Woher sollen all die Data Scientists kommen?

Von Thomas Kestler, 14.10.2019

Der zukünftige Bedarf an neuen Berufsbildern wie Data Scientist, Data Engineer, Data Analyst etc. ist gewaltig. Welche Anforderungen müssen Kandidaten dafür erfüllen und woher sollen diese überhaupt kommen?

Digitalisierung und Automatisierung verändern gerade die gesamte Wirtschaft. Nicht nur die Automobilindustrie bekommt dies mit voller Wucht zu spüren. China demonstriert, wie Techniken, die vor wenigen Jahren als Science Fiction erschienen, heute Alltag sind. Mehr und mehr Arbeitsschritte können automatisiert werden, so dass sich die Arbeitswelt in den nächsten Jahren massiv verändern wird. Künstliche Intelligenz (AI – Artificial Intelligence) ist ein Kernthema dieser Entwicklungen. Das autonome Fahren zeigt gerade auf, was bereits möglich ist, wo aber auch die Hürden liegen, die genommen werden müssen. In erster Linie werden immens viele Daten zum Training der AI benötigt. Real können diese Daten gar nicht mehr gefahren werden, deshalb findet das Training der Fahrsituationen längst bereits größtenteils in der virtuellen Realität (Simulation) statt. Dafür sind immense Investitionen an HW und SW nötig. Und an Datenexperten, welche diese Daten erzeugen, verarbeiten und analysieren können. Das bringt uns zum Thema.

Neue Job-Rollen

Es entstehen gerade neue Berufsbilder (Job-Rollen), die sich allerdings noch nicht stabilisiert haben. Aktuell zeichnen sich drei Job-Rollen im Kern ab:

Job-Rolle	**Beschreibung**	**Gehaltsniveau**
Data Scientist	Analyse großer, komplexer Datenmengen, fund. Math. Wissen, Programmier-kenntnisse, Kenntnis stat.Tools	+++++
Data Engineer	Verarbeitung, Transport, Filterung, Speicherung, Konvertierung von Daten. Stark auf Programmierung konzentriert, aber nicht nur	++++
Data Analyst	Analyse von Daten und dere Visualisierung, Erstellung und Bereitstellung von Dashboards, Scorecards, Reports, …	+++

Diese Job-Rollen gliedern sich allerdings in einen größeren Kontext von Job-Rollen rund um das Thema datengetriebenes Management ein, siehe Studie der Royal Society[1].

Es muss hervorgehoben werden, dass die Rollenbeschreibungen noch fließend sind, es gibt derzeit keine Standardisierung der Berufsbilder (auch keine Industriestandards o.ä.). Data Science hat sich auch noch nicht als eigenständige Kategorie etabliert, oft finden sich die benötigten Themen unter Big Data, Database Management, Programming und weiteren verteilt.

Im Folgenden sollen die Anforderungen an die drei Job-Rollen näher ausgeführt werden. Die Reihenfolge stellt auch kein Ranking dar, auch die obigen Gehaltsklassen sind nur ungefähr, teilweise werden Data Engineers höher eingestuft als Data Scientists. Am Ende kommt es

1https://royalsociety.org/-/media/policy/projects/dynamics-of-data-science/dynamics-of-data-science-skills-report.pdf

darauf an, wie dringend ein Unternehmen[2] einen Experten mit dem benötigten Wissen braucht.

Im Anhang findet sich eine Tabelle der meistgeforderten Skills für Data Scientists (und implizit Engineers).

Data Scientist

Der Data Scientist hat ein fundiertes mathematisches Wissen, insbesondere Stochastik und Statistik. Er muss die erfolgten Analysen kritisch beurteilen können, um nicht Fehlinterpretationen aufzusitzen. Darüber hinaus benötigt er umfassendes Wissen über Datenhaltung (Big Data, RDBMS) und Verarbeitung (Hadoop, Spark, Kafka, …). Um handlungsfähig zu sein, muss er entsprechende Programmierkenntnisse (Python, R, SQL) haben. Tools wie SAS, Tableau, etc. sollte er auch gut beherrschen. Um neue Erkenntnisse aus großen Datenmengen zu erarbeiten muss er maschinelles Lernen in Theorie und Praxis beherrschen.

Die ermittelten Erkenntnisse muss er dann auch entsprechend kommunizieren können (Communication Skills). Zudem muss er die Anforderungen der Fachabteilungen erfassen und in Analyseaufträge umsetzen können.

