

# Data Science für Lehre, Forschung und Praxis

Kurt Stockinger · Thilo Stadelmann

Eingegangen: 6. Februar 2014 / Angenommen: 31. März 2014  
© Springer Fachmedien Wiesbaden 2014

**Zusammenfassung** Data Science ist in aller Munde. Nicht nur wird an Konferenzen zu Big Data, Cloud Computing oder Data Warehousing darüber gesprochen: Glaubt man dem McKinsey Global Institute, so wird es alleine in den USA in den nächsten Jahren eine Lücke von bis zu 190.000 Data Scientists geben (Manyika et al. Big data: the next frontier for innovation, competition, and productivity, Report. [www.mckinsey.com/insights/business\\_technology/big\\_data\\_the\\_next\\_frontier\\_for\\_innovation](http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation), 2011). In diesem Kapitel beleuchten wir daher zunächst die Hintergründe des Begriffs Data Science. Dann präsentieren wir typische Anwendungsfälle und Lösungsstrategien auch aus dem Big Data Umfeld. Schließlich zeigen wir am Beispiel des Diploma of Advanced Studies in Data Science der ZHAW Möglichkeiten auf, selber aktiv zu werden.

**Schlüsselwörter** Datenmanagement · Data Warehousing · Datenanalyse

## 1 Einleitung

Data Science ist als neuer Begriff seit einigen Jahren an fast jeder Konferenz zu Datenmanagement, Data Warehousing oder Datenanalyse zu hören. Gemäß Harvard Business Review (Davenport und Patil 2012) gilt der Beruf des Data Scientists sogar als der „sexiest Job“ des 21. Jahrhunderts. Unternehmen suchen verzweifelt nach dieser begehrten Spezies, die schon fast den Charakter eines Universalgenies zu verkörpern scheint (Stockinger 2013).

---

K. Stockinger (✉) · T. Stadelmann  
Winterthur, Schweiz  
E-Mail: [stog@zhaw.ch](mailto:stog@zhaw.ch)

Manche Skeptiker behaupten, alles sei nur eine trendige Wortneuschöpfung, die ausgehend von den USA langsam auf Europa überschwappt. Andere wiederum meinen, dass Data Scientists nur für Big Data wirklich notwendig seien – und Big Data gäbe es nur in einigen wenigen Internetfirmen an der US-Westküste. Hingegen sieht man bereits auch im deutschsprachigen Raum Stelleninserate von großen Unternehmen, die nach Data Scientists Ausschau halten.

Wie soll man nun mit diesen widersprüchlichen Meinungen umgehen, und was verbirgt sich wirklich hinter dem Begriff Data Science sowie dem neuen Berufsbild des Data Scientists? Was sind die Kernpunkte dieser neuen Disziplin, und wie kann man sich zum Data Scientist aus- bzw. weiterbilden?

In diesem Artikel wollen wir uns diesen Fragestellungen widmen und entsprechende Antworten geben. Wir geben einen historischen Überblick über die Entstehung von immer größeren Datenmengen und definieren danach den Begriff Data Science. Im Anschluss daran zeigen wir typische Data Science Use Cases auf und besprechen aktuelle Trends, die sowohl für die Forschung als auch für die Wirtschaft von Interesse sind. Im letzten Schritt skizzieren wir den Inhalt eines Curriculums für Data Science, welches Data Scientists der Zukunft für die Herausforderungen der Unternehmen in Richtung Daten-basiertes Entscheiden vorbereitet.

