Ich möchte meinen Vortrag über Fair Machine Learning heute gleich mit einem Beispiel beginnen. Wir stellen uns vor, dass wir als Data Scientists damit beauftragt wurden, der New Yorker Polizei dabei zu helfen, die Kriminalitätsraten herunterzubringen. Und wir als Data Scientists behelfen uns damit, indem wir einen Machine Learning Algorithmus trainieren wollen, der vorhersagt, ob eine Person eine Straftat begehen würde ja oder nein. Und als Trainingsdaten stellt uns die New Yorker Polizei vergangene Polizeistopps zur Verfügung. Nun gibt es ein Problem, das dabei entstehen kann Die vergangenen Polizeistopps basieren eben auf die Polizeistopps. Früher basieren Spiegeln natürlich die vergangenen die Entscheidungen der Polizei in der Vergangenheit wieder und damit auch mögliche Diskriminierungen, die damit einhergeht, wie zum Beispiel Racial Profiling. Unser Algorithmus wird. Diese Strukturen wird dann sehr gut lernen, diese Strukturen zu reproduzieren und damit könnte mögliche rassistische Diskriminierung in unseren Algorithmus hereingebracht werden. Wie können wir als Data Scientists jetzt also sicherstellen, dass unser Algorithmus gerecht ist? Genau damit befasst sich das Forschungsfeld Fair Machine Learning und ich habe mir in der letzten Zeit einen Überblick darüber verschafft und möchte euch jetzt mitgeben, die wichtigsten Sachen mitgeben damit. Deshalb beginne ich meinen Vortrag damit, euch vorzustellen, wie wir im Machine Learning eigentlich über Fairness denken. Darin wird auch der Hauptfokus meines Vortrags liegen, weil wir merken werden, dass das gar nicht so einfach ist. Dann werde ich noch auf ein paar Fairnessmethoden eingehen und dann noch eine etwas kritische Anmerkung machen, indem ich auf Bias und die Feedbackloop eingehe. Im Fair Machine Learning gibt es jetzt im Prinzip zwei grundlegende Ideen über Fairness zu denken.

Gruppen Fairness und individuelle Fairness. Gruppen Fairness möchte Gleichheit zwischen Gruppen herstellen und individuelle Fairness möchte Gleichheit zwischen zwei Individuen innerhalb einer Gruppe herstellen. Und auf Basis dieser zwei grundlegenden Ideen haben sich jetzt wirklich eine zahlreiche Vielfalt von Formalisierungen ergeben, auf die wir nicht alle im Detail eingehen werden, sondern ich möchte euch eher die wichtigsten Untergruppen mitgeben und vor allem deren Intuition dahinter. Und wenn ich mal auf eine wirklich explizite Definition eingehe, werde ich einen englischen Begriff nennen, einfach um Verwirrung zu vermeiden. Genauer und einen Überblick über wirklich genaue Definitionen findet ihr dann noch im Handout, das ich euch später austeile. Wir nehmen jetzt das STANDARD. Also wir. Wir behalten bei all dem unser unser unser Beispiel im Hinterkopf. Wir wollen einen binären Klassifizierungsalgorithmus trainieren, der vorhersagt, ob die Person eine Straftat begehen wird ja oder nein? Und wir machen uns besonders Sorgen über rassistische Diskriminierung. Wir können jetzt also dahergehen und sagen Okay, wir teilen. Wir nehmen jetzt mal an, einfach um hier jetzt simpel zu bleiben, dass wir unterscheiden zwischen weißen Personen und People of Colour, also praktisch Leuten mit Migrationshintergrund. Und wir wollen eben sichergehen und können jetzt da mit dieser Idee von Gruppen Fairness herangehen und sagen Wir wollen nicht, dass Personen aufgrund ihrer Gruppenzugehörigkeit Diskriminierung erfahren und fordern deshalb Unabhängigkeit zwischen dem der Vorhersage und der Gruppenzugehörigkeit. Das nennt man sogenannte Independents und das eins der drei großen Untergruppen der Gruppen Fairness und Fair Machine Learning.

