Positionspapier Deutsche Statistische Gesellschaft

Vorsitzender Prof. Dr. Wolfgang Schmid
Geschäftsführerin Anna-Liesa Otto
Schatzmeister Prof. Dr. Philipp Otto

Die Rolle der Statistik

für Big Data, Data Literacy, Machine Learning, KI, Analytics und Data Science

Warum die digitalisierte Informations- und Wissensgesellschaft statistische Kompetenzen braucht

Datenverständnis und die Extraktion von Wissen aus Daten sind von wachsender Bedeutung für Wissenschaft, Wirtschaft und Gesellschaft. Defiziten soll mit Initiativen und Programmen zu Themen wie Data Literacy, Statistical Literacy, Künstliche Intelligenz und Data Engineering begegnet werden. Data Science etabliert sich als neues Wissenschaftsgebiet, das Teilgebiete der Fächer Informatik, Statistik und Mathematik umspannt und Lösungsansätze für neue digitale Daten, Big Data, Datenanalyse und Wissensextraktion verfolgt.

Die Statistik muss als etablierte Wissenschaft der Datenanalyse und Inferenz eine zentrale Position in diesen Prozessen und Strukturen einnehmen. Ohne statistisches Know-How laufen die Bemühungen Gefahr, ihre Ziele zu verfehlen.

Die Deutsche Statistische Gesellschaft versteht Data Literacy und Data Science als integrale Teilgebiete der Statistik, hat diese in ihre Systematik übernommen und fördert Forschung und Lehre über diese Themen.

Der gestiegene Bedarf an Kompetenzen und Fähigkeiten in den Bereichen Data Literacy, Digitalisierung, Big Data, Datenanalyse und Data Science erfordert substanzielle Beiträge der Statistik. Die Entwicklung und Vermittlung benötigter Kenntnisse gehört in die Hände qualifizierter Statistikerinnen und Statistiker.

Die Deutsche Statistische Gesellschaft fordert daher einen umfassenden Ausbau der Statistik in Forschung und Lehre (incl. statistischer Beratung) an Universitäten und Hochschulen.

Daten und die stark wachsende Nachfrage, diese zu verstehen und aus ihnen Informationen, Wissen und Erkenntnis zu extrahieren, sind heutzutage allgegenwärtiger als je zuvor. Die hier von Gesellschaft, Wirtschaft und Politik wahrgenommenen – und zum Teil realen - Probleme und Defizite hinsichtlich der Verfügbarkeit von Analysekompetenz werden als kritische Faktoren eingestuft. Probleme und Lösungsansätze werden unter den Begriffen *Data Literacy*, *Statistical Literacy, Data and Business Analytics*, *Künstliche Intelligenz (KI), Data Engineering, Knowledge Engineering* und *Data Science* diskutiert und haben im Wissenschaftssystem bereits zu umfangreichen Aktivitäten und Initiativen unter diesen Schlagworten geführt, von 100 zusätzlichen KI-Professuren über neue Forschungsstrukturen bis hin zu neuen

Qualifizierungsmöglichkeiten und Studiengängen, letztere oftmals mit ausgeprägtem online- und E-Learning Bezug.

Dieser Prozess, der stark von Akteuren der Nachfrageseite bestimmt wird, die oftmals keine einschlägige statistische Fachausbildung besitzen, wirft Fragen hinsichtlich der Rolle und Bedeutung der Statistik auf, insbesondere im Verhältnis zu *Data Science*.

Die *Datenwissenschaften* (*Data Science*) formen sich international als wissenschaftliches Gebiet, interdisziplinär und getrieben durch Anwendungen in der Wirtschaft, großen Teilen der Wissenschaften und der amtlichen Statistik. Anliegen sind die Erforschung und Anwendung von wissenschaftlichen Methodiken für die Informations- und Erkenntnisgewinnung aus Daten und datengesteuerte Problemlösungen durch Verarbeitung, Aufbereitung, Analyse und Inferenz von sehr großen, hochdimensionalen Datenbeständen (*Big Data, neue digitale Daten*). Das Gebiet der *Künstlichen Intelligenz* verfolgt dies speziell mit dem Ziel, intelligentes Verhalten aus Daten für selbstoptimierende KI-Systeme zu extrahieren, vor allem mit Methoden des *Machine Learnings* und *Deep Learning Networks*. Dies erfordert Kompetenzen, die in der Vergangenheit oft nur verteilt über die Fächer Informatik, Statistik und Mathematik vorlagen.

Auch wenn das Verständnis von *Data Science* aus dieser Problemlage entstanden und noch im Fluss ist, so werden doch stets Modelle, Methoden und Erkenntnisse aus der Statistik und Informatik, sowie Optimierung, Numerik und dem Anwendungsgebiet (*Domain Knowledge*) studiert, eingesetzt und problembezogen weiterentwickelt. Zwar können wichtige Spezialprobleme in den einzelnen Disziplinen sinnvoll eigenständig beforscht werden und haben diese bereits substantiell und nachhaltig befruchtet und in Teilen auch neu geformt. Dennoch ist *Data Science* als wissenschaftliches Thema durch interdisziplinäres Denken geprägt und bezieht hieraus sein besonderes Profil.

