DISKRIMINIERENDE ALGORITHMEN: TECHNISCHE ODER SOZIALE FRAGE?

Winterkongress Digitale Gesellschaft Schweiz 2021
26.02.2021
Linus Petermann & David Sommer

Automatisierte Selektion in Bewerbungsverfahren

- Einfaches Beispiel:
  - *Algorithmus: Vorselektion Bewerbungsgespräche*
  - *Trainiert an alten Bewerbungsunterlagen, sucht Muster für «eingestellt»*
  - *Optimales Muster «eingestellt»: Schweizer Pass, Wohnort Zürich, Zweisprachig*
- Aktuelle Bewerber*innen:

Nachname	*Nationalität*	*Letzter Bildungsabschluss*	*Sprachen*	*Wohnort*	*#calc Distanz zu Arbeitsort (km)*
Pasquinelli	Schweiz	MA Universität Zürich	Deutsch, Englisch	Zürich	2
Uljanow	Schweiz, Russland	MA ZHAW	Deutsch, Englisch, Russisch	Spreitenbach	10
Lopez	Spanien	MA Universität Madrid	Deutsch, Englisch, Spanisch	Spreitenbach	10

Direkte und indirekte Diskriminierung

- Selektion nach Nationalität → Direkte Diskriminierung
  - *Lösung: Attribut entfernen*
- Selektion nach Wohnort → Indirekte Diskriminierung
  - *Starker Zusammenhang mit nicht-schweizerischer Nationalität*
  - *Lösung?*

→ Reproduktion sozialer Normen

Nachname	*Nationalität*	*Letzter Bildungsabschluss*	*Sprachen*	*Wohnort*	*#calc Distanz zu Arbeitsort (km)*
Pasquinelli	Schweiz	MA Universität Zürich	Deutsch, Englisch	Zürich	2
Uljanow	Schweiz, Russland	MA ZHAW	Deutsch, Englisch, Russisch	Spreitenbach	10
Lopez	Spanien	MA Universität Madrid	Deutsch, Englisch, Spanisch	Spreitenbach	10

Wandelnde Normen, unkritische Algorithmen

- Früher: Ungleichbehandlung von Frauen in jungem Alter
  - *Gerechtfertigt und mit «sachlichem Grund»*
- Heute: Rechtfertigung wird zunehmend prekär

Nachname	*Nationalität*	*Letzter Bildungsabschluss*	*Sprachen*	*Geschlecht*	*Alter*
Pasquinelli	Schweiz	MA Universität Zürich	Deutsch, Englisch	w	30
Uljanow	Schweiz, Russland	MA ZHAW (Fachhochschule)	Deutsch, Englisch, Russisch	m	32
Lopez	Spanien	MA Universität Madrid	Deutsch, Englisch, Spanisch	w	43

Soziale Prägung der Technik

- Diskriminierung nicht «nur» technisches Problem
- Bestimmung von Diskriminierung ist offen und wird gesellschaftlich ausgehandelt
  - *Was ist «sachlich»?*
- Vorstellungen und Ordnung von Welt prägen Algorithmen
  - → *Immer eine Übertragung von Werten und Normen in Technik*

Aufbau

- KI-Systeme: Begriff, Einsatz, Verbreitung
- Technische Grundlagen: Funktionsweise, Problematiken und Widersprüche
- Rechtliche Grundlagen und mögliche Rahmenbedingungen
- Was tun? Perspektiven für Gesellschaft und Politik
- Fazit

ADM Systeme: Begriff, Einsatz

- KI, AI, Machine Learning, Neuronale Netzwerke…
  - *Unklare / verschwommene Definitionen*

→ Automated Decision Making (ADM Systeme)

- Heutiger Haupteinsatzzweck
  - *Effizienzsteigerung in Prozessen*
    - Ermöglicht Verarbeitung viel grösserer Datenmengen
  - *"Faire" Beurteilung*
    - Anstatt viele varierende Expertenbeurteilungen
    - Z.B.: Matura

