KAPITEL 2

Was die Technologie nicht verbessern kann

> *Trainingsdaten für Machine Learning ohne Ground Truth — eine dokumentierte, reproduzierbare Wahrheit — sind ein Weg ins Scheitern. Was ein Modell leisten kann, beginnt erst dort, wo die Wahrheit dokumentiert ist.*

Es gibt eine Hoffnung, die in fast jedem KI-Projekt-Briefing auftaucht. Sie wird selten ausgesprochen, aber sie strukturiert die Erwartungen aller Beteiligten. Die Hoffnung lautet: Wenn das Modell groß genug ist, korrigiert es schlechte Eingangsdaten.

Das tut es nicht. Kein Modell der letzten zehn Jahre hat das geleistet, und die nächste Generation wird es ebenfalls nicht tun. Was Machine Learning leistet, ist eine Annäherung an die Funktion, die in den Trainingsdaten angelegt ist. Wenn die Trainingsdaten widersprüchlich sind, reproduziert das Modell die Widersprüche. Wenn sie unvollständig sind, reproduziert es die Lücken. Wenn sie schlicht falsch sind, reproduziert es eine falsche Realität: präzise, schnell und in industrieller Skalierung.

Was Ground Truth eigentlich ist

Ground Truth ist der Begriff für eine dokumentierte Wahrheit, an der ein Modell gemessen werden kann. In einem Werk bedeutet das: Ein Defekt wurde identifiziert, kategorisiert und beschriftet, und zwar so, dass eine zweite Person mit dem gleichen Werkzeug zur gleichen Beschriftung kommen würde.

Der erste Satz ist unstrittig. Der zweite Satz ist der eigentliche Engpass. In den meisten Industrie-Projekten sind Defekte nicht reproduzierbar beschriftet. Was heute als Riss klassifiziert wird, ist morgen ein Lackfehler. Was eine Schicht als Ausschuss markiert, akzeptiert die nächste Schicht. Die Inkonsistenz ist nicht die Ausnahme, sondern der Normalfall. Sie entsteht systembedingt: Defekte am Rand der Toleranz sehen visuell ähnlich aus, und keine Person hält rund um die Uhr dieselbe Wahrnehmung.

Warum ein größeres Modell hier nicht hilft

Auf solche Daten lässt sich ein Modell trainieren. Das Training läuft durch, produziert eine fallende Fehlerkurve und liefert am Ende eine Genauigkeitskennzahl. Diese Kennzahl beschreibt, wie gut das Modell die widersprüchlichen Etiketten reproduziert. Sie beschreibt nicht, wie gut es Defekte erkennt.

Im Grenzfall — dem Edge Case — bleibt das Modell genauso unzuverlässig wie die Person, die es trainiert hat. Ein Grenzfall ist ein Defekt, den der Labeler heute erfasst und morgen übersieht. Genau dort sollte das Modell helfen. Genau dort versagt es.

Das ist die Grenze, an der Technologie nicht weiterhilft. Mehr Trainingsdaten verbessern nichts, wenn die Daten widersprüchlich sind. Größere Modelle verbessern nichts, wenn das Signal selbst unklar ist. Aufwendigere Architekturen verbessern nichts, wenn das, was angenähert werden soll, gar nicht stabil definiert wurde.

Ein Fall aus der Werkstoffprüfung

In der Werkstoffprüfung wiederholt sich dieser Engpass in typischer Form. Ein Hersteller will Defekte in einer Oberfläche automatisiert erkennen. Die ersten Trainingsdaten kommen aus der bestehenden Qualitätssicherung. Drei Schichten labeln parallel. Ein Modell wird trainiert. Die Akzeptanzrate

auf produktiven Daten liegt bei 73 Prozent: auf dem Papier brauchbar, in der Werkshalle ein Albtraum. Die 27 Prozent verteilen sich nicht zufällig, sondern auf genau jene Grenzfälle, an denen das Modell hätte helfen sollen.

Die Lösung liegt nicht in einem größeren Modell. Sie liegt in einer Reformulierung des Ground-Truth-Problems. Statt die menschliche Inkonsistenz zu reproduzieren, wird die zugrundeliegende physikalische Eigenschaft eines Defekts dokumentiert: Form, Tiefe, Reflexionscharakter. Diese Eigenschaften lassen sich simulieren. Aus der Simulation entstehen synthetische Defekte mit bekannten Parametern. Auf diese synthetischen Daten lässt sich Ground Truth schreiben, weil die Wahrheit bei der Erzeugung definiert wurde.

In einem Pilotaufbau dieser Art wurden zwanzig Modelle auf rein synthetischen Daten trainiert. Die Stabilität war über alle zwanzig Modelle vergleichbar — eine Eigenschaft, die kein auf menschlich gelabelten Daten trainiertes Modell erreicht hatte. Die zwanzig Modelle stimmten miteinander überein, weil die Realität, die sie annäherten, stabil dokumentiert war.

Wo synthetische Daten nicht helfen

Synthetische Daten sind keine Universallösung. Sie funktionieren dort, wo das physikalische Phänomen verstanden ist und sich in einer Simulation reproduzieren lässt. Sie funktionieren nicht in Domänen, in denen die Wahrheit selbst noch nicht entdeckt ist, etwa in komplexen sozialen Daten oder bei Phänomenen, deren Mechanik nicht modellierbar ist.

Was sie aber leisten: Sie verschieben den Engpass von der Datenbeschaffung zur Phänomen-Modellierung. Wer das physikalische Verhalten eines Defekts beschreiben kann, kann beliebig viele Trainingsbeispiele erzeugen. Wer es nicht kann, hatte ohnehin nie eine echte Ground Truth. Machine Learning hätte das Projekt nicht gerettet.

Fazit

Nicht jedes Datenproblem ist ein Datenproblem. Manche sind Definitions-Probleme. Wer ein KI-Projekt beauftragt, ohne vorher zu prüfen, ob eine Ground Truth überhaupt existiert, beauftragt eine teure Wiederholung dessen, was die Werkshalle schon hat: dieselbe Inkonsistenz, in höherem Tempo, mit GPU-Stromrechnung obendrauf.

Drei Fragen für den eigenen Kontext

1. **Existiert eine reproduzierbare Ground Truth?** Würden zwei unabhängige Labeler mit denselben Daten zu denselben Etiketten kommen? Ist das mit einer Stichprobe von 100 Fällen je geprüft worden, oder wird es nur angenommen?
2. **Falls nicht: Lässt sich das zugrundeliegende Phänomen modellieren?** Können physikalische, regelbasierte oder simulative Eigenschaften so dokumentiert werden, dass synthetische Daten mit bekannter Wahrheit erzeugt werden können?

3. **Falls auch das nicht geht: Ist Machine Learning der richtige Hebel?** Oder ist das Problem ein Definitions-Problem, das organisatorisch gelöst werden muss, bevor jede Modellierung beginnt?