CHEATSHEET: DIE 95 WICHTIGSTEN BEGRIFFE RUND UM DATA & AI

Dein kompakter Überblick zu Datenarchitektur, Statistik, Machine Learning und mehr.

Datenmanagement & Architektur

1. Big Data
Sehr große, komplexe und schnell wachsende Datenmengen, die sich durch Volumen, Vielfalt und Geschwindigkeit auszeichnen und spezielle Technologien und Methoden für ihre Analyse erfordern.

2. Data Catalog
Ein detailliertes, durchsuchbares Verzeichnis aller Datenbestände einer Organisation (inklusive Metadaten wie Beschreibungen, Herkunft und Qualität), das Datenexperten hilft, schnell die passenden Daten zu finden.

3. Data Fabric
Ein integriertes Datenarchitekturkonzept, das als vereinheitlichte Schicht ("Fabric") Daten und Prozesse über verschiedene Systeme hinweg verbindet, um nahtlosen Datenzugriff und -management zu ermöglichen.

4. Data Integration
Das Zusammenführen von Daten aus mehreren heterogenen Quellen in eine einheitliche Sicht, um konsolidierte, konsistente Datensätze für operative und analytische Zwecke bereitzustellen.

5. Data Lake
Ein unstrukturiertes Datenrepository, das Rohdaten in ihrem ursprünglichen Format aufnimmt und speichert, um flexible Analysen zu ermöglichen.

6. Data Lakehouse
Ein hybrides Datenarchitekturmodell, das Merkmale von Data Lakes und Data Warehouses kombiniert – es unterstützt unstrukturierte Rohdaten sowie strukturierte Abfragen in einem einheitlichen System.

7. Data Lineage
Die Historie und der Weg der Daten von ihrer Entstehung bis zu ihrer aktuellen Form – zeigt, woher Daten stammen, welche Transformationen sie durchlaufen haben und wohin sie fließen.

8. Data Mesh
Ein dezentraler Datenarchitektur-Ansatz, bei dem domänenorientierte Teams Verantwortung für ihre eigenen Datenprodukte tragen und Daten als Produkte verwalten, unterstützt durch föderierte Governance.

9. Data Modeling
Der Prozess des strukturierten Entwurfs einer Datenstruktur (Datenmodells), der definiert, wie Daten organisiert, verknüpft und gespeichert werden – meist in Form von Entitäten und Beziehungen.

10. Data Quality
Die Güte von Daten, gemessen an Kriterien wie Genauigkeit, Vollständigkeit, Konsistenz, Aktualität und Relevanz – entscheidend dafür, ob Daten für ihren vorgesehenen Zweck geeignet sind.

11. Data Warehouse
Eine zentrale, strukturierte Datenbank, die integrierte, historische Daten aus verschiedenen Quellen für Analyse und Reporting speichert und bereitstellt.

12. Extract, Transform, Load (ETL)
Ein Datenintegrationsprozess, bei dem Daten aus Quellen extrahiert, in ein geeignetes Format umgewandelt (gereinigt, angereichert) und dann in ein Zielspeicher (z. B. Data Warehouse) geladen werden.

13. Knowledge Graph
Ein Wissensgraph, der Wissen über reale Entitäten (z. B. Personen, Orte, Dinge) und deren Beziehungen untereinander in Form eines grafbasierten Modells darstellt und maschinell interpretierbar macht.

14. Master Data Management (MDM)
Ein unternehmensweiter Ansatz, der Prozesse und Tools einsetzt, um zentrale Geschäftsobjektdaten (z. B. Kunden, Produkte) über Systeme hinweg einheitlich, konsistent und aktuell zu halten.

15. Metadata Management
Die Verwaltung von Metadaten – also Daten, die andere Daten beschreiben – durch Organisieren, Strukturieren und Speichern dieser Informationen, um Datenbestände leichter auffindbar, verständlich und nutzbar zu machen.

16. NoSQL Database
Eine nichtrelationale Datenbank, die Daten in flexiblen Formaten (z. B. Dokument-, Schlüssel-Wert-, Spalten- oder Graphenstruktur) speichert und damit hohe Skalierbarkeit und schnelle Abfragen ermöglicht.

