

The background features several abstract, overlapping geometric shapes in shades of blue, red, and dark blue, creating a modern, data-driven aesthetic.

Data Science

Daten verstehen im Zeitalter von Machine Learning, Künstlicher Intelligenz & Co.

Die richtigen Entscheidungen treffen durch gekonnte Analyse und Auswertung von Daten

Die Digitalisierung bringt neue Herausforderungen und neue Berufsbilder hervor. Auf dieser Seite erfahren Sie Genaueres über den unternehmerischen Umgang mit Big Data aus der Sicht von Data Science.

Wir zeigen auf, welche Teilbereiche die Datenwissenschaft hat und welche Kompetenzen benötigt werden, um in diesem Berufsfeld erfolgreich zu sein.



Inhaltsverzeichnis

Data Science: im Zentrum der vierten industriellen Revolution	4
Data Sourcing – der integrale Bestandteil	6
Data Cleansing – Ordnung und Struktur	7
Explorative Datenanalyse (EDA)	8
Data Mining: Wissen entdecken	9
Künstliche Intelligenz (KI/AI) in der Datenwissenschaft.	12
Coding die wichtigsten Programmiersprachen der Datenwissenschaft	18
Berufsprofil: Data Scientist/Data Analyst	20

Data Science

Im Zentrum der vierten industriellen Revolution

Big Data. Im Zuge der digitalen Transformation sehen wir uns heute mit einer neuen Art Ressource konfrontiert – den Daten.

Unternehmen, die sich der Datenökonomie verwehren, können nicht mehr unbesorgt in die Zukunft blicken. Denn dank moderner Technologien, computergestützter Anwendungen, der Cloud-Anwendungen usw. sind heute **große Mengen anfallender Daten für alle Unternehmen verfügbar**, vom Kleinunternehmen bis zum internationalen Konzern. Zumindest sind diese Daten theoretisch verfügbar.

Denn in diesem Kontext **entstand** auch die Notwendigkeit, die nun erhaltenen riesigen **Datenmengen auszuwerten**. Big Data allein erlaubt nämlich noch keinerlei Einsichten.

Und damit man **Rückschlüsse ziehen** kann, müssen Spezialistinnen und Spezialisten die Daten auf Zusammenhänge hin untersuchen, sie analysieren und interpretieren sowie entsprechend aufbereitet zur weiteren Nutzung zur Verfügung stellen.

Diese Interpretation bzw. Visualisierung hilft dem Management oder der Unternehmensführung, begründete Entscheidungen zu treffen – oder aber weitere Fragen zu stellen. Nur **so wird Big Data greifbar, begreifbar und nutzbar: mit Data Science**.

Was bedeutet also Data Science?

Data Science ist der englische Begriff für die Datenwissenschaft, also eine **Wissenschaft, die Daten sammelt, studiert und auswertet**, um daraus Erkenntnisse und mehr Wissen zu gewinnen.

Dadurch ist es Unternehmen möglich, **faktenbasierte Entscheidungen zu treffen** und oft wettbewerbsentscheidende Optimierungen anzustreben, beispielsweise Betriebsabläufe zu optimieren, Kunden besser anzusprechen, Logistiklösungen zu vereinfachen und einiges mehr.

Data Science ist eine angewandte Wissenschaft, die interdisziplinär betrieben wird: **Statistik, Informatik, Mathematik** und ihr Teilgebiet **Stochastik** sowie nicht zuletzt das Wissen um die jeweilige Branche spielen eine immens wichtige Rolle.



Data Sourcing

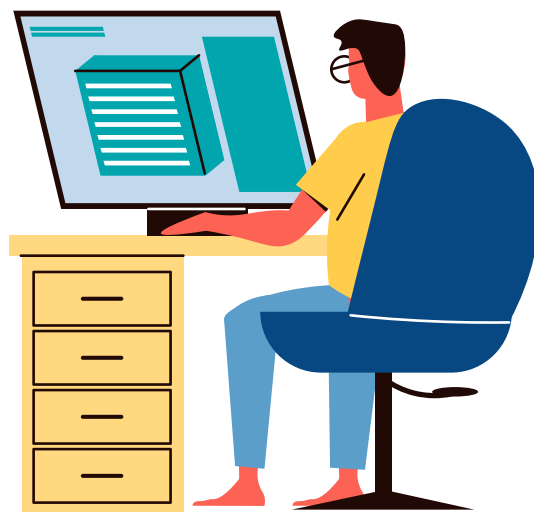
Der integrale Bestandteil

Ohne **Data Sourcing** (Deutsch: Beschaffung von Daten) ist die Datenwissenschaft nicht möglich.

