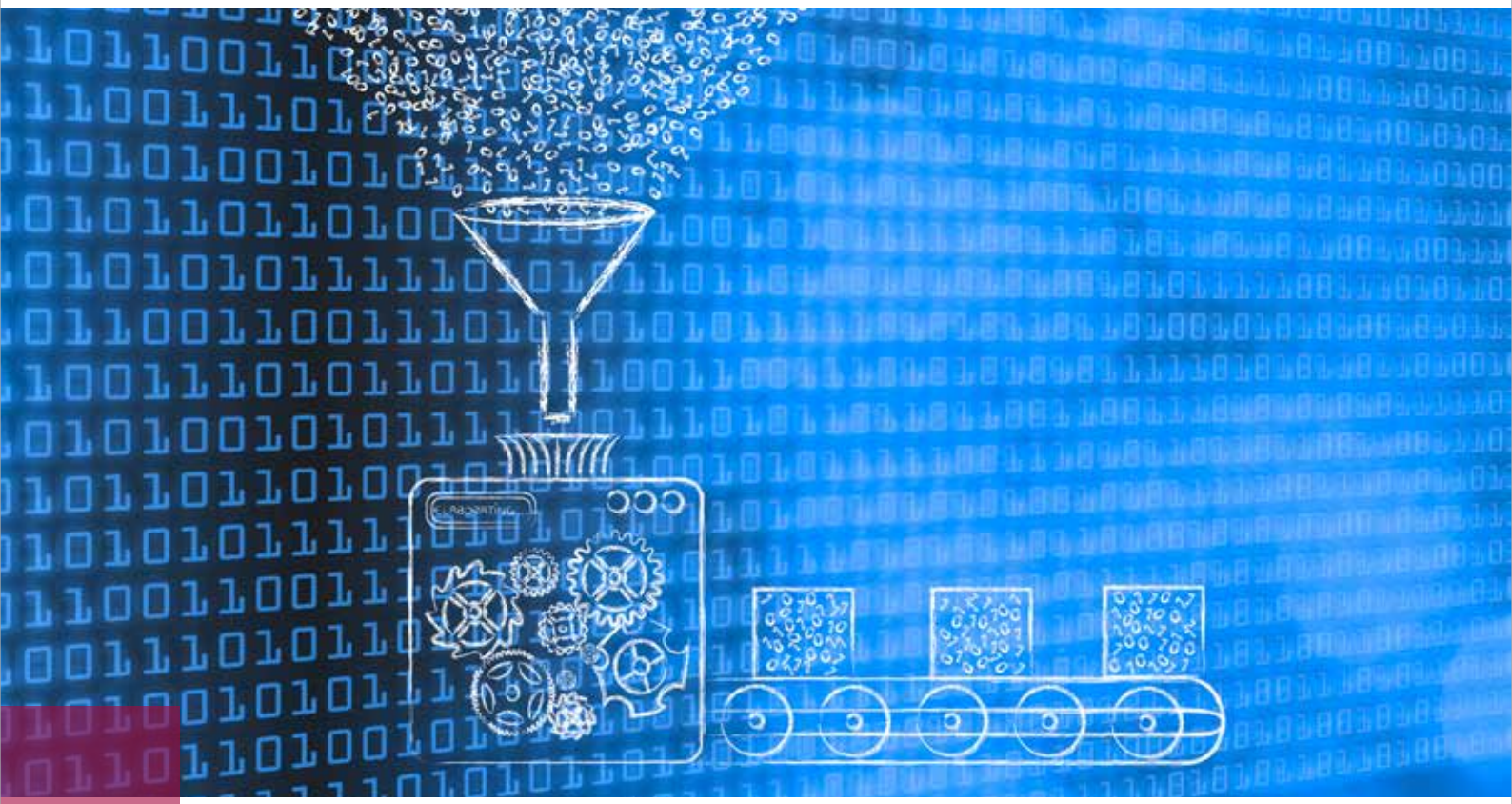
Lernende Systeme

DIE PLATTFORM FÜR KÜNSTLICHE INTELLIGENZ



Von Daten zu KI

Intelligentes Datenmanagement als Basis für Data Science und den Einsatz Lernender Systeme

GEFÖRDERT VOM

Bundesministerium für Bildung und Forschung

acatech
DEUTSCHE AKADEMIE DER TECHNIKWISSENSCHAFTEN

WHITEPAPER

Daniel Keim, Kai-Uwe Sattler
AG Technologische Wegbereiter und Data Science

Inhalt

Zusammenfassung

Egal ob Satellitenbilder als Datenquellen für Navigationssysteme oder Urlaubsfotos auf Social Media-Plattformen – täglich werden unvorstellbar große Mengen an neuen Daten generiert. Daten sind daher in unserer zunehmend digitalisierten Welt zu einem zentralen Rohstoff geworden. Ein umfassendes Datenmanagement sowie die Fähigkeit, die Daten überhaupt erst für die Analyse zugänglich zu machen, stellt daher eine wichtige Voraussetzung dar, um wertvolle Erkenntnisse in der Wissenschaft gewinnen und nutzenbringende Anwendungen für Wirtschaft und Gesellschaft generieren zu können. Der interdisziplinäre Forschungszweig Data Science, also das Management und die Analyse von Daten, gilt daher schon heute als eine der wichtigsten Schlüsseldisziplinen für Wissenschaft und Wirtschaft. Auch für die weitere Anwendung von Künstlicher Intelligenz (KI) und Lernenden Systemen stellt die Verfügbarkeit von Daten und die Datenverwaltung eine zentrale Voraussetzung dar.

Der Fokus bei Data Science liegt auf der Art und Weise, wie Daten verarbeitet, aufbereitet und analysiert werden. Durch wissenschaftlich fundierte Methoden, Prozesse, Algorithmen und Systeme können Erkenntnisse und Muster aus strukturierten und unstrukturierten Daten abgeleitet werden. Damit gelten Data Science-Methoden als Wegbereiter für wissenschaftliche Erkenntnisse in vielfältigen Forschungsfeldern, etwa in der Klimaforschung, Astronomie, Materialwissenschaft, Chemie oder Medizin. Sie machen zudem Anwendungen erst möglich, die unser Alltagsleben erleichtern, etwa die Nutzung von Navigations- oder Sprachassistenten. Darüber hinaus gelten sie als Voraussetzung, um die Potentiale von KI-Anwendungen ausschöpfen zu können, etwa in der Produktfertigung, der Logistik oder im Kundenmanagement.

Das Whitepaper, das von Expertinnen und Experten der Arbeitsgruppen Technologische Wegbereiter und Data Science der Plattform Lernende Systeme erarbeitet wurde, beleuchtet die Bedeutung, gesellschaftliche Relevanz und Nutzenpotentiale dieser Disziplin, benennt Beispiele für die Anwendung von Data Science-Methoden auf große Datenmengen (Kapitel 1) und beleuchtet auch Prozessketten von Data Science-Anwendungen (Kapitel 2). Auf Basis der Analyse von Data Science-Prozessen und Datenmanagementtechnologien werden zudem verschiedene Grundlagen dargestellt, die den Einsatz von Maschinellem Lernen und KI ermöglichen. Das Papier adressiert auch Herausforderungen für die weitere Entwicklung von Data Science, wozu etwa der aufwendige Prozess der Erschließung der Daten und der Sicherstellung der notwendigen Datenqualität gehört.

Darauf aufbauend werden im Anschluss wichtige Datenmanagementtechnologien für Data Science (Kapitel 3) erläutert. Neben der Datenverwaltung und -aufbereitung werden Datenbanken inzwischen nicht mehr nur für die Speicherung von Daten, sondern zunehmend auch für die Sicherung berechneter Modelle verwendet. Insgesamt werden Maschinelles Lernen und Datenbanken zunehmend integriert konzipiert. In Anknüpfung daran werden im Whitepaper Berufsfelder und wichtige Expertise-Felder für Data Scientists

erläutert (Kapitel 4), wozu etwa Kenntnisse im Datenmanagement oder im Bereich der Statistik und des Maschinellen Lernens zählen. Anschließend (Kapitel 5) werden Perspektiven und Ansätze aufgezeigt, um Daten für die Gesellschaft künftig noch effizienter und effektiver nutzbar zu machen und das Verständnis von Data Science-Prozessen und Datenmanagementtechnologien in unserer Gesellschaft zu fördern – einer Gesellschaft, in der die Erfassung, Verarbeitung und Analyse von Daten eine Grundlage für Wohlstand, Alltagserleichterungen und wissenschaftlichen Fortschritt darstellt. Als mögliche Handlungsoptionen gelten hier etwa die weitere Forschungsförderung sowie die Förderung im Bereich der Aus- und Weiterbildung, um so das Nutzenpotential der Schlüsseldisziplin Data Science ausschöpfen zu können (Kapitel 6).

1. Bedeutung und gesellschaftliche Relevanz

Tagtäglich werden unglaubliche Mengen an Daten produziert. In jeder einzelnen Minute eines Tages versenden Nutzerinnen und Nutzer hunderttausende Kurznachrichten, stellen Millionen von Suchanfragen und sehen sich Millionen von Videos an (Martin 2019). Intel beziffert die Datenmenge, die beim automatisierten Fahren generiert wird, auf 4.000 Gigabyte pro Tag und Fahrzeug (Intel 2016). Der Teilchenbeschleuniger der europäischen Organisation für Kernforschung (CERN) in Genf produzierte 88 Petabyte an Daten im Jahr 2018, was in etwa 22 Millionen Filmen in hoher Bildqualität entspricht (WLCG 2020). **Diese Daten geben ihren Mehrwert jedoch nicht ohne Weiteres preis – dazu müssen sie ausgewertet werden**. Für die Datenauswertung sind allerdings nicht nur Methoden der Statistik und des Maschinellen Lernens erforderlich, sondern ebenfalls ein umfassendes **Datenmanagement, um die Daten überhaupt erst für die Analyse zugänglich zu machen**. In den Diskussionen um Künstliche Intelligenz (KI) wird jedoch häufig wenig beachtet, dass Datenmanagement eine Voraussetzung darstellt, um KI erfolgreich einzusetzen. Dieses Whitepaper widmet sich dieser Beobachtung, zeigt die Bedeutung von Data Science-Prozessen sowie Datenmanagementtechnologien für KI auf und identifiziert wichtige Expertisefelder.

