DAGStat Stellungnahme: Die Rolle der Statistik in der Künstlichen Intelligenz

Version 03.03.2020

Einleitung

Die Entwicklung und Anwendung von Künstlicher Intelligenz (KI) hat eine umfassende wissenschaftliche, wirtschaftliche, gesellschaftliche und politische Diskussion ausgelöst. Die Statistik als fächerübergreifendes Wissensgebiet spielt eine substantielle Rolle sowohl für das theoretische und praktische Verständnis von KI als auch für deren weitere Entwicklung. Als Beitrag für die aktuelle Diskussion hat die Deutsche Arbeitsgemeinschaft Statistik (DAGStat) ihre Position zu KI in diesem Papier näher beschrieben. Neben den statistischen Methoden und der Betrachtung von oft vermeidbaren Fehlern bei Planung und Ausführung von KI-Anwendungen wird auch auf die ebenso notwendige wie sinnvolle Erweiterung der Curricula im schulischen und universitären Bereich eingegangen.

Entgegen der öffentlichen Wahrnehmung ist KI kein neues Phänomen. So wurde KI bereits 1956 auf der Dartmouth Conference erwähnt [53, 54]. Die ersten datengetriebenen Algorithmen wie Perceptron [77] und Backpropagation [78] wurden in den 50er und 60er Jahren entwickelt. Der Lighthill-Report 1973 fällte ein überwiegend negatives Urteil zur KI-Forschung in Großbritannien und führte dazu, dass die finanzielle Unterstützung für KI-Forschung weitgehend eingestellt wurde (sogenannter erster KI-Winter). Die folgende Phase überwiegend wissensbasierter Entwicklung endete 1987 mit dem sogenannten zweiten KI-Winter. 1988 publizierte Judea Pearl sein Buch „Probabilistic Reasoning in Intelligent Systems“, für das er 2011 mit dem Turing Award ausgezeichnet wurde [80]. Ab Anfang der 90er Jahre entwickelte sich KI wieder datengetrieben mit großen Durchbrüchen wie Support Vector Machines [81], Random Forest [82], Bayesianischen Methoden [136], Boosting und Bagging [97, 98], Deep Learning [83] und Extreme Learning Machines [99].

Inzwischen spielt die KI in vielen Lebensbereichen eine zunehmend bedeutendere Rolle. Internationale Organisationen und nationale Regierungen haben sich aktuell positioniert bzw. den ordnungspolitischen Rahmen beschrieben. Beispielhaft seien hier die KI-Strategie der Bundesregierung [1] und die Stellungnahme der Datenethikkommission [107] aus dem Jahr 2019 genannt. Des Weiteren beschäftigen sich nun auch Zulassungsbehörden wie die US Food and Drug Administration (FDA) mit KI-Themen. So wurde 2018 mit der EKG-Funktion der Apple Watch zum ersten Mal eine KI-Anwendung durch die FDA genehmigt [84].

Eine einheitliche Definition des Begriffs Künstliche Intelligenz existiert bis heute nicht. Die DAGStat orientiert sich daher am Verständnis der Bundesregierung, das der KI-Strategie Deutschlands zugrunde liegt [1]. In diesem Sinne bezeichnet KI (genaugenommen: schwache KI) ein künstliches, maschinelles System, das ein konkretes Anwendungsproblem so gut oder besser lösen kann wie ein Mensch. Ein wichtiger Aspekt eines KI-Systems ist, dass es selbstlernend ist. Im vorliegenden Papier werden wir uns auf die datengetriebenen Aspekte der KI fokussieren. Darüber hinaus gibt es in der KI vielfältige Bereiche, die sich mit der Verarbeitung und Inferenz aus symbolischen Daten beschäftigen, auf die wir hier nicht näher eingehen [55].

Wie für die KI, so gibt es auch für den Bereich des maschinellen Lernens (ML) in Literatur und Praxis weder eine einheitliche Definition noch eine einheitliche Zuordnung von Methoden zu diesem

Bereich. Legt man Simons Definition von 1983 zugrunde [2], so bezeichnet Lernen Veränderungen eines Systems derart, dass eine gleichgelagerte Aufgabe bei der nächsten Durchführung effektiver oder effizienter durchgeführt werden kann.

Häufig werden die Begriffe KI und ML in einem Atemzug mit Big Data und Data Science genannt und manchmal sogar synonym verwendet. Hier gilt allerdings: weder sind KI-Methoden zwingend nötig, um Big Data Probleme zu lösen noch sind Methoden aus dem KI-Kontext lediglich bei Vorliegen von Big Data anwendbar. Data Science andererseits wird häufig als Schnittbereich von Informatik, Statistik und der jeweiligen Fachwissenschaft betrachtet und ist somit nicht an die Nutzung bestimmter Methoden oder das Vorliegen bestimmter Datensituationen gebunden.

Je nach Aufgabe des KI-Systems kommen verschiedene ML-Ansätze zum Einsatz, die zum Beispiel auf Regression (für Fragen im stetigen Spektrum) oder Klassifikation (für die Zuordnung zu einer von endlich vielen Klassen) basieren. Wichtige Kategorien von ML-Ansätzen sind Supervised Learning, Unsupervised Learning und Reinforcement Learning [56]. Viele KI-Systeme verarbeiten Trainingsbeispiele mit vorgegebenen Lösungen. Man spricht dann von Supervised Learning, während beim Unsupervised Learning keine Lösungen gegeben sind. Die Inputdaten können Messwerte, Börsenkurse, Audiosignale, Klimadaten oder Texte sein, aber auch sehr komplexe Zusammenhänge beschreiben wie z.B. Schachspiele. Tabelle 1 gibt eine vereinfachte Übersicht über Beispiele statistischer Verfahren und Modelle, die in KI-Systemen Anwendung finden. Hierbei wurde der Übersichtlichkeit halber bei der Klassifikation auf die Unterscheidung zwischen datenbasierten und theorie-basierten Modellen verzichtet.

Auch wenn die Entwicklung von KI-Systemen vor allem in der Informatik beheimatet ist, hat die Statistik schon früh eine wichtige Rolle gespielt, nachdem beispielsweise Zusammenhänge zwischen Backpropagation und nichtlinearen Kleinste-Quadrate-Methoden erkannt wurden. Wichtige Methoden des maschinellen Lernens, die eine ausgezeichnete Rolle in der KI spielen, wie etwa Random Forests oder Boosting wurden von Statistikern entwickelt. Andere, z.B. Radial Basis Function Networks [106], können auch als nichtlineare Regressionsverfahren betrachtet und studiert werden. Auch neuere Entwicklungen wie die Extreme Learning Machines oder Broad Learning Systems [104] weisen enge Bezüge zu statistischen Methoden wie der multiplen, multivariaten Regressionsschätzung und Ridge Regression auf. Auch die theoretische Untermauerung der Validität von maschinellen Lernmethoden, etwa durch Konsistenzaussagen und Generalisierungsschranken [57, 105], erfordert substantiell Erkenntnisse der mathematischen Statistik.

Soll eine Fragestellung empirisch untersucht werden, sind eine Reihe von Schritten nötig, wie Abbildung 1 veranschaulicht. Der Prozess beginnt mit der präzisen Formulierung der Fragestellung und führt dann weiter über das Studiendesign (inklusive Fallzahlplanung und Bias-Kontrolle). Daran schließt sich dann die Analyse an. Abschließend sind die Ergebnisse der Analysen zu interpretieren. Die KI fokussiert sich häufig auf den Schritt der Datenanalyse, während relevante Schritte in Vor- und Nachbereitung weniger Aufmerksamkeit bekommen.

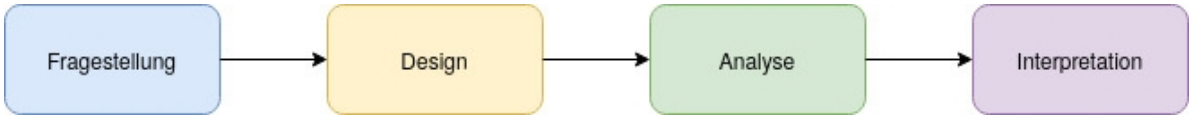

Abbildung 1: Flussdiagramm von Studienplanung, Design, Analyse und Interpretation.

Die Inputdaten für die KI-Techniken sind in vielen Anwendungen sehr hochdimensional, d.h. es werden viele Variablen (auch Features genannt) beobachtet, wobei jede Variable einen Bereich möglicher Werte aufweist. Zudem werden für die Prognosen oft nichtlineare Zusammenhänge mit komplexen Interaktionen berücksichtigt. Dadurch tritt jedoch selbst bei Stichprobenumfängen in den Größenordnungen von Millionen das Problem des Fluches der hohen Dimension auf [4], da die Daten dünn im hochdimensionalen Raum sind („Sparsity“). ML-Verfahren, die von diesen dünnen Daten die Struktur des hochdimensionalen Raums lernen müssen, benötigen daher typischerweise eine enorme Menge an Trainingsdaten. Lokale Methoden haben das Problem, dass lokale Umgebungen mit wachsender Dimension sehr groß werden müssen, um einen festen Anteil der Daten einzufangen.

Tabelle 1: Übersicht über statistische Verfahren und Modelle, die in KI-Systemen Anwendung finden [86 – 89]

Zweck	**ML-Ansatz**	**Statistische Verfahren & Modelle (Beispiele)**	**Anwendungsbeispiele**
Erkennen von Ähnlichkeiten in Daten	Unsupervised Learning	Clusteranalyse, Faktorenanalyse	Personalisierte Medizin [65], Kundenanalyse, Entwicklung psychometrischer Tests [76]
Vorhersage von Ereignissen/ Zuständen	Supervised Learning: Regressions-ML-Systeme	Daten getriebene Modellwahl	Absatzprognose, Wirtschaftsentwicklung [68], Wetter-/Klimavorhersage [66], Vorhersage von Migrationsbewegungen [67]
Erklären von Ereignissen/ Zuständen	Supervised Learning: Regressions-ML-Systeme, interpretierbar	Theorie basierte Modellwahl	Risikofaktoren Epidemiologie [69, 70], Theorienprüfung
Erkennen von bestimmten Objekten	Supervised Learning: Klassifikations-ML-Systeme	Klassifikation	Sprach- und Gesichtserkennung [71, 72], Diagnose und Differentialdiagnose [73 - 75]

Die KI hat Fortschritte in verschiedenen Anwendungsgebieten erreicht. Zu nennen sind hier die automatisierte Gesichtserkennung, die automatische Spracherkennung und -übersetzung [100], die Objektverfolgung in Filmmaterial, das autonome Fahren sowie das Feld der Strategiespiele wie Schach oder Go, wo inzwischen Computerprogramme die besten menschlichen Spieler schlagen [95, 96].