Bei diesem Prozess **extrahieren** die Unternehmen **Daten aus verschiedenen so genannten primären und sekundären Quellen** und integrieren diese in unternehmenseigene Dateninfrastrukturen. So ist es möglich, die Daten auszuwerten und für die betrieblichen Arbeitsabläufe zu nutzen.

Primäre Datenquellen sind vom Unternehmen selbst generierte Daten: Ein klassisches Beispiel ist die Kundenumfrage.

Sekundäre Datenquellen stammen von Drittanbietern und können auf unterschiedlichen Wegen abgerufen werden – hier kommen allerdings unter Umständen Disziplinen wie Data Governance und natürlich Datenschutz-Bestimmungen ins Spiel. Die sekundären Datenquellen können intern sein (zum Beispiel Kundendaten im CRM-Programm) oder extern (vom Datenanbieter).



Data Cleansing

Ordnung und Struktur

Wer bis jetzt aufmerksam gelesen hat und schon einmal einen CRM-Auszug in den Händen hielt – oder eine Kundenumfrage auswerten durfte – wird sich spätestens jetzt fragen, **ob alle gewonnenen Daten ungefiltert aufgenommen und verarbeitet werden**. Die kurze Antwort ist: **nein**.

Data Cleansing, auch als Data Cleaning oder Data Scrubbing bekannt, ist der durch Tools gesteuerte Prozess, **bei dem Inkonsistenzen dokumentiert** (z. B. Ausreißer-Werte) **oder behoben werden**. Fehler, die man zu beheben sucht, können unterschiedliche strukturelle Fehler, doppelte Datensätze, irrelevante Daten u. Ä, beinhalten.



Das Ziel von Data Cleansing ist Data Integrity oder die Datenkonsistenz. Denn nur solche bereinigten Daten lassen sich sauber einer Data Analyse zuführen.

Explorative Datenanalyse (EDA)

Bevor es an die eigentliche Analyse geht, kommt die Explorative Datenanalyse (EDA) ins Spiel. Diese **Methode der Datenerkennung** wurde vom US-amerikanischen Mathematiker John Tukey in den 1970er Jahren entwickelt.

EDA kann dabei helfen, die Daten **mit visuellen Mitteln verständlich** zu machen. So können die Datenprofis noch vor den ersten Fragestellungen **interessante Muster oder optimale Auswertungsmethoden erkennen**.

Data Mining

Wissen Entdecken

Unter **Data Mining** versteht man den eigentlichen, sehr weit gefassten und umfangreichen computergestützten **Prozess der Datenanalyse** – also der **Mustererkennung in den Datenbeständen**.

Wichtig: Data Mining ist nicht mit Data Sourcing zu verwechseln, auch wenn es ebenfalls der Datengewinnung dient – **Sourcing** bezieht sich dabei auf die **Gewinnung von Rohdaten**, während **Mining** die **Daten aus der gesammelten, konsistenten Datenbank beschafft**. Man spricht beim Data Mining auch von Knowledge Discovery in Databases.

Beim Data Mining kommen Methoden und Expertise aus der IT, der Statistik und der Mathematik zusammen.



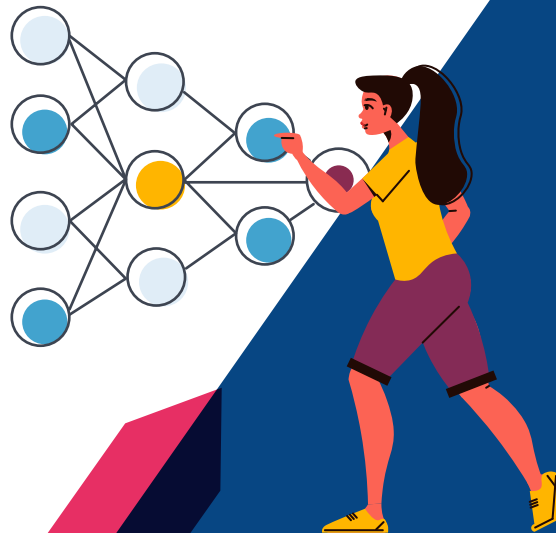
Data Mining

Methoden und -Algorithmen

Es werden allgemein vier Methoden unterschieden, mit denen Erkenntnisse für eine Vielzahl von Anwendungen gefördert werden – von Medizin und Forschung, Produktion und Logistik bis hin zu Handel, Bank- und Finanzwesen sowie dem Versicherungswesen:

Klassifikation

Anwendungsbeispiel:
Vorhersagen zum
Kundeninteresse

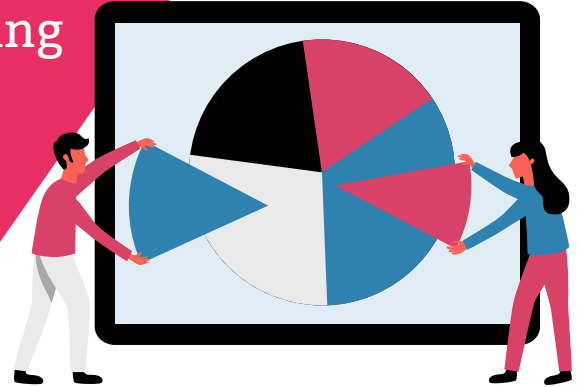


Prognose

Anwendungsbeispiel:
Umsatzprognose

Gruppierung (Segmentierung und Clustering)

Anwendungsbeispiel:
Segmentierung der
Datenbank der Newsletter-
Abonnenten



Abhängigkeitsentdeckung (Assoziation und Sequenz)

Anwendungsbeispiel:
„Kunden, die X kauften, kauften auch Y.“

Wichtiger Teilbereich: Predictive (Data) Analytics

Als Teildisziplin von Data Mining fokussiert sich **Predictive Analytics** darauf, **Vorhersagen über künftige Ereignisse zu treffen**. Das Ziel ist, Datenmuster zu erkennen, die unternehmerische Risiken zu senken und unternehmerische Chancen dadurch steigern zu können.

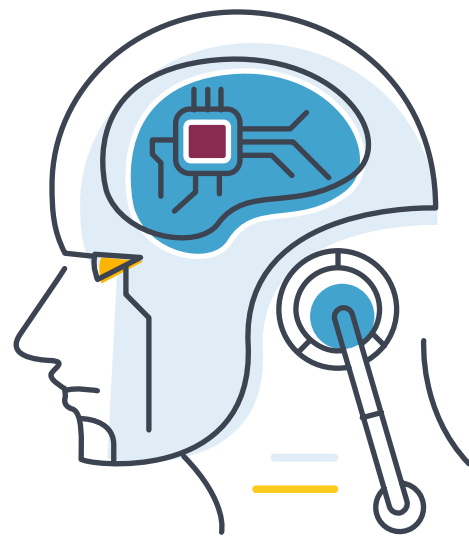


Künstliche Intelligenz (KI/AI)

In der Datenwissenschaft

Der Begriff **künstliche Intelligenz** (KI) oder auch **Artificial Intelligence** (AI) beschreibt die **Fähigkeit von Maschinen/ Computern, selbstständige Entscheidungen menschenähnlich zu treffen**.

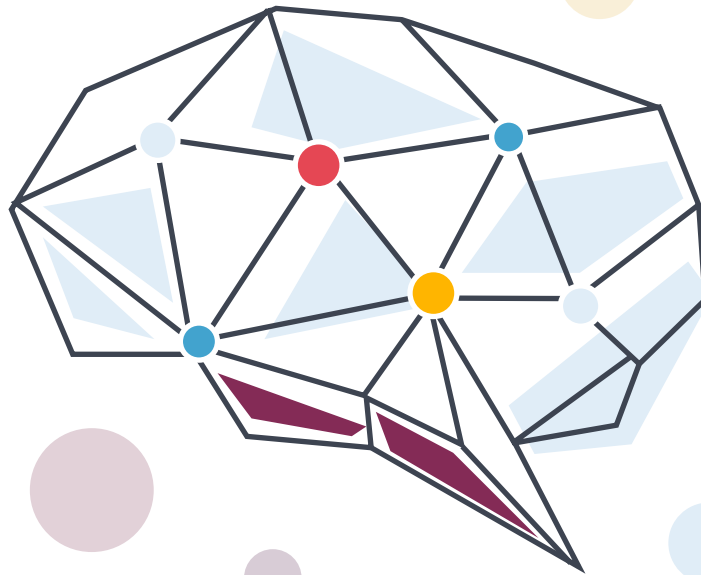
Wenn KI-Experten heute davon sprechen, meinen sie allerdings meistens den bis dato mit künstlichen neuronalen Netzen ermöglichten **Teilbereich des Machine Learning**. Denn noch ist kein Computer dazu in der Lage, eigenständig Muster zu erkennen und zu entscheiden.



Machine Learning (Unsupervised Learning)

Beim Unsupervised (Machine) Learning sind die Endwerte hingegen nicht bekannt.