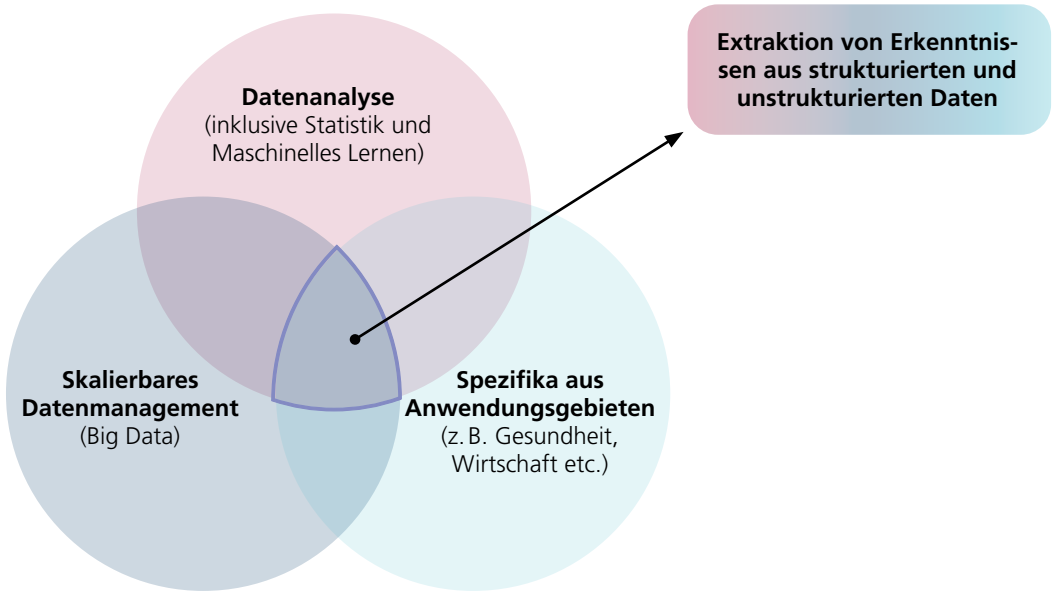
Data Science als Disziplin

Der interdisziplinäre Forschungszweig Data Science befasst sich mit dem Management und der Analyse von Daten. In den Naturwissenschaften hat dies zum Begriff des „vierten Paradigmas“ geführt. Dabei handelt es sich um einen Begriff, der auf die Verbreitung datengetriebener Forschung in den Naturwissenschaften verweist und Methoden für große Datenmengen hervorhebt – neben den empirischen Methoden, theoretischen Modellen und wissenschaftlichem Rechnen bzw. der Simulation (Hey, Tolle & Tansley 2009). Zwei Dimensionen können hinsichtlich solcher Methoden unterschieden werden. Erstens benötigen viele moderne Methoden der Datenanalyse enorme Datenmengen, um einen gewünschten Mehrwert zu erzielen. Dies ist gegenwärtig bei Methoden aus der Forschung zu Künstlicher Intelligenz der Fall, etwa beim Deep Learning. Hier kann beispielsweise eine automatische Klassifizierung von Bildern, Sprache oder Audiosignalen angestrebt werden. Zweitens werden Methoden für das Management von Daten benötigt, um die Daten für die Analyse und Auswertung zugänglich zu machen, das heißt Methoden für die Erfassung, Aufbereitung und Verarbeitung von großen Datenmengen. Solche Methoden stellen daher heutzutage in vielerlei Hinsicht eine Notwendigkeit dar, um das Potential von Daten ausschöpfen zu können, wie zum Beispiel ihr Vermögen, zur Realisierung des autonomen Fahrens oder zur Aufklärung von Betrugsfällen beizutragen (vgl. Anwendungsbeispiele in Tabelle 1, Seite 8). Es braucht dafür stets eine Kombination von Methoden des Datenmanagements und der Datenanalyse. Aus diesem Grund sind

diese Methoden technologische Wegbereiter für neue Entwicklungen in allen Forschungs- und Anwendungsgebieten, in denen große Datenmengen potentiell zur Verfügungen stehen. Dies gilt im besonderen Maße für die Forschung zu Künstlicher Intelligenz und Robotik, da hier häufig sowohl große Datenmengen benötigt werden (z. B. für Maschinelles Lernen) als auch große Datenmengen anfallen (z. B. durch Sensorik).

Methoden des Datenmanagements und der Datenanalyse sind jedoch nicht nur technologischer Wegbereiter für neue Entwicklungen in der Forschung und in Anwendungsdomänen. Sie sind auch maßgeblicher Motor für die Etablierung von Data Science als disziplinenübergreifende Wissenschaft. So liefern Informatik und Statistik die Methoden, die von Domänenexpertinnen und -experten aus der Betriebswirtschaft oder der Medizin in ihrem jeweiligen Bereich eingesetzt werden. Während Methoden des Datenmanagements sowohl ein Fundament für die Datenanalyse darstellen als auch für die Umsetzung von Projekten in den Anwendungsdomänen, können wiederum Methoden aus der Forschung zu Künstlicher Intelligenz, wie etwa Maschinelles Lernen, oder aus dem Bereich Statistik zur Verbesserung des Datenmanagements beitragen. Schließlich kann domänenspezifisches Wissen das Methodenwissen mit konkreten Herausforderungen in der Anwendung konfrontieren, wie zum Beispiel die Erfordernisse des Datenschutzes in der Medizin. Data Science vereint somit Methoden, Verfahren und Algorithmen aus dem skalierbaren Datenmanagement (Big Data) und der Datenanalyse (inkl. Statistik und Maschinelles Lernen) sowie Anwendungsspezifika bestimmter Domänen (z. B. Wirtschafts- und Gesundheitssektor) (Markl 2015) zur Extraktion von Erkenntnissen aus strukturierten und unstrukturierten Daten (vgl. Abbildung 1).

Abbildung 1: Data Science als disziplinübergreifende Wissenschaft



Nutzenpotential von Daten heben – Werkzeuge der Data Science als Voraussetzung

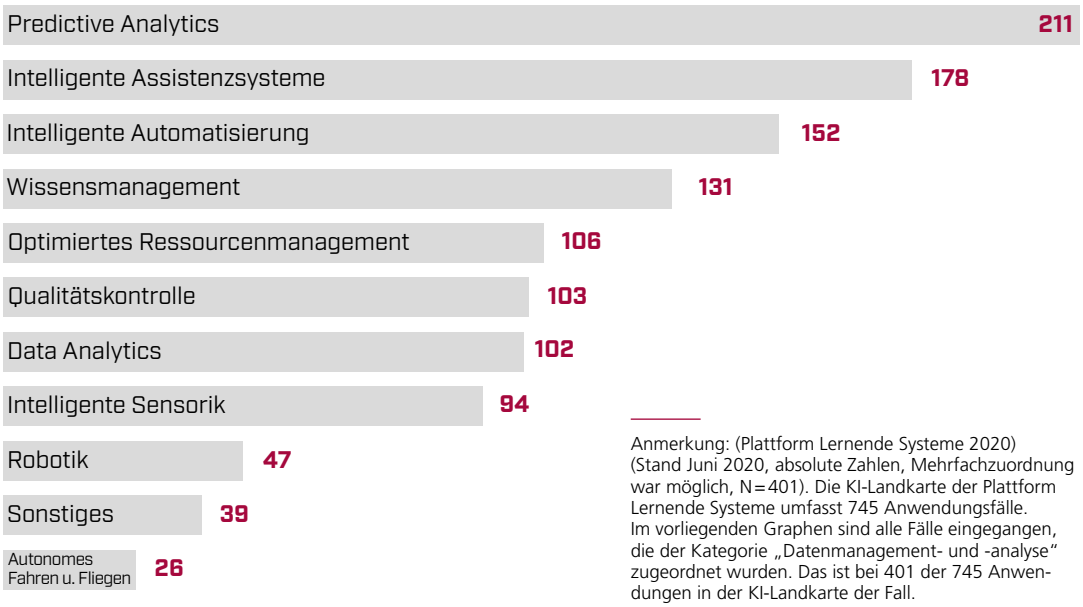
Die gegenwärtigen und künftigen Methoden, Verfahren und Algorithmen aus dem Werkzeugkasten der Data Science sind der Schlüssel, um das Potential der Daten ausschöpfen zu können und ihren Nutzen für die Wissenschaft, den Alltag und die Wirtschaft zu realisieren. So sind sie, erstens, ein Wegbereiter für wissenschaftliche Erkenntnisse in vielfältigen Forschungsfeldern, wie der Klimaforschung, der Astronomie, der Materialwissenschaft bzw. Chemie oder der Medizin (The Royal Society 2019). Die neuesten KI-Algorithmen helfen dabei, die Evolution von Galaxien zu erforschen, neue chemische Verbindungen zu entdecken oder wirkungsvolle Antibiotika zu identifizieren (Göpel 2020 Falk, 2019).
Sie machen, zweitens, Anwendungen erst möglich, die wir heute wie selbstverständlich nutzen und die unser Alltagsleben erleichtern, wie etwa die nutzerzentrierte Suche im Internet, die Navigation per Kartendienst auf dem Smartphone oder die Nutzung von Sprachassistenten, wie Siri und Alexa. Nicht zuletzt stellen Data Science-Werkzeuge, drittens, die Voraussetzung dar, um die vielfältigen Chancen, die Künstliche Intelligenz heute schon bietet, zu ergreifen, so etwa in der Produktfertigung, der Logistik oder im Kundenmanagement. Diese Chancen sind mit einem enormen Potential für künftiges wirtschaftliches Wachstum und Wohlstand verknüpft. So prognostiziert eine Studie für Deutschland ein Wachstum des Bruttoinlandsproduktes von 11,3 Prozent bis 2030 – dies entspräche ca. 430 Milliarden Euro (PWC 2018). Es wundert daher nicht, dass der Beruf des Data Scientists als einer der wichtigsten und vielversprechendsten Berufe gilt (DeNisco Rayome 2019; World Economic Forum 2018). Data Science-Kompetenzen werden demnach heute schon in vielen Berufsbildern gefordert.

Ebenso wie dem Alltagsnutzen sowie dem wissenschaftlichen und wirtschaftlichen Nutzen, der sich aus Daten ergibt, ist auch den Daten selbst ein hoher Wert beizumessen, genauso wie der Fähigkeit, diese zu erschließen, auszuwerten und zu analysieren. Die Datengrundlage kann hierbei sehr vielfältig sein: Informationen über Nutzerinnen und Nutzer und ihr Verhalten, Sensordaten zu Ereignissen der realen Welt (z. B. Erdbeobachtungen von Satellitenmissionen, Bewegungsdaten von Fahrzeugen) oder auch Trainings- und Modelldaten für Lernende Systeme. Dementsprechend sind auch die Anwendungsmöglichkeiten über viele Einsatzfelder und Branchen hinweg äußerst divers (vgl. für einen Überblick über Anwendungsbeispiele, Einsatzfelder und Branchen: Tabelle 1 und Abbildung 2 sowie Abbildung 3; S. 8 und 9). So wurden in der KI-Landkarte der Plattform Lernende Systeme beispielsweise hunderte von KI-Anwendungsfällen mit Bezug zu Datenmanagement und -analyse zusammengetragen. Diese KI-Anwendungen werden beispielsweise im Wissensmanagement, in intelligenten Assistenzsystemen oder in der prädiktiven Analyse eingesetzt (vgl. Abbildung 2) und in verschiedenen Branchen, etwa dem Gesundheitswesen, Energie, Umwelt, Mobilität und Logistik (Abbildung 3).

