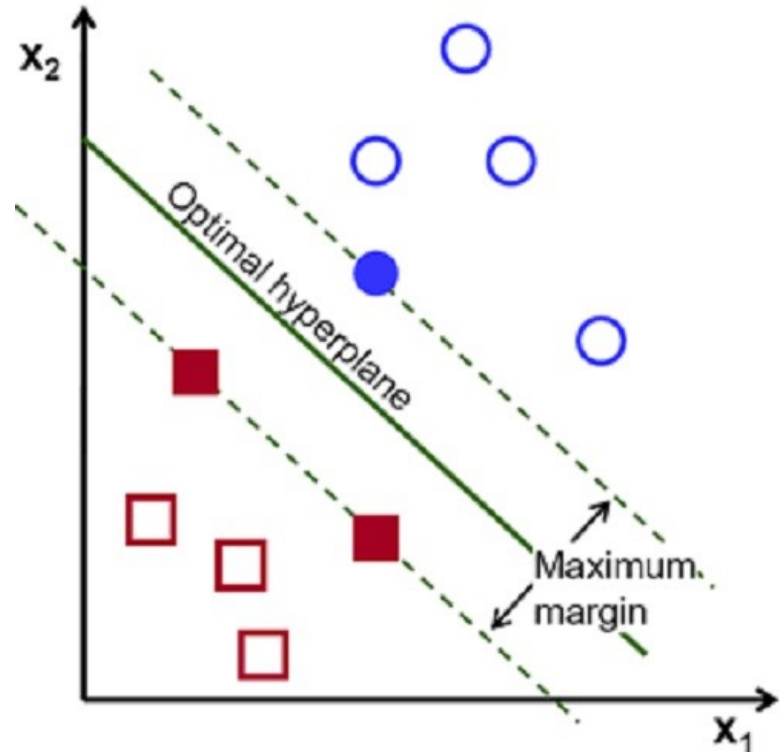
Die SVM unterstützt sowohl die lineare als auch die nicht-lineare Klassifizierung. Für die nicht-lineare Klassifizierung kommt der sogenannte Kernel-Trick zum Einsatz, der den Objektbereich zur Abbildung nicht-linearer Trennflächen um zusätzliche Dimensionen (Hyperebenen) erweitert.

Wie funktioniert der Support Vector Machines

Die von der Support Vector Machine ermittelten Klassengrenzen sind sogenannte Large Margin Classifier und lassen um die Klassengrenzen herum einen möglichst breiten Bereich frei von Objekten.

Die Support Vector Machine versucht in verschiedenen Trainingsdurchläufen die Hyperplane so zu wählen, dass die Gap maximal wird.

Diese misst den Abstand vom nächsten Element jeder Gruppe zu der Hyperplane..

