Exkurs: Data Engineering

Nach all den linearen Modellen, Residuenanalysen und statistischen Tests ist es Zeit für eine kleine Abwechslung! In den nächsten drei Kapiteln machen wir einen Exkurs in die Welt des **Data Engineering** - ein Bereich, der sich von der reinen Datenanalyse unterscheidet, aber für Data Scientists durchaus relevant sein kann.

Was ist Data Engineering?

Während wir uns bisher hauptsächlich damit beschäftigt haben, bereits vorhandene und aufbereitete Datensätze zu analysieren, kümmert sich Data Engineering unter anderem um die **Beschaffung, Aufbereitung und Bereitstellung** von Daten. Data Engineers bauen die Infrastruktur und Pipelines, die dafür sorgen, dass Data Scientists überhaupt mit sauberen, strukturierten Daten arbeiten können.

Warum dieser Exkurs?

Streng genommen ist Data Engineering ein eigenständiger Beruf mit eigenen Technologien und Herausforderungen. Dennoch gibt es gute Gründe, warum Data Scientists die Grundlagen kennen sollten:

- **Datenquellen verstehen**: Nicht alle Daten kommen als ordentliche CSV-Dateien daher
- **Bessere Zusammenarbeit**: Wenn ihr wisst, wie Data Engineers arbeiten, könnt ihr besser kommunizieren
- **Mehr Autonomie**: In kleineren Teams seid ihr vielleicht selbst für die Datenbeschaffung zuständig
- **Erweiterte Möglichkeiten**: APIs, Datenbanken und Big Data eröffnen völlig neue Datenquellen

Was euch erwartet

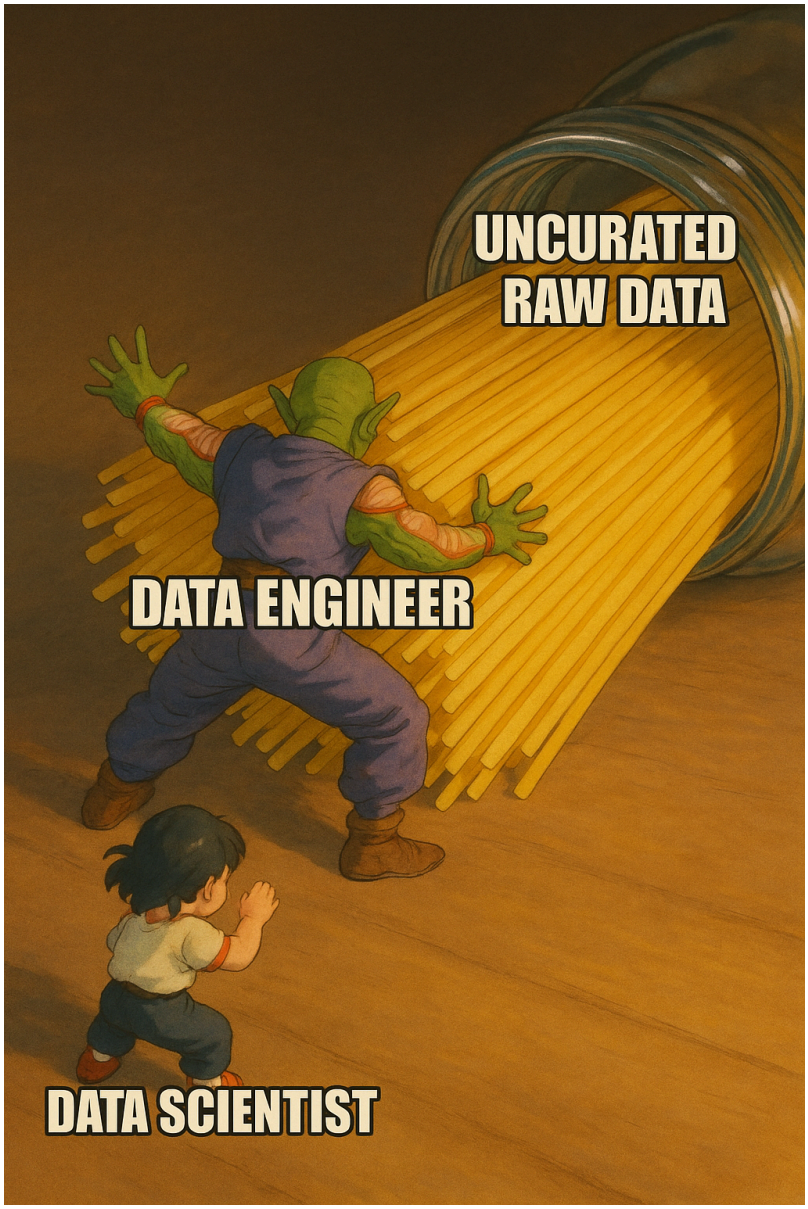
In den kommenden drei Kapiteln schauen wir uns an:

1. **HTTP APIs**: Wie ihr Daten von Webservices abruft
2. **SQL**: Die Sprache der Datenbanken und wie ihr sie mit Python nutzt
3. **Big Data**: Ein Überblick über Technologien für riesige Datenmengen

> ℹ Oberflächlich, aber praxisnah
>
> Diese Themen sind so umfangreich, dass man Jahre damit verbringen könnte. Wir kratzen bewusst nur an der Oberfläche - das Ziel ist, dass ihr wenigstens einmal eine API angezapft, eine SQL-Abfrage geschrieben und von Hadoop gehört habt. Je nach eurem späteren Job werdet ihr vielleicht täglich damit arbeiten oder nie wieder - das hängt stark davon ab, ob es Data Engineers in eurem Team gibt und woher eure Daten kommen.

Zurück zur Statistik

Nach diesem kleinen Ausflug in die Welt der Datenbeschaffung kehren wir dann wieder zu unserem Hauptthema zurück: der statistischen Datenanalyse und den Methoden des Machine Learning. Aber mit einem erweiterten Werkzeugkasten für die Datenbeschaffung werdet ihr deutlich flexibler in der Wahl eurer Datenquellen sein.


UNCURATED
RAW DATA
DATA ENGINEER
DATA SCIENTIST