

## ***AI210 Einführung in die Datenextraktion und Datenaufbereitung***

### **Kurzbeschreibung:**

Im Workshop **AI210 Einführung in die Datenextraktion und Datenaufbereitung** erlernen Sie, wie Sie Daten aus den ungewöhnlichsten Quellen extrahieren, aufbereiten und für Ihre Projekte nutzbar machen. Ob Textdateien, Datenbanken, Webseiten oder sogar PDFs und Word-Dokumente - entdecken Sie die Möglichkeiten und werden Sie zum Meister der Datenmanipulation.

Vor jeder Datenanalyse steht zunächst die Datenbeschaffung. Doch Datenquellen beschränken sich nicht auf relationale Datenbanken oder Excel-Dateien. Oft liegen unsere Daten in unterschiedlichsten Formaten vor: Von PDF-Dateien über Webseiten bis hin zu Bildern kann alles dabei sein. Dieser Einführungskurs vermittelt anhand anschaulicher Beispiele Techniken, um Daten aus verschiedensten Quellen zu extrahieren und für die Weiterverwendung aufzubereiten. Der Kurs beinhaltet die Verarbeitung von Textdateien wie Logs, CSV und Excel, das Einlesen von Daten aus Datenbanken, das Auslesen von Webseiten und die Datenextraktion aus weiteren Formaten wie PDF und Word. Zudem werden Techniken zur Datenaufbereitung und Bereinigung vorgestellt.

### **Zielgruppe:**

- Entwickler
- IT-Fachkräfte

### **Voraussetzungen:**

- AI200 Einführung in Python für Data Science und KI (alternativ Grundkenntnisse in Python)

### **Sonstiges:**

**Dauer:** 1 Tage

**Preis:** 450 Euro plus Mwst.

### **Ziele:**

Daten aus verschiedensten Quellen extrahieren, Daten aufbereiten und bereinigen, Textdateien verarbeiten, Daten aus Datenbanken einlesen, Webseiten auslesen, Daten aus weiteren Formaten extrahieren können.

#### Inhalte/Agenda:

- ♦ Einführung in die Datenextraktion und -aufbereitung
- ♦ Verarbeiten von Textdateien wie Logs, CSV und Excel mit Pandas und Regex
- ♦ Einlesen von Daten aus Datenbanken, inklusive SQL und NoSQL
- ♦ Auslesen von Webseiten mit BeautifulSoup und Requests
- ♦ Datenextraktion aus weiteren Formaten wie PDF, Word und aus Bildern mit OCR-Techniken
- ♦ Techniken zur Datenaufbereitung und Bereinigung mit Pandas (einschließlich Datenmanipulation, Datentransformation und Fehlerbehandlung)
- ♦ Zusammenfassung, Best Practices und Abschlussdiskussion