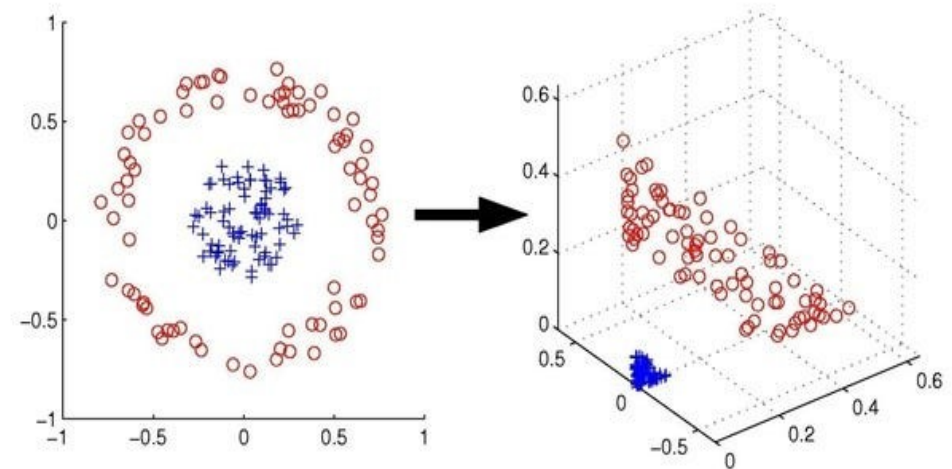


Support Vector Machines bei Nicht-linearen Daten

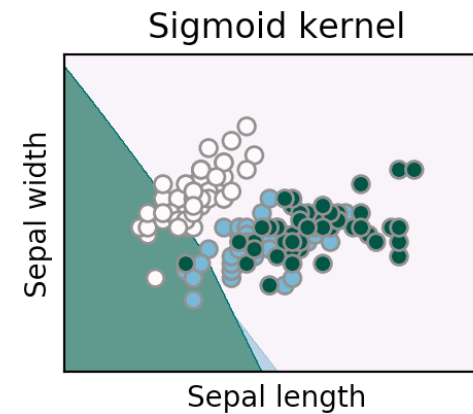
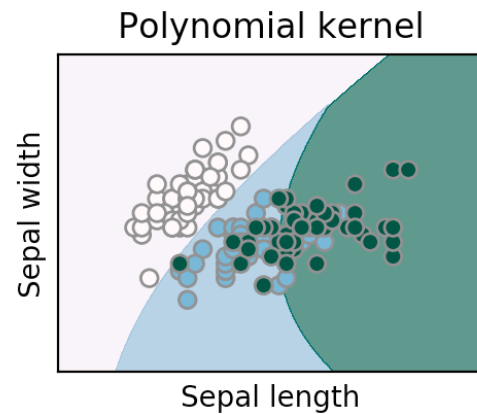
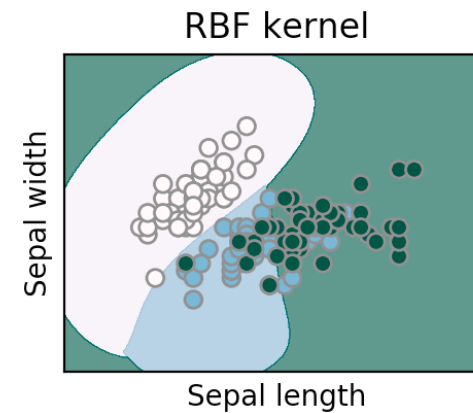
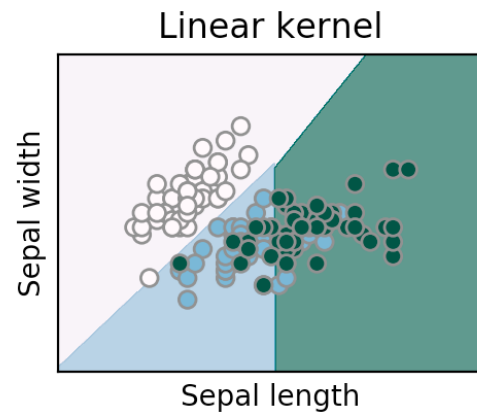
Obwohl wir die beiden Datengruppen mit dem Auge deutlich unterscheiden können, lassen sie sich nicht mit einer einfach linearen Hyperplane abgrenzen.

Eine Möglichkeit um die Support Vector Machines trotzdem nutzen zu können ist, eine neue Dimension einzuführen und diese so zu erstellen, dass die Datenpunkte in einem höherdimensionalen Raum durch eine Hyperplane trennbar sind.

Hier kommt der sogenannte Kernel-Trick zum Einsatz!



Support Vector Machines Kernels



Random Forest

Random Forest ist ein supervised Machine Learning Algorithmus, welcher sich aus einzelnen Decision Trees zusammensetzt.

Eine solche Art von Modell wird als Ensemble Modell bezeichnet, da ein „Ensemble“ aus unabhängigen Modellen genutzt wird, um ein stärkeres Modell zu erhalten.

Das daraus abgeleitete Modell ist robuster und genauer und kann mit Überanpassungen besser umgehen als konstitutive Modelle.

