
CONSULTING
enpit

DOKUMENTENKLASSIFIKATION MIT MACHINE LEARNING

Andreas Nadolski
Softwareentwickler

andreas.nadolski@enpit.de

Twitter: @enpit
Blogs: enpit.de/blog
medium.com/enpit-developer-blog

CONSULTING enpit

AGENDA

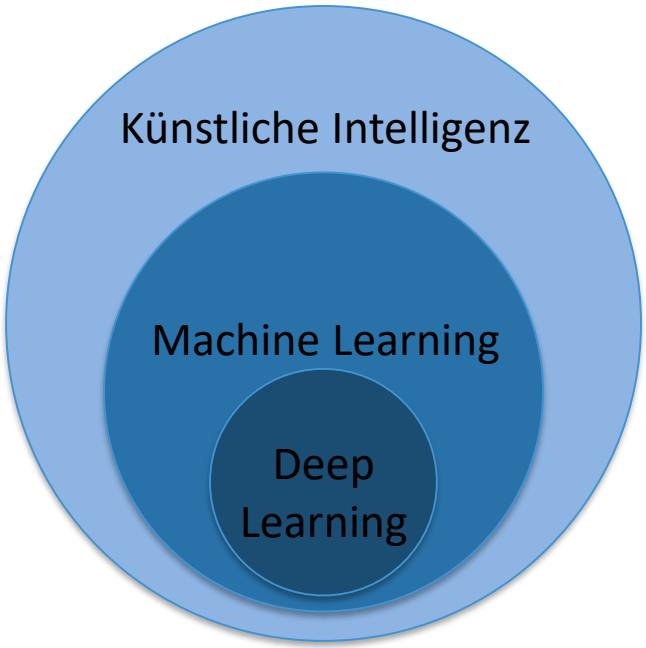
- Machine Learning
- Dokumentenklassifikation
- Live-Demo
- Herausforderungen und Fazit

AGENDA

- **Machine Learning**
- Dokumentenklassifikation
- Live-Demo
- Herausforderungen und Fazit

WAS IST
MACHINE LEARNING?

Einordnung Künstliche Intelligenz, Machine Learning und Deep Learning



BEISPIELE FÜR MACHINE LEARNING

- SPAM Erkennung
- Spracherkennung
- Bildklassifikation (z.B. Gesichtserkennung)
- Objekterkennung (z.B. Straßenschilder)
- Klassifizierung von Dokumenten

Heute schon im Einsatz!

“Machine learning is the science of getting computers to act without being explicitly programmed.”

Andrew Ng

WAS IST MACHINE LEARNING?

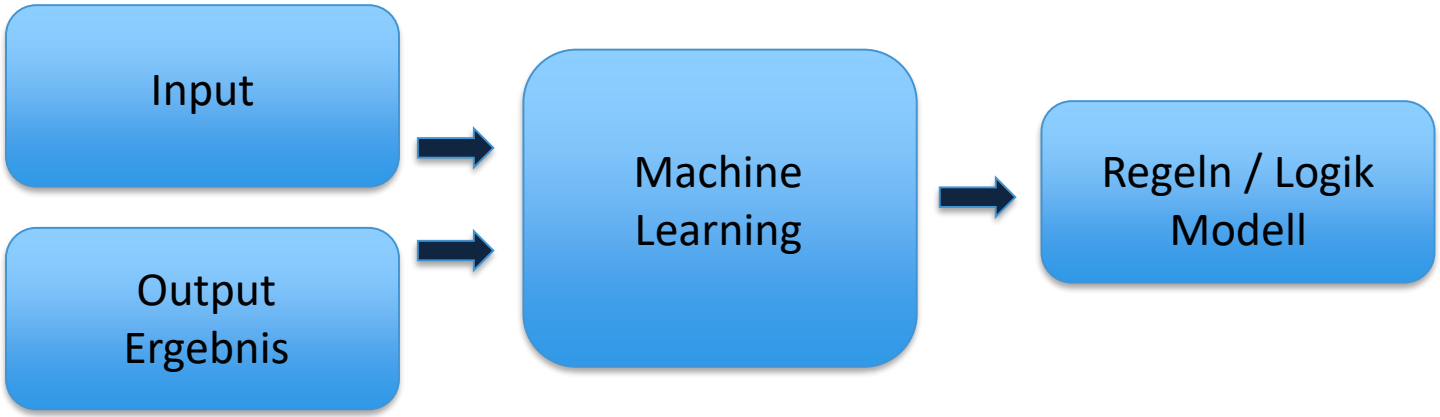
WAS IST
MACHINE LEARNING?

Klassische Vorgehensweise

Regeln müssen richtig sein und mit der Zeit angepasst werden

WAS IST
MACHINE LEARNING?