Die Gruppenzugehörigkeit, also hier als Variable a gekennzeichnet, wird jetzt definiert über das sogenannte protected Attribute. Ein protected Attribut ist im Fair Machine Learning einfach ein Merkmal, anhand dessen nicht diskriminiert werden soll. Also zum Beispiel die Ethnie. Und wir wollen jetzt eben, dass zwischen sowohl weißen als auch People of Colour gleiche Vorhersagerraten vorhanden sind. Wir können uns das ganz leicht visualisieren, indem wir einfach sagen, dass also wir nehmen jetzt einfach mal an, wir hätten 15 weiße Personen und 15 People of Color, anhand derer wir unseren Datensatz, anhand der wir jetzt Fairness definieren. Und wir wollen eben, dass sowohl 3/15 Weiße als auch 3/15 Leute mit Migrationshintergrund Als als vorher. Als kriminell vorhergesagt werden. Dann sagen wir, unser Algorithmus ist fair. Das nennen wir dann zum Beispiel Statistical parity und fordern eben gleiche Wahrscheinlichkeit für Epsilon eins zwischen unseren beiden Gruppen. Jetzt eine weitere Untergruppe von von Gruppen. Fairness ist Separation, die auch Unabhängigkeit zwischen Y, HUD und A fordert, aber dann aber bedingt auf das wahre Label und somit fokussiert sich Separation auf gleiche Fehlerraten zwischen den Gruppen. Also zum Beispiel, wenn wir sehen jetzt hier, dass wir unter den Weißen zwei fälschlicherweise als kriminell vorhergesagt haben. Und genauso wollen wir auch, dass zwischen dass unter den People of Colour zwei Leute fälschlicherweise als kriminell vorhergesagt werden. Das kann dazu führen, dass wir unterschiedliche Threshold setzen müssen, je nach Gruppe. Was uns vielleicht irgendwie intuitiv auch erstmal komisch vorkommt. Nicht so mit mit intuitiver Fairness einhergeht, aber so, dass die Folge davon.

Wir können uns dementsprechend. Hier fordern wir also zum Beispiel gleiche false positive straight, aber wir können uns eigentlich Separation und auch sufficiency die dritte große Gruppen Fairness Metrik einfach anhand der der Fehlermatrix herleiten, indem wir einfach hier vertikal bedingen und können nicht nur die gleiche false positive Rate fordern, sondern nicht die gleiche false false positive fordern, sondern auch gleiche falls negative Rate, true, positive Rate Usw. Und Predictive Parity formalisiert das zum Beispiel, indem wir eben die gleiche false positive Freiheit fordern. Die gleiche Wahrscheinlichkeit für Y gleich eins, wenn man eigentlich null ist, soll gleich sein zwischen beiden Gruppen und zu fischen. Sie lässt sich genauso selbes Prinzip anhand der der Fehlermatrix konstruieren. Indem wir jetzt aber horizontal auf die Prediction hier bedingen und damit praktisch wollen, dass die Zuverlässigkeit einer Vorhersage gleich ist. Zwischen beiden Gruppen, also zum Beispiel Predictive Parity, fordert, dass die Wahrscheinlichkeit für Y gleich eins gegeben. Man wurde als eins vorhergesagt, also geben man wurde als kriminell vorhergesagt. Wie wahrscheinlich ist es dann, dass man wirklich kriminell ist? Das soll gleich sein zwischen beiden Gruppen. Somit sowohl Separation als auch sufficiency lassen sich jetzt anstatt mit Y hud auch mit es formulieren, sofern wir das haben. Es ist ja einfach unser Score. Den haben wir ja meistens, zum Beispiel beim logistischen Modell. Ist ja nicht so, dass wir nur die Vorhersage eins oder null kriegen, sondern wir kriegen die Wahrscheinlichkeit der positiven Gruppe angehören. Das wäre jetzt in dem Fall S und so können wir zum Beispiel Kalibration fordern und eben anstatt auf Y Hut zu Bedingungen wie in zu fischen sie auf S bedingen und hier Gleichheit zwischen den Gruppen fordern.

