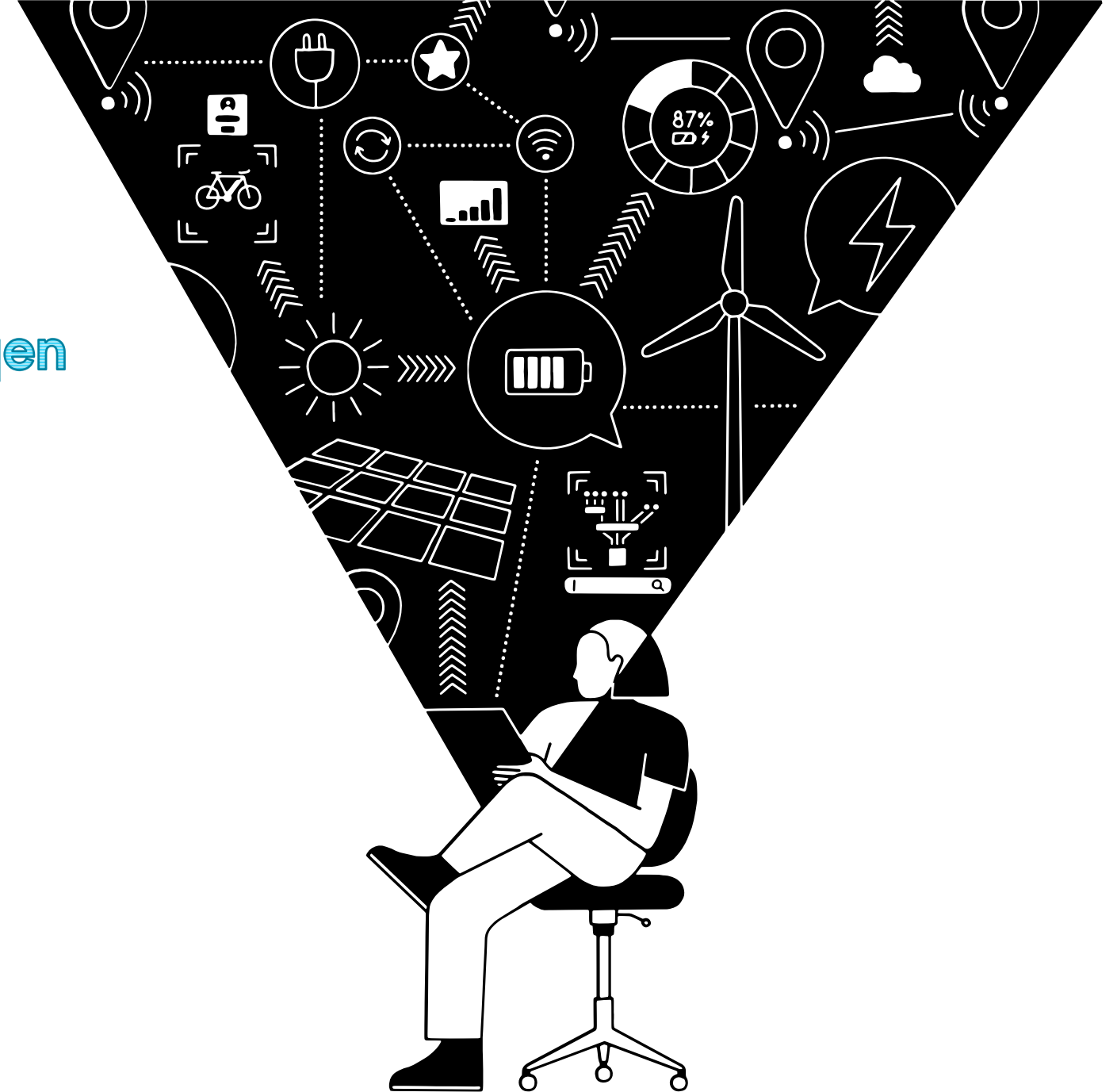


Schulung Data Science

Business Akademie: **Grundlagen**
mit Zertifizierung

Dr. André Bojahr
Michael Aydinbas

exeta





ANDRÉ BOJAHR

Lead Data Scientist

Dr. André Bojahr hat über 9 Jahre Erfahrung in Disziplinen Datenanalyse und -modellierung. Während seiner Forschungszeit entwickelte er Methoden zur statistischen Analyse von großen Datenmengen. Danach bearbeitete er verschiedenste Data Science Projekte im Bereich Zeitreihenanalyse, Computervision und NLP. Zuletzt konzipierte, entwickelte er Analyseverfahren von Radarsattelitenaufnahmen im Bahnkontext.

Biografie

- Tätigkeiten im Bereich Beratung, Automotive und Forschung
- Dr. rer. nat. in Physik, Universität Potsdam und Max-Born Institut Berlin

Beratungskompetenz

- Data Science, Maschinelles Lernen, Deep Learning, Datenanalyse, Statistik, Computer Vision, Zeitreihenanalyse, NLP, Projektmanagement, Ausbildung, Datenbanken
- Technische Expertise in PyTorch, Tensorflow, Scikit-learn, Pandas, Python, C++, Embedded Systems, Jetson, Cloud-Computing, AWS, Azure

Sprachen

- Deutsch, Englisch

Selected Relevant Project Experience

Projektleiter und Senior Data Scientist, SAR-satellite images for tree detection, Eisenbahnbundesamt

- Anforderungsmanagement, Dokumentation, Entwicklung
- Bildverarbeitung (Bildtransformationen, Geomatching, Rauschreduktion mit Deep Learning)
- Änderungsdetektion (statistische Analyse/Tests, Deep Learning Ansätze (U-Net))

Senior Data Scientist, NLP-Matching von Projekt und Mitarbeiter Profilen, Intern (vorheriger Arbeitgeber)

- NLP-Methoden (TF-IDF, Word2Vec, Doc2Vec)
- Ähnlichkeitsmetriken (Kosinus-Ähnlichkeit, euklidische Distanz)
- Graphendatenbank Neo4j

Senior Data Scientist, Fehlerursachendetektion bei Fahrzeugen, Deutscher Fahrzeughersteller

- Zeitreihenanalyse (geführte Merkmalsauswahl mittels Statistik und Ingenieurswissen)
- Machine Learning Methoden (Random Forrest, Gradient Boosted Trees, Neural Networks)
- Explainable AI (LIME, SHAP, Surrogate Models)



MICHAEL AYDINBAS

Senior Consultant

Michael Aydinbas verfügt über 11 Jahre IT-Erfahrung sowie 4 Jahre Berufserfahrung als Experte für Datenplattformen. Darüber hinaus ist er als Projektleiter tätig. Sein Schwerpunkt liegt auf skalierbaren, performanten Big Data Anwendungen. Er hat zahlreiche Projekte in verschiedenen Branchen begleitet und umgesetzt. Als Data Engineer konzipiert und gestaltet er Messdatenanalyse-Pipelines – von der Datenaufbereitung über deren Analyse bis zur Auslieferung.

Biografie

- Tätigkeiten im Bereich Automotive, IT-Dienstleistungsunternehmen, Finance, Lehre
- M.Sc. Psychologie in IT, TU Darmstadt.
B.Sc. Umweltingenieurwissenschaften, TU Darmstadt

Beratungskompetenz

- Datenstrategie und Systemlandschaft, Schulungen zu Python, Machine Learning und Deep Learning, Forschung und Entwicklung
- Zertifizierungen: PSM I, Machine Learning mit Python, CPRE FL
- IT-Expertise u.a. in Data Science, Large-Scale Data Processing, System Integration, AWS

Sprachen

- Deutsch, Englisch

Selected Relevant Project Experience

Projektleiter/ Senior Data Engineer, Betreiben eines Data Mesh, Finance

- Einarbeitung in die gesamte Datenplattform (AWS) und Übernahme von (Legacy) ETL-Datenstrecken
- Aufbau eines neuen Data Teams, Recruiting und Onboarding Prozesse

Projektleiter/ Senior Solution Engineer, Emissionsdatenberechnung und Datenintegration

- Projektplanung und Mitarbeiterführung für drei Teilgebiete, zentraler Ansprechpartner für den Kunden
- Erarbeitung von serverless Cloud-Lösungsarchitekturen und Softwarekonzepten

Projektleiter/ Senior Consultant, Big Data im Entwicklungsprozess, Automotive

- Begleitung/Umsetzung des Deployments von entwickelten Analysetools auf eine Big Data Analyse-Plattform
- Entwicklung, Kundenkommunikation, Projektplanung, Dokumentation

Data Scientist/ Requirements Engineer, Entwicklung eines Sequencing-Frameworks, Automotive

- Anforderungsanalysen
- Entwicklung einer Python Toolbox für Datenexploration/ Datenanalyse mit Apache Spark
- Entwicklung eines Python Sequencing Frameworks zur automatisierten Erkennung von Fahrszenen

Schulung Data Science Business Akademie: Grundlagen mit Zertifizierung

Data Science im Unternehmen

- Einführung und Definition
- Abgrenzung zu anderen Unternehmensbereichen
- Typische Use Cases
- Betriebswirtschaftliche Anwendungsgebiete und konkrete Data Science Ziele
- Betriebswirtschaftliche Erfolgsmessung

Statistische Grundlagen

- Deskriptive Statistik und statistische Verteilungen
- Grundlagen der Wahrscheinlichkeitsrechnung
- Explorative Analyseverfahren

