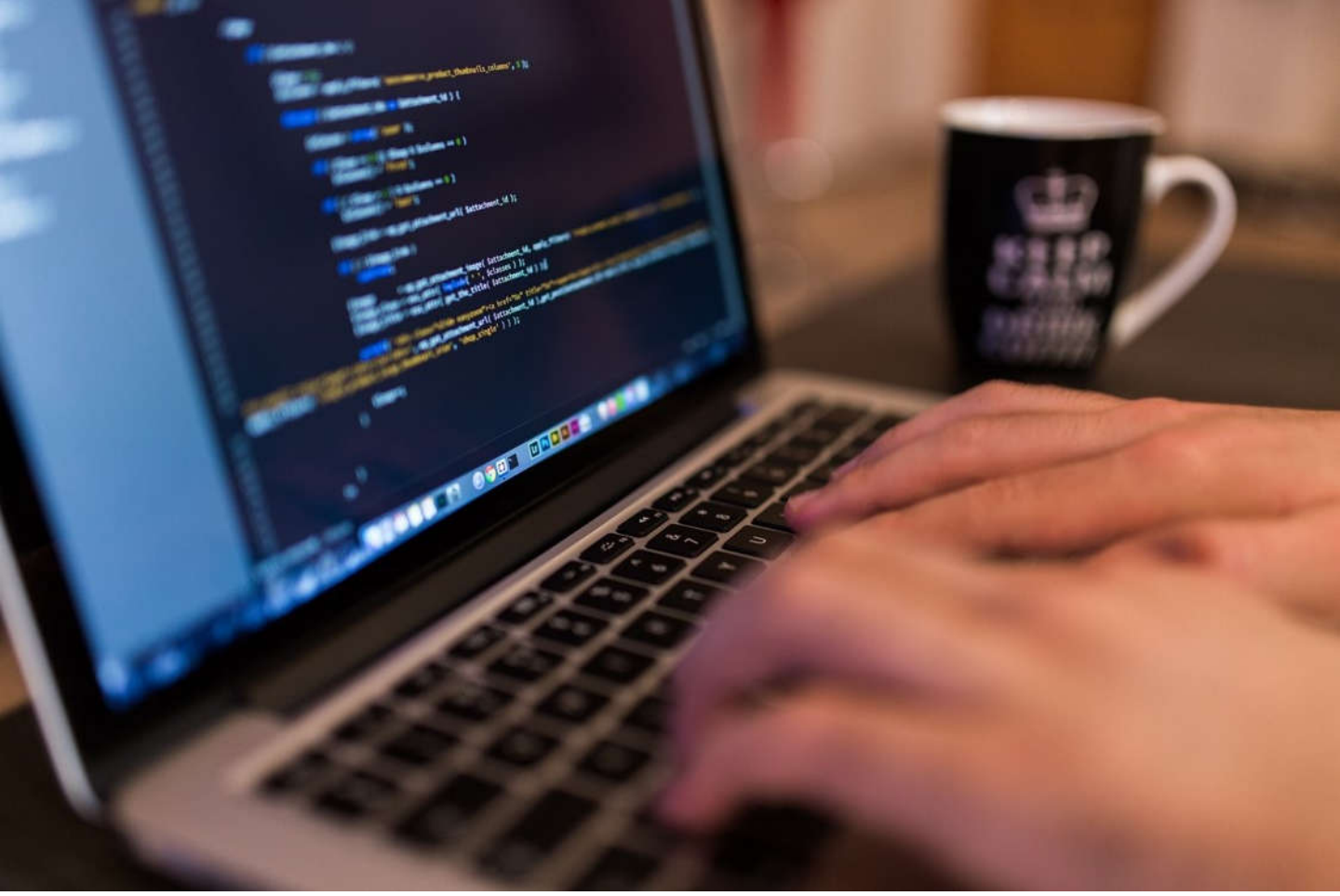

# Data Science Bootcamp in Python

## Intensive Schulung über Data Science und Machine Learning mit pandas und scikit-learn

ENABLE AI

MASTER YOUR DATA

# Data Science Bootcamp in Python

## Data Science, Machine Learning, pandas, scikit-learn

### Über den Kurs

**Dauer:** 5 Tage

**Gruppengröße:** 3-10

**Level:** Anfänger mit Programmiererfahrung*

**Anteil Coding:** 60%

**Sprache:** Python

**Bibliotheken:** pandas, numpy, matplotlib, seaborn, scikit-learn

### Auf einen Blick

- Wichtige Datenstrukturen
- pandas Data.Frame
- Statistiken berechnen
- Eine eigene Funktion schreiben
- Grafiken mit seaborn erstellen
- Daten einlesen/schreiben (csv, pickle, Datenbank)
- Machine Learning, Train-Test-Split, Kreuzvalidierung
- Entscheidungsbaum, K-Means, DBSCAN, AdaBoost,
- Random Forest, SVM, Neuronales Netz, Ensembles

#### Überblick über das Seminar

Diese einwöchige Schulung fängt bei den notwendigen Grundlagen von Python für die Datenanalyse (data analytics) an und es werden alle Voraussetzungen unterrichtet, um am Ende Machine Learning Algorithmen umzusetzen. Bei den Algorithmen wird das notwendige theoretische Verständnis geschult. Das Ziel liegt auf der praktischen Umsetzung der Datenanalyse und Algorithmen in Python. Am Ende des Seminars können Sie selbstständig erste Datenanalysen umsetzen, Machine Learning Algorithmen einsetzen und deren Ergebnisse interpretieren.

Das Paket pandas ist ein Schwerpunkt, da pandas speziell für Data Science entwickelt wurde. Die wichtigsten Schritte für die Datenaufbereitung werden eingeübt. Zur Erstellung von Plots und Grafiken wird das Paket seaborn verwendet mit einer kurzen Einführung in matplotlib. Matplotlib überzeugt durch die Fülle an Möglichkeiten einen Plot anzupassen, seaborn durch die Einfachheit auch komplexere Plots mit wenig Code zu erstellen. Es werden die Grundlagen in numpy gelehrt, um dieses wichtige Paket im Data Mining einsetzen zu können.