Vorgehensweise bei Machine Learning



VERSCHIEDE ARTEN DES MACHINE LEARNING

Überwachtes Lernen
Supervised Learning

- Daten sind gekennzeichnet
- Ziel: Ergebnis vorhersagen

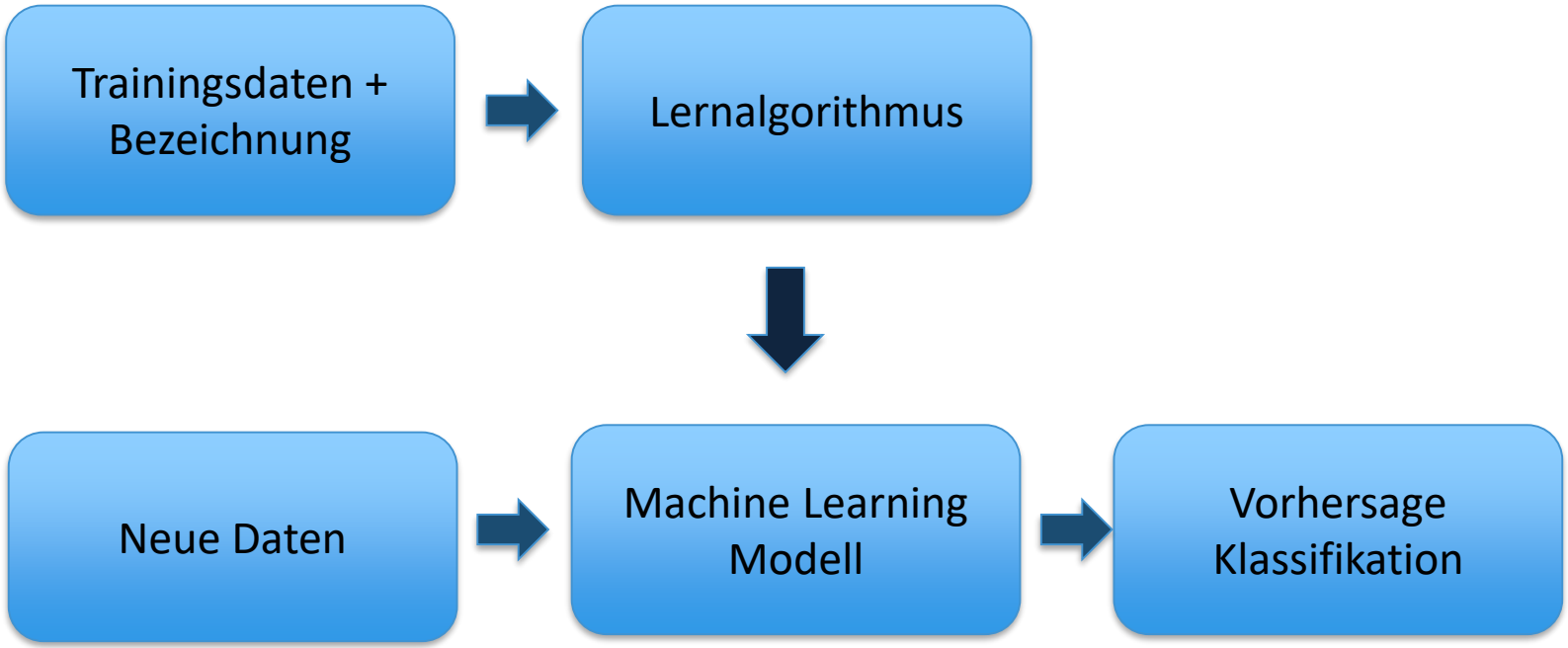
Unüberwachtes Lernen
Unsupervised Learning