Die Deutsche Statistische Gesellschaft hat diesen Entwicklungen, die auch eine Entwicklung der Statistik waren und sind, Rechnung getragen und *Data Science* neben *Computational Statistics* und *Statistical Literacy* in ihre Fachsystematik und Jahrestagungen fest integriert. Die Entwicklung und Einordnung von *Data Science* und die besondere Rolle der Statistik wurden in Ausschüssen der Gesellschaft und Jahrestagungen ebenso diskutiert wie Literacy-Fragestellungen.

Die Statistik nimmt eine zentrale Position in den Datenwissenschaften ein. Sie ist die Wissenschaft und praktische Disziplin, die die Lösung des ersten *Big Data*-Problems der Menschheit, der Volkszählungen und Bevölkerungsstatistik, ermöglichte. Sie erforscht basierend auf der Wahrscheinlichkeitsrechnung seit jeher essentielle Kernfragen für Datenverständnis und Wissensextraktion, nämlich Datendeskription, Datenexploration und Datenanalyse sowie Stichprobentheorie und Inferenzstatistik. Theoretische Grundlagen wurden bereits vor Beginn des Computerzeitalters gelegt und praktisch angewendet. Die Statistik hat nicht nur wesentliche Grundlagen einer theoretischen Fundierung vieler Verfahren des *Machine Learnings* durch die *Statistical Learning Theory* geliefert, sondern auch mit *Random-Forrest*-Klassifizierern und *Bagging*-Methoden einige der heutzutage meist verwendeten *Machine-Learning*-Verfahren für Datenanalyse und Prädiktion entwickelt. Diese, wie auch Weiterentwicklungen klassischer statistischer Methoden, werden heutzutage in Gebieten wie der medizinischen Diagnostik, Business Analytics, Bildverarbeitung und in autonomen Systemen intensiv eingesetzt. Mit interpretierbaren Modellen, fundierten Ansätzen zur Quantifizierung von Unsicherheiten und Bewertung von Replizierbarkeit sowie substantiellen Fortschritten bei der statistischen Inferenz für Big-Data-Analysen tragen die

moderne Statistik und Stochastik auch zu aktuellen Entwicklungen und Forschungstrends entscheidend bei. Statistische Expertise ist ebenfalls an vielen Stellen relevant für verbesserte Algorithmen und deren Verständnis. So ist die statische Kreuzvalidierung ein wichtiges Instrument für die Trainingsphase von *Deep Learnern*, um eine gute Generalisierungsfähigkeit zu erreichen.

Die Spezialisierung der Statistik als Disziplin auf einzelne Wissenschaftsfelder und der Einbezug von Expertenwissen über diese Felder hat zur Etablierung von Teildisziplinen wie der *Biometrie/Biostatistik*, *Umweltstatistik*, *Industriestatistik* oder der *Ökonometrie* geführt. Die technologischen Entwicklungen in der Rechentechnik und die Digitalisierung von Gesellschaft, Wirtschaft und empirischen Wissenschaften wurden in der Statistik frühzeitig aufgegriffen und haben viele Teilgebiete nachhaltig verändert. Insbesondere entstand die neue Teildisziplin der *Rechnergestützten Statistik*. Ebenso haben sich *Hochdimensionale Statistik* und *Statistisches Maschinelles Lernen* als Forschungsgebiete der Statistik etabliert und in Form eines erweiterten Methodenspektrums Eingang in anwendungsbezogene Gebiete (insbes. *Ökonometrie*, *Empirische Wirtschaftsforschung*, *Biostatistik*, *Technische Statistik*) gefunden. Die neuen Möglichkeiten durch Hardware und Software erlauben ebenfalls deutlich komplexere stochastische Modellierungen und Methoden zur Etablierung und Anwendung der benötigten statistischen Theorien.

Viele der unter dem Begriff *Data Science* adressierten Themen und Probleme finden in diesen Entwicklungen der Statistik ihre natürlichen Anknüpfungspunkte und wissenschaftliche Zitationsbasis, auch wenn sie durchaus neue Herausforderungen formulieren. Dies sowohl für die disziplinäre Forschung in der *Mathematischen* und *Angewandten Statistik* und *Stochastik*, als auch für interdisziplinäre Forschung und Datenanalyse. Hier sind insbesondere die Ökonometrie, Industriestatistik, Ausbildung und Lehre und die Amtliche Statistik zu nennen. Es ist festzustellen, dass eigenständige Initiativen zu *Data Science* oder *Data Literacy* besonders in denjenigen Wissenschaftsfeldern entstehen, in denen der Siegeszug der Statistik im letzten Jahrhundert nicht nachhaltig Eingang gefunden hat in Form der Etablierung von Statistikprofessuren oder der Herausbildung einer selbstständigen statistischen Teildisziplin.

