

DATA SCIENCE 1

VORLESUNG 2 – VOM BUSINESS-PROBLEM
ZUR DATENFRAGE

PROF. DR. CHRISTIAN BOCKERMANN

HOCHSCHULE BOCHUM

WINTERSEMESTER 2026/2027

RÜCKBLICK ÜBUNG 1

Bin ich besser geworden?

Was war beim Radtraining das eigentliche Problem?

Chat: Ihre Antwort?



Übung 1: Ihre vier Felder

Ziel	Verfügbare Daten
Offene Fragen	Wen müsste man fragen?

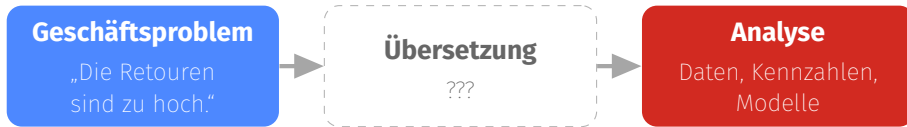


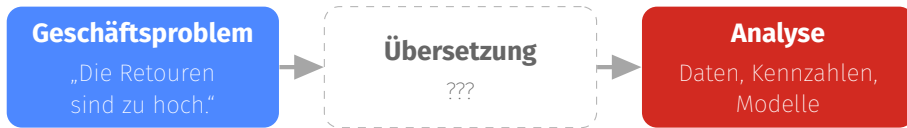
Übung 1: Ihre vier Felder

Ziel	Verfügbare Daten
Offene Fragen	Wen müsste man fragen?

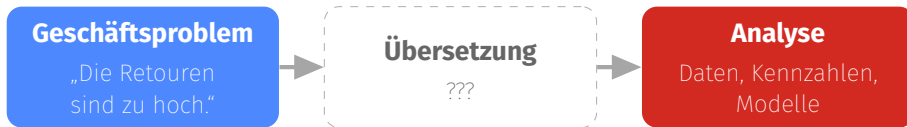
Die Daten waren vollständig da – die Frage war trotzdem **nicht beantwortbar**.

DAS ÜBERSETZUNGSPROBLEM





- Das Management spricht in **Problemen und Zielen**.
- Daten und Algorithmen brauchen **präzise Fragen und Definitionen**.
- Dazwischen liegt eine Übersetzung – und die ist **der schwierigste Schritt**.



- Das Management spricht in **Problemen und Zielen**.
- Daten und Algorithmen brauchen **präzise Fragen und Definitionen**.
- Dazwischen liegt eine Übersetzung – und die ist **der schwierigste Schritt**.

Genau das ist Ihre Rolle als **BWLER im Datenprojekt**.

Problem

Was stört?

Geschäftsziel

Was soll besser werden?

Entscheidung

Was würden wir anders machen?

Analysefrage

Was müssten wir dafür wissen?

Zielgröße

Was genau wollen wir messen oder vorhersagen?

Daten

Was wissen wir **zum Zeitpunkt der Entscheidung?**

Erfolgskriterium

Woran merken wir, dass es geklappt hat?

Problem

„Ich trainiere viel, aber komme nicht richtig voran.“

Geschäftsziel

Beim Radmarathon im Mai schneller sein als im letzten Jahr

Entscheidung

???

Chat: Was ist hier die Entscheidung?

Problem

„Ich trainiere viel, aber komme nicht richtig voran.“

Geschäftsziel

Beim Radmarathon im Mai schneller sein als im letzten Jahr

Entscheidung

Trainingsplan für die nächsten Wochen anpassen: mehr Intensität oder mehr Erholung?

Problem

„Ich trainiere viel, aber komme nicht richtig voran.“

Geschäftsziel

Beim Radmarathon im Mai schneller sein als im letzten Jahr

Entscheidung

Trainingsplan für die nächsten Wochen anpassen: mehr Intensität oder mehr Erholung?

Analysefrage

Bin ich ausgeruht genug, um mich noch zu steigern – oder übertrainiert?

Zielgröße

z. B. Leistung in Watt bei gleichem Puls

Problem	„Ich trainiere viel, aber komme nicht richtig voran.“
Geschäftsziel	Beim Radmarathon im Mai schneller sein als im letzten Jahr
Entscheidung	Trainingsplan für die nächsten Wochen anpassen: mehr Intensität oder mehr Erholung?
Analysefrage	Bin ich ausgeruht genug, um mich noch zu steigern – oder übertrainiert?
Zielgröße	z. B. Leistung in Watt bei gleichem Puls
Daten	Leistung, Puls, Schlaf der <i>letzten</i> Wochen – vor der nächsten Trainingswoche
Erfolgskriterium	Zeit beim Radmarathon; vorher: Leistung bei gleichem Puls steigt

1. **Erst die Entscheidung, dann die Daten.**

Eine Analyse ohne Entscheidung ist ein Bericht, kein Projekt.