Nach den Grundlagen für Python (Datenstrukturen, eigene Funktionen schreiben) und der Erläuterung von pandas für die Auswertung von Daten, erhalten Sie einen vertieften Überblick über Machine Learning Algorithmen, welche wir in Python mit dem Paket scikit-learn selbst coden werden. Hierbei wird eine Auwahl der bekanntesten Algorithmen erklärt (Lineare und logistische Regression, Entscheidungsbaum, Random Forest, SVM, Ensemble Learning, AdaBoost, K-Means, DBSCAN Clustering). Ein wichtiger Bestandteil des Seminars ist das eigenständige Arbeiten und Lösen von Übungsaufgaben, so dass Sie mit direkter Hilfe des Trainers das Besprochene direkt in der Praxis umsetzen und anwenden können.

#### Inhalte des Seminars

Sie erhalten einen Überblick über die im Data Science, Data Mining, Machine Learning und Deep Learning populärste Programmiersprache Python. Wir verwenden die Anaconda Distribution (nach eigenen Angaben “The World’s most popular data science platform“) und als Entwicklungsumgebung/IDE wird spyder verwendet.

ENABLE AI

* Programmierkenntnisse in Python sind nicht notwendig. Erste Erfahrung mit einer anderen Sprache (z.B. R, VBA, C) ist ausreichend.

Nach dem Seminar können Sie Daten aus verschiedenen Formaten und von Datenbanken (mit den Paketen SQLAlchemy und pandas) einlesen, Daten mit seaborn/matplotlib plotten bzw. Daten mit pandas bereinigen (fehlende Werte ersetzen, Zeilen und Spalten anpassen, neue Spalten erzeugen) und Berechnungen mit numpy durchführen.