17. Relationale Datenbank
Eine Datenbank, die Daten in Tabellen mit Zeilen und Spalten organisiert und Beziehungen zwischen diesen Tabellen definiert – SQL wird zur Verwaltung und Abfrage genutzt.

Statistik & Data-Science-Grundlagen

18. Bayessche Inferenz
Ein Ansatz der statistischen Schlussfolgerung, der Bayes' Theorem verwendet, um die Wahrscheinlichkeit einer Hypothese zu aktualisieren, sobald neue Evidenz vorliegt.

19. Deskriptive Statistik
Statistische Methoden zur Zusammenfassung und Beschreibung von Datensätzen durch Kennzahlen (z. B. Mittelwert, Median, Modus) und Streuungsmaße (z. B. Varianz, Standardabweichung).

20. Datenbereinigung
Die systematische Identifikation und Korrektur (oder Entfernung) fehlerhafter, unvollständiger oder inkonsistenter Daten in einem Datensatz, um die Datenqualität zu erhöhen.

21. Explorative Datenanalyse (EDA)
Ein Ansatz der Datenanalyse, bei dem Daten visuell und statistisch untersucht werden, um Muster, Hauptmerkmale und Ausreißer aufzudecken.

22. Feature Engineering
Der Prozess, aus Rohdaten aussagekräftige Merkmale (Features) abzuleiten oder auszuwählen, die von einem Machine-Learning-Modell genutzt werden können.

23. Hypothesentest
Ein statistisches Verfahren, bei dem eine Annahme (Nullhypothese) über einen Parameter einer Population anhand von Stichprobendaten überprüft wird.

24. Induktive Statistik
Auch „schließende Statistik" genannt; umfasst Verfahren, um von Stichprobendaten auf die gesamte Grundgesamtheit zu schließen.

25. Korrelation
Ein Maß für den statistischen Zusammenhang zwischen zwei Variablen; positive oder negative Korrelationen bedeuten gemeinsame oder gegenläufige Bewegungen.

26. Konfidenzintervall
Ein Bereich von Werten, von dem angenommen wird, dass er mit einer bestimmten Wahrscheinlichkeit den wahren Wert eines Populationsparameters enthält.

27. p-Wert
Die Wahrscheinlichkeit, unter der Annahme der Nullhypothese mindestens so extreme Beobachtungswerte zu erhalten wie die tatsächlich gemessenen.

28. Regressionsanalyse
Ein Verfahren zur Modellierung des Zusammenhangs zwischen einer abhängigen Variablen und einer oder mehreren unabhängigen Variablen.

29. Stichprobenverzerrung
Eine systematische Verzerrung, die entsteht, wenn eine Stichprobe nicht repräsentativ für die Grundgesamtheit ist.

30. Wahrscheinlichkeitsverteilung
Eine Funktion, die alle möglichen Werte einer Zufallsvariablen und deren Wahrscheinlichkeiten beschreibt, z. B. Normalverteilung.

31. Cross-Validation
Ein Verfahren zur Bewertung der Verallgemeinerungsfähigkeit eines Modells, indem die verfügbaren Daten mehrfach in Trainings- und Testdaten aufgeteilt werden.

32. Overfitting
Ein Modellphänomen, bei dem ein Machine-Learning-Modell die Trainingsdaten „zu gut" lernt und dadurch an neuen Daten schlecht generalisiert.

33. Ausreißererkennung
Die Identifikation von Datenpunkten, die deutlich von den übrigen Beobachtungen abweichen und möglicherweise Fehler oder Besonderheiten darstellen.

Machine Learning & AI-Verfahren

34. Bestärkendes Lernen
Ein Lernansatz, bei dem ein Agent durch Belohnungen und Strafen in einer Umgebung lernt, optimale Handlungsstrategien zu entwickeln.

35. Clustering
Ein unüberwachtes Lernverfahren, das ähnliche Datenpunkte automatisch in Gruppen (Cluster) zusammenfasst.

36. Deep Learning
Ein Teilgebiet des Machine Learning, das mit tiefen neuronalen Netzen komplexe Muster und Strukturen in großen Datenmengen lernt.

