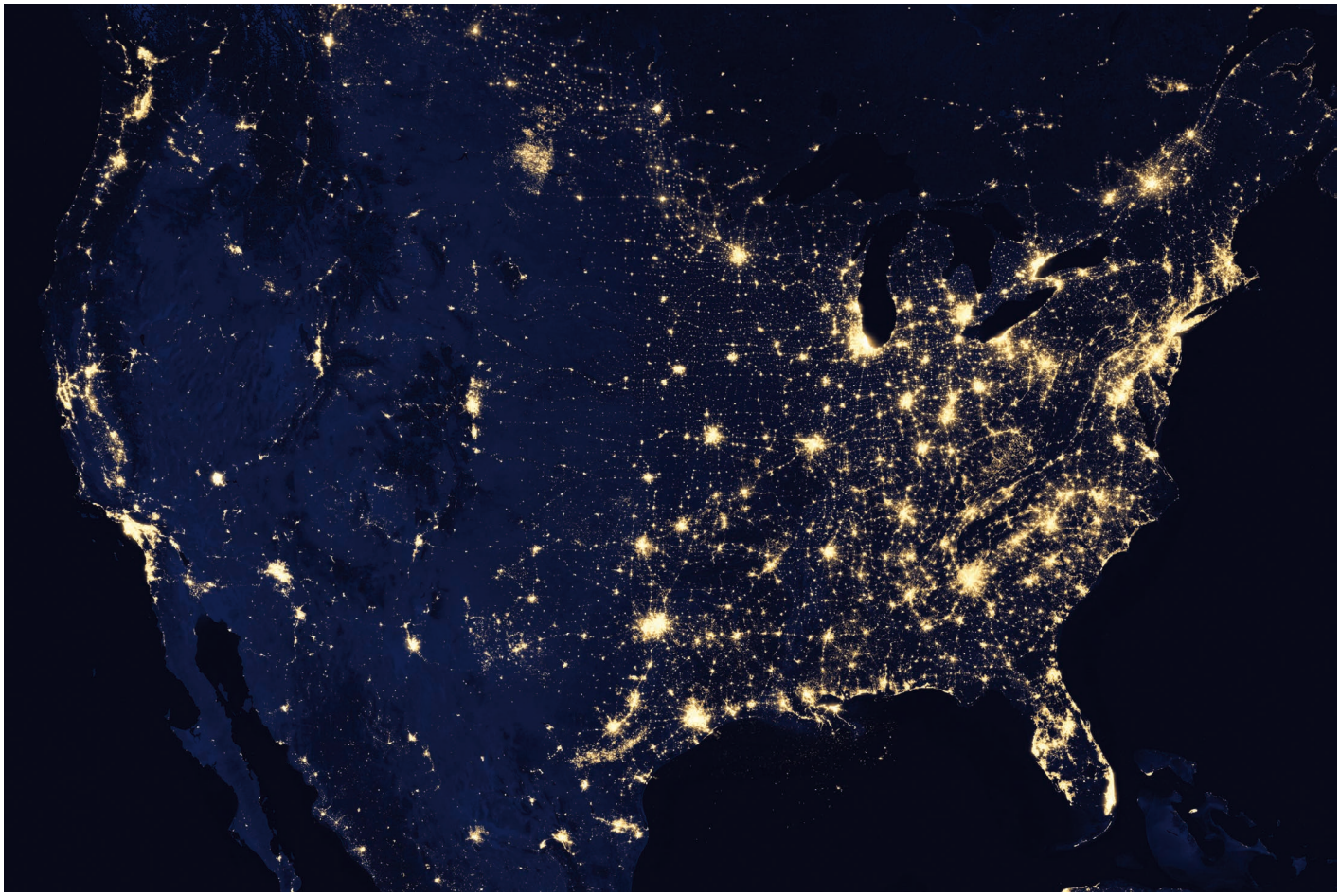
Datenwissenschaft

Was versteht man eigentlich unter „Data Science”?



Unternehmen sammeln Daten auf der ganzen Welt, beispielsweise durch geschäftliche Transaktionen, intelligente Geräte (IoT), Industrieanlagen, Videos oder Social Media

E-Mails, Social Media, Apps – jeden Tag werden Unmengen an Daten generiert, sowohl im privaten als auch im beruflichen Bereich. Bereits im Dezember 2012 kam eine Studie der International Data Corporation (IDC) zu dem Schluss, dass sich das Datenwachstum weltweit alle zwei Jahre verdoppelt. Die Existenz dieser Daten und die Notwendigkeit, sie zu speichern, zu verarbeiten und zu nutzen, sind die Grundlage von Datenwissenschaft.

