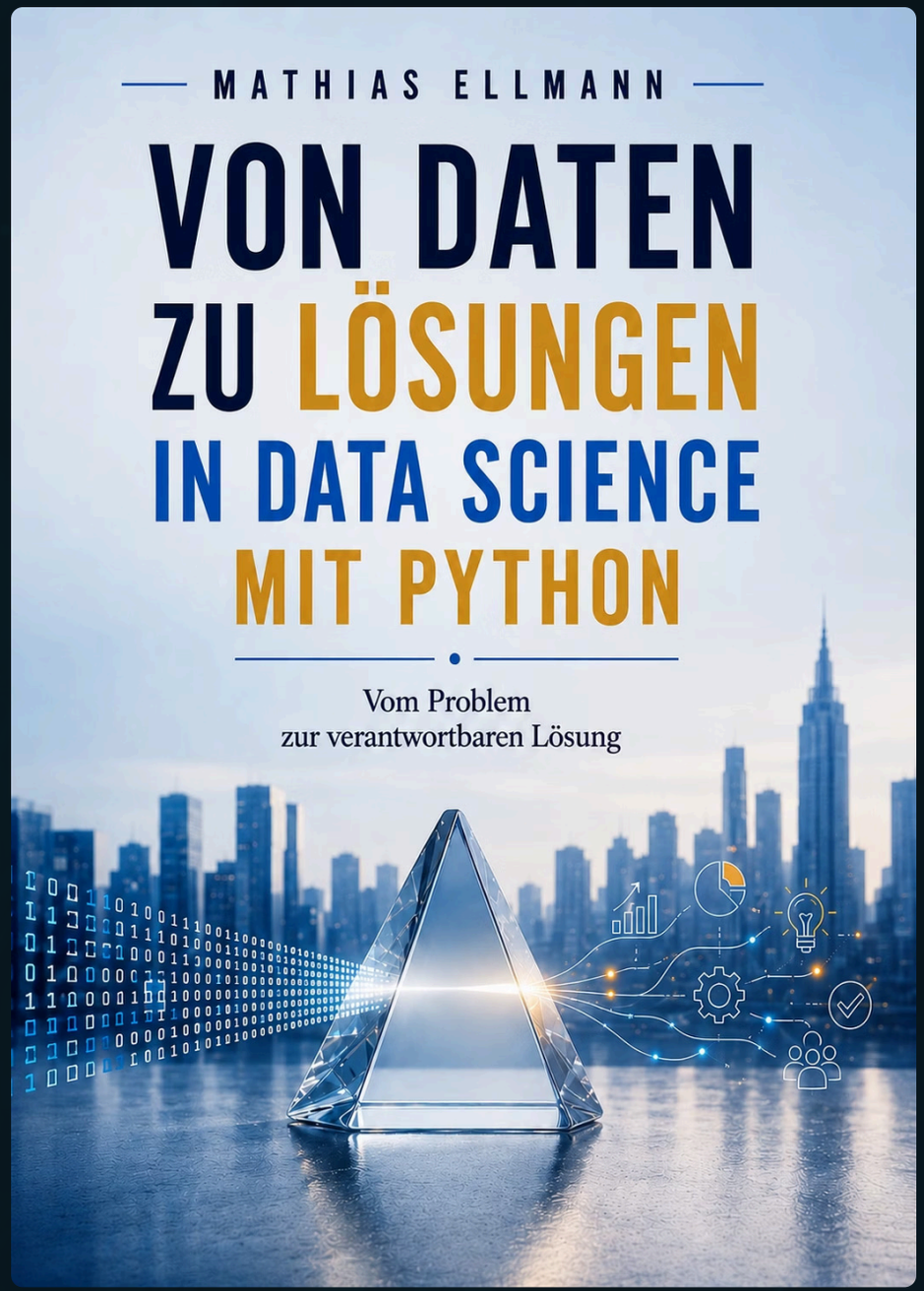


Von Daten zu Lösungen in Data Science mit Python

Vom Problem zur verantwortbaren Lösung

Mathias Ellmann

ISBN: 978-3-6952-6121-5

Warum Data Science heute unverzichtbar ist

Daten wachsen exponentiell

Organisationen erzeugen mehr Daten als je zuvor — doch Daten allein schaffen keinen Wert. Der Rohstoff ist vorhanden; die Fähigkeit, ihn zu nutzen, fehlt häufig.

Entscheidungsdruck steigt

Unternehmen müssen schneller, fundierter und transparenter entscheiden. Datengestützte Entscheidungsprozesse sind kein Vorteil mehr — sie sind Voraussetzung.

Komplexität nimmt zu

Märkte, Prozesse und Systeme werden vielschichtiger. Intuitive Entscheidungen reichen nicht mehr aus, wenn Wechselwirkungen und Abhängigkeiten zunehmen.

Die eigentliche Frage

Nicht: *Wie verarbeiten wir Daten?* Sondern: *Wie lösen wir Probleme mit Daten?* Diese Verschiebung ist der Kern des gesamten Buches.