Tabelle 1: Beispiele für die Anwendung von Data Science-Methoden auf große Datenmengen

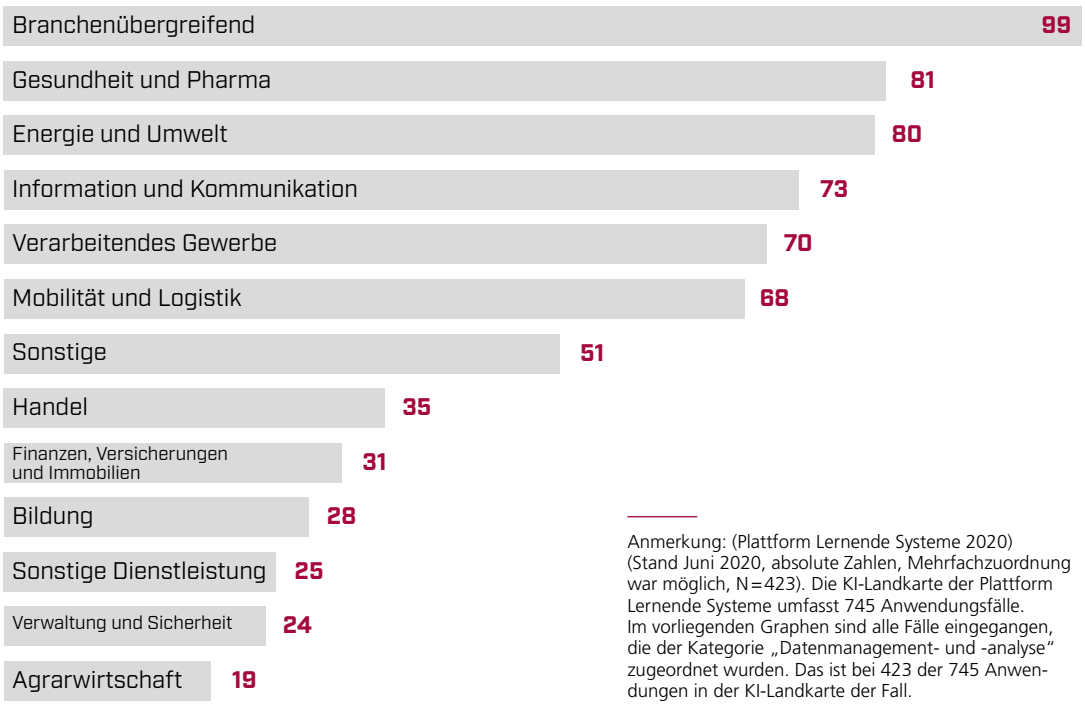
Einsatzgebiet	Beschreibung
Autonomes Fahren	Projekte im Bereich autonomes Fahren u. a. zur Klassifikation und Erkennung von Verkehrssituationen anhand der Sensordaten
Betrugsaufklärung	Betrugserkennung u.a. bei Banken und Kreditkartenunternehmen durch Analyse der Transaktionsdaten
Logistik	Entwicklung neuer Verfahren zum lebenslangen Lernen mit flexiblen Datenmodalitäten für die Produktions-, Transport- und Logistikautomation (vgl. Still AG)
Medizin	Systeme zur Unterstützung von Medizinern bei der Diagnose etwa durch Analyse von Röntgenbildern oder Hautbildern zur Krebserkennung
Physik	Forschung im Bereich der Hochenergiephysik wie das IceCube-Projekt zum Nachweis von Neutrino-Ereignissen oder Arbeiten am Teilchenbeschleuniger der Europäischen Organisation für Kernforschung (CERN) in Genf zur Erforschung der Natur der Elementarteilchen und ihrer Wechselwirkungen (Cid & Cid 2020)
Sprachassistenten und Übersetzungsdienste	Erkennung und Übersetzung natürlicher Sprache durch das Training mit großen Datenmengen, die Anwendungen wie Apple Siri, Amazon Alexa und Google Speech oder Übersetzungsdienste wie DeepL erst ermöglicht haben
Vorausschauende Wartung	Zustandsüberwachung und vorausschauende Wartung von Produktionsmitteln mit Hilfe von Sensordaten und maschinellem Lernen (vgl. relayr GmbH, Lufthansa, ABB AG)
Werbemarkt/nutzerzentrierte Erlebnisse und Angebote im Internet	Aktivitäten der großen Internetfirmen wie Google, Facebook oder Twitter etwa zur Modellierung und Vorhersage von Nutzerverhalten

Abbildung 2: KI-Anwendung mit Datenmanagement und -analyse Bezug nach Einsatzfeldern



Anmerkung: (Plattform Lernende Systeme 2020) (Stand Juni 2020, absolute Zahlen, Mehrfachzuordnung war möglich, N=401). Die KI-Landkarte der Plattform Lernende Systeme umfasst 745 Anwendungsfälle. Im vorliegenden Graphen sind alle Fälle eingegangen, die der Kategorie „Datenmanagement- und -analyse" zugeordnet wurden. Das ist bei 401 der 745 Anwendungen in der KI-Landkarte der Fall.

Abbildung 3: KI-Anwendungen mit Datenmanagement und -analyse Bezug nach Branche



Anmerkung: (Plattform Lernende Systeme 2020) (Stand Juni 2020, absolute Zahlen, Mehrfachzuordnung war möglich, N=423). Die KI-Landkarte der Plattform Lernende Systeme umfasst 745 Anwendungsfälle. Im vorliegenden Graphen sind alle Fälle eingegangen, die der Kategorie „Datenmanagement- und -analyse" zugeordnet wurden. Das ist bei 423 der 745 Anwendungen in der KI-Landkarte der Fall.

Der lange Weg von den Daten zur Anwendung – Datenmanagement und Qualitätssicherung

Die Anwendungsbeispiele (vgl. Tabelle 1, S. 8) bieten einen Einblick in die enorme Vielfalt, mit der Data Science-Methoden nutzenbringend eingesetzt werden können. Obwohl aus Daten schon relativ bald nach ihrer Erhebung gute und verwertbare Einsichten gewonnen werden können, ist es meist ein langer Weg, bis das volle Nutzenpotential der Daten ausgeschöpft wird, denn Daten müssen zunächst einmal erschlossen werden. Gerade dieses Erschließen von Daten und nicht zuletzt die Sicherstellung der notwendigen Datenqualität ist häufig ein aufwendiger Prozess. So wird für KI-Projekte der Aufwand für die Datenerfassung und -aufbereitung auf bis zu 80 Prozent geschätzt (Yaddow 2019).

Bei der Erschließung der Daten stehen Lernende Systeme und Datenmanagement oft in einem wechselseitigen Verhältnis. Einerseits sind Lernverfahren oft wichtige Werkzeuge im Prozess der Datenerfassung und -aufbereitung, beispielsweise wenn es darum geht, mit fehlenden Daten umzugehen, Daten zusammenzuführen (Record Linkage) oder Abhängigkeiten aufzudecken (Ilyas & Chu 2019). Andererseits ist die wichtigste Voraussetzung für Lernende Systeme und KI die Verfügbarkeit von Daten und somit auch die Datenverwaltung. So sind ohne geeignete Trainingsdaten viele Lernverfahren nicht einsetzbar. Gleichzeitig müssen diese Trainingsdaten effizient und effektiv verwaltet werden. Dies umfasst neben dem effizienten und sicheren Zugriff auf die eigentlichen Daten auch Metadaten, um beispielsweise Nachvollziehbarkeit gewährleisten zu können.

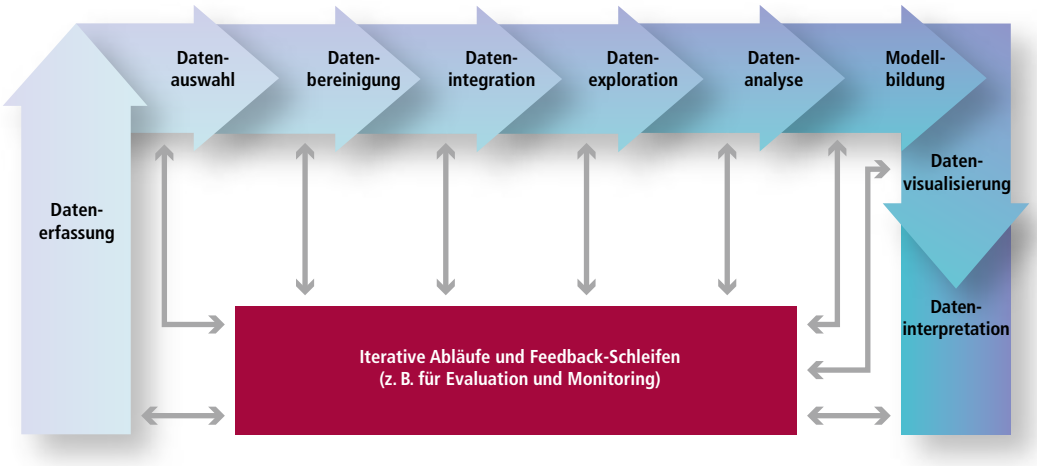
Weiterhin ist die Qualität der Eingangsdaten für die Korrektheit und Aussagekraft der Analyseergebnisse von großer Bedeutung (Rat für Informationsinfrastrukturen 2019). Datenqualität muss dabei in verschiedenen Dimensionen betrachtet werden: Neben Aspekten wie Vollständigkeit, Widerspruchsfreiheit, Konsistenz oder Aktualität spielen gerade für Machine Learning-Verfahren auch Fragen von Fairness und das Vermeiden von Verzerrungen (Bias) bereits bei der Datenauswahl eine wichtige Rolle (Getoor 2019), um etwa Diskriminierung zu vermeiden. Dieses wichtige Thema wird auch im Whitepaper „Künstliche Intelligenz und Diskriminierung“ der Plattform Lernende Systeme ausführlich behandelt (Beck et al. 2019).

