# Auf dem Weg der angewandten Statistik in die Data-Science-Zukunft

# Herausforderungen und Chancen für die Alpen-Adria Universität

Vortrag, gehalten an der Zehnjahresfeier des Instituts für Statistik an der AAU am 1. Dezember 2017

Spektabilität, Vorstand und Gründungsvorstand, geschätzte Kolleginnen und Kollegen, werte Absolventinnen und Absolventen, sehr geehrte Festgäste!

Es ist mir eine große Ehre aus Anlass des zehnjährigen Bestehens des Institutes für Statistik an der Alpen-Adria Universität zu Ihnen sprechen zu dürfen. Meine Verbundenheit mit dem Institut ist nicht nur geographischer Natur, Graz ist die nächstgelegene österreichische Universitätsstadt, sondern auch persönlicher Natur: berufliche Kontakte mit den Wegbereitern des heutigen Institutes, insbesondere Emeritusprofessor Haro Stettner und Prof. Manfred Borovcnik.

Auf den explizit ersten Lehrstuhl für Statistik in Klagenfurt wurde im Jahr 1994 Prof. Jürgen Pilz berufen. Zuvor, von 1993 bis 1994, war ich als Gastprofessor für Angewandte Statistik am Institut für Mathematik der damaligen Universität für Bildungswissenschaften tätig. Dies war kurz nach meiner Habilitation für das Gesamtfach Statistik an der Karl-Franzens-Universität Graz.

Der Begriff *Data Science* war damals noch nicht bekannt. Aber *Data* waren zu dieser Zeit bereits im Zentrum meiner Tätigkeit: Ich war jene Person in Klagenfurt, die die statistische Programmiersprache S einführte. S war eine bahnbrechende Entwicklung der Bell Laboratories in den USA, ab 1988 New S genannt, und ist heute in ihrer modernen Implementierung als R bekannt. In meine damalige Lehre, aber auch Forschung, baute ich noch einen weiteren Aspekt ein: statistische Modellierung in der Abfolge Modellidentifikation, -schätzung und -kritik auf der Basis des sogenannten „White Book“ (Chambers, J. M. und Hastie, T. J. (1992): Statistical Models in S. CRC Press, Boca Raton). Dieses Werk leitete eine nachhaltige Entwicklung ein, die noch heute in der *Big Data*-Analytik unverzichtbar ist: die einheitliche Formelnotation (erkennbar am Tilde-Zeichen) bei der Modellspezifikation und *Data Frames* als Datenobjekte. Mit dem genannten Rüstzeug machten wir sofort computergestützte Übungen. Meine Studierenden der Mathematik und Informatik waren begeistert. So entstand eine solide Basis für den bis heute andauernden umfassenden Einsatz von R am Institut für Statistik.

Von 1991 -1993 arbeitete ich in einem FWF-Grundlagenprojekt mit Prof. Haro Stettner zusammen. Sein Titel war „Asymptotik, Numerik und Implementation des Backfitting-Algorithmus bei Verallgemeinerten Additiven Modellen“. Damals galt international das Interesse den nonparametrischen Alternativen zu den Generalisierten Linearen Modellen. Der Backfitting-Algorithmus war das erste Schätzverfahren für rein nonparametrische Regressionsmodelle auf der Basis linearer Glätter. Über seine theoretischen Grundlagen und sein numerisches Verhalten, abhängig von der Art der beteiligten Glätter (Kern, Spline oder lokale Polynome) war damals noch wenig bekannt. Prof. Wolfgang Härdle (Humboldt Universität zu Berlin) und ich gaben 1996 ein Buch mit dem Titel „*Statistical Theory and Computational Aspects of Smoothing*“ (Physica, Heidelberg) heraus. Es enthält auch einen Beitrag zu dem genannten FWF-Projekt.

In den frühen 1990er Jahren gab es auch statistische Kolloquien und Seminare in Klagenfurt, teilweise in Kooperation mit der Sektion Steiermark-Kärnten der Region Österreich-Schweiz der Internationalen Biometrischen Gesellschaft, die oft von Prof. Manfred Borovcnik organisiert wurden. Bedeutende Sprecher waren hier zum Beispiel Prof. Angelika van der Linde (Universität Bremen) und Prof. Wilfried Grossmann (Universität Wien). Van der Linde war damals eine ausgewiesene Expertin in glättenden Splines und Bayes-Methodik und Grossmann ein Innovator in Datenstrukturen wie sie heute im Kontext von *Big Data* aktuelles Forschungsthema sind. Überdies war Grossmann ein Brückenbauer zwischen Statistik und Informatik wie wir ihn bei den aktuellen *Big Data*-Herausforderungen mehr denn je brauchen würden. Klagenfurt war ohne Zweifel bereits auf der Landkarte der Statistik verzeichnet lange bevor es zur Gründung des heute gefeierten Institutes kam.