37. Generative Adversarial Network (GAN)
Ein Modell bestehend aus zwei neuronalen Netzen – Generator und Diskriminator – die gegeneinander trainiert werden, um realistisch wirkende Daten zu erzeugen.

38. Generative KI
Künstliche Intelligenz, die neue Inhalte (Texte, Bilder, Audio, Videos) erzeugen kann, anstatt nur vorhandene Daten zu analysieren.

39. Few-Shot Learning
Ein Lernansatz, bei dem ein Modell mit nur sehr wenigen Trainingsbeispielen pro Klasse auskommt und trotzdem gut generalisiert.

40. Large Language Model (LLM)
Ein sehr großes Sprachmodell (z. B. GPT-4, BERT), trainiert auf riesigen Textmengen, das Sprache verstehen und generieren kann.

41. Machine Learning
Ein Teilgebiet der künstlichen Intelligenz, bei dem Computer selbstständig Muster aus Beispieldaten lernen und Vorhersagen treffen.

42. Neuronales Netz
Ein von biologischen Gehirnen inspiriertes Rechenmodell aus Schichten von Neuronen, das Muster und Zusammenhänge in Daten lernt.

43. Natural Language Processing (NLP)
Ein KI-Feld, das Computern ermöglicht, menschliche Sprache zu verstehen, zu verarbeiten und zu erzeugen.

44. Random Forest
Ein Ensemble-Lernverfahren, bei dem viele Entscheidungsbäume trainiert und deren Ergebnisse für bessere Vorhersagen aggregiert werden.

45. Support Vector Machine (SVM)
Ein Algorithmus, der optimale Entscheidungsgrenzen (Hyperplanes) im Merkmalsraum findet, um Datenpunkte zu klassifizieren.

46. Transfer Learning
Ein Verfahren, bei dem ein bereits vortrainiertes Modell als Ausgangspunkt für eine neue, verwandte Aufgabe genutzt wird.

47. Transformer-Architektur
Eine Netzarchitektur mit Self-Attention-Mechanismen, die effizient Sequenzdaten verarbeitet – Grundlage moderner Sprachmodelle.

48. Überwachtes Lernen
Ein Lernverfahren, bei dem ein Modell mit gelabelten Trainingsdaten lernt, Eingaben korrekt auf Ausgaben abzubilden.

49. Unüberwachtes Lernen
Ein Lernverfahren, bei dem ein Modell Muster oder Strukturen in nicht gelabelten Daten entdeckt, z. B. durch Clustering.

Governance, Ethik & Trust

50. Accountability
Die klare Zuordnung von Verantwortung und Haftung für die Entwicklung, den Betrieb und die Folgen eines KI-Systems.

51. AI Audit
Eine systematische Überprüfung eines KI-Systems auf Kriterien wie Sicherheit, Fairness, Transparenz und Rechtskonformität.

52. AI Ethik
Das Set von Prinzipien und Werten, die sicherstellen sollen, dass KI-Systeme im Einklang mit menschlichen Werten entwickelt und eingesetzt werden.

53. AI Governance
Strategien, Regularien und interne Kontrollen, die sicherstellen, dass KI-Systeme eines Unternehmens verantwortungsvoll und regelkonform eingesetzt werden.

54. Bias (algorithmische Voreingenommenheit)
Eine systematische Verzerrung in einem KI-System, die zu unfairen oder ungleichen Ergebnissen führt.

55. Data Governance
Ein Rahmenwerk aus Richtlinien, Prozessen und Verantwortlichkeiten für den verantwortungsvollen Umgang mit Unternehmensdaten.

56. Datenschutz
Der Schutz personenbezogener Daten vor unerlaubtem Zugriff und Missbrauch, etwa durch Einhaltung von Datenschutzgesetzen wie der DSGVO.

57. Differential Privacy
Ein Datenschutzverfahren, bei dem absichtlich Rauschen zu aggregierten Daten hinzugefügt wird, um Rückschlüsse auf Einzelpersonen zu verhindern.

