Veranstaltung | Workshop

Data Science für Big Data mit Python, Spark und AWS

Am 28. März 2023 spricht Prof. Dr. Schlesinger im Rahmen der Methodenwerkstatt Statistik über »Data Science für Big Data mit Python, Spark und AWS«. Die Veranstaltung richtet sich an alle Promovierenden und wissenschaftlichen Mitarbeiter*innen der HWR Berlin.

Di
28.03.

Uhrzeit
14.00–18.00

Kosten
kostenfrei

Anmeldung
Details auf Moodle

Zu Moodle

Veranstalter/in
Institute for Data-Driven Digital Transformation (d-cube) in Kooperation mit der Methodenwerkstatt Statistik

Zum Institut

Im Rahmen der „Methodenwerkstatt Statistik“ möchten wir zusammen mit einer Reihe von HWR-Professor*innen mit Expertise in verschiedenen quantitativen Forschungsmethoden eine stärker zielgerichtete Methodenausbildung von Promovierenden und Nachwuchswissenschaftler*innen an der HWR fördern. Diese Input-Veranstaltungen im interaktiven Formate gestaltet strebt an, dass im gemeinsamen Austausch eine Annäherung an das Thema ausgearbeitet und erreicht wird.

Die Methodenwerkstatt Statistik steht allen Promovierenden, aber auch wissenschaftlichen Mitarbeitenden sowie Interessierten an methodischen Fragestellungen offen.

Data Science für Big Data mit Python, Spark und AWS

Data Science, d.h. das Extrahieren oder Extrapolieren von Wissen und Erkenntnissen aus strukturierten und unstrukturierten Daten spielt eine bedeutende Rolle in Wissenschaft und Industrie. Das Forbes Magazine bezeichnete Data Scienctists als sexiest jobs on earth.

Python spielt wiederum eine herausragende Rolle, um Data Science oder Statistik programmatisch zu unterstützen. Das liegt u.a. an der leichten Zugänglichkeit und der überragenden Menge und Qualität an verfügbaren Bibliotheken, die in Python zur Verfügung stehen.

Hat man es darüber hinaus mit einer großen Datenmenge zu tun, wie das bei Machine Learning oft der Fall ist oder auch generell bei vielen unternehmerischen Use Cases, dann reicht die sequentielle Verarbeitung auf einem einzelnen Rechner oft nicht aus. Man könnte nun auf unterster Ebene mit parallelen Strukturen programmieren, aber es gibt glücklicherweise einige Frameworks, die einem das erleichtern.

Dazu gehört Spark, das aus der Hadoop-Familie entstammt. Im Kern formuliert man seinen Code in deklarativer, funktionaler Weise, der dann wiederum parallelisiert und auf einem Cluster, der aus mehreren Rechnern ausgeführt wird.

Oft hat man nicht viele eigene Rechner, die einen Cluster bilden können zur Verfügung.
Um dem zu begegnen, bieten sich Cloudanbieter wie Amazon Web Services (AWS) an.

In dieser Session machen wir einen Rundumschlag um Data Science mit Python und gehen hier auf Ansätze ein, die einem das Leben mit großen Datenmengen erleichtern. Wir fangen mit Python an, gehen auf ein paar nützliche Frameworks ein, die direkt auf einzelnen Rechnern laufen, und dann über auf Spark und das Aufsetzen der Infrastruktur in AWS.