Insbesondere für Aufgaben in der Spracherkennung und -übersetzung sowie der Textanalyse und -übersetzung werden mit großem Erfolg die aus der Statistik stammenden Hidden-Markov-Modelle angewendet und weiterentwickelt [94, 101], da sie in einer gewissen Art und Weise Grammatiken

abbilden können. Automatische Sprachübersetzungssysteme können selbst Sprachen wie Chinesisch in Echtzeit in Sprachen der europäischen Sprachfamilie übersetzen und sind z.B. bei der EU im Einsatz [102]. Ein weiteres potentielles Anwendungsgebiet ist die Medizin, wo KI zum Beispiel zur besseren Früherkennung von Krankheiten, für akkuratere Diagnosen, oder zur Vorhersage von akuten Ereignissen zum Einsatz kommen könnte [23, 24]. Die amtliche Statistik nutzt unter anderem diese Methoden zur Klassifikation sowie zur Erkennung, Zuschätzung und/oder Imputation relevanter Merkmalsausprägungen von statistischen Einheiten. In der Wirtschaft und Ökonometrie werden ferner KI-Methoden angewandt und weiterentwickelt, um aus großen Datenmengen individuellen Konsumverhaltens Rückschlüsse auf gesamtmakroökonomische Entwicklungen zu ziehen [129, 130].

Trotz dieser positiven Entwicklungen, die auch die öffentliche Debatte zu großen Teilen bestimmen, ist jedoch auch Vorsicht geboten. So wird immer wieder von Grenzen der KI berichtet, zum Beispiel im Fall eines tödlich verlaufenen Unfalls mit einem autonom fahrenden Fahrzeug [6]. Aufgrund der möglicherweise folgenschweren Konsequenzen von falsch positiven oder falsch negativen Entscheidungen bei KI-Anwendungen ist eine sorgfältige Prüfung dieser Systeme notwendig [85], insbesondere bei einer flächendeckenden Anwendung wie der Videoüberwachung des öffentlichen Raums. So hat ein Versuch der Bundespolizei am S-Bahnhof Südkreuz in Berlin gezeigt, dass Systeme der automatisierten Gesichtserkennung zur Fahndung nach Gewalttätern aktuell Falschakzeptanzraten von durchschnittlich 0,67% (Testphase 1) bzw. 0,34% (Testphase 2) aufweisen [103]. Das heißt, dass nahezu einer von 150 (bzw. einer von 294) Passanten fälschlicherweise als Gewalttäter klassifiziert wird. In der Medizin können Fehlentscheidungen ebenfalls negative Auswirkungen haben, etwa überflüssige Operationen und Chemotherapien bei falschen Krebsdiagnosen. Entsprechende Prüfverfahren für die Medizin werden aktuell von Regulatoren wie der US FDA entwickelt [7].

Es stellen sich auch ethische Fragen hinsichtlich der Anwendung von KI-Systemen [107]. Neben grundsätzlichen Erwägungen (Welche Entscheidungen sollen Maschinen für uns treffen und welche nicht?) kann auch eine gesellschaftlich gewollte Anwendung mit akzeptierten Fehlentscheidungsraten ethische Fragen aufwerfen, etwa wenn das verwendete Verfahren Bevölkerungsgruppen diskriminiert oder keine hinreichende Kausalität gegeben ist.

Die Statistik kann einen entscheidenden Beitrag zum erfolgreicheren und sichereren Einsatz von KI-Systemen leisten, z.B. in Bezug auf:
1. **Planung und Design**: Bias-Reduktion; Validierung; Repräsentativität
2. **Beurteilung von Datenqualität und Datenerhebung**: Standards für die Qualität von diagnostischen Tests und Audits; Umgang mit fehlenden Werten
3. **Unterscheidung zwischen Kausalität und Assoziation**: Berücksichtigung von Drittvariableneffekten; Beantwortung kausaler Fragestellungen; Simulation von Interventionen
4. **Einschätzung von Sicherheit bzw. Unsicherheit in Ergebnissen**: Erhöhung der Interpretierbarkeit; mathematische Validitätsbeweise oder theoretische Eigenschaften in bestimmten KI-Kontexten; Bereitstellung stochastischer Simulationsdesigns; genaue Analyse der Gütekriterien von Algorithmen im KI-Kontext

Diese Punkte werden in den folgenden Abschnitten näher betrachtet. Darüber hinaus entstehen durch den beschriebenen Wandel auch Herausforderungen und ein erhöhter Bedarf an Lehre und Weiterbildung in KI sowie an Kommunikation von KI-Methoden und Analyseergebnissen.

Planung und Design

Die mögliche Validität der Ergebnisse wird schon beim Design der Studie entscheidend mitgeprägt. Häufig werden in KI-Anwendungen Datensätze analysiert, die ursprünglich für eine andere Fragestellung gesammelt wurden (sogenannte Sekundärdaten). Die Situation ist typischerweise

gegeben, wenn Routinedaten zu wissenschaftlichen Zwecken ausgewertet werden. Zum Beispiel basieren in einer kürzlich veröffentlichten Studie KI-unterstützte Prädiktionsmodelle zur Vorhersage eines bestimmten medizinischen Ereignisses (z.B. eine Krankenhauseinweisung) auf medizinischen Abrechnungsdaten [60].
Ein weiterer Punkt sind Convenience Samples, bei denen die zu untersuchende Stichprobe nicht zufällig gezogen wird, sondern aus „verfügbaren“ Probanden besteht, beispielsweise Onlinefragebögen, welche nur Leute erreichen, die die entsprechende Homepage besuchen und sich die Zeit nehmen, die Fragen zu beantworten.

In der Statistik unterscheidet man zwei Arten von Validität [8]:

1. Interne Validität, d.h. die Gültigkeit des Schlusses, dass eine innerhalb einer Studie eingetretene Veränderung im Outcome tatsächlich auf die untersuchten Ursachen zurückgeführt werden kann. Diese wird z.B. im Kontext der klinischen Forschung durch randomisierte kontrollierte Experimente sichergestellt. Im KI- und ML-Kontext kann sich interne Validität auch auf die Vermeidung von systematisch verzerrten Prognosen (z.B. eine systematische Unterschätzung von Risiken) und Assoziationsabschätzungen beziehen.

2. Externe Validität, d.h. die Übertragbarkeit der beobachteten Effekte und Zusammenhänge auf umfassendere bzw. andere Populationen, Umgebungen, Situationen, ... Diese wird beispielsweise in den Sozialwissenschaften im Kontext der Meinungsforschung unter anderem durch Survey Sampling zu erreichen versucht, d.h. es werden spezielle Sampling Methoden verwendet, um repräsentative Stichproben zu bekommen. Repräsentativität ist dabei im Sinne von [63] zu verstehen.

Der naive Glaube, dass hinreichend viele Daten schon für Repräsentativität sorgen werden, erfüllt sich dabei nicht [9, 10]. Ein prominentes Beispiel hierfür ist Google Flu [11], wobei Google anhand von Suchanfragen versuchte Grippewellen vorherzusagen, die jedoch die tatsächliche Prävalenz stark überschätzten. Ein weiteres Beispiel ist Microsofts Chatbot ‘Tay’ [12, 13]: Tay sollte die Sprachmuster eines 19-jährigen amerikanischen Mädchens nachahmen und aus der Interaktion mit menschlichen Nutzern von Twitter lernen. Nach kurzer Zeit begann der Bot jedoch, anzügliche und beleidigende Tweets zu verfassen, was Microsoft zwang, den Dienst nur 16 Stunden nach seinem Start wieder abzuschalten. Auch die kürzlich veröffentlichte Apple Heart Study [14], die die Fähigkeit der Apple Watch, Vorhofflimmern zu erkennen, an über 400.000 Teilnehmern untersuchte, ist ein Beispiel unzureichender Studienplanung: Das durchschnittliche Alter der Probanden betrug 41 Jahre. Vorhofflimmern tritt fast ausschließlich bei über 65-Jährigen auf. Damit ist die Studie trotz einer hohen Anzahl an Probanden nicht repräsentativ für die Population, für die ein solches Monitoring interessant wäre.

Wenn die Daten nicht repräsentativ für die zu untersuchende Population sind, können Scheinkorrelationen und Bias (in seiner vielfältigen Ausprägung wie z.B. Selektions-, Attrition-, Performance und Detectionbias) die Ergebnisse verfälschen. Ein klassisches Beispiel hierfür ist Simpsons Paradoxon [15], bei dem die Nicht-Beachtung von Subgruppen zu einer Umkehrung der Resultate führen kann, siehe Abbildung 2. Weitere Beispiele sind typische, durch die Art der Datenerhebung verursachte Verzerrungspotentiale wie z.B. Lead-Time- und Length-Bias [64].

In der Statistik sind Methoden und Prinzipien für den Umgang mit Bias bekannt (z.B. Risk of Bias Assessment in systematischen Reviews in der Medizin [16]). Beispiele dafür sind Stratifizierung, marginale Analysen, Berücksichtigung von Interaktionen und Meta-Analysen, aber auch spezifische Techniken bei der Datenerhebung wie z.B. das (Teil-)Randomisieren und die (Teil-)Verblindung sowie Methoden der sogenannten optimalen Designs. Um die Repräsentativität von Ergebnissen zu überprüfen, ist außerdem die Validierung an externen Daten essentiell. In diesem Zusammenhang

stellt die Statistik auch Designs zur Verfügung, die Schlüsse hinsichtlich der internen und externen Validität erlauben [125 - 127].

Zudem spielt die Fallzahl eine entscheidende Rolle für die Aussagekraft und Verlässlichkeit der Ergebnisse [9]. Bei hochdimensionalen Fragestellungen ergibt sich zudem das Problem der „Sparsity“ (siehe Einleitung). Die Statistik kann durch das Bilden von stochastischen Modellen und zugehörigen approximativen Berechnungen oder numerischen Simulationen Aussagen über die Möglichkeiten und Grenzen einer KI-Anwendung (bei gegebener Fallzahl) machen bzw. die notwendigen Fallzahlen in der Planung einer Studie abschätzen. Das ist keine Routinearbeit und Bedarf einer entsprechend fundierten und fortgeschrittenen statistischen Ausbildung, Kompetenz und Erfahrung.