## 2 Data Science – Hintergrund und Definition

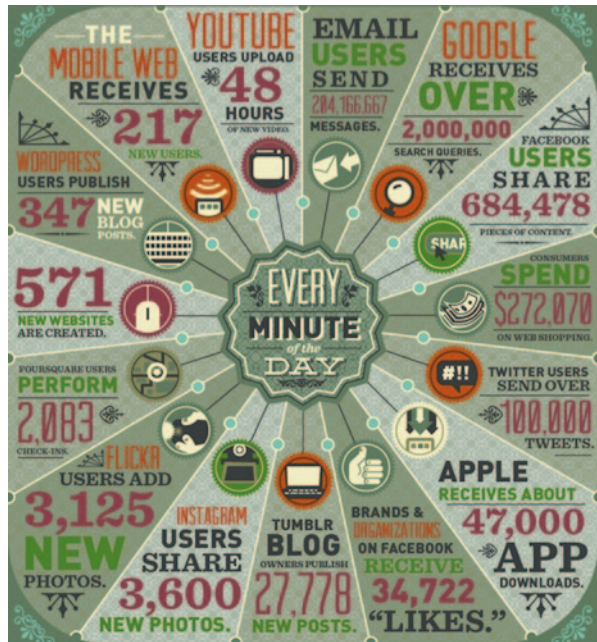
### 2.1 Wie entstand Big Data

Data Science gründet in der Verfügbarkeit von Daten. Diese nehmen nicht erst seit dem Aufkommen des Datendursts von Google und der Prägung des Begriffs „Big Data“ rapide an Menge, Vielfalt und Geschwindigkeit zu (siehe die „drei V’s“ in der üblichen Definition von Big Data: Volume, Variety und Velocity (Soubra 2012)). Vorläufer auf dem Gebiet Big Data und somit die **erste Welle des Datentsunami** sind **Grossforschungsanlagen** wie etwa CERN oder das Hubble Space Telescope. Beispielsweise würden Experimente am CERN in nur einer Sekunde hunderte von Terabytes an Daten produzieren, würde nicht ein Hardwarefilter eine Grundselektion durchführen.

Um derart große Datenmengen zu verwalten, wurde bereits Ende des 20. Jahrhunderts im Rahmen des RD45 Projektes (RD45 2001) ein Datenbanksystem entwickelt, das Petabytes an Daten speichern sollte. Zu diesem Zeitpunkt gab es auf dem Markt noch kein Datenbanksystem, das derartige Datenmengen verarbeiten konnte (Düllmann 1999). Aus Sicht der Daten handelt es sich bei wissenschaftlichen Experimenten oftmals um numerische Daten (z. B. Temperatur, Geschwindigkeit, Anzahl an Teilchenkollisionen) oder Bilder (z. B. von Sternen oder Galaxien).

Ein weiterer Grund des Datenwachstums in den **Wissenschaften** ist auf einen **Paradigmenwechsel** (Hey et al. 2009) zurückzuführen. So bestand das erste Paradigma der Wissenschaften vor allem darin, theoretische Studien durchzuführen. Im zweiten Paradigma wurden die Theorien experimentell nachgewiesen. Großexperimente wie am CERN sind jedoch sehr komplex und vor allem teuer. So entstand als drittes Paradigma die Computersimulation als Grundlage für wissenschaftliche Erkenntnisse (Computational Science). Als 4. Paradigma gelten Daten-intensive Wissenschaften.

**Abb. 1** Generiertes Datenvolumen im Internet pro Minute nach (James 2012) (Stand: Juni 2012)



Nach den Großforschungsexperimenten wurde die **zweite Welle** von Big Data durch **Internetfirmen** des angehenden 21. Jahrhunderts wie etwa Google, Yahoo oder Amazon eingeläutet. Im Unterschied zu den wissenschaftlichen Daten verwalteten diese Firmen zu Beginn vor allem Textdaten. Durch das Indizieren von Bildern und Videos kam es später zu einer zusätzlichen Datenexpansion, die weiter anhält.

Als **dritte Welle** des Datensunamis gelten **soziale Netzwerke** wie etwa Facebook oder LinkedIn. Hier können nun einzelne Personen – im Unterschied zu Grossforschungsanlagen – zum Anwachsen des Datenvolumens beitragen.

Als **vierte Welle** beobachten wir aktuell die Zunahme von **Maschinen-generierten Daten**, etwa Logfiles oder Sensordaten im Internet of Things.

Einen Überblick darüber, wie viele Daten pro Minute im Internet bereits im Juni 2012 generiert worden sind, liefert Abb. 1 (James 2012).

Letztlich steht der Begriff Big Data jedoch nicht mehr nur für die schiere Menge, Variabilität oder Ausbreitungsgeschwindigkeit von Daten, sondern eher für ein neues Paradigma im Umgang mit ihnen: Daten gelten als verfügbar (zu jedem Thema und aus unterschiedlichsten Perspektiven gesammelt), und es setzt sich die Ansicht durch, dass in deren Auswertung leicht zu schöpferischer Wert für ein Unternehmen oder die Gesellschaft als ganzes liegt – „Data“ ermöglicht „Big Gain“. Im Sinne der 3 V's überwältigend „Big“ ausfallende Daten wird es hingegen immer geben.

## 2.2 Was ist Data Science?

Nachdem wir uns mit der Entstehungsgeschichte der Datenflut auseinandergesetzt haben, wollen wir uns nun dem Begriff **Data Science** widmen.