Was ich seither sehe, ist eine Profilbildung in der technikorientierten angewandten Statistik durch Prof. Jürgen Pilz. Schwerpunkt sind eindeutig Bayes'sche Methoden zur räumlichen Analyse, insbesondere die Entwicklung neuer Kriging-Verfahren, heute in der GIS Software *Geostatistical Analyst* implementiert. Typische Anwendungen sind Klimamodellierung, Umweltstatistik (z. B. Erfassen und Kartieren kritischer Umweltvariablen) und Industriestatistik (z. B. automatische Defektklassifikation). Beeindruckend ist die Fülle von drittmittel- und industriefinanzierten Projekten, welche es erlaubten langfristig Absolventen/innen an das, an sich personalschwache, Institut zu binden.

Des Weiteren identifiziere ich am Institut für Statistik wesentliche Beiträge zur „Statistics Education" von Prof. Manfred Borovcnik. Wichtige Arbeitsbereiche sind statistische Inferenz aus unterschiedlichen methodologischen Perspektiven, Modellbildung als gestalterisches Element im statistischen Curriculum, Untersuchungen zur Verbindung von Wahrscheinlichkeit und Risiko, sowie e-Learning und technologiegestütztes Lernen. Neben kritischen Diskussionsbeiträgen in Form von Publikationen und bei Fachtagungen sind die Entwicklung dynamischer Applets zum Lernen komplizierter statistischer Begriffe besonders hervorzuheben.

Als besondere Leistung aller Lehrenden am Institut sehe ich die äußerst erfolgreiche und praxisnahe Betreuung von Studierenden. Insgesamt gab es bisher 38 Diplomabschlüsse und 24 Promotionen (bereits zum zweiten Mal eine „Promotio sub auspiciis praesidentis rei publicae"). Derzeit werden 8 Masterarbeiten und 5 Dissertationen betreut.

Es stellt sich nun die Frage: Wie kann die Zukunft des Institutes für Statistik aussehen?

Faktum ist, dass fast alle neuen Technologien ohne das hier vertretene Fach nicht denkbar wären:

- keine Mobiltelephonie ohne Statistik,
- keine Webdienste ohne Statistik
- kein Satellitenmonitoring ohne Statistik
- keine biometrische Erkennung ohne Statistik
- keine Smarttechnologien ohne Statistik
- keine autonom fahrenden Autos ohne Statistik
- keine vernetzten Maschinen ohne Statistik
- keine personalisierte Medizin ohne Statistik

Diese Liste könnte fast beliebig fortgesetzt werden.

Welche gemeinsamen Merkmale haben nun all diese High-Tech-Anwendungen? Es sind in erster Linie:

- enorme Datenmengen, oft mit einer räumlichen und/oder zeitlichen Komponente
- Echtzeitdaten (data streams)
- hohe Komplexität, also schwach oder nicht strukturierte Daten
- hohe Dimensionalität, also um viele Größenordnungen mehr unbekannte Parameter als Beobachtungseinheiten oder aber niedrige Dimensionalität bei unerschöpflich vielen Beobachtungseinheiten
- schwierige Separierbarkeit von relevanter Information und Noise

Hier greifen klassische statistische Paradigmen nicht mehr. Ohne Zweifel werden neue Analysemethoden benötigt. Doch die statistische Grundlagenforschung hat bereits jetzt viele Fundamente gelegt, auf denen *Data Science*-Ansätze aufgebaut werden können. Denken wir nur an die schon erwähnten Glättungsmethoden, aber auch an die große Klasse der Penalisierungsverfahren. Damit haben wir Werkzeuge um mit schwach besetzten Systemen (data sparsity) umgehen zu können. Eine Kombination dieser beiden Verfahren eröffnet bereits jetzt Möglichkeiten für die Analyse hoch komplexer Daten.

Spätestens nach meiner Gastprofessur am Institut für Statistik im letzten Wintersemester sollten diese Verfahren hier nicht mehr unbekannt sein. Das Interesse und die Motivation der Studierenden waren groß und die Lehre auf diesen neuen Gebieten sollte aus den genannten Gründen unbedingt fortgesetzt werden.

Wenden wir uns nun weiter der aktuellen fachlichen Entwicklung zu:

*Big Data* ist in aller Munde. Nicht wenige Entscheidungsträger in Politik, Wirtschaft und Wissenschaft vertreten die Meinung, dass die Herausforderungen der neuen Technologien nur durch eine neue Disziplin namens *Data Science* gemeistert werden könnten. Somit stellt sich die Frage nach den Inhalten von *Data Science*.

Um diese zu beantworten, möchte ich folgende objektive Betrachtung anstellen: Ein Hauptreferenzwerk ist der aktuelle Sammelband „*Handbook of Big Data",* herausgegeben von Peter Bühlmann (ETH Zürich) gemeinsam mit anderen (CRC Press, Boca Raton, 2016). An diesem Werk haben 34 Autoren mitgearbeitet. Ich habe mir nun deren fachliche Herkunft angesehen und zwischen den akademischen Fächern Statistik, Biostatistik, Mathematik und Informatik (Computerwissenschaften) unterschieden. Das Ergebnis sieht folgendermaßen aus:

- Statistik: 15 Personen
- Biostatistik: 8 Personen
- Mathematik: 2 Personen
- Informatik: 9 Personen

Somit sind 23 Autoren Statistikinstituten zuzuordnen. Das ist eine klare Mehrheit von fast 70% der insgesamt 34 Autoren.