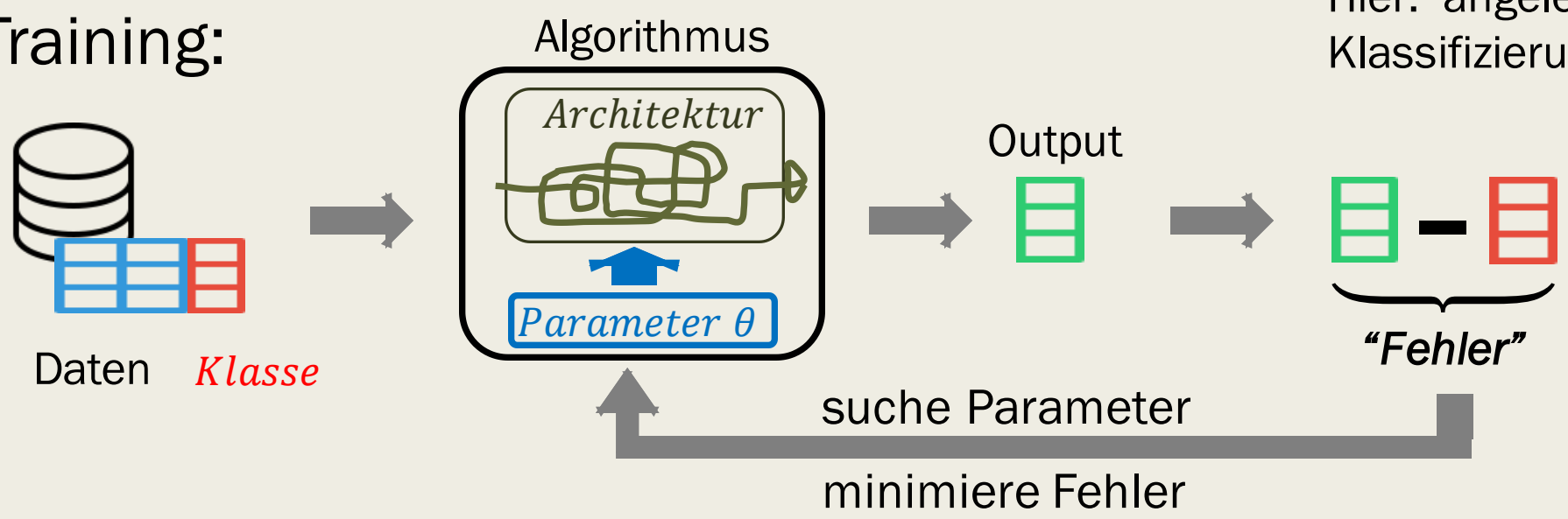
ADM Systeme: Verbreitung

- Grosse Bandbreite
  - Automatisierung von Entscheidungs und Dienstleistungsprozessen
- Unterschiedliche Folgen und Diskriminierungseffekte
- Beispiele:
  - *Amazon Bewerbungsverfahren*
    - Ausschluss nicht-männlicher Personen (2018)
  - *Kindsgefährdungsanalysen New York*
    - Unterstützung von Präventiveingriffen
  - Zuteilung von Schulhäusern
    - Bessere soziale Durchmischung
    - Derzeit in Zürich