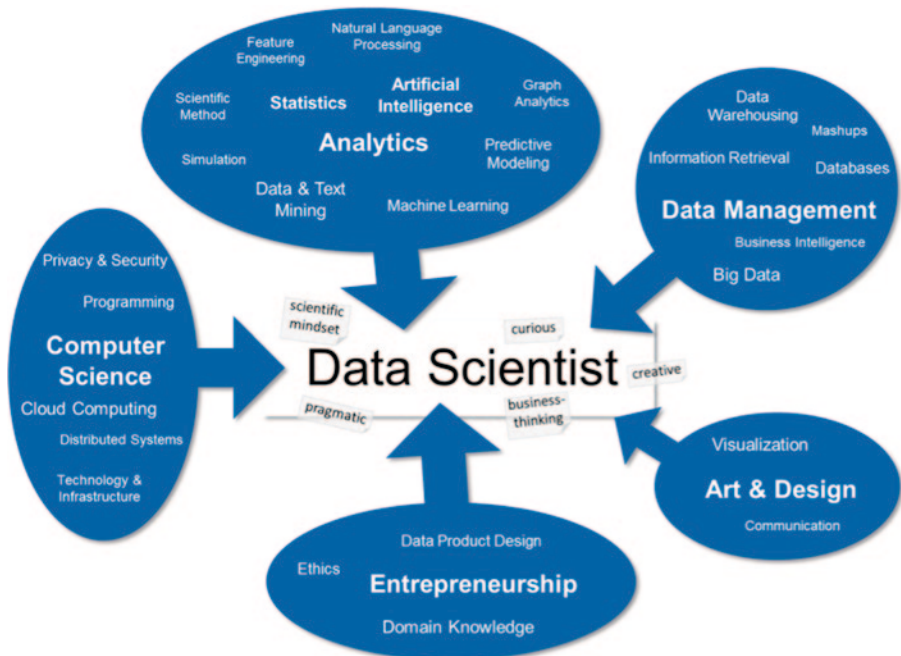


Abb. 2 Die Data Science Skill Set Map

Data Science ist eine **interdisziplinäre Wissenschaft**, die Methoden und Ansätze zur Auswertung *unterschiedlichster* Arten von Daten mit *verschiedensten* Mitteln bündelt. Ausgehend von konkreten Fragestellungen wird ein **Data Product** entwickelt, d. h. eine neue Information oder ein neuer Service, der Wertschöpfung aus der Analyse bestehender Daten betreibt (Loukides 2010).

Abbildung 2 gibt einen Überblick über die Fähigkeiten und Eigenschaften eines Data Scientists in Form einer Landkarte (Stadelmann et al. 2013). Die blauen Gebiete auf dieser Karte entsprechen dabei wichtigen Kompetenzclustern, die sich der Data Scientist aus dem Repertoire teils mehrerer etablierter Teildisziplinen aneignet.

Drei Punkte lassen sich anhand dieser Skill Set Map festhalten, die wir in den nächsten drei Abschnitten genauer analysieren werden:

- In welchem Verhältnis steht Data Science zu anderen Teildisziplinen?
- Welche Kompetenzcluster zeichnen den Data Scientist aus?
- Welche Eigenschaften besitzen Data Scientists?

### 2.2.1 Auf den Schultern von Riesen

Eine humorvolle Definition beschreibt den Data Scientist als besser im Programmieren als der typische Statistiker sowie besser in Statistik als der typische Informatiker (Wills 2012). Das Data Science Venn Diagramm (Conway 2010) führt neben Mathematik und Statistik „Hacking Skills“ als Zutat für Data Science auf. DJ Patil, der den Begriff Data Science mitgeprägt hat, nennt Naturwissenschaften als häufigsten

Hintergrund seiner Teammitglieder (Patil 2011). Abbildung 2 schließlich nennt nur Begriffe, die bislang eindeutig einem oder mehreren etablierten Fachgebieten zuzuordnen sind.

Es ist schwierig, sich mit Data Science auseinanderzusetzen und nicht allenthalben diese inhärente Interdisziplinarität wahrzunehmen. Gleichzeitig wird nirgends der Anspruch erhoben, diese Begriffe seien ihren Ursprungsdisziplinen zu enteignen und dieser neuen „Disziplin-in-Entstehung“ einzuverleiben.