Data Science Prozess

- Zieldefinition
- Datenauswahl
- Datenbereinigung und Datenvorverarbeitung
- Modellbildungsprozess
- Werkzeuge
- Interpretation und Darstellung der Ergebnisse
- Nutzung im Unternehmen

ML Algorithmen

- Supervised/Unsupervised Machine Learning
- Clustering
- Klassifikation und Regressionsmethoden
- Entscheidungsbäume und Ensemble Verfahren
- Lineare und logistische Regression, Support Vector Machines
- Neuronale Netzwerke


Praktische Anwendungen

- Arbeiten mit Data Science Werkzeugen
- Grafische Werkzeuge
- Implementierung unterschiedlicher Ziel-funktionen
- Beispielhafte Methodenanwendung für Data Science Datensätze
- Übungen zum selbständigen Anwenden der Methoden

Vorgehensmodelle

- Übersicht verschiedener Modelle
- Cross Industry Standard Process of Data Mining (Crisp-DM)
- Knowledge Discovery in Databases (KDD)

Schulungsüberblick

TAG 1	TAG 2	TAG 3	TAG 4	TAG 5
Data Science im Unternehmen <ul style="list-style-type: none">• Use Cases• Vorgehens- modelle• KI-Projekte einführen	Data Science Prozess Teil I <ul style="list-style-type: none">• Statistische Grundlagen• Explorative Datenanalys e	Data Science 	Machine Learning Algorithmen unsupervised & Supervised Learning	xAI Anwendung Abschlusstest Ausblick
Übungen				

Schu

Abschl

THEORI

Wir möch
bringen u
daher gilt

- Fragen
 - Es ist f
 - keine
 - Keine
 - Fragen
- auf ze



ntnisse
ng die
rten
haben
e sich am
m an Tag

Lernziele

Data Science Aufgaben, Rollen, Prozesse

Use Case Identifikation und Bewertung im Unternehmen

Einordnung von Buzzwords

Data Science Terminologie & Grundlagen

Durchspielen eines Data Science Prozesses anhand von Beispielen und selbständigen Übungen

Statistische Grundlagen (Deskriptive Statistik)

Explorative Datenanalyse (EDA)

Supervised und Unsupervised Machine Learning

Zentrale Algorithmen

Vorstellungsrunde

<https://miro.com/app/board/uXjVP7h6SRU=/>

Pass: DHL_2022

Wer bin ich?

Habe ich schon Erfahrung mit Data Science?

Was ist mein Hintergrund?

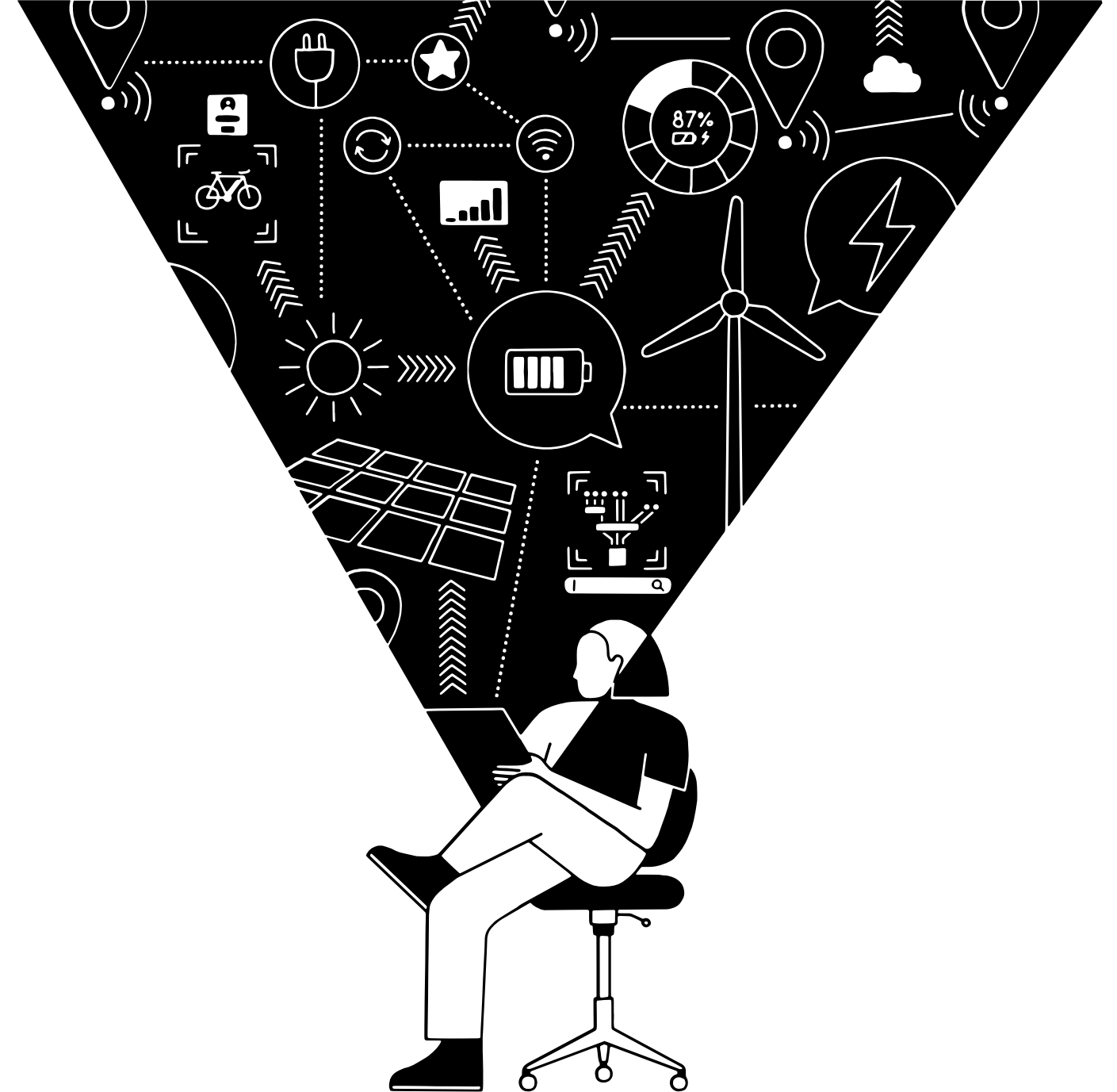
Habe ich schon Hands-on-Erfahrung mit der Entwicklung von Data Science Lösungen und welche Tools habe ich benutzt?

Was erwarte ich von dem Workshop? Wann ist der Workshop für mich ein Erfolg?

Data Science im Unternehmen

Dr. André Bojahr
Michael Aydinbas

exeta



Einstieg Data Science und Artificial Intelligence

Data Science ist ein wichtiger Schlüssel um Künstliche Intelligenzen, wie sie heute existieren, möglich zu machen.

Zum Einstieg ein paar lockere Beispiele aktueller KI-Systeme:

<https://labs.openai.com>

<https://chat.openai.com/chat>

<http://www.deepl.com>

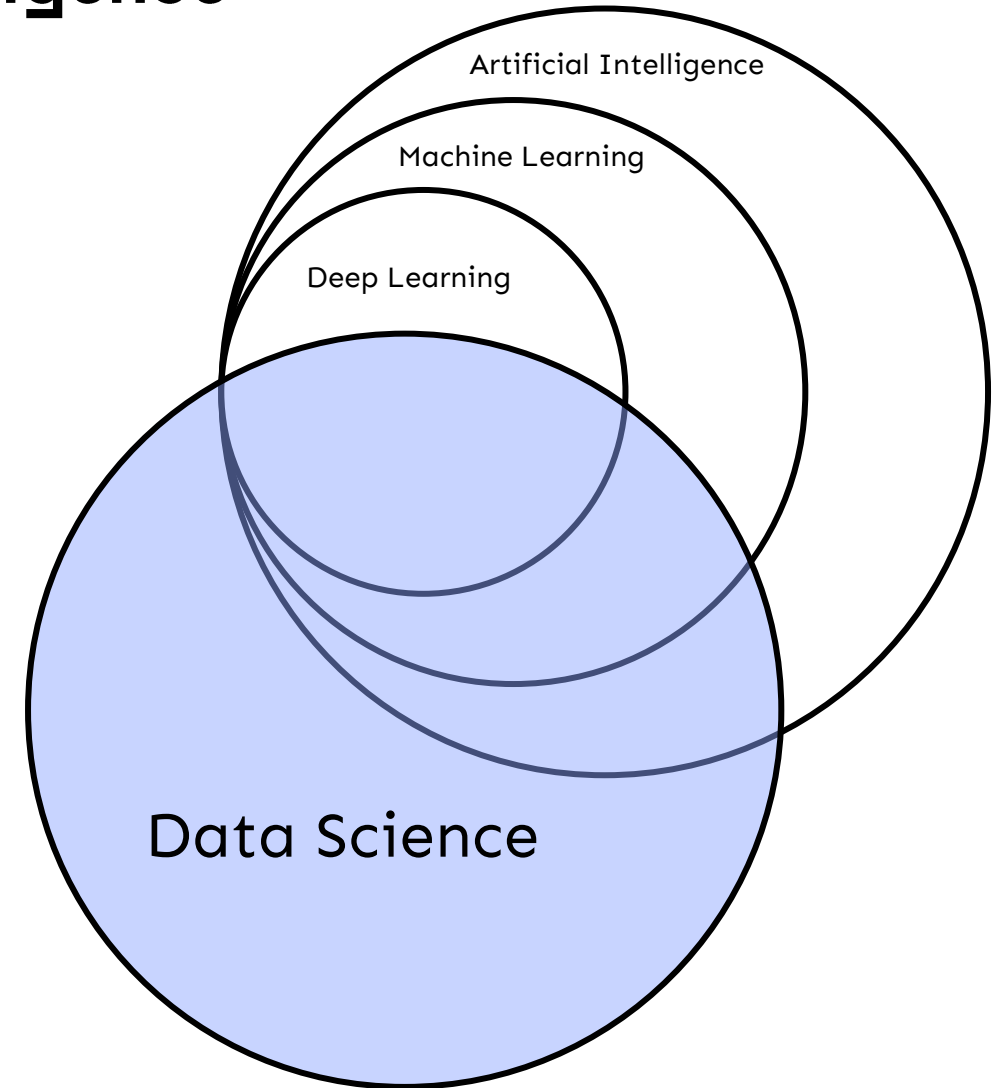
<http://imagineaire.cc/gaugan2/>

<https://huggingface.co/spaces/akhaliq/yolov7>

<https://reininakano.com/arbitrary-image-stylization-tfjs>

<https://transcranial.github.io/keras-js>

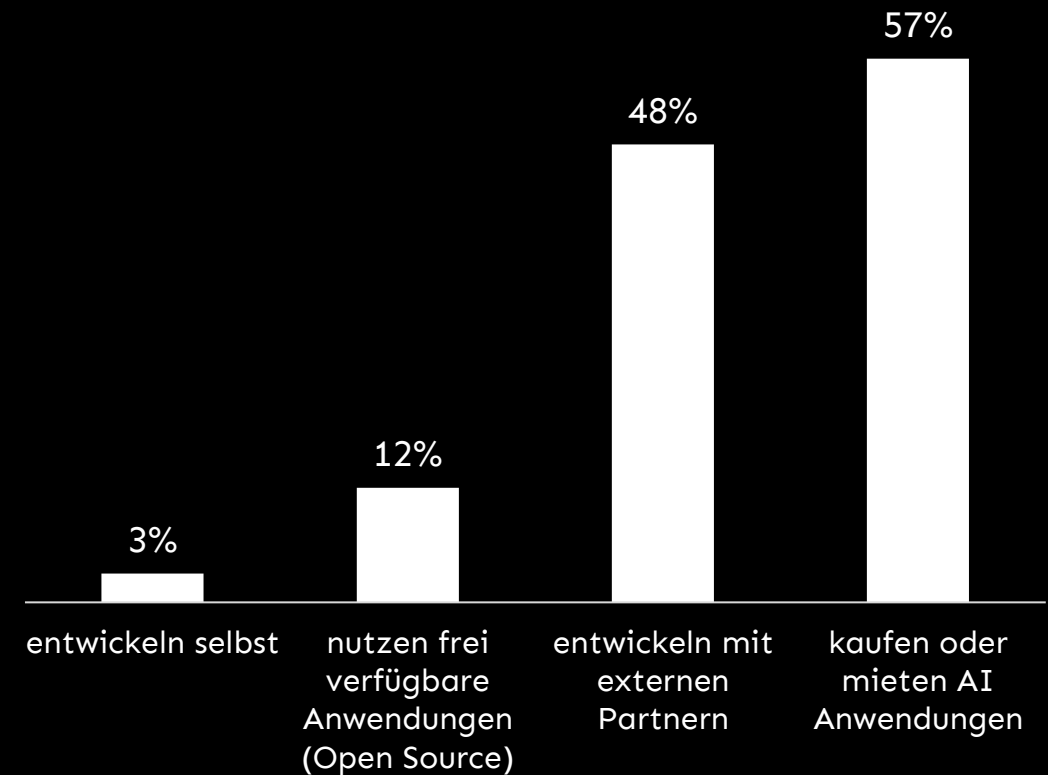
Technisch werden AI-Systeme in dieser Schulung nicht behandelt.
Data Science ist aber eine wichtige Grundlage dieser Systeme.



AI in Zahlen

- Im Jahr 2022 beträgt die Vorhersage des weltweiten **Umsatzes mit AI-Software 62,5 Milliarden US-Dollar**. Das ist im Vergleich zu 2021 eine Steigerung von 21.3%.
(Gartner 2021)
- Für Deutschland wurde eine potenzielle **Steigerung des BIP durch AI basierte Dienstleistungen von 480 Milliarden Euro** bis 2025 prognostiziert, was einem Wachstum von 13% entspricht.
(BMWI 2021)
- Untersuchungen aus 2021 zeigen, dass nur **4% des deutschen Mittelstands** AI in ihren Unternehmen nutzen. Dies ist insbesondere auf fehlendes Knowhow zurückzuführen.
(KfW 2021)

Unternehmen die AI nutzen



Daten: Bitkom Research 2021

Was ist eigentlich Data Science?

Einführung und Definition

Data Science ist ein Feld, welches wissenschaftlich fundierte Methoden, Prozesse, Algorithmen zur Extraktion von Erkenntnissen, Mustern und Schlüssen sowohl aus strukturierten als auch unstrukturierten Daten ermöglicht.

Der unternehmerische Zweck ist, basierend auf Daten, Prozesse zu verschlanken, Geld und Zeit zu sparen oder neue Services anzubieten.

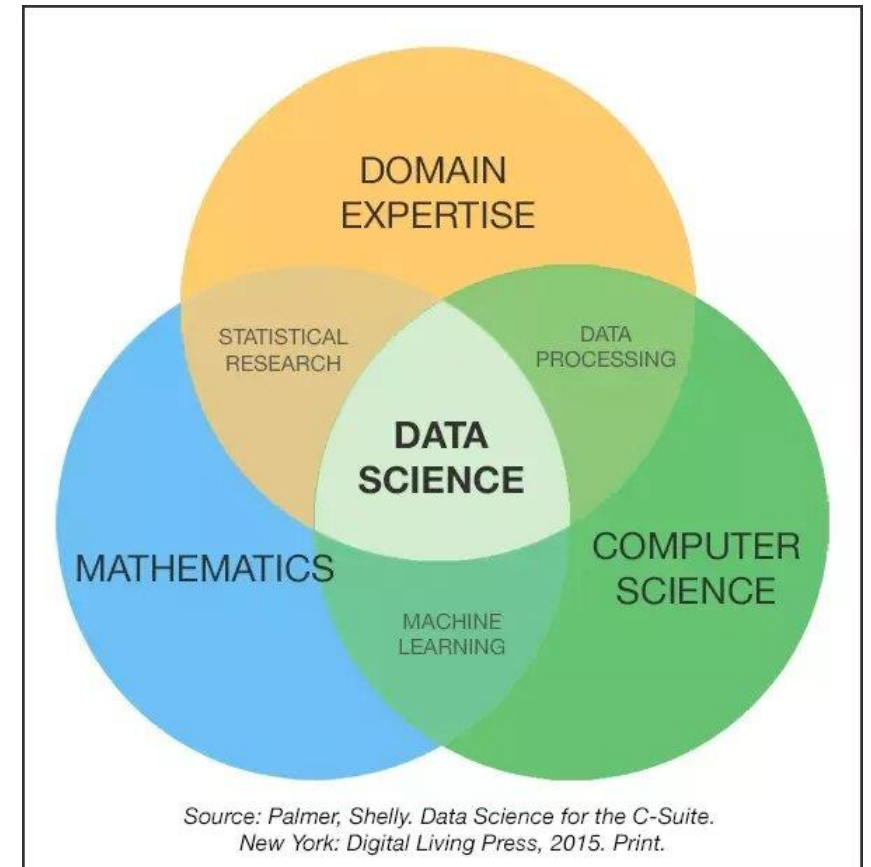
Die Rolle eines Data Scientisten

The Sexiest Job of the 21st Century*

Data Scientisten besitzen ein tiefes und fundiertes mathematisches Verständnis. Zusätzlich erlaubt ihnen ihr Informatikwissen mathematisch sehr anspruchsvolle Algorithmen umzusetzen. Darüber hinaus benötigen Data Scientisten ein sehr gutes Systemverständnis für die Domäne, in der sie arbeiten. Data Scientisten, die oft die Domäne wechseln, benötigen besonders gute Kommunikationsfähigkeiten und eine besonders schnelle Auffassungsgabe, um schnell wieder Systemverständnis zu erlangen.