Beides, das überwachte und unüberwachte Lernen, sind Lernalgorithmen, die die künstliche Intelligenz auf eine bestimmte Aufgabe hin trainieren sollen.



Machine Learning (Supervised Learning)

Unterschieden wird dabei in Unsupervised und Supervised (Machine) Learning.

Beim Überwachten Lernen geht es darum, den maschinellen Lernprozess anhand von richtigen, bereits bekannten Ergebnissen zu gestalten.



Pattern Recognition

An der Stelle ist es also unerlässlich, von der Pattern Recognition zu sprechen.

Pattern Recognition beschreibt die Fähigkeit eines sogenannten kognitiven Systems (z. B. des menschlichen Verstandes), **Gesetzmäßigkeiten, Wiederholungen als auch Ähnlichkeiten zu registrieren.**

Menschen erkennen Sprache, Bilder/Gesichter intuitiv. Der menschliche Verstand sucht wahrnehmungspsychologisch auch dort nach Mustern, wo gar keine vorhanden sind – denn die Mustererkennung, ob visuell, ob sprachlich, ist für Menschen überlebenswichtig. Mitunter entstehen dadurch aber auch Fehlannahmen, die sich beispielsweise in Aberglauben oder Vorurteilen manifestieren können.

Wenn wir also davon sprechen, dass die Computer nicht vollends über die menschliche Fähigkeit der [Mustererkennung](#) verfügen, meinen wir damit eine Mustererkennung in einem unregulierten, breiten, „real life“-Kontext. In einem definierten Datenkontext jedoch wird ein entsprechend trainierter Computer **viel effizienter und schneller Muster und Zusammenhänge erkennen und auswerten.** Das ist eine der Kernherausforderungen bei der KI-Entwicklung.

Künstliche neuronale Netze (KNN)

Der Einfachheit halber spricht man bei KI- und **Datenverarbeitungsprozessen** von „Computern“. Faktisch sind damit **künstliche neuronale Netze bzw. künstliche neuronale Netzwerke gemeint**. Die Anfänge der KNN gehen in die 1940er Jahre zurück und sind seit 2009 erneut auf dem wissenschaftlich-wirtschaftlichen Vormarsch.

Wie es der Name schon verrät, sind diese Netze in Bezug auf ihre Topologie nach dem Vorbild der biologischen neuronalen Netze aufgebaut.

Künstliche Neuronen sind die Basis der KNN. Sie werden bei den meisten KNN-Modellen in hintereinander angeordneten Schichten verbunden. Mehrschichtige Netze bestehen aus der Eingabeschicht (Input Layer), einer oder mehreren verborgenen Schichten (Hidden Layer) und der sichtbaren Ausgabeschicht (Output Layer).

Künstliche neuronale Netze werden u. a. zur Text- und Bilderkennung, zur Fehlererkennung wie beispielsweise bei Frühwarnsystemen, zur Bildverarbeitung, zur maschinellen Übersetzung, zur Synthese von Bild und Sprache, bei der medizinischen Diagnostik sowie natürlich beim Data Mining eingesetzt.

Mehrschichtige KNN und Deep Learning

Diese spezielle Methode gilt als Teilbereich des maschinellen Lernens und kommt zum Tragen, wenn die KNN Zwischenschichten (also **Hidden Layers**) **zwischen Eingabe- und Ausgabeschicht** aufweist. Die hierarchische Anordnung von Rechenkonzepten erlaubt die Verarbeitung von Rohdaten im wahrsten Wortsinne Schicht für Schicht. Die Anzahl von Schichten, ab der man von **Deep Learning** sprechen kann, ist derzeit nicht festgelegt.



Coding

Die wichtigsten Programmiersprachen der Datenwissenschaft

Ohne computergestützte Prozesse gäbe es keine Datenwissenschaft. So viel ist sicher. Aber welche **Programmiersprachen** sind wichtig und warum?

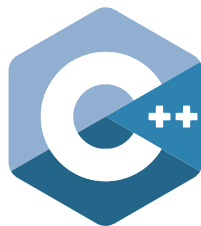
Das schnelle Tempo, das der Arbeit und auch der Entwicklung der Datenwissenschaft immanent ist, verlangt vom Data Scientist **mehr als nur die Kenntnis einer Programmiersprache.**

Es gibt aber mindestens **fünf Data-Science-Programmiersprachen**, die für typische Data-Science-Aufgaben wie Datenverarbeitung, Datenanalysen oder Datenaufbereitung unerlässlich sind:



R

wichtig bei der Entwicklung von Machine-Learning-Modellen



C++

kommt zum Einsatz beim Entwickeln skalierbarer Big-Data-Bibliotheken

SQL

SQL

die Datenbank-Programmiersprache



Java

ergänzend zu Python, sehr vielseitig

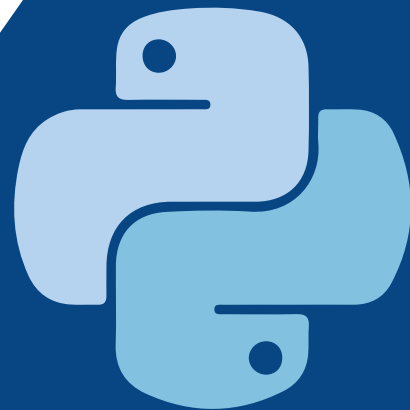
Darüber hinaus werden weitere Programmiersprachen wie Haskell, Matlab und Perl gebraucht, um unterschiedliche Arbeitsbereiche der Datenwissenschaft abzudecken. Die Wichtigkeit/Beliebtheit der einzelnen Sprachen ist zudem auch je nach Region (Deutschland, USA, China, Indien) unterschiedlich.

Python

Diese 1991 entwickelte, dynamische „Allzweck“-Programmiersprache wird u. a. bei der **Entwicklung von Web-Applikationen und Spielen genutzt** und häufig als sogenannte Skriptsprache verwendet. Die Bibliotheken und die klare Syntax der nach der britischen Comedy-Truppe Monty Python's Flying Circus benannten Sprache machen sie für Data Scientists so wichtig.

Da **Python** mit dem Anspruch entwickelt wurde, einen besonders **„aufgeräumten“, klaren Code zu ermöglichen**, unterstützt es mehrere Programmierparadigmen (z. B. objektorientiert, aspektorientiert), was den Programmierenden eine große Freiheit bei der Problemlösung bietet.

Weitere Vorteile von Python: Die Programmiersprache **läuft plattformunabhängig und erlaubt**, richtig angewandt, ein **besonders schnelles Handling** großer Datenmengen.



Berufsprofil

Data Scientist/Data Analyst

Gute **Datenwissenschaftlerinnen und Datenwissenschaftler sind derzeit** tatsächlich Mangelware. Zudem gilt der Job nicht nur dank Harvard Business Review als der „**Sexiest Job**“ **des 21. Jahrhunderts**.

Was muss man also mitbringen, um auf diesem spannenden, zukunftsweisenden Gebiet erfolgreich zu sein?

Wir haben hauptsächlich über den **Informatik-Part** der Datenwissenschaft gesprochen. Dabei bleibt sie ein interdisziplinäres Gebiet: Als Datenexpertin oder Datenexperte wird man **nicht allein am Code tüfteln, sondern vor allem auch mit Menschen interagieren** müssen.

Denn die Beschaffung von relevantem Wissen in einem Unternehmen kann einem Detektivspiel gleichen und **verlangt nach Kenntnissen der Branche**, nach einem betriebswirtschaftlichen Mindset und nach hervorragenden Kommunikationsfähigkeiten. Denn die Interpretation und Aufbereitung komplexer Daten auf eine Art und Weise, die ohne Programmier- und Mathematikkenntnisse möglich ist, ist dabei genauso wichtig, wie das eigentliche Data Sourcing und Data Mining.




Dennoch sind Programmierkenntnisse ganz offensichtlich eine Grundvoraussetzung für den Berufsweg als Datenprofi. Das Beschaffen, das Bereinigen und Aufbereiten von Daten verlangt mindestens nach Grundkenntnissen in den oben erwähnten Programmiersprachen. Fürs Anlegen von Datenmodellen und weitere Prozesse sind zudem Mathematik- und Statistikkenntnisse notwendig.

Wie werde ich Data Scientist?

Der klassische Weg in den Beruf führt über ein **Studium – Universitätsabschlüsse in Mathematik, Statistik oder Informatik** sind dabei die klaren Favoriten.

Doch da sich auch die Data Scientists oft mit spezifischen Anforderungen einer bestimmten Branche auskennen müssen, ist ein **Quereinstieg über eine Zusatzausbildung ebenso möglich**. [Präsenzkurse oder E-Learning](#): Die Optionen sind vielfältig.

Mit am wichtigsten ist jedoch der Wunsch, Neues zu lernen und mit interdisziplinären Teams zu arbeiten.



Überblick
zu unsere
Weiterbildungen

Data Science-Weiterbildungen

Certified Data
Science



XDi

Die Zukunft gestalten lernen

 xd-i.com

 intouch@xd-i.com

 +49 30 5200 1310