An das
Rektorat der Universität Konstanz
über den
Ausschuss für Lehre und Weiterbildung

Antrag auf einen „Freiraum für die Lehre“ aus Mitteln des Projekts „b³ - beraten, begleiten, beteiligen“

Datum: 23.10.2017

Antragsteller/-in

Name	Prof. Dr. Nils B. Weidmann
Tel. Nr.	5676
Email-Adresse:	nils.weidmann@uni-konstanz.de
Fachbereich:	Politik und Verwaltungswissenschaft
Thema des Freiraumprojekts:	Kurs: Data Management for Social Scientists
Zeitraum (max. 2 Semester):	WS 2018/19
Reduktion des Lehrdeputats:	9 SWS

Kurze Zusammenfassung der geplanten Maßnahme (max. 1.000 Zeichen):

Die zunehmende Digitalisierung unserer Gesellschaft führt zu einer Verfügbarkeit von neuen und umfangreicheren sozialwissenschaftlichen Daten über menschliches Verhalten. Dies stellt unsere Disziplin allerdings vor neue Herausforderungen in der Verwaltung und Verarbeitung dieser neuen Datensätze. Bisher unterstützt das Lehrprogramm am Fachbereich Politik und Verwaltungswissenschaft diese neuen Anforderungen nur unzureichend. Ziel dieses Projektes ist es, durch die Entwicklung eines neuen interaktiven Kurses diese Lücke zu füllen. Der Kurs „Data Management for Social Scientists“ soll Studierende mit den Grundkonzepten und den wichtigsten Software-Tools zur Datenverwaltung und –verarbeitung vertraut machen. Er verwendet dabei Lernbeispiele mit sozialwissenschaftlichen Anwendungen, die mit einem interaktiven Mechanismus automatisch auf Korrektheit getestet werden können. Dieses Verfahren wird bereits auf Lernplattformen wie z.B. http://exercism.io/ eingesetzt und soll für diesen Kurs adaptiert werden.

Projektskizze

Ausgangssituation

Die zunehmende Digitalisierung unserer Gesellschaft führt zu einer Verfügbarkeit von neuen und umfangreicheren Daten über menschliches Verhalten. Beispielsweise erlauben Daten

über Kommunikationsverbindungen im Internet Rückschlüsse auf die Internetnutzung in bestimmten Regionen (Weidmann et al. 2016). Daten aus sozialen Medien können genutzt werden, um die Stimmung der jeweiligen Nutzer zu messen (Golder und Macy 2011). Diese Medien können auch verwendet werden, um über einfache Experimente politische und soziale Präferenzen von Nutzern zu erfassen (Nisser und Weidmann 2016).

Trotz der Chancen, die sich aus der Verfügbarkeit dieser Daten für die Sozialwissenschaften ergeben, stellt die digitale Revolution unsere Forschung auch vor große praktische Herausforderungen. Diese Herausforderungen finden sich primär in der Verwaltung und Verarbeitung dieser neuen Datensätze in der sozialwissenschaftlichen Forschung. Im Wesentlichen sind es drei Merkmale, welche diese neuen Datensätze auszeichnen: (1) *Die Datenmenge.* Im Zeitalter von „Big Data" haben sozialwissenschaftliche Datensätze nicht mehr dutzende oder hunderte von Einträgen, sondern Millionen oder Milliarden von Zeilen. (2) *Neue Datentypen.* Viele Datensätze beinhalten neben herkömmlichen Formaten wie Text oder Zahlen auch neue Datentypen wie z.B. geographische Koordinaten, die speziell enkodiert und verarbeitet werden müssen. (3) *Kompliziertere Datenstrukturen.* Viele über digitale Kanale gesammelte Daten werden nicht in einer einfachen tabellarischen Form geliefert, wie dies meistens in den Sozialwissenschaften der Fall ist. Häufig sind Datensätze „genestet", haben also Eintrage auf verschiedenen Aggregationsstufen wie z.B. Staaten, Wahlkreise in diesen Staaten, und Städte in den jeweiligen Wahlkreisen. Andere Datensätze wiederum haben keine vorgegebene Struktur, so wie dies häufig bei Texten oder Webseiten der Fall ist.

Daten, die durch eine oder mehrere dieser Charakteristiken gekennzeichnet sind, lassen sich nicht mehr mit einfachen Methoden verarbeiten. Die in den Sozialwissenschaften leider noch häufig anzutreffende Office-Software „Excel" kann nur max. 1 Million Einträge in einer Tabelle verarbeiten. Schwerwiegender ist jedoch das Problem, dass diese Art von Software zu einer manuellen Datenverarbeitung einlädt, welche die erledigten Datenmanipulationen nicht replizierbar (und deshalb bei Bedarf korrigierbar) macht. Wenn Sozialwissenschaftler die Chancen von „Big Data" im digitalen Zeitalter nutzen wollen, müssen sie lernen, die in der Informatik schon lange existierenden Konzepte von Daten und Datenbanken zu nutzen, und komplexe Daten systematisch mit Hilfe von Software – und nicht von Hand – zu verwalten.