1. **Erst die Entscheidung, dann die Daten.**

Eine Analyse ohne Entscheidung ist ein Bericht, kein Projekt.

2. **Begriffe müssen definiert werden.**

„besser“, „Retoure“, „Kunde“ – was genau ist gemeint?

1. **Erst die Entscheidung, dann die Daten.**

Eine Analyse ohne Entscheidung ist ein Bericht, kein Projekt.

2. **Begriffe müssen definiert werden.**

„besser“, „Retoure“, „Kunde“ – was genau ist gemeint?

3. **Es zählt nur, was man zum Zeitpunkt der Entscheidung weiß.**

Wer heute entscheidet, kennt die Zukunft nicht – auch nicht in den Daten.

DAS MACHINE LEARNING CANVAS

The Machine Learning Canvas (v0.4) Designed for: _____ Designed by: _____ Date: _____ Iteration: _____

Decisions  How are predictions used to make decisions that provide the proposed value to the end-user?	ML task  Input, output to predict, type of problem.	Value Propositions  What are we trying to do for the end-user(s) of the predictive system? What objectives are we serving?	Data Sources  Which raw data sources can we use (internal and external)?	Collecting Data  How do we get new data to learn from (inputs and outputs)?
Making Predictions  When do we make predictions on new inputs? How long do we have to featurize a new input and make a prediction?	Offline Evaluation  Methods and metrics to evaluate the system before deployment.	Building Models  When do we create/update models with new training data? How long do we have to featurize training inputs and create a model?		
Live Evaluation and Monitoring  Methods and metrics to evaluate the system after deployment, and to quantify value creation.				

machinelearningcanvas.com by Louis Dorard, Ph.D. Licensed under a Creative Commons Attribution-ShareAlike 4.0 International License 

Feld im Canvas	Frage	Übersetzungskette
Value Propositions	Welchen Nutzen soll das System stiften?	Problem, Ziel
Decisions	Wie werden Vorhersagen zu Entscheidungen?	Entscheidung
ML Task	Was ist Eingabe, was wird vorhergesagt?	Analysefrage, Zielgröße
Data Sources	Welche Datenquellen gibt es?	Daten
Features	Welche Merkmale leiten wir ab?	Daten
Making Predictions	Wann wird vorhergesagt?	Zeitpunkt
Offline Evaluation	Wie prüfen wir das Modell vorab?	Erfolgskriterium
Live Evaluation	Wie messen wir den Nutzen im Betrieb?	Erfolg → KPI
Collecting Data, Building Models	Neue Daten, neu trainieren	(technisch, später)

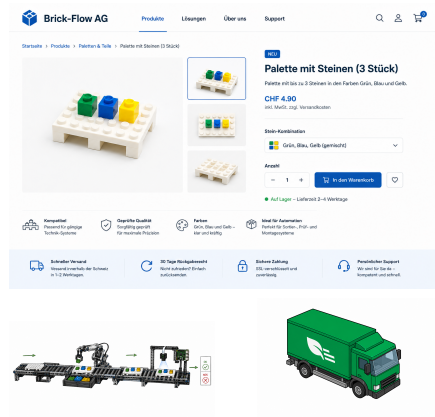
- In der Mitte steht **Value Propositions**: Wozu das Ganze?
- Links davon die **Anwendung**: Entscheidungen, Vorhersagen, Erfolgsmessung
- Rechts davon das **Lernen**: Datenquellen, Merkmale, Modelle

- In der Mitte steht **Value Propositions**: Wozu das Ganze?
- Links davon die **Anwendung**: Entscheidungen, Vorhersagen, Erfolgsmessung
- Rechts davon das **Lernen**: Datenquellen, Merkmale, Modelle

Der Algorithmus ist nur **ein** Feld von zehn.
Die **KPI** steht in **Live Evaluation and Monitoring** –
daran wird später der Erfolg einer Datenstrategie
gemessen.

USE CASE: RETOUREN BEI DER BRICK- FLOW AG

- Fertigt Klemmbausteine in Paletten
- Fertige Sets und Individual-Paletten mit freiem Farbmix
- Verkauf über den eigenen Online-Shop
- Endkunden (B2C) und Fachhandel (B2B)
- Versand über einen Logistik-Dienstleister





Der Vorstand ist alarmiert:

Die Retourenkosten steigen.

„Unsere Retouren kosten uns zu viel Geld. Verstehen Sie das Problem, sagen Sie uns, wie wir es messen sollen – und schlagen Sie vor, was wir tun können.“

„Unsere Retouren kosten uns zu viel Geld. Verstehen Sie das Problem, sagen Sie uns, wie wir es messen sollen – und schlagen Sie vor, was wir tun können.“

Wie hoch schätzen Sie die Retourenquote bei Brick-Flow?

