**Intro**

"Stellen Sie sich vor, wir als Data Scientists sollen der New Yorker Polizei helfen, die Kriminalitätsraten zu senken. Dafür trainieren wir einen Machine-Learning-Algorithmus, der vorhersagt, ob eine Person eine Straftat begehen könnte. Die Polizei stellt uns historische Daten zu Polizeistopps zur Verfügung.

Doch hier entsteht ein Problem: Diese Daten spiegeln die Entscheidungen der Polizei in der Vergangenheit wider – und möglicherweise auch deren Vorurteile, wie etwa Racial Profiling. Wenn wir diese Daten nutzen, könnte unser Algorithmus solche diskriminierenden Strukturen übernehmen und verstärken.

Wie können wir also sicherstellen, dass unser Algorithmus fair bleibt? Genau damit beschäftigt sich das Forschungsfeld Fair Machine Learning."

Damit beschäftigt sich FairML. Ich habe mich über die letzte Zeit mit dem Bereich FairML beschäftigt, euch überblick geben : Agenda

Zurück zu unserer Problematik: wir wollen gerecht Krinminalität vorhersagen, aber haben Daten, die vergangene rassistische Diskriminierung widerspiegeln. Bevor wir irgendwas tun können, müssen wir erstmal für uns definieren, was Fairness denn ist.

**2.1 Gruppen Fairness - Independence**

Verständnis von Fairness, das zugrunde liegt: Es ist unfair, wenn Personne aufrgrund ihrer Gruppenzugehörigkeit (definiert durch PA) Diskriminierungen erfahren

Gruppenzugehörigkeit soll unabhängig von Vorhersage sein

z.B. Demographic parity fordert gleiche positive Vorhersageraten zwischen Gruppen d.h. in Gruppe a wird der gleiche Anteil als als positiv vorhergesagt wie in Gruppe b

Bild Beschreibung: unterschiedliche thresholds für jede gruppe könnte nötig sein

**2.2 Gruppen Fairness – Separation + Sufficiency**

Sowohl Separation als auch Sufficiency können anhand Fehlermatrix hergeleitet werden

Bei Separation bedingen wir auf wahres Label Y 🡪 Fokus auf Fehlerraten

Für Sufficiency bedingen wir auf Vorhersage 🡪 Fokus auf Zuverlässigkeit der Vorhersage

In Software packeten umgesetzt

Als Differenz oder Quotient

Simpel, praktisch, Gleicheit zwischen Gruppen zwingt uns vereinzelte Personen ungerecht zu behandeln 🡪 Individuelle Fairness

**3. Individuelle Fairness**

FTA

FTU: An sich erstmal eher eine Vorgehensvorschrift

Messbar machen: wenn modell access

Komplexe Strukturen 🡪 kausale Methode

In unserem Szenario vermutlich historischer Bias im Spiel: höhere Kriminalitätsraten weil mehr Polizei in Afro amerikanischen Nachbarschaften, strengere Kontrollen, niedriger ökonomischer Status, weniger staatliche Hilfe 🡪 wahres label, also von historischem bias infiziert

Zusätzlich auch selection bias: unsere stichprobe nicht repräsentative gezogen

Algortihmischer bias: je nach modell