Datenbank Spektrum
https://doi.org/10.1007/s13222-020-00354-8

SCHWERPUNKTBEITRAG

Ein Data Engineering Kurs für 10.000 Teilnehmer

Nicolas Alder[1] · Tobias Bleifuß[1] · Leon Bornemann[1] · Felix Naumann[1] · Tim Repke[1]

Eingegangen: 15. April 2020 / Angenommen: 20. August 2020
© Der/die Autor(en) 2020

Zusammenfassung
Im Januar und Februar 2020 boten wir auf der openHPI Plattform einen Massive Open Online Course (MOOC) mit dem Ziel an, Nicht-Fachleute in die Begriffe, Ideen, und Herausforderungen von Data Science einzuführen. In über hundert kleinen Kurseinheiten erläuterten wir über sechs Wochen hinweg ebenso viele Schlagworte. Wir berichten über den Aufbau des Kurses, unsere Ziele, die Interaktion mit den Teilnehmerinnen und Teilnehmern und die Ergebnisse des Kurses.

Schlüsselwörter MOOC · Massive Open Online Course · Data Science · openHPI

1 Data Engineering und Data Science: Klarheit in den Schlagwort-Dschungel

Schlagworten wie „AI“, „Data Science“ und „Machine Learning“ begegnet man in den Medien und im Alltag fast täglich. Vielen Menschen fehlt allerdings eine Vorstellung, was genau sich hinter diesen und weiteren Begriffen verbirgt. Im Idealfall wird eine für das Verständnis notwendige Erklärung mitgeliefert. Mit jeder neuen Begegnung wächst jedoch der Wunsch nach einem Überblick über die größeren Zusammenhänge und Implikationen auf die eigene Lebens- und Arbeitswelt.

In der akademischen Ausbildung an Universitäten erhalten mittlerweile wohl fast alle Studierenden der Informatik und verwandten Disziplinen einen Einblick in die Bedeutung dieser Begriffe und eine Vorstellung über die verwendeten Methoden und Techniken. Gerade durch ihre fortschreitende Einwirkung auf unterschiedlichste Bereiche unseres Lebens fällt es aber leicht zu argumentieren, dass jeder und jede unabhängig von fachlicher Spezialisierung eine Grundkenntnis dieser Themen haben sollte [1]. Idealerweise würde dieser Einblick schon in der Schule vermittelt [2]. Selbst wenn solche Themen flächendeckend in die Lehrpläne aufgenommen werden, bleiben ältere Generationen zurück, die den Einfluss der modernen datengetriebenen Technologien ebenso erleben.

Doch wie kann man dieses Wissen an einen größeren Teil der Bevölkerung vermitteln und dabei auch ältere Mitmenschen oder zum Beispiel Mitarbeiter kleinerer und mittelständischer Betriebe, sowie andere Interessierte erreichen? Eine Möglichkeit sind sogenannte MOOCs (Massive Open Online Courses), wie sie etwa auf der Plattform openHPI[1] angeboten werden. Diese MOOCs sind also offene, frei-zugängliche Kurse, die online einem breiten Publikum angeboten werden. Kurse auf der openHPI Plattform sind kostenlos. Dabei werden Vorlesungen in Form von aufgezeichneten Videos durch Quizze zur Selbstüberprüfung und Foren zum Austausch ergänzt. Der Erfolg der Teilnahme wird dabei durch Hausaufgaben und eine Klausur überprüft und durch ein Zertifikat am Ende des Kurses bescheinigt. Unsere Forschungsgruppe hat bereits 2013 als einen der ersten Kurse auf der Plattform eine Einführung in SQL angeboten [4].

Der besonders rege Zulauf unseres neuen Kurses zum Thema „*Data Engineering und Data Science: Klarheit in*

Aus Gründen der Lesbarkeit wählen wir im folgenden Text die männliche Form, nichtsdestoweniger beziehen sich die Angaben auf Angehörige aller Geschlechter.

Nicolas Alder
Nicolas.Alder@student.hpi.de

Tobias Bleifuß
Tobias.Bleifuss@hpi.de

Leon Bornemann
Leon.Bornemann@hpi.de

✉ Felix Naumann
Felix.Naumann@hpi.de

Tim Repke
Tim.Repke@hpi.de

[1] Hasso-Plattner-Institut, Universität Potsdam, Potsdam, Deutschland

[1] https://open.hpi.de.

