Data Science – Analytics and Applications

Peter Haber • Thomas Lampoltshammer
Manfred Mayr
Editors

Data Science – Analytics and Applications

Proceedings of the 2nd International Data Science Conference – iDSC2019

Springer Vieweg

Editors
Peter Haber
Informationstechnik & System-Management
Fachhochschule Salzburg
Puch/Salzburg, Österreich

Manfred Mayr
Informationstechnik & System-Management
Fachhochschule Salzburg
Puch/Salzburg, Österreich

Thomas Lampoltshammer
Dept. für E-Governance in Wirtschaft
und Verwaltung
Donau-Universität Krems
Krems an der Donau, Österreich

ISBN 978-3-658-27494-8 ISBN 978-3-658-27495-5 (eBook)
https://doi.org/10.1007/978-3-658-27495-5

Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://dnb.d-nb.de abrufbar.

Springer Vieweg

Springer Vieweg ist ein Imprint der eingetragenen Gesellschaft Springer Fachmedien Wiesbaden GmbH und ist ein Teil von Springer Nature.
Die Anschrift der Gesellschaft ist: Abraham-Lincoln-Str. 46, 65189 Wiesbaden, Germany

Preface

It is with deep satisfaction that we write this foreword for the proceedings of the 2nd International Data Science Conference (iDSC) held in Salzburg, Austria, May 22nd - 24th 2019. The conference programme and the resulting proceedings represent the efforts of many people. We want to express our gratitude towards the members of our program committee as well as towards our external reviewers for their hard work during the reviewing process.

iDSC proofed itself – again – as an innovative conference, which gave its participants the opportunity to delve into state-of-the-art research and best practice in the fields of data science and data-driven business concepts. Our research track offered a series of presentations by researchers regarding their current work in the fields of data mining, machine learning, data management, and the entire spectrum of data science.

In our industry track, practitioners demonstrated showcases of data-driven business concepts and how they use data science to achieve organizational goals, with a focus on manufacturing, retail, and financial services. Within each of these areas, experts described their experience, demonstrated their practical solutions, and provided an outlook into the future of data science in the business domain.

Our sponsors – The MathWorks GmbH and Cognify KG – had their own, special platform via workshops to provide our participants hands-on interaction with tools or to learn approaches towards concrete solutions. In addition, an exhibition of products and services offered by our sponsors took place throughout the conference, with the opportunity for our participants to seek contact and advice.

Completing the picture of our programme, we proudly presented keynote presentations from leaders in data science and data-driven business, both researchers and practitioners. These keynotes provided all participants the opportunity to come together and share views on challenges and trends.

Keynote presentations were given by: Josef Waltl (Amazon Web Services), Bodo Hoppe (IBM Research & Development GmbH), Christian D. Blakely (PricewaterhouseCoopers Switzerland), Peter Parycek (Danube University Krems, ÖFIT/Fraunhofer Fokus Berlin), Stefan Wegenkittl (Salzburg University of Applied Sciences) and David C. Anastasiu (San José State University).

We thank all speakers for sharing their insights with our community. We also thank Michael Ruzicka for moderating the industry track and for his help and support for its realisation. Especially we like to thank our colleagues, Nicole Siebenhandl, Maximilian Tschuchnig and Dominik Vereno, for their enormous and contructive commitment to organizing and conducting the iDSC.

We are convinced that the iDSC proceedings will give scientists and practitioners an excellent reference to the current activities in the field of data science and also impulses for further studies, research activities and applications in all discussed areas. The visibility is ensured by the support of our publishing house Springer / Vieweg Wiesbaden Germany.

Peter Haber, Thomas Lampoltshammer and Manfred Mayr

Conference Chairs

Data Science is everywhere

The 2nd International Data Science Conference (iDSC2019) combined state-of-the-art methods from academia with industry best practices.

From 22nd to 24th May 2019, representatives of the academic and industry world met for the 2nd International Data Science Conference at Salzburg University of Applied Sciences, Austria. Over 40 top-class speakers from industry and academia presented trends and technologies, as well as new methods and analytical approaches; overall, more than 120 international experts took the opportunity to discuss new challenges and developments regarding data science in industry, research, education, and society.

The technological progress in IT and the resulting changes in the business and social environment ("digitalization") have led to an increase of the amount of data. Several billion terabyte of data have been produced up until now, and the number is growing exponentially. "Data collection, permeates all areas of our lives: from standard use cases such as the analysis of our shopping behaviour followed by purchase suggestions, up to competitive sports", explains conference organizer Peter Haber, senior lecturer at the degree programme Information Technology & Systems Management, who has initiated the conference together with Manfred Mayr (Salzburg University of Applied Sciences) and Thomas Lampoltshammer (Danube University Krems).

