

# Explainable AI - Stand der Forschung und Technik

Studiengang: MAS | Vertiefung: MAS Data Science

Machine Learning ist zu einem regelrechten Hypethema geworden. Die Qualität der Ergebnisse ist für viele Anwendungsgebiete ausreichend geworden, so dass Menschen immer öfters mit den Entscheidungen von Algorithmen konfrontiert werden. Aber wie transparent arbeiten diese eigentlich? Weshalb wird mir ein bestimmter Film empfohlen oder, gravierender, ein Kredit verweigert? Explainable AI versucht bisher unverständliche Systeme für Menschen verständlich zu machen.

## Ausgangslage

Die immer besseren Machine Learning (ML) Algorithmen in Kombination mit oftmals gigantischen Datenmengen ergeben Resultate, die für den produktiven Einsatz genügen und teilweise sogar die Leistungen von Menschen übertreffen. Aus dieser Sicht funktioniert ML einwandfrei. Problematisch wird deren Einsatz jedoch, wenn eine Entscheidung begründet werden soll. Der Verweis auf ein „durch ein computer-gestütztes Entscheidungssystem“ ist nicht befriedigend und verhindert manchmal den Einsatz von ML, beispielsweise in den Gebieten Medizin oder Justiz. Explainable AI (XAI) ist ein neues Forschungsgebiet, in welchem Methoden und Werkzeuge erarbeitet werden um das Verstehen und Begründen von ML basierten Entscheidungen zu ermöglichen.

## Ziele

Diese Arbeit soll einen Überblick über vorhandene Methoden und Werkzeuge bieten, welche ML Modelle erklären und analysieren. In einigen Beispielen werden die Anwendungen dieser Techniken erläutert. Ein weiterer Aspekt sind die Rahmenbedingungen für zukünftige ML Anwendungen, die von XAI profitieren oder diese sogar voraussetzen.

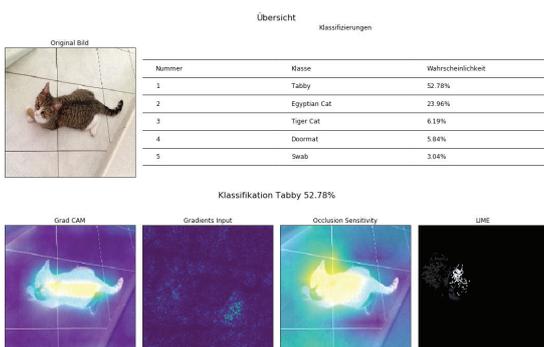
## Ergebnis

Durch die starke Nachfrage nach XAI, zum Teil gefordert durch Vorstösse von Politikern und Regierungsbehörden, ist die Anzahl neu entwickelter Methoden gross. Diese Arbeit hat einige bereits länger vorhandene Methoden an den Fallbeispielen Bilderkennung und Textanalyse angewendet. Dadurch konnte gezeigt werden, dass XAI das Verständnis für die Funktionsweise eines neuronalen Netzes und eines Random Forest Klassifikators erhöht. Durch ein absichtlich mit ungünstigen Daten erzeugtes Modell konnte auch das Prinzip des „Kluger Hans“-Effekt gezeigt werden und wie man solche Probleme erkennen und vermeiden kann.

Eine weitere Erkenntnis meiner Arbeit ist auch die Schwierigkeit eines Einsatzes dieser Werkzeuge im alltäglichen Betrieb. Zum einen sind die erzeugten Erklärungen (Visualisierungen) nicht immer einfach verständlich, zum anderen ist die Anwendung auf ein existierendes Modell schwierig aufgrund der starken Abhängigkeit von der verwendeten Software (Machine Learning Bibliotheken). Ebenso hat sich gezeigt, dass obwohl prinzipiell unabhängig, oftmals implizite Anforderungen an den Aufbau eines Modells gestellt werden, damit die verschiedenen Implementierungen der XAI Werkzeuge mit diesen Modellen umgehen können.



Marc Habegger  
marc.h@begger.ch



Visualisierte Klassifizierung in verschiedenen Verfahren

### ypres (probability 0.692) top features

Contribution	Feature
+0.005	BEAR'S
+0.005	post
+0.005	it
+0.005	most
-0.004	more positive ...
-0.004	more negative
-0.007	was
-0.008	Highlighted in text (sum)

It's an angrier review of "the **perfect** storm" (warner bros - ) in "more people die on fishing boats ... per capita, than working in any other **job** in the u.s. ... "every journey a fishing boat makes can be an all-or **nothing** risk, not a **bit** at its most exhilarating and its most terrifying." says **blaise** wolfgang petersen ("das boot"), "and that's just what he captures in this **best** story of struggle and humanity aboard a seafishing boat, the andrea gall, sailing out of Gloucester, Massachusetts, in late October, 1991. "nearly in bill willitt's screenplay, based on sebastian junger's **best**-seller, we meet the crew of six. "the veteran captain ( **best** ) is frustrated because he can't find fish on the grand banks. **best** a rival skipper (many elizabeth mastrantonio) brings in huge hauls. "this right-hand man (mark wallberg) needs money to build a new **big** with his girl-friend (diane lane). "there's a devoted dad (john c. reilly) with an estranged wife and son, a free-spirited jamaican (allen Payne), a lonely guy (john hawkes), and a last **best** replacement with a **best** attitude (william baltzer). "the skipper's command he can **best** his **best** lock straps in remote french gips, and the does "but then trouble begins. "there's a rogue wave, a man overboard and the sea machine breaks - with 60, 000 lb. "not fish that **best** scold "but that's minor compared with a deadly monster storm approaching which a boston meteorologist describes as "a disaster of epic proportions" that **best** threatens the lives of a coast guard helicopter rescue team trying to save three people stranded on a sailboat on the high seas. "it's formulaic and there are clichés, but the walls of water, created by fluid dynamics simulating real-**best** phenomena, are awesome. "on the granger movie gauge of 1 to 10, "the **perfect** storm" is a terrifying, suspenseful 8. "hanging on for the white-knuckled thrill ride of the **summer** 11".

Visualisierung Stimmungsanalyse von Filmkritiken