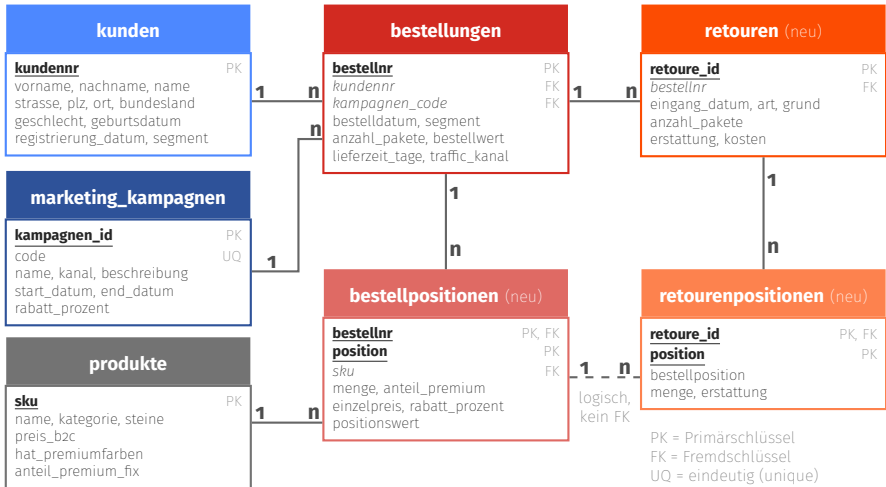
unter 5 %

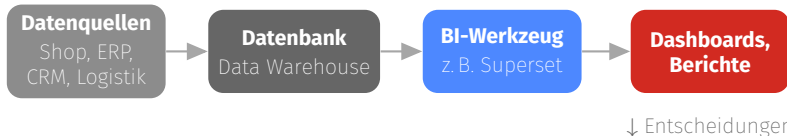
5–10 %

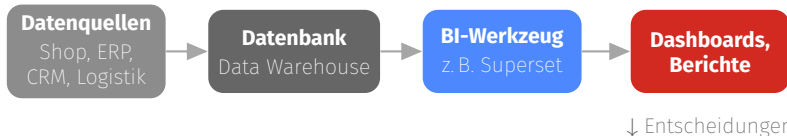
10–20 %

über 20 %

Umfrage: Ihre Schätzung?

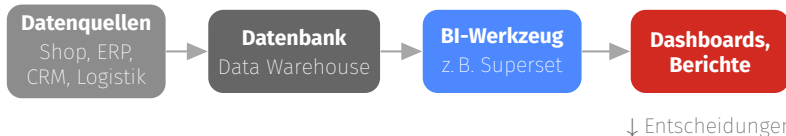






Business Intelligence (BI)

- Daten aus dem Unternehmen **sammeln, auswerten und darstellen**
- *Self-Service*: Fachabteilungen bauen Auswertungen selbst – ohne zu programmieren
- Gemeinsame Sicht auf die **Kennzahlen** des Unternehmens



Business Intelligence (BI)

- Daten aus dem Unternehmen **sammeln, auswerten und darstellen**
- *Self-Service*: Fachabteilungen bauen Auswertungen selbst – ohne zu programmieren
- Gemeinsame Sicht auf die **Kennzahlen** des Unternehmens

In den Analytics-Stufen: vor allem **descriptive** und **diagnostic** – „Was ist passiert?“ und „Warum?“

Metrik

Was wird gemessen?

Eine Zahl, die aus vielen Zeilen *zusammengefasst* wird

Anzahl Bestellungen,
Summe Umsatz, Ø Lieferzeit

Dimension

Wonach wird aufgeteilt?

Ein Merkmal, nach dem die Metrik *gruppiert* wird

Jahr, Segment, Bundesland, Kanal

Filter

Welche Daten zählen?

Eine Bedingung, die Zeilen *ein- oder ausschließt*

nur B2C, ab 2020, ohne Testbestellungen

Metrik

Was wird gemessen?

Eine Zahl, die aus vielen Zeilen *zusammengefasst* wird

Anzahl Bestellungen,
Summe Umsatz, Ø Lieferzeit

Dimension

Wonach wird aufgeteilt?

Ein Merkmal, nach dem die Metrik *gruppiert* wird

Jahr, Segment, Bundesland, Kanal

Filter

Welche Daten zählen?