Vielmehr ist Data Science eine *einzigartige* Mischung dieser Skills aus Analytics, Engineering und Kommunikation, um ein spezifisches Ziel zu erreichen, nämlich die Erzeugung von (gesellschaftlichem oder betrieblichem Mehr-) Wert aus Daten. Als angewandte Wissenschaft lässt sie den Teildisziplinen ihren Wert und ist dennoch eigenständig und notwendig (siehe die Diskussion in (Provost und Fawcett 2013)).

Ob die Entwicklung dabei ähnlich verläuft zur Bildung der Subdisziplin Data Mining, die bis heute etwa in Statistik- und Informatik-Curricula zu finden ist, oder analog zum Herausschälen der Informatik aus den Fachgebieten Elektrotechnik und Mathematik, ist noch nicht auszumachen. Es ist anzunehmen, dass die Bündelung analytischen Wissens aller Fachgebiete in Forschungszentren und Ausbildungscurricula voranschreitet, wenn sich die prognostizierte Omnipräsenz analytischer Fragestellungen in Wirtschaft und Gesellschaft entsprechend entwickelt. Gleichzeitig beobachten wir momentan eher interdisziplinäre Initiativen und Kooperationen anstatt grundlegende Neuordnungen der akademischen Landkarte bezüglich der Fachgebiete.

### 2.2.2 Übergeordnete Kompetenzcluster

Data Science begrüßt Wissen und Praxis aller Teildisziplinen, die dem Ziel der Erschaffung von Mehrwert durch Datenanalyse dienen. Trotzdem sind die blauen Gebiete der Karte in Abb. 2 nicht nach Fachgebieten aufgeteilt, sondern nach den wichtigsten Kompetenzbereichen des Data Scientist:

**Computer Science und Data Management.** Der Umgang mit Daten ist so entscheidend, das Datenmanagement-Fähigkeiten als eigenes Kompetenzgebiet auftauchen, auch (aber keineswegs nur) „at scale“ im Umfeld von Big Data. Doch auch andere IT-Fähigkeiten aus der (Wirtschafts-)Informatik sind in der Praxis des Data Scientist wichtig, allen voran das Programmieren – jedoch eher im Sinne von Scripting als der Architektur großer Softwaresysteme.

**Analytics.** Analytische Fähigkeiten aus dem Bereich der Statistik, des maschinellen Lernens und der künstlichen Intelligenz zur Extraktion von Wissen aus Daten und zur Generierung von (Vorhersage-)Modellen sind die Kernfähigkeit des Data Scientists. Hierbei ist der unterschiedliche Zugang der Teildisziplinen, etwa von Statistikern und IT'ern bzgl. Modellierung (Breiman 2001), besonders wertvoll.

**Entrepreneurship.** Der Data Scientist hat nicht nur die Verantwortung zur Implementierung einer analytischen Lösung für ein gegebenes Problem, sondern er benötigt auch die Fähigkeit zum Stellen der richtigen Fragen bzgl. Business-Value sowie

Folgen für Betrieb und Gesellschaft, um sich seine analytischen Fragestellungen selbst aufzustellen. Dies bedeutet auch den Aufbau substanziellen Wissens aus der jeweiligen Fachdomäne.

**Art und Design.** Als Verantwortlicher für den gesamten analytischen Workflow kommuniziert der Data Scientist selbst seine Ergebnisse auf (Senior) Management Ebene. Dies benötigt neben adressatengerechter Kommunikation die Fähigkeit zur korrekten grafischen Aufbereitung komplexester Zusammenhänge mit Mitteln der Informationsvisualisierung. Auch als Teil analytischer Lösungen und Services für den Kunden sind angemessene grafische Darstellungen bedeutend. Gleichzeitig spielt die grafische Aufbereitung von (Zwischen-)Ergebnissen unter dem Stichwort Visual Analytics eine große Rolle.

Ein erfahrener Data Scientist sollte etwa 80% dieser Kompetenzlandkarte abdecken, verteilt über alle fünf blauen Ovale. Die notwendigen Fähigkeiten können trainiert werden. Ein typischer Werdegang beginnt mit einem Studium etwa in Statistik, Informatik oder datenintensiven Wissenschaften, von dem aus Fähigkeiten in den anderen Bereichen durch interdisziplinäre Arbeit und Weiterbildung hinzugewonnen werden.

### 2.2.3 Eigenschaften eines Data Scientists

Die zielgerichtete analytische Arbeit an Datensätzen erfordert bestimmte Eigenschaften auf Seiten des Data Scientists: Kreativität, Neugier und wissenschaftliche Denkweise fördern neuartige Erkenntnisse zu Tage. Unternehmerisches Denken hält dabei ein klares Ziel vor Augen. Pragmatismus sorgt für die notwendige Effizienz in einer komplexen Tool-Landschaft. Diese Eigenschaften sind schwer trainierbar, aber wichtig für den praktischen Erfolg.