Üblicherweise arbeitet der Data Scientist eng mir einem Data Engineer zusammen, der für die Bereitstellung der Daten zuständig ist (siehe dort).

Data Engineer

Der Data Engineer ist verantwortlich für die Bereitstellung der Daten aus unterschiedlichsten Systemen. Heute sind Data Warehouses und Data Lakes üblich. Daten müssen aus unterschiedlichen Systemen gesammelt, normiert, konsolidiert und bereinigt werden. ETL-Techniken muss er bestens beherrschen, ebenso wie SQL und Big-Data-Systeme. Die zunehmende Digitalisierung macht eine direkte Verarbeitung durch Data Streaming (z.B: Apache Kafka) nötig. Selbstredend muss er fundierte Programmierkenntnisse besitzen.

Bei solch riesigen Datenmengen ist es zwingend erforderlich die Qualität der Daten beurteilen zu können. Deshalb ist Mathematik zwingend nötig und die Kenntnis von Data Quality Methoden und Tools.

Wer Programme erstellt muss dies immer auch im Rahmen der bestehenden Policies tun, also in Hinblick auf Sicherheit, Wartbarkeit, etc. Dazu gehört, dass jeglicher Source-Code in ein zentrales Repository gehört und das Deployment den etablierten Standards genügt.

Data Analyst

Der Data Analyst soll vorwiegend auf Anforderung Daten visualisieren, z.B. in Form von Dashboards. Daher muss er sich sehr gut mit Tools wie Tableau, Qlik oder anderen auskennen und Excel sicher beherrschen (weil viele Daten aus den Abteilungen in Excel verarbeitet werden). SQL muss auf hohem Niveau beherrscht werden, Programmierkenntnisse sind von Vorteil. Er muss diese Daten auch so darstellen können, dass die Konsumenten diese sicher interpretieren und einordnen können.

Zwischenfazit

Nach kurzer Vorstellung der drei Job-Rollen wird schnell ersichtlich, dass gerade die Anforderungen an Data Scientist und Data Engineer schon fast utopisch erscheinen. Man wird sicher Kandidaten finden, welche ein oder mehrere der geforderten Skills mitbringen, aber nur sehr schwer jemanden, der alle Skills in der entsprechenden Tiefe (Berufspraxis) mitbringt.

2 Nachfolgend sind gleichbedeutend Unternehmen und Organisationen gemeint

Für alle drei Job-Rollen gilt zudem, dass ein tiefes Domain-Wissen nötig ist. Die heutige Industrie ist hochkomplex und es dauert viele Jahre im Beruf, bis man Technik und Prozesse wirklich verstanden hat.

Der Umgang mit Daten erfordert immer auch Verantwortung, ein Data Scientist trägt hier teilweise eine erhebliche Verantwortung, z.B. wenn er Fehlverhalten anhand einer Datenanalyse vermutet. Gerade Data Scientist und Data Engineer haben meist Zugriffe auf sämtliche Daten eines Unternehmens.

Die erzielten Ergebnisse müssen immer kritisch hinterfragt werden. Die Datenanalyse hält viele Stolperstricke bereit. Alle Ergebnisse sollten nachvollziehbar dokumentiert werden, so dass sie auch später verifiziert werden können. Wissenschaftliches Arbeiten ist hier angesagt, keine Schnellschüsse.

Potenziale für Kandidaten

Woher sollen nun also Kandidaten für obige Positionen gewonnen werden? Das McKenzie Global Institute hatte 2013 einen Fehlbedarf von 190.000 Data Scientists bis 2020 prognostiziert (Links siehe Weblinks am Ende).

Möglicherweise hilft die Suche bei Job-Portalen wie Pivigo, Monster oder StepStone. Es mag sein, das große Unternehmen die wenigen Talente gewinnen können, alle anderen dürften aber Schwierigkeiten haben, denn es gibt einfach nicht genügend Kandidaten mit den nötigen Skills.

Laut SpiegelOnline ist die derzeit Nachfrage nach Datenexperten gemäß Stellenanzeigen noch sehr verhalten[3]. Möglicherweise wollen Unternehmen die Themen zunächst mit externen Beratern angehen, das würde die höhere Anzahl an Projekten als Job-Angeboten auf GULP erklären[4]. Insgesamt scheint die aktuelle Nachfrage gemessen an der strategischen Bedeutung aber noch zu gering – Deutschland verschläft evtl. mal wieder wichtige Innovationen.