In many places, data mining and machine learning have become part of the occupational routine in order to generate value out of the flood of data. "The permeation in all areas was evident at the conference. The presented use cases ranged from market and trend analysis, the energy, health and sports sector, to predictive maintenance in the industrial sector", summarizes Manfred Mayr, co-organizer and academic programme director of the degree programme Business Informatics and Digital Transformation, the key areas of the conference.

Besides talks about state-of-the-art methods of data analytics by researchers from the international academic community, business experts gave insights into best practices and obstacles from the practice. The conference highlights:

- **Artificial intelligence and machine learning gain importance**

  More and more manufacturing companies place greater emphasis on machine learning, a form of artificial intelligence that enables systems to learn independently from data. They, for example, optimize their maintenance with this method. Monitors, among others, get equipped with intelligent sensors to collect data. For analysing the data, so-called predictive anticipation algorithms give information about the projected maintenance need. With this, a lot of money can be saved.

- **Developing the business models of the future with blockchain technology**

  Bodo Hoppe, distinguished engineer at IBM Research & Development GmbH, explained in his keynote that the potential for new business models lies in the continuous monitoring of the production and supply chain. With "IBM© Food Trust" the software giant IBM has developed a new trust and transparency system for the food industry. Hoppe: "The solution is based on blockchain technology and grants authorised users immediate access to meaningful data of the food supply chain, from the farm via the stores to the end consumer."

- **Small Data: Solutions for SME**

  The 2nd International Data Science Conference showed possibilities for the optimization of business models, especially for small and medium sized enterprises (SME). "Often they are lacking the special knowledge and the supposed data to evaluate the chances and risks for their own company", knows conference organizer Peter Haber. "Recent improvements in the area of deep learning, however, provide solutions for these cases with only limited data available", says David Anastasiu, assistant professor at the Department of Computer Engineering at San José State University. Thus, SME could improve their business operations significantly.

  Josef Waltl, Global Segment Lead at Amazon Web Services and alumnus of the degree programme Information Technology and Systems Management, presented, therefore, in his keynote a cloud-based machine-learning system that could help companies with their use cases.

- **Open data & scientific cooperation as chances for SME**

  Peter Parycek, head of the Department for E-Governance and Administration at Danube University Krems, wants to open the access to the data of the big players. SME could use these open data as a reference to generate transfer possibilities for their use cases. Experts, therefore, advice SME to invest in data scientists.

  Stefan Wegenkittl, academic programme director of the degree programme Applied Image & Signal Processing, as well as head of the applied mathematics and data mining department at the degree programme Information Technology and Systems Management, emphasized the possibility to intensify the exchange with higher education institutions: "Suitable solutions can be found if agile development and management processes are connected with current data science research questions."

- **Social benefits: detecting autism & revealing insider trade**

  Data science offers exciting possibilities not only for business purposes. David Anastasiu from the San José University presented his research project about the computer-based detection of autism. Anastasiu and his team received the respective data from electrocardiograms, which measure the activity of all heart muscle fibres, and data about the skin conductance.

  With this information, they determined the course of the relaxation times which – as they could show in their research project – can give an indication about whether a person is suffering from autism or not. Anastasiu: "Our model reached an accuracy of 99.33 percent. Diagnosis by doctors are around 82.9 percent accurate." Jad Rayes and Priya Mani from the George Mason University presented another field of application with social benefit. They are able to detect insider trade activities. With this information, the police could reduce the crime on the capital market.

Positive feedback for the conference

The high number of participants, as well as their positive feedback proved that the concept of the conference is successful. "The participants liked the format as well as the size. The great mix of academic inputs and practical examples from the industry was appreciated", so Haber proudly. The next conference will take place from 13th to 14th May 2020. "But this time in Dornbirn, in cooperation with FH Vorarlberg" – another indicator for the successful concept.

Data Science at Salzburg University of Applied Sciences

With a particular specialization in Data Science & Analytics in the master's programme, the degree programme Information Technology & Systems Management at Salzburg University of Applied Sciences offers prospective data scientists a perfect education. Alumni have, besides mathematical and statistical core competences, fundamental knowledge in the areas machine learning, data mining and deep learning. Data science is also one of the cross-cutting topics taught in the Business Informatics and Digital Transformation study degree.