## 3 Data Science Use Cases

In diesem Abschnitt stellen wir drei Use Cases vor, die unterschiedliche Aspekte von Data Science abdecken und somit diverse Data Science Skills benötigen.

Use Case 1 fällt in den Bereich Predictive Analytics und erfordert Fähigkeiten aus dem Bereich maschinelles Lernen bzw. Statistik. Use Case 2 hingegen erfordert Fähigkeiten unter anderem aus dem Bereich Information Retrieval, während Use Case 3 eine Kombination aus Data Warehousing, Information Retrieval und Semantic Web Know-How erfordert.

### 3.1 Use Case 1: Predictive Maintenance

Zwei Branchen stehen aktuell besonders im Fokus des professionellen Diskurses zu Data Science Anwendungsfällen: Zum einen der **(e-)Commerce** mit seinem Wunsch nach aussagekräftigen Benutzerprofilen und **analytischem CRM**; zum anderen die Finanzindustrie im Bereich automatischer **Transaktionsdurchführung** und **Betrugserkennung**.

Oft vergessen geht dabei, dass die Branchen des **Maschinen- und Anlagenbaus** sowie generell **produzierende Betriebe** zukünftig den größten Gewinn aus der Nutzung von Daten ziehen könnten (Gartner 2012): Die „alte“ Industrie, in der Europa und insbesondere der deutschsprachige Raum führend auf dem Weltmarkt ist, kann durch das, was teilweise als **4. Industrielle Revolution** oder Industrie 4.0 bezeichnet wird, große Optimierungspotentiale und neue Geschäftsmodelle umsetzen (Wahlster 2013). Grundlage hierfür sind sogenannte cyber-physikalische Systeme, d. h. mit dem Internet der Dinge verbundene Sensoren, die neue logistische Prozesse in der Herstellung und damit höhere Grade der Fabrikautomation und Individualfertigung in kleinsten Stückzahlen erlauben.

Ein weiterer Fall, wie Sensorendaten gerade im Maschinenbau eingesetzt werden können, ist **Predictive Maintenance**: Eine Maschine, ausgestattet mit Möglichkeiten zur Überwachung ihrer selbst, kann durch präzise Vorhersagemodelle den richtigen Zeitpunkt ihrer Wartung selbst bestimmen und darauf hinweisen. Dies führt zu neuen Kalkulationsmöglichkeiten für das After-Sales-Geschäft. Grundlage ist die kontinuierliche Überwachung bestimmter Betriebsparameter (Rundlauf von Achsen, Geräusch von Kugellagern; etc.) und eine Erkennung von substanziellen Veränderungen in den Zeitreihen der Messwerte.

### 3.2 Use Case 2: Data in Context

Viele Unternehmen werten bereits erfolgreich interne Daten beispielsweise in einem Data Warehouse aus, leiden aber sozusagen an Blind- und Taubheit, denn externe Daten (Wirtschaftsnachrichten, Wetterberichte, Social Media Eindrücke; etc.) werden nicht hinzugezogen. Somit fehlt der Kontext zur Interpretation etwa der eigenen Verkaufszahlen.

Das Zürcher Startup **Squirro** hat mit Hilfe von Technologie der **Zürcher Hochschule für Angewandte Wissenschaften (ZHAW)** ein System aufgebaut, das es ermöglicht, die eigenen Daten im Kontext dessen auszuwerten, was „draussen in der Welt“ geschieht und gemeldet wird.

Abbildung 3 zeigt einen Screenshot der Integration dieser Technologie in Qlik-View, in dem gerade Finanznachrichten aus dem Web als Hintergrundinformation zum gezeigten Börsenkurs dargestellt werden.

Die Technologie im Hintergrund basiert auf Information Retrieval Verfahren und Open Source Software zum Bau von Suchsystemen: Squirro durchsucht Quellen wie Twitter oder News-Aggregatoren im Web bezüglich aus den Geschäftsdaten extrahierter Stichwörter. Dabei wird tiefgehendes Sprach-Know-How angewandt, um die „semantische Lücke“ zu überbrücken und ein Matching zwischen Webmeldungen und eigenen Metadaten herzustellen.