Das Missverständnis über Data Science

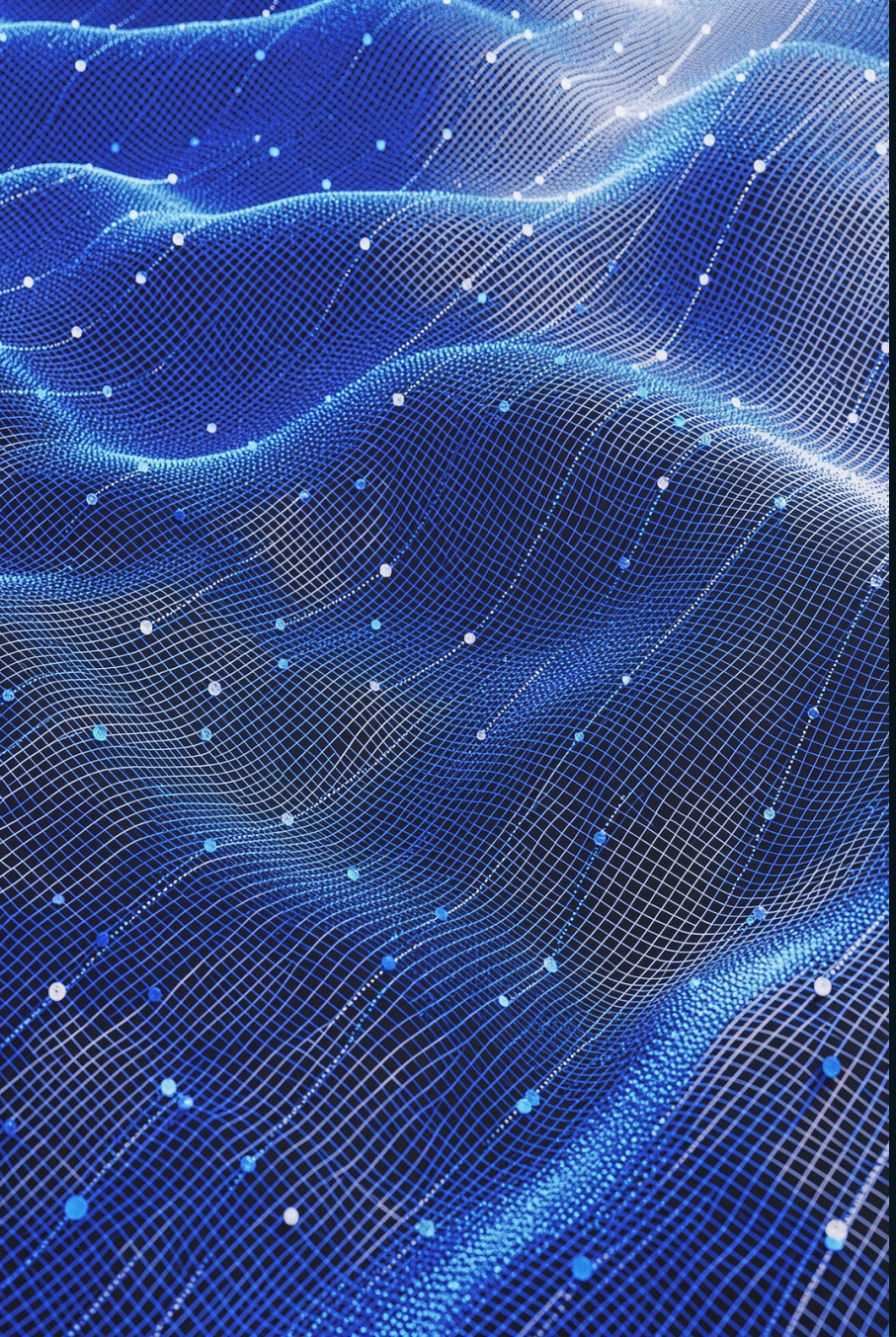
Was viele glauben

- Data Science besteht aus Algorithmen, Modellen, Tools und Technologien
- Wer die besten Bibliotheken beherrscht, löst die besten Probleme
- Mehr Rechenleistung und mehr Daten führen automatisch zu besseren Ergebnissen

Was tatsächlich zählt

Probleme lösen — Kein Unternehmen investiert in Data Science, um Modelle zu trainieren. Der Auftrag ist immer ein geschäftliches oder organisationales Problem.

Viele Projekte scheitern nicht an fehlenden Algorithmen. Sie scheitern an unklaren Fragestellungen, unzureichender Datenqualität, falsch gewählten Zielgrößen, ungeeigneten Metriken oder Missverständnissen zwischen technischen und fachlichen Beteiligten.



Die zentrale These des Buches

> **Data Science ist Problemlösung — nicht Algorithmenoptimierung.**

Daten sind nicht das Ziel. Daten sind der Ausgangspunkt. Lösungen sind das Ziel von Data Science — und der Weg dorthin führt über Erkenntnisse und Entscheidungen, nicht über Modellkomplexität.

Von Daten zu Erkenntnissen

1

Rohdaten

Besitzen keinen unmittelbaren Wert. Erst durch Analyse entstehen Informationen.

2

Informationen

Werden Informationen in Zusammenhänge gebracht, entstehen Erkenntnisse.

3

Erkenntnisse

Nutzbares Wissen, das Entscheidungen ermöglicht und Handlungen begründet.

Beispiel: Millionen Bestellungen eines Onlinehändlers zeigen erst durch Analyse, welche Produkte gemeinsam gekauft werden, welche Kundengruppen aktiv sind und welche Lieferprobleme auftreten. Die Daten lagen vor — der Wert entstand durch Analyse.

Kernaufgabe: Umwandlung von Daten in nutzbares Wissen. Dieser Schritt erfordert analytisches Denken, Domänenwissen und methodische Sorgfalt — nicht nur technische Werkzeuge.

Von Erkenntnissen zu Entscheidungen

Der oft übersehene Schritt

Erkenntnisse allein reichen nicht aus. Zwischen Analyse und Wirkung liegt eine Lücke, die aktiv geschlossen werden muss.