Nicole Siebenhandl, Sandra Lagler
Local Organizer, Communications

Organization

Organizing Institution

Salzburg University of Applied Sciences

Conference Chairs

Peter Haber	Salzburg University of Applied Sciences
Thomas J. Lampoltshammer	Danube University Krems
Manfred Mayr	Salzburg University of Applied Sciences

Local Organizers

Nicole Siebenhandl	Salzburg University of Applied Sciences
Maximilian E. Tschuchnig	Salzburg University of Applied Sciences
Dominik Vereno	Salzburg University of Applied Sciences

Program Committee

David Anastasiu	San José State University
Arne Bathke	University of Salzburg
Markus Breunig	University of Applied Science Rosenheim
Frank Danielsen	University of Agder
Eric Davis	Industrial Labs
Günther Eibl	Salzburg University of Applied Sciences
Süleyman Eken	Kocaeli University
Karl Entacher	Salzburg University of Applied Sciences
Mohammad Ghoniem	Luxembourg Institute of Science and Technology
Elmar Kiesling	Vienna University of Technology
Michael Gadermayr	Salzburg University of Applied Sciences
Manolis Koubarakis	National and Kapodistrian University of Athens
Maria Leitner	Austrian Institute of Technology
Elena Lloret Pastor	University of Alicante
Giuseppe Manco	University of Calabria
Robert Merz	Vorarlberg University of Applied Sciences
Edison Pignaton de Freitas	Federal University of Rio Grande do Sul
Florina Piroi	Vienna University of Technology
Kathrin Plankensteiner	Vorarlberg University of Applied Sciences

Siegfried Reich	Salzburg Research
Peter Reiter	Vorarlberg University of Applied Sciences
Michael Ruzicka	Cockpit-Consulting
Marta Sabou	Vienna University of Technology
Johannes Scholz	Graz University of Technology, Institute of Geodesy
Axel Straschil	pmIT Consult
Lőrinc Thurnay	Danube University Krems
Andreas Unterweger	Salzburg University of Applied Sciences
Gabriela Viale Pereira	Danube University Krems
Stefan Wegenkittl	Salzburg University of Applied Sciences
Karl-Heinz Weidmann	Vorarlberg University of Applied Sciences
Anneke Zuiderwijk	van Eijk – Delft University of Technology

Reviewer

David Anastasiu	San José State University
Frank Danielsen	University of Agder
Eric Davis	Industrial Labs
Günther Eibl	Salzburg University of Applied Sciences
Süleyman Eken	Kocaeli University
Karl Entacher	Salzburg University of Applied Sciences
Michael Gadermayr	Salzburg University of Applied Sciences
Peter Haber	Salzburg University of Applied Sciences
Manolis Koubarakis	National and Kapodistrian University of Athens
Thomas J. Lampoltshammer	Danube University Krems
Maria Leitner	Austrian Institute of Technology
Robert Merz	Vorarlberg University of Applied Sciences
Elena Lloret Pastor	University of Alicante
Edison Pignaton de Freitas	Federal University of Rio Grande do Sul
Florina Piroi	Vienna University of Technology
Kathrin Plankensteiner	Vorarlberg University of Applied Sciences
Siegfried Reich	Salzburg Research
Peter Reiter	Vorarlberg University of Applied Sciences
Johannes Scholz	Graz University of Technology, Institute of Geodesy
Lőrinc Thurnay	Danube University Krems
Maximilian E. Tschuchnig	Salzburg University of Applied Sciences
Andreas Unterweger	Salzburg University of Applied Sciences
Gabriela Viale Pereira	Danube University Krems
Stefan Wegenkittl	Salzburg University of Applied Sciences

Table of Content

German Abstracts

Data Analytics | Complexity

Exploring Insider Trading Within Hypernetworks
Erforschung von Insiderhandel innerhalb von Hypernetzwerken

Insiderhandel kann lähmende Auswirkungen auf die Wirtschaft haben, und seine Verhinderung ist entscheidend für die Sicherheit und Stabilität der globalen Märkte. Es wird angenommen, dass Insider, die zu ähnlichen Zeiten handeln, Informationen austauschen. Wir analysieren 400 Unternehmen und 2.000 Insider und identifizieren interessante Handelsmuster in diesen Netzwerken, die auf illegale Aktivitäten hinweisen können. Insider werden entweder als routinemäßige oder opportunistische Händler eingestuft, sodass wir uns auf gut getaktete und hochprofitable Handelsaktivitäten des letzteren Typs konzentrieren können. Durch die Einstufung des Handels und der Analyse der Rolle jedes Händlers in einem Hypernetzwerk zeigen sich Gruppen von opportunistischen und routinemäßigen Händlern. Diese Idee bildet die Grundlage eines graphenbasierten Erkennungsalgorithmus, der darauf abzielt, Händler zu identifizieren, die zu opportunistischen Gruppen gehören. Die Handelseinstufung und Handelsgruppen bieten interessante Möglichkeiten, zuverlässigere Überwachungssysteme zu entwickeln, die automatisch illegale Aktivitäten auf den Märkten erkennen und vorhersagen können, mit welcher Wahrscheinlichkeit diese Aktivitäten in Zukunft eintreten werden.