Autor:
Prof. Dr. Mohammad Madhavi
Professor für Data Science
GISMA University of Applied Sciences
www.gisma.com

Die meisten Unternehmen erzeugen jeden Tag eine große Menge an strukturierten und unstrukturierten Daten – kurz: „Big Data”. Aber was passiert mit diesen ganzen Informationen? Sie müssen analysiert und nach Verwertbarkeit gefiltert werden. Das ist aus verschiedenen Gründen gar nicht so einfach: Neben der extremen Datenmenge (Volume) stellen die schnelle Datenveränderung (Velocity), die verschiedenen Datenformate (Variety), der Mehrwert, der sich durch Daten generieren lässt (Value) und die Unsicherheit bzw. der Wahrheitsgehalt der Daten (Veracity) eine Herausforderung dar. Data Scientists wenden unterschiedliche Techniken an, sodass Unternehmen von ihren Datenschätzen profitieren können.

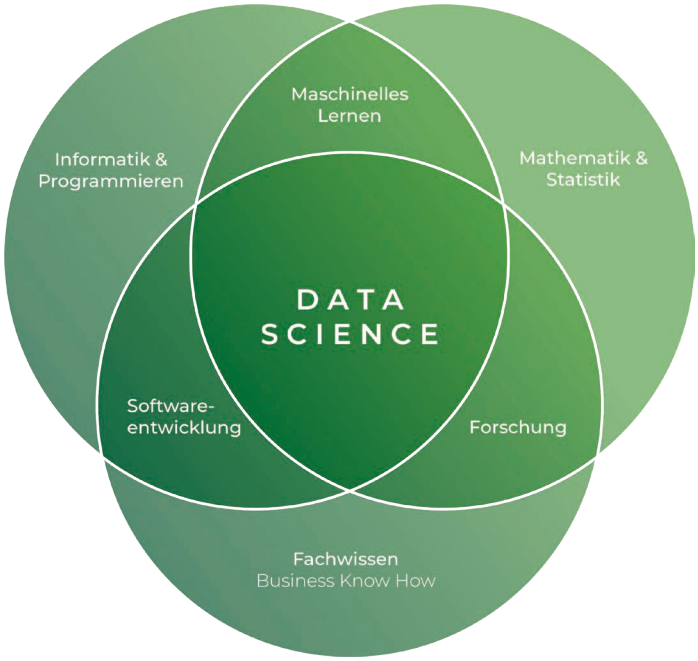
Data Science: eine neue Wissenschaft aus dem 21. Jahrhundert

Auch wenn der Begriff Data Science schon seit den 1960er-Jahren existiert, hat sich das Fachgebiet mit seinen Teilgebieten „Maschinelles Lernen” (ML) und „Deep Learning” erst Anfang des 21. Jahrhunderts entwickelt. Im Jahr 2001 etablierte William S. Cleveland die Datenwissenschaft als eigenständige Disziplin in einem wissenschaftlichen Artikel. Ein Jahr später erschien erstmals die Fachpublikation The Data Science Journal, in dem seither regelmäßig Artikel über die Verwaltung, Verbreitung, Nutzung und Wiederverwendung von Daten veröffentlicht werden. Das Berufsbild des Data Scientists wurde erst einige Jahre später eingeführt.

Definition

Data Science bezeichnet die Schnittmenge zwischen den Wissenschaftsbereichen Mathematik, Informatik und branchenspezifischem Wissen. Dabei handelt es sich um einen Sammelbegriff für Systeme, Algorithmen, Methoden und Prozesse, mit denen Wissen aus vorhandenen Daten extrahiert wird: Aus den Informationen, die aus den großen Datenmengen von Unternehmen generiert werden, sollen Handlungsempfehlungen für das Unternehmensmanagement abgeleitet werden.

Data Science-Projekte verbinden Unternehmensdaten und unternehmerische Fragestellungen. Zu den Aufgaben von Data Scientists zählen unter anderem die Datenerhebung (Data Sourcing), die Datenbereini-



Data Science bezeichnet die Schnittmenge zwischen den Wissenschaftsbereichen Mathematik, Informatik und branchenspezifischem Wissen

gung (Data Cleansing), die Datenaufbereitung sowie die Datenanalyse. Ziel dabei ist stets die Verbesserung von Unternehmensprozessen, wie zum Beispiel die Kosten- und Umsatzoptimierung, Steigerung der Vertriebserfolge oder die Prognose von Kaufverhalten und Trends.

Techniken und Methoden der Datenanalyse

Die Grundlage der Datenwissenschaft bilden Theorien und Techniken aus den Bereichen Informationstechnologie, Mathematik, Wahrscheinlichkeitsrechnung und Statistik. Eine Datenanalyse setzt sich in der Regel aus verschiedenen Schritten zusammen. Zunächst müssen die Daten eingelesen werden. Für die weitere Verarbeitung wird der Datensatz an das Projekt bzw. das spezifische Projekt angepasst und entsprechend aufbereitet (Datenvorverarbeitung). Das heißt, er muss in das richtige Format umgewandelt werden, damit er analysiert werden kann. Anschließend werden die Daten bereinigt und geprüft und fehlende Daten gegebenenfalls ersetzt.