Im Folgenden werden durch die Analyse von Data Science-Prozessen und Datenmanagementtechnologien die Grundlagen dargestellt, auf denen Maschinelles Lernen und KI aufsetzen kann. Darauf aufbauend werden im Anschluss die wichtigsten Expertisefelder für Data Scientists und Berufe mit Bezug zu Data Science im Hinblick auf das Datenmanagement definiert. Nach einem knappen Blick auf das Thema Data Management und Data Science in Deutschland werden schließlich Perspektiven aufgezeigt, um künftig Daten für die Gesellschaft noch effizienter und effektiver nutzbar machen zu können und allgemein das Verständnis von Data Science-Prozessen und Datenmanagementtechnologien in unserer Gesellschaft zu fördern – einer Gesellschaft, in der die Erfassung, Verarbeitung und Analyse von Daten eine Grundlage für Wohlstand, Alltagserleichterungen und wissenschaftlichen Fortschritt darstellt.

2. Data Science-Prozesse

Die Basis vieler Data Science-Anwendungen sind Prozessketten, welche die Schritte der Datenerfassung, Auswahl, Bereinigung, Integration, Exploration, Analyse und Modellbildung bis hin zur Visualisierung und Interpretation umfassen (siehe Abbildung 4). Diese Prozesse werden entweder explizit (prozedural oder deklarativ) spezifiziert und dann automatisiert ausgeführt oder eher implizit in interaktiver und explorativer Weise vollzogen. Häufig handelt es sich hierbei nicht um statische Abläufe, wie etwa Extraktions-Transformations-Lade (ETL)-Prozesse im Data Warehousing, sondern um interaktive Prozesse, die menschliche Interventionen und Entscheidungen („Human-in-the-Loop") erfordern. Teilweise bestehen diese aber auch aus iterativen Abläufen mit Feedback-Schleifen, um gegebenenfalls Daten, Methoden oder Parameter zu wechseln, neue Trainingsdaten zu beschaffen oder Modelle mit neuen Daten zu aktualisieren. Derartige Interventionen erfordern ein kontinuierliches Monitoring und eine Evaluation der Ergebnisse der einzelnen Schritte, etwa zur Qualität der Eingangsdaten oder der Modellgüte.

Abbildung 4: Data Science-Prozesse



Im Zuge derartiger Prozesse kommen eine Vielzahl von Methoden aus unterschiedlichen Bereichen zum Einsatz. Beispiele hierfür sind Signalverarbeitungsmethoden, etwa zum Filtern der Daten, Verfahren zur Datenintegration, Statistik-Methoden für die Charakterisierung der Daten, die Ableitung von Kennzahlen oder Features, Zeitreihenanalyse etc., Datenvisualisierungsmethoden zur visuellen Datenexploration, Data Mining- und Machine Learning-Methoden zum Bereinigen der Daten (z. B. Ersetzen fehlender Werte, Duplikaterkennung) sowie zur eigentlichen Modellbildung. Eine wichtige Rolle spielen in diesen Prozessen auch die Auswahl und Erfassung der Daten – beispielsweise als Trainingsdaten – sowie die Sicherstellung der Datenqualität. Neben Data Profiling und Data Cleaning kommt hierbei der Datenintegration und -anreicherung eine wichtige Rolle zu, um die oft heterogenen Ausgangsdaten erschließen und

fusionieren zu können. In diesem Kontext kommen auch semantische Technologien wie Wissensgraphen zum Einsatz, deren Fakten unter anderem zur Validierung, Bewertung und Annotation von Daten genutzt werden. So liefern Datensammlungen und Wissensbasen wie Wikidata, Freebase, DBpedia oder YAGO umfangreiche Fakten beispielsweise für die Datenaufbereitung.

Speziell mit Methoden und Prozessen zur Erfassung, Verwaltung, Speicherung, Aufbereitung, Anreicherung und Bereitstellung der Daten beschäftigt sich das Gebiet Data Engineering. Im Mittelpunkt stehen dabei Fragen der Bereitstellung von performanten und zuverlässigen Infrastrukturen für das Datenmanagement, die fundamental für die effiziente Unterstützung von Data Science-Prozessen sind, sowie Methoden für die Verwaltung und Aufbereitung der Daten und Modelle. Data Engineering wird daher im Folgenden als Oberbegriff für Datenmanagement, Datenintegration und Datenaufbereitung verwendet.

Darüber hinaus ist eine visuelle Exploration und Analyse der Daten und Modelle unabdingbar (Visual Analytics), um die Qualität der Daten und Modelle zu beurteilen, mit den Daten und Modellen effektiv interagieren zu können oder neue Trainingsdaten abzuleiten.

Um Data Science-Anwendungen nachvollziehen, ihre Qualität beurteilen und somit ihre Vertrauenswürdigkeit attestieren zu können, ist allerdings nicht nur ein tiefes Verständnis der einzelnen Komponenten eines Data Science-Prozesses notwendig, sondern auch ein Verständnis des Zusammenspiels der Komponenten und damit der Prozesskette als Ganzes.

3. Datenmanagementtechnologien für Data Science

Die Datenmanagement-Community inklusive der Industrie hat in den vergangenen 30 bis 40 Jahren eine Vielzahl von Methoden und Techniken entwickelt, die fundamental für Data Science – und damit auch die Künstliche Intelligenz – sind und weit über klassische relationale Datenbanksysteme hinausgehen. Beispiele hierfür sind:

- **Datenverwaltung**: Hierzu zählen Speichertechnologien und Datenstrukturen sowie Techniken zum physischen Design von Datenbanken wie Caching, Replikation, Indexierung, Datenkompression, Partitionierung für verteilte und parallele Datenhaltung sowie Vorberechnung und Materialisierung von häufig genutzten Daten. Dies betrifft auch die Verwaltung von Trainingsdaten und Modellen.
- **Datenaufbereitung**: Techniken zum Data Profiling und Data Cleaning mittels statistischer Verfahren und Methoden des Maschinellen Lernens, Techniken der Datenintegration über die Kombination und Verknüpfung von Daten bis hin zu heterogenen Systemen.
- **Performance und Skalierung**: Datenmanagementsysteme sind für hochperformante und skalierbare Verarbeitung großer Datenmengen optimiert. Dies wird zum einen durch Verteilung und Parallelisierung der Verarbeitung erreicht. Neben verteilten bzw. parallelen Datenbanksystemen zählen hierzu auch Plattformen für die verteilte Verarbeitung in Cluster-Umgebungen wie Apache Spark oder der erfolgreichen deutschen Entwicklung Flink sowie Dateninfrastrukturen für das Cloud Computing. Hierbei besteht auch eine enge Verbindung zum verteilten Machine Learning. Zum anderen umfasst dies den effizienten Einsatz moderner Hardware (z. B. in Form von Main-Memory-Datenbanken, die Nutzung von GPU- oder FPGA-basierten Beschleunigern sowie Multicore- und Cluster-Systemen), aber auch die Online-Verarbeitung von Datenströmen, wie sie zum Beispiel im Internet der Dinge, bei Industrie 4.0-Szenarien oder Social-Media-Plattformen anfallen. Zunehmend spielen auch spezielle Hardware-Architekturen wie Tensor-Recheneinheiten und neuromorphe Systeme eine wichtige Rolle.
- **Optimierung und Ausführung komplexer Prozesse**: Die Entwicklung und Ausführung komplexer Datenverarbeitungspipelines wird durch deklarative Anfrageverarbeitung inkl. Analytics-Methoden (z. B. Online Analytical Processing, OLAP) und Datenflussmodelle (wie in Spark oder Flink) unterstützt. Dies betrifft auch die Unterstützung kontinuierlichen Lernens, das heißt von iterativen Prozessen, in denen zusätzliche Trainingsdaten aus der Nutzung der KI-Anwendung berücksichtigt werden.
- **Gewährleistung von Wiederholbarkeit und Nachvollziehbarkeit**: Gerade für datengetriebene Lernprozesse ist es wichtig, die Modellbildung wiederhol- und nachvollziehbar zu gestalten, beispielsweise um fehlerhafte Vorhersagen, Änderungen der Trainingsdaten oder der Verarbeitungsprozesse nachvollziehen zu können. Hierfür bieten Datenmanagementtechniken wie Daten-Versionierung, Time-Travel-Anfragen oder Auditing Lösungswege an.

Datenbanken und Maschinelles Lernen werden im Sinne von „Machine Learning Systems“ (Ratner 2019) zunehmend als Teile eines Systems betrachtet. So werden Datenbanken nicht nur zur Speicherung der eigentlichen Daten, sondern auch zur Speicherung von berechneten Modellen und Programmcodes genutzt und dienen damit als Analysewerkzeug. Dies betrifft auch die Integration von Methoden des Maschinellen Lernens, entweder indem Lernverfahren in Datenbanksysteme integriert werden oder indem Datenmanagementtechniken (wie etwa Indexing) zur Beschleunigung von Machine Learning- und Analyseaufgaben genutzt werden. Deutlich wird dies unter anderem auch an Entwicklungen wie etwa bei der Software-Plattform für Big Data-Analysen Apache Spark, das inzwischen auch (wieder) eine Schnittstelle für die Datenbanksprache Structured Query Language (SQL) mit einem Anfrageoptimierer anbietet, also einer Anwendung, die nach dem effizientesten Weg sucht, um auf Daten zuzugreifen. Ein weiteres Beispiel ist die populäre Programmbibliothek Pandas-Framework für die Programmiersprache Python. Es unterstützt als zentrale Datenabstraktion, das heißt eine vereinfachte Repräsentation von Daten, Data Frames (d. h. Tabellen). Pandas realisiert auf Data Frames unter anderem SQL-ähnliche Operationen – auch wenn diese noch weit von der automatischen Anfrageoptimierung und Skalierung von SQL-Datenbanksystemen entfernt sind. Natürlich lassen sich die Anforderungen moderner Data Science-Anwendungen nicht mehr allein durch klassische SQL-Systeme erfüllen:

- Das relationale Datenmodell ist oft zu restriktiv, insbesondere für schwach strukturierte Daten, aber auch für komplexe Strukturen wie Graphen oder hierarchische Strukturen. Moderne SQL-Systeme bieten zwar Unterstützung für Datenaustauschformate wie JSON, erfordern jedoch spezielle Funktionen zur Verarbeitung.
- Nicht alle Operationen lassen sich komfortabel durch SQL-Anfragen oder benutzerdefinierte Funktionen (UDF) ausdrücken. Iterative Abläufe, komplexe Data Mining- und Lernverfahren können zwar SQL-Anfragen nutzen, erfordern aber wiederholten Datentransfer zwischen Datenbank und Analysewerkzeug.
- Der Import großer Datenmengen oder gar von potentiell unendlichen Datenströmen in ein Datenbanksystem, bevor die Daten überhaupt analysiert werden können, ist nicht immer möglich oder sinnvoll. Dies ist insbesondere dann der Fall, wenn Daten nur einmalig, etwa für eine Aufbereitung, eine Analyse bzw. ein Training benötigt werden oder wenn beispielsweise Sensordaten kontinuierlich eintreffen, aber als Rohdaten nur begrenzte Gültigkeit haben.