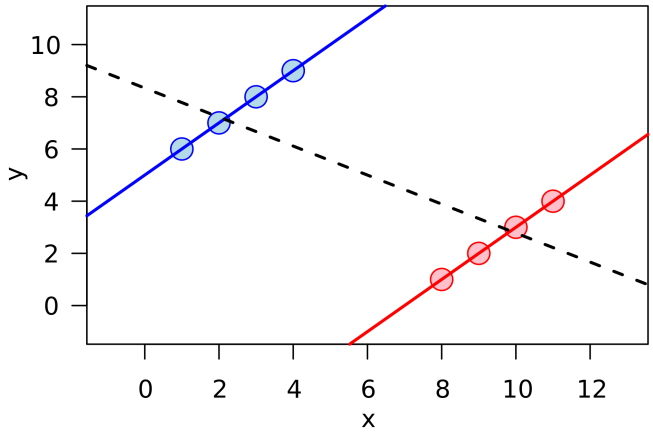

Abbildung 2: Simpsons Paradoxon für stetige Daten: Ein positiver Trend ist sichtbar für beide Gruppen (rot und blau). Werden die Daten gemeinsam betrachtet, erscheint dagegen ein negativer Trend (gestrichelte Linie) [124].

Die Statistik kann deshalb dazu beitragen, die Datenerhebung bzw. Aufbereitung (Fallzahlen, Sampling Design, Gewichtung, Einschränkung des Datensatzes [17], etc.) für die anschließende Auswertung mit KI-Methoden zu optimieren. Grundlegende statistische Techniken sind hier zum Beispiel die Zugrundelegung eines Modells zur Datengenerierung (Data Model) oder Methoden der faktoriellen Versuchsplanung, d.h. einen Faktor gezielt und geschickt zu variieren, um möglichst viel über seinen Einfluss aussagen zu können.

Aus der Statistik sind weiterhin die verschiedenen Phasen in der Entwicklung eines diagnostischen Tests bekannt [18]. In vielen KI-Anwendungen dagegen wird die finale Evaluationsphase an externen Daten nie erreicht, da sich die Algorithmen in der Zwischenzeit wieder verändert haben. Auch die klassischen Gütemaße der Statistik wie Sensitivität, Spezifität und ROC-Kurven können bei der Bewertung von KI-Methoden helfen. Zudem ist die Bewertung der Unsicherheit in den Ergebnissen ein zentraler Gesichtspunkt, auf den wir später zurückkommen werden.

Beurteilung von Datenqualität und Datenerhebung

„Daten sind das neue Öl der Weltwirtschaft.“ Unaufhörlich hallt laut Tagesspiegel dieses Credo durch die Startup-Konferenzen und Gründerforen (Lutz Maicher in Der Tagesspiegel, 21.03.2016). Diese Metapher ist dabei ebenso populär wie falsch. Zunächst einmal handelt es sich bei diesen Daten um Rohöl, das einer weiteren Raffinierung bedarf, ehe es genutzt werden kann. Hinzu kommt, dass die Ressource Rohöl begrenzt ist. „Daten hingegen kann man im Prinzip unendlich oft nutzen, die Vorräte

schwinden nicht durch Gebrauch, sondern vergrößern sich täglich.“ (Philipp Hübel in Philosophie Magazin, April / Mai 2019). Umso wichtiger ist ein verantwortlicher Umgang bei der Aufbereitung.

Abbildung 3: Datenrelevanz und Datenqualität sind gleichwertige Bestandteile eines Fit-for-Purpose Datensatzes. Abbildung nach Duke-Margolis [61].

Die Sicherstellung von Datenqualität ist in allen Analysen von großer Bedeutung, gemäß dem Motto „Garbage in, Garbage out“. Wie im vorigen Kapitel bereits angesprochen kommen im Kontext von KI überwiegend Sekundärdaten zum Einsatz. Bei deren Erhebung und Zusammenstellung liegt in der Regel keine spezifische Fragestellung zugrunde, sondern sie werden zu anderen Zwecken gesammelt, etwa aus Abrechnungs- und Lagerhaltungsgründen. Der Begriff der Wissenserkennung in Datenbanken (Knowledge Discovery in Data Bases [121]) spiegelt sehr deutlich die Annahme wider, dass Daten als gegebene Grundlage betrachtet werden, aus denen durch die Verfahren der KI Information und Wissen zu gewinnen sind. Diese Vorstellung steht konträr zum traditionellen empirischen Forschungsprozess, bei dem aus theoretischen Fragestellungen durch Konzeptionalisierung und Operationalisierung empirisch testbare Forschungsfragen abgeleitet werden, für die gezielt entsprechende Messgrößen erhoben werden. Der Prozess der Operationalisierung wird im Kontext der KI ersetzt durch den ETL-Prozess: „Extract, Transform, Load“ [122]. Aus dem Datensee (Data Lake) sollen die relevanten Messgrößen extrahiert, danach transformiert und schlussendlich in die (automatisierten) Analyseverfahren geladen werden. Dabei wird von den Verfahrend der KI erwartet, dass sie in der Lage sind, aus hochdimensionalen Daten relevante Einflussgrößen zu destillieren.

Der Erfolg dieses Verfahrens hängt fundamental von der Qualität der Daten ab. Datenqualität sei hier in Anlehnung an Karr et al. (2006) definiert als die Eigenschaft von Daten für die Entscheidungsfindung und -bewertung schnell, ökonomisch und effektiv nutzbar zu sein [19]. In diesem Sinne ist Datenqualität ein multi-dimensionales Konzept, das weit über die Messgenauigkeit hinaus geht und Aspekte wie Relevanz, Vollständigkeit, Verfügbarkeit, Aktualität, Metainformation, Dokumentation und vor allem kontext-abhängiges Fachwissen umfasst [61,62]. Im Bereich der amtlichen Statistik sind Relevanz, Genauigkeit und Zuverlässigkeit, Aktualität und Pünktlichkeit, Kohärenz und Vergleichbarkeit, Zugänglichkeit und Klarheit als Dimensionen der Datenqualität definiert [128].

Eine zunehmende Automatisierung der Datenerhebung, etwa durch Sensorik, mag die Messgenauigkeit auf kostengünstige und einfache Art erhöhen. Ob dadurch die erwartete Verbesserung der Datenqualität erreicht wird, muss sich im jeweiligen Anwendungsfall erst zeigen. Eine klassische Problematik der Datenqualität stellen fehlende Werte dar, für die in der Statistik eine Vielzahl von Methoden zum Umgang mit fehlenden Werten, z.B. Imputationsverfahren, oder Methoden der Datenanreicherung entwickelt wurden. Der KI-Ansatz des ubiquitären Datensammelns ermöglicht die Existenz redundanter Daten, die mit entsprechendem Kontextwissen zur Vervollständigung lückenhafter Datenbestände genutzt werden können. Dies erfordert aber eine entsprechende Einbindung des Kontextwissens in den Datenextraktionsprozess.

Die datenhungrigen Entscheidungsverfahren der KI und Statistik unterliegen bezüglich Relevanz und Aktualität einem großen Risiko, basieren sie doch implizit auf der Annahme, dass sich die in den Daten verborgenen Muster auch für die Zukunft perpetuieren sollen. Dies führt in vielen

Anwendungen zu einer unerwünschten Festschreibung bestehender Stereotype und resultierenden Benachteiligungen, etwa bei der automatischen Kreditvergabe oder der automatischen Bewerberauswahl (siehe z.B. Gender Bias in Amazons AI Recruiting Tool [20]).

In der Trias „Experiment – Beobachtungsstudie – Convenience Sample (Datensee)“ bewegt sich das Feld der KI mit Blick auf seine Datengrundlage immer weiter weg vom klassischen Ideal der kontrollierten Experimentaldatenerhebung, die eine Untersuchung kausaler Fragestellungen gewährleistet, hin zur auf reinen Assoziationen beruhenden Exploration gegebener Daten. Auf diese Thematik wird im nachfolgenden Abschnitt noch näher eingegangen.

Die explorative Datenanalyse stellt ein breites Spektrum an Werkzeugen zur Verfügung, die empirischen Verteilungen der Daten zu visualisieren und entsprechende Kennzahlen abzuleiten. Dies kann im Preprocessing dazu genutzt werden, Anomalien festzustellen oder Bereiche typischer Werte festzulegen, um Eingabe- oder Messfehler zu korrigieren und Normwerte zu bestimmen. Im Zusammenspiel mit Standardisierungen in der Datenablage können hier Datenfehler im Messprozess frühzeitig erkannt und korrigiert werden. Auf diese Weise kann die Statistik bei der Beurteilung der Datenqualität im Hinblick auf systematische, standardisierte und vollständige Erfassung helfen. Methodische Untersuchungen bei Umfragen (Survey Methodology) haben primär die Datenqualität im Fokus. Die im Rahmen der statistischen Umfrageforschung gewonnenen Erkenntnisse zur Sicherung der Datenqualität im Hinblick auf interne und externe Validität stellen ein profundes Fundament für entsprechende Entwicklungen im Kontext der KI bereit. Weiterhin sind in der Statistik diverse Verfahren zur Imputation fehlender Daten bekannt, die je nach vorhandenem Kontext und Fachwissen zur Vervollständigung der Daten genutzt werden können. Statistiker haben sich intensiv mit der Behandlung von fehlenden Werten unter verschiedenen Entstehungsprozessen (Non-response, Missing Not At Random, Missing At Random, Missing Completely At Random [108]), der Auswahlverzerrung (Selection Bias) und Messfehlern (Measurement Error) auseinandergesetzt.

Ein weiterer Punkt ist das Parametertuning, d.h. die Bestimmung sogenannter Hyperparameter, die das Lernverhalten der ML-Algorithmen kontrollieren: umfassendes Parametertuning von Methoden im KI-Kontext benötigt häufig sehr große Datenmengen. Für kleineres Datenaufkommen ist dies fast nicht möglich. Bestimmte modellbasierte (statistische) Methoden können aber dennoch verwendet werden.