Das Wirkungslosigkeitsproblem

Ein Unternehmen kann wissen, dass eine bestimmte Kundengruppe eine hohe Kündigungswahrscheinlichkeit besitzt. Solange daraus keine Maßnahme entsteht, bleibt dieses Wissen wirkungslos.

Typische Entscheidungsfragen

Welche Kunden ansprechen? Welche Produkte empfehlen? Welche Risiken priorisieren? Diese Fragen verbinden Analyse mit Handlung.

Fazit

Data Science endet nicht bei der Analyse. Ziel ist die Unterstützung von Entscheidungen — und damit die Vorbereitung konkreter Maßnahmen.

Von Entscheidungen zu Lösungen

Entscheidungen ≠ Lösungen

Selbst gute Entscheidungen sind noch keine Lösung — sie müssen umgesetzt werden. Der letzte Schritt im Transformationsprozess ist der schwierigste: die Überführung von Erkenntnissen in veränderte Handlungen.

⚠ Ein Vorhersagemodell für Maschinenausfälle besitzt keinen Nutzen, solange die Vorhersagen nicht in Wartungsprozesse einfließen. Eine Kundensegmentierung bleibt wertlos, wenn Marketingmaßnahmen nicht darauf abgestimmt werden.

Was eine Lösung ausmacht

- **Analyse:** Erkenntnisse aus Daten gewinnen
- **Modellierung:** Muster und Zusammenhänge formalisieren
- **Umsetzung:** Ergebnisse in Prozesse integrieren
- **Verbesserung:** Kontinuierliches Lernen aus der Praxis

Erst wenn Entscheidungen zu veränderten Handlungen führen, entsteht eine Lösung. Data Science ist damit ein vollständiger Problemlösungsprozess — nicht nur ein analytischer Teilschritt.

Die Rolle der Fragestellung

> Der Erfolg eines Projekts hängt häufig stärker von der Qualität der Fragestellung als von der Wahl eines bestimmten Algorithmus ab. Eine unklare Fragestellung führt fast zwangsläufig zu unklaren Ergebnissen.

Schlechte Fragestellung

„Wir wollen Machine Learning einsetzen."

Beschreibt kein Problem. Definiert kein Ziel. Ermöglicht keine Auswahl geeigneter Daten, Merkmale oder Modelle.

Gute Fragestellung

„Wie können wir frühzeitig erkennen, welche Kunden kündigen werden?"

Definiert ein konkretes Problem. Ermöglicht die Auswahl geeigneter Daten, Merkmale und Modelle. Schafft Grundlage für messbare Ergebnisse.

Gute Fragen führen zu besseren Lösungen. Präzise Fragestellungen sind der Schlüssel — nicht die Wahl des Algorithmus.

Probleme verstehen — der Ausgangspunkt jedes Projekts

Kunden kündigen

Häufiger als erwartet — warum? Welche Muster liegen vor? Welche Maßnahmen sind möglich?

Maschinen fallen aus

Ungeplant und kostspielig — wann? Welche Signale gehen voraus? Wie früh ist Erkennung möglich?

Betrugsfälle

Verursachen hohe Kosten — wie erkennen? Welche Merkmale unterscheiden Betrug von normalem Verhalten?

Grundprinzip: Data-Science-Projekte beginnen selten mit Daten. In den meisten Fällen steht zunächst ein Problem oder ein Wunsch nach Verbesserung im Mittelpunkt. Erst wenn das Problem verstanden ist, stellt sich die Frage, ob Daten helfen können — und welche.

Probleme verändern sich im Projektverlauf. Flexibilität und kontinuierliche Rückkopplung zwischen Problemverständnis und Datenlage sind notwendig.