Oder auch Well. Kalibration ist eine Abwandlung von von Kalibration, die einfach noch ein bisschen stärker ist und eben auch auf das S bedingt. Zusammengefasst können wir also sagen, dass Zufficiency besonders jetzt die Perspektive des Entscheidenden annimmt. Also zum Beispiel der Polizei, die zu dem Zeitpunkt, zu dem sie entscheiden soll, ob jemand eine kriminell werden würde, ja noch nicht das wahre Label kennt und deshalb noch nicht das wahre Label kennt. Separation ist Bei Separation bedingen wir ja oft das wahre Y und deshalb ist Separation besonders gut, wenn wir sicher gehen können, dass dieses wahre Y durch einen objektiven, wahren Prozess entstanden ist und keinerlei Bias mit sich trägt, was in unserem Fall vielleicht auch ein bisschen kritisch werden könnte, worauf ich später nochmals zu sprechen komme. Und Independence ist vor allem gut, wenn eine Form der Gleichheit zwischen Gruppen erzwungen werden soll. Also egal, egal was. Es sollen immer 50 % als als kriminell vorhergesagt werden. Indem wir. Die Gruppen Metriken sind. In gewisser Weise die einfachsten Metriken. Sie sind leicht zu verstehen anhand der Fehlermatrix ist jedem Statistiker, jeder Statistikerin normalerweise bekannt. Und damit sind die groben Metriken auch in bekannten Softwarepaketen implementiert, wie zum Beispiel in Matrix Fairness und werden dort dann oft nicht einzeln herausgegeben. Also nicht die einzelnen true positive Trades zwischen den Gruppen, sondern als Differenzen.

Und diese sollen dann halt null sein. Wir wollen da null oder auch als Quotienten. Wir wollen eins erreichen. Sie sind simpel, praktisch. Indem wir aber Gleichheit zwischen Gruppen in gewisser Weise erzwingen oder herstellen wollen, kann es dazu kommen, dass vereinzelte Personen ungerecht behandelt werden. Wir hatten das vorher in der Grafik in gewisser Weise schon gesehen. Wir müssen eben unterschiedliche Thresholds setzen und so, irgendwie werden jetzt vereinzelte Personen, die eigentlich negativ sind, als positiv vorhergesagt. Einfach nur, weil wir eben gleiche falsch positiv Raten zwischen unseren Gruppen haben wollen. Independence ist jetzt. Deshalb gibt es in gewisser Weise die, die die individuellen Fairness Metriken, die jetzt sagen. Fairness ist für mich, wenn zwei ähnliche Individuen ähnlich behandelt werden. Fairness through awareness fta formalisiert es im Prinzip, indem wir einen Lipschitz Kriterium fordern. Und das macht nichts anderes, als uns einfach die Distanz zwischen der Vorhersage von zwei Individuen nach oben hin zu beschränken, und zwar in Abhängigkeit von unseren Features. Also wenn sich zwei Personen sehr ähnlich sind, wenn sie aus demselben Grund irgendwie von der Polizei gestoppt wurden, wenn sie dieselbe Criminal History haben, wenn Sie beide eine Waffe besitzen, dann aber eine unterschiedlichen Ethnie angehören. Die einen sind weiß, die andere ist es schwarz, dann wollen wir nicht, dass diese beiden eine unterschiedliche Vorhersage kriegen. Sonst würden wir den Algorithmus als unfair betrachten. Die Schwierigkeit, die FDA mit sich bringt, ist die Definition genau dieser Gleichheit, also des Distanzmaßes im Feature Space.

Wann sagen wir, dass zwei Individuen gleich sind? Reicht es, wenn sie am gleichen Wohnort leben? Oder ist es sinnvoller, auch die Criminal History in Betracht zu ziehen? Das ist nicht eindeutig und sollte in jedem Fall von FDA auf die bestimmte Situation abgestimmt werden und auf das bestimmte protected Attribute, je nachdem, was eben am meisten Sinn macht in diesem Kontext. Und somit steht FDA in gewisser Weise konträr zu dem zweiten großen individuellen Fairnessansatz FTu fairness through own awareness oder auch Blinding genannt, weil hier wollen wir im Prinzip, dass das protected Attribute einfach gar keine Rolle in unserem ganzen Entscheidungsprozess spielt und auch darin, wie wir diesen Algorithmus jetzt designen. Also es ist eher eine Vorgehensvorschrift als wirklich eine mathematische Definition, die es zu erfüllen gilt, in dem uns FTU sagt, dass das protected Attribute eben nicht im Entscheidungsprozess verwendet werden soll. Also würden wir jetzt zum Beispiel, wenn wir eine logistische Regression oder ein Random Forest trainieren, auf unseren Trainingsdaten würden wir dort die Ethnie der Person löschen oder halt irgendwie verbergen während dem Training, so dass unser Klassifizierungsalgorithmus gar nicht darauf zurückgreifen kann. Ftu hat das Problem der Proxies. Das sind Variablen, die stark mit dem Protected Attribute korreliert sind. Im Normalfall reicht es nämlich nicht, einfach zu sagen ich lösche jetzt die Ethnie aus meinem Datensatz, dann kann mein Algorithmus ja gar nicht anhand dessen entscheiden, weil es eben sein kann, dass der Wohnort extrem stark damit korreliert ist, welche Ethnie die Person hat. Und so bleibt praktisch eine indirekte Diskriminierung anhand der Ethnie bestehen, obwohl wir die Ethnie ja aus unserem Datensatz gelöscht haben.

