Datenbank Spektrum (2021) 21:175–177
https://doi.org/10.1007/s13222-021-00393-9

EDITORIAL

Editorial

Ralf Schenkel[1] · Stefanie Scherzinger[2] · Marina Tropmann-Frick[3] · Theo Härder[4]

Angenommen: 4. Oktober 2021 / Online publiziert: 20. Oktober 2021
© Der/die Autor(en) 2021

1 Schwerpunktthema „Data Engineering for Data Science“

Data Engineering ist ein entscheidender Teil eines jeden Data-Science-Projekts. Datenerfassung und Metadatenmanagement sind Voraussetzungen für jede sinnvolle Analyse und erfordern in der Praxis den größten Teil des Zeitaufwands in diesen Projekten. In der öffentlichen Diskussion kommen diese Aspekte dagegen leider zu kurz, man versteht oft unter Data Science nur die eigentlichen Analyse- und Lernverfahren, die ohne das zugehörige Data Engineering aber in der Regel keine sinnvollen Ergebnisse liefern können. Dieses Themenheft des Datenbankspektrums, das eine Initiative des neu gegründeten DBIS-Arbeitskreises „Data Engineering für Data Science“ ist, legt daher den Schwerpunkt auf Aspekte des Data Engineering, enthält aber auch Beiträge über Datenanalyse.

Angenommene Beiträge zum Themenheft konnten erstmals im Rahmen des Datenbank-Workshops bei der Konferenz „Lernen, Wissen, Daten, Adaptivität“ (LWDA) Anfang September 2021 vorgestellt werden. Was als Experiment gestartet war, hat aus unserer Sicht hervorragend funktioniert. Mit in der Spitze mehr als 30 Teilnehmenden war der Workshop sehr gut besucht, die zehn Vorträge deckten ein breites Spektrum des Data Engineering ab. Dieses Modell könnte bei weiteren Auflagen der Konferenz fortgesetzt werden. Möglicherweise auch wegen dieses Angebots war die Resonanz auf den Call for Papers so groß, dass nicht alle angenommenen Beiträge in dieses Heft aufgenommen werden können. Sie werden in den folgenden Ausgaben erscheinen. Dies gilt leider auch für einen Beitrag, der die von der Gesellschaft für Informatik vorgeschlagenen Empfehlungen zur Gestaltung von Data-Science-Studiengängen vorstellt.

In diesem Heft sind sechs wissenschaftliche Langbeiträge und zwei Kurzbeiträge zum Schwerpunktthema enthalten, außerdem die Vorstellung des Sonderforschungsbereichs FONDA als Community-Beitrag, der sehr gut zum Thema des Hefts passt.

Den Anfang machen Alexander Schoenenwald (OTH Regensburg), Simon Kern, Josef Viehhauser (beide BMW Group) und Johannes Schildgen (OTH Regensburg), die in ihrem Beitrag „Collecting and Visualizing Data Lineage of Spark Jobs“ ein System vorstellen, das Lineage von Datensätzen visualisieren kann, die mit Hilfe von Spark-Pipelines generiert wurden, und das somit die Auswahl von Datensätzen für Data-Science-Projekte vereinfacht.

Im folgenden Beitrag „Performance Evaluation of Policy-Based SQL Query Classification for Data-Privacy Compliance“ beschäftigen sich Peter K. Schwab, Jonas Röckl, Maximilian S. Langohr und Klaus Meyer-Wegener (Friedrich-Alexander-Universität Erlangen-Nürnberg) mit der Generierung von Metadaten über SQL-Anfragen, um diese Anfragen automatisiert hinsichtlich der Einhaltung anwendungsspezifischer Privacy-Regeln zu klassifizieren. Der Fokus liegt dabei auf einer möglichst effizienten Implementierung dieser Klassifikation.