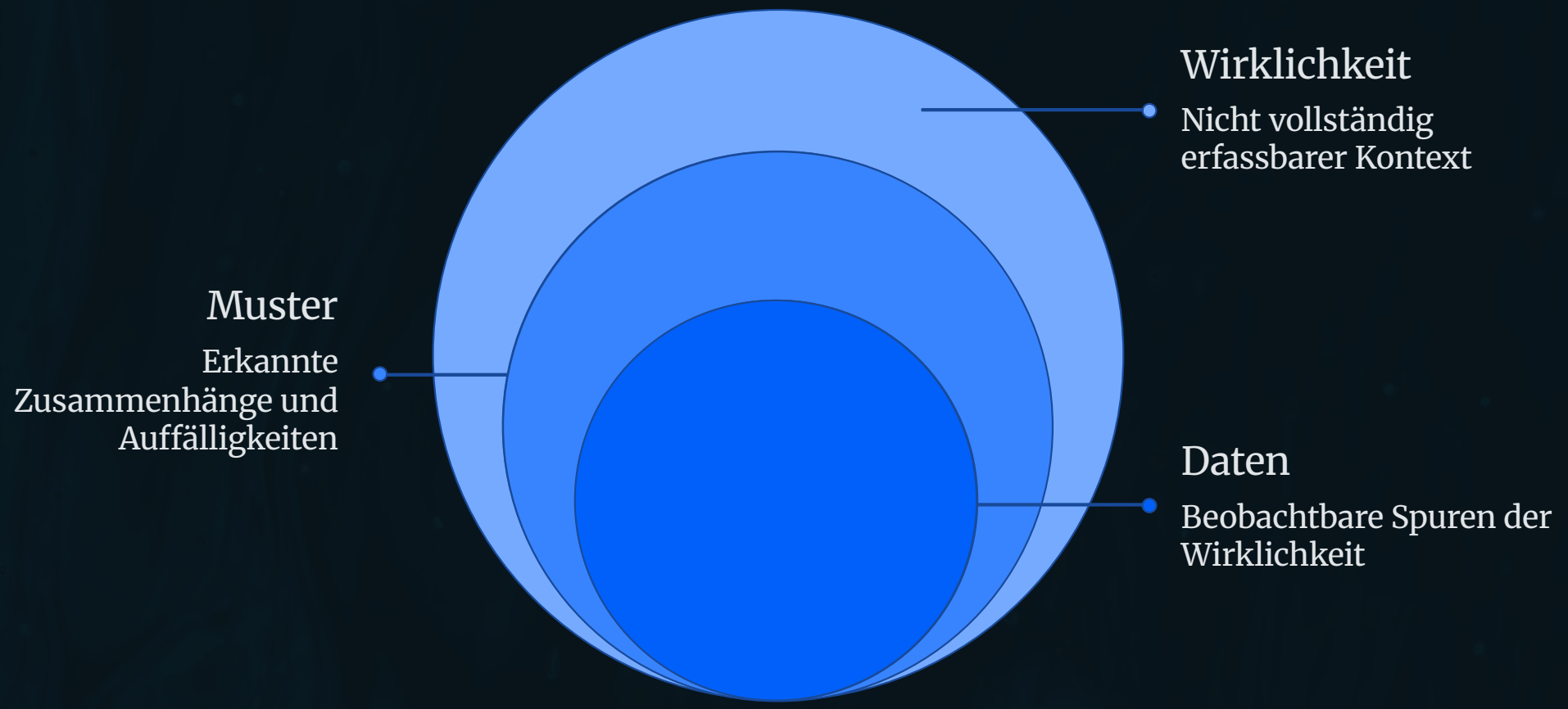
Daten als Spuren der Wirklichkeit

Was Daten zeigen

- Hinweise auf Zusammenhänge, Muster und Auffälligkeiten
- Grundlage für Hypothesen und Analysen
- Beobachtbare Ausschnitte eines komplexen Systems

Was Daten nicht zeigen

- **Ursachen und Kausalitäten** — Korrelation ist keine Kausalität
- **Den vollständigen Kontext** — Daten sind immer eine Vereinfachung der Realität
- **Die richtige Entscheidung** — das bleibt menschliche Aufgabe



Daten sind Spuren der Wirklichkeit — keine vollständige Abbildung. Wer Daten als objektive Wahrheit behandelt, übersieht die Grenzen jeder Messung und jedes Erhebungsprozesses.

Die Gefahr vorschneller Interpretationen

Daten sprechen nicht für sich selbst

Sie müssen interpretiert werden. Nicht jede Auffälligkeit in Daten ist ein Fehler — und nicht jeder Zusammenhang ist bedeutsam.

Bestätigungsfehler

Wer eine Hypothese sucht, findet sie — auch in zufälligen Mustern. Selektive Wahrnehmung ist ein systematisches Risiko in der Datenanalyse.

Goodharts Gesetz

Wenn eine Kennzahl zum Ziel wird, hört sie auf, eine gute Kennzahl zu sein. Optimierung auf Metriken kann das eigentliche Ziel verfehlen.

Reflexion als Qualitätsmerkmal

Kritisches Denken schützt vor Fehlschlüssen. Die Fähigkeit, eigene Interpretationen zu hinterfragen, ist eine Kernkompetenz in Data Science.

Datenqualität als Grundlage

Kernthese: Datenqualität ist keine technische Nebensache — sie ist eine Diagnoseaufgabe. Qualitätsprobleme müssen verstanden, dokumentiert und begründet behandelt werden.

Fehlende Werte

Warum fehlen sie? Zufällig, systematisch oder strukturell? Die Antwort bestimmt die Behandlungsstrategie.

Dubletten

Mehrfacheinträge verzerren Analysen und Modelle. Erkennung und Bereinigung erfordern klare Kriterien.

Ausreißer

Fehler oder echte Ausnahmen? Die Antwort verändert die Lösung grundlegend.

Dokumentation

Entscheidungen über Daten müssen nachvollziehbar sein. Qualitätsprobleme dokumentieren ist Pflicht, nicht Option.

ETL als Problemlösung

> **ETL ist kein technischer Prozess — ETL ist ein Denkprozess.** Jede Transformation ist eine Entscheidung — und muss begründet werden.

1

Extract

Welche Datenquellen sind geeignet? Welche Daten fehlen? Welche Qualitätsanforderungen gelten?

2

Transform

Wie werden Daten in eine nutzbare Form gebracht? Welche Entscheidungen sind dabei nötig? Welche Annahmen werden getroffen?

3

Load

Wie werden Ergebnisse bereitgestellt und reproduzierbar gemacht? Wer hat Zugriff, wann und in welcher Form?

ETL ist die Grundlage aller weiteren Schritte. Ein kontinuierlicher ETL-Prozess sichert Reproduzierbarkeit und schafft die Voraussetzung für nachvollziehbare Analysen.

Von Rohdaten zu Informationen

01

Diagnostischer Blick

Der erste Blick auf einen Datensatz ist immer ein diagnostischer Blick: Was ist vorhanden? Was fehlt? Was ist auffällig?

02

Datenbereinigung mit Python

pandas als Werkzeug zur systematischen Aufbereitung. Datentypen, Formate, Kategorien: Jede Korrektur ist eine Entscheidung mit Konsequenzen.

03

Reproduzierbarkeit als Ziel

Bereinigungsschritte müssen dokumentiert und automatisierbar sein. Rohdaten sind selten sauber — sie spiegeln die Realität von Prozessen wider.