Bisher unterstützt das Lehrprogramm am Fachbereich Politik und Verwaltungswissenschaft diese neuen Anforderungen nur unzureichend. Die Methodenausbildung im BA-Bereich besteht aus den Einführungsvorlesungen „Empirische Methoden" und „Statistik", welche theoretische Grundlagen von Forschungsdesign und statistischen Methoden bereitstellen. Auch im MA-Programm gibt es entsprechende Veranstaltungen („Research Design") mit einem ähnlichen Schwerpunkt. In Vertiefungskursen können sich Studierende praktische Kenntnisse in den wichtigsten Softwaretools für statistische Analyse (*R* oder *Stata*) aneignen. Diese Kurse setzten jedoch auf der Ebene der Datenanalyse an und behandeln dabei nicht, wie potentiell große und diverse Datenmengen erst einmal in ein für die statistische Analyse passendes Format gebracht werden. Aus eigener Betreuungserfahrung weiß der Antragsteller, dass es häufig Studierende gibt, die in statistischen Methoden hervorragend ausgebildet sind, bei der Datenverwaltung allerdings auf manuelle Tools wie MS Excel zurückgreifen. Diese Lücke soll der neue Kurs füllen.

Ziele der Maßnahme

Ziel des Projektes ist die Entwicklung eines interaktiven Vertiefungskurses, der den Studierenden die Grundkonzepte der Datenhaltung und -verarbeitung erklärt. Dies beinhaltet sowohl theoretische Konzepte, als auch deren technische Umsetzung mit entsprechenden Software-Tools. Er basiert dabei auf etablierten Konzepten aus der Informatik wie relationale Datenbanken, die jedoch in den Sozialwissenschaften nicht breit verwendet werden (primär die Sprache SQL). Der Kurs (2 SWS) soll auf Englisch entwickelt und unterrichtet werden und sowohl für BA- als auch für MA-Studierende zugänglich sein. Da Methodenkurse am Fachbereich programmübergreifend angeboten werden, wird sich der Kurs problemlos in das existierende Lehrangebot im BA und MA-Programm integrieren lassen.

Der Kurs soll neben einem umfangreichen Kursskript (welches die Grundlage für ein späteres Buch ist) eine Sammlung von begleitenden Übungsaufgaben bereitstellen. Studierende können diese Aufgaben in der Sprache SQL lösen. Die Lernumgebung gibt dann Rückmeldung dazu, ob die entsprechende Lösung korrekt ist und wenn nicht, wo der Fehler liegt. Dieses Konzept des interaktiven Testens von Programmcode wird bereits erfolgreich auf Lernplattformen wie http://exercism.io/ verwendet und wird für den Kurs übernommen. Es handelt sich um eine Software, die auf den Laptops der Kursteilnehmerinnen und –teilnehmer installiert wird und deshalb auch offline verfügbar ist. Diese Software ist frei erhältlich und mit keinen Lizenzkosten verbunden. Im Rahmen des Projekts müssen nur die Aufgaben und die Lösungsdefinitionen erstellt werden. Der Kurs soll natürlich nicht ausschließlich in diesem Modus unterrichtet werden – gerade, wenn eine Aufgabe nicht selbst gelöst werden kann, muss der Dozent oder ein/e Tutor/in Hilfestellung geben. Dennoch ermöglicht die interaktive Lernumgebung, dass die Kursteilnehmerinnen und –teilnehmer mit deutlich weniger Hilfestellung auskommen, als dies in einem regulären Übungsbetrieb der Fall wäre.

Die im Kurs abzudeckende Thematik stellt sich wie folgt dar: Teil 1 des Kurses beschäftigt sich mit Grundkonzepten des relationalen Datenmodells (Tabellen, Felder, Zeilen, Schlüssel, Joins, Aggregationen). In Teil 2 werden diese Konzepte dann mit Hilfe der Sprache SQL praktisch implementiert. Dabei stehen die *Definition* der Datenstruktur, die *Extraktion* von Informationen, und die *Modifikation* von Daten im Vordergrund. In Teil 3 geht der Kurs auf unstrukturierte Daten (wie z.B. Texte oder Websites) ein. Für diese Daten gibt es neue Datenbankkonzepte, die gemeinhin unter dem Begriff „NoSQL“ zusammengefasst werden. Aus diesem breiten Feld greift sich der Kurs solche Anwendungen (wie MongoDB) heraus, die besonders für sozialwissenschaftliche Projekte geeignet erscheinen.