- Keine Kennzeichnung
- Ziel: Unbekannte Strukturen finden

Verstärkendes Lernen
Reinforcement Learning

- Erlernen von Aktionen
- Belohnung

ÜBERWACHTES LERNEN

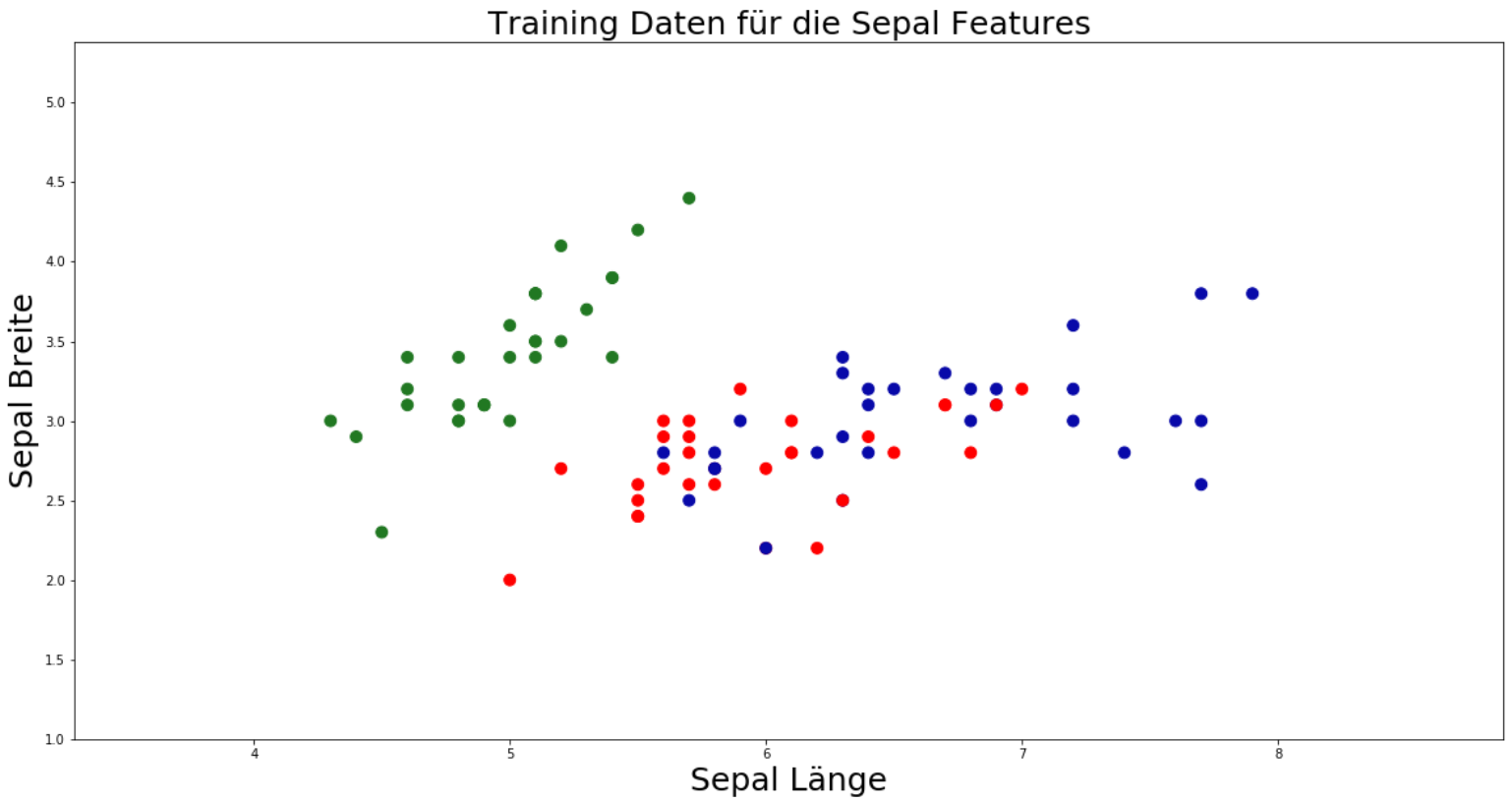


BEISPIEL FÜR ÜBERWACHTES LERNEN

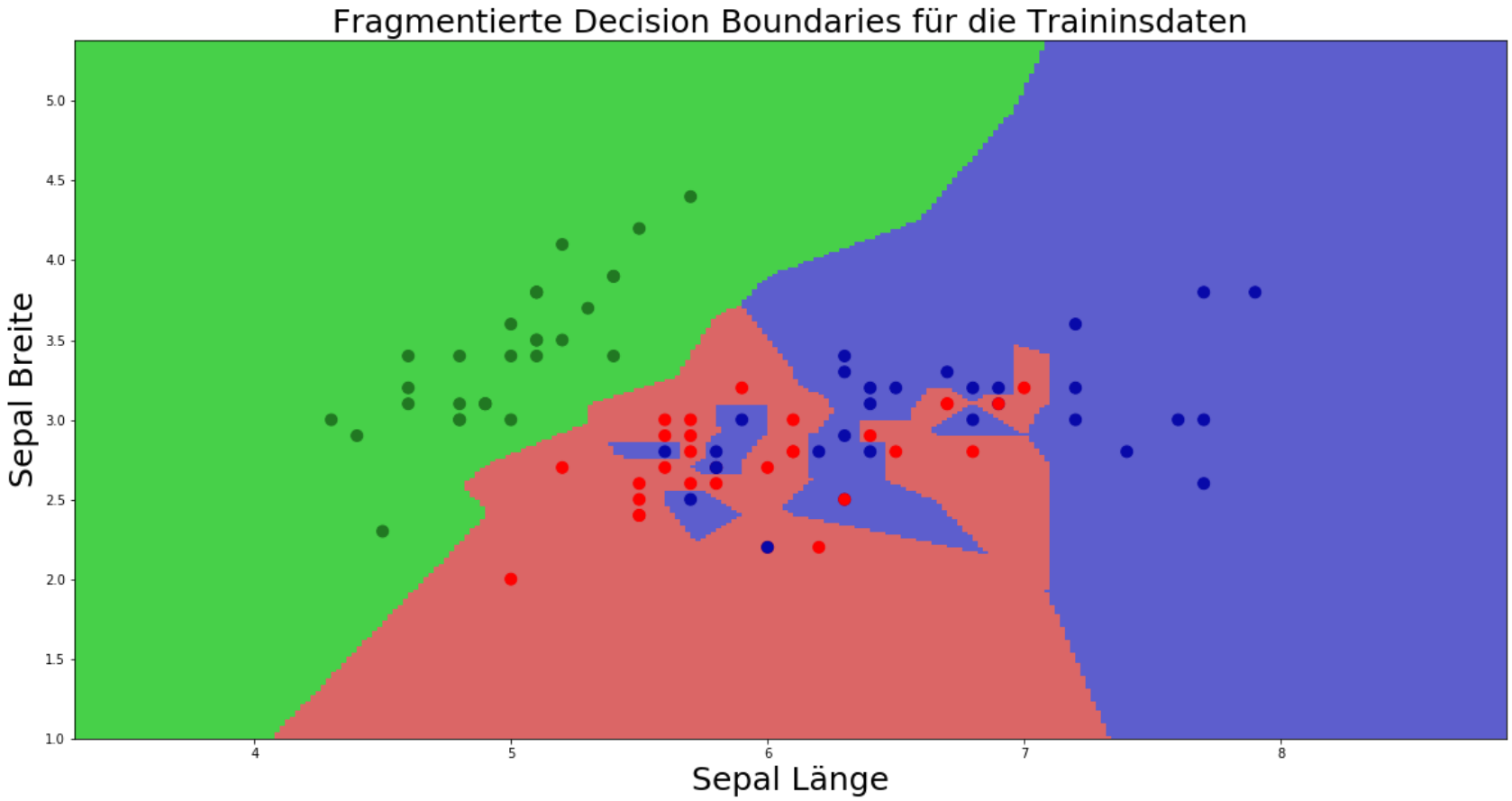
Irisdatensatz (Schwertlilien) mit vier Features

sepal length	sepal width	petal length	petal width	class
50	20	35	10	Iris-versicolor
60	22	40	10	Iris-versicolor
62	22	45	15	Iris-versicolor
60	22	50	15	Iris-virginica
45	23	13	3	Iris-setosa
50	23	33	10	Iris-versicolor
55	23	40	13	Iris-versicolor
63	23	44	13	Iris-versicolor

