Das Berufsbild „Data Scientist“

Angi Voss, Fraunhofer IAIS

Dieser Artikel stellt den Data Scientist vor und gibt einen Überblick über dessen Aufgaben, Kompetenzen, Werkzeuge sowie Aus- und Weiterbildungsmöglichkeiten.

Parallel zum neuen Technologiefeld „Big Data“ entwickelt sich derzeit das Berufsbild der Data Scientists. Diese machen „Big Data“ zu „Smart Data“. Mit Erfindungsreichtum halten sie Ausschau nach verschiedensten Datenquellen, führen sie zusammen, explorieren die Daten und entwickeln Ideen, um sie gewinnbringend zu nutzen. Dazu benötigen sie einschlägiges Fachwissen – zum Beispiel in Handel, Gesundheitswesen, Finanzwirtschaft, Energiewirtschaft oder Produktion und Logistik. Data Scientists müssen ihr Unternehmen gut kennen, um richtig einschätzen zu können, wo datenbasierte Erkenntnisse zu mehr Effizienz im operativen Betrieb, zu besseren Plänen, besseren Produkten oder gar neuen Dienstleistungen und Geschäftsmodellen führen können.

Mit den identifizierten Datenquellen – seien sie noch so ungewöhnlich, vielfältig, umfangreich oder schnelllebig – müssen Data Scientists effizient umgehen können. Hier sind IT-Kenntnisse im skalierbaren Daten-Management und in der Parallelverarbeitung gefragt. Cloud Computing, Stream-Processing, MapReduce, SQL- und NoSQL-Datenbanken sind Technologi-

en, die sie einschätzen und einsetzen können.

Zur Entwicklung einer Anwendung gehört die Analyse der verfügbar gewordenen Daten. Data Scientists explorieren die Daten und ihre Qualität und konzipieren die Verarbeitungsschritte, die in der Anwendung später vollautomatisch durchgeführt werden. Dazu gehören Daten-Bereinigung, -Vorverarbeitung, -Fusion und -Anreicherung. Mit den vorhandenen Daten trainieren und evaluieren sie prädiktive statistische Modelle.

In der Anwendung erlauben sie, in den neuen Daten Muster zu erkennen, bestimmte Arten von Informationen zu extrahieren, Entscheidungen zu treffen und Prognosen aufzustellen. Hierzu beherrschen Data Scientists klassische statistische Verfahren und Methoden aus dem Data Mining und dem maschinellen Lernen. Sie nutzen SQL und Skriptsprachen wie R oder Python. Darüber hinaus verlangen digitalisierte Daten, Bilder und Videos nach speziellen Methoden; Textdaten erfordern zum Beispiel eine mehrstufige Vorverarbeitung, natürliche Sprachverarbeitung und Modellierungstechniken für hochdimensionale Merkmalsräume.

Bei der Analyse und Interpretation unüberschaubarer Datenmengen spielt die interaktive Visualisierung eine wichtige Rolle. Gleich am Anfang hilft dem Data Scientist die visuelle Exploration der Daten, später das visuelle Debugging bei der Beurteilung und Verbesserung der Modelle und schließlich das visuelle Reporting bei der Kommunikation der Ergebnisse.

Neben Informationsgewinn und Skalierbarkeit sind bei der Bewertung einer Big-Data-Anwendungsidee Datensicherheit und Datenschutz wichtige Faktoren. Für individualisierte Kundenansprache, Produkte und Dienstleistungen finden sich bereits heute viele Anwendungsbeispiele aus den USA, aber längst nicht alle lassen sich mit dem deutschem Datenschutz in Einklang bringen. Hierzu wird künftig in Deutschland eine Reihe von Forschungsaktivitäten gefördert, über deren Ergebnisse sich Data Scientists auf dem Laufenden halten sollten. Zunehmend etablieren sich auch Datenanbieter, die proprietäre oder offene Daten bereinigen und gesetzeskonform aggregieren. Solche Plattformen und Dienste sollten Data Scientists kennen, da die Kombination von externen und internen Daten besondere Potenziale bietet.

Data Science ist Teamarbeit und macht Spaß

Kaum jemand wird sich in allen Gebieten gleich gut auskennen, sodass Data Scientists in der Regel in Teams arbeiten, die interdisziplinär zusammengestellt sein können. Hier finden sich neben Informatikern, Betriebswirten, Statistikern und Mathematikern auch Physiker und Sozialwissenschaftler sowie Wissenschaftler aus weiteren Anwendungsdisziplinen.

Der Harvard Business Review sprach im Jahr 2012 vom „Sexiest Job of the 21st Century"[1]. Es fallen Begriffe wie Künstler, Hacker-Mentalität, Daten-Jiujitsu, Kreativität, Unternehmertum und holistische Herangehensweise. Interdisziplinäre Data-Scientist-Teams bilden hoch attraktive Arbeitsumgebungen.

Das Schulungsangebot wächst

Eine Studie von McKinsey [2] hat bereits im Jahr 2011 eine große Nachfrage nach Data Scientists in den USA prognostiziert. Als Reaktion darauf entstand eine Vielzahl von Studienangeboten im Bereich der Datenanalyse. In Europa rechnet man mit einem Bedarf von 300.000 in den nächsten Jahren [3].

Zwar gibt es in Deutschland noch keine speziellen Studiengänge für Data Scientists, jedoch inzwischen eine Reihe von Seminaren an den Hochschulen und kostenpflichtige Fortbildungsangeboten für Berufstätige. Die Lehrinhalte umfassen inhaltlich meist drei Bereiche:

- *Infrastruktur und Datenmanagement*
  Typische Inhalte im Bereich „Infrastruktur und Datenmanagement" sind Cloud Computing, verteilte Systeme und Parallelisierung, MapReduce und das Hadoop-Ecosystem, Stream- und Event-Processing, approximative Algorithmen, Graphenverarbeitung, Datenintegration, -ex- und -import.
- *Analytik*
  Im Bereich „Analytik" geht es um Daten-Aufbereitung und -Qualität, Statistik, fortgeschrittene Analytik, Modellierung, Data Mining und maschinelles Lernen, R, Mahout, Bezug zu BI, Textanalyse und Visualisierung.
- *Anwendung*
  Themen im anwendungsspezifischen Teil der Seminare sind Einsatzmöglichkeiten, Fallbeispiele, Praxisberichte, Wirtschaftlichkeit, Geschäftsmodelle, Datensicherheit, -governance und -schutz sowie Vorgehensweisen im Unternehmen.

Das Fraunhofer-Institut für Intelligente Analyse- und Informationssysteme IAIS bietet seit Anfang 2013 Schulungen zu Big Data Architecture, Basic Analytics, Text Analytics und Visual Analytics an [4]. Das Angebot wird im Jahr 2014 im Rahmen der Fraunhofer-Initiative „Big Data" (siehe www.big-data.fraunhofer.de) erweitert.

Der Auf- und Ausbau von Qualifizierungsprogrammen für Data Scientists wird in den nächsten Jahren öffentlich gefördert, zum Beispiel als Teil eines Fünf-Punkte-Plans für Big Data in Frankreich, im Horizon-2020-Programm der Europäischen Kommission und voraussichtlich auch durch die Big-Data-Kompetenzzentren, die das Bundesministerium für Bildung und Forschung in Deutschland ausgeschrieben hat.

Quellen:

[1] http://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century
[2] www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation
[3] www.journaldunet.com/solutions/emploi-rh/plan-numerique-big-data-0713.shtml
[4] www.iais.fraunhofer.de/data-scientist.html

Angi Voss
angelika.voss@iais.fraunhofer.de