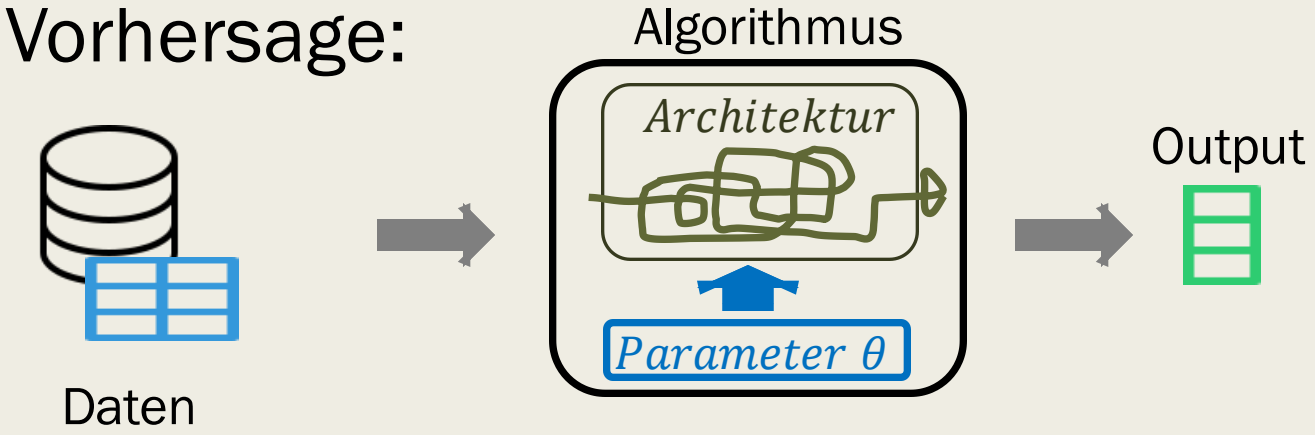
Wie funktionieren diese Algorithmen?

Hier: angeleitete (supervised) Klassifizierung

Training:



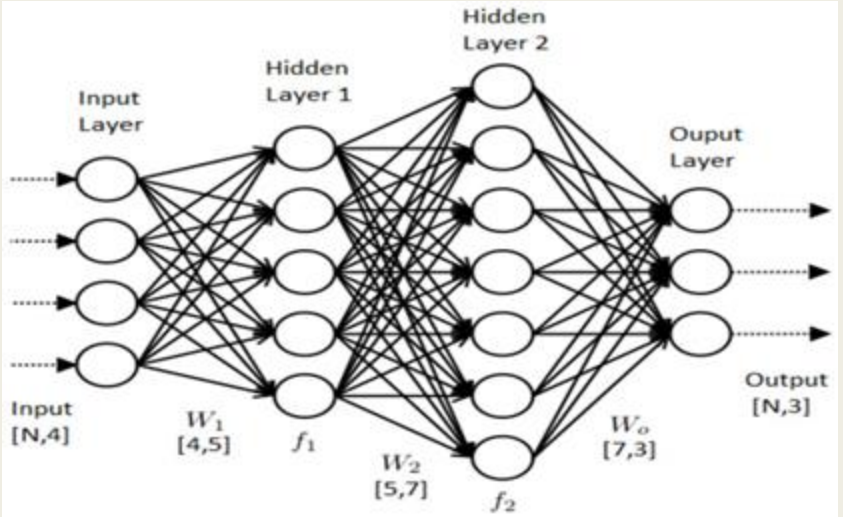
Vorhersage:



Qualitätsmessung, oft: Vorhersagegenauigkeit (im Schnitt)