4. Notwendige Expertise

Data Scientists benötigen Expertise, die über die reine Analyse von Daten hinausgeht. Die Auswahl der Daten kann einen signifikanten Einfluss auf Analyse- und Lernergebnisse haben, teilweise sogar mehr als die Wahl des eigentlichen Algorithmus (Pereira, Norvig & Halevy 2009; Banko & Brill 2001). Daher besteht in vielen Data Science-Anwendungen die Herausforderung, nicht nur einzelne Komponenten eines Data Science-Prozesses umsetzen und nachvollziehen zu können, sondern den Gesamtprozess und das Methoden-Set sowie den effizienten und effektiven Umgang mit großen Datenmengen zu beherrschen. Hierfür sind Kenntnisse aus verschiedenen Bereichen notwendig:

- Fähigkeiten im Datenmanagement: Kenntnisse zur Datenmodellierung, -transformation und -integration sowie zu Methoden der effizienten Speicherung (Indexierung, Partitionierung) und Verarbeitung durch Datenbankoperationen.
- Kenntnisse aus den Bereichen Maschinelles Lernen und Künstliche Intelligenz im Sinne der Methodenkenntnisse und der Randbedingungen bzw. Anforderungen an Eingangsdaten.
- Kenntnisse aus dem Bereich Statistik: Sowohl bezüglich grundlegender Eigenschaften (wie Signifikanz und Güte von Ergebnissen) als auch konkreter Methoden für Sammlung, Erschließung und Qualitätsbewertung von Daten und Analyseergebnissen.
- Kenntnisse aus der Visualisierung zur interaktiven Exploration der Daten und Modelle, um die Daten, Modelle und ihre Eigenschaften (Qualität, Bias etc.) verstehen zu können und mit den Methoden des Maschinellen Lernens bzw. der Künstlichen Intelligenz interagieren zu können.
- Kenntnisse aus den Bereichen Ethik und Recht zum verantwortungsvollen Umgang mit Daten, von der Erfassung der Daten über die Beurteilung der Qualität beziehungsweise Eignung, um etwa Voreingenommenheit („Bias“) zu vermeiden, bis hin zum Verständnis für die Relevanz und Rolle von Persönlichkeitsrechten und weiteren relevanten rechtlichen Vorgaben (wie z. B. der Datenschutzgrundverordnung, Allgemeines Gleichbehandlungsgesetz, Fair Decision Making etc.). Allerdings kann auch jenseits einer optimalen Datenqualität ein Bias in den Daten vorliegen, wenn diese schlicht bestehende Diskriminierung in der Gesellschaft abbilden, so dass Data Scientists ein Bewusstsein für die Möglichkeit solcher Verzerrungen in den Daten haben sollten.

Für die Entwicklung vertrauenswürdiger Data Science-Anwendungen bzw. die Einschätzung ihrer Vertrauenswürdigkeit ist nicht nur ein tiefes Verständnis für die einzelnen Komponenten des zugrundeliegenden Data Science-Prozesses notwendig, sondern ebenso für den gesamten Prozess.

Derartige Fähigkeiten sollten nicht mehr nur auf Informatikabsolventinnen und Informatikabsolventen oder „Datenwissenschaftlerinnen und Datenwissenschaftler“ beschränkt sein, sondern betreffen alle Bereiche von den Ingenieurwissenschaften über Naturwissenschaften und Medizin bis hin zu Geistes- und Sozialwissenschaften. Die Gesellschaft für

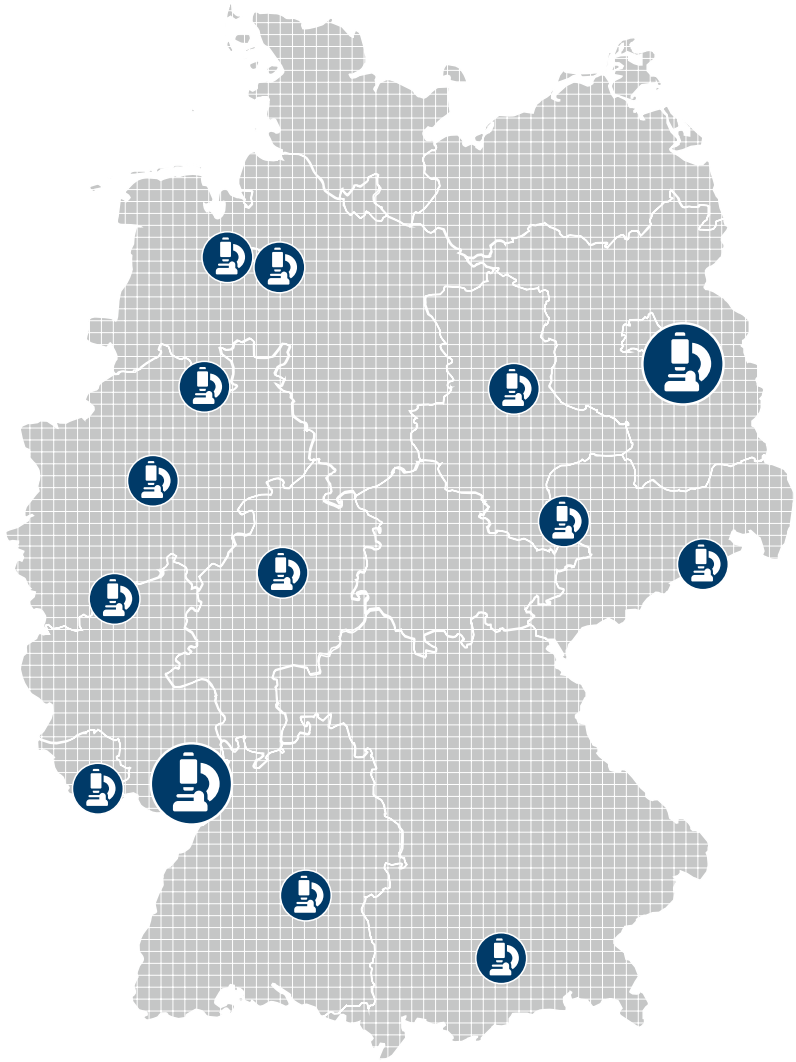
Informatik hat dies etwa in ihrem Data Literacy-Papier (Gesellschaft für Informatik 2018) formuliert. Die inhaltliche Ausrichtung und Prioritäten von Studiengängen und Weiterbildungen für Data Science hat die Gesellschaft für Informatik unter Mitarbeit der Plattform Lernende Systeme dargelegt und spezielle Optionen zur Verbindung von Data Science und Domänenwissenschaften erläutert (Gesellschaft für Informatik, 2019). Umgekehrt ist es, je nach Neigung, für künftige Data Scientists sinnvoll, sich Kenntnisse mit Blick auf Anwendungsdomänen anzueignen. Hier können beispielsweise Kenntnisse im Produktmanagement genauso wie Einblicke in Spezifika des Gesundheitswesens oder der Industrieproduktion wichtig sein.

Um also das Potential der Data Science für die Forschung ausschöpfen zu können, ist es notwendig, die genannten Fähigkeiten auch über die Data Science hinaus in anderen Disziplinen zu stärken, so dass auch dort vermehrt neue Erkenntnisse mit Hilfe von Data Science-Werkzeugen generiert werden können. Über die Forschung hinaus ist die breite Vermittlung der genannten Data Science-Fähigkeiten jedoch eine Grundlage, um Talente auszubilden, die schließlich dazu beitragen können, KI-Wissen in die Unternehmen zu tragen und dort den Wissenstransfer in die Anwendung zu unterstützen.

5. Data Engineering und Data Science in Deutschland

Deutschland verfügt über eine starke Industrie- und Forschungslandschaft in den Bereichen Data Engineering und Data Science: Neben großen Unternehmen wie SAP SE oder der Software AG bieten eine ganze Reihe kleinerer, innovativer Unternehmen wie Exasol oder RapidMiner Lösungen und Dienste für Datenmanagement und -analyse an. Institutionen wie das Deutsche Forschungszentrum für Künstliche Intelligenz (DFKI) und die weiteren vom Bund eingerichteten KI-Kompetenzzentren in Berlin, Dresden/Leipzig, München, Tübingen sowie Rhein-Ruhr in Bonn/Dortmund bündeln in Verbindung mit verschiedenen lokalen Initiativen in den Ländern die Forschungskompetenzen.

Abbildung 5: Landkarte der KI-Kompetenzzentren



Quelle: www.ki-landkarte.de

Vier der Kompetenzzentren weisen einen besonderen Forschungsbezug zum Datenmanagement auf und machen daher exemplarisch einige der Schwerpunkte in diesem Forschungsfeld in Deutschland deutlich. So forscht das Kompetenzzentrum in Berlin (BIFOLD) an der Schnittstelle zwischen Big Data, Datenmanagement und Maschinellem Lernen. Es arbeitet beispielsweise daran, die Effizienz von Datenmanagement und -verarbeitung zu steigern und untersucht Architekturen für skalierbare Verarbeitung großer Datenmengen. Das Zentrum in Leipzig und Dresden (ScaDS.AI) fokussiert unter anderem auf Datenqualität und -integration sowie Big Data-Architekturen und Big Data Lifecycle Management und Workflows, im Zentrum in Bonn und Dortmund (ML2R) steht neben der Forschung zu Lernenden Systemen generell das Maschinelle Lernen unter Ressourcenbeschränkungen im Fokus, wie z. B. auf Smartphones oder direkt auf Sensoren, und im Münchner Kompetenzzentrum (MCML) engagieren sich viele Wissenschaftlerinnen und Wissenschaftler als Dozentinnen und Dozenten im Zertifikatskurs „Data Science" und forschen an räumlichen, zeitlichen und raum-zeitlichen Daten, die man vor allem in der Medizin oder in der Bild- und Videoanalyse findet.

Die deutsche KI-Forschung ist gegenwärtig noch in einer guten Position. Mit Blick auf die Treiber der aktuellen KI-Innovationswelle, das Maschinelle Lernen und das Datenmanagement, hat Deutschland hervorragende Wissenschaftlerinnen und Wissenschaftler vorzuweisen (Markl 2019). Der Wettbewerbsdruck hat jedoch zugenommen, da sowohl Regierungen als auch große Unternehmen enorm in diese Bereiche investieren und einige Länder mutiger und konsequenter voranschreiten (ebd.).