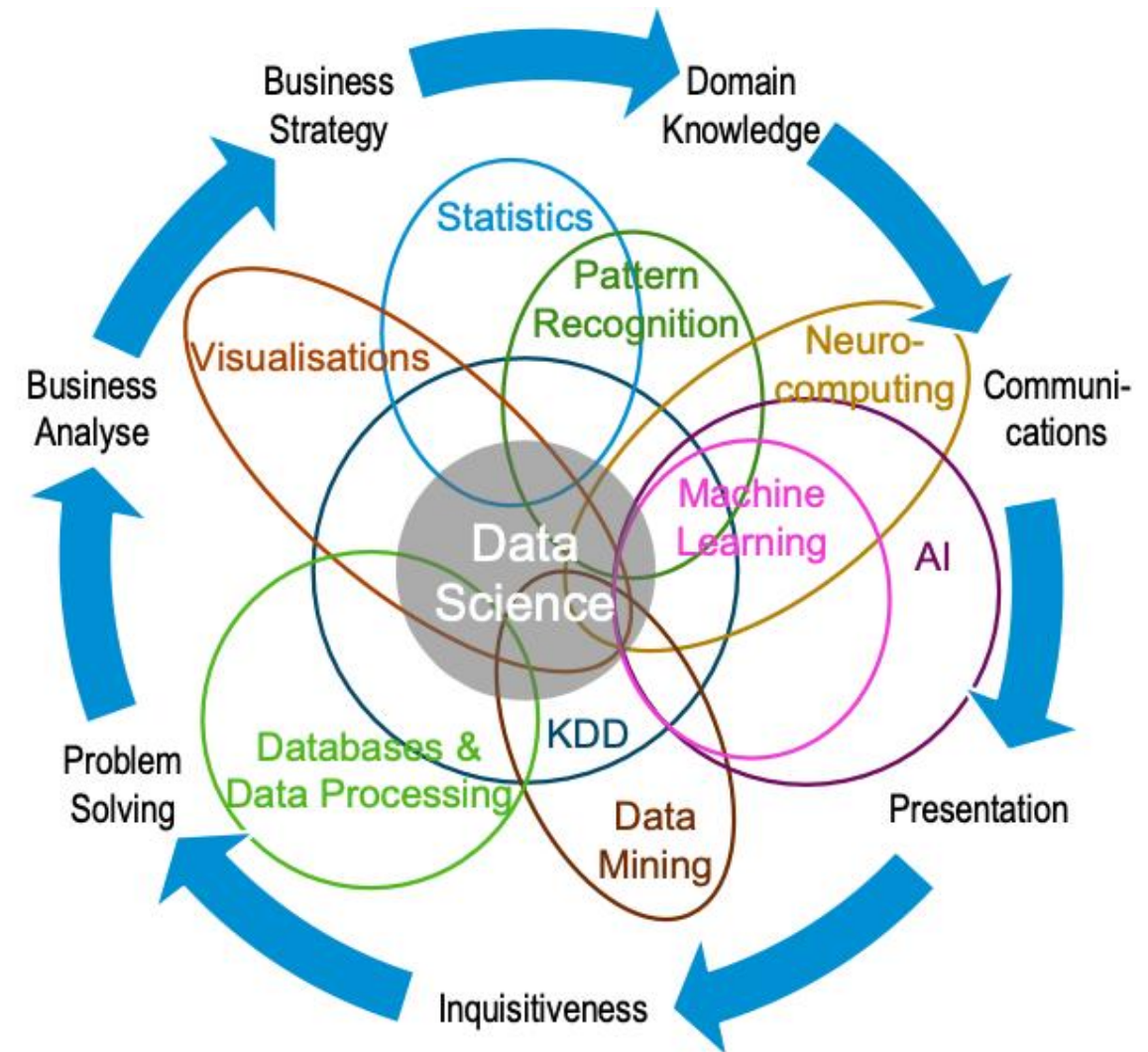
*Harvard Business Review Oct 2012

<https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century>



Data Science ist Multidisziplinär

- **Softskills:** Kommunikativ, Halten von Präsentationen, einfaches erklären von komplexen Sachverhalten, Empathisch, Neugierde
- **Technisches Wissen:** Datenbanken (SQL, NoSQL, Graphendatenbanken), Statistik, Datenverarbeitung, Machine Learning, AI, Neuronale Netze, Mustersuche, Cloud Umgebungen
- **Problemlösekompetenz:** Strukturiert <-> Kreativität
- **Business Understanding**
- Verständnis für die Domäne z. B.:
 - Verständnis der Funktionsweise eines Sensors
 - Bei vorhersage der Batteriekapazität (über Lebenszeit)
→ Verständnis von der Batteriephysik / chemie



Data Science Tätigkeitsfelder

Data Science ist eine interdisziplinäre Disziplin, die sich mit der Extraktion von Wissen aus Daten befasst, und je nach Problemstellung verschiedene Ausprägungen in der analytischen Vorgehensweise haben kann.



DESCRIPTIVE ANALYTICS

Was ist passiert?

Wann ist das Gleitlager überhitzt?



DIAGNOSTIC ANALYTICS

Warum ist es passiert?

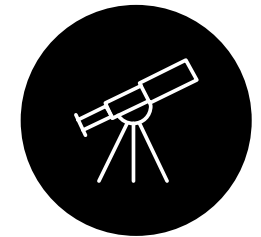
Warum ist das Gleitlager überhitzt?



PREDICTIVE ANALYTICS

Was wird passieren?

Wann wird das Gleitlager wahrscheinlich überhitzen?



PRESCRIPTIVE ANALYTICS

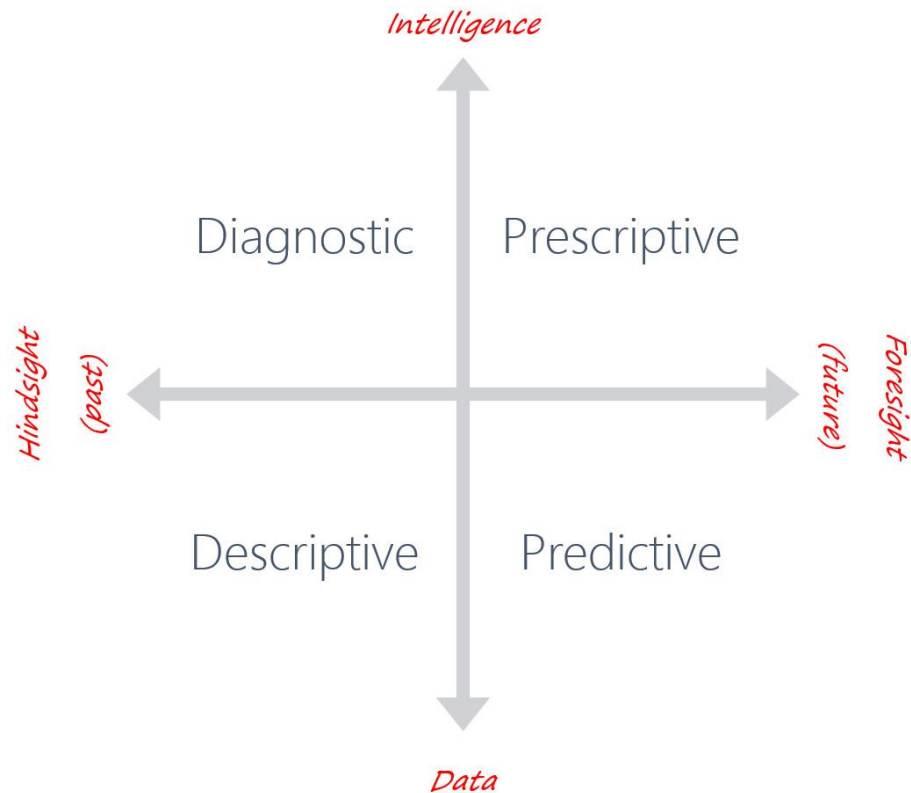
Was sollen wir tun?

Wie kann verhindert werden, dass das Gleitlager überhitzt?

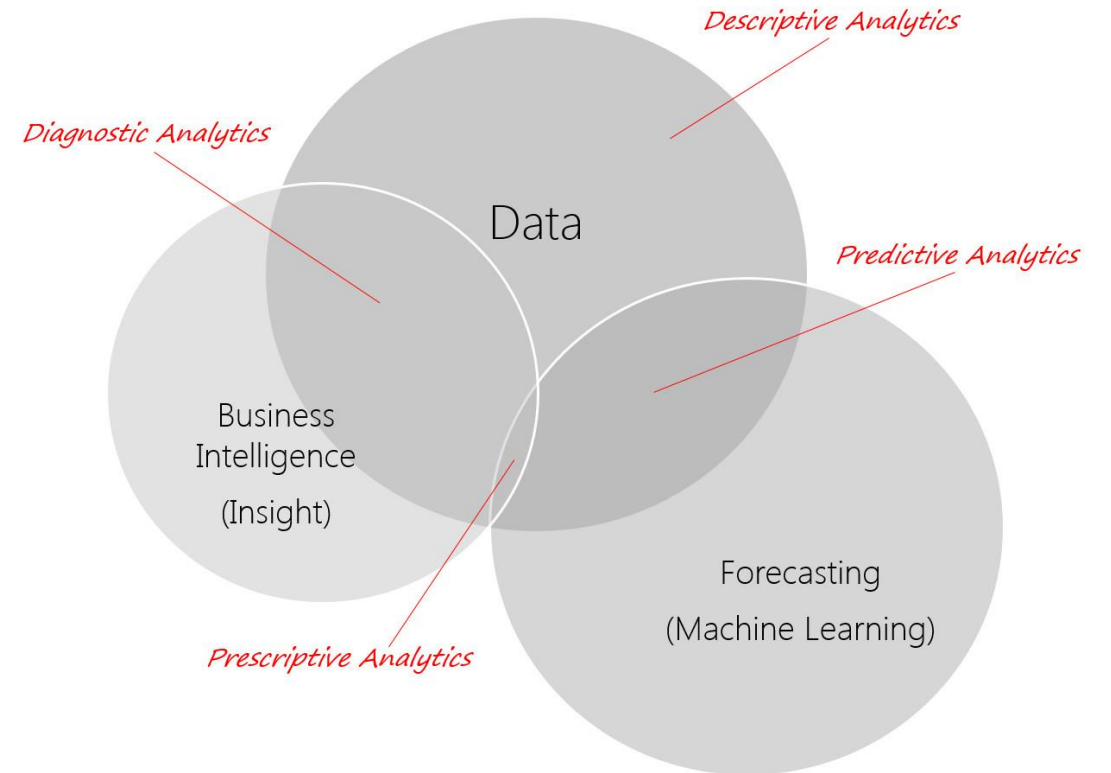
Data Science Tätigkeitsfelder

Data Science ist eine interdisziplinäre Disziplin, die sich mit der Extraktion von Wissen aus Daten befasst, und je nach Problemstellung verschiedene Ausprägungen in der analytischen Vorgehensweise haben kann