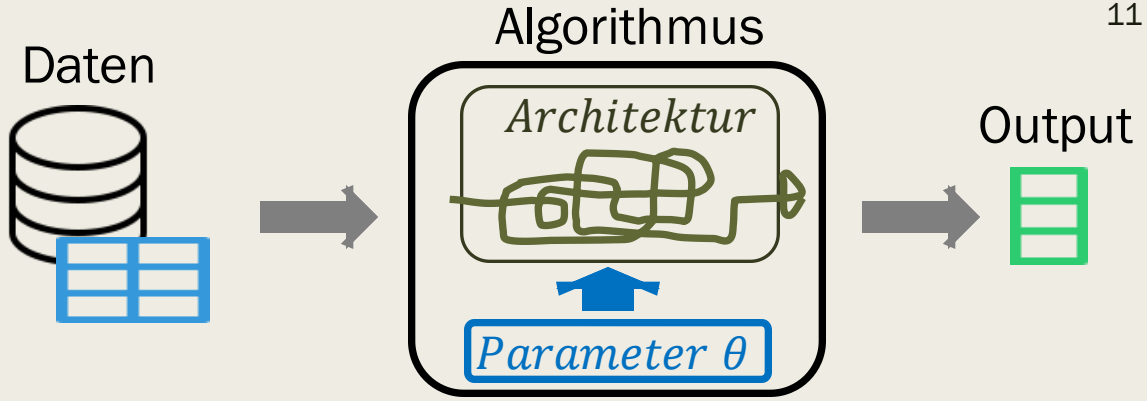
Kausalität

- Kausalität:
  → *Warum ein Algorithmus wie klassifiziert*
- Sinn ist nicht immer ersichtlich
  → Algorithmus findet unbekannte Korrelation
- Kann auch falsche Schlüsse ziehen
  → Wolf wird an Schnee im Hintergrund erkannt
- Zunahme Komplexität: Z.B. Neuronale Netze:
  → Milliarden an Parameter
  → Kausalität schwierig



Quelle Bilder:
https://s3.amazonaws.com/media.eremedia.com/wp-content/uploads/sites/4/2019/12/16105353/AI-wolf-or-husky.jpg
https://miro.medium.com/max/700/1*ZB6H4HuF58VcMOWbdpcRxQ.png

Bias (Verzerrung)



- Rohdaten beinhaltet bereits Diskriminierung
  - *Früher wurden mehr Schweizer eingestellt*
    ⟶ Algorithmus empfiehlt eher Schweizer
- *Population Bias*: Untergruppen waren nicht Teil des Trainings:
  - *Seifenspender funktioniert bei dunkler Hautfarbe nicht*
- *Representation Bias*: Untergruppen werden vernachlässigt
  - *Niedriger Anteil ⟶ grosser "Fehler" Einzelner im Schnitt vernachlässigbar*
- *Algorithmic Bias*: Architektur-/Design-Entscheidungen
  - *Algorithmus schafft es nicht, auf wichtige Details einzugehen*
  - *Definition von Fehler hat enorme Auswirkungen auf Resultat*

Fairness?

- Angesprochene Algorithmen unterscheiden zwingend
- Wie bekommt man diese “fair”
  - *Anpassung der Rohdaten*
  - *Korrektur beim „Fehler“*
- Was ist fair?
- Individuelle vs. Gruppen-Fairness:
  - *Bsp: Priorisierte Einstellung von Frauen in Männerdomäne*
  - *Für Parität im Schnitt: Frau hat deutlich höhere individuelle Einstellungs-Chancen*
- Generell: Prioritisierung eines Zusammenhangs schwächt meistens einen anderen
  - Grund: andere Daten-Ausgangsstruktur –> andere Dynamik
- Trade-off zwischen Effizienz und Fairness
  - *Schwächt Nützlichkeit des Algorithmus*
- Algorithmen zwar neutral, reproduzieren aber Datenstruktur und "künstliche" Zwänge

→ Fairness ist technisch nicht eindeutig, benötigt gesellschaftliche Diskussion

Stand Regulierung Schweiz und Europa

- Brandaktuelles Thema: Noch kein Grundkonsens
- Lösung über Datenschutz?
- Schweiz: Neues Datenschutzgesetz, Art. 21 (ab 2022)
  - Recht auf Informierung und Überprüfung bei / von vollautomatisierten Entscheiden
- Europäische Union, DS-GVO Art. 22 Abs. 1
  - *Recht auf nicht ausschließlich automatisierte Verarbeitung bei rechtlicher Wirkung*

    → Enge Definition, schränkt nur Extremfälle ein
- Aber: Datenschutz reicht nicht als Rechtsrahmen
  - *Bsp.: Optimale Platzierung von Feuerwehrwachen (basierend auf öffentlichen Daten)*
- ADM unabhängig vom Datenschutz: Leitlinien für Bundesverwaltung
  - Spricht wichtige Punkte an, guter Schritt
  - Schwammig formuliert (da noch unbekanntes Terrain)
  - Gilt nur für Bundesverwaltung!

