Trainingsdaten

Die Grenze des Modells ist die Grenze seiner Daten

Rohdaten sind nur Spuren der Realität: Protokolle, Formulare, Chats, Messwerte. Trainingsdaten sind daraus etwas geworden, aus dem sich lernen lässt. Der Unterschied ist der zwischen einer willkürlichen Zettelsammlung und einem strukturierten Lehrbuch. Über Erfolg entscheidet nicht die Menge, sondern ob die Daten zur Aufgabe passen.

DER WEG VON DER ROHMASSE ZUM LERNMATERIAL

1. **Sammeln.** Erst die Aufgabe klären, dann den Datenbedarf ableiten. Ein Spamfilter braucht Betreff, Inhalt und Metadaten der Mails, nicht die Systemprotokolle des Routers.
2. **Bereinigen.** Dubletten raus, Tippfehler korrigieren, Datumsformate vereinheitlichen. Sonst lernt das Modell Unsinn, etwa Tippfehler als Spam-Signal, weil sie zufällig gehäuft auftraten.
3. **Strukturieren.** In Zeilen und Spalten bringen: eine Zeile je Fall, eine Spalte je Merkmal. Kategorien und Skalen festlegen, Bezeichnungen vereinheitlichen.
4. **Kennzeichnen.** Jedes Beispiel braucht seine richtige Antwort: Spam oder kein Spam, Betrug oder kein Betrug. Die Trennschärfe des Modells hängt direkt an der Sorgfalt hier.
5. **Aufteilen.** Ein Teil wird zurückgehalten. Ohne diese Trennung misst man nur, ob das Modell auswendig gelernt hat.

DREI EBENEN DER DATENQUALITÄT

INHALTLICH
Bilden die Daten die Wirklichkeit ab, in der das Modell später arbeiten soll?

STRUKTURELL
Liegen sie in einer Form vor, die das Modell verarbeiten kann?

SEMANTISCH
Bedeutet das Label wirklich das, was die Fachabteilung darunter versteht?

WORAN ES TATSÄCHLICH SCHEITERT

- **Unklare Fragestellung.** Es wird gesammelt, bevor feststeht, was beantwortet werden soll.
- **Falsche Quelle.** Die Daten passen nicht zur Aufgabe oder decken sie nur zur Hälfte ab.
- **Uneinheitliche Labels.** Zwei Personen haben denselben Fall verschieden eingeordnet.
- **Vermischte Datensätze.** Ohne saubere Trennung ist die gemessene Leistung eine Illusion.

ZWEI FÄLLE, ZWEI LERNFORMEN

Fall	Was in den Daten steht	Was danach zu tun ist
Spamfilter, überwacht	Jede Mail trägt die Markierung Spam oder kein Spam	Markierungen einheitlich halten, sonst wird der Filter unzuverlässig
Kundensegmente, unüberwacht	Käufe, Klicks, Reaktionszeiten, ohne vorgegebene Antwort	Werte normieren, und die entstandenen Gruppen fachlich deuten. Sonst bleibt es Mathematik

WOHIN ES SICH GERADE BEWEGT

- **Weg von der Masse.** Kleinere, genauer umrissene und konsistent beschriftete Datensätze sind besser kontrollierbar als riesige Sammlungen unklarer Herkunft.
- **Erzeugte Daten.** Mit KI erzeugtes Material bildet seltene Fälle nach oder anonymisiert sensible Angaben. Es braucht dafür aber hochwertiges echtes Ausgangsmaterial.

DIE REIHENFOLGE, DIE ZÄHLT

1. Erst die Aufgabe, dann die Daten
2. Erst die Daten verstehen, dann das Modell wählen
3. Erst sauber trennen, dann die Leistung messen

MERKSÄTZE

Viel hilft nicht viel. Passung schlägt Menge.

Ein schwaches Lernmaterial lässt sich nicht durch einen anderen Algorithmus ausgleichen. Das bleibt Kosmetik.

Synthetische Daten helfen bei seltenen Fällen, erben aber jede Schieflage des echten Materials.