Unterscheidung zwischen Kausalität und Assoziationen

Noch vor wenigen Jahrzehnten war die größte Herausforderung der KI, Maschinen dazu zu bringen, eine mögliche Ursache mit einer Reihe von beobachtbaren Bedingungen in Verbindung zu bringen. Die rasante Entwicklung der letzten Jahre (sowohl was Theorie und Methodik statistischer Lernverfahren betrifft als auch in Bezug auf die Rechenleistung von Computern) hat zu einer Vielzahl an Algorithmen und Methoden geführt, die das inzwischen meistern. Ein Beispiel hierfür sind Deep-Learning-Methoden, die in der Robotik [21] und beim autonomen Fahren [22] genauso Anwendung finden wie in computergestützten Erfassungs- und Diagnosesystemen (beispielsweise zur Brustkrebsdiagnose [23]), in der Wirkstoffentdeckung in der pharmazeutischen Forschung [24] oder in der Landwirtschaft [25].

Mit hoher Prädiktionsgüte sind KI-Methoden in der Lage, Strukturen und Zusammenhänge in großen Datenmengen aufzudecken – basierend auf Assoziationen. Aufgrund der Leistungsfähigkeit von KI-Methoden in großen Datensätzen werden diese in der Medizin auch häufig zur Analyse von Register- und Beobachtungsdaten verwendet, die nicht im strengen Rahmen einer klassischen randomisierten Studie erhoben wurden (siehe auch Kapitel Design). Die Aufdeckung von Korrelationen und Assoziationen ist (speziell in diesem Rahmen) nicht mit Kausalität gleichzusetzen.

Judea Pearl, von vielen als „Vater der KI“ bezeichnet, sagte kürzlich in einem Interview [26]: *„As much as I look into what’s being done with deep learning, I see they’re all stuck there on the level of associations.“*

Ein wichtiger Schritt in der weiteren Entwicklung von KI ist daher, Argumentation basierend auf Assoziationen mit kausaler Argumentation zu ersetzen. Den Unterschied beschreibt Pearl [27] wie folgt: *„An associational concept is any relationship that can be defined in terms of a joint distribution of observed variables, and a causal concept is any relationship that cannot be defined from the distribution alone.“*

Schon die formale Definition eines kausalen Effektes ist nicht trivial. In der Statistik und der klinischen Epidemiologie werden dazu beispielsweise die Bradford-Hill-Kriterien [28] sowie das von Rubin [29] eingeführte Counterfactual Framework zu Hilfe genommen.

Das zentrale Problem in Beobachtungsdaten sind Drittvariableneffekte, die im Gegensatz zum Randomized Controlled Trial nicht designbedingt ausgeschlossen werden und deren (Nicht-) Berücksichtigung zur verzerrten Schätzung kausaler Effekte führt. Hierbei gilt es zwischen *Confoundern*, *Collidern* und *Mediatoren* zu unterscheiden [30]. Confounder sind unbeobachtete oder unberücksichtigte Variablen, die sowohl Exposition (Exposure) als auch die Zielgröße (Outcome) beeinflussen (siehe Abbildung 4(a)). Dadurch können Effekte der Exposition verfälscht werden, wenn naiv korreliert wird. Bereits 1935 wies Fisher in seinem Buch „The Design of Experiments“ auf dieses Problem hin. Die formale Definition wurde in der Epidemiologie bereits in den 1980er-Jahren von Greenland und Robins [31] entwickelt. Später wurden auch grafische Kriterien wie das Back-Door-Criterion [32, 33] zur Definition von Confounding entwickelt.

Das Confounding-Problem wird in der Statistik entweder im Design (z. B. randomisierte Studie, Stratifizierung, ...) oder in der Auswertung (Propensity Score Methoden [34], Marginal Structural Models [35], Graphical Models [36]) berücksichtigt. Interessant ist in diesem Zusammenhang die Beobachtung, dass randomisierte Studien (die in der Medizin eine lange Tradition haben) in letzter Zeit auch verstärkt in ökonometrischen Studien eingesetzt werden [37, 38, 119]. Bei Beobachtungsdaten hat die Ökonometrie viele methodische Beiträge geleistet, um Behandlungseffekte zu identifizieren, z.B. durch den Potential Outcome Approach assoziiert mit den Arbeiten von Donald Rubin, Paul Rosenbaum und Koautoren [29, 131 - 134] und Arbeiten zur Politikevaluation von Heckman und Koautoren [135].

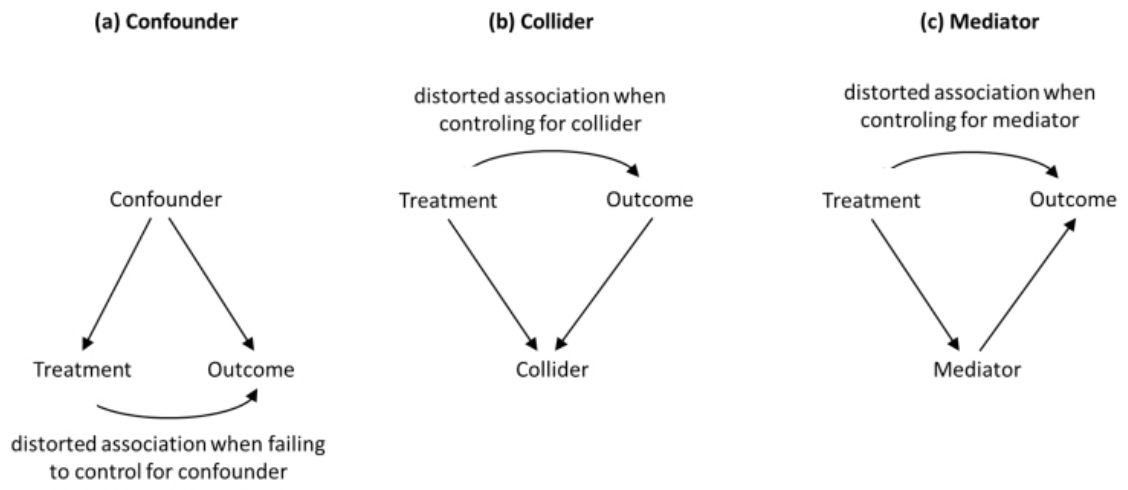

Abbildung 4: Drittvariableneffekte in Beobachtungsdaten, nach [123].

Im Gegensatz zu Confoundern führen Collider und Mediatoren genau dann zu verzerrten Schätzern kausaler Effekte, wenn sie bei deren Bestimmung berücksichtigt werden. Während Collider gemeinsame Folgen von Treatment und Outcome darstellen (Abbildung 4(b)), handelt es sich bei Mediatoren um Variable, die einen Teil des kausalen Mechanismus repräsentieren, über den das Treatment auf das Outcome wirkt (Abbildung 4(c)). Insbesondere bei Längsschnittdaten gilt es folglich theoretisch informiert zu differenzieren, welches Verhältnis in den Beobachtungsdaten vorhandene Drittvariable zu Treatment und Outcome haben, um Bias in den Schätzern kausaler Effekte durch deren (Nicht-) Berücksichtigung zu vermeiden.

Nur durch Integration entsprechender statistischer Theorien und Methoden in die KI wird es möglich, kausale Fragen zu beantworten und Interventionen zu simulieren. In der Medizin beispielsweise können dann Fragen wie „Was wäre der Effekt eines allgemeinen Rauchverbots auf das deutsche Gesundheitssystem?“ untersucht und belastbare Aussagen gemacht werden – auch ohne randomisierte Studien, die hier natürlich nicht möglich wären. Pearls Vorstellung geht dabei über den Einsatz von ML-Methoden in kausalen Analysen hinaus (diese werden z. B. im Zusammenhang mit Targeted Learning [39] oder Causal Random Forest [40] eingesetzt). Seine Vision ist es vielmehr, das von ihm beschriebene kausale Framework [41] in die ML-Algorithmen zu integrieren, um die Maschinen in die Lage zu versetzen, kausale Schlüsse zu ziehen und Interventionen zu simulieren.

Die Integration kausaler Methoden in die KI trägt auch dazu bei, die Transparenz und damit die Akzeptanz von KI-Methoden zu stärken, da der Verweis auf Wahrscheinlichkeiten oder statistische Zusammenhänge in der Erklärung nicht so effektiv ist wie der Verweis auf Ursachen und kausale Wirkungen [42].

Das Ziel von Statistik im Zusammenhang mit KI muss es sein, die Interpretation von Daten zu vereinfachen. Daten allein sind kaum eine Wissenschaft, egal wie groß sie werden und wie geschickt sie manipuliert werden. Wichtig ist der Erkenntnisgewinn, der zukünftige Interventionen ermöglicht [43].

Einschätzung von Unsicherheit, Interpretierbarkeit und Validierung

In der KI wird die Bewertung der Unsicherheit häufig vernachlässigt. Zum einen beruht dies auf der oben diskutierten Fehleinschätzung, dass viele Daten automatisch zu exakten Ergebnissen führen und damit eine Bewertung der Unsicherheit überflüssig machen. Zum anderen ist die Bewertung der Unsicherheit mit den benutzen Methoden häufig auch nicht ganz einfach. Wie wichtig diese Bewertung dennoch ist, verdeutlicht das folgende Zitat des amerikanischen Statistikers David B. Dunson: „*It is crucial to not over-state the results and appropriately characterize the (often immense) uncertainty to avoid flooding the scientific literature with false findings.*“ [44]

Um das Hauptziel einer möglichst genauen Vorhersagegüte zu erreichen, werden in KI-Anwendungen bewusst Annahmen an zugrundeliegende Verteilungen und funktionale Zusammenhänge fallengelassen. Dies ermöglicht eine höhere Flexibilität der Verfahren, erschwert jedoch die genaue Quantifizierung der Unsicherheit in den Schätzungen, d.h. die Angabe valider Prädiktions- und Unsicherheitsbereiche für die Zielgrößen und relevanten Parameter. Wie Bühlmann und Kollegen es formulieren: „*The statistical theory serves as guard against cheating with data: you cannot beat the uncertainty principle.*“ [45]

In den letzten Jahren wurden bereits Vorschläge zur Unsicherheitsquantifizierung bei KI-Methoden entwickelt (z.B. durch Verknüpfung mit Bayesianischen Approximationen, Bootstrapping, Jackknifing und anderen Kreuzvalidierungstechniken, Gaußprozessen, Monte-Carlo-Dropout, vgl. z.B. [110 – 113, 120]). Deren theoretische Validität (d. h., dass es sich z.B. tatsächlich um 95% Prädiktionsbereiche

handelt) ist aber bisher in vielen Situationen noch ungeklärt bzw. nur unter sehr restriktiven oder z.T. unrealistischen Annahmen nachgewiesen.