Dann gibt es Ansätze, die sagen gut, dann löscht man eben auch alle Proxys aus dem Datensatz, aber das hat dann wieder andere Probleme, wie das der da, das unser Algorithmus dann immer schlechter wird, dass wir irgendwann auch sehr viele Variablen einfach löschen Und das alles ist eben sehr uneindeutig und ist nicht leicht zu lösen. In gewisser Weise wird diese dieser Fakt, dass wir oftmals komplexe Strukturen in unseren Daten haben, komplexe Zusammenhänge. Jetzt durch die dritte große Kategorie von Fairness Definitionen aufgenommen. Das sind die kausalen Definitionen von Fairness. Diese Definitionen werden Literatur als dritte große Kategorie festgestellt, was auch so passt. Es gibt sowohl kausale Gruppen, Fairnessdefinitionen als auch kausale individuelle Fairnessdefinitionen. Aber sie sind noch mal eine dritte große Kategorie insofern, als dass sie im Prinzip wirklich versuchen, als dass sie versuchen, praktisch diese Abhängigkeitsstrukturen in den Daten zu verstehen und auch modellieren zu wollen. Und in und das wollen jetzt sowohl die bisher vorgestellten individuellen und Gruppen Fairness Metriken nicht, sondern sie sind eher observierend und deskriptiv in dem Sinne, dass wir hier jetzt keine Annahmen über die Daten machen. Ich kann jetzt an der Stelle nicht noch tiefer, nicht tiefer ins Detail, in die kausalen Methoden gehen, weil sie einfach doch noch mal sehr umfangreich sind. Wir widmen uns jetzt eher den Fairnessmethoden. Was ist jetzt also, wenn wir festgestellt haben okay, unser Algorithmus erfüllt Keine dieser erfüllt eine dieser gewählten Fairnessmethoden nicht.

Wie machen wir ihn jetzt fairer? Es gibt im Prinzip drei Ansätze, die sich unterteilen lassen in Preprocessing, Inprocessing oder Postprocessing. Je nachdem, wann sie in der Machine Learning Pipeline stattfinden. Preprocessing Methoden haben die Idee, dass die Daten vor dem Training bearbeitet werden sollen, so dass unser Algorithmus zum Beispiel praktisch auf korrigierten Daten lernt und somit keine Diskriminierung stattfinden kann. Das kann zum Beispiel durch Sampling oder auch Transformationen entstehen. Wir haben in Processing Methoden, für die wir Zugriff auf den Algorithmus haben brauchen, weil hier der Ansatz ist, dass wir wirklich das Optimierungsproblem an sich modifizieren, indem wir wirklich die Loss Function nehmen und zum Beispiel einen Regularisierungsterm anhängen und somit unser Algorithmus nicht nur für hohe Treffsicherheit optimiert, sondern auch für eine der oben genannten Fairness Metriken. Postprocessing klappt jetzt wieder mit Blackboxalgorithmen, genau wie Preprocessing, weil wir hier einfach die fertigen Labels nehmen, die unser Algorithmus vorhergesagt haben und auch die Scores und zum Beispiel individuelle Gruppen Thresholds setzen, um dann Predictive Equality oder was auch immer gerecht zu werden. Also hier ist es auch wieder so, dass es wirklich jetzt verschiedenste Methoden gibt, manche auf bestimmte Algorithmen angepasst, manche eher universeller. Es gibt bei all diesen Methoden und all diesen Fairnessdefinitionen ist es wirklich, kann man sich leicht verlieren, vor allem eben, wenn man das Ganze anwendet. Es ist deshalb ganz wichtig. Es ist aber bei diesem Thema Fairness ganz wichtig, dass man sich nicht in Details, dass man immer noch mal einen Schritt zurücktritt und seine Daten und seinen Algorithmus in einem größeren Kontext betrachtet.