DARSTELLUNG NACH ZEIT UND METHODIK



DARSTELLUNG ALS VENN-DIAGRAMM



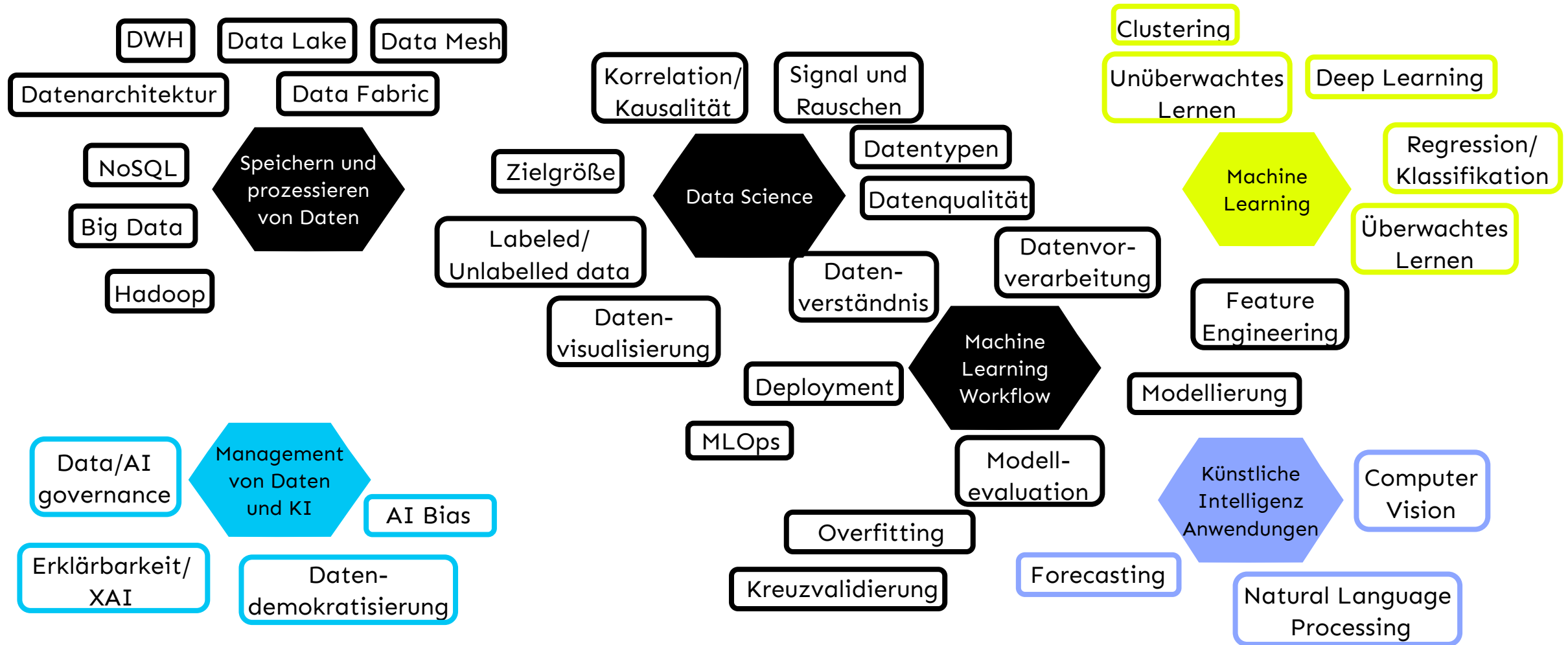
https://miro.com/app/board/uXjVP7h6SRU=

Pass: DHL_2022

**Interaktive Übung
zu bekannten Begriffen
aus Data Science Umfeld**



Data Science Word Cloud



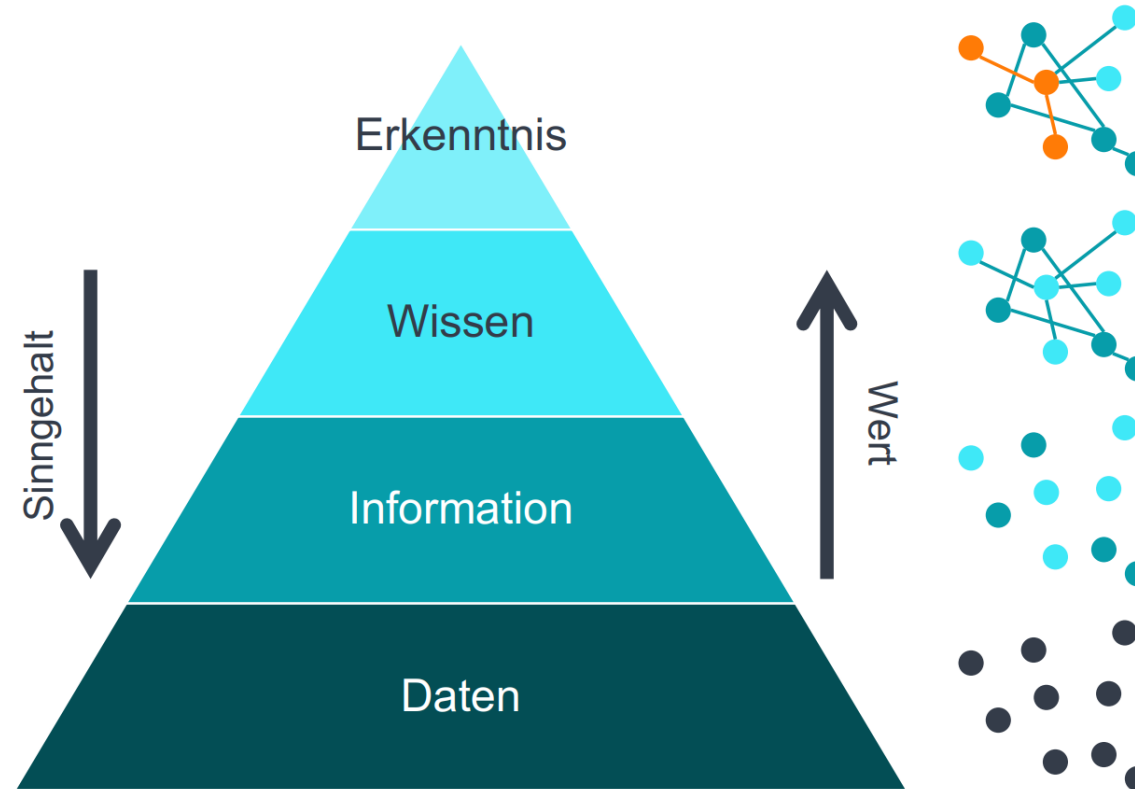
Data Science im Unternehmen

Daten und Information

Verarbeitbarkeit und Sinngehalt

Information beschreibt den Verarbeitbarkeitsgrade und den Sinngehalt

Daten bezieht sich auf die Art und Weise wie Information gespeichert werden



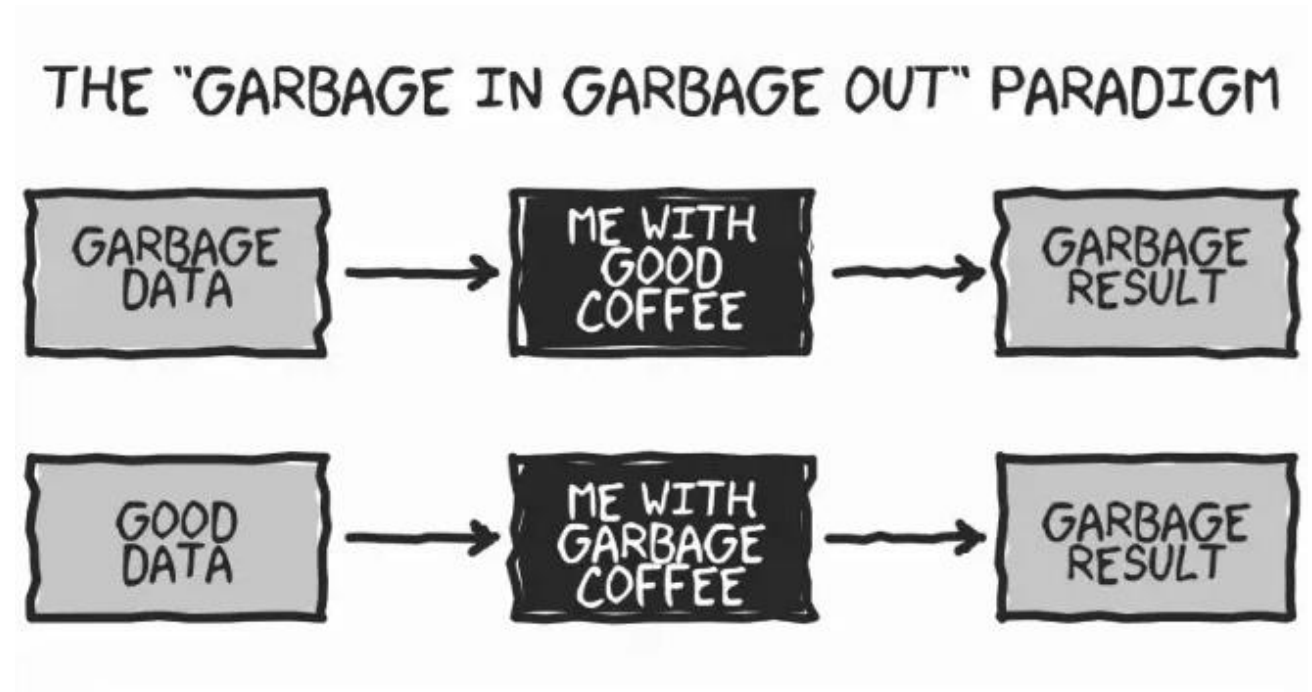
Was sind Daten?