Chance influence in datasets with a large number of features
Zufallseinfluss bei Datensätzen mit einer großen Anzahl von Merkmalen

Die maschinelle Lernforschung, z. B. Genomforschung, basiert oft auf wenigen Datensätzen, die zwar sehr viele Merkmale, aber nur kleine Stichprobengrößen enthalten. Diese Konfiguration fördert den Zufallseinfluss auf den Lernprozess und die Auswertung. Frühere Forschungen konzentrierten sich auf die Verallgemeinerung von Modellen, die auf Grundlage solcher Daten erhalten wurden. In diesem Beitrag untersuchen wir den Zufallseinfluss auf die Klassifizierung und Regression. Wir zeigen empirisch, wie groß der Zufallseinfluss auf diese Datensätze ist. Damit werden die daraus gezogenen Schlussfolgerungen in Frage gestellt. Wir verknüpfen die Beobachtungen der Zufallskorrelation mit dem Problem der Methodengeneralisierung. Schließlich besprechen wir die Zufallskorrelation und nennen Richtlinien, die den Zufallseinfluss verringern.

Data Analytics | NLP and Semantics

Combining Lexical and Semantic Similarity Methods for News Article Matching
Kombination von lexikalischen und semantischen Ähnlichkeitsmethoden beim Abgleich von Nachrichtenartikeln

Der Abgleich von Nachrichtenartikeln verschiedener Quellen mit unterschiedlichen Schilderungen ist ein entscheidender Schritt für die verbesserte Verarbeitung des Online-Nachrichtenflusses. Obwohl es Studien zum Auffinden gleicher oder nahezu gleicher Dokumente in verschiedenen Bereichen gibt, beschäftigt sich keine dieser Studien mit der Gruppierung von Nachrichtentexten auf Grundlage ihrer Ereignisse oder Quellen. Ein bestimmtes Ereignis kann aufgrund der unterschiedlichen politischen Ansichten der Verlage aus sehr unterschiedlichen Perspektiven mit unterschiedlichen Wörtern, Konzepten und Meinungen geschildert werden. Wir entwickeln eine neue Methode zum Abgleich von Nachrichtendokumenten, die mehrere verschiedene lexikalische Abgleichswerte mit Ähnlichkeitswerten aufgrund von semantischen Darstellungen von Dokumenten und Wörtern kombiniert.

Unser experimentelles Ergebnis zeigt, dass diese Methode beim Nachrichtenabgleich sehr erfolgreich ist. Wir entwickeln darüber hinaus einen überwachten Ansatz, indem wir Nachrichtenpaare als gleich oder ungleich kennzeichnen und danach strukturelle und zeitliche Merkmale extrahieren. Das Einstufungsmodell lernte anhand dieser Merkmale, insbesondere der zeitlichen Merkmale, und konnte so angelernt werden. Unsere Ergebnisse zeigen, dass das überwachte Modell eine höhere Leistung erzielen kann und somit besser geeignet ist, die oben genannten Schwierigkeiten beim Abgleich von Nachrichten zu lösen.

The Effectiveness of the Max Entropy Classifier for Feature Selection
Die Wirksamkeit des Max-Entropy-Klassifikators für die Merkmalsauswahl

Die Merkmalsauswahl ist die Aufgabe, die Anzahl der Eingangsmerkmale für eine Klassifizierung systematisch zu reduzieren. Bei der Verarbeitung natürlicher Sprache wird die grundlegende Merkmalsauswahl in der Regel durch das Auslassen gängiger Stoppwörter erreicht.

Um die Anzahl der Eingangsmerkmale noch weiter zu reduzieren, werden bei einer zahlenbasierten Eingangsdarstellung tatsächliche Merkmalsauswahlverfahren wie Transinformation oder Chi-Quadrat verwendet. Wir schlagen einen aufgabenorientierten Ansatz zur Auswahl von Merkmalen auf Grundlage von Gewichten vor, die von einem Max-Entropy-Klassifikator gelernt wurden, der für die Klassifizierung trainiert wurde.

Die restlichen Merkmale können dann von anderen Klassifikatoren genutzt werden, um die eigentliche Klassifizierung durchzuführen. Experimente mit verschiedenen Aufgaben der natürlichen Sprachverarbeitung bestätigen, dass die gewichtsbasierte