Der Kurs soll ohne Vorkenntnisse in R oder einer anderen Sprache zu absolvieren sein. Er benutzt zwar die Sprache R zur Kommunikation mit Datenbanksystemen, vertieft diese Sprache aber über einige grundlegende Kommandos hinaus nicht. Von zentraler Bedeutung ist die Verwendung sozialwissenschaftlicher Anwendungen und *use cases* in den Übungsaufgaben. Es gibt zahlreiche Einführungskurse in das Datenmanagement, die aber fast ausschließlich auf Anwendungen aus der Informatik zielen.

Kenntnisse im Bereich Datenmanagement sind für den Großteil der Absolventen des FB Politik und Verwaltungswissenschaft nützlich: Wer weiter wissenschaftlich arbeiten will, kann sich dadurch neue Datenquellen und Analysemethoden erschließen und den Forschungsprozess teilweise automatisieren und damit transparenter und replizierbarer gestalten. Andererseits sind diese Fähigkeiten von großem Nutzen für Studierende, die im Bereich „Data Science“ in der freien Wirtschaft oder der öffentlichen Verwaltung arbeiten

wollen. Hier sind die im Kurs unterrichteten Konzepte und Tools von zentraler Bedeutung und können das Spektrum der Berufe, die unseren Absolventinnen und Absolventen zugänglich gemacht werden, deutlich erweitern.

Eckpunkte und Meilensteinplanung

Das Projekt soll in drei Phasen durchgeführt werden, an deren Ende die jeweiligen Meilensteine stehen.

1. *Vorbereitung der technischen Infrastruktur*
   Vor Beginn der eigentlichen Kursentwicklung wird ein Online-Repository für die interaktive Lernumgebung eingerichtet. Diese Website stellt eine Anleitung bereit, welche Software für den Kurs benötigt wird und wie diese zu installieren ist.
2. *Erstellung der Übungsaufgaben und der Beispiellösungen*
   Der Hauptteil der Arbeit besteht in der Erstellung des Kursskripts, der Übungsaufgaben und der Beispiellösungen. Die Arbeit wird dabei zwischen dem Antragsteller und einem Forschungsassistenten (stud. Hilfskraft) aufgeteilt, so dass ersterer für das Skript und die Aufgaben, und letzterer für die Erarbeitung der Lösungsdefinitionen verantwortlich ist. Durch die Einbindung von studentischen Hilfskräften als Tester wird sichergestellt, dass diese Dokumentation auf die Bedürfnisse der Studierenden eingeht und von ihnen benutzt werden kann.
3. *Fertigstellung des Kursskripts und Veröffentlichung des Repositorys*
   Am Ende des Projektzeitraums steht die Veröffentlichung des Kursskripts und des Online-Repositorys und die Verwendung in einem entsprechenden Kurs am Fachbereich. Das Feedback in diesem Kurs wird zur Überarbeitung des Lehrmaterials verwendet, bevor es zu einem Lehrbuch weiterentwickelt wird.

Angaben zur Nachhaltigkeit

Das Kursmaterial wird in einer Form bereitgestellt, die es auch anderen Lehrenden möglich macht, den entsprechenden Kurs zu unterrichten. Dies geschieht auch im Hinblick auf die Weiterentwicklung zu einem Lehrbuch, wo diese Voraussetzung ja zwingend gegeben sein muss. Der Antragsteller hat bereits mit dem SAGE Verlag bezüglich des Buchprojekts Kontakt aufgenommen und ist auf großes Interesse gestoßen. Wie bei anderen Kursen zu Forschungsmethodik stellt sich auch hier die Frage der zukünftigen Anpassung des Kurses an neue Software-Versionen. Da allerdings die verwendeten Software-Konzepte und Tools (SQL) lange in der Informatik etabliert sind, sind diese Änderungen überschaubar und können sehr einfach vom Antragsteller ausgeführt werden, so dass der Kurs auch in der Zukunft mit aktueller Software gelehrt werden kann.

Literatur

- Golder, Scott A und Michael W Macy. 2011. “Diurnal and Seasonal Mood Vary with Work, Sleep, and Daylength across Diverse Cultures.” *Science* 333(6051):1878–1881.
- Nisser, Annerose und Nils B Weidmann. 2016. “Measuring Ethnic Preferences in Bosnia and Herzegovina with Mobile Advertising.” *PloS one* 11(12):e0167779.
- Weidmann, Nils B., Suso Benitez-Baleato, Philipp Hunziker, Eduard Glatz und Xenofontas Dimitropoulos. 2016. “Digital Discrimination: Political Bias in Internet Service Provision across Ethnic Groups.” *Science* 353(6304):1151–1156.

Kostenkalkulation

Siehe beiliegende Excel-Tabelle.

Von der zuständigen Sektion auszufüllen:		
Der Antrag ist mit dem Fachbereich und der zuständigen Studienkommission abgestimmt?	ja	nein
Der beantragte Freiraum kann aus Sicht der Sektion gewährt werden?	ja	nein