Repräsentationen von Fakten, welche in unterschiedlichen digitalen Formen gespeichert werden kann, was eine Verarbeitung mittels Computer und Algorithmen erlaubt.

Data Science im Unternehmen

Grenzen von Data Science

- Die wachsende Bedeutung von Data Science ist durch die Verfügbarkeit von **Big Data** und **billiger Rechenleistung** begründet.
- **Kleine Datensätze, Daten von schlechter Qualität, inkonsistente Daten und fehlerhafte Daten** sind für Data Scientists ein täglich wiederkehrendes Problem und können Zeit verschwenden und zu Analysen führen, die irreführend sind.
- Dieser Kurs wird Sie in einige grundlegende Techniken der Data Science einführen, **aber denken Sie immer daran, dass Data Science auf Daten beruht!**



https://miro.com/app/board/uXjVP7h6SRU=

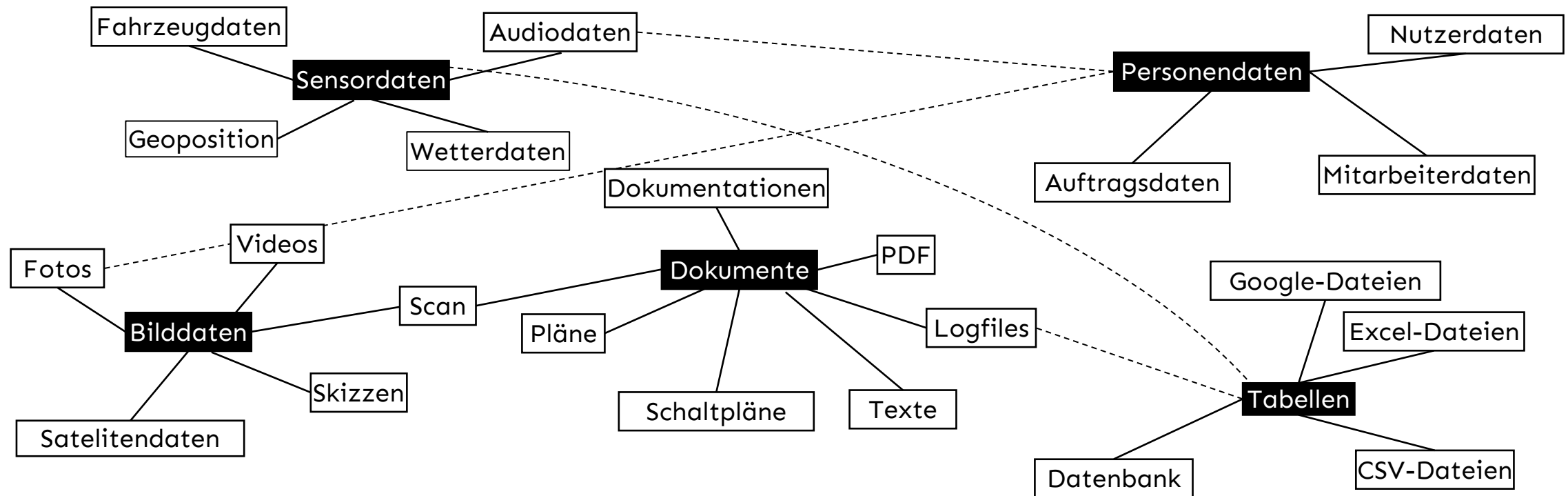
Pass: DHL_2022

Interaktive Übung zu Datenquellen bei DHL



Datenquellen und Datenarten

Daten – das Öl des 21. Jahrhunderts



Überwachtes und Unüberwachtes Lernen

Typen von Maschinellem Lernen

Die meisten Data Science Use Cases können in einer der beiden Kategorien einsortiert werden. Typischerweise sind die allermeisten in der ersten Kategorie (Überwachtes Lernen) einsortiert.

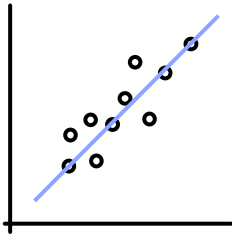


Warum ist die linke Kategorie dominant/beliebter/wichtiger?

Supervised Learning
(Überwachtes Lernen)

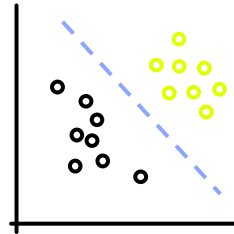
Unsupervised Learning
(Unüberwachtes Lernen)

Regression



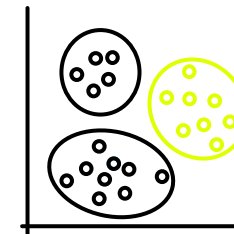
Vorhersage einer
kontinuierlichen
Zielgröße, z.B.
Immobilienpreis

Klassifikation



Vorhersage einer
kategorischen
Zielgröße, z.B.
Schadensklasse

Clustering



Einteilen von
Beobachtungen
in **Gruppen**, z.B.
Kundensegmente

Data Science im Unternehmen

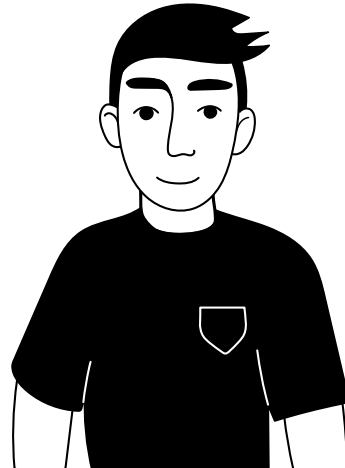
Komplexität der heutiger Data Science Anwendungen und Anforderungen an das Personal sorgt für weitere Rollen und Spezialisierungen:

Data Scientist



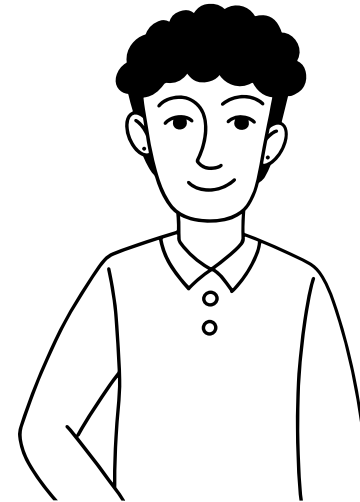
Fokus auf Modellierung,
Algorithmen und
Mathematik

Machine Learning Engineer



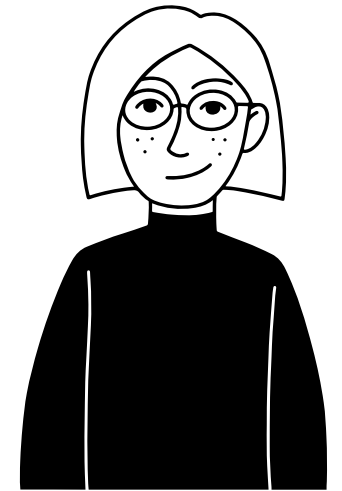
Fokus auf sauberen
Produktiv-Code für die
Algorithmen und
Modelle

Data Engineer



Fokus auf Architektur
und Werkzeuge der
Datenverarbeitung
(Cloud, Datenbanken)

Data Analyst

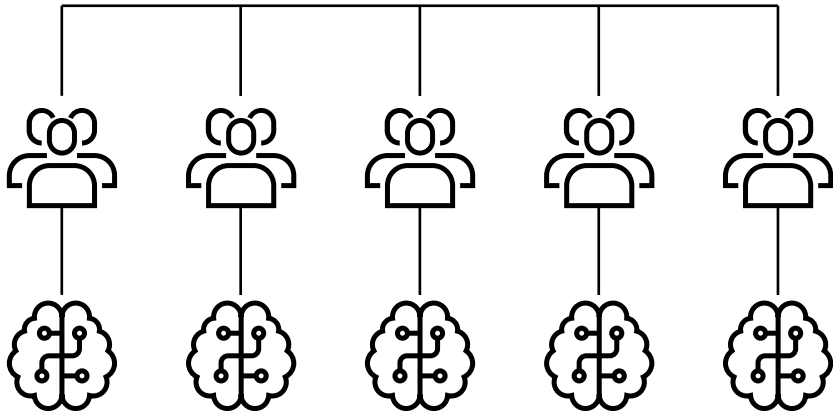


Fokus auf Business und
Visualisierung
(Business
Understanding)