Aus- und Weiterbildung

Der Bereich der (dualen) Ausbildung ist streng reglementiert und kurzfristig kaum zu adaptieren. Es scheint auch illusorisch, im straffen Ausbildungsprogramm solche Themengebiete wie oben beschrieben unterzubringen. Bleibt also die Weiterbildung übrig.

Jedes einzelne Themengebiet ist für sich bereits sehr anspruchsvoll. Daher müssen die Themengebiete weiter segmentiert und aufbereitet werden.

Die fundamentale Basis stellt die angewandte Mathematik dar, mit Schwerpunkt auf Statistik. Leider finden sich hier Defizite auf breiter Basis, außer der Kandidat hat Mathematik studiert. Selbst bei Ingenieuren bestehen teilweise Defizite in Grundlagen und Anwendung.

Der Großteil der Daten in Unternehmen wird in relationalen Datenbank gespeichert, daher ist eine sichere Anwendung von SQL wichtig. Reine Grundkurse reichen hier nicht aus, sie müssen ergänzt werden durch angewandte Praxis. Sofern Big Data zum Einsatz kommt, sind hier weitere Qualifikationen nötig.

Programmierkenntnisse sind vor allem für den Data Engineer essentiell, aber auch für den Data Scientist. Programmiersprachen wie Python, R, Java, etc. müssen sicher beherrscht

3 https://www.spiegel.de/netzwelt/web/stellenausschreibungen-von-dax-konzernen-ki-kenntnisse-kaum-gefragt-a-1290292.html
4 https://www.gulp.de/gulp2/g/projekte?query=big%20data

werden. Dazu muss auch das gesamte Ökosystem um eine Srache herum (Entwicklungswerkzeuge, Bibliotheken, Frameworks) sicher beherrscht werden (Stichwort: Nutzung OpenSource).

Das maschinelle Lernen gewinnt zunehmend an Bedeutung bei der Datenanalyse. Es existieren bereits zahlreiche Lösungen für die praktische Anwendung. Cloud-Provider (z.B: Microsoft Azure, Ámazon AWS) stellen Services dafür bereit.

Für die reine Nutzung von Analysetools wie z.B: Qlik oder Tableau existieren ausreichend Schulungs- und Informationsangebote.

Innerbetriebliche Weiterbildungspfade

Für eine innerbetriebliche Weiterbildung sollten definierte Ausbildungspfade erstellt werden, so dass Mitarbeiter den Weg und das Ziel absehen können. Dabei sollte es nicht nur den Mount Everest (Data Scientist) als Ziel geben, sondern eine stufenweise Qualifikation entlang der Leiter der datengetriebenen Berufsbilder. Dies kann der IT-Projektmanager sein, der Projektdaten analysieren muss, der Marketing Manager, der die ihm vorgelegten Analysen auch bewerten und prüfen können muss. Kurzum, Datenanalyse wird für immer mehr Berufsbilder immer wichtiger und Unternehmen sollten sich darauf einstellen und entsprechende Weiterbildungsangebote bieten. Große Unternehmen sollte überlegen die Expertise in einem Center of Competence zu bündeln.

Gewisse Commodities können extern zugekauft werden, z.B. SQL-Schulung, Programmierschulung (Grundlagen), etc. Ob außer Haus, intern, ob als Webinar oder MOOC, das bleibt zur Auswahl. Gerade Grundkenntnisse der Programmierung und deren Anwendung im beruflichen Alltag bieten enormes Potenzial (alleine das Mindset: „wenn ich etwas nicht zur Verfügung habe, dann schaffe ich es mir selbst“).

Wichtig erscheint es, die Weiterbildungspfade transparent und in machbaren Schritten zu planen und realisieren. Berufsbegleitende Weiterbildung fordert den Mitarbeitern viel ab und wird in kleinen Schritten meist besser angenommen. Während der beruflichen Entwicklung können sich Neigungen und Bedürfnisse zeigen, welche den Pfad in eine bestimmte Richtung lenken, z.B. verstärkte Vertiefung im Bereich Oracle bis letztlich zum Oracle Certified DBA (Datenbankadministrator). Selbst wenn hier kein Data Scientist gewonnen werden konnte, ist der Gewinn eines qualifizierten DBA trotzdem sehr wertvoll.