Sie kennen die wichtigsten Datentypen in Python, können eigene einfache Funktionen schreiben und kennen die Umsetzung von Control Flows (For-Schleife, If-Else). Sie verstehen das Grundkonzept eines pandas DataFrames und können damit Data Wrangling und Data Cleaning durchführen. Die Vorverarbeitung von Daten (data preprocessing) für die Umsetzung von Algorithmen mit scikit-learn wird angesprochen.

Die Einteilung von Machine Learning in supervised-unsupervised (überwachtes-unüberwachtes Lernen) und Reinforcement Learning ist Ihnen bekannt und Sie können mit scikit-learn eigenständig Algorithmen in Python trainieren, validieren, einen Train-Test Datensplit durchführen und Gütekriterien zur Beurteilung eines Algorithmus berechnen und interpretieren.

Sie wissen, was Overfitting (Überanpassung) bedeutet, wie dies nach dem Training eines Algorithmus identifiziert werden kann und welche Anpassungen es bei einzelnen Algorithmen gibt, um Overfitting zu verringern.

Ein großer Schwerpunkt liegt auf der Bibliothek scikit-learn und der Umsetzung und das intuitive Verständnis von bekannten Algorithmen des Machine Learning. Das Seminar umfasst Algorithmen für die Regression (Lineare Regression, Random Forest, Neural Network, Decision Tree), für die Klassifikation (Logistische Regression, Enscheidungsbaum, Random Forest, AdaBoost, K-Nearest Neighbor) und dem Clustering (K-Means, DBSCAN). Desweiteren wird das Erstellen eines Ensembles erläutert und die Konzepte von Grid-Search zur automatischen Optimierung von Hyperparametern und die Umsetzung einer Kreuzvalidierung (Cross-Validation) an Stelle eines klassischen Train-Test-Datensplits.

Am Ende ist die Einstiegshürde für die Benutzung von Python für Machine Learning/Data Science/Data Mining/Business Intelligence/Data Analytics genommen und eine Vertiefung in scikit-learn ist vorhanden, so dass Sie eigenständig Ihr Wissen nach dem Seminar Stück für Stück eigenständig erweitern können. Der Schwerpunkt während des Seminars liegt auf der eigenen Umsetzung auf Ihrem Laptop mit Unterstützung des Dozenten.

### Wer sollte teilnehmen?

Dieser Kurs mit Python richtet sich an data scientists, angehende Machine Learning engineers, Datenanalysten, Busines Intelligence Analysts, Data Analysts, o.ä. welche die Programmiersprache Python für Data Science/Data Mining kennenlernen möchten, um Datenanalysen unter Verwendung von Machine Learning Algorithmen eigenständig mit Python und den Pakten scikit-learn umzusetzen.

### Methode des Seminars

Dieses Seminar ist sehr praxisorientiert. Die Teilnehmer arbeiten direkt und selbstständig mit der Programmiersprache Python in der Entwicklungsumgebung Spyder mit der Anaconda Distribution, so dass das Erlernte direkt geübt und vertieft werden kann. Der Trainer moderiert dabei verschiedene Aufgaben und begleitet die Teilnehmer unterstützend durch die einzelnen Lehreinheiten.

## Voraussetzungen

Dieser Python Kurs setzt keine grundlegenden Kenntnisse in Python voraus. Notwendig ist jedoch Vorerfahrung mit einer Programmiersprache, damit Konzepte einer Variablen, Zuweisung von Werten zu einer Variablen, eine Funktion bzw. eine for-Schleife bekannt sind.

Notwendig sind außerdem grundlegende Vorkenntnisse im Bereich der Statistik (Begriffsdefinitionen wie bspw. Mittelwert, Median, Standardabweichung, Normalverteilung), Kenntnisse grundlegender mathematischer Symbole und Begriffe (Summenzeichen, Integral, Funktion, Ableitung, Exponentialfunktion) und Kenntnis der booleschen Algebra mit den logischen Operatoren (UND, ODER, NICHT) sind sehr empfohlen.