Im Bereich Datenmanagement ist die deutsche Forschung systemorientiert. Sie hat in den letzten Jahren zu mehreren, sehr erfolgreichen Ausgründungen geführt. Beispiele sind unter anderem das Münchner Start-up Hyper-DB, das 2016 von Tableau, einem Hersteller für Visualisierungs- und Reportingsoftware, erworben wurde, sowie das Berliner Start-up Data Artisans mit Apache Flink, das 2019 vom chinesischen Konzern Alibaba übernommen wurde. Hinzu kommen mehrere koordinierte Programme der Deutschen Forschungsgemeinschaft (DFG) in Form von Sonderforschungsbereichen und Schwerpunktprogrammen, die sich den Themen Data Engineering und Data Science widmen, und zielgerichtete Forschungsprogramme, wie etwa die Richtlinie zur Förderung von Projekten zum Thema „Erzeugung von synthetischen Daten für Künstliche Intelligenz" des Bundesministeriums für Bildung und Forschung (Bundesministerium für Bildung und Forschung 2020).

6. Perspektiven und Ansätze

Vor dem Hintergrund des großen Nutzenpotentials von Daten für die Wirtschaft, die Wissenschaft, aber auch für das Alltagsleben, ist Data Science eine Schlüsseldisziplin. Sie bietet die Möglichkeit, diesen Nutzen abzuschöpfen. Hierbei sind jedoch nicht nur die Statistik und Maschinelles Lernen von großer Bedeutung, sondern auch Data Science-Prozesse als Gesamtaufgabe sowie Datenmanagementtechnologien. Datenmanagement stellt, wie dieses Whitepaper veranschaulicht hat, das zentrale Fundament sowohl für Data Science als wissenschaftliche Disziplin als auch für Lernende Systeme dar. Welche Handlungsfelder ergeben sich aus diesem Befund? Was bedeutet dies für Aus- und Weiterbildung sowie für die Forschung im Zeitalter der Digitalisierung, in dem Big Data-Technologien und Maschinelles Lernen als Schlüsselgebiete angesehen werden?

- **Data Literacy**: Wie bereits von der Gesellschaft für Informatik (GI) gefordert, muss Data Literacy-Kompetenzen ein breiterer Raum in Schule und Studium eingeräumt werden. Dies gilt weit über den Informatikunterricht in den Schulen oder Studiengänge mit Informatikbezug hinaus und betrifft neben Fähigkeiten zur Erschließung, Sammlung und Qualitätsbewertung von Daten auch grundlegende Kompetenzen zum Einsatz von Werkzeugen zur Datenverarbeitung und -analyse sowie zur Visualisierung und kritischen Interpretation der Ergebnisse.
- **Data Science-/Data Engineering-Ausbildung**: In Studiengängen im Bereich Data Science, aber auch in Informatikstudiengängen sollte mehr Wert auf Data Engineering-Themen gelegt werden. Dies geht beispielsweise über die Inhalte klassischer Datenbank-Vorlesungen hinaus und betrifft neben Data Literacy-Kompetenzen etwa Aspekte von Datenintegration und -qualität, Datenexploration und -visualisierung, aber auch alternative Datenmodelle und Verarbeitungsparadigmen. Für Studiengänge und Weiterbildungsangebote an Hochschulen im Bereich Data Science hat der Arbeitskreis „Data Science/Data Literacy“ unter Mitarbeit der Plattform Lernende Systeme Empfehlungen zur inhaltlichen Ausgestaltung erarbeitet (Gesellschaft für Informatik 2019).
- **Infrastruktur und Daten**: Die eindrucksvollen Beispiele und Anwendungen der KI-Labore der Internet-Konzerne dürfen nicht zu einer unreflektierten Übernahme von scheinbar erfolgreichen Modellen verführen. Neben dem Verständnis für die Lernmethoden und ihre Grenzen sowie der Kenntnis der genutzten Trainingsdaten erfordert dies aber auch, überhaupt die Möglichkeit zu haben, vergleichbare aufwendige Lernverfahren durchführen zu können. Hierfür werden geeignete Infrastrukturen mit ausreichender Speicher- bzw. Rechenkapazität und Datensammlungen (z. B. für Trainingsdaten) benötigt.
  Gleichzeitig müssen auch „Small Data“-Methoden berücksichtigt werden. Anwendungsbereiche, die gerade in Deutschland wichtig sind, wie etwa Medizintechnik oder Maschinenbau, zeichnen sich speziell im Umfeld von KMU durch deutlich kleinere Datenmengen aus, die oft nur schwach integriert sind. Infrastrukturen, Forschungs- und Ausbildungsprogramme für den KI-Bereich sollten dieser Situation Rechnung tragen. Insgesamt spielen Datenmanagementtechnologien somit für Unternehmungen wie europäische Datenräume, die beispielsweise die Europäische Kommission in ihrer

Datenstrategie vorantreiben möchte (Europäische Kommission 2020), und Cloud-basierte Dateninfrastrukturen wie GAIA-X oder ganz allgemein Datenökosysteme eine zentrale Rolle. Da Recheninfrastrukturen auch Gegenstand der Forschung sind, ist auch ein Ausbau solcher Infrastrukturen notwendig, um zugrundeliegende Prozesse des Datenmanagements selbst gestalten und Experimente durchführen zu können.

- **Forschung**: In künftigen Forschungsprogrammen, etwa der KI-Strategie der Bundesregierung oder in Forschungsprojekten von Unternehmen sowie Forschungs- und Entwicklungseinrichtungen, sollte der Bedeutung des Data Engineering als Gesamtprozess in Verbindung mit Maschinellen Lernverfahren noch mehr Rechnung getragen werden. So können Stärken sowohl im Bereich der Erfassung und Auswahl der Daten, ihrer Exploration und Visualisierung als auch im Einsatz von KI- und Datenmanagement-Methoden in Data Science-Prozessen weiter ausgebaut werden. In Anbetracht der wesentlichen Rolle von Datenmanagementtechnologien für die Datenräume und -ökosysteme der Zukunft ist eine noch stärkere Förderung einschlägiger Forschungsfelder zielführend.
- **Anwendungsorientierung**: Data Engineering spielt eine bedeutende Rolle, um Lernende Systeme in die Anwendungen zu bringen. Der Gesamtprozess aus Data Engineering und Lernenden Systemen sollte daher auch bei der Entwicklung und Umsetzung von KI-Anwendungen in Unternehmen noch mehr Berücksichtigung finden, indem besonderes Augenmerk auf die Erfassung, Vorverarbeitung und sichere Speicherung sowohl von Trainingsdaten als auch von Prozessdaten gelegt wird. Speziell in technischen bzw. industriellen Anwendungsfeldern wie Automatisierung und Predictive Analytics, aber auch im Medizinbereich bilden qualitativ hochwertige Daten eine wesentliche Basis für die erfolgreiche Anwendung Lernender Systeme, gleichzeitig sind derartige Daten aber auch hochsensitiv. Infrastrukturen und Datenökosysteme müssen daher anwendungs- bzw. branchenspezifische Lösungen in geeigneter Weise berücksichtigen.

Über dieses Whitepaper

Die Autoren des Whitepapers sind Mitglieder der Arbeitsgruppe Technologische Wegbereiter und Data Science der Plattform Lernende Systeme. Als eine von insgesamt sieben Arbeitsgruppen befasst sie sich mit den technologischen Grundlagen und Enablern von Künstlicher Intelligenz. Dabei geht es beispielsweise um die Anforderungen an die Forschung, die Ausbildung von KI-Fachleuten oder den Transfer von Forschungsergebnissen in erfolgreiche Anwendungen. Die produktive Diskussion in der Arbeitsgruppe trug maßgeblich zur Entwicklung des Whitepapers bei.

Autoren
Prof. Dr. Daniel A. Keim, Universität Konstanz
Prof. Dr. Kai-Uwe Sattler, Technische Universität Ilmenau

Die Arbeitsgruppe wird geleitet von
Prof. Dr. Katharina Morik, Technische Universität Dortmund
Prof. Dr. Volker Markl, Technische Universität Berlin

Redaktion
Maximilian Hösl, Geschäftsstelle der Plattform Lernende Systeme
Dr. Ursula Ohliger, Geschäftsstelle der Plattform Lernende Systeme

Über die Plattform Lernende Systeme

Lernende Systeme im Sinne der Gesellschaft zu gestalten – mit diesem Anspruch wurde die Plattform Lernende Systeme im Jahr 2017 vom Bundesministerium für Bildung und Forschung (BMBF) auf Anregung des Fachforums Autonome Systeme des Hightech-Forums und acatech – Deutsche Akademie der Technikwissenschaften initiiert. Die Plattform bündelt die vorhandene Expertise im Bereich Künstliche Intelligenz und unterstützt den weiteren Weg Deutschlands zu einem international führenden Technologieanbieter. Die rund 200 Mitglieder der Plattform sind in Arbeitsgruppen und einem Lenkungskreis organisiert. Sie zeigen den persönlichen, gesellschaftlichen und wirtschaftlichen Nutzen von Lernenden Systemen auf und benennen Herausforderungen und Gestaltungsoptionen.

Literatur

Banko, M., & Brill, E. (2001): Scaling to Very Very Large Corpora for Natural Language Disambiguation. In Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics, Université Des Sciences Sociales, Toulouse, France S. 26–33.

Beck, S., Grunwald, A., Jacob, K., Matzner, T. (2019): Künstliche Intelligenz und Diskriminierung – Whitepaper aus der Plattform Lernende Systeme: https://www.plattform-lernende-systeme.de/files/Downloads/Publikationen/AG3_Whitepaper_250619.pdf (abgerufen am 25.09.2020).

Bundesministerium für Bildung und Forschung (2020): Richtlinie zur Förderung von Projekten zum Thema „Erzeugung von synthetischen Daten für Künstliche Intelligenz". https://www.bmbf.de/foerderungen/bekanntmachung-3068.html (abgerufen am 25.09.2020).

Cid, X., & Cid, R. (2020): LHC Data Analysis – Taking a closer look at LHC. https://www.lhc-closer.es/taking_a_closer_look_at_lhc/0.lhc_data_analysis (abgerufen am 25.09.2020).

DeNisco Rayome, A. (2019): Why data scientist is the most promising job of 2019. https://www.techrepublic.com/article/why-data-scientist-is-the-most-promising-job-of-2019 (abgerufen am 25.09.2020).

Europäische Kommission. (2020): Eine europäische Datenstrategie. https://ec.europa.eu/info/sites/info/files/communication-european-strategy-data-19feb2020_de.pdf (abgerufen am 25.09.2020).

Falk, D. (2019): How Artificial Intelligence Is Changing Science. https://www.quantamagazine.org/how-artificial-intelligence-is-changing-science-20190311/ (abgerufen am 25.09.2020).

Gesellschaft für Informatik (2018): Data Literacy und Data Science Education: Digitale Kompetenzen in der Hochschulausbildung. https://gi.de/themen/beitrag/data-literacy-und-data-science-education-digitale-kompetenzen-in-der-hochschulausbildung/ (abgerufen am 25.09.2020).