Universitäten

Mehrere deutsche Universitäten Universitäten bieten bereits Studiengänge zum Thema Data Science an[5]. Stellvertretend sei der Masterstudiengang „Data Science“ an der LMU in München genannt[6]. Das Curriculum[7] umfasst Statistik, Informatik. Grundlagen der Data Science, Predictive Modelling, Human Computation and Analytics, Data Ethics and Security sowie den aktuellen Stand der Data Science wie weitere Wahlfächer. Im dritten Semester kommen praktische Übungen dazu, wobei angesichts des Themenumfangs fraglich ist, ob ein Semester hier ausreichend ist. Daher werden bereits entsprechende Vorkenntnisse in Data Mining, Wahrscheinlichkeitstheorie, Maschinellem Lernen, etc gefordert[8].

---

5 https://studieren.de/data-science.hochschulliste.t-0.c-41509.html https://de.wikipedia.org/wiki/Data_Science Abschnitt Ausbildungsmöglichkeiten

6 https://www.m-datascience.mathematik-informatik-statistik.uni-muenchen.de/index.html

7 https://www.m-datascience.mathematik-informatik-statistik.uni-muenchen.de/program/curriculum/index.html

8 https://www.m-datascience.mathematik-informatik-statistik.uni-muenchen.de/requirements/index.html

Forschung versus Industrie

Im universitären Forschungsbetrieb finden sich zahlreiche Wissenschaftler mit den Skills eines Data Scientists. Die Gewinnung von solchen Forschern für die Industrie ist aber oft nicht einfach, da Forscher fürchten dauerhaft aus dem Wissenschaftsbetrieb auszuscheiden (ohne laufende Publikationen keine Anerkennung und keine Rückkehr).

Große Unternehmen können Wissenschaftlern sicher attraktive Konditionen anbieten, so dass sie nicht fürchten müssen, den Anschluss an den Wissenschaftsbetrieb zu verlieren.

Seminare und Schulungsangebote

Es gibt bereits etliche Anbieter von Schulungsangeboten mit und ohne Zertifizierung, so z.B. das Fraunhofer Institut[9]. Doch selbst bei einem Umfang von 20 Seminartagen können die Themen hier nur oberflächlich gestreift werden.

Wenn schon der Zeitraum eines Masterstudiums zum Beherrschen der nötigen Skill sehr knapp erscheint, können externe Seminare nur sehr begrenzt nützen. Dann dafür ein „offizielles“ Zertifikat zu vergeben mutet mindestens gewagt an. Dann sollte man sich doch lieber für das Masterstudium entscheiden.

Zielführender scheint es, die einzelnen Skills gezielt über Jahre hinweg – und dann gerne auch auf Basis externer Seminare als Einstieg – zu entwickeln. Dafür braucht es einen langen Atem.

Kongresse

Mittlerweile wurden und werden auch Kongresse zu dem Thema veranstaltet. Hier eine willkürliche Auswahl:

- German Data Scientist Days, Uni München - https://www.gdsd.statistik.uni-muenchen.de/2020/index.html
- Data Driven Business, Berlin - https://datadrivenbusiness.de/
- Data Festival, München - https://www.datafestival.de/
- Data2day, Ludwigshafen - https://www.data2day.de/
- Datascience Ruhr, Bochum - https://www.data-science.ruhr/kongress-2019/
- Big Data Minds, Berlin - https://www.big-data-minds.com/
- Gartner Analytics Conference, London - https://www.gartner.com/en/conferences/emea/data-analytics-uk

9 https://www.bigdata.fraunhofer.de/de/datascientist/zertifizierungen/machine_learning_zertifizierung.html

Zusammenfassung

Daten und deren Analyse werden immer wichtiger. Die digitale Revolution ist nicht mehr zu stoppen. Länder wie China demonstrieren was bereits möglich ist. Deutschland scheint bei Big Data und Data Science den Anschluss zu verpassen.

Es gibt noch keine klaren Definitionen der Berufsbilder für Datenexperten. Trotzdem ist es bereits heute möglich die benötigten Skills zu identifizieren und aufzubauen. Mehr und mehr Berufsbilder werden diese Skills zunehmend benötigen. Unternehmen und Organisationen sollten Weiterbildungspfade dafür bereitstellen.

Die Anzahl und Komplexität der geforderten Skills für einen Datenexperten ist erheblich, es dauert viele Jahre zum Aufbau solcher Kompetenzen. Schnelle Erfolge sind nicht zu erwarten. Eine stufenweise, berufsbegleitende Weiterbildung scheint am ehesten zielführend.