Im Gegensatz dazu besteht bei vielen Methoden die Möglichkeit, sie auf konkrete stochastische Modelle zu stützen, die etwas weniger flexibel sind. Dafür erlauben sie aber eine einfache Berücksichtigung der zugrundeliegenden Unsicherheit durch Angabe von validen Prädiktions- oder Vertrauensintervallen und somit eine bessere Interpretierbarkeit der Ergebnisse.

Im Vergleich dazu sind die geschätzten Parameter bei vielen KI-Ansätzen (wie z. B. Deep Learning) nur schwer zu interpretieren. Vorreiterarbeiten aus der Informatik zu diesem Thema sind beispielsweise [138, 139], für die Leslie Valiant 2010 mit dem Turing-Preis ausgezeichnet wurde. Zur Verbesserung der Interpretierbarkeit inklusive Quantifizierung der Unsicherheit der durch eine KI-Methode identifizierten Muster ist also noch weitere Forschung notwendig, welche sich auch stark auf statistische Ansätze stützten wird. So wurde beispielsweise vorgeschlagen, Hilfsmodelle zu verwenden, d.h. relativ einfache statistische Modelle, welche nach Anpassung eines Deep-Learning-Ansatzes die wichtigsten dadurch repräsentierten Muster beschreiben und potentiell auch zur Quantifizierung von Unsicherheit dienen können [90 - 93]. Statistische Verfahren und KI-Lernansätze können sich gegenseitig ergänzen, wie sich an den relativ jungen und aktiven Bereichen der Computational und Statistical Learning Theory [57, 58, 105] zeigt. Ein wichtiger Aspekt ist dabei die Modellkomplexität, die z.B. durch Entropien (wie VC-Dimensionen) oder mittels Kompressionsschranken erfasst wird [59]. Diese Konzepte wie auch verschiedene Formen der Regularisierung [114 - 116], d. h. der Beschränkung auf einen Teilbereich des Parameterraumes, ermöglichen es, eine Überanpassung der Lernverfahren (Overfitting) an eine gegebene Datengrundlage zu erkennen oder sogar zu korrigieren. Die Anwendung komplexitätsreduzierender Konzepte kann in diesem Zusammenhang als direkte Umsetzung des Sparsamkeitsprinzips (Lex Parsimoniae) angesehen werden und erhöht häufig die Interpretierbarkeit der resultierenden Modelle [117, 118]. Regularisierung und komplexitätsreduzierende Konzepte sind integraler Bestandteil vieler KI-Methoden. Sie gehören aber auch zu den grundlegenden Prinzipien der modernen Statistik, die dort bereits vor der Einführung von KI-Methoden vorgeschlagen wurden, z.B. im Zusammenhang mit Shrinkage- und empirischen Bayesverfahren. Es gibt mittlerweile auch zahlreiche gemeinsame Konzepte aus der KI und Statistik. Damit finden sich in diesem Bereich viele Möglichkeiten zu einem Austausch der Methoden.

Die oben angesprochenen Aspekte der Validierung der Methoden sind dabei über Interpretierbarkeit und Quantifizierung von Unsicherheit hinaus enorm wichtig. Im KI-Kontext erfolgt diese Validierung häufig nur an einzelnen, oftmals mehrfach verwendeten „etablierten“ Datensätzen. Dabei kann die Stabilität bzw. Variabilität der Ergebnisse aufgrund der fehlenden Generalisierbarkeit nicht zuverlässig beurteilt werden. Auch hier können statistische Konzepte effizient eingesetzt werden: Um der Vielfalt realer Möglichkeiten gerecht zu werden, bedient sich die Statistik nämlich wiederum stochastischer Modelle. Neben mathematischen Validitätsbeweisen und theoretischen Untersuchungen werden dabei auch ausführliche Simulationsstudien durchgeführt, mit denen die Grenzen der Methoden (durch Überschreitung der gemachten Annahmen) evaluiert werden. Hierbei kann der von der Statistik eingenommene stochastische Blickwinkel äußerst gewinnbringende Erkenntnisse liefern. Der Aspekt der Validierung gilt auch für das Gütekriterium (z.B. Accuracy, Sensitivität und Spezifität) eines KI-Algorithmus, dessen Schätzer ebenfalls zufallsbehaftet sind und deren Unsicherheit in der Regel gar nicht quantifiziert wird.

Eine besondere Herausforderung stellen die immer schneller werdenden Entwicklungszyklen der KI-Systeme dar, mit denen die Möglichkeiten der Validierung nicht immer mithalten können. Erschwerend kommt hinzu, dass die Entwicklungsprozesse von mobilen Apps oder Online-Learning-Systemen (wie beispielsweise den Recommender-Systemen von Amazon) de facto nie enden und dadurch eine fortlaufende Validierung benötigen.

Die Statistik kann hier beitragen, die Validität und Interpretierbarkeit von KI-Methoden zu erhöhen, indem sie Beiträge zur Quantifizierung der Unsicherheit liefert. Dies kann für manche Verfahren durch mathematische Untersuchungen sowie Zusammenhänge für Methoden unter der Annahme spezifischer stochastischer Modelle [46–50, 105] erfolgen (wie z.B. asymptotische Konsistenzbeweise oder (finite) Angaben von Fehlerschranken oder Robustheitsuntersuchungen). Andererseits beinhaltet es auch die Bereitstellung stochastischer Simulationsdesigns [51] und die Angabe leicht interpretierbarer statistischer Hilfsmodelle. Schließlich ermöglicht sie die genaue Analyse der Gütekriterien von Algorithmen im KI-Kontext.

Bildung, Weiterbildung und Öffentlichkeitsarbeit

Die KI ist seit Jahren ein Wachstumsbereich, dessen Entwicklung noch lange nicht abgeschlossen ist. Neben vielen ethischen und rechtlichen Problemstellungen hat sich gezeigt, dass bei der Erhebung und Verarbeitung der Daten noch viele offene Fragen beantworten werden müssen. Deshalb bieten sich statistische Methoden bei Fragestellung, Design, Analyse und Interpretation noch stärker als bisher als integraler Bestandteil der KI- Systeme an. Gerade bei der Weiterentwicklung der Methoden kann die Statistik z.B. den wissenschaftlichen Austausch mit der Gründung von Netzwerken zwischen Anwendern und Experten stärken. Mit ihrem Spezialwissen ist sie ein natürlicher Partner für andere Disziplinen in Lehre, Forschung und Praxis.

Bildung

Die DAGStat hält eine Verankerung der KI und der zugrundeliegenden statistischen Methoden in den verschiedenen Bildungswegen für dringend geboten. Dies beginnt bereits mit der schulischen Ausbildung, wo Statistik ebenso wie Informatik fest in den Lehrplänen verankert sein sollte. Grundlagen der KI können grundsätzlich altersgerecht schon in den verschiedenen Schulformen (Grundschule, Sekundarstufe 1, Sekundarstufe 2 und natürlich auch Berufsbildende Schulen) vermittelt werden. Hierzu müssen schulbezogene Projekte und Lehrerfortbildungsinitiativen unter wissenschaftlicher Begleitung durch die Didaktiken der Informatik und der Statistik initiiert und gefördert werden, die die komplexen Inhalte für Schüler/innen angemessen elementarisieren und interessant machen und dabei sowohl fachliche wie gesellschaftliche Aspekte berücksichtigen. Ein Pilotprojekt dazu ist das „Projekt Data Science und Big Data in der Schule“ (www.prodabi.de) [137]. International hat Projekt IDS (www.introdatascience.org) eine Vorreiterrolle bei schulbezogenen Projekten sowie das „International Data Science in Schools Project“ (www.idssp.org). Bei der Entwicklung der schulischen Curricula sollten Statistiker ihr Expertenwissen einbringen können. Bei digitalisierten Lehr- und Ausbildungsangeboten (z.B. E-Learning) sollten in Statistik qualifiziert ausgebildete Lehrende beteiligt werden.

Für den Hochschulbereich gilt es, sowohl die methodischen Fächer im Bereich Data Science wie Mathematik, Statistik und Informatik als auch Querschnittsbereiche wie Medizin, Ingenieurs-, Sozial- und Wirtschaftswissenschaften etc. mit wissenschaftlichen Grundlagen für die Forschung und für die Anwendung von KI auszustatten. Dazu zählen angepasste Stellenpläne an den Universitäten ebenso wie Förderprogramme, ggf. neue Studiengänge, Doktorandenprogramme, Forschungsverbünde und Forschungsprogramme. Hier erscheint die Qualifikation der Lehrenden in der statistischen Methodik aufgrund der stärker werdenden Nachfrage besonders ausbaufähig.

Weiterbildung

Weiterbildung für Berufstätige sollte auf verschiedensten Ebenen qualifiziert ausgebaut werden. Hier sind Weiterbildungsprogramme zu KI in unterschiedlichen Formen und Formaten denkbar: Workshops/Sommerschulen, Webinars, Programme zur beruflichen Weiterbildung, Mentoring, Laborbesuche, etc. Insbesondere sollte es auch hier sowohl Angebote für Methodiker wie

Informatiker, Statistiker und Mathematiker geben, die noch nicht im Bereich KI arbeiten als auch für Anwender wie z. B. Kliniker, Ingenieure, Sozialwissenschaftler und Ökonomen.

Die DAGStat setzt es sich daher zum Ziel, zusammen mit anderen Fachgesellschaften Workshops und Weiterbildungsangebote zu entwickeln. Hierbei sollen bestehende Strukturen genutzt und ein Schulterschluss mit anderen Berufsfeldern erreicht werden. Des Weiteren kann die DAGStat Verbindungen zwischen Mitgliedsgesellschaften und Anwendern vermitteln.