Published online: 16 September 2020

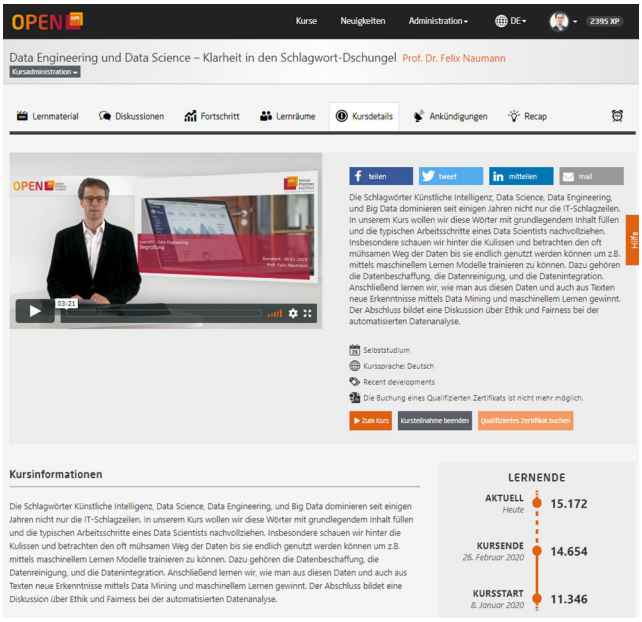

Abb. 1 Startseite des openHPI MOOCs

den Schlagwort-Dschungel" (siehe Abb. 1) gibt uns zumindest recht, dass der oben genannte Wunsch nach einem Überblick über die im Titel genannten Themenblöcke existiert. Zu Kursbeginn am 8. Januar 2020 waren mehr als 11.000 Teilnehmer in dem 6-wöchigen Kurs, der 141 Videos mit einer Gesamtlänge von 13:20 Stunden und 217 Fragen in Selbsttests und 71 Fragen als wöchentliche Hausaufgaben und 30 Fragen in einer Klausur umfasste, eingeschrieben. Nach dem Ende des Kurses ist die Teilnehmerzahl sogar auf über 17.000 Teilnehmer gewachsen, von denen 74 % (12.576) mindestens einmal auf die Kursinhalte zugegriffen haben. Er gehört damit zu den erfolgreichsten Kursen der Plattform.

2 Themen des Kurses

In der Planung des Kurses war es für uns von vornherein wichtig den Kurs in deutscher Sprache anzubieten, da die englische Sprache durchaus eine Einstiegshürde ist, die viele potentiell interessierte Teilnehmer abschrecken kann. Auch war uns kein vergleichbarer deutscher Kurs bekannt.

Da wir für den Kurs explizit planten, nicht tief in technische Details einzusteigen, sondern nur einen Überblick über Grundlagen zu vermitteln, war es uns möglich ein sehr breites Gebiet über die sechs Wochen hinweg abzudecken. Die Titel und Inhalte der Wochen waren:

1. **Big Data und Data Science:** Datenspeicherung, Datenmodelle, Anfragen, Big Data, Datenvielfalt
2. **Data Science Anwendungen und Text Mining:** Privatsphäre und Ethik, Datenkompetenz, Word Embeddings, NLP
3. **Skalierbares Datenmanagement:** OLTP & OLAP, ACID/BASE, Parallelisierung, Verteilung, Map/Reduce, Partitionierung, Cloud
4. **Datenaufbereitung:** Datenqualität, Datenreinigung, Duplikaterkennung, Datenfusion
5. **Informationsintegration:** Heterogenität, Schema Mapping & Matching, Data Warehouses, ETL, Data Lakes
6. **Statistik, Data Mining, Machine Learning:** Risikokompetenz, Visualisierung, Assoziationsregeln, Klassifizierung, neuronale Netze, erklärbare KI

Die Themenkomplexe zeigen den bewussten Fokus auf Data Engineering anstatt speziell auf mathematisch orientierteres Data Science. Auch wenn Data Engineering weniger stark in der öffentlichen Wahrnehmung steht, sind dessen Aspekte unserer Ansicht nach ebenso wichtig wie die Kernthemen von Data Science [3, 6]. Data Scientists verbringen mitunter einen Großteil ihrer Zeit mit Themen des Data Engineerings, z. B. Datenverwaltung, Datenreinigung und Aufbereitung [5].