Eine Bedingung, die Zeilen *ein- oder ausschließt*

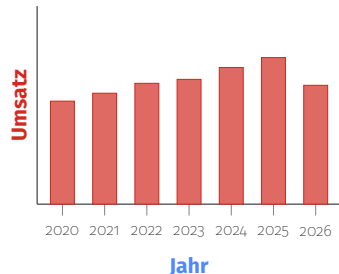
nur B2C, ab 2020, ohne Testbestellungen

In Superset heißen sie *Metrics*, *Dimensions* (bzw. *Group by*) und *Filters*.
Typische Metriken: **COUNT**, **SUM**, **AVG**, **MIN**, **MAX**.

„Wie hat sich der **Umsatz mit Endkunden seit 2020 pro Jahr** entwickelt?“

„Wie hat sich der **Umsatz mit Endkunden seit 2020 pro Jahr** entwickelt?“

Metrik Summe `bestellwert`
Dimension Jahr des `bestelldatum`
Filter `segment` = B2C
`bestelldatum` ab 2020



schematisch

```
SELECT EXTRACT(YEAR FROM bestelldatum) AS jahr,  
       SUM(bestellwert) AS umsatz  
FROM   bestellungen  
WHERE  segment = 'B2C'  
       AND bestelldatum >= '2020-01-01'  
GROUP BY jahr  
ORDER BY jahr;
```

```
SELECT EXTRACT(YEAR FROM bestelldatum) AS jahr,  
       SUM(bestellwert) AS umsatz  
FROM bestellungen  
WHERE segment = 'B2C'  
      AND bestelldatum >= '2020-01-01'  
GROUP BY jahr  
ORDER BY jahr;
```

- **Metrik** → Aggregatfunktion im SELECT
- **Dimension** → GROUP BY
- **Filter** → WHERE

Ein BI-Werkzeug wie Superset schreibt diese Abfrage für Sie – aus Klicks.

$$\text{Retourenquote} = \frac{\text{Anzahl ...}}{\text{Anzahl ...}}$$

- Eine Quote ist eine **Metrik** aus **zwei** Zählungen.
- Was genau im Zähler und im Nenner steht, ist **nicht** vorgegeben.
- **Filter** und **Dimensionen** verändern das Ergebnis.

$$\text{Retourenquote} = \frac{\text{Anzahl ...}}{\text{Anzahl ...}}$$

- Eine Quote ist eine **Metrik** aus **zwei** Zählungen.
- Was genau im Zähler und im Nenner steht, ist **nicht** vorgegeben.
- **Filter** und **Dimensionen** verändern das Ergebnis.

Welche Bestellungen zählen? Welche Retouren? Ab wann?
Das legen Sie in Ihrer Hausaufgabe fest – und begründen es.

- **Apache Superset:** Open-Source-Werkzeug für *Business Intelligence*
- Datenbanken abfragen, Diagramme und Dashboards bauen – im Browser
- Die Brick-Flow-Datenbank ist dort angebunden (**brickflow_db**)

In der Demo:

- Anmelden
- Tabelle auswählen und ansehen
- Ein erstes Diagramm

Diese Woche **optional** – am **22.10.** arbeiten alle damit.

Adresse

analytics.brick-flow.de

Benutzername

Ihre HSBO-E-Mail-Adresse

Kennwort

kommt in der Vorlesung

Probieren Sie den Zugang jetzt aus!

Probleme bitte direkt im Chat melden.



Chat: Login geklappt?

IHR ARBEITSAUFTRAG

1 – Verstehen

- Was ist eine Retoure, was kostet sie?
- Welche Retourenquoten sind im Online-Handel üblich?
- Welche Fragen hätten Sie an den Vorstand?

2 – Die Daten lesen

- Welche Tabellen und Spalten im ER-Modell brauchen Sie?
- Welche Fragen lassen sich mit den Daten beantworten – und welche nicht?

3 – Messen: eine KPI definieren

- Definieren Sie **genau eine** KPI.
- Zähler? Nenner? Zeitraum? Was wird ausgeschlossen?
- **Begründen Sie jede Entscheidung.**

4 – Handeln: das ML-Canvas

- Füllen Sie das Machine Learning Canvas für Brick-Flow aus.
- Ihre KPI gehört in *Live Evaluation and Monitoring*.

Gruppen wie in Übung 1 · Vorstellung am **22.10.** · höchstens **5 Folien** · optional: Apache Superset

Höchstens 5 Folien / 5 Minuten

1. Das Problem in einem Satz
2. Ihre KPI: Definition und Begründung
3. Die relevanten Daten (Tabellen, Spalten)
4. Ihr ML-Canvas
5. Offene Fragen

Pflichtfelder im ML-Canvas

- Value Propositions
- Decisions
- ML Task
- Data Sources
- Features
- Making Predictions
- Live Evaluation and Monitoring

Die übrigen Felder sind optional.

Diese Woche definieren – nächste Woche messen.

Am 22.10. setzen wir Ihre KPIs in Apache Superset um.