Interdisziplinärer Wissensaustausch

Durch Gründung von und Beteiligung an Netzwerken können Methodiker mit Anwendern/Fachexperten zusammengebracht werden, um einen kontinuierlichen Austausch zwischen den Disziplinen zu begründen bzw. aufrecht zu erhalten. Die DAGStat kann hier als erster Ansprechpartner zur Verfügung stehen. Neben KI-Methoden sollten in diesen Veranstaltungen insbesondere auch die Themen Datenkuration, Qualitätsmanagement und Datenintegration abgedeckt werden. Eine weitere Möglichkeit zum Wissenstransfer besteht in der Erstellung von an Anwender gerichteten Publikationen und Richtlinien, wie es sich beispielsweise die STRATOS Initiative für medizinische Beobachtungsstudien zum Ziel gesetzt hat.

Öffentlichkeitsarbeit

Die Vertretung der Statistik in der Öffentlichkeit gehört zu den originären Aufgaben der DAGStat [109]. Dazu zählen die DAGStat-Homepage, das halbjährlich erscheinende Bulletin sowie das jährlich von der DAGStat organisierte Symposium.

Ein weiterer Schritt für eine offensivere Öffentlichkeitsarbeit ist das DAGStat Symposium 2020. Erstmals ist es gelungen, das Deutschen Zentrum für Herzkreislauf Forschung (DZHK) als Kooperationspartner zu gewinnen. Die Veranstaltung findet unter dem Titel „Künstliche Intelligenz in der Medizin: Aufbruch in eine neue Ära oder leeres Versprechen?“ am 26. März 2020 in der Urania in Berlin statt. Teilnehmer sind namhafte Wissenschaftler, Vertreter der Wirtschaft, des Datenschutzbeauftragten und des Bundesgesundheitsministeriums sowie ein Mitglied der Ethikkommission.

Die DAGStat-Tagung soll wie bisher im Rhythmus von drei Jahren veranstaltet werden. Neben dem Austausch der Statistiker wird auch hier eine verstärkte Öffentlichkeitsarbeit angedacht. Vorbild ist u.a. die bereits im Jahr 2019 durchgeführte Session Statistik für die Öffentlichkeit, in der Methoden und Anwendungen vorgestellt wurden.

Die DAGStat bietet außerdem an, sich aktiv an Initiativen im Kontext von KI zu beteiligen. Aufgrund ihres Spezialwissens hat sie ein Alleinstellungsmerkmal, das sie in die Politikgestaltung einbringen kann und will.

Zusammenfassung und Diskussion

Die Statistik ist eine breit angelegte wissenschaftsübergreifende Fachrichtung. Mit ihrem Spezialwissen über Datenauswertungen angefangen von der Fragestellung über Design und Analyse bis hin zur Interpretation hat sie eine wichtige und besondere Rolle. Als Kernelement der KI ist sie der natürliche Partner für andere Disziplinen in Lehre, Forschung und Praxis. Insbesondere lassen sich folgende Beiträge der Statistik für die künstliche Intelligenz zusammenfassen:

1. Methodische Entwicklung: Die Entwicklung von KI-Systemen und ihre theoretische Unterfütterung hat sehr stark von Forschungen in den Computerwissenschaften und der Statistik profitiert und so manches Verfahren wurde von Statistikern entwickelt. Neuere Entwicklungen wie Extreme Learning Machines zeigen, dass die Statistik auch für die

Konzeption von KI-Systemen wichtige Beiträge liefert, z.B. durch verbesserte Lernalgorithmen basierend auf penalisierten oder robustifizierten Schätzverfahren.

2. Planung und Design: Die Statistik kann dazu beitragen, die Datenerhebung bzw. Aufbereitung (Fallzahlen, Sampling Design, Gewichtung, Einschränkung des Datensatzes, Design of Experiments, etc.) für die anschließende Auswertung mit KI-Methoden zu optimieren. Außerdem können die Gütemaße der Statistik und ihre zugehörigen Inferenzmethoden bei der Bewertung von KI-Modellen helfen.
3. Beurteilung von Datenqualität und Datenerhebung: Die explorative Datenanalyse stellt ein breites Spektrum an Werkzeugen zur Verfügung, die empirischen Verteilungen der Daten zu visualisieren und entsprechende Kennzahlen abzuleiten, die sowohl dazu genutzt werden können Anomalien festzustellen oder Bereiche typischer Werte festzulegen, um Eingabefehler zu korrigieren, Normwerte zu bestimmen und fehlende Werte zu imputieren. Im Zusammenspiel mit Standardisierungen in der Datenablage können hier Datenfehler im Messprozess frühzeitig erkannt und korrigiert werden. Mit Hilfe modellbasierter statistischer Methoden ist außerdem ein umfassendes Parametertuning auch für kleines Datenaufkommen möglich.
4. Unterscheidung von Kausalität und Assoziationen: In der Statistik sind Methoden zum Umgang mit Drittvariableneffekten bekannt. Hier gilt es, theoretisch informiert zu differenzieren, welches Verhältnis in den Beobachtungsdaten vorhandene Drittvariable zu Treatment und Outcome haben, um Bias in den Schätzern kausaler Effekte zu vermeiden. Pearls kausales Framework ermöglicht dabei die Analyse kausaler Effekte sowie die Simulation von Interventionen. Die Integration kausaler Methoden in die KI trägt auch dazu bei, die Transparenz und Akzeptanz von KI-Methoden zu stärken.
5. Einschätzung von Sicherheit bzw. Unsicherheit in Ergebnissen: Die Statistik kann dazu beitragen, die Quantifizierung von Unsicherheit in und die Interpretierbarkeit von KI-Methoden zu ermöglichen oder zu verbessern. Durch Annahme spezifischer stochastischer Modelle können außerdem mathematische Validitätsbeweise geliefert werden. Zudem werden durch die Bereitstellung stochastischer Simulationsdesigns Grenzen der Methoden ausgelotet.
6. Eine gewissenhafte Umsetzung der Punkte 2 bis 5 inklusive vorher festgelegtem Auswertungsplan wirkt zudem der Replikationskrise [5] in vielen Wissenschaftsbereichen entgegen. In dieser seit Beginn der 2010er Jahre andauernden methodischen Krise hat sich herausgestellt, dass viele Studien insbesondere in der Medizin und den Sozialwissenschaften nur schwer oder gar nicht reproduzierbar sind.
7. Bildung, Weiterbildung und Öffentlichkeitsarbeit: Mit ihrem Spezialwissen ist die Statistik der natürliche Partner für andere Disziplinen in Lehre und Weiterbildung. Gerade bei der Weiterentwicklung der Methoden der Künstlichen Intelligenz kann die Statistik den wissenschaftlichen Austausch stärken.

Dabei unterstützt die DAGStat ohne Einschränkung die von der Datenethikkommission der Bundesregierung im Oktober 2019 veröffentlichten Grundsätze. Künstliche Intelligenz muss sich in all ihren Anwendungen ethisch, rechtlich, kulturell und institutionell in die Gesellschaft einbetten. Dies dient dazu, eine verantwortungsvolle und gemeinwohlorientierte Entwicklung und Nutzung zu erreichen. Die DAGStat setzt sich auch dafür ein, dass ein für alle Disziplinen und Anwender verbindlicher Ordnungsrahmen national wie international festgelegt wird und bietet an, diesen fachspezifisch unter Einbringung von Expertenwissen mitzugestalten.

Referenzen

[1] https://www.bmbf.de/files/Nationale_KI-Strategie.pdf, accessed 31.01.20

[2] Simon, H. A. (1983). Why should machines learn?. In Michalski R. S., Carbonell J. G., Mitchell T. M. (Eds.) *Machine learning* (pp. 25-37). Morgan Kaufmann.

[3] Samuel, A. L. (1959). Some studies in machine learning using the game of checkers. *IBM Journal of research and development*, 3(3), 210–229.

[4] Bellman, R. (1957). *Dynamic programming*. Princeton University Press.

[5] Pashler, H., & Wagenmakers, E. J. (2012). Editors' introduction to the special section on replicability in psychological science: A crisis of confidence?. *Perspectives on Psychological Science*, *7*(6), 528-530.

[6] https://www.wired.com/story/ubers-self-driving-car-didnt-know-pedestrians-could-jaywalk/, accessed 30.01.20

[7] https://www.fda.gov/media/122535/download, accessed 31.01.20

[8] Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference/William R. Shedish, Thomas D. Cook, Donald T. Campbell*. Boston: Houghton Mifflin,.

[9] Meng, X. L. (2018). Statistical paradises and paradoxes in big data (I): Law of large populations, big data paradox, and the 2016 US presidential election. *The Annals of Applied Statistics*, *12*(2), 685-726.

[10] Meng, X. L., & Xie, X. (2014). I got more data, my model is more refined, but my estimator is getting worse! Am I just dumb?. *Econometric Reviews*, *33*(1-4), 218-250.

[11] Lazer, D., Kennedy, R., King, G., & Vespignani, A. (2014). The parable of Google Flu: traps in big data analysis. *Science*, *343*(6176), 1203-1205.

[12] Davis, E. (2016). AI amusements: the tragic tale of Tay the chatbot. *AI Matters*, *2*(4), 20-24.

[13] Wolf, M. J., Miller, K., & Grodzinsky, F. S. (2017). Why we should have seen that coming: comments on Microsoft's tay experiment, and wider implications. *ACM SIGCAS Computers and Society*, *47*(3), 54-64.

[14] Perez, M. V., Mahaffey, K. W., Hedlin, H., Rumsfeld, J. S., Garcia, A., Ferris, T., ... & Hung, G. (2019). Large-scale assessment of a smartwatch to identify atrial fibrillation. *New England Journal of Medicine*, *381*(20), 1909-1917.

[15] Simpson, E.H., 1951. The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, *13*(2), pp.238-241.

[16] Higgins, J. P., Altman, D. G., Gøtzsche, P. C., Jüni, P., Moher, D., Oxman, A. D., ... & Sterne, J. A. (2011). The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. *BMJ*, *343*, d5928.

[17] Rubin, D. B. (2008). For objective causal inference, design trumps analysis. *The Annals of Applied Statistics*, *2*(3), 808-840.

[18] Pepe, M. S. (2003). *The statistical evaluation of medical tests for classification and prediction*. Medicine.

[19] Karr, A. F., Sanil, A. P., & Banks, D. L. (2006). Data quality: A statistical perspective. *Statistical Methodology*, *3*(2), 137-173.