Unser Ziel war es, den Teilnehmern ein allgemeines technisches Verständnis zu den wichtigsten Themen von Data Science und Data Engineering zu vermitteln. Insbesondere wollten wir die Fähigkeit zur selbstständigen, kritischen Bewertung von Aussagen, z. B. in der Presse, zu diesen Themen fördern.

Der zeitliche Aufwand des Kurses war für Teilnehmer mit vier Stunden pro Woche angesetzt: durchschnittlich zwei Stunden Videomaterial, eine Stunde für Selbsttests und maximal eine Stunde pro Hausaufgabe. Teilnehmer haben mitunter festgestellt, dass man dem Sprecher bei einer 1,3-fachen Abspielgeschwindigkeit gut folgen konnte.

Der Zeitaufwand für uns als Anbieter des Kurses war natürlich ungleich größer: Themenwahl, Folienerstellung und Videoaufnahmen entsprachen in etwa der Erstellung einer neuen Vorlesung des gleichen Umfangs. Der Umstand, dass wir nur an der Oberfläche der vielen Themen kratzten erleichterte die Vorbereitung des Materials übrigens nicht, da die Vielfalt entsprechend größer war und wir darauf achten mussten, Begriffe und Konzepte zu vermeiden, die Informatikstudenten allgemein bekannt sind. Nach einer Eingewöhnungsphase mussten Videoaufnahmen nur selten neu gestartet werden – der Zeitaufwand entsprach in etwa dem 1,5-fachen des fertigen Videomaterials. Die Aufzeichnung, Überarbeitung und Bereitstellung der Videos wurde von einem professionellen Team am HPI übernommen.

Die Konzeption der Selbsttest-, Hausaufgaben- und Klausurfragen hingegen war ungleich aufwändiger als bei vergleichbaren Übungseinheiten traditioneller Vorlesungen. Richtige Antworten sollten zusätzliches Wissen transportie-

ren. Falsche Antworten sollten nicht zu offenkundig falsch sein. Negation oder gar doppelte Negation über Frage und Antwort hinweg sollte vermieden werden. Jedes Detail muss stimmen und jede denkbare Interpretation der Multiple-Choice-Aufgaben und derer möglichen Antworten muss bedacht werden: Schließlich werden tausende Teilnehmer über der jeweiligen Frage brüten, eine Erklärung ihrer Lösungsidee finden und diese dann gegebenenfalls verteidigen, sofern sie aus unserer Sicht falsch war. In einigen Fällen mussten wir eine Aufgabe neu formulieren oder gar ganz zurückziehen.

Insgesamt hat die Erstellung der Fragen pro Woche in etwa drei Tage in Anspruch genommen. Dabei wurden zunächst etwa 50–80 mögliche Fragen vorgeschlagen, die von den anderen Teammitgliedern getestet und anschließend diskutiert und überarbeitet wurden. Hinzu kamen 1–2 Stunden pro Tag und Teammitglied für die Forumsbetreuung. Auch wenn die Mehrheit der Teilnehmerfragen von anderen Teilnehmern beantwortet wurden, war es nötig, diese Antworten zu prüfen oder selbst Antworten zu formulieren – von kurzen Richtigstellungen bis hin zu tiefergehenden Erklärungen, Verweisen auf weiterführende Literatur oder der Moderation allzu hitziger Diskussionen.