### 3.3 Use Case 3: Search over Data Warehouse (SODA)

**SODA** ermöglicht eine Google-ähnliche Suche für Data Warehouses (Blunski et al. 2012) und wurde in einer Kollaboration von **Credit Suisse** und **ETH Zürich** im Rahmen des **Enterprise Computing Centers** entwickelt. Somit können auch End-User,



Abb. 3 Squirrel integriert beispielsweise Geschäftszahlen und Social Media Streams in QlikView

die weder SQL beherrschen noch das Datenmodell des Data Warehouses kennen, große Datenbanken einfach durchsuchen.

Vereinfacht dargestellt funktioniert SODA wie folgt. Zunächst gibt der End-User eine Suchanfrage ein, z. B. „Customers Zürich Financial Instruments“. Unter Zuhilfenahme von Metadaten (Datenmodell, Ontologien, Indizes sowohl auf Basisdaten als auch auf Datenmodelle, etc.) stellt SODA fest, welche Tabellen die gefragten Daten enthalten. Falls eine Query mit mehreren Suchworten auch mehrere Tabellen als Resultat liefert, werden die Beziehungen erkannt (Primary Key – Foreign Key Beziehungen zwischen Tabellen) und es wird ausführbares SQL generiert.

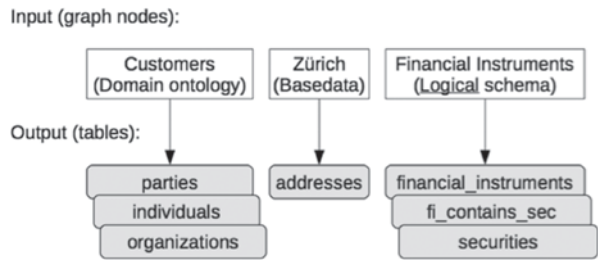
Nehmen wir als Beispiel wiederum unsere ursprüngliche Query an, nämlich „Customers Zürich Financial Instruments“. In diesem Fall werden die Suchbegriffe auf unterschiedliche Datenbanktabellen gemäß „**longest match**“ abgebildet, wie in Abb. 4 dargestellt ist. Beispielsweise bilden die beiden Suchbegriffe „Financial Instruments“ den longest match. Andererseits werden die Suchbegriffe „Customers“ und „Zürich“ direkt auf die entsprechenden Tabellen abgebildet.

SODA generiert nun automatisch ein korrektes SQL-Statement, das alle beteiligten Tabellen mittels eines Verbundoperators (join operator) verbindet.

Danach führt SODA die SQL-Anweisungen aus. SODA wurde erfolgreich mit dem Data Warehouse der Credit Suisse getestet, welches aus Hunderten von Tabellen besteht. Mehr Details sind im Artikel (Blunski et al. 2012) zu finden.



**Abb. 4** SODA analysiert die Suchanfragen und erkennt, welche Datenbanktabellen referenziert werden



## 4 Data Science Curriculum

Nachdem wir uns einige Use Cases aus dem Bereich Data Science angesehen haben, wollen wir nun besprechen, wie man sich zum Data Scientist aus- oder weiterbilden lassen kann.

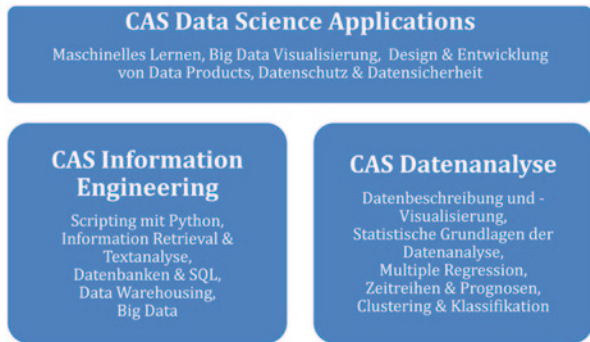
Data Science birgt große Chancen für Europa als Lebens- und Wirtschaftsraum, wenn es gelingt, ausreichend gut ausgebildete Fachleute mit dem nötigen Rüstzeug auszustatten. Als interdisziplinäres Fachgebiet ist Data Science inhärent komplex, und wie der „Engineering“-Anteil viel Erfahrung zur Beherrschung benötigt, erwartet der „Science“-Anteil tiefgehendes Verständnis wissenschaftlicher Zusammenhänge. Data Science auf die Erstellung und Interpretation von Business Intelligence Dashboards zu reduzieren, wäre sicher zu kurz gegriffen, auch wenn Business Intelligence Techniken sicher ein Teil des Pakets sind.