Es ist wichtig, dass zukünftig **alle** abgehenden Ingenieure (pardon: Bachelor, Master, …) über Grundkenntnisse und Anwendungswissen zur Datenanalyse verfügen. Darüber hinaus sollte das Thema auch im System der (dualen) Berufsausbildung soweit integriert werden, dass die Abgänger Möglichkeiten, Potenziale und Grenzen von Data Science und AI grob einschätzen können und wissen, wie und wo sie sich bei Bedarf weiter informieren können.

Die Nutzung datengetriebener Prozesse wird zukünftig für immer mehr Mitarbeiter im Unternehmen wichtig. Eine möglichst breite Qualifizierung ist daher nötig.

Anhang

Nennung der wichtigsten Skills für Data Scientist verschiedener Quellen. X=explizit gennant, I=implizit.

Skill	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
Machine Learning / AI	X	X	X		X		X		X	I	X		X	X	
Python	X	X		I	X		I	X	X	X		X	X	X	X
R	X	X		I	X	X	I	X	X	X	X		X		X
Programming/Coding (incl. Java, Scala)	I	X		X	I	X	X	I	I	I	I			X	I
SAS or Other Analytical Tools				X		X			X						X
SQL	X	X			X			X	X	X	X		X	X	X
Hadoop/Spark	X	I		X				X	X	X		X	X	X	X
Math / Statistics		X	X	X	X	X	X	X	X	X	X			I	X
Critical Thinking		X													
Communication Skills		X	X		X	X			X		X			X	I
Data Viszualization		X			X		X	X				X			X
Unstructured Data		X		X							X				
Data Architecture			X						X						
Risk Analysis, processes/system eng.			X												
Data Mining and Processing				X					X						
Apache Kafka									X						
Business / Domain Knowledge						X									
Big Data							X						X		
Data Ingestion/Wrangling							X		X						
TensorFLow										X			X		
Jupyter/Zeppelin										X					
GPU & CUDA												X			

1 https://www.techrepublic.com/article/top-5-tech-skills-data-scientists-need-and-how-to-learn-them/
2 https://www.techrepublic.com/article/top-5-tech-skills-data-scientists-need-and-how-to-learn-them/
3 https://www.cio.com/article/3263790/the-essential-skills-and-traits-of-an-expert-data-scientist.html
4 https://www.simplilearn.com/what-skills-do-i-need-to-become-a-data-scientist-article
5 https://blog.udacity.com/2014/11/data-science-job-skills.html
6 https://analyticstraining.com/5-must-skills-need-become-data-scientist/
7 https://www.edureka.co/blog/how-to-become-a-data-scientist/
8 https://www.mastersindatascience.org/data-scientist-skills/
9 https://www.ubuntupit.com/best-20-data-scientist-skills-that-you-need-to-get-data-science-jobs/
10 https://www.skillbyte.de/must-have-ressourcen-skills-und-techniken-fuer-data-engineers-und-data-scientists/
11 https://www.crampete.com/resources/blogs/11-top-data-science-skills/
12 https://analyticsindiamag.com/9-skills-a-data-scientist-must-have-to-land-a-job-aim-skills-study-2019/
13 https://cvcompiler.com/blog/how-to-become-more-marketable-as-a-data-scientist/
14 https://www.techrepublic.com/article/top-5-tech-skills-data-scientists-need-and-how-to-learn-them/
15 https://www.gulp.de/knowledge-base/17/i/skills-fur-big-data.html

Weblinks

Scikit-learn – Python-basierter Toolkit für maschinelles Lernen:
https://scikit-learn.org/stable/

WEKA – Java-basierter OpenSource Toolkit für maschinelles Lernen der Waikatao Universtiy, New Zealand:
https://ai.waikato.ac.nz/weka/

Und der komplette MOOC dazu:
https://www.cs.waikato.ac.nz/ml/weka/mooc/dataminingwithweka/

Die Waikato University bietet mit MOA und ADAMS weitere Toolkits für streambasierte Analyse an.

Data Science auf Basis von Open Source als Einstieg, in der Praxis gibt es aber nicht nur OpenSource-Produkte. Trotzdem gut zur Orientierung:
http://datasciencemasters.org/

McKinsey Gobal Institute – US Game Changers July 2013. Nach sechs Jahren interessant zu sehen, inwieweit sich die damalig Prognosen zu bis zu 190.000 Data Scientists manifestiert haben:
https://www.mckinsey.com/featured-insights/americas/us-game-changers

Wikipdia-Artikel zu Data Science:
https://de.wikipedia.org/wiki/Data_Science