Datenmodellierung

Der nächste Schritt ist die Datenmodellierung: Ein maschinelles Lernmodell wird trainiert, um die Beziehung zwischen Eingabe- und Ausgabevariablen zu modellieren. Nehmen wir zum Beispiel die Daten einer Hauspreisprognose in Tabellenform. Jede Zeile steht für ein Haus, und jede Spalte steht für ein Eingabeattribut wie die Anzahl der Zimmer oder die Fläche des Gebäudes. Ein spezielles Attribut ist der Preis des Hauses, der die gewünschte Ausgabe für ein bestimmtes Haus darstellt. Ein für diesen Datensatz trainiertes maschinelles Lernmodell wird eine Funktion sein, die Eingabeattribute für ein bestimmtes Haus empfangen kann und das Ausgabeattribut für dieses Haus, d. h. den Preis, zurückgibt.

Modellvalidierung

Nachdem das Modell erstellt worden ist, folgt die Modellvalidierung. In diesem Schritt wird überprüft, ob das trainierte Modell gut gewählt und abgestimmt ist oder nicht. Danach müssen die Unternehmen bewerten, ob die Vorhersageergebnisse des trainierten Modells zutreffend sind (endgültige Modellbewertung). Anschließend kann das Vorhersagemodell in der Produktion verwendet werden, um Vorhersagen für den Preis eines neuen Hauses zu treffen.

Anwendungsbereiche von Data Science

Datenwissenschaft bietet zahlreiche Einsatzmöglichkeiten: Sie lässt sich insbesondere in Unternehmensbereichen einsetzen, in denen eine Vielzahl an Daten aufkommt und Unternehmen aus ihren Datensätzen lernen und sich verbessern können. Neben der korrekten Umsetzung der Methoden und Techniken der Datenanalyse ist daher auch spezifisches Fachwissen jener Branche obligatorisch. Während früher Data Science beispielsweise in den Bereichen Astronomie, Biologie oder diversen Sozialwissenschaften zum Einsatz kam, werden Datenanalysen heute vermehrt im Online-Handel bzw. E-Commerce, in der Logistik, im Gesundheitswesen, im Finanzwesen, bei Banken und Versicherungen sowie in der Industrie und Produktion angewandt.

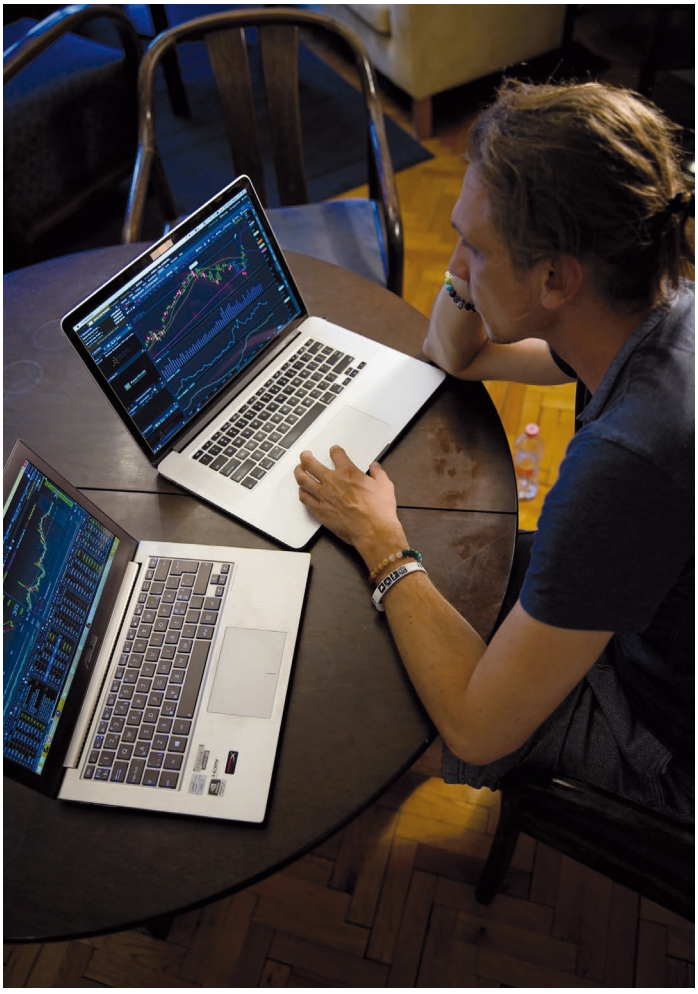
Im E-Commerce wird unter anderem das Kaufverhalten von Kunden analysiert, um etwa Prognosen zu erstellen, welche Artikel in Zukunft vermehrt produziert werden sollen. Auch die Ursachen für die vermehrte Rückgabe von bestimmten Waren können mithilfe von Data Science identifiziert werden. Eine weitere Möglichkeit von Datenanalysen ist die Nutzung im Bereich des Marketings für zielgerichtete Werbung. So werden Kunden passende Produkte vorgeschlagen, wenn sie den Onlineshop besuchen oder auch in den sozialen Medien unterwegs sind. Dafür werden historische Daten zu Transaktionen, Verhalten und Demografie der Kunden untersucht, um zukünftige Produktpräferenzen besser vorherzusagen.