Daher sind solide Aus- und Weiterbildungsangebote von großer normativer Bedeutung: Sie setzen Standards für den Inhalt und definieren das Gebiet. Dadurch können Unternehmen planen und bekommen eine bessere Vorstellung darüber, was sie von ihren Data Scientists erwarten können. In den USA sind spezielle Data Science Angebote seit etwa 2012 zu beobachten wie z. B. an der Columbia University oder der online-Kurs an der UC Berkeley. In Europa gibt es wenige Masterstudiengänge in Großbritannien sowie durch Firmen oder Forschungsorganisationen organisierte Fortbildungen im Umfang weniger Stunden oder Tage.

Die **Zürcher Hochschule für Angewandte Wissenschaften (ZHAW)** hat einen kompletten Weiterbildungsstudiengang „**Diploma of Advanced Studies (DAS) in Data Science**“ entwickelt, der im Herbst 2014 das erste Mal vollumfänglich angeboten wird. Zuvor wurde in den vergangenen Semestern bereits mehrfach erfolgreich der Teilkurs „Certificate of Advanced Studies (CAS) in Datenanalyse“ durchgeführt.

Im Unterschied zu bereits existierenden Data Science Kursen an anderen Hochschulen richtet sich der DAS Data Science an der ZHAW vor allem an Personen, die eine Weiterbildung anstreben und somit bereits eine gewisse Praxiserfahrung im Berufsalltag mitbringen. Darüber hinaus liegt der Fokus auf starker Interdisziplinarität und Praxisausbildung. Dies wird dadurch gewährleistet, dass die Studierenden Fallbeispiele aus ihrer Unternehmung oder ihrem Institut mitbringen können, die dann in die Weiterbildung zum Data Scientist mit einfließen.

Der DAS Studiengang der ZHAW ist aufgebaut aus zwei Konvergenzmodulen (**CAS Datenanalyse** und **CAS Information Engineering**) sowie einer „Meisterklasse“ zur Vertiefung in ausführlichen Praxisthemen. Tabelle 1 gibt hier einen Überblick. Die beiden Konvergenz-CAS können unabhängig voneinander belegt und

**Tab. 1** Aufbau des DAS Data Science aus 3 CAS**Tab. 2** Inhalte eines Data Science Curriculums nach Abstraktions-Schichten

Schicht	Inhalt
 <p>Infrastruktur</p>	<ul style="list-style-type: none"> <li>• Datenbanken</li> <li>• Cloud Computing</li> <li>• Big Data Technologien</li> </ul>
 <p>Algorithmen</p>	<ul style="list-style-type: none"> <li>• Data Mining, Statistik &amp; Predictive Modeling</li> <li>• Maschinelles Lernen &amp; Graphenanalyse</li> <li>• Information Retrieval &amp; Sprachverarbeitung</li> <li>• Business Intelligence &amp; Visual Analytics</li> <li>• Data Warehousing &amp; Entscheidungsunterstützung</li> </ul>
 <p>Geschäft</p>	<ul style="list-style-type: none"> <li>• Visualisierung &amp; Kommunikation der Ergebnisse</li> <li>• Privatheit, Sicherheit &amp; Ethik</li> <li>• Unternehmertum &amp; Data Product Design</li> </ul>

besucht werden und vermitteln Data Science Grundlagen aus der Statistik und IT. Da jeder Teilnehmer seine eigene Erfahrung mitbringt, werden individuelle Vorkenntnisse nach Möglichkeit angerechnet.

Nach erfolgreicher Absolvierung der Konvergenzmodule kann der **CAS Data Science Applications** besucht werden. Hier wird intensiv und unter Zuhilfenahme aller Methoden und Technologien an Praxisbeispielen gearbeitet. Auch der nicht-technische Aspekt von Data Science wird aufgegriffen: Wie identifiziert man beispielsweise erfolgversprechende Use Cases (technisch und monetär gesehen)? wie wahrt man den gesetzlichen und ethischen Rahmen für die automatische Auswertung beliebiger Daten?

Tabelle 2 zeigt die Inhalte der Data Science Ausbildung aus einem anderen Blickwinkel, nämlich aufgeteilt in Schichten nach Abstraktionsgrad vom Business Case. Die Inhalte des Geschäfts-Layers liegen nah an den Use Cases der Praxis, die vom Data Scientist nicht nur oberflächlich verstanden werden müssen. Dies spielt in die Auswahl der Verfahren aus dem Algorithmen-Layer hinein, abstrahiert aber weitgehend von der technischen Infrastruktur. In der Ausbildung zusammengehalten werden die Inhalte mit durchgehenden Praxisbeispielen, die zur Erkennung und Umsetzung durchschlagender Business Cases beitragen sollen.