Man kann sich das Ganze als eine Art Kreislauf vorstellen, auch Feedbackloop genannt, der aus Personendaten und Algorithmus besteht. Weil es ist ja jetzt so, dass unser Algorithmus von Daten lernt, diese Daten aber natürlich praktisch unsere Realität, also unsere Gesellschaft irgendwie in gewisser Weise widerspiegeln und und wir dann anhand dieses trainierten Algorithmus auch wieder Entscheidungen treffen. Und so haben wir praktisch eine Art Feedbackloop, eine Art Kreislauf, an jeder dieser drei Stellen kann kann Bias in den Prozess gebracht werden und vor allem kann Bias auch verstärkt werden im Laufe dieses Prozesses. Also zum Beispiel schauen wir uns das Ganze nochmal ein bisschen genauer an. Also schauen wir uns das Ganze nochmal. Veranschaulicht, wie wir uns das Ganze vielleicht noch mal anhand unseres Beispiels von vorher auch. Also es ist ja jetzt so, also es ist an also wir können uns vorstellen, dass die Polizei eben das in der Vergangenheit eben viel Diskriminierung, dass viel Diskriminierung auch der Polizei gegenüber, gegenüber People of Colour entsteht und und somit zum Beispiel People of Color viel öfter gestoppt werden von der Polizei, viel mehr Polizeipräsenz, einfach in bestimmten Regionen, wo People of Color leben stattfindet und wir somit viel höhere Kriminalitätsraten bei People of Colour eben aufzeichnen. Und das spiegelt sich dann. Und so bringen wir als Personen, als Gesellschaft, in dem wir praktisch diskriminierende Entscheidungen treffen, wobei es in diesen Kreislauf in Form von zum Beispiel historischem Bias Es kann auch sein, dass durch die Daten Bias in diesen Kreislauf gebracht wird, in dem zum Beispiel unsere Daten durch einen verzerrten Prozess gesampelt werden und nicht unsere Grundgesamtheit, unsere wahre Population eigentlich widerspiegeln, also wir überproportional weiße Leute in unserem Datensatz haben und deshalb die Gesichter dann von Weißen später besser erkannt werden als von Schwarzen, einfach weil unser Sample nicht die eigentlich die Weltbevölkerung widerspiegelt.

Oder eben der Algorithmus verwendet schlicht einfach verzerrte Schätze und bringt somit selbst bei korrekten Daten und keinem historischen Bias noch Diskriminierung in den Kreislauf. Also wir haben uns jetzt das Thema Fair Machine Learning an dem einfachsten Fall einer binären Klassifikation und einem Protected Attribute im Prinzip veranschaulicht. Natürlich ist es in der Realität oftmals komplexer. Personen gehören nicht nur. Wir haben mehrere protected attributes Personen gehören verschiedensten Gruppen an und wir haben natürlich vielfältigere Aufgaben wie Regressionen und Supervised Learning, wo wir auch Fairness sicherstellen wollen. In der Forschung gibt es also noch viele weitere, Gibt es also noch viele Erweiterungen davon. Und in der Forschung gibt es also noch viele Erweiterungen von den jetzt bekannten Fairnessdefinitionen. Und dazu kommt eben auch, dass wir eben immer unseren Kontext und die Daten in. Und deshalb gibt es eben noch nicht die eine Definition von Fairness. Und es ist immer wichtig, die den bestimmten Kontext, die Aufgabe und die Daten im Blick zu behalten. Damit hoffe ich, dass ich eure Neugierde auf das Thema wecken konnte und bedanke mich für eure Aufmerksamkeit.

(Transcribed by [Sonix.ai](https://sonix.ai/?ref=mkt_docx_export) - Remove this message by [upgrading your Sonix account](https://sonix.ai/pricing?ref=mkt_docx_export))