Die Teilnehmer sollten Vorerfahrung mit dem Umgang von Daten haben, z.B. in Excel oder einer BI-Software, damit Konzepte einer spaltenweisen Berechnung bzw. einfache Statistiken (Mittelwert, Varianz) bekannt sind.

Das Seminar wird auf Deutsch gehalten. Englischkenntnisse (lediglich im Verstehen von englischen Texten) sind sehr empfehlenswert, da die Programmiersprache, Fachbegriffe und die Dokumentationen im Internet auf Englisch sind. Aus diesem Grund sind auch die erstellten Folien in der Schulung auf Englisch.

## Technische Voraussetzungen der Teilnehmer (Laptop, etc.)

- Die Teilnehmer benötigen für die Übungsaufgaben Laptops. Wir empfehlen, Ihren eigenen Laptop mit der vorab installierten Software mitzubringen. Eine genaue Installationsanleitung für die Software wird Ihnen vor dem Seminar per E-mail zugesandt. Auf Anfrage stellen wir auch Schulungslaptops zur Verfügung.
- Bitte prüfen Sie, ob Ihr Firmenlaptop Zugangsbeschränkungen im Internet hat. Die digitalen Unterlagen (Skript, Code, Dateien) werden im Seminar online zum Download zur Verfügung gestellt. Sie erhalten vor dem Seminar per E-Mail den Link zu einer Testdatei zum Download, um dies überprüfen zu können.
- Sie sollten sich in firmenfremde WLAN-Netze registrieren können.
- Als Backup Lösung ist es möglich, dass der USB Port bei Ihrem Laptop freigeschalten ist, um damit verwendete Dateien oder sonstige Unterlagen übertragen zu können.
- Im Seminar wird das Betriebssystem Windows verwendet. Der Umgang mit Ihrem verwendeten Betriebssystem und Laptop sollte bekannt sein. Insbesondere sollten Sie ohne Schwierigkeiten Sonderzeichen auf der Tastatur finden (insbesondere bei Apple Geräten werden auf manchen Tastaturen nicht immer runde, eckige bzw. geschweifte Klammern dargestellt).

# Tagesübersicht

Tag 1

**10.00-10.15**
## Begrüßung und Organisatorisches
- Vorstellungsrunde
- Erwartungen der Teilnehmer

**10.15-11.45**
## Grundlagen von Python
- Neue Pakete installieren
- Spyder als Entwicklungsumgebung/IDE
- Funktionen und Methoden
- Wichtige Aspekte von Python im Vergleich zu anderen Programmiersprachen

**11.45-12.00**
## Kaffeepause

**12.00-13.30**
## Grundlegende Datenstrukturen
- Überblick über die basic data types (string, integer, float, NaN)
- Erläuterung der wichtigsten Datenstrukturen: list, tuple, dictionary
- List comprehension

**13.30-14.30**
## Mittagspause

**14.30-16.00**
## Das Paket pandas – Data.Frame
- Struktur (Zeilen, Spalten) eines DataFrames
- Auswahl einer Zeile/Spalte
- Zeilen/Spalten erstellen, löschen, ändern
- Boolean indexing: eine logische Abfrage zur Selektion von Zeilen
- Daten zusammenfassen, um einen Überblick zu erhalten.