[20] Dastin, J. (2018). Amazon scraps secret AI recruiting tool that showed bias against women. *San Fransico, CA: Reuters. Retrieved on October*, *9*, 2018. https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G, accessed 27.11.2019.

[21] Levine, S., Pastor, P., Krizhevsky, A., Ibarz, J., & Quillen, D. (2018). Learning hand-eye coordination for robotic grasping with deep learning and large-scale data collection. *The International Journal of Robotics Research*, *37*(4-5), 421-436.

[22] Teichmann, M., Weber, M., Zoellner, M., Cipolla, R., & Urtasun, R. (2018,). Multinet: Real-time joint semantic reasoning for autonomous driving. In *2018 IEEE Intelligent Vehicles Symposium (IV)* (pp. 1013-1020). IEEE.

[23] Burt, J. R., Torosdagli, N., Khosravan, N., RaviPrakash, H., Mortazi, A., Tissavirasingham, F., ... & Bagci, U. (2018). Deep learning beyond cats and dogs: recent advances in diagnosing breast cancer with deep neural networks. *The British journal of radiology*, *91*(1089), 20170545.

[24] Chen, H., Engkvist, O., Wang, Y., Olivecrona, M., & Blaschke, T. (2018). The rise of deep learning in drug discovery. *Drug discovery today*, *23*(6), 1241-1250.

[25] Kamilaris, A., & Prenafeta-Boldú, F. X. (2018). Deep learning in agriculture: A survey. *Computers and electronics in agriculture*, *147*, 70-90.

[26] Quanta Magazine, May 2018, https://www.quantamagazine.org/to-build-truly-intelligent-machines-teach-them-cause-and-effect-20180515/#, accessed 31.01.2020

[27] Pearl, J. (2010). The foundations of causal inference. *Sociological Methodology*, *40*(1), 75-149.

[28] Hill, A. B. (1965). The environment and disease: association or causation?

[29] Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, *66*(5), 688.

[30] Pearl, J. (2009). *Causality*. Cambridge University Press.

[31] Greenland, S., & Robins, J. M. (1986). Identifiability, exchangeability, and epidemiological confounding. *International journal of epidemiology*, *15*(3), 413-419.

[32] Greenland, S., Robins, J. M., & Pearl, J. (1999). Confounding and collapsibility in causal inference. *Statistical science*, 29-46.

[33] Pearl, J. Aspects of Graphical Models Connected With Causality. In *Proceedings of the 49th Session of the International Statistical Science Institute*, 1993.

[34] Cochran, W. G., & Rubin, D. B. (1973). Controlling bias in observational studies: A review. *Sankhyā: The Indian Journal of Statistics, Series A*, 417-446.

[35] Robins, J. M., Hernan, M. A., & Brumback, B. (2000). Marginal structural models and causal inference in epidemiology.

[36] Didelez, V. (2007). Graphical models for composable finite Markov processes. *Scandinavian Journal of Statistics*, *34*(1), 169-185.

[37] Duflo, E., Glennerster, R., & Kremer, M. (2007). Using randomization in development economics research: A toolkit. *Handbook of development economics*, *4*, 3895-3962.

[38] Athey, S., & Imbens, G. W. (2017). The econometrics of randomized experiments. In *Handbook of economic field experiments* (Vol. 1, pp. 73-140). North-Holland.

[39] Van der Laan, M. J., & Rose, S. (2011). *Targeted learning: causal inference for observational and experimental data*. Springer Science & Business Media.

[40] Athey, S., & Imbens, G. W. (2015). Machine learning methods for estimating heterogeneous causal effects. *stat*, *1050*(5), 1-26.

[41] Pearl, J. (2010). The foundations of causal inference. *Sociological Methodology*, *40*(1), 75-149.

[42] Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, *267*, 1-38.

[43] Pearl, J. (2018). Theoretical impediments to machine learning with seven sparks from the causal revolution. *arXiv preprint arXiv:1801.04016*.

[44] Dunson, D. B. (2018). Statistics in the big data era: Failures of the machine. *Statistics & Probability Letters*, *136*, 4-9.

[45] Bühlmann, P., & van de Geer, S. (2018). Statistics for big data: A perspective. *Statistics & Probability Letters*, *136*, 37-41.

[46] Bartlett, P. L., Bickel, P. J., Bühlmann, P. et al. (2004). Discussions of boosting papers, and rejoinders. *Annals of Statistics*, *32*(1), 85-134.

[47] Scornet, E., Biau, G., & Vert, J. P. (2015). Consistency of random forests. *The Annals of Statistics*, *43*(4), 1716-1741.

[48] Wager, S., & Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, *113*(523), 1228-1242.

[49] Athey, S., Tibshirani, J., & Wager, S. (2019). Generalized random forests. *Annals of Statistics*, *47*(2), 1148-1178.

[50] Devroye, L., Györfi, L., & Lugosi, G. (2013). *A probabilistic theory of pattern recognition* (Vol. 31). Springer Science & Business Media.

[51] Morris, T. P., White, I. R., & Crowther, M. J. (2019). Using simulation studies to evaluate statistical methods. *Statistics in Medicine*, 38(11), 2074-2102.

[52] https://www.faz.net/aktuell/wirtschaft/daten-sind-das-neue-oel-15739406.html, accessed 04.02.2020

[53] Solomonoff, R. J. (1985). The time scale of artificial intelligence: Reflections on social effects. *Human Systems Management*, *5*(2), 149-153.

[54] Moor, J. (2006). The Dartmouth College artificial intelligence conference: The next fifty years. *AI Magazine*, *27*(4), 87-87.

[55] Russell, S., & Norvig, P. (2005). AI a modern approach. *Learning*, 2(3), 4.

[56] Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction*. MIT press.

[57] Vapnik V. (1998) *Statistical learning theory*. Wiley, New York

[58] Kearns M. J., Vazirani U.V. (1994) *An introduction to computational learning theory*. MIT, Cambridge

[59] Langford, J. (2005). Tutorial on practical prediction theory for classification. *Journal of machine learning research*, *6*(Mar), 273-306.

[60] Lin, E. J. D., Hefner, J. L., Zeng, X., Moosavinasab, S., Huber, T., Klima, J., ... & Lin, S. M. (2019). Currently Reading A Deep Learning Model for Pediatric Patient Risk Stratification. *The American Journal of Managed Care.*

[61] https://healthpolicy.duke.edu/sites/default/files/atoms/files/characterizing_rwd.pdf, accessed 31.01.2020

[62] https://healthpolicy.duke.edu/sites/default/files/u31/rwd_reliability.pdf, accessed 31.01.2020

[63] Gabler S, Häder S (2019) Repräsentativität: Versuch einer Begriffsbestimmung. In Häder S, Häder M, Schmich S (Hrsg.) *Telefonumfragen in Deutschland*, 2019, Springer, S. 35-43.

[64] https://www.iqwig.de/de/glossar.2727.html, accessed 02.02.2020

[65] Ortega, H., Li, H., Suruki, R., Albers, F., Gordon, D., & Yancey, S. (2014). Cluster analysis and characterization of response to mepolizumab. A step closer to personalized medicine for patients with severe asthma. *Annals of the American Thoracic Society*, *11*(7), 1011-1017.

[66] Feng, Q. Y., Vasile, R., Segond, M., Gozolchiani, A., Wang, Y., Abel, M., ... & Dijkstra, H. A. (2016). ClimateLearn: A machine-learning approach for climate prediction using network measures. *Geoscientific Model Development*.

[67] Robinson, C., & Dilkina, B. (2018). A machine learning approach to modeling human migration. In *Proceedings of the 1st ACM SIGCAS Conference on Computing and Sustainable Societies* (pp. 1-8).

[68] Shen, S., Jiang, H., & Zhang, T. (2012). Stock market forecasting using machine learning algorithms. *Department of Electrical Engineering, Stanford University, Stanford, CA*, 1-5.

[69] Hsich, E., Gorodeski, E. Z., Blackstone, E. H., Ishwaran, H., & Lauer, M. S. (2011). Identifying important risk factors for survival in patient with systolic heart failure using random survival forests. *Circulation: Cardiovascular Quality and Outcomes*, *4*(1), 39-45.

[70] Lee, B. J., & Kim, J. Y. (2015). Identification of type 2 diabetes risk factors using phenotypes consisting of anthropometry and triglycerides based on machine learning. *IEEE journal of biomedical and health informatics*, *20*(1), 39-46.

[71] Li, Y., Gong, S., & Liddell, H. (2000). Support vector regression and classification based multi-view face detection and recognition. In *Proceedings Fourth IEEE International Conference on Automatic Face and Gesture Recognition (Cat. No. PR00580)* (pp. 300-305). IEEE.

[72] Patil, P. B. (1998). Multilayered network for LPC based speech recognition. *IEEE Transactions on Consumer Electronics*, *44*(2), 435-438.

[73] Foster, K. R., Koprowski, R., & Skufca, J. D. (2014). Machine learning, medical diagnosis, and biomedical engineering research-commentary. *Biomedical engineering online*, *13*(1), 94.

[74] Fakoor, R., Ladhak, F., Nazi, A., & Huber, M. (2013, June). Using deep learning to enhance cancer diagnosis and classification. In *Proceedings of the international conference on machine learning* (Vol. 28). New York, USA: ACM.

[75] Zacharaki, E. I., Wang, S., Chawla, S., Soo Yoo, D., Wolf, R., Melhem, E. R., & Davatzikos, C. (2009). Classification of brain tumor type and grade using MRI texture and shape in a machine learning scheme. *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine*, *62*(6), 1609-1618.

[76] Sese, A., Palmer, A. L., & Montano, J. J. (2004). Psychometric measurement models and artificial neural networks. *International Journal of Testing*, 4(3), 253-266.

[77] Rosenblatt, F. (1958). The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6), 386.

[78] Kelley, H. J. (1960). Gradient theory of optimal flight paths. *Ars Journal*, 30(10), 947-954.

[79] Lighthill, I. (1973). Artificial Intelligence: A General Survey. In *Artificial Intelligence: A Paper Symposium*. London: Science Research Council.

[80] Pearl, J. Probabilistic reasoning in intelligent systems. 1988. *San Mateo, CA: Kaufmann*, *23*, 33-34.

[81] Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine learning*, 20(3), 273-297.