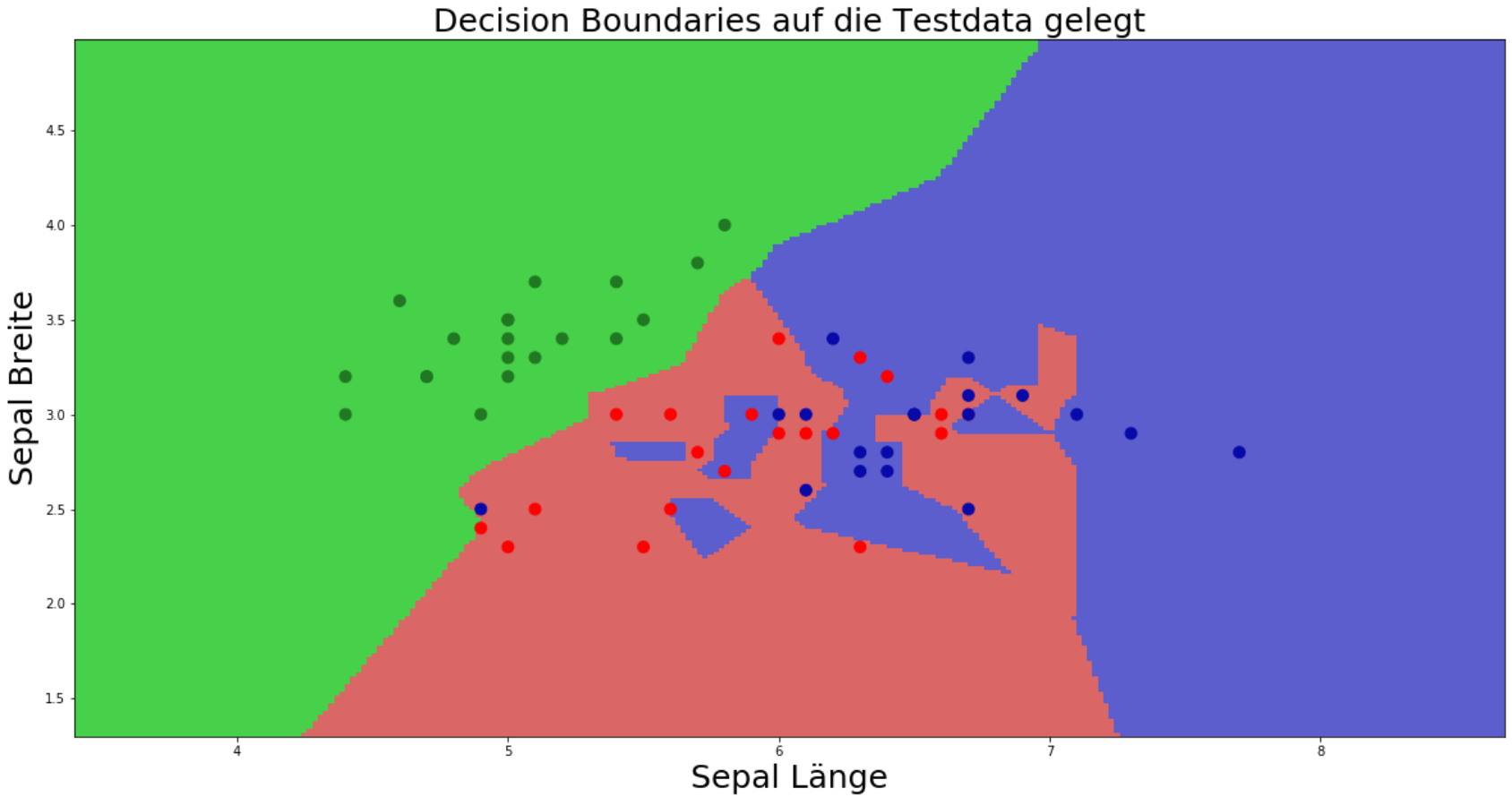
VERTEILUNG DER TRAININGSDATEN



DESCISION BOUNDARIES
TRAINGSDATEN



DESCISION BOUNDARIES TESTDATEN



CONSULTING enpit

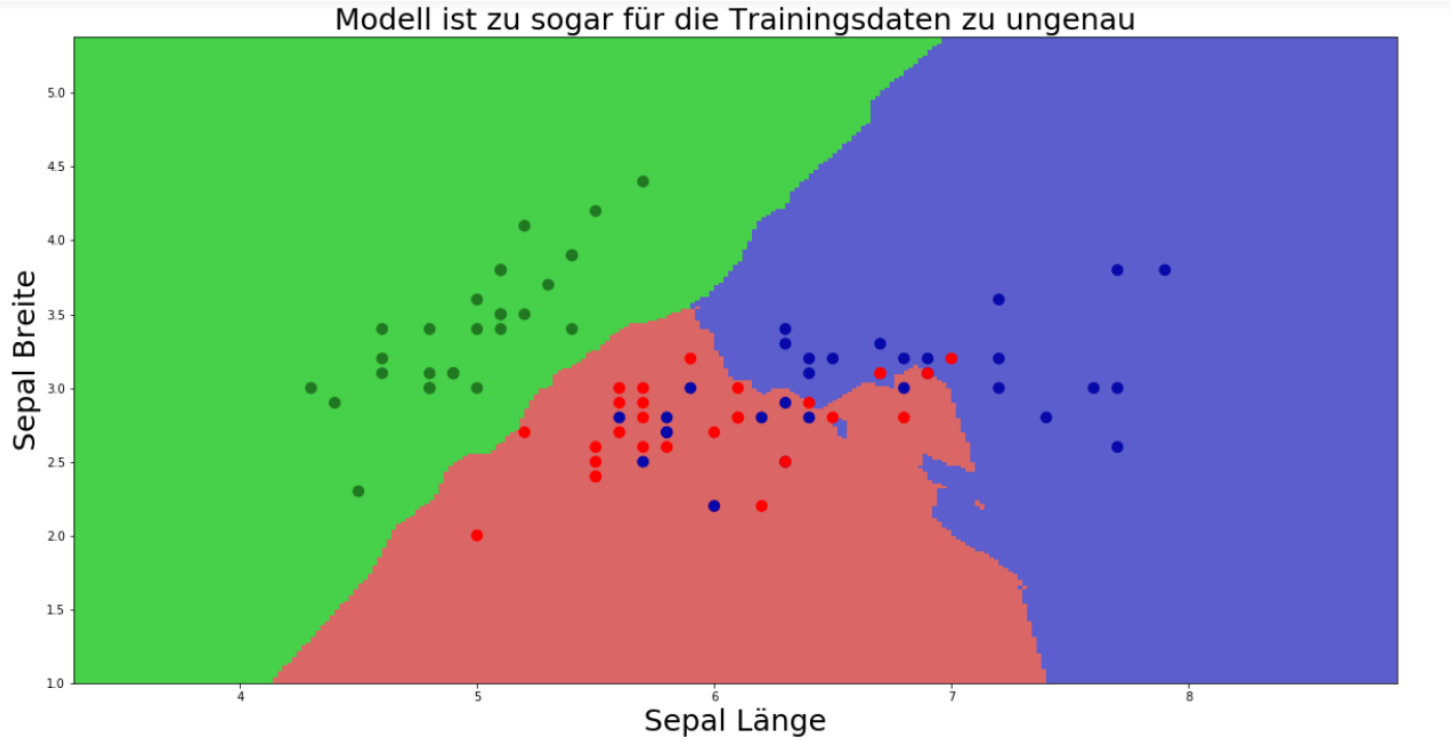
BEISPIEL FÜR ÜBERWACHTES LERNEN

Overfitting!

- Model ist zu speziell und zu komplex
- **Low Bias** und **High Variance**
- Keine Generalisierung

Das Modell reagiert nur schlecht auf neue unbekannte Daten

DESCISION BOUNDARIES
TRAININGSDATEN (10 NEIGHBOURS)



BEISPIEL FÜR ÜBERWACHTES LERNEN

Underfitting!