Gesellschaft für Informatik (2019): Data Science: Lehr- und Ausbildungsinhalte [unter Mitwirkung der Plattform Lernende Systeme]. https://www.plattform-lernende-systeme.de/files/Downloads/Publikationen/GI_Arbeitspapier_Data-Science_2019-12_01.pdf (abgerufen am 25.09.2020).

Getoor, L. (2019): Responsible Data Science. IEEE Big Data 2019: https://users.soe.ucsc.edu/~getoor/Talks/IEEE-Big-Data-Keynote-2019.pdf (abgerufen am 25.09.2020).

Göpel, G. (2020): Forscher entdecken dank KI neues Antibiotikum. https://background.tagesspiegel.de/gesundheit/forscher-entdecken-dank-ki-neues-antibiotikum (abgerufen am 25.09.2020).

Hey, T., Tolle, K. M., & Tansley, S. (2009): The Fourth Paradigm: Data-intensive Scientific Discovery. Microsoft Research.

Ilyas, I. F., & Chu, X. (2019): Data Cleaning. ACM Press.

Intel (2016): Data is the New Oil in the Future of Automated Driving. https://newsroom.intel.com/editorials/krzanich-the-future-of-automated-driving (abgerufen am 25.09.2020).

Markl, V. (2015): Gesprengte Ketten – Smart Data, deklarative Datenanalyse, Apache Flink. Informatik Spektrum, 38(1), S. 10–15.

Martin, N. (2019): Forbes. https://www.forbes.com/sites/nicolemartin1/2019/08/07/how-much-data-is-collected-every-minute-of-the-day/#3941e8453d66 (abgerufen am 25.09.2020).

Pereira, F., Norvig, P., & Halevy, A. (2009): The Unreasonable Effectiveness of Data. IEEE Intelligent Systems, 24 (March/April), S. 8–12.

Plattform Lernende Systeme (2020): https://www.plattform-lernende-systeme.de/ki-landkarte.html (abgerufen am 25.09.2020).

PWC (2018): Auswirkungen der Nutzung von künstlicher Intelligenz in Deutschland. München: PricewaterhouseCoopers GmbH.

Rat für Informationsinfrastrukturen (2019): Herausforderung Datenqualität. http://www.rfii.de/?p=4043 (abgerufen am 25.09.2020).

Ratner, A. et al. (2019): MLSys: The New Frontier of Machine Learning Systems. https://arxiv.org/abs/1904.03257 (abgerufen am 25.09.2020).

The Royal Society (2019): The AI revolution in scientific research. https://royalsociety.org/-/media/policy/projects/ai-and-society/AI-revolution-in-science.pdf?la=en-GB&hash=5240F21B56364A00053538A0BC29FF5F (abgerufen am 25.09.2020).

WLCG (2020): Worldwide LHC Computing Grid. https://wlcg-public.web.cern.ch/about (abgerufen am 25.09.2020).

World Economic Forum (2018): The Future of Jobs Report 2018. http://www3.weforum.org/docs/WEF_Future_of_Jobs_2018.pdf (abgerufen am 25.09.2020).

Yaddow, W. (2019): AI and BI Projects Are Bogged Down With Data Preparation Tasks. https://tdwi.org/Articles/2019/08/16/DIQ-ALL-AI-and-BI-Data-Preparation-Tasks.aspx (abgerufen am 25.09.2020).

Glossar

Apache Spark – Software-Plattform für Big Data-Analysen in Cluster Computing-Systemen (vgl. unten).

Benutzerdefinierte Funktionen (engl.: user-defined function, UDF) – Funktionen, die vom Anwender selbst erstellt werden. „Verschiedene Programmierumgebungen und Datenbankmanagementsysteme erlauben die Definition und Nutzung von User Defined Functions. Die Funktionen müssen der Syntax der zugrundeliegenden Programmierumgebung entsprechen. Häufig werden UDFs beispielsweise in SQL-Datenbankumgebungen genutzt."[1]

Big Data – Datenmengen, die sich auszeichnen durch ihr Volumen (Volume), die Vielfalt der Datentypen und Quellen (Variety), die Geschwindigkeit, mit der sie anfallen (Velocity), sowie die Unsicherheit bezüglich der Qualität der Daten (Veracity). Oft handelt es sich dabei um größtenteils unstrukturierte Daten, die etwa von sozialen Netzwerken oder mobilen Geräten stammen (Internet of Things). Ein weiterer Aspekt von Big Data umfasst die Lösungen und Systeme, die dabei helfen, mit diesen Datenmengen umzugehen, um darin beispielsweise neue Muster und Zusammenhänge zu erkennen[2].

Big Data-Architekturen – Architekturen für große Datenmengen, die dafür ausgelegt sind, die Aufnahme, Verarbeitung und Analyse von Daten zu bewältigen, die für traditionelle Datenbanksysteme zu groß oder komplex sind.

Caching – Zwischenspeicherung einer Datenteilmenge zur Überwindung einer Zugriffslücke.

Cloud Computing – „Bereitstellung von Computingressourcen (z.B. Server, Speicher, Datenbanken, Netzwerkkomponenten, Software, Analyse- und intelligente Funktionen) über das Internet, um schnellere Innovationen, flexible Ressourcen und Skaleneffekte zu bieten."[3]

Cluster-Umgebungen/-Systeme – Verbund von vernetzten Rechnern, die nach außen hin als ein einziger Rechner erscheinen[4].

Data Cleaning (deutsch: Datenbereinigung) – Verfahren zum Entfernen oder Korrigieren von Datenfehlern wie Dopplungen, Formatierungsfehlern oder fehlerhaften, unvollständigen Datensätzen in Datenbanken.

1 BigData-Insider (2020): Definition. Was ist eine User Defined Function? https://www.bigdata-insider.de/was-ist-eine-user-defined-function-a-919802/ (abgerufen am 13.05.2020).
2 Plattform Lernende Systeme (2020): Glossar. https://www.plattform-lernende-systeme.de/glossar.html (abgerufen am 13.05.2020).
3 Microsoft: Was ist Cloud Computing? https://azure.microsoft.com/de-de/overview/what-is-cloud-computing/ (abgerufen am 13.05.2020).
4 IT-Administrator Magazin (2014): Server/Client. https://www.it-administrator.de/themen/server_client/grundlagen/172792.html (abgerufen am 13.05.2020).

Data Literacy – „(…) Fähigkeit, planvoll mit Daten umzugehen und sie im jeweiligen Kontext bewusst einsetzen und hinterfragen zu können."[5]

Data Lifecycle Management (DLCM) – „kontinuierliche Bewertung aller erfassten Daten und der daraus abgeleiteten Zwischenergebnisse, um jeweils die optimale Verwaltung der Daten zu gewährleisten."[6]

Data Profiling – „(…) überwiegend automatisierte Prozesse, mit denen sich die Qualität von Daten im Hinblick auf Struktur, Eindeutigkeit, Konsistenz und Logik analysieren und bewerten lässt."[7]

Data Mining – „(…) systematische Anwendung computergestützter Methoden, um in vorhandenen Datenbeständen Muster, Trends oder Zusammenhänge zu finden."[9]

Data Warehousing – „eine für Analysezwecke optimierte zentrale Datenbank, die Daten aus mehreren, in der Regel heterogenen Quellen zusammenführt."[8]

Datenexploration – iterativer Prozess „mit weitgehend automatisierten Verfahren zur Analyse mehrdimensionaler Daten und Bestimmung der in ihnen enthaltenen neuen verwertbaren Informationen mittels intelligenter Methoden aus Statistik, Visualisierung, maschinellem Lernen, Knowledge Discovery und Datenbanktechnologien."[10]

Datenmanagement – „(…) Sammlung von Maßnahmen, Verfahren und Konzepten. Ziel ist die Bereitstellung von Daten für eine optimale Unterstützung der verschiedenen Prozesse in den Unternehmen. Zum Datenmanagement gehören Maßnahmen zur Sicherstellung der Datenqualität, -konsistenz und -sicherheit sowie das Data Lifecycle Management."[11]

DBpedia – Gemeinschaftsanstrengung, um strukturierte Informationen aus Wikipedia zu extrahieren und diese Informationen im Web verfügbar zu machen. DBpedia ermöglicht es, anspruchsvolle Anfragen an Wikipedia zu stellen und andere Datensätze im Web mit Wikipedia-Daten zu verknüpfen.[12]

Deep Learning – Methode des Maschinellen Lernens in künstlichen neuronalen Netzen. Diese umfassen mehrere Schichten – typischerweise eine Eingabe- und Ausgabeschicht sowie mehr als eine „versteckte" dazwischenliegende Schicht. Die einzelnen Schichten bestehen aus einer Vielzahl künstlicher Neuronen, die miteinander verbunden sind und

5 Gesellschaft für Informatik e.V. (2018): Studie: Ansätze zur Vermittlung von Data-Literacy-Kompetenzen. https://gi.de/themen/beitrag/studie-ansaetze-zur-vermittlung-von-data-literacy-kompetenzen (abgerufen am 13.05.2020).

6 Center For Scalable Data Analytics And Artificial Intelligence (2020): https://www.scads.de/de/projekt/methodenwissenschaften/119-data-life-cycle-management-und-workflows (abgerufen am 29.07.2020).

7 BigData Insider (2018): Definition. Was ist Data Profiling? https://www.bigdata-insider.de/was-ist-data-profiling-a-691538/ (abgerufen am 13.05.2020).

8 Rahm, Erhard (2015): Data Warehouse. https://dbs.uni-leipzig.de/file/dw-kap1.pdf (abgerufen am 29.05.2020)

9 BigData-Insider (2016): Definition. Was ist Data Mining? https://www.bigdata-insider.de/was-ist-data-mining-a-593421/ (abgerufen am 13.05.2020).

10 Spektrum.de (2001): Datenexploration. https://www.spektrum.de/lexikon/kartographie-geomatik/datenexploration/844 (abgerufen am 13.05.2020).

11 Storage Insider (2019): Definition. Was ist Data Management/Datenmanagement? https://www.storage-insider.de/was-ist-data-management-datenmanagement-a-850258/ (abgerufen am 13.05.2020).