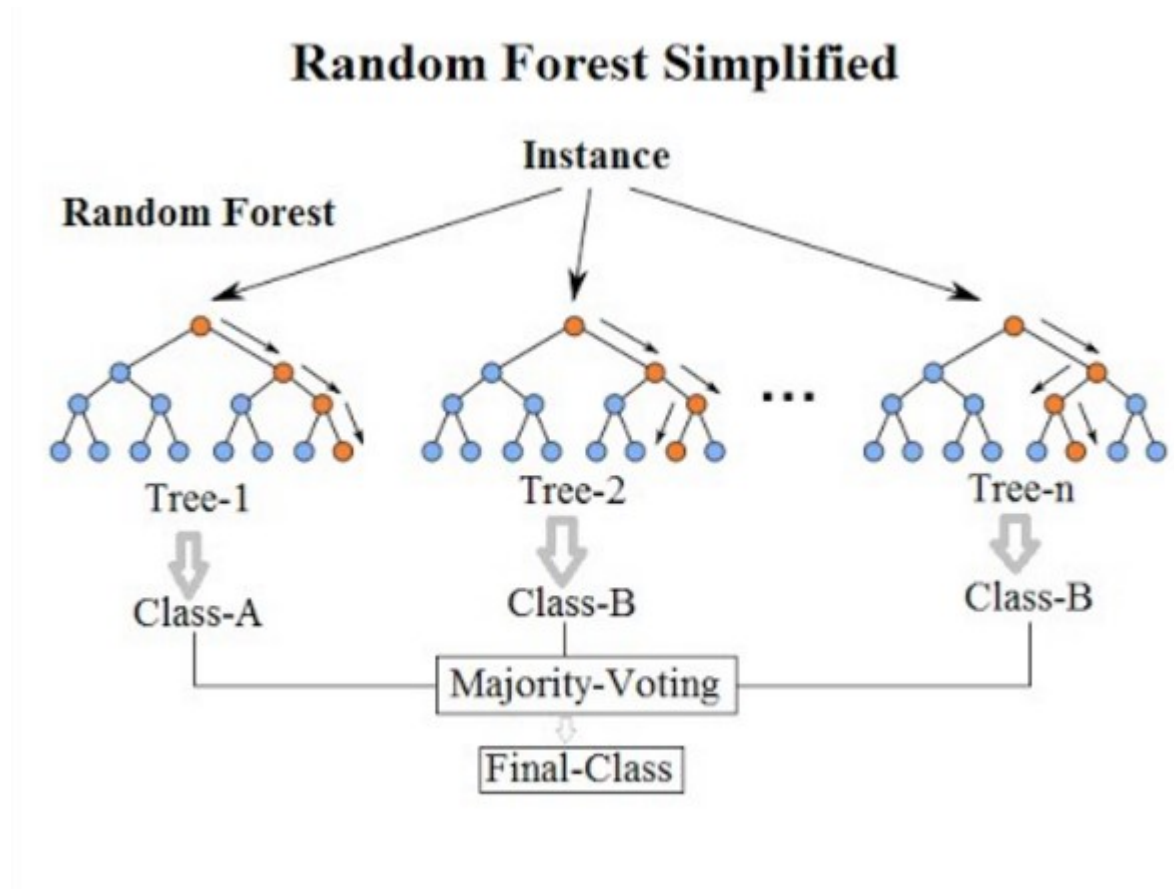
In der Praxis wird dieser Algorithmus für verschiedene Klassifikationsaufgaben eingesetzt.

Wie funktioniert der Random Forest?

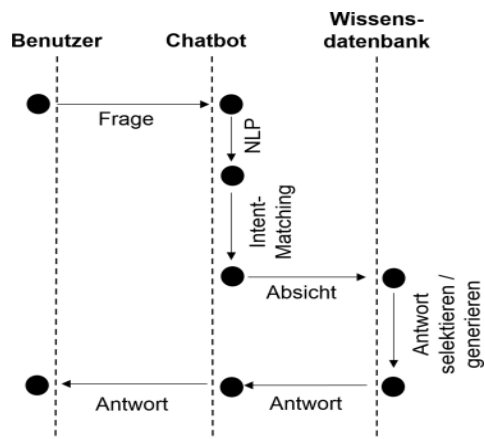
Der Random Forest besteht aus einer Vielzahl dieser [Decision Trees](#), welche als ein sogenanntes [Ensemble](#) zusammenarbeiten. Jeder einzelne Entscheidungsbaum gibt eine Vorhersage, beispielsweise ein Klassifizierungsergebnis ab, und der Forest nutzt das Ergebnis, das von den meisten [Decision Trees](#) unterstützt wird, als Vorhersage des gesamten Ensembles.