3 Teilnehmer und Forum

Ein wichtiger Bestandteil beim Lernen ist der Austausch zwischen den Lernenden. Sämtliche Diskussionen in den letztlich 732 Themen im Forum wurden von uns moderiert, um die Korrektheit der Antworten auf fachliche Fragen sicherzustellen. Das Forum wurde von fast 7.000 Teilnehmern genutzt, von denen sich 615 aktiv mit insgesamt nahezu 4.200 Posts eingebracht haben. Dabei schrieben die 20 aktivsten Teilnehmer (Teammitglieder ausgeschlossen) etwa 37 % aller Beiträge. Im Vordergrund der Themen stand oft die Diskussion der Selbsttests zwischen den Videos und der Ergebnisse der wöchentlichen Hausaufgaben. Viele davon ergaben sich aus Missverständnissen bei Begrifflichkeiten, etwa wenn diese in anderen Domänen andere Bedeutungen haben oder umgangssprachlich anders verwendet werden. Bei der Konzeption der Fragen für Selbsttests und Hausaufgaben haben wir uns bemüht, potenziell missverständliche oder kontroverse Antwortmöglichkeiten zu vermeiden, um so den Diskurs im Forum fachlich und konstruktiv zu halten. Im allgemeinen Sprachgebrauch, beispielsweise, wird „Durchschnitt“ oft nur als das arithmetische Mittel verstanden. Um zu verdeutlichen, dass beispielsweise das harmonische Mittel und der Median weitere Maßzahlen für den Mittelwert darstellen, haben wir eine entsprechende Frage gestellt, die im Nachgang zu viel Diskussion geführt hat.

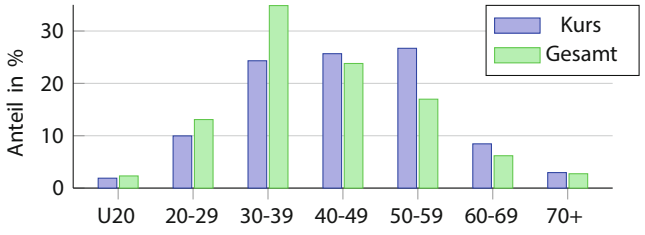

Abb. 2 Altersverteilung in unserem Kurs und Durchschnitt aller openHPI-Kurse

Die Spannweite des Vorwissens der Teilnehmer war sehr breit. Neben Fragen zu Grundlagen von Einsteigern gab es besonders engagierte Nutzer, die ihre fundierte fachliche Kenntnis in Diskussionen und vor allem auch Hilfestellungen mit anderen teilten. Beispielsweise gab es Fragen zur korrekten Zählung von falsch-positiven Vorhersagen zur Berechnung von Genauigkeit oder zum Unterschied zwischen Prozent und Prozentpunkten, während andere Teilnehmer über Anwendungsszenarien von erklärbarer KI diskutierten. Wie Abb. 2 zeigt, haben sich nicht nur Jugendliche und Studierende für den Kurs interessiert – weit über ein Drittel der Teilnehmer war (gemäß freiwilliger Angabe) über 50 Jahre alt. Damit hat dieser Kurs nicht nur Teilnehmerrekorde auf der openHPI Plattform gebrochen, sondern auch ein demographisch breiter verteiltes Publikum als im Durchschnitt anderer Kurse angesprochen.

4 Folien und Prüfungsfragen

Da das Lernziel für den Kurs ein Grundverständnis wesentlicher Begriffe war, und nicht etwa vertieftes Fachwissen, haben wir die Folien der Videos visuell zurückhaltend konzipiert – nur die wichtigsten Stichpunkte und unterstützende Grafiken fanden sich in den Folien und wurden ausführlich besprochen. Informationen wurden vor allem über das Gesprochene vermittelt. Im Diskussionsforum stieß dieser Stil auf gemischte Reaktionen. Bei einigen Nutzern führte er zu mehr Kollaboration: Ein Teil jener, die sich mehr Textinhalte wünschten, gründete einen offenen Lernraum und teilte dort Notizen und Skripte. Manche Teilnehmer gaben an, dass sie durch das Erstellen der Notizen mehr lernten als etwa bei anderen Kursen, die ausführlichere Materialien zur Verfügung stellten. Andere Nutzer nutzten ausschließlich die Tonspur und konnten die Vorlesung gleichsam als Podcast konsumieren.

Wir bemühten uns, für alle Themen weiterführende wissenschaftliche Grundlagenliteratur und Verlinkungen sowohl in deutscher als auch englischer Sprache bereitzustellen. Dies war für deutsche Materialien aufgrund der Dominanz der englischen Sprache in der Fachwelt, stellenweise herausfordernd. Auch verwendeten wir englisch-

sprachige Begriffe für die erläuterten Konzepte, insofern uns keine geläufige deutsche Übersetzung bekannt war. Einige Teilnehmer kritisierten diesen Einsatz englischer Bezeichnungen innerhalb eines deutschsprachigen Kurses.