Sieht man sich darüber hinaus Master- und Doktoratsstudien in *Data Science* von nordamerikanischen Forschungsuniversitäten an, die hier die Vorreiter waren, dann dominieren auch dort die Statistikdepartments. An der University of California, Berkeley, kann man zum Beispiel ein PhD in Statistik absolvieren mit dem Zusatz "with a Designated Emphasis in Computational Science and Engineering" und damit dem *Data Science*-Aspekt entsprechend Rechnung tragen.

Lassen Sie mich nun die wesentlichen fachlichen Schwerpunkte des Themenbereichs *Data Science* aufzeigen. Es sind dies:

1. Datenhaltung (Datenbanken)
2. Datenzentrierte explorative Methoden (Überlappung mit Data Mining von einst)
3. Entwicklung effizienter Algorithmen (Numerik für enorm große Datensätze)
4. Angewandte Graphentheorie (Netzwerktheorie)
5. Modellschätzung (einschließlich Variablenselektion)
6. Regularisierungsmethoden (Penalisierungsverfahren)
7. Inferenzmethoden („computer age statistical inference“)
8. Statistisches Lernen (Weiterentwicklung des maschinellen Lernens)
9. Targeted Learning
10. Ensemble Methoden (z. B. „divide and recombine“)

Die Datenhaltung (1) ist eine informatische Vorleistung aber keine spezifische Aufgabe von *Data Science*. Die datenzentrierten explorativen Methoden (2) sind interdisziplinär. Die Entwicklung effizienter Algorithmen (3) ist primär der numerischen Mathematik geschuldet. Zur angewandten Graphentheorie (4) gibt es Beiträge von der Informatik als auch der Statistik. Modellschätzung (5) ist ein ureigenes Thema der Statistik. Dasselbe gilt für die Regularisierungsmethoden (6) und die Inferenzmethoden (7). Das maschinelle Lernen hat im statistischen Lernen (8) eine bedeutende Weiterentwicklung erfahren. Viele dieser hier genannten statistischen Forschungsaspekte werden in dem epochalen Buch von Trevor Hastie, Robert Tibshirani und Jerome Friedman (alle Stanford University) mit dem Titel *The Elements of Statistical Learning. Data Mining, Inference, and Prediction* (Springer, New York, 2009, 2. Auflage) behandelt. Es ist meines Wissens der einzige Statistiktitel, der regelmäßig in den Buchregalen von Informatikern zu finden ist. Der rezente Ansatz des Targeted Learning (9) ist eine eigenständige statistische Entwicklung für kausale Inferenz von Mark van der Laan (University of California, Berkeley) mit erheblichem Potential für große komplexe Datenbestände. Zu den Ensemble Methoden (10) gibt es sowohl statistische als auch informatische Forschungsansätze.

Was sagt uns nun diese Liste? Die überwiegende Zahl der fachlichen Schwerpunkte von *Data Science* betrifft statistische Arbeitsgebiete. Vielen von ihnen, wie zum Beispiel die Modellschätzung oder die Inferenzmethoden, basieren auf theoretischen Konzepten, die seit Jahrzehnten in der statistischen Datenanalyse bewährt sind. Vieles davon ist in der Sprache R implementiert und frei zugänglich.

An dieser Stelle sei angemerkt, dass bei *Big Data* oft nur an Data Mining, also die Exploration von Datenstrukturen, gedacht wird. In Wahrheit geht es jedoch fast immer um Modellbildung und Entscheidungsfindung. Selbst eine vergleichsweise einfache Aufgabe wie die maßgeschneiderte personalisierte Werbung im Web erfordert Eigenschaftsmodelle typischer Nutzer und Konsumenten. Die in naher Zukunft auf uns zukommende Industrie 4.0, informations- und kommunikationsgesteuerte industrielle Produktion, ist nicht denkbar ohne neue inferenzstatistische Methoden für Entscheidungen in Netzwerken, um nur eine Aufgabe zu nennen.

Man darf zusammenfassend sagen dass das Fach Statistik eine große Zukunft hat und der statistische Beitrag zu *Data Science* wesentlich größer ist als üblicherweise kommuniziert wird. *Big Data* ist zu einem guten Teil angewandte Statistik von morgen. Entsprechend ausgestattet, mit neuen Aufgaben versehen, sollte das Institut für Statistik an der Alpen-Adria Universität Klagenfurt diese Zukunft maßgeblich mitgestalten und einen wichtigen Beitrag zu einer zeitgemäßen technischen Ausbildung leisten können.

Mit den besten Wünschen für die Zukunft, auf viele weitere Jahrzehnte,

Ihr Prof. Dr. Dr. Michael G. Schimek