Jeder Baum innerhalb des Forests wird auf einem [Sample](#) des Trainingsdatensatzes trainiert, was verhindert, dass die Bäume untereinander korreliert sind.

Wie funktioniert der Random Forest?



Was ist ein Chatbot



Ein Chatbot ist ein **textbasiertes Dialogsystem**, das in der Lage ist, mit Menschen in **menschlicher Sprache zu kommunizieren**.

Ausprägungsarten von Chatbot-typen

Wissensdomäne

Geschlossene-Domain Chatbots

Antworten werden von Wissensdatenbank erzeugt
Fragen werden rational beantwortet

Offene-Domain Chatbots

- Fragen werden nicht rational beantwortet
- Antworten werden aus Häufigkeiten generiert

Methode zur Erzeugung von Antworten

„Intelligente“ Chatbots

Folgen Klassifizierungsalgorithmen und basieren auf
Algorithmen des maschinellen Lernens
Sind weniger fehleranfällig
Werden „intelligenter“, je mehr Menschen diese nutzen

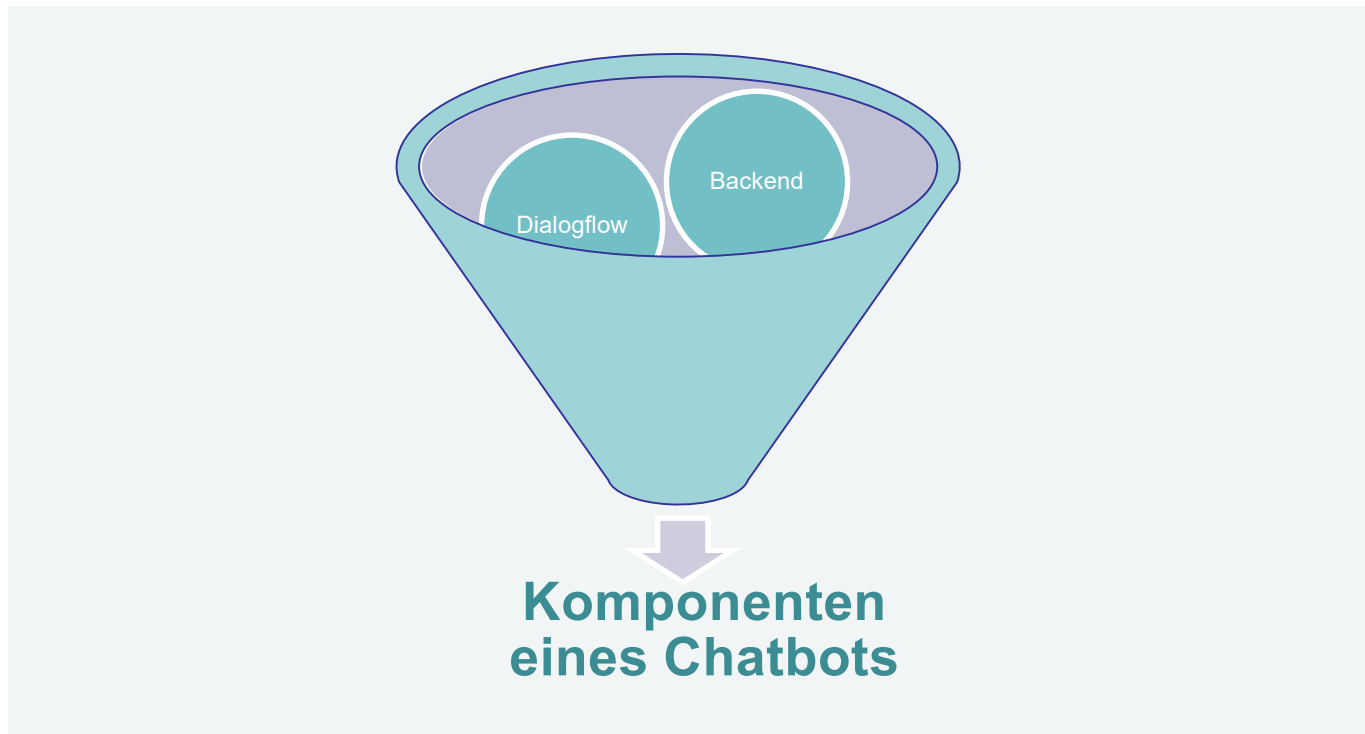
Regelbasierte Chatbots

- Sehr limitiert, da diese nach fest vorgegebenen Mustern und Befehlen bzw. festgelegten Regeln funktionieren
- Sind nur so „intelligent“, wie die Programmierung dahinter

Eigenschaften eines „guten“ Chatbots

- Gut durchdachte **Gesprächsabläufe**—>idealerweise nicht vom Entwickler-Team modelliert wurden