58. EU AI Act
Ein vorgeschlagenes EU-Gesetz zur Regulierung von KI-Systemen nach Risikostufen, inklusive Transparenz- und Sicherheitsanforderungen.

59. Erklärbare KI (XAI)
KI-Modelle, deren Entscheidungen und Vorhersagen für Menschen nachvollziehbar und verständlich gemacht werden.

60. Fairness
Im KI-Kontext die Abwesenheit von systematischen Benachteiligungen oder Bevorzugungen bestimmter Gruppen.

61. Human-in-the-Loop
Ein KI-Entwurfsprinzip, bei dem Menschen bewusst in wichtige Entscheidungsphasen eines automatisierten Systems eingebunden werden.

62. Transparenz
Das Prinzip, dass die Funktionsweise, Daten und Entscheidungslogik eines KI-Systems offen und nachvollziehbar gemacht werden.

63. Verantwortungsvolle KI
Ein Ansatz zur KI-Entwicklung und -Nutzung, der ethische, rechtliche und gesellschaftliche Aspekte gezielt berücksichtigt.

64. Vertrauenswürdige KI
KI-Systeme, die zuverlässig, robust, ethisch korrekt und gesetzeskonform gestaltet sind und so das Vertrauen der Nutzer gewinnen.

65. Datenhoheit
Das Prinzip, dass Daten den gesetzlichen Regelungen des Landes unterliegen, in dem sie erhoben oder gespeichert werden.

Visualisierung & Storytelling

66. Augmented Analytics
Durch KI und maschinelles Lernen unterstützte Datenanalyse, die Teile des Analyseprozesses (wie Datenaufbereitung und Mustersuche) automatisiert.

67. Business Intelligence (BI)
Verfahren und Technologien, die Rohdaten in aussagekräftige Informationen überführen, um fundierte Geschäftsentscheidungen zu ermöglichen.

68. Dashboard
Ein Übersichtsdisplay, das die wichtigsten Kennzahlen, Metriken und Trends zu einem Thema auf einen Blick darstellt.

69. Small Multiples
Eine Serie identisch aufgebauter, nebeneinander gereihter Mini-Charts, die denselben Messbereich für unterschiedliche Kategorien oder Zeiträume zeigen. Dadurch lassen sich Muster, Trends und Abweichungen schnell vergleichen, ohne dass das Auge sich an neue Achsen gewöhnen muss.

70. Data-Driven Decision Making
Ein Managementansatz, bei dem Entscheidungen primär auf Basis von Datenanalysen und Fakten getroffen werden.

71. Data Journalism
Ein journalistischer Ansatz, bei dem große Datenmengen gesammelt, analysiert und visualisiert werden, um komplexe Sachverhalte verständlich zu machen.

72. Data Storytelling
Die Kunst, Analysen und Erkenntnisse in eine nachvollziehbare, überzeugende Geschichte einzubetten, oft unterstützt durch Visualisierungen.

73. Datenvisualisierung
Die grafische Darstellung von Daten, um Informationen verständlich zu vermitteln und Muster oder Ausreißer sichtbar zu machen.

74. Geodatenvisualisierung
Die Darstellung von Daten mit geografischem Bezug – typischerweise auf Karten – um räumliche Muster und Zusammenhänge zu erkennen.

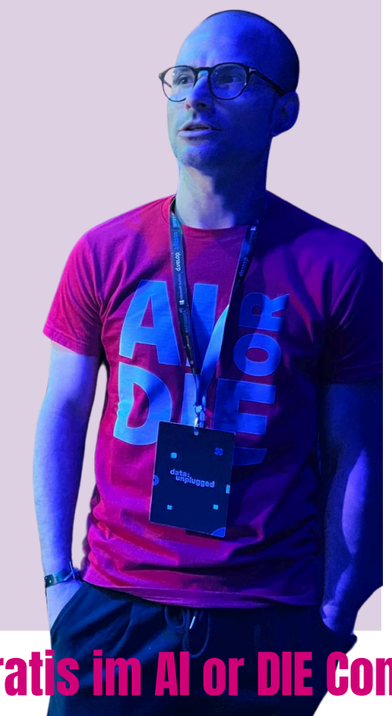
75. Infografik
Eine grafische Zusammenfassung von Daten und Informationen, die komplexe Sachverhalte auf einer einzigen übersichtlichen Grafik verständlich darstellt.