Ioannis Prapas (Technische Universität Berlin), Behrouz Derakhshan, (DFKI), Alireza Rezaei Mahdiraji (DKFI) und Volker Markl (Technische Universität Berlin) zeigen in ihrem Beitrag „Continuous Training and Deployment of Deep Learning Models“, wie man komplexe Deep-Learning-Mo-

✉ Ralf Schenkel
schenkel@uni-trier.de

Stefanie Scherzinger
Stefanie.Scherzinger@uni-passau.de

Marina Tropmann-Frick
marina.tropmann-frick@haw-hamburg.de

Theo Härder
haerder@cs.uni-kl.de

[1] Professur für Datenbanken und Informationssysteme, Universität Trier, 54296 Trier, Deutschland

[2] Lehrstuhl für Informatik mit Schwerpunkt skalierbare Datenbanksysteme, Uni Passau, 94032 Passau, Deutschland

[3] Professur für Datenanalyse, HAW Hamburg, 20099 Hamburg, Deutschland

[4] AG Datenbanken und Informationssysteme, TU Kaiserslautern, 67663 Kaiserslautern, Deutschland

delle inkrementell auf neu eingetroffenen Daten trainieren kann, ohne dabei an Vorhersagequalität gegenüber einem regelmäßigen Neutrainieren mit der gesamten Datenmenge einzubüßen.

Die Analyse und mögliche Verbesserung automatisch generierter Clusterings anhand interner Evaluierungsmaßen stehen im Mittelpunkt des Beitrags von Daniyal Kazempour (Christian-Albrechts-Universität zu Kiel), Johannes Winter (Ludwig-Maximilians-Universität München), Peer Kröger (Christian-Albrechts-Universität zu Kiel) und Thomas Seidl (Ludwig-Maximilians-Universität München) mit dem Titel „On Methods and Measures for Inspection and Evaluation of Arbitrarily Oriented Subspace Clusters".

Im folgenden Beitrag „Season- and Trend-aware Symbolic Approximation for Accurate and Efficient Time Series Matching" stellen Lars Kegel, Claudio Hartmann, Maik Thiele und Wolfgang Lehner (TU Dresden) eine Lösung vor, um Zeitreihen besser approximieren zu können, indem zyklische Muster und Trends berücksichtigt werden. Dadurch wird auch die Effizienz der Suche in einem Datensatz mit Zeitreihen signifikant verbessert.

Chris-Marian Forke und Marina Tropmann-Frick (Hochschule für Angewandte Wissenschaften Hamburg) zeigen in ihrem Beitrag „Feature Engineering as a Part of Data Processing for Spatio-Temporal Data" die Wichtigkeit von Feature Engineering bei der Verarbeitung von Daten mit zeitlichen und räumlichen Komponenten.

In ihrem Beitrag „Measuring Data Changes in Data Engineering and their Impact on Explainability and Algorithm Fairness" aus der Reihe „Kurz erklärt" gehen Meike Klettke (Universität Rostock), Andrian Lutsch (Universität Rostock) und Uta Störl (Fernuniversität in Hagen) auf verschiedene Methoden ein, mit denen man die Auswirkungen von Data-Engineering-Schritten auf die Daten messen kann, und diskutieren mögliche Auswirkungen auf Erklärbarkeit und Fairness von Analysen in solchen Pipelines.

Data Science und insbesondere auch Data Engineering spielen zunehmend auch in der Hochschullehre eine Rolle, sei es innerhalb der regulären Informatikstudiengänge oder als Teil von Spezialangeboten. Um hier einen Überblick zu erhalten, haben die Gastherausgeber dieses Heftes im Mai 2021 eine Onlineumfrage unter Lehrenden durchgeführt, deren Ergebnisse im Beitrag „Data Engineering in der Hochschullehre" vorgestellt werden.

2 Community-Beiträge

Die Rubrik „Datenbankgruppen vorgestellt" enthält den Beitrag „The Collaborative Research Center FONDA" von Ulf Leser et al. (21 Koautoren von verschiedenen Berliner Organisationen). Dieser Artikel bietet einen Überblick über die Konzepte und Ziele der Forschungsarbeiten des von der Deutschen Forschungsgemeinschaft geförderten Forschungszentrums „FONDA – Foundations of Workflows for Large-Scale Scientific Data Analysis" (CRC 1404). In den interdisziplinären Projekten, an denen etwa 50 Forscher beteiligt sind, sollen neue Technologien, Algorithmen und Modelle erforscht und entwickelt werden, um Portabilität, Anpassbarkeit und Zuverlässigkeit von Workflows, die über verteilten Infrastrukturen ausgeführt werden, zu verbessern. Der Beitrag skizziert auch die Vision der Berliner Wissenschaftler zur Zukunft der Workflow-basierten wissenschaftlichen Datenanalyse.