Die valide Analyse von sehr großen Datenbeständen erfordert substantielle Beiträge der Statistik und somit qualifizierte Statistiker, die umfassende Kompetenzen auch in Bereichen wie *Machine Learning*, *Data Privacy and Literacy*, *paralleles Rechnen*, *Algorithmik* und *Optimierung* besitzen. Eine rein algorithmische Sichtweise, die *Data Science* als Ingenieursfach oder Teilgebiet der Informatik versteht und auf eine datenverarbeitende, algorithmische Sicht reduziert, greift deutlich zu kurz und wird selbst gesteckte Ziele wie Unsicherheitsquantifizierung oder Erklärbarkeit von KI nicht erreichen können. Sie läuft insbesondere auch Gefahr, die grundsätzliche wissenschaftstheoretische Erkenntnis zu ignorieren, dass eine Ergebnisinterpretation und die Bewertung von Datenunsicherheit Statistik benötigt in Form von wissenschaftlichen, falsifizierbaren Modellen, die Mechanismen der Datengenerierung berücksichtigen.

Statistik zeigt die Möglichkeiten und Grenzen der Wissensextraktion aus Daten auf und stellt somit auch die Grundlage für einen kritischen Umgang mit Daten dar. Statistik war und ist die Wissenschaft für die Erkenntnisgewinnung aus Daten und jegliche Datenwissenschaft ist ohne Statistik nicht denkbar.

Vor dem Hintergrund einer Analyse in jüngerer Zeit national und international aufgelegter Programme und etablierter Studiengänge sowie neu geschaffener Forschungsstrukturen an

renommierten Einrichtungen spricht die Deutsche Statistische Gesellschaft die folgenden Empfehlungen und Forderungen aus:

Positionen und Empfehlungen der Deutschen Statistischen Gesellschaft im Einzelnen:

- Als Fachgesellschaft für theoretische, angewandte und praktische Statistik, die auch Statistikerinnen und Statistiker mit Expertenwissen aus Anwendungsgebieten vertritt, versteht die Deutsche Statistische Gesellschaft Data Science und Data Literacy als integrale Teilgebiete.

- Die Deutsche Statistische Gesellschaft befürwortet und fordert den Ausbau der Statistik an Hochschulen durch Einrichtung neuer Professuren und Stellen für wissenschaftliche Mitarbeiter/innen, um dem substantiell gestiegenen Bedarf an Kompetenzen und Fähigkeiten im Bereich Statistik, Digitalisierung, Data Literacy und Data Science in Lehre und Forschung nachzukommen.

  Dies ist insbesondere notwendig, damit sich der substantiell wachsende Lehr- und Schulungsbedarf nicht zu Lasten der Forschungsqualität auswirkt.

  Ein Stellenaufwuchs ist insbesondere in Hinblick auf wirtschafts- und ingenieurwissenschaftliche Fakultäten und Studiengänge notwendig, um der gestiegenen Bedeutung der Thematik für den Wirtschaftsstandort Deutschland gerecht zu werden.

- Die statistische Beratung und nicht-curriculare statistische Schulungsangebote an Hochschulen müssen bedarfsgerecht ausgebaut werden.

- An Lehr- und Forschungsstrukturen zu *Data Science*, *Machine Learning*, *KI* und *Data Literacy* (z.B. Förderprogramme, Studiengänge, Doktorandenprogramme, Forschungsverbünde, Forschungsprogramme) sowie digitalisierte Lehr- und Ausbildungsangebote (E-Learning) müssen in Statistik qualifiziert ausgebildete Lehrende und Forschende beteiligt werden. Sie sollten in die Leitung und Koordinierung eingebunden sein.

  Der Aufbau von Strukturen in Lehre und Forschung ohne Verbindung zu bestehenden Statistikinstituten bzw. –professuren und ohne maßgebliche statistische Fachkompetenz wird den Anforderungen nicht gerecht. Die Entwicklung und Vermittlung von Datenanalyse-Kompetenzen gehört in die Hände von qualifizierten Statistiker/innen. Die Deutsche Statistische Gesellschaft empfiehlt hier dringend, dies in der weiteren Entwicklung zu berücksichtigen und ggfs. nachzubessern.

- Standardisierte, breitenorientierte (etwa fächerübergreifende) Online-Learning Angebote zu Themen wie *Data Science* und *Data Literacy* sind im Sinne eines ergänzenden Angebots zu begrüßen. Sie dürfen Lehrende und die Vielfalt der Lehre und Didaktik jedoch nicht einschränken, sondern sollten diese bereichern, und können fachspezifische Lehre nicht ersetzen.