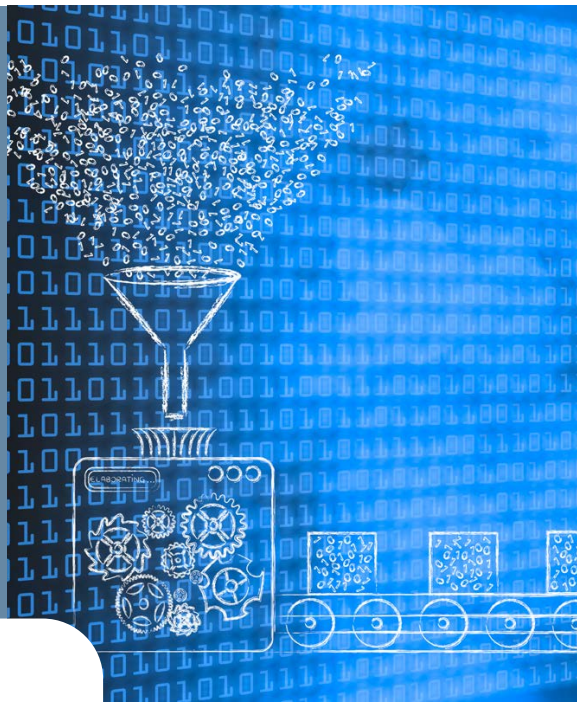
Lernende Systeme
DIE PLATTFORM FÜR KÜNSTLICHE INTELLIGENZ

Von Daten zu KI

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



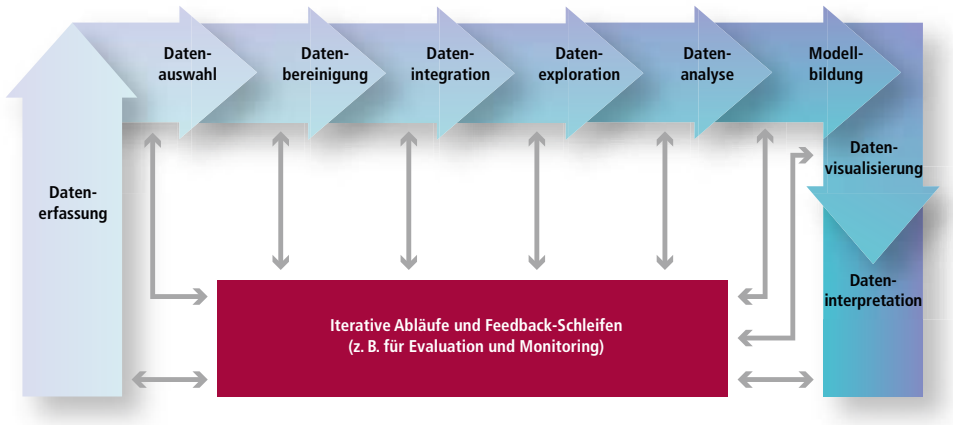
Kurzfassung

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. Gerade für die Anwendung von Lernenden Systemen stellt die Verfügbarkeit von Daten und das Datenmanagement 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 gewonnen 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, wie beispielsweise 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.

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 der Daten umfassen. 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.

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 oder 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)
- Methoden 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 kommen 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 zu 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.

Zu den zentralen Methoden und Techniken, die fundamental für Data Science sind, – und damit zugleich auch für den Einsatz von Künstlicher Intelligenz – gehören:

- **Datenverwaltung:** Hierzu zählen Speichertechnologien, Datenstrukturen, 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[1] oder der erfolgreichen deutschen Entwicklung Flink[2] und 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.

1 Apache Spark – Software-Plattform für Big Data-Analysen in Cluster Computing-Systemen.
2 Flink – Framework der Apache Software Foundation für das sogenannte Stream Processing. Es ermöglicht die kontinuierliche Verarbeitung von Datenströmen mit geringer Verzögerung.

- **Optimierung und Ausführung komplexer Prozesse:** Die Entwicklung und Ausführung komplexer Datenverarbeitungspipelines werden durch deklarative Anfrageverarbeitung inkl. Analytics-Methoden (z. B. Online Analytical Processing, OLAP[3]) 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.

Es gibt verschiedene Ansätze, 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. Dazu gehören unter anderem:

- **Data Literacy:** Wie bereits von der Gesellschaft für Informatik (GI) gefordert, muss den 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 generell 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 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.

3 OLAP - Online Analytical Processing: darunter versteht man einen Ansatz zur schnellen Beantwortung mehrdimensionaler analytischer Anfragen in der Datenverarbeitung.

Gleichzeitig müssen auch Small-Data-Methoden berücksichtigt werden. Anwendungsbereiche, die gerade in Deutschland wichtig sind, wie 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 daher 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.

Impressum
Herausgeber: Lernende Systeme – Die Plattform für Künstliche Intelligenz | Geschäftsstelle | c/o acatech | Karolinenplatz 4 | D-80333 München | kontakt@plattform-lernende-systeme.de | www.plattform-lernende-systeme.de |
Folgen Sie uns auf Twitter: @LernendeSysteme | Stand: Oktober 2020 | Bildnachweis: faithie/Adobe Stock/Titel

Diese Kurzfassung entstand auf Grundlage des Whitepapers *Von Daten zu KI – Intelligentes Datenmanagement als Basis für Data Science und den Einsatz Lernender Systeme*, München, 2020. Es wurde erstellt von der Arbeitsgruppe Technologische Wegbereiter und Data Science der Plattform Lernende Systeme. Die Originalfassung der Publikation ist online verfügbar unter: https://www.plattform-lernende-systeme.de/publikationen.html

GEFÖRDERT VOM

Bundesministerium für Bildung und Forschung

acatech
DEUTSCHE AKADEMIE DER TECHNIKWISSENSCHAFTEN