Die Rubrik „Dissertationen" enthält in diesem Heft sieben Kurzfassungen von Dissertationen aus der deutschsprachigen DBIS-Community, die seit März 2021 erfolgreich abgeschlossen wurden.

Die Rubrik „Community" berichtet unter „News" über aktuelle Informationen, welche die DBIS-Gemeinde betreffen.

3 Künftige Schwerpunktthemen

3.1 Best Papers of BTW 2021

This special issue of the „Datenbank-Spektrum" is dedicated to the Best Papers of the BTW 2021 at the TU Dresden (running in a virtual format during the summer of 2021). The selected contributions should be extended to match the format of regular DASP papers.
Paper format: 8–10 pages, double-column

Selection of the Best Papers by the program chairs and the guest editor: Oct. 1st, 2021
Deadline for submissions: Nov. 1st, 2021
Issue delivery: DASP-1-2022 (March 2022)

Guest editor:
Theo Härder, University of Kaiserslautern
haerder@cs.uni-kl.de

3.2 Berlin Institute for the Foundations of Learning and Data (BIFOLD)

Das Berlin Institute for the Foundations of Learning and Data (BIFOLD) ist ein von BMBF und dem Land Berlin gefördertes Kompetenzzentrum, das aus der Fusion des Berlin Big Data Center (BBDC) und dem Berliner Zentrum für Maschinelles Lernen (BZML) hervorgegangen ist. BIFOLD hat sich zum Ziel gesetzt, hochinnovative Technologien zu entwickeln, die riesige Datenmengen organisieren und mit deren Hilfe fundierte Entscheidungen getroffen werden können, um wirtschaftlichen und gesellschaftlichen Mehrwert zu schaffen. Zu diesem Zweck werden die

bislang isoliert voneinander existierenden Gebiete Datenmanagement und Maschinelles Lernen verschmolzen. Die Technologien des Zentrums sollen den Stand der Technik in der Erforschung von Methoden des Datenmanagements, des maschinellen Lernens und deren Schnittstelle vorantreiben und die führende Stellung Deutschlands in Wissenschaft und Wirtschaft im Bereich der KI ausbauen. Als Technologietreiber stehen mehrere wirtschaftlich, wissenschaftlich und gesellschaftlich relevante Anwendungsbereiche im Fokus: Fernerkundung, digitalisierte Geisteswissenschaften, die Medizin sowie Informationsmarktplätze.

Aufbauend auf weltweit anerkannten Forschungsergebnissen sollen eine automatische Optimierung, Parallelisierung sowie eine skalierbare und adaptive Verarbeitung von Algorithmen in heterogenen, verteilten Umgebungen unter Einsatz moderner Rechnerarchitekturen ermöglicht werden. Daneben stehen Erklärbarkeit, verantwortungsvolles Datenmanagement und innovative Anwendungen der Datenanalyse im Fokus. Behandelt werden dabei Bereiche des Datenmanagements, des maschinellen Lernens, der linearen Algebra, der Statistik, der Wahrscheinlichkeitstheorie, der Computerlinguistik sowie der Signalverarbeitung. Durch Entwicklung und Bereitstellung von Open-Source-Systemen sowie von Algorithmen und Methoden zur Datenanalyse wird das Zentrum die Ausbildung, Forschung, Entwicklung, Innovation und kommerzielle Nutzung von Big Data Analytics und KI-Anwendungen in Deutschland fördern und so deutschen Firmen einen Wettbewerbsvorteil sichern.

Wir erbitten Einreichungen in Deutsch oder Englisch mit einem Umfang von 8 bis 10 Seiten (zweispaltig) gemäß den Layoutvorgaben (siehe https://www.springer.com/13222).

Frist zur Einreichung: 1. Feb. 2022

Erscheinen des Themenheftes: DASP-2-2022 (Juli 2022)

Gasteditor:
Dr. Alexander Borusan, TU Berlin
alexander.borusan@tu-berlin.de

3.3 N.N.

Wir erbitten Einreichungen in Deutsch oder Englisch mit einem Umfang von 8 bis 10 Seiten (zweispaltig) gemäß den Layoutvorgaben (siehe https://www.springer.com/13222).

Frist zur Einreichung: 1. Juni 2022
Erscheinen des Themenheftes: DASP-3-2022 (November 2022)

Gasteditor: N.N.

Funding Open Access funding enabled and organized 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.