- Der Benutzer muss eine erste **Antwort** erhalten, nach **max. 2 Nachfragen** des Chatbots
- Textblöcke (mit **mehr als 60 Zeichen**) sind als Antwort **tabu**
- Verwendung von **Verlinkungen**
- Vorschläge für **weitere interessante Themen**
- Chatbots müssen eine **Persönlichkeit** haben, jedoch muss erkennbar sein, dass es ein Chatbot ist und **nicht mit einem Menschen assoziiert** wird

Aufbau des Chatbots



Verwendet Technologien des Natural Language Processing



Beispiel NLP (Natural Language Processing)

Satz als Text:

„Ich heirate, was kostet ein neuer Personalausweis?“

Ergebnis nach NLP-Anwendung:

heirat kost Personalausweis

“Bag of Words” Modell (BoW)

Satz als Text

1. [Ein Hund ist ein Tier]
2. [Eine Katze ist auch ein Tier]

Vokabular

[auch, ein, eine, Hund, ist, Katze, Tier]

Satz als Vektor

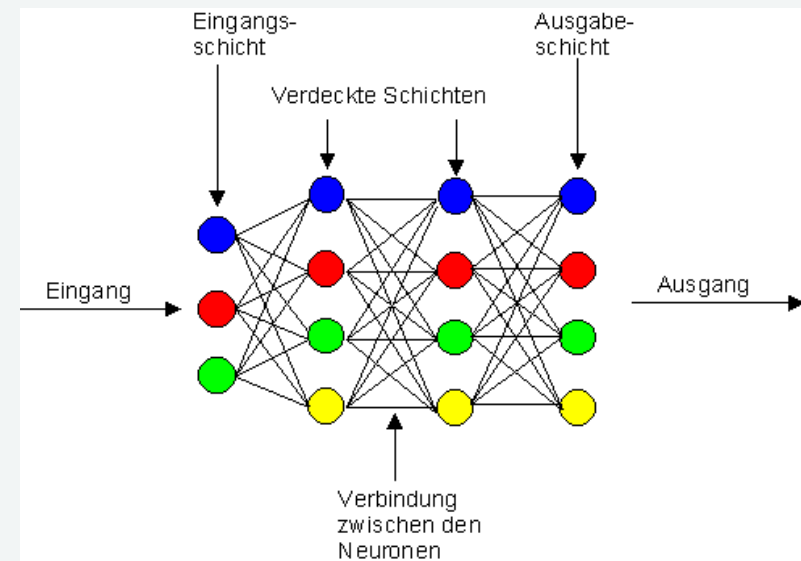
1. [0, 1, 0, 1, 1, 0, 1]
2. [1, 0, 1, 0, 1, 1, 1]

Klassifikator

Der **Klassifikator** identifiziert die **Intention** des Users.

Die **Intention** ist die Absicht, die ein User mit seiner Anfrage ausdrücken möchte.

Neuronale Netze



***Vielen Dank
Für Ihre
Aufmerksamkeit!***