76. Interaktive Visualisierung
Visualisierungen, die dem Nutzer erlauben, in Echtzeit mit den Daten zu interagieren (z. B. durch Filtern, Zoomen, Drilldowns).

77. Narrative Visualisierung
Eine Technik, bei der Visualisierungen in eine erzählerische Struktur eingebettet werden, sodass die Datengeschichte schrittweise nachvollzogen werden kann.

78. Self-Service Analytics
Datenanalysetools und -prozesse, die es Fachanwendern ermöglichen, eigenständig Daten zu analysieren und Visualisierungen zu erstellen, ohne Unterstützung der IT.

79. Visual Analytics
Die Kombination analytischer Methoden mit interaktiven Visualisierungen, um aus großen und komplexen Datensätzen schneller Erkenntnisse zu gewinnen



Engineering, Ops & Automation

80. AIOps
Der Einsatz von KI und maschinellem Lernen zur Automatisierung und Verbesserung von IT-Betriebsaufgaben wie Überwachung, Anomalieerkennung und Fehlerbehebung.

81. AutoML
Automated Machine Learning – Tools, die Schritte wie Datenvorbereitung, Modellauswahl und Hyperparameteroptimierung automatisieren.

82. CI/CD (Continuous Integration/Continuous Deployment)
Entwicklungsmethoden, bei denen Codeänderungen häufig integriert, getestet und automatisch in die Produktionsumgebung überführt werden.

83. Cloud Computing
Ein Modell der IT-Bereitstellung, bei dem Rechenressourcen über das Internet flexibel als Dienst verfügbar gemacht werden.

84. Concept Drift
Eine Veränderung in den Daten über die Zeit, wodurch ein ursprünglich gut trainiertes Modell an Genauigkeit verliert.

85. Containerisierung
Eine Virtualisierungsmethode, bei der Anwendungen samt ihrer Abhängigkeiten isoliert in Containern ausgeführt werden.

86. Data Engineering
Der Aufbau und die Pflege der technischen Dateninfrastruktur (z. B. Pipelines, Datenbanken), um Daten für Analyse und Nutzung bereitzustellen.

87. DataOps
Eine Methodik, die Prinzipien aus DevOps auf den Bereich Datenanalyse überträgt, um Datenpipelines schneller und zuverlässiger zu betreiben.

88. DevOps
Eine Kultur und Sammlung von Praktiken, die Entwicklung (Dev) und IT-Betrieb (Ops) eng verzahnt, um Software schneller und zuverlässiger bereitzustellen.

89. Edge AI
Das Ausführen von KI-Algorithmen direkt auf Endgeräten oder am Netzwerkrand (Edge), um Latenzzeiten und Datenschutzrisiken zu minimieren.

90. Feature Store
Ein zentrales Repository, in dem Merkmale (Features) gespeichert und versioniert werden, um sie konsistent für ML-Training und Inferenz zu nutzen.

91. Modellbereitstellung
Die Implementierung eines trainierten Machine-Learning-Modells in einer Produktionsumgebung, um Vorhersagen für echte Nutzerdaten zu liefern.

92. Modellüberwachung
Die laufende Beobachtung der Performance eines produktiven ML-Modells, um Qualitätseinbußen frühzeitig zu erkennen und gegenzusteuern.

93. Modellversionierung
Das systematische Nachverfolgen und Verwalten verschiedener Versionen eines Machine-Learning-Modells während seines Lebenszyklus.

94. Serverless Computing
Ein Cloud-Modell, bei dem Entwickler Code ausführen können, ohne sich um die zugrundeliegende Serverinfrastruktur kümmern zu müssen.

95. Datenpipeline
Eine definierte Abfolge von Verarbeitungsschritten, die Rohdaten in eine verarbeitbare Form überführt und in Zielsysteme lädt.

Folge Kai-Uwe Stahl jetzt auf LinkedIn - und hol dir das hochauflösende PDF gratis im AI or DIE Content Hub!