Der DAS Data Science der ZHAW orientiert sich an der in Abb. 2 vorgestellten Skill Set Map und berücksichtigt alle Schichten der Ausbildung nach Tab. 2. Er lässt sich berufsbegleitend je nach gewähltem Modell (Konvergenzmodule parallel oder sequentiell) in ca. ein bis zwei Jahren absolvieren, wobei pro CAS ein Präsenznachmittag in der Hochschule anfällt.

## Literatur

- Blunski L, Jossen C, Kossmann D, Mori M, Stockinger K (2012) SODA: generating SQL for business users. *Proceedings of very large databases*. PVLDB 5(10):932–943
- Breimann L (2001) Statistical modeling: the two cultures. *Stat Sci* 16(3):199–309
- Conway D (2010) The data science venn diagram, blog post. [drewconway.com/zia/2013/3/26/the-data-science-venn-diagram](http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram)
- Davenport TH, Patil DJ (2012) Data scientist: the sexiest job of the 21st century, Oktober 2012. [hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century/ar/1](http://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century/ar/1)
- Düllmann D (1999) Petabyte databases. SIGMOD Conference, Philadelphia
- Gartner (2012) Big data opportunity heat map by industry, Juli 2012. [b-i.forbesimg.com/louiscolombus/files/2013/08/big-data-heat-map-by-industry.jpg](http://b-i.forbesimg.com/louiscolombus/files/2013/08/big-data-heat-map-by-industry.jpg)
- Hey T, Tansley S, Tolle K (2009) The forth paradigm, microsoft research, Oktober 2009
- James J (2012) How much data is created every minute? Blog Post, Juni 2012. [www.domo.com/blog/2012/06/how-much-data-is-created-every-minute/?dkw=socf3](http://www.domo.com/blog/2012/06/how-much-data-is-created-every-minute/?dkw=socf3)
- Loukides M (2010) What is data science? Blog Post, Juni 2010. [radar.oreilly.com/2010/06/what-is-data-science.html](http://radar.oreilly.com/2010/06/what-is-data-science.html)
- Manyika J, Chui M, Brown B, Bughin J, Dobbs R, Roxburgh C, Byers AH (2001) Big data: the next frontier for innovation, competition, and productivity, Report, Mai 2001. [www.mckinsey.com/insights/business\\_technology/big\\_data\\_the\\_next\\_frontier\\_for\\_innovation](http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation)
- Patil DJ (2011) Building data science teams. Blog Post, September 2011. [radar.oreilly.com/2011/09/building-data-science-teams.html](http://radar.oreilly.com/2011/09/building-data-science-teams.html)
- Provost F, Fawcett T (2013) Data science and its relationship to big data and Data-Driven decision making, big data vol. 1, no. 1, March 2013
- RD45 (2001) A persistent object manager for HEP. [wwwasd.web.cern.ch/wwwasd/cernlib/rd45/](http://wwwasd.web.cern.ch/wwwasd/cernlib/rd45/)
- Stadelmann T, Stockinger K, Braschler M, Cieliebak M, Baudinot G, Dürr O, Ruckstuhl A (2013) Applied Data Science in Europe – Challenges for academia in keeping up with a highly demanded topic. In: European Computer Science Summit. ECSS 2013. August 2013, Amsterdam, The Netherlands, Informatics Europe
- Soubra D (2012) The 3Vs that define Big Data. Blog Post, Juli 2012. [www.datasciencecentral.com/forum/topics/the-3vs-that-define-big-data](http://www.datasciencecentral.com/forum/topics/the-3vs-that-define-big-data)
- Stockinger K (2013) Data Scientists – Die neuen Helden des 21. Jahrhunderts? Tagesanzeiger, Oktober 2013, Zürich
- Wahlster W (2013) Industry 4.0: the semantic product memory as a basis for cyber-physical production systems. [www.dfki.de/~wahlster/Vortrag\\_SGAICO\\_Zuerich\\_27\\_05\\_13/](http://www.dfki.de/~wahlster/Vortrag_SGAICO_Zuerich_27_05_13/). Zugegriffen: 27. Mai 2013
- Wills J (2012) Data Scientist (n.), Tweet, Mai 2012. [twitter.com/josh\\_wills/status/198093512149958656](https://twitter.com/josh_wills/status/198093512149958656)