Rohdaten sind selten sauber. Sie spiegeln die Realität von Prozessen wider — mit all ihren Unvollkommenheiten, Ausnahmen und Inkonsistenzen. Bereinigung ist keine Korrektur der Realität, sondern eine begründete Entscheidung über die Darstellung.

Feature Engineering — Hypothesen in Merkmale übersetzen

Kernprinzip

Features sind keine neutralen Datenpunkte — sie sind Hypothesen.

Feature Engineering fragt: Welche Merkmale enthalten entscheidungsrelevante Informationen? Diese Frage erfordert Domänenwissen, nicht nur technisches Können.

Domänenwissen ist entscheidend — Modelle lernen nicht von selbst, was fachlich relevant ist.

Typische Feature-Typen

- **Zeitmerkmale:** Wochentag, Monat, Saison, Zeitabstand
- **Verhältnisgrößen:** Quoten, Raten, Normierungen
- **Aggregationen:** Summen, Mittelwerte, Maxima über Zeitfenster
- **Kaufhistorien:** Häufigkeit, Recency, Monetary Value

⚠ **Zu viele Features können schaden.** Qualität vor Quantität — irrelevante Merkmale erhöhen Rauschen und Overfitting-Risiko.

Features als Hypothesen

Jedes Merkmal enthält eine Annahme

Feature-Ideen entstehen aus Fragestellungen, Domänenwissen und Hypothesen — nicht aus Daten allein. Wer Features baut, trifft Entscheidungen über die Wirklichkeit.

Datenleckage als Risiko

Features, die zukünftige Informationen enthalten, erzeugen täuschend gute Modelle — die in der Praxis versagen. Datenleckage ist eine der häufigsten und gefährlichsten Fehlerquellen.

Feature-Qualität bewerten

Fachliche Relevanz, Verteilung, Zusammenhang mit der Zielgröße — drei Dimensionen der Qualitätsbewertung, die vor Modellierung geprüft werden müssen.

Dokumentation von Features

Annahmen sichtbar machen. Welche Hypothese steckt hinter diesem Merkmal? Warum wurde es aufgenommen? Dokumentation schützt vor stillen Fehlern.

Vom Problem zum Modell

> **Modelle lösen keine Geschäftsprobleme — Menschen lösen Geschäftsprobleme.** Modelle sind Werkzeuge, die diesen Prozess unterstützen.

Typische Fehler

- Modell entwickeln, bevor das Problem verstanden wurde
- Zielvariable ohne fachliche Begründung wählen
- Klassifikation oder Regression nach Präferenz, nicht nach Problem

Richtiger Ansatz

Von der Geschäftsfrage zur Datenfrage: Was soll vorhergesagt werden? Welche Zielvariable ist sinnvoll? Die Wahl hängt vom Problem ab — nicht von der Methode.

Modelle als Werkzeuge — nicht als Wahrheit

Modelle sind Vereinfachungen

Sie bilden Realität ab, ersetzen sie nicht. Kein Modell erfasst alle relevanten Einflussfaktoren — das ist keine Schwäche, sondern eine Eigenschaft.

Komplexität ≠ Nutzen

Mehr Komplexität bedeutet nicht automatisch mehr Nutzen. Ein einfaches, erklärbares Modell kann wertvoller sein als ein hochgenaues Black-Box-Modell.

Interpretierbarkeit

Ein erklärbares Modell kann Vertrauen schaffen, das ein Black-Box-Modell nicht kann. Interpretierbarkeit ist ein Entscheidungskriterium — kein Kompromiss.

Modelle als Hypothesen

Jedes Modell ist ein Experiment — und muss als solches dokumentiert werden. Modellauswahl ist Problemlösung, nicht Optimierung einer Kennzahl.

Baselines schützen vor Selbsttäuschung

Eine Baseline ist der ehrlichste Test eines Modells. Ohne Baseline ist unklar, ob ein Modell tatsächlich besser ist als eine einfache Regel.

Warum Baselines unverzichtbar sind

Ohne Vergleichsmaßstab ist jede Modellleistung bedeutungslos. Eine Accuracy von 90 % klingt gut — bis man feststellt, dass die naive Mehrheitsklasse 92 % erreicht.

Eine schlechte Baseline ist besser als keine

Selbst eine einfache Heuristik schafft Orientierung. Baselines machen sichtbar, was ein Modell tatsächlich leistet — und was nicht.

Fachliche Baselines

Manchmal ist das Expertenwissen die stärkste Baseline. Wenn ein erfahrener Mitarbeiter mit einfachen Regeln bessere Ergebnisse erzielt als ein Modell, ist das ein wichtiges Signal.

Ockhams Rasiermesser

Einfache Lösungen werden systematisch unterschätzt. Das Prinzip gilt auch in Data Science: die einfachste Lösung, die das Problem löst, ist oft die beste.

Metriken als Entscheidungshilfen

Das Metrik-Dilemma

Eine Metrik beantwortet nie alle Fragen — sie beleuchtet einen Aspekt der Modellleistung. Metriken und Projektziele müssen übereinstimmen: technische Güte ≠ fachlicher Nutzen.

> ⚠ **Goodharts Gesetz:** Wenn eine Kennzahl zum Ziel wird, hört sie auf, eine gute Kennzahl zu sein. Optimierung auf eine einzelne Metrik kann das eigentliche Ziel verfehlen.

Wichtige Metriken im Überblick

Metrik	Stärke	Schwäche
Accuracy	Einfach interpretierbar	Irreführend bei Klassenungleichgewicht
Precision	Fokus auf False Positives	Ignoriert False Negatives
Recall	Fokus auf False Negatives	Ignoriert False Positives
F1-Score	Balance aus P und R	Kein Kontext zu Kosten
ROC-AUC	Schwellenwert-unabhängig	Schwer kommunizierbar

Mehrere Metriken kombinieren — kein einzelner Wert entscheidet. Die Wahl der Metrik ist eine fachliche Entscheidung, keine technische.