- Nicht einmal die Trainingsdaten passen
- **High Bias** und **Low Variance**

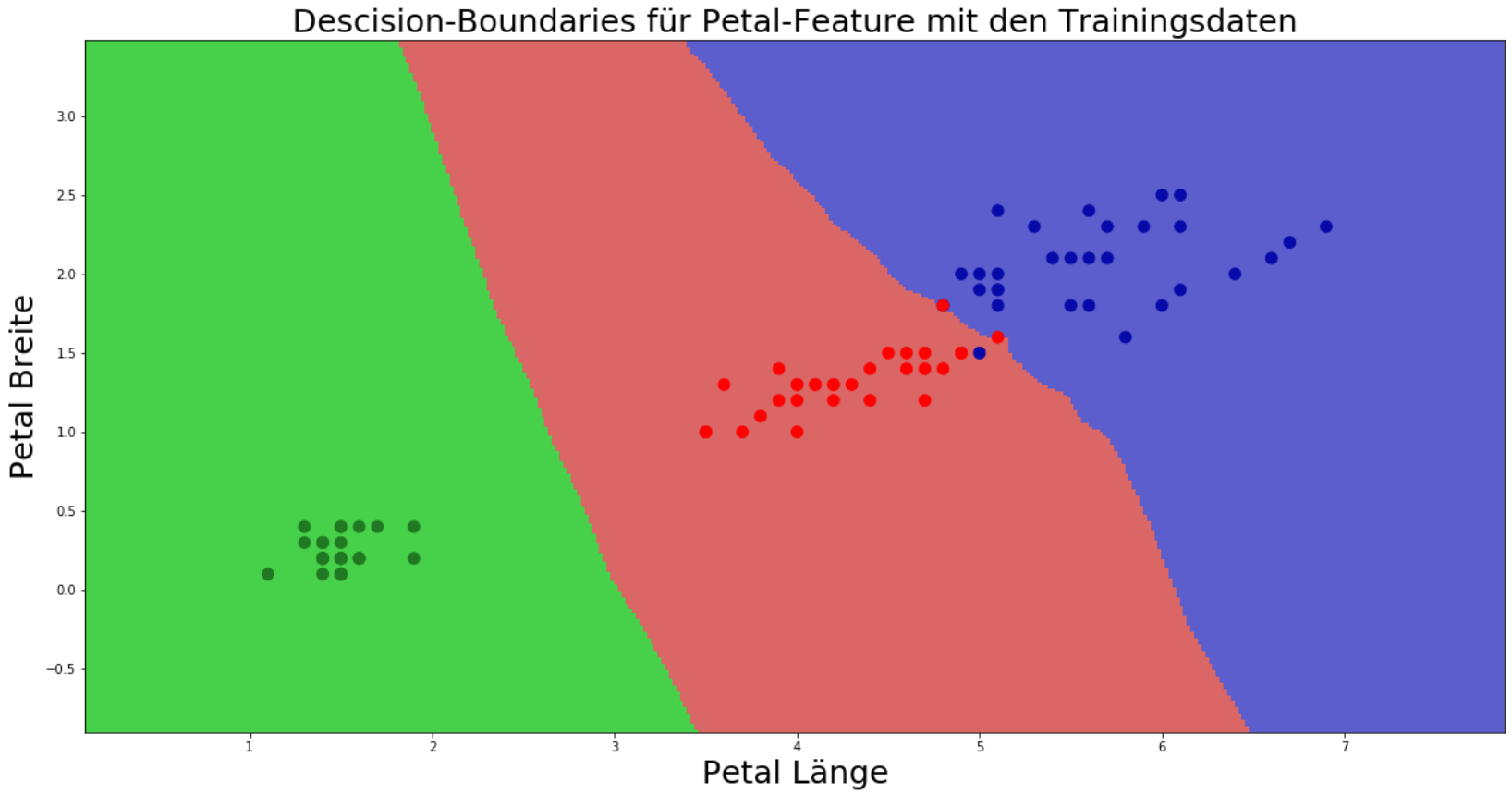
Nicht genug an die Testdaten angepasst

BEISPIEL FÜR ÜBERWACHTES LERNEN

- Zu wenig Daten
- Schlechte Daten
- Nicht die richtigen Daten

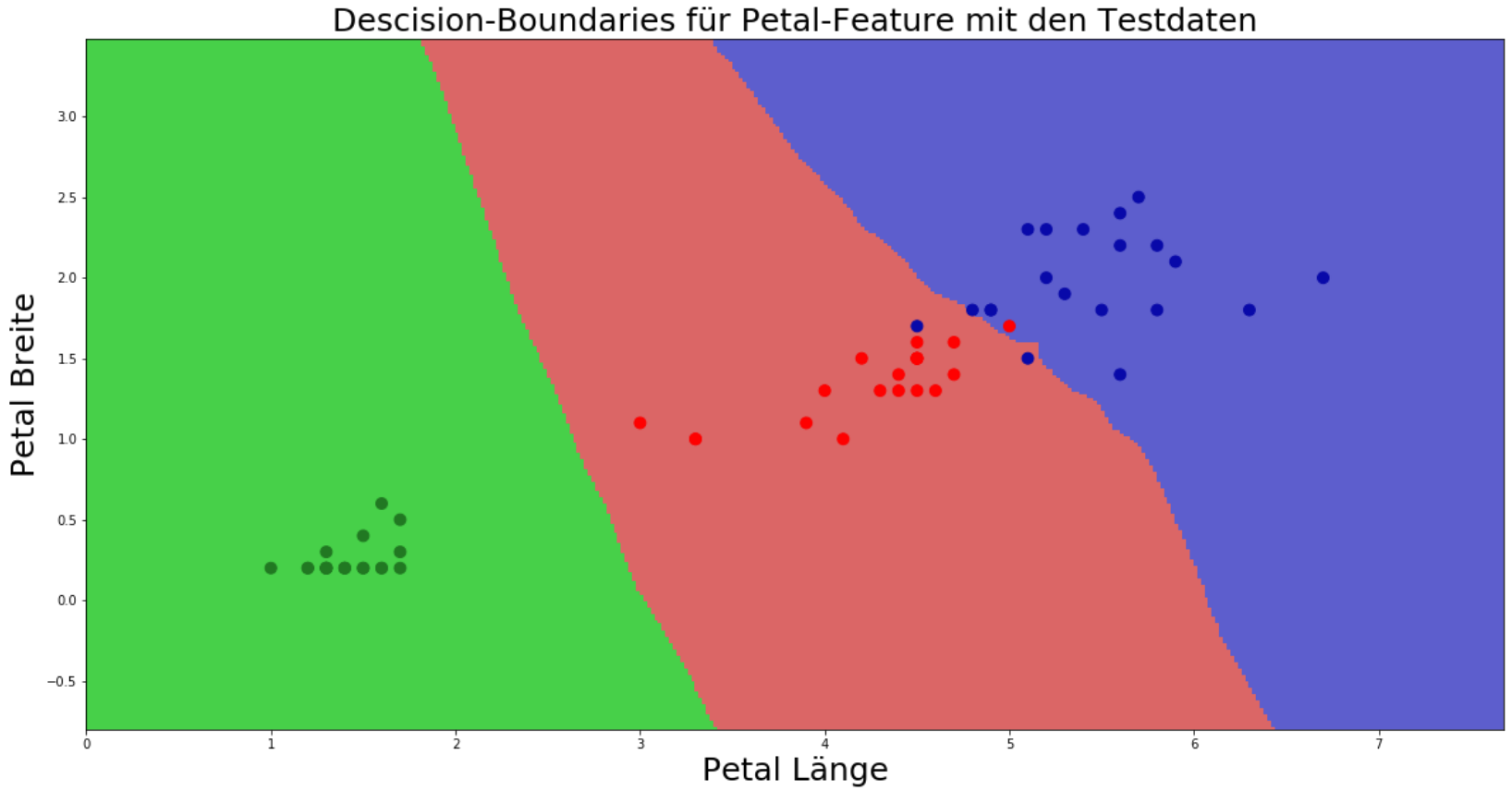
Feature Auswahl verbessern, um den Sweet-Spot zu finden!