Im Kurs wurden drei verschiedene Arten der Wissens- und Leistungskontrolle durchgeführt. Themenbezogene Selbsttests (unbepunktet und beliebig oft wiederholbar), wöchentliche Hausaufgaben (bepunktet) und eine abschließende Klausur (bepunktet). Alle Kontrollen wurden vollautomatisch durch die openHPI Plattform korrigiert und bewertet. Im Anschluss an fast alle Videos wurde ein kurzer Selbsttest angeboten. Dieser bestand, wie auch Hausaufgaben und Klausur, aus Multiple Choice Fragen (eine richtige Antwort) oder Multiple Answer Fragen (beliebig viele richtige Antworten) und diente ausschließlich der Selbstkontrolle. Gleichzeitig orientierten sich auch die Selbsttests in Themen und Aufgabenformat an den Hausaufgaben und der Klausur und bereiteten Nutzer so darauf vor. Da Selbsttestaufgaben nicht für das abschließende Zeugnis oder Zertifikat herangezogen wurden, boten wir den Nutzern dort auch Rechercheaufgaben zur eigenständigen Vertiefung in die Themen an, die auf eine überwiegend positive Resonanz bei den Nutzern stießen. So konnten wir Nutzer mit der Frage „Welche der folgenden alternativen Tests zum Turing-Test existieren tatsächlich?“ auf Lovelace- und Metzger-Test sowie die CBMM Turing++ Questions aufmerksam machen und anschließend eine rege Diskussion über deren Inhalte im Forum verfolgen. Durch die Frage „In welchen der folgenden Beispiele lässt sich das Benfordsche Gesetz beobachten?“ erarbeiteten sich Nutzer selbstständig, weshalb niedrigere Zahlen bei Hausnummern häufiger vorkommen, das gleiche Phänomen aber bei Lottozahlen nicht vorzufinden ist (inkl. eigener Implementierungen).

Die sechs wöchentlichen Hausaufgaben setzten sich aus der gesamten Themenwoche zusammen und mussten innerhalb von 60 Minuten beantwortet werden. Für die Abschlussklausur hatten Nutzer 120 Minuten Zeit und mussten 30 Fragen beantworten. Das Schwierigkeitsniveau wurde über alle Kontrollformate gleich gehalten.

5 Resümee

Nach der zweiten Kurswoche haben wir den Kursteilnehmern die Möglichkeit gegeben, die Risiken und Chancen von Big Data Anwendungen aus ihrer Perspektive zu benennen. Zum Ausklang des Kurses präsentierten wir eine Auswertung. Aus den über 2.200 Antworten haben wir eine nach Themen eingefärbte interaktive Übersicht erstellt (siehe Abb. 3 und https://hpi.de/naumann/sites/openhpi2020/). Diese Momentaufnahme zeigt beispielsweise, dass Chancen vor allem für medizinische Anwendungen gesehen werden, aber die Bedrohung der Privatsphäre als ein besonderes Ri-

Abb. 3 Visualisierung der Meinungen der Teilnehmer zu Chancen und Risiken von Data Science (hochauflösend)

siko aufgefasst wird. Ein Studium der einzelnen Beiträge auf der Webseite (oder durch Hineinzoomen in die hochauflösende Abbildung) fördert immer wieder Überraschendes zutage und zeigt deutlich, dass sich in jedem Risikobereich auch Chancen verbergen.

Eine Abschlussbefragung zum Ende des Kurses haben 1.286 Nutzer genutzt, um den Kurs überwiegend positiv zu bewerten. Viele wünschen sich für einzelne Themengebiete weitere Vertiefungskurse. Die wichtigsten Kritikpunkte waren die knapp gehaltenen Folien und der Mangel eines Skripts zum Üben und Lernen für die Hausaufgaben und die Klausur. Wir sind dem begegnet, indem wir vermehrt Verweise auf weitere Literatur zur Verfügung stellten. Eine viel bessere Maßnahme jedoch ergriffen einige Kursteilnehmer, die ihre teils sehr ausführlichen Mitschriften und Notizen allen anderen in virtuellen Lernräumen und im Forum zur Verfügung stellten.