Trade-offs verstehen

> **Jede Modellentscheidung ist ein Trade-off** — es gibt keine optimale Lösung ohne Kontext.

Interpretierbarkeit vs. Genauigkeit

Ein erklärbares Modell kann Vertrauen schaffen, das ein Black-Box-Modell nicht kann. Vertrauen hat einen Wert — auch wenn er sich nicht in Accuracy messen lässt.

Trainingszeit vs. Modellqualität

Ressourcen sind begrenzt. Mehr Rechenzeit führt nicht immer zu besseren Ergebnissen — der Grenznutzen nimmt ab.

Kosten von Fehlern

False Positives und False Negatives haben unterschiedliche Konsequenzen — je nach Anwendungsfall. Ein Krebstest und ein Spamfilter haben entgegengesetzte Prioritäten.

Pareto-optimal denken

Keine Lösung dominiert in allen Dimensionen. Die beste Lösung ist die, die den relevanten Trade-off am besten adressiert.

Unsicherheit als Normalzustand

Kernthese

Prognosen liefern niemals Gewissheit — nur Wahrscheinlichkeiten.

Modelle sind Vereinfachungen: Sie können nicht alles erfassen, was in der Realität geschieht. Unsicherheit ist kein Versagen — sie ist der Normalzustand.

Unsicherheit kommunizieren ist professionell — nicht ein Zeichen von Schwäche.

Quellen von Unsicherheit

- **Datenqualität:** Fehlende, verzerrte oder unrepräsentative Daten
- **Modellgrenzen:** Vereinfachungen und Annahmen im Modell
- **Verteilungsverschiebungen:** Die Welt verändert sich nach dem Training
- **Unbekannte Einflussfaktoren:** Was nicht gemessen wird, kann nicht modelliert werden

Szenarien statt Punktvorhersagen — Bandbreiten zeigen Realität besser als einzelne Zahlen. Entscheidungen unter Unsicherheit erfordern Szenarien, nicht Gewissheit.

Risiken sichtbar machen

Datenrisiken

Verzerrungen, fehlende Repräsentativität, historische Fehler. Wenn Trainingsdaten vergangene Diskriminierung widerspiegeln, lernt das Modell diese Diskriminierung.

Modellrisiken

Überanpassung, Datenleckage, instabile Features. Modelle, die auf Trainingsdaten perfekt funktionieren, können in der Praxis versagen.

Betriebsrisiken

Data Drift, Concept Drift, technische Schulden. Die Welt verändert sich — Modelle, die nicht überwacht werden, veralten still.

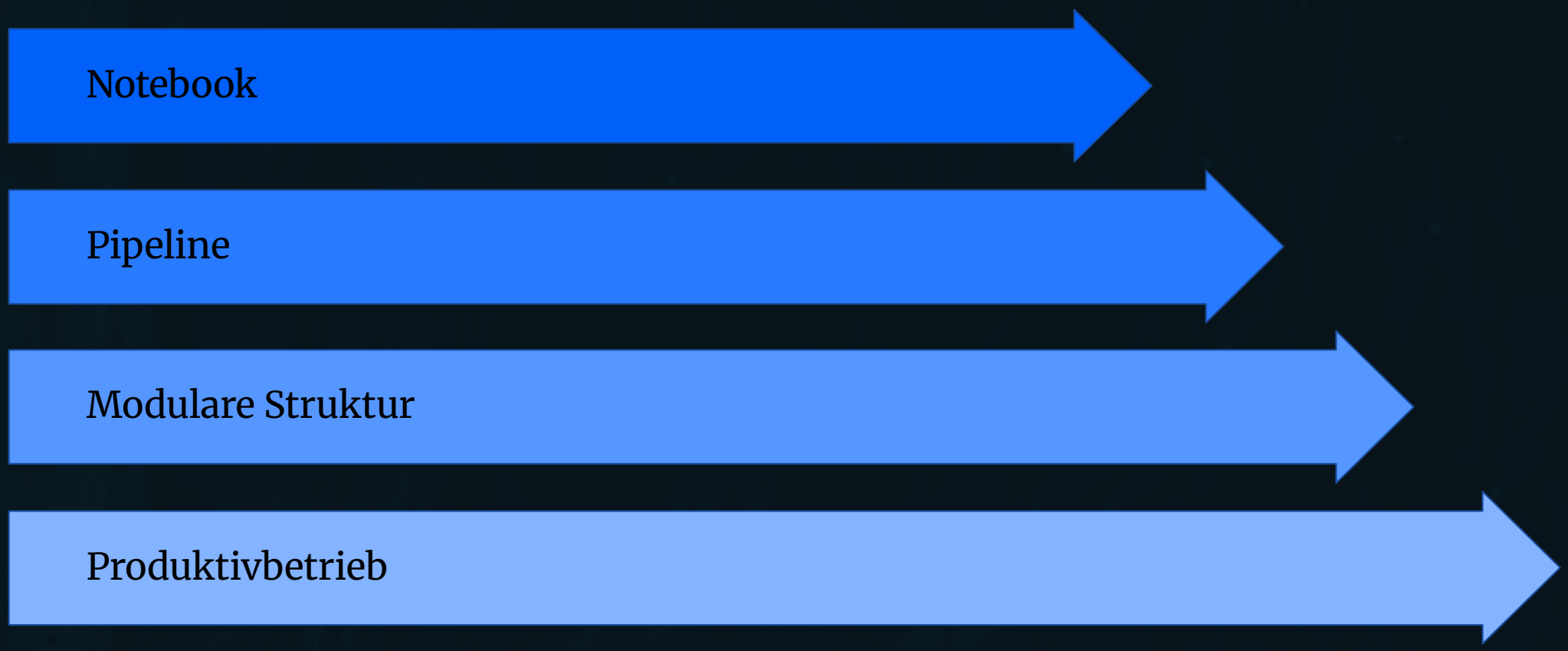
Ethische Risiken

Diskriminierung, mangelnde Erklärbarkeit, Verantwortungsdiffusion. Wer entscheidet, wenn ein Algorithmus entscheidet? Diese Frage muss vor der Umsetzung beantwortet werden.