FEATURE AUSWAHL
ÜBERWACHTES LERNEN



CONSULTING enpit

FEATURE AUSWAHL
ÜBERWACHTES LERNEN

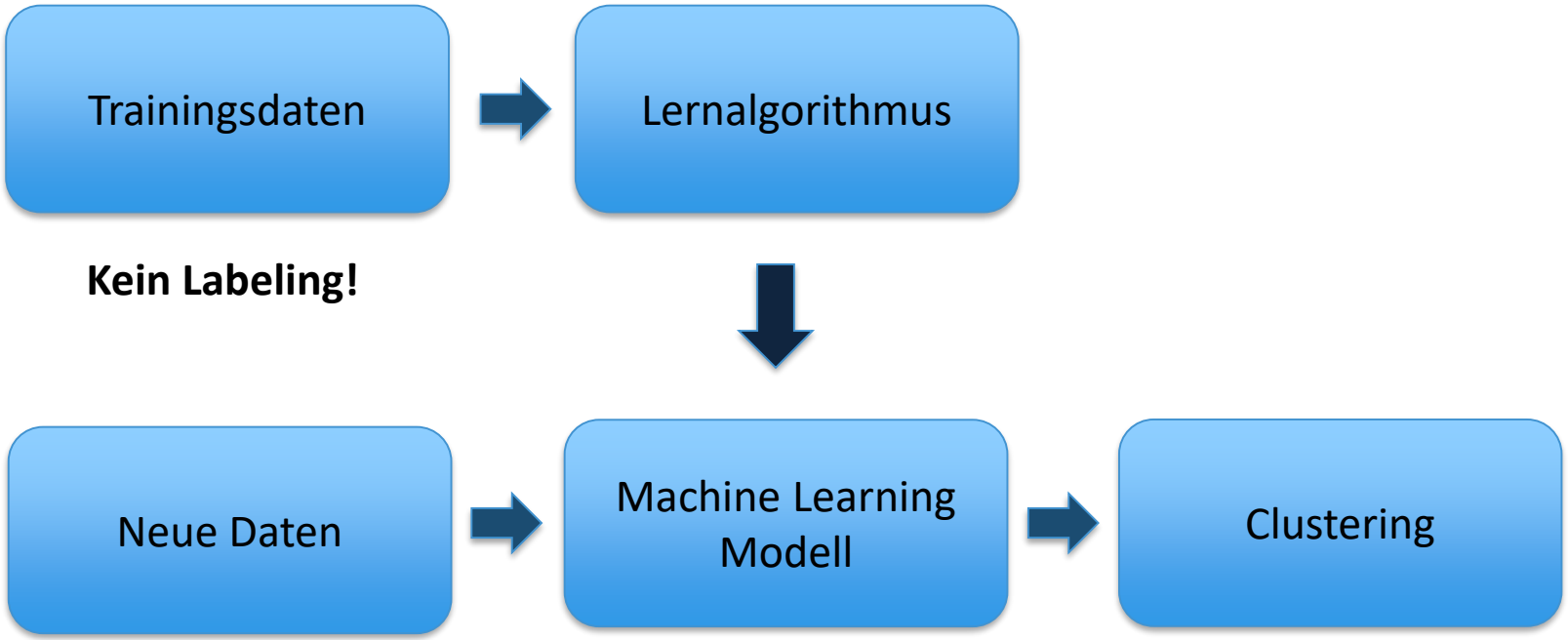


BEISPIEL FÜR ÜBERWACHTES LERNEN

Vergleich des Accuracy Scores - je nach Featureauswahl

- **Sepal** 0.76 – 0.8
- **Petal** 0.96 – 0.97
- **Sepal und Petal** 0.95

UNÜBERWACHTES LERNEN



AGENDA

- Einordnung ML
- **Dokumentenklassifikation**
- Live-Demo
- Herausforderungen und Fazit

KLASSIFIZIERUNG VON DOKUMENTEN

„One word for it. Hilarious. I haven't watched at movie like this in a long time. At points in the movie, I totally forgot it was a movie, I just felt like I was back watching Viva La Bam, or even watching say, my own friend going through something like this. It was realistic and I liked how Bam, Ryan, Raab, Rake, and Brandon and the rest of the guys didn't try to hard too actually act. They, to me, were just acting like their famous idiot selves. There were a few scenes that I adored more than others, like Raab in the shower, holy, I laughed so hard. He honestly was probably my favourite character besides Bam's. He really, in my opinion, made the movie just a bit more hilarious. It's basically a must see for any fans of the CKY crew:]"