12 Agile Knowledge Engineering and Semantic Web: DBpedia. Querying Wikipedia like a Semantic Database. http://aksw.org/Projects/DBpedia.html (abgerufen am 13.05.2020).

auf Eingaben von Neuronen aus der jeweils vorherigen Schicht reagieren. In der ersten Schicht wird etwa ein Muster erkannt, in der zweiten Schicht ein Muster von Mustern und so weiter. Je komplexer das Netz (gemessen an der Anzahl der Schichten von Neuronen, der Verbindungen zwischen Neuronen sowie der Neuronen pro Schicht), desto höher ist der mögliche Abstraktionsgrad – und desto komplexere Sachverhalte können verarbeitet werden. Angewendet wird Deep Learning bei der Bild-, Sprach- und Objekterkennung sowie dem verstärkenden Lernen.[13]

Extraktions-Transformation-Lade (ETL)-Prozess – „(…) mehrere Einzelschritte, durch die sich Daten aus verschiedenen Datenquellen per Extrahieren und Aufbereiten in ein Data Warehouse integrieren lassen. Der Prozess kommt häufig zur Verarbeitung großer Datenmengen im Big Data- und Business-Intelligence-Umfeld zum Einsatz."[14]

Flink – „(…) Framework der Apache Software Foundation für das sogenannte Stream Processing. Es ermöglicht die kontinuierliche Verarbeitung von Datenströmen mit geringer Verzögerung. Das Framework ist einsetzbar für Big Data-Anwendungen und Echtzeitanalysen."[15]

FPGA (engl.: Field Programmable Gate Array) – Programmierbarer digitaler integrierter Schaltkreis, in den („vor Ort") eine logische Schaltung geladen werden kann.

GAIA-X –Projekt zum Aufbau einer leistungs- und wettbewerbsfähigen, sicheren und vertrauenswürdigen Dateninfrastruktur für Europa, das von Vertretern der deutschen Bundesregierung, Wirtschaft und Wissenschaft getragen wird.[16]

GPU (engl.: Graphics Processing Unit; deutsch: Grafikprozessor) – Spezialprozessor, der auf die Berechnung von Grafiken und Bilddaten spezialisiert und optimiert ist. Im Umfeld von Datenanalysen und KI kommen üblicherweise General-Purpose GPUs zum Einsatz, die aufgrund ihres hohen Parallelisierungsgrades Berechnungen wie etwa Lernverfahren deutlich beschleunigen können.

Indexierung – Aufbau eines Datenbankindex. Ein Datenbankindex ist eine zusätzliche Datenstruktur, die den Zugriff auf die eigentlichen Daten beschleunigen soll. Ein Eintrag in einem Datenbankindex besteht typischerweise aus einem (ein- oder mehrdimensionalen) Schlüssel zur Suche und einem Verweis auf den eigentlichen Datensatz.

Industrie 4.0 – Transformation „klassischer" Industrien durch das Internet der Dinge, Daten und Dienste.

13 Plattform Lernende Systeme (2020): Deep Learning. https://www.plattform-lernende-systeme.de/glossar.html. (abgerufen am 13.05.2020).

14 BigData Insider (2018): Definition. Was ist ETL (Extract, Transform, Load)? https://www.bigdata-insider.de/was-ist-etl-extract-transform-load-a-776549/ (abgerufen am 13.05.2020).

15 BigData Insider (2019): Definition. Was ist Apache Flink? https://www.bigdata-insider.de/was-ist-apache-flink-a-812389/ (abgerufen am 13.05.2020).

16 Bundesministerium für Wirtschaft und Energie (Hrsg.), Bundesministerium für Bildung und Forschung (2019): Das Projekt GAIA-X. https://www.bmwi.de/Redaktion/DE/Publikationen/Digitale-Welt/das-projekt-gaia-x.pdf?__blob=publicationFile&v=16 (abgerufen am 29.05.2020).

Internet der Dinge – „(…) Vernetzung von Gegenständen mit dem Internet, damit diese Gegenstände selbstständig über das Internet kommunizieren und so verschiedene Aufgaben für den Besitzer erledigen können. Der Anwendungsbereich erstreckt sich dabei von einer allg. Informationsversorgung über automatische Bestellungen bis hin zu Warn- und Notfallfunktionen."[17]

JSON (engl.: JavaScript Object Notation) – ein lesbares Datenaustauschformat.

Main-Memory-Datenbanken (auch In-Memory-Datenbanken genannt) – „(…) Datenbankmanagementsystem, das seine Daten nicht auf herkömmlichen Festplattenspeichern ablegt, sondern direkt den Arbeitsspeicher (RAM) hierfür nutzt. Dadurch lassen sich wesentlich höhere Zugriffsgeschwindigkeiten realisieren."[18]

Maschinelles Lernen – Methode der Künstlichen Intelligenz (KI). Sie zielt darauf, dass Maschinen ohne explizite Programmierung eines konkreten Lösungswegs automatisiert sinnvolle Ergebnisse liefern. Spezielle Algorithmen lernen aus den vorliegenden Beispieldaten Modelle, die dann auch auf neue, zuvor noch nicht gesehene Daten angewendet werden können. Dabei werden drei Lernstile unterschieden: überwachtes Lernen, unüberwachtes Lernen und verstärkendes Lernen. Maschinelles Lernen mit großen neuronalen Netzen wird als Deep Learning bezeichnet.[19]

Metadaten – Strukturierte Daten, die Informationen über Merkmale anderer Daten enthalten.[20]

Neuromorphe Systeme – „Neuromorphe Hardware basiert auf spezialisierten Rechnerarchitekturen, die die Struktur (Morphologie) Neuronaler Netze (NN) von Grund auf widerspiegeln: Dedizierte Verarbeitungseinheiten bilden direkt in der Hardware die Funktionsweise von Neuronen nach, zwischen denen ein physisches Verbindungsnetz (Bus-System) für den schnellen Austausch von Informationen sorgt."[21]

OLAP (Online Analytical Processing) – Ansatz zur schnellen Beantwortung mehrdimensionaler analytischer Anfragen in der Datenverarbeitung.

Pandas-Framework – Programmbibliothek für die Programmiersprache Python, die Hilfsmittel für die Verwaltung von Daten und deren Analyse anbietet.

Predictive Analytics – Methoden der Datenanalyse, mit denen auf zukünftige Ereignisse und Trends geschlossen werden kann. Grundlage für diese Analyseform sind im Allgemeinen historische Daten, auf deren Basis mathematische Vorhersagemodelle berechnet bzw. gelernt werden. Solche Modelle werden wiederum auf aktuelle Daten angewandt, um auf künftige Trends und Entwicklungen zu schließen.

17 Springer Gabler Wirtschaftslexikon (2018): Internet der Dinge. https://wirtschaftslexikon.gabler.de/definition/internet-der-dinge-53187/version-276282 (abgerufen am 13.05.2020).

18 BigData-Insider (2017): Definition. Was ist eine In-Memory-Datenbank? https://www.bigdata-insider.de/was-ist-eine-in-memory-datenbank-a-655470/ (abgerufen am 13.05.2020).

19 Plattform Lernende Systeme (2020): Glossar. https://www.plattform-lernende-systeme.de/glossar.html (abgerufen am 13.05.2020).

21 Böker, Elisabeth (2020): https://www.forschungsdaten.info/themen/beschreiben-und-dokumentieren/metadaten-und-metadatenstandards/ (abgerufen am 29.05.2020).

Record Linkage – Methode, um Datensatzeinträge in einem Datensatz zu finden, die sich auf dieselbe Entität in verschiedenen Datenquellen beziehen.[22]

SQL (engl.: Structured Query Language; deutsch: Strukturierte Abfrage-Sprache) – standardisierte Datenbanksprache zur Erstellung von Datenbankstrukturen in relationalen Datenbanken sowie zum Bearbeiten und Abfragen der darauf basierenden Datenbestände.[23]

Tensor-Recheneinheiten – Spezieller Chip, der von Google zur Optimierung von Maschinellem Lernen und KI entwickelt wurde. Mit TPU kann das Training von neuronalen Netzwerken wesentlich beschleunigt werden. Die Chips werden etwa für die Google-Suche oder für Übersetzungen mit Google Translate angewendet.

Verzerrungen (Bias) – Verzerrungseffekte: „Die Psychologie versteht darunter Einstellungen oder Stereotypen, welche die Wahrnehmung unserer Umwelt, Entscheidungen und Handlungen positiv oder negativ beeinflussen. Diese Beeinflussung kann unbewusst (impliziter Bias) oder bewusst (expliziter Bias) geschehen. In der Statistik wird ein Bias als Fehler im Rahmen der Datenerhebung und -verarbeitung (z. B. Fehler in der Stichprobenauswahl) oder die bewusste oder unbewusste Beeinflussung von Probandinnen und Probanden verstanden."[24]

Visuelle Analyse (Visual Analytics) – interdisziplinärer Ansatz, der Methoden aus den Gebieten Visualisierung und Datenanalyse verbindet, um Erkenntnisse aus großen und komplexen Datenmengen zu gewinnen. Der Ansatz kombiniert die Stärken der automatischen Datenanalyse mit den Fähigkeiten des Menschen, Muster oder Trends schnell visuell erfassen zu können.

Wikidata – freie und offene Wissensbasis, die sowohl von Menschen als auch von Maschinen gelesen und bearbeitet werden kann. Wikidata fungiert als zentraler Speicher für die strukturierten Daten ihrer Wikimedia-Schwesterprojekte, darunter Wikipedia, Wikivoyage, Wiktionary, Wikisource und andere.[25]

YAGO (Yet Another Great Ontology) – Open-Source-Wissensdatenbank, die am Max-Planck-Institut für Informatik in Saarbrücken entwickelt wurde.

22 Christen, Peter (2012): Data Matching. Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection. Springer-Verlag. Berlin.

23 Datenbanken-verstehen.de: SQL Einführung – Was ist SQL? https://www.datenbanken-verstehen.de/sql-tutorial/sql-einfuehrung/ (abgerufen am 13.05.2020).

24 Susanne Beck et al. (2019): Künstliche Intelligenz und Diskriminierung – Whitepaper aus der Plattform Lernende Systeme. https://www.plattform-lernende-systeme.de/publikationen-details/kuenstliche-intelligenz-und-diskriminierung-herausforderungen-und-loesungsansaetze.html?file=files/Downloads/Publikationen/AG3_Whitepaper_250619.pdf (abgerufen am 13.05.2020).

25 Wikidata: Welcome to Wikidata. https://www.wikidata.org/wiki/Wikidata:Main_Page (abgerufen am 13.05.2020).

Impressum

Herausgeber
Lernende Systeme –
Die Plattform für Künstliche Intelligenz
Geschäftsstelle | c/o acatech
Karolinenplatz 4 | 80333 München
www.plattform-lernende-systeme.de

Gestaltung und Produktion
PRpetuum GmbH, München

Stand
Oktober 2020

Bildnachweis
faithie/Adobe Stock/Titel

Bei Fragen oder Anmerkungen zu dieser Publikation kontaktieren Sie bitte Johannes Winter (Leiter der Geschäftsstelle):
kontakt@plattform-lernende-systeme.de

Folgen Sie uns auf Twitter: @LernendeSysteme

Empfohlene Zitierweise
Daniel Keim, Kai-Uwe Sattler: Von Daten zu KI – Intelligentes Datenmanagement als Basis für Data Science und den Einsatz Lernender Systeme. Whitepaper aus der Plattform Lernende Systeme, München 2020.