[82] Breiman, L. (2001). Random forests. *Machine learning*, 45(1), 5-32.

[83] Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural networks*, *61*, 85-117.

[84] https://www.medtechintelligence.com/news_article/apple-watch-4-gets-fda-clearance/, accessed 02.02.2020

[85] https://ainowinstitute.org/, accessed 02.02.2020

[86] Friedman, J., Hastie, T., & Tibshirani, R. (2001). *The elements of statistical learning* (Vol. 1, No. 10). New York: Springer series in statistics.

[87] James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 112, pp. 3-7). New York: springer.

[88] Libbrecht, M. W., & Noble, W. S. (2015). Machine learning applications in genetics and genomics. *Nature Reviews Genetics*, *16*(6), 321-332.

[89] Einav, L., & Levin, J. (2014). The data revolution and economic analysis. *Innovation Policy and the Economy*, *14*(1), 1-24.

[90] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Model-agnostic interpretability of machine learning. *arXiv preprint arXiv:1606.05386*.

[91] Peltola, T. (2018). Local interpretable model-agnostic explanations of Bayesian predictive models via Kullback-Leibler projections. *arXiv preprint arXiv:1810.02678*.

[92] Molnar, C. (2019). *Interpretable machine learning*. https://christophm.github.io/interpretable-ml-book/

[93] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1135-1144).

[94] Juang, B. H., & Rabiner, L. R. (1991). Hidden Markov models for speech recognition. *Technometrics*, *33*(3), 251-272.

[95] Koch, C. (2016). How the computer beat the go player. *Sci Am Mind*, *27*, 20-23.

[96] Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., ... & Lillicrap, T. (2018). A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play. *Science*, *362*(6419), 1140-1144.

[97] Freund and Schapire (1997). A decision-theoretic generalization of online learning and an application to boosting, *Journal of Computer and System Sciences*, 55, 119-139.

[98] Breiman (1996). Bagging predictors, *Machine Learning*, 26, 123-140.

[99] Huang G.-B., Zhou Q.-Y. and Siew C.K. (2006). Extreme learning machine: theory and applications, *Neurocomputing*, 70, 1-3, 489-501.

[100] Barrachina S. et al. (2009). Statistical Approaches to computer –assisted translation. *Computational Linguistics*, 35, 1, 3-28.

[101] Kozielski M., Doetsch P. and Ney H. (2013). Improvements in RWTH´s system for off-line handwritten recognition. *Proceedings of the International Conference on Document Analysis and Recognition*.

[102] https://ec.europa.eu/info/resources-partners/machine-translation-public-administrations-etranslation_en#translateonline, accessed 07.02.2020

[103]https://www.bundespolizei.de/Web/DE/04Aktuelles/01Meldungen/2018/10/181011_abschlussbericht_gesichtserkennung_down.pdf?__blob=publicationFile&v=1, accessed 07.02.2020

[104] Chen P. And Liu Z. (2018). Broad Learning Systems: An effective and efficient incremental learning system without the need for deep architecture, *IEEE Transactions on neural networks and learning sytems*, 29, 1, 10-25.

[105] Györfi L., Kohler M., Krzyzak A. and Walk, H. (2002). *A Distribution-Free Theory of Nonparametric Regression*, Springer.

[106] Chen, S., Cowan, C. F., & Grant, P. M. (1991). Orthogonal least squares learning algorithm for radial basis function networks. *IEEE Transactions on neural networks*, 2(2), 302-309.

[107] https://www.bmi.bund.de/SharedDocs/downloads/DE/publikationen/themen/it-digitalpolitik/gutachten-datenethikkommission.pdf;jsessionid=6B37F4B2D6F0875D6DB26D190B85F5C0.2_cid373?__blob=publicationFile&v=6, accessed 10.02.2020

[108] Molenberghs, G., Fitzmaurice, G., Kenward, M. G., Tsiatis, A., & Verbeke, G. (2014). *Handbook of missing data methodology*. Chapman and Hall/CRC.

[109] https://www.dagstat.de/ueber-uns/ziele/, accessed 10.02.2020

[110] Wager, S., Hastie, T., & Efron, B. (2014). Confidence intervals for random forests: The jackknife and the infinitesimal jackknife. *The Journal of Machine Learning Research*, *15*(1), 1625-1651.

[111] Gal, Y., & Ghahramani, Z. (2016). Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *International Conference on Machine Learning* (pp. 1050-1059).

[112] Garnelo, M., Rosenbaum, D., Maddison, C. J., Ramalho, T., Saxton, D., Shanahan, M., ... & Eslami, S. M. (2018). Conditional neural processes. *arXiv preprint arXiv:1807.01613*.

[113] Osband, I., Blundell, C., Pritzel, A., & Van Roy, B. (2016). Deep exploration via bootstrapped DQN. In *Advances in neural information processing systems* (pp. 4026-4034).

[114] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(1), 267-288.

[115] Wager, S., Wang, S., & Liang, P. S. (2013). Dropout training as adaptive regularization. In *Advances in neural information processing systems* (pp. 351-359).

[116] Zaremba, W., Sutskever, I., & Vinyals, O. (2014). Recurrent neural network regularization. *arXiv preprint arXiv:1409.2329*.

[117] Tibshirani, R. (1997). The lasso method for variable selection in the Cox model. *Statistics in Medicine*, *16*(4), 385-395.

[118] Ross, A., Lage, I., & Doshi-Velez, F. (2017). The neural lasso: Local linear sparsity for interpretable explanations. In *Workshop on Transparent and Interpretable Machine Learning in Safety Critical Environments, 31st Conference on Neural Information Processing Systems*.

[119] Kohavi, R., Tang, D., Xu, Y., Hemkens, L. G., & Ioannidis, J. P. (2020). Online randomized controlled experiments at scale: lessons and extensions to medicine. *Trials*, *21*(1), 1-9.

[120] Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, *15*(1), 1929-1958.

[121] Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI magazine*, *17*(3), 37-37.

[122] Theodorou, V., Abelló, A., Thiele, M., & Lehner, W. (2017). Frequent patterns in ETL workflows: An empirical approach. *Data & Knowledge Engineering*, *112*, 1-16.

[123] Catalogue of bias collaboration, Lee H, Aronson JK, Nunan D. Collider bias. In Catalogue Of Bias. 2019. https://catalogofbias.org/biases/collider-bias/, accessed 12.02.2020.

[124] https://en.wikipedia.org/wiki/Simpson%27s_paradox#/media/File:Simpson's_paradox_continuous.svg

[125] Bartels, Daniel M., Hastie Reid & Urminsky, Oleg. (2018). Connecting laboratory and field research in judgement and decision making: causality and the breadth of external validity. *Journal of Applied Research in Memory & Cognition*, 7, 1, 11-15.

[126] Braver, Sanford L. & Smith, Melanie C. (1996). Maximizing both internal and external validity in longitudinal true experiments with voluntary treatments: the “combined modified” design. *Evaluation & Program Planning*, 19, 4, 287-300.

[127] Roe, Brian E. & Just, David R. (2009). Internal and external validity in economics research: tradeoffs between experiments, field experiments, natural experiments, and field data. *American Journal of Agricultural Economics*, 91, 5, 1266-1271.

[128] https://www.destatis.de/DE/Methoden/Qualitaet/sicherung-datenqualitaet.html, accessed 14.02.2020

[129] McCracken, M. W., & Ng, S. (2016). FRED-MD: A monthly database for macroeconomic research. *Journal of Business & Economic Statistics*, *34*(4), 574-589.

[130] Ng, S (2018). Opportunities and Challenges: Lessons from Analyzing Terabytes of Scanner Data. Advances in Economics and Econometrics: Eleventh World Congress of the Econometric Society Monographs, p.1-34. In B. Honor´e, A. Pakes, and L. Samuelson (eds). Cambridge University Press.

[131] Rubin, D. B. (2006). *Matched sampling for causal effects*. Cambridge University Press.

[132] Rosenbaum, P. R. (2017). *Observation and Experiment: An Introduction to Causal Inference*. Harvard University Press.

[133] Rosenbaum, P. R. (2002). *Observational Studies.* Springer.

[134] Rosenbaum, P. R. (2010). *Design of observational studies* (Vol. 10). New York: Springer.

[135] Heckman, J. J. (2001). Micro data, heterogeneity, and the evaluation of public policy: Nobel lecture. *Journal of political Economy*, *109*(4), 673-748.

[136] Zhu, J., Chen, J., Hu, W., & Zhang, B. (2017). Big learning with Bayesian methods. *National Science Review*, *4*(4), 627-651.

[137] Biehler, R., Budde, L., Frischemeier, D., Heinemann, B., Podworny, S., Schulte, C., & Wassong, T. (Eds.). (2018). Paderborn Symposium on Data Science Education at School Level 2017: The Collected Extended Abstracts. Paderborn: Universitätsbibliothek Paderborn. http://dx.doi.org/10.17619/UNIPB/1-374.

[138] Valiant, L. G. (1984). A theory of the learnable. *Communications of the ACM*, *27*(11), 1134-1142.

[139] Valiant, L. (2013). *Probably Approximately Correct: Nature's Algorithms for Learning and Prospering in a Complex World*. Basic Books (AZ).

Autoren:

Gerd Antes	Cochrane Deutschland	GMDS – Biometrie
Sigrid Behr	Novartis	DGEpi (AG Statistische Methoden in der Epidemiologie)
Harald Binder	Universität Freiburg	GMDS – Biometrie
Werner Brannath	Universität Bremen	IBS-DR
Florian Dumpert	Destatis	Destatis
Tim Friede	Universitätsmedizin Göttingen	DAGStat
Sarah Friedrich	Universitätsmedizin Göttingen	DAGStat
Johannes Lederer	Ruhr Universität Bochum	FG Stochastik
Heinz Leitgöb	Universität Eichstätt-Ingolstadt	Sektion Methoden der emp. Sozialforschung DGS
Katja Ickstadt	Technische Universität Dortmund	IBS-DR
Hans Kestler	Universität Ulm	GfKl - Data Science Society
Markus Pauly	Technische Universität Dortmund	FG Stochastik
Lydia Spies	Destatis	Destatis
Ansgar Steland	RWTH Aachen	DStatG
Adalbert Wilhelm	Jacobs University Bremen	GfKl - Data Science Society

Redaktionelle Bearbeitung: Reiner Latsch