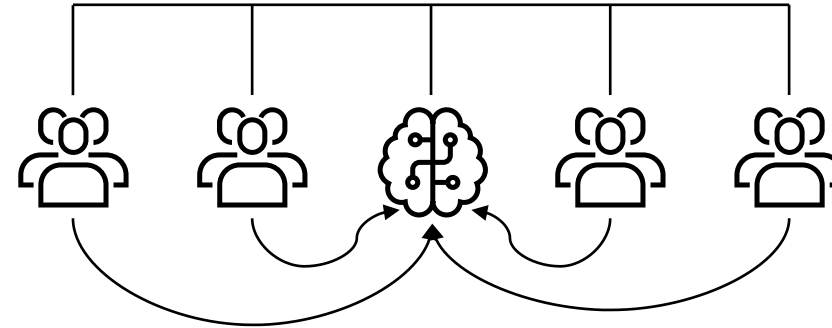
Data Science im Unternehmen

Organisatorische Struktur Unterschied zu anderen Business-Units

Verteilter bzw. dezentraler Ansatz



Zentralisierter Ansatz



Betriebswirtschaftliche Erfolgsmessung

UNTERNEHMERISCHE ERFOLGSKRITERIEN

Beschreiben die Kriterien für ein erfolgreiches oder nützliches Projektergebnis aus der Sicht des Unternehmens

Was fällt Ihnen dazu ein?

DATA SCIENCE ERFOLGSKRITERIEN

Beschreiben die Kriterien für ein erfolgreiches Ergebnis des Projekts in technischer Hinsicht

Was fällt Ihnen dazu ein?

https://miro.com/app/board/uXjVP7h6SRU=

Pass: DHL_2022

Interaktive Übung zu Erfolgsbemessung



Data Science im Unternehmen

Realität in der Industrie

How Do You Measure Success?

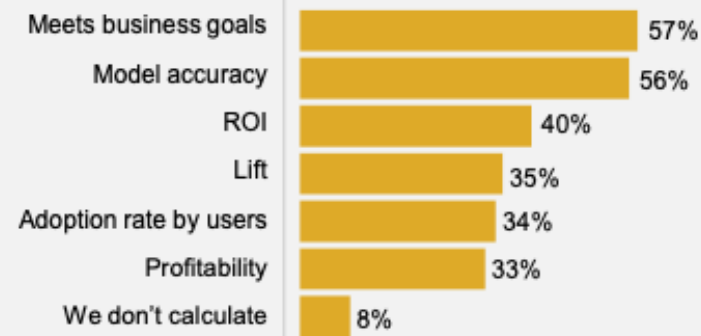


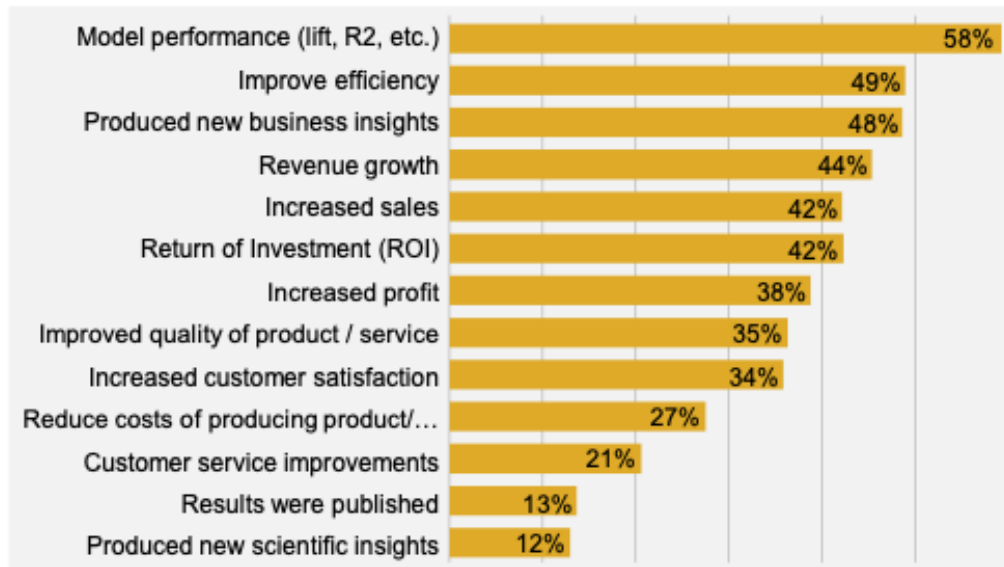
Figure 5. Based on 110 users who have implemented predictive analytics initiatives that offer "very high" or "high" value. Respondents could select multiple choices.

PREDICTIVE ANALYTICS

Extending the Value of Your Data Warehousing Investment
By Wayne W. Eckerson

See: <https://www.rexeranalytics.com/>

© 2020 SAP SE or an SAP affiliate company. All rights reserved. | PUBLIC



In their Third Annual Data Miner Survey, Rexer Analytics, an analytics and renowned CRM consulting firm based in Winchester, Massachusetts, asked the BI community "How do you evaluate project success in data mining?" Out of 14 different criteria, a massive 58% ranked "model performance" (lift, R2, etc.) as the primary factor.

Data Science im Unternehmen

Was macht einen guten Data Scientist aus?

- Denken, Denkweise und die Fähigkeit zu analytischem, kreativem, kritischem und neugierigem Denken
- Methoden und Wissen über komplexe Systeme und Ansätze zur Durchführung von Top-down- und Bottom-up-Problemlösungen
- Fundierte Kenntnisse der gängigen Methoden und Modelle in den Bereichen Statistik, Data Mining und maschinelles Lernen
- Fähigkeit zur Implementierung, Wartung und Fehlerbehebung von Big-Data-Infrastrukturen, wie z. B. Cloud Computing, Hochleistungsrecheninfrastrukturen, verteilte Verarbeitungsparadigmen, Stream Processing und Datenbanken
- Kenntnisse in den Bereichen Mensch-Computer-Interaktion, Visualisierung und Wissensrepräsentation Darstellung und Verwaltung von Wissen
- Erfahrung in der Softwareentwicklung (einschließlich Systementwurf und -analyse), Qualitätssicherung
- Erfahrung im Umgang mit großen Datenmengen und gemischten Datentypen und -quellen in einem vernetzten und verteilten Umfeld
- Erfahrung in der Datenextraktion und -verarbeitung, dem Verständnis von Merkmalen und der Analyse von Beziehungen
- Aktives Interesse an und Kenntnisse über multidisziplinäre und transdisziplinäre Studien und Methoden in den Natur-, Technik-, Sozial- und Lebenswissenschaften
- Umfassende Erfahrung mit modernem analytikorientiertem Scripting, Datenstrukturen, Programmiersprachen und Entwicklungsplattformen in einer Linux-, Cloud- oder verteilten Umgebung
- Theoretischer Hintergrund und Fachwissen für die Bewertung der technischen und geschäftlichen Vorzüge der analytischen Ergebnisse
- Hervorragende schriftliche und mündliche Kommunikationsfähigkeiten [Matsudaira 2015] und organisatorische Fähigkeiten, die Fähigkeit, analytische Materialien und Berichte für verschiedene Zielgruppen zu verfassen und zu bearbeiten, und die Fähigkeit, analytische Konzepte und Ergebnisse in geschäftsfreundliche Interpretationen umzuwandeln; die Fähigkeit, verwertbare Erkenntnisse an nichttechnische Zielgruppen zu vermitteln, und Erfahrung in der datengestützten Entscheidungsfindung