**16.00-16.15**
## Kaffeepause

**16.15-18.00**
## Berechnen von Statistiken direkt im pandas Data.Frame
- Einfache Statistiken direkt auf einem Data.Frame (Mittelwert, Min, Max, Summe, Median, Varianz,…)
- Zusammenfassen und Filtern von Daten
- Fehlende Werte ersetzen
- Kreuztabelle (Kontingenztafel)

**18.00**
## Ende

# Tagesübersicht

Tag 2

**09.00-09.15**
**Rückblick und offene Fragen von Tag 1**

**09.15-10.45**
Control Flows
- Eine eigene Funktion schreiben
- Default Parameter in einer Funktion setzen: position arguments und keyword arguments
- For Schleifen
- If-Else Bedingungen
- List comprehension mit if-else

**10.45-11.00**
Kaffeepause

**11.00-12.30**
Datenvisualisierung mit seaborn / matplotlib
- Wichtige Grundlagen von matplotlib
- Achsen-Beschriftung, Legende, Titel ändern
- Einen Plot speichern
- In seaborn Linienplot, Boxplot, Histogram, Scatterplot, Barplot erstellen
- Darstellungen (Größe der Datenpunkte, Farbe, Gruppierung) mit einer Variable variieren oder festsetzen

**12.30-13.30**
Mittagspause

**13.30-15.15**
Daten einlesen und schreiben
- Das Arbeitsverzeichnis in Python und der IDE spyder setzen
- Ein CSV bzw. Excel einlesen und schreiben
- Von einer URL einlesen
- Überblick über nützliche Parameter
- Lesen und Schreiben vom Python Format pickle
- Umgang mit großen Daten

**15.15-15.30**
Kaffeepause

**15.30-17.00**
Datenbank
- Das Paket SQLAlchemy, um mit einer Datenbank zu verbinden
- Einzelne Tabellen extrahieren bzw. schreiben
- SQL Befehle an die Datenbank schicken, um Daten zu ändern
- Datenabfragen per SQL Statement direct aus Python heraus

**17.00**
Ende

# Tagesübersicht

Tag 3

**09.00-09.15**
**Rückblick und offene Fragen von Tag 2**

**09.15-10.45**
**Numpy**
- Ein numpy array und dessen Attribute
- Arrays erstellen und mit Daten befüllen (bzw. mit Zufallszahlen)
- Mathematische Operationen mit Numpy
- Funktionen der Statistik umsetzen

**10.45-11.00**
Kaffeepause

**11.00-12.30**
Datenaufbereitung
- Daten für die Analyse normalisieren
- Fehlende Werte ergänzen
- Dummy Variablen erstellen bzw. One-Hot Encoding

**12.30-13.30**
Mittagspause

**13.30-15.15**
**Überblick über Machine Learning**
- Einführung in Machine Learning (ML)
- Anwendungsbeispiele von ML
- Künstliche Intelligenz – Machine Learning – Deep Learning
- Unterschied Supervised – Unsupervised Learning (überwachtes – unüberwachtes Lernen)
- Overfitting und Train-Test-Split

**15.15-15.30**
Kaffeepause

**15.30-17.00**
**Lineare Regression mit scikit-learn**
- Aufteilung der Daten in Test-Train, ein Model erstellen und validieren
- Grundlagen des Algorithmus
- Umsetzung in Python mit scikit-learn
- Validieren der Ergebnisse (mean squared error)

**17.00**
Ende

# Tagesübersicht

Tag 4

**09.00-09.15**
**Rückblick und offene Fragen von Tag 3**

**09.15-10.45**
Logistische Regression mit statsmodels
- Statsmodels – ein Paket für statistische Modelle und Analysen
- Aufteilung der Daten in Test-Train, ein Model erstellen und validieren
- Grundlagen des Algorithmus
- Validieren der Ergebnisse

**10.45-11.00**
Kaffeepause

**11.00-12.30**
Entscheidungsbaum mit scikit-learn
- Aufteilung der Daten in Test-Train, ein Model erstellen und validieren
- Grundlagen des Algorithmus
- Validieren der Ergebnisse (u.a. confusion matrix, sensitivity, accuracy)
- Anpassen von Hyperparametern im Training

**12.30-13.30**
Mittagspause

**13.30-15.15**
Ensembles (+ Ada-Boost)
- Ein ensemble mit scikit-learn erstellen und trainieren
- Bagging (Bootstrap Aggregating)
- Boosting
- Grundlagen des AdaBoost Algorithmus
- AdaBoost für Klassifikation und Regression
- Umsetzung in Python mit scikit-learn
- Validieren der Ergebnisse