KLASSIFIZIERUNG VON DOKUMENTEN

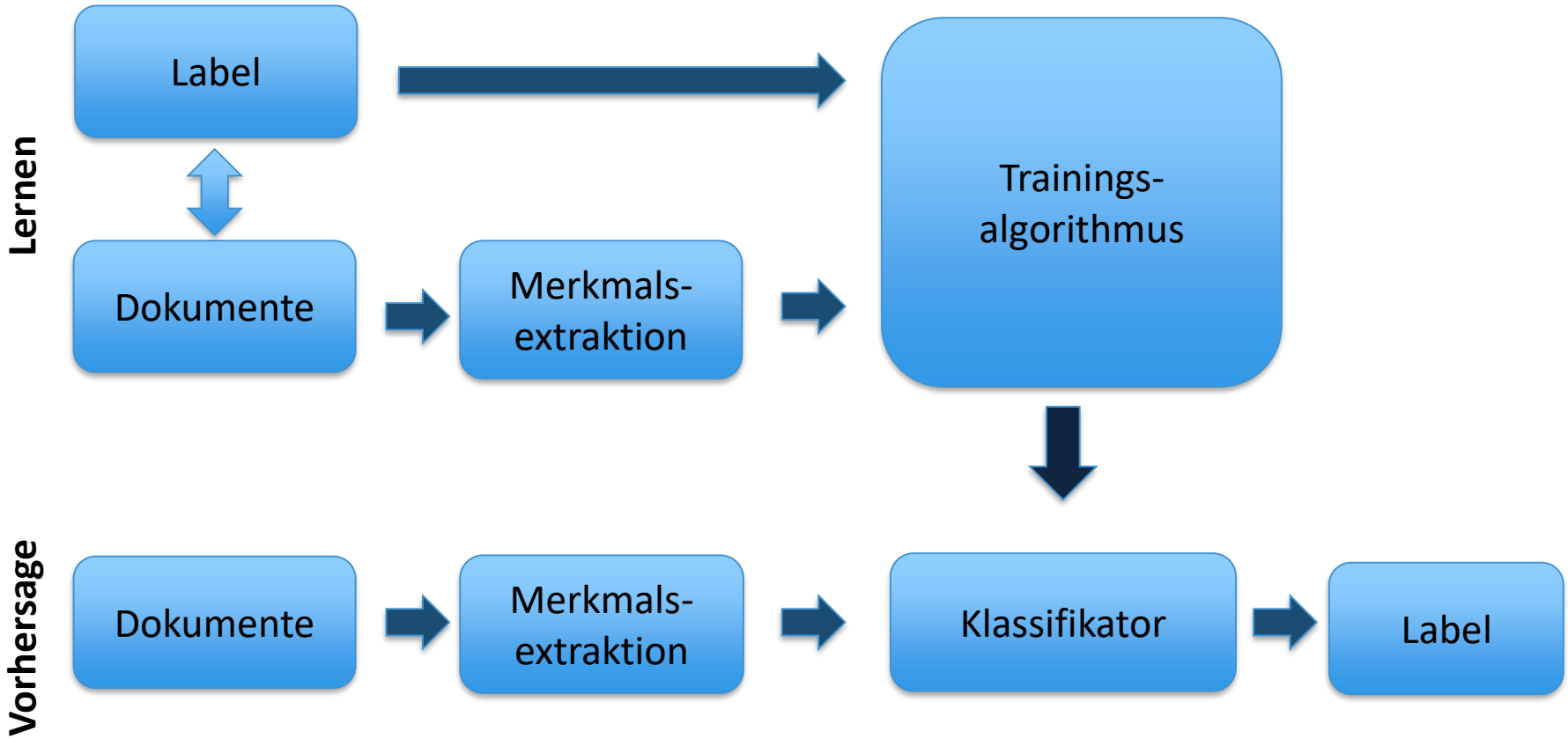
Besonderheit: Viele Features

Wort	
Film	1
wenn	2
Erde	3
…	…
neu	n

KLASSIFIZIERUNG VON DOKUMENTEN

- Schritt 1: Erstellung von Trainingsdaten für das Klassifikationstraining
- Schritt 2: Training des Klassifikationsmodells
- Schritt 3: Einsatz der Klassifikation

KLASSIFIZIERUNG VON DOKUMENTEN



KLASSIFIZIERUNG VON DOKUMENTEN

Vorgehen

- Datenbeschaffung
- Daten bereinigen
- Daten einlesen

KLASSIFIZIERUNG VON DOKUMENTEN

Datenqualität

- Überblick über die Qualität verschaffen
- Probleme erkennen
- Daten bereinigen

KLASSIFIZIERUNG VON DOKUMENTEN

Qualität beurteilen und Probleme erkennen

- Ausreichend Testdaten pro Klasse vorhanden?
- Alle Features immer vorhanden?
- Leere Einträge oder Dubletten?
- Falsche oder fehlende Werte / Labels

Korrigieren (z.B. Mittelwert), löschen oder mittels ML bestimmen

KLASSIFIZIERUNG VON DOKUMENTEN

Besonderheiten Dokumentenklassifikation

- Feature-Raum mit sehr hoher Dimension
- Viele Wörter und mehrere Varianten für ein Wort
- man braucht viele Daten um statistisch signifikante Aussagen zu treffen

Reduktion der Features

KLASSIFIZIERUNG VON DOKUMENTEN

Reduktion der Features

- Reduktion der Trainingszeit
- Vermeidung von Overfitting
- Interpretierbarkeit der Daten

Reduktion der Features, ohne wichtige Informationen zu verlieren (oder minimaler Informationsverlust)

KLASSIFIZIERUNG VON DOKUMENTEN

Featureauswahl

Welche Features werden benötigt?

Welche eignen sich gut?

Welche optimalen Kombinationen gibt es?

KLASSIFIZIERUNG VON DOKUMENTEN

Umgang mit vielen Wörtern

- Stopwords und unnötige Zeichen entfernen
- Wortreduktion Tf-idf (Wortfrequenz)
- Stemming (Stammformreduktion)
- Lemmatrization (Lemmatisierung)

Ziel: Reduzierung des Featureraumes auf ca. 10.000 Wörter

DOKUMENTENKLASSIFIKATION
IM UNTERNEHMESKONTEXT

Szenarien im Unternehmenskontext

- Erkennen von Bescheinigungen
- Erkennen von SPAM
- Kunden bei Formulareingaben unterstützen

AGENDA

- Einordnung ML
- Dokumentenklassifikation
- **Live-Demo**
- Herausforderungen und Fazit

DOKUMENTENKLASSIFIKATION
LIVE DEMO

Stimmungsanalyse anhand der IMDb-Filmdatenbank

Wurde ein Film positiv oder negativ bewertet?

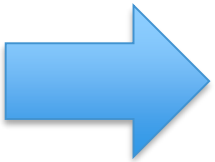
- Daten sind frei verfügbar
- Klassifizierung ist gegeben

AGENDA

- Einordnung ML
- Dokumentenklassifikation
- Live-Demo
- **Herausforderungen und Fazit**

HERAUSFORDERUNGEN UND FAZIT

- Datenbeschaffung und Datenqualität
- Feature engineering
- Rechenpower



Mit qualitätsgesicherten Daten sind gute Erfolge möglich!

VIELEN DANK
FÜR EURE AUFMERKSAMKEIT

FRAGEN?

Andreas Nadolski
Softwareentwickler

andreas.nadolski@enpit.de

Twitter: @enpit
Blogs: enpit.de/blog
medium.com/enpit-developer-blog

Karriere: www.enpit.de/karriere/

CONSULTING

enpit