Intensive Diskussionen ergaben sich ob der zugegebenermaßen häufigen „ähm“’s des Sprechers in den Videos. Hier tritt der Unterschied zwischen Vorlesungen, die er schon häufig lieferte, und Stoff, den er erst selten oder noch gar nicht einem breiten Publikum erklärte, zutage. Unangemes-

sene, teils persönliche Kritik ergab sich selten und wurde oft und zügig von anderen Teilnehmern beanstandet. Das Lob gegen Ende des Kurses hingegen war überwältigend und zerstreute jeden Verdruss.

Die openHPI Plattform vergibt drei Arten von Abschlussbestätigungen. So konnten wir nach Auswertung der Klausur folgende vergeben:

- 4.570 *Teilnahmebescheinigungen* an Teilnehmer, die mindestens 50 % der Inhalte angesehen haben. Dies sind 43 % derjenigen, die überhaupt einmal einen Inhalt anschauten.
- 3.546 *Zeugnisse* an Teilnehmer, die mindestens 50 % der Punkte aus Hausaufgaben und Klausur erzielten. Dies sind 38 % der Teilnehmer, die bis zur Halbzeit des Kurses in den Kurs hineingeschaut haben.
- 150 *qualifizierte Zertifikate* mit Foto an Teilnehmer, die gegen eine Gebühr unter Online-Aufsicht durch einen Drittanbieter die Hausaufgaben und Klausur erfolgreich meisterten.

Der Kurs steht weiterhin allen Interessierten unter https://open.hpi.de/courses/data-engineering2020 offen. Die Inhalte des Forums stehen nur noch im Lesemodus zur Verfügung — die Selbsttests können weiter bearbeitet werden. Tatsächlich haben sich seit Ende des Kurses bereits mehr als zweitausend weitere Nutzer eingeschrieben.

Funding Open Access funding provided by Projekt DEAL.

Open Access Dieser Artikel wird unter der Creative Commons Namensnennung 4.0 International Lizenz veröffentlicht, welche die Nutzung, Vervielfältigung, Bearbeitung, Verbreitung und Wiedergabe in jeglichem Medium und Format erlaubt, sofern Sie den/die ursprünglichen Autor(en) und die Quelle ordnungsgemäß nennen, einen Link zur Creative Commons Lizenz beifügen und angeben, ob Änderungen vorgenommen wurden.

Die in diesem Artikel enthaltenen Bilder und sonstiges Drittmaterial unterliegen ebenfalls der genannten Creative Commons Lizenz, sofern sich aus der Abbildungslegende nichts anderes ergibt. Sofern das betreffende Material nicht unter der genannten Creative Commons Lizenz steht und die betreffende Handlung nicht nach gesetzlichen Vorschriften erlaubt ist, ist für die oben aufgeführten Weiterverwendungen des Materials die Einwilligung des jeweiligen Rechteinhabers einzuholen.

Weitere Details zur Lizenz entnehmen Sie bitte der Lizenzinformation auf http://creativecommons.org/licenses/by/4.0/deed.de.

Literatur

1. Abedjan Z, Anuth H, Esmailoghli M, Mahdavi M, Neutatz F, Chen B (2020) Data Science für alle: Grundlagen der Datenprogrammierung. Informatik Spektrum. https://doi.org/10.1007/s00287-020-01253-8
2. Heinemann B, Opel S, Budde L, Schulte C, Frischemeier D, Biehler R, Podworny S, Wassong T (2018) Drafting a data science curriculum for secondary schools. In: Proceedings of the 18th Koli Calling International Conference on Computing Education Research, S 1–5
3. Microsoft Azure Team data science process. https://docs.microsoft.com/en-us/azure/machine-learning/team-data-science-process/overview. Zugegriffen: 31. März 2020
4. Naumann F, Jenders M, Papenbrock T (2014) Ein Datenbankkurs mit 6000 Teilnehmern. Informatik Spektrum 37(4):333–340
5. NVIDIA blog Accelerated data science. https://blogs.nvidia.com/blog/2018/11/15/accelerated-data-science-hpc/. Zugegriffen: 31. März 2020
6. Udacity blog Data science. https://blog.udacity.com/2014/11/data-science-job-skills.html. Zugegriffen: 31. März 2020