Erhöhen der Effizienz

Im Versicherungs- und Finanzsektor wird Data Science eingesetzt,

Ein typischer Datenanalyse-Prozess setzt sich aus mehreren Schritten zusammen. Dazu gehört das Einlesen der Daten, die Datenvorverarbeitung und die Datenmodellierung

Data Science lässt sich in allen Unternehmensbereichen einsetzen, in denen eine Vielzahl an Daten generiert werden



In der Datenwissenschaft sind Fachkräfte aus den Bereichen Mathematik, Informatik, Programmierung, Physik, Betriebswissenschaften und Statistik tätig

um die Effizienz zu erhöhen, Produkte zu etablieren und zu verbessern und Verkaufspotenziale vollständig auszuschöpfen. Die Analyse der Daten unterstützt so unter anderem die Vertriebsmitarbeiter und Kundenberater dabei, ihre Kunden mit individuell passenden Versicherungsverträgen oder Finanzprodukten zu versorgen. Denn gerade bei Banken und Versicherungsunternehmen sind Cross- und Upselling-Potenziale im Rahmen bereits bestehender Verträge besonders wichtig.

Big Data in der Medizin

Auch die Medizin kann große Vorteile aufgrund des Einsatzes von Data Science verzeichnen. Datenanalysen können beispielsweise angewandt werden, um Körperfunktionen oder den Blutdruck zu überwachen und Unregelmäßigkeiten festzustellen, wodurch Krankheiten frühzeitig erkannt und effektiver behandelt werden können. Ebenso kann die Auswertung bestimmter Daten dazu dienen, die Abläufe in medizinischen Einrichtungen zu optimieren: Durch eine präzise Abstimmung mit dem Schichtplan und Daten zu vorherigen Untersuchungen können zum Beispiel Wartezeiten verkürzt werden – so profitieren sowohl das medizinische Fachpersonal als auch die Patienten.

Der Weg zum Data Scientist

Für Data Science braucht es nicht nur die richtigen Daten, Ressourcen, IT-Systeme, Tools und Verfahren, sondern auch qualifizierte Mitarbeiter. Viele Data Scientists sind Mathematiker, Informatiker, Programmierer, Physiker, Betriebswirtschaftler oder Statistiker, die sich auf datenwissenschaftliche Analysen spezialisiert haben. Data Scientists sollten sowohl Grundlagen der Softwareentwicklung, wie Programmiersprachen, beherrschen und Erfahrung im Umgang mit KI-Anwendungen gesammelt haben als auch betriebswirtschaftliches Wissen und Kommunikations- und Teamfähigkeit mitbringen.

Mittlerweile kann man Data Science aber auch studieren. Während im Bachelor Grundlagen der Mathematik, Informatik und Statistik unterrichtet werden, werden im Master unter anderem Kenntnisse in den Bereichen „Machine Learning", „Natural Learning Processing" und „Business Analytics" vertieft. Überdies werden beispielsweise an der GISMA University of Applied Science in dem Studiengang „Data Science, AI, and Digital Business" Bereiche wie Personalführung, Projekt- und Innovationsmanagement thematisiert, was auf dem Arbeitsmarkt weitere Vorteile bietet.

Potenziale und Herausforderungen für die Datenwissenschaft

Data Science lässt sich in fast allen Unternehmensbereichen und Branchen nutzen, um Prozesse zu optimieren. Big Data und die Analyse dieser Daten bergen große Potenziale für Unternehmen. An Daten wird es den meisten Unternehmen heutzutage nicht mehr mangeln – die Herausforderung ist vielmehr die Verfügbarkeit von qualitativ hochwertigen Datensätzen, das Überschätzen der Datenqualität und das Unterschätzen der negativen Konsequenzen bei schlechter Qualität. Hinzu kommt die Notwendigkeit eines gut aufgebauten, strukturierten Teams von Experten, die sich in die Verantwortungs- und Aufgabenbereiche aufteilen und miteinander arbeiten. Die Umsetzung von Data Science-Projekten ist folglich nicht trivial, bietet Unternehmen, die sich der Herausforderung stellen, allerdings große Chancen. ◄