**15.15-15.30**
Kaffeepause

**15.30-17.00**
Random Forest
- Aufteilung der Daten in Test-Train, ein Model erstellen und validieren
- Grundlagen des Algorithmus
- Validieren der Ergebnisse
- Out-of-bag error
- Random Forest für die Regression
- Anpassen von Hyperparametern im Training

**17.00**
Ende

# Tagesübersicht

Tag 5

**09.00-09.15**
**Rückblick und offene Fragen von Tag 4**

**09.15-10.45**
Einblick in weitere Algorithmen

- Grundlagen der folgenden Algorithmen, sowohl für Klassifikation, als auch für Regression:
  - K-nearest Neighbor
  - Einfaches Neuronales Netzwerk (Multi-Layer Perceptron) in scikit-learn
- Umsetzen der Algorithmen mit scikit-learn
- Validieren der Ergebnisse

**10.45-11.00**
Kaffeepause

**11.00-12.30**
Grid Search & Cross Validation

- Cross-validation (Kreuzvalidierung). Idee und Umsetzung in scikit-learn
- Grid Search: Automatische Suche nach den besten Hyperparametern. Umsetzung in Python
- Wie trainiere ich nach der Kreuzvalidierung das finale Modell?

**12.30-13.30**
Mittagspause

**13.30-15.15**
Clustering (K-Means, DBSCAN)

- Ein Cluster-Model erstellen und validieren
- Grundlagen der Algorithmen
- Umsetzung in Python mit scikit-learn
- Validieren der Ergebnisse (Sillhouette Score, Calinski-Harabasz)

**15.15-15.30**
Kaffeepause

**15.30-17.00**
Abschließende Aspekte, **Fallbeispiel zum Wiederholen**

- Wann wähle ich welchen Algorithmus?
- Welche Entscheidungsgrenzen bilden Algorithmen (intuitive 2D Darstellung)?
- Fallbeispiel, um eigenständig das Gelernte zu wiederholen und auftretende Fragen zu besprechen

**17.00**
Ende

# Ihre Dozenten

Einer unserer folgenden Experten leitet das Seminar

**Dr. Rolf Köhler**

Nach dem Studium der Mathematik und BWL promovierte er im Cyber Valley am Max-Planck Institut Tübingen. Sein Forschungsschwerpunkt war im Bereich Machine Learning und Bildverarbeitung. Seit 2015 arbeitet er bei der Robert Bosch GmbH im Bereich Deep Learning und implementiert und adaptiert verschiedene Algorithmen für industrielle Anwendungsfälle. Daraus sind mehrere Patentanmeldungen entstanden. Seit 7 Jahren verwendet er die Programmiersprache Python.


**Jan Köhler**

Vom Hintergrund Statistiker und Wirtschaftsingenieur hat er über 7 Jahre an den neuesten Technologien in Machine Learning, Deep Learning und Data Science im Bosch Center for Artificial Intelligence (BCAI) gearbeitet, hat in der Praxis bei über 25 Patentanmeldungen (meist als Haupterfinder) beigetragen und ist Mitautor bei Veröffentlichungen auf Machine Learning Konferenzen neben Veröffentlichungen im Bereich der Medizinstatistik bzw. des Operations Research. In vielen Praxisprojekten unterstützte er bisher als Data Scientist und hat verschiedene Teilnehmer, vom Projektmitarbeiter bis zum Konzern-Vorstand geschult.

# Zusammenfassung

**Preise**

**2630 € zzgl. MwSt.**

**Termin und Ort**

Termine und Orte finden Sie unter https://enable-ai.de

## Haben Sie Fragen? Wir helfen Ihnen. Versprochen.

Enable AI, Stuttgart

0711 96881553

info@enable-ai.de

ENABLE AI