Quelle: <https://dl.acm.org/doi/pdf/10.1145/3076253>

Wir bauen ein Glossar

https://miro.com/app/board/uXjVP7h6SRU=

Pass: DHL_2022

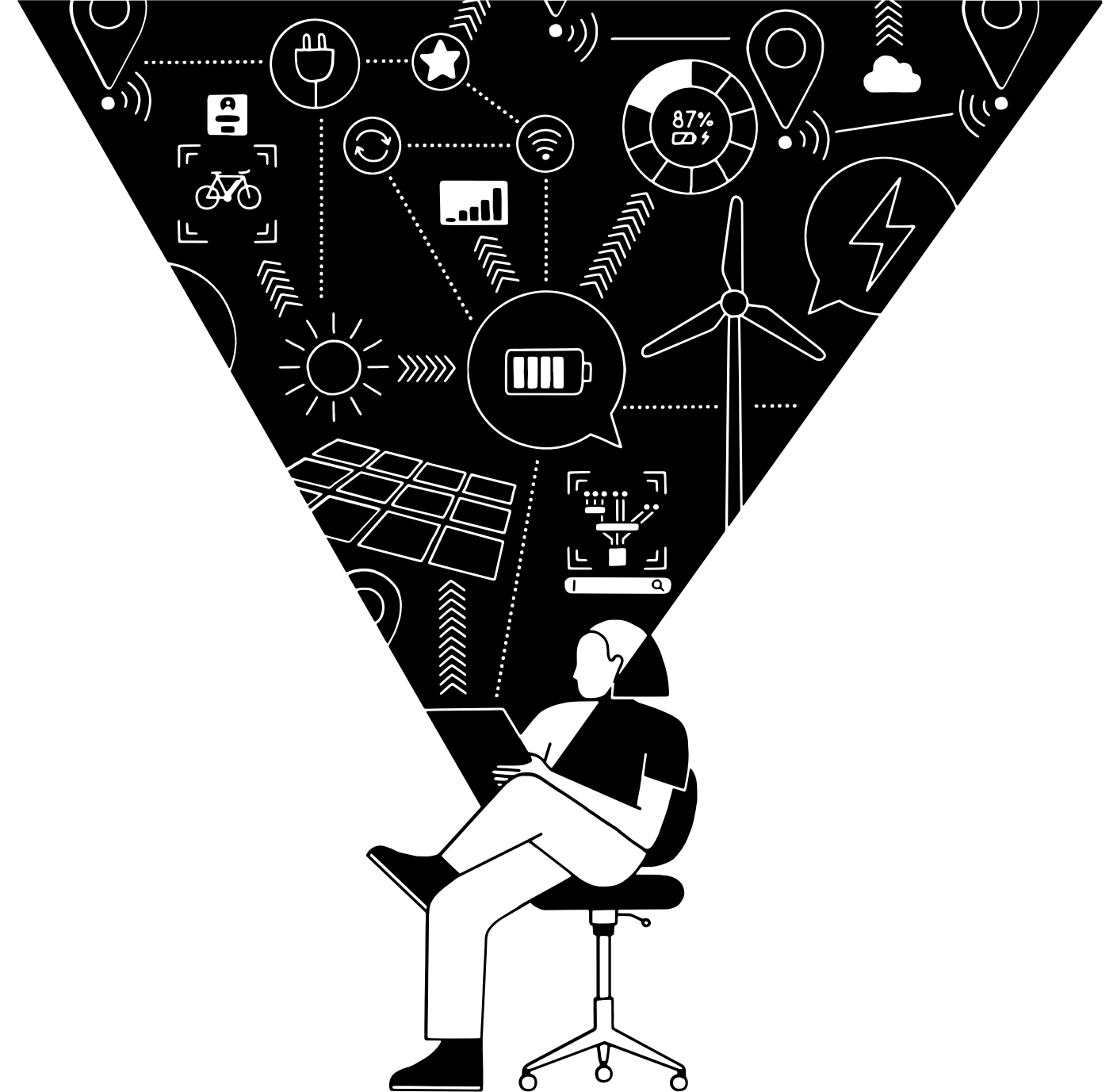
Was war das denn?



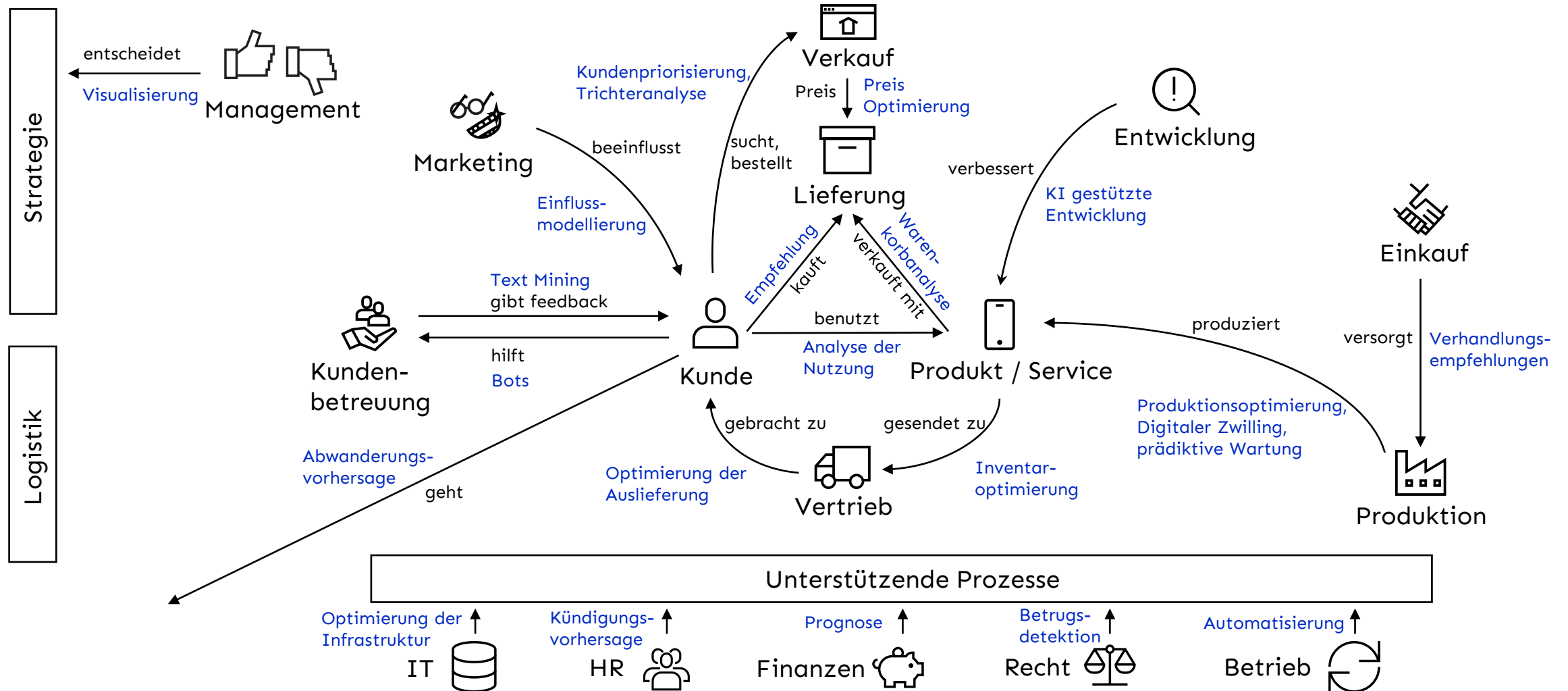
Data Science Use Cases

Dr. André Bojahr
Michael Aydinbas

exeta



Typische Use Cases, Betriebswirtschaftliche Anwendungsgebiet – konkrete Ziele



Aus Daten Mehrwert erzeugen - Geschäftsanwendungen von Data Analytics



Sales & Marketing

- *Churn Reduction*
- Kundengewinnung
- Lead Bewertung
- Produktempfehlungen
- Kampagnenoptimierung
- Kundensegmentierung
- *Next Best Action*



Operations

- *Predictive Maintenance*
- Auslastungsvorhersage
- Bedarfsoptimierung
- Preisoptimierung
- Prozessoptimierung
- Qualitäts Management



Fraud & Risk

- *Fraud Detection*
- *Claims Analyse*
- Credit Score
- Risikobewertung
- Bilanzanalyse



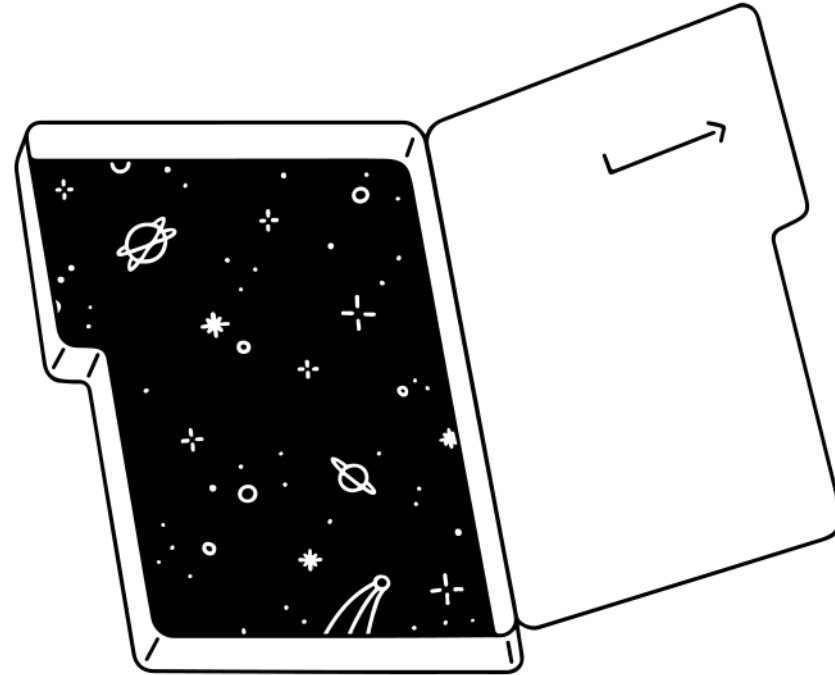
Finance & HR

- *Forecasting*
- Budget Simulationen
- Profitanalyse
- Risikomodellierung

Übung zu Use Cases

1. Usecase Erarbeitung auf Basis der Produkte und Dienstleistungen (intern, extern) bei DHL oder der bekannten/vorhandenen Daten
2. Ausfüllen des Data Science Canvas

PRAXIS



https://miro.com/app/board/uXjVP7h6SRU=
Pass: DHL_2022