Allgemeine Regulierungsansätze

- Transparenz beim Einsatz
  - *Kennzeichnung, komplette Veröffentlichung?*
  - *Komplett für staatliche Systeme? Strafverfolgung?*
  - *Nur vollautomatisierte Systeme?*
- Kritikalitätsansatz: Anforderungen je nach Schadensrisiko
  - *Wie evaluieren?*
  - *Zulassungsbewilligungen wie in Pharmaindustrie?*
- Klare Haftung, von Menschen oder Unternehmen
  - *Keine Unterscheidung zwischen Entscheiden von Mensch oder Maschine?*
  - *Keine Auslagerung an andere Juristriktionen (Google vs. Spanien,* CJEU, 2014*)*
- Forderung nach Möglichkeiten zum (öffentlichen) Testen solcher Systeme.
  - Standartisierte Schnittstelle?

Gesellschaftlicher Umgang mit ADM Systemen

- Weiterentwicklung und schnelle Verbreitung von ADM Systemen absehbar
- **Schleier der Neutralität und Technikgläubigkeit aufbrechen:**
  - *Werte- und Normübertragungen in Technik sichtbar machen und reflektieren*
- Gefahr: Verstärkung sozialer Normen, versteckte Diskriminierung
  - *Algorithmen ohne «Gewissen» → Gefahr der Verstärkung sozialer Normen*
  - *Konventionsdruck*
- ADM-Systeme zum Teil von Debatten machen:
  - *Datenbasis: Welche/Wessen Daten, welche Filter*
  - *Struktur/Architektur: Art des Algorithmus, Erneuerungszeitpunkte*
  - *Einsatz: Wann, Wofür, Wie? → Ob überhaupt?*
- Seiteneffekt: verbesserte statistische Erkennbarkeit von Diskriminierung

Fazit

- Reproduktion und Verstärkung sozialer Normen und bestehender Ungleichbehandlung
  - *Diskriminierende Gesellschaft → diskriminierende Daten & Algorithmen*
- Technische Lösungen beschränkt
  - *Diskriminierung technisch nicht komplett auflösbar*
  - *Widersprüche in Fairnessdefinitionen*
  - *→ Gesellschaftliche „Begleitung“ und Diskussion technischer Entwicklung notwendig*
- Technikgläubigkeit verführt zur Verantwortungsabgabe
- Regulierungsrahmen ungenügend / nicht vorhanden
  - *Intensive Diskussionen, Regulierungen absehbar*
  - *Zivilgesellschaft vs. Unternehmens- und Regierungsbestrebungen stark asymmetrisch*
  - *Rechte müssen erkämpft werden*

- Hauptsächlich politische Frage

Weiterführende Links und Literatur

Hagendorff, Thilo (2019): *Maschinelles Lernen und Diskriminierung: Probleme und Lösungsansätze*, in: *Österreichische Zeitschrift für Soziologie*, Jg. 44, Nr. 1, S. 53–66, doi: 10.1007/s11614-019-00347-2.

Orwat, Carsten (2019): *Diskriminierungsrisiken durch Verwendung von Algorithmen*, (Studie) Berlin: Antidiskriminierungsstelle des Bundes.

Automating Society Report Schweiz (2020) - AlgorithmWatch /CH

Sommer, David (2021): *Machine-Learning-Leitlinien des Bundes*, [online] https://www.digitale-gesellschaft.ch/2020/12/17/machine-learning-leitlinien-des-bundes-entscheidende-gefahren-nicht-benannt/

Backups

Example representation bias

Quelle: Mehrabi, Ninareh, et al. "A survey on bias and fairness in machine learning." *arXiv preprint arXiv:1908.09635* (2019).