⊗ **Risiken gehören zu jeder Data-Science-Lösung — sie müssen benannt werden.** Unbenannte Risiken sind keine abwesenden Risiken.

Vom Notebook zur produktiven Lösung

> **Ein Notebook ist kein Produkt — es ist ein Experiment.** Der Weg von der Analyse zur produktiven Lösung erfordert strukturierte Entwicklung.



Typische Probleme großer Notebooks

- Fehlende Reproduzierbarkeit
- Schwer wartbarer Code
- Keine Versionierung
- Inkonsistenz zwischen Training und Produktivbetrieb

Lösungsansätze

- **Pipelines:** Konsistenz zwischen Training und Produktivbetrieb sicherstellen
- **Modulare Strukturen:** Wiederverwendbarkeit und Testbarkeit
- **Versionierung:** Code und Daten nachvollziehbar halten
- **Reproduzierbarkeit** ist ein Qualitätsmerkmal — nicht ein optionales Feature

Monitoring und Lernen

Data Drift

Die Verteilung der Eingabedaten verändert sich im Laufe der Zeit. Ein Modell, das auf historischen Daten trainiert wurde, kann auf veränderte Realität nicht automatisch reagieren.

Concept Drift

Die Beziehung zwischen Features und Zielgröße verändert sich. Was gestern ein gutes Vorhersagemerkmal war, kann heute irrelevant sein.

Monitoring als Grundlage

Monitoring ist kein Selbstzweck — es ist die Grundlage für kontinuierliche Verbesserung. Welche Kennzahlen überwacht werden, ist eine fachliche Entscheidung.

Lernen aus der Umsetzung

Fehler sind Lernquellen — nicht Versagen. Lösungen veralten — Daten und Realität verändern sich. Kontinuierliches Lernen ist Teil des Prozesses.

Ergebnisse verständlich kommunizieren

Kommunikation ist Teil der Problemlösung

Nicht ihr Anhang. Wer Ergebnisse nicht kommunizieren kann, hat keine Lösung — er hat eine Analyse.

Von Daten zu Entscheidungen: Ergebnisse müssen in Alltagssprache übersetzt werden. Von Entscheidungen zu Handlungen: Handlungsorientierte Kommunikation schafft Wirkung.

Vertrauen durch Nachvollziehbarkeit — Unsicherheit sichtbar machen schafft mehr Vertrauen als falsche Gewissheit.

Zielgruppengerechte Kommunikation

Zielgruppe	Bedarf
Management	Entscheidungsrelevanz, Risiken, Handlungsoptionen
Fachbereich	Praktische Implikationen, Einschränkungen, Anwendungsfälle
Technisches Team	Methodik, Metriken, Implementierungsdetails
Regulatoren	Nachvollziehbarkeit, Dokumentation, Fairness

Der Informationseisberg

Ein zentrales Framework nach Karpman im Buch: vier Ebenen vollständiger Kommunikation. Wer nur Daten präsentiert, kommuniziert unvollständig.

Punkt über der Wasseroberfläche

Die sichtbare Aussage, Kennzahl oder Feststellung — verständlich erst durch Information, Bedeutung und Absicht.

Information

Daten im Zusammenhang — was zeigt sich? Muster, Trends, Auffälligkeiten.

Bedeutung

Was bedeutet das für unser Problem? Welche Schlüsse sind zulässig?

Absicht

Was soll entschieden oder verändert werden? Die handlungsrelevante Ebene.

Verantwortung in Data Science

> **Technologie unterstützt — Menschen entscheiden — Menschen tragen Verantwortung.**

Verantwortung beginnt früh

Verantwortung beginnt bei der Fragestellung — nicht erst bei der Umsetzung. Wer die falsche Frage stellt, trägt Verantwortung für die Konsequenzen.

Diskriminierung durch Daten

Modelle können diskriminieren, wenn Daten verzerrt sind. Verantwortung liegt beim Menschen, nicht beim Algorithmus — auch wenn der Algorithmus entscheidet.

Nachvollziehbarkeit als Pflicht

Entscheidungen, die auf Modellen basieren, müssen erklärbar sein. Nachvollziehbarkeit ist eine ethische Pflicht — nicht ein technisches Feature.

Verantwortbare Lösung

= wirksam + nachvollziehbar + transparent + begründet. Alle vier Dimensionen müssen erfüllt sein — keine ist optional.

Was Leser aus dem Buch mitnehmen

Systematisches Problemlösen

Ein Denkrahmen für den gesamten Data-Science-Prozess — von der Fragestellung bis zur verantwortbaren Lösung.

Bessere Analysen

Durch präzise Fragestellungen, saubere Daten und begründete Entscheidungen — nicht durch mehr Algorithmen.

Bessere Entscheidungen

Durch Metriken, Trade-off-Analyse und Risikobetrachtung. Entscheidungen, die auf Evidenz basieren und Unsicherheit einschließen.

Mehr Transparenz

Durch Dokumentation, Reproduzierbarkeit und Nachvollziehbarkeit. Transparenz als Qualitätsmerkmal, nicht als Bürde.

Mehr Verantwortung

Durch den Blick auf ethische Risiken und die Grenzen von Modellen. Verantwortung als integraler Bestandteil des Prozesses.

Praktische Python-Kompetenz

Als Werkzeug, nicht als Selbstzweck. Python im Dienst der Problemlösung — nicht als Ziel an sich.

Data Science beginnt mit Problemen

Nicht Algorithmen

Data Science beginnt nicht mit Algorithmen. Data Science beginnt mit Problemen.

Nicht Daten allein

Daten allein lösen keine Probleme. Modelle allein schaffen keinen Nutzen.

Entscheidungen

Erst durch Entscheidungen entstehen Lösungen. Der Weg führt über Erkenntnisse, Entscheidungen und Verantwortung.

> Gute Data Science entsteht nicht allein durch bessere Algorithmen. Sie entsteht durch gute Fragen, saubere Daten, begründete Entscheidungen, nachvollziehbare Prozesse und die Fähigkeit, technische Ergebnisse in praktische Wirkung zu übersetzen.

Quellen und theoretische Grundlagen I

Data Science als Problemlösung und Erkenntnisprozess

Fayyad, Usama; Piatetsky-Shapiro, Gregory; Smyth, Padhraic: From Data Mining to Knowledge Discovery in Databases. AI Magazine 17(3), 1996, S. 37–54.

Grundlage für die Einordnung von Data Science als Prozess, in dem aus Daten schrittweise Wissen entsteht.

Chapman, Pete; Clinton, Julian; Kerber, Randy; Khabaza, Thomas; Reinartz, Thomas; Shearer, Colin; Wirth, Rüdiger: CRISP-DM 1.0. Step-by-step Data Mining Guide. 2000.

Grundlage für den prozessorientierten Blick auf Data-Science-Projekte: vom Geschäftsverständnis über Datenverständnis und Modellierung bis zur Umsetzung.

Provost, Foster; Fawcett, Tom: Data Science for Business. O'Reilly Media, Sebastopol, 2013.

Grundlage für die Perspektive, dass Data Science wirtschaftliche und organisatorische Entscheidungen vorbereitet und deshalb vom Problem her gedacht werden muss.

Quellen und theoretische Grundlagen II

Datenqualität, Feature Engineering und Python

Olson, Jack E.: Data Quality. The Accuracy Dimension. Morgan Kaufmann, San Francisco, 2003.

Grundlage für die Bedeutung von Datenqualität als Voraussetzung belastbarer Analysen, Modelle und Entscheidungen.

McKinney, Wes: Python for Data Analysis. Data Wrangling with pandas, NumPy, and Jupyter. 3. Auflage. O'Reilly Media, Sebastopol, 2022.

Grundlage für die praktische Datenaufbereitung mit Python, insbesondere für Bereinigung, Transformation und explorative Analyse.

Kuhn, Max; Johnson, Kjell: Feature Engineering and Selection. A Practical Approach for Predictive Models. CRC Press, Boca Raton, 2019.

Grundlage für Feature Engineering als methodischen Schritt, in dem aus Rohdaten entscheidungsrelevante Merkmale entwickelt und bewertet werden.

Quellen und theoretische Grundlagen III

Modelle, Metriken und Entscheidungen unter Unsicherheit

Hastie, Trevor; Tibshirani, Robert; Friedman, Jerome: The Elements of Statistical Learning. Data Mining, Inference, and Prediction. 2. Auflage. Springer, New York, 2009.

Grundlage für Modellbildung, Generalisierung, Overfitting und die kritische Bewertung statistischer Lernverfahren.

Goodhart, C. A. E.: Problems of Monetary Management: The UK Experience. In: Monetary Theory and Practice: The UK Experience. Macmillan Education UK, London, 1984, S. 91–121. ISBN: 978-1-349-17295-5. DOI: 10.1007/978-1-349-17295-5_4.

Grundlage für Goodharts Gesetz: Wenn eine Kennzahl zum Ziel wird, verliert sie ihre Aussagekraft.

Hammond, John S.; Keeney, Ralph L.; Raiffa, Howard: Smart Choices. A Practical Guide to Making Better Decisions. Harvard Business Review Press, Boston, 2015.

Grundlage für das PrOACT-Framework zur strukturierten Entscheidungsfindung unter Unsicherheit.

Quellen und theoretische Grundlagen IV

Kommunikation, Verantwortung und Informationseisberg

Karpman, Stephen B.: Ein Leben ohne Spiele. Die neue Transaktionsanalyse der Vertrautheit, der Offenheit und der Zufriedenheit. Process Training und Consulting, Weilheim, 2016. ISBN: 978-3-937471-03-7.

Grundlage für den Informationseisberg mit den vier Ebenen Punkt, Information, Bedeutung und Absicht.

Toulmin, Stephen: The Uses of Argument. Cambridge University Press, Cambridge, 1958.

Grundlage für die argumentative Struktur datenbasierter Empfehlungen: Daten werden erst durch Begründungen und Schlussfolgerungen zu tragfähigen Entscheidungsvorlagen.

O’Neil, Cathy: Weapons of Math Destruction. Crown, New York, 2016.

Grundlage für die gesellschaftlichen Risiken datenbasierter Entscheidungen, insbesondere bei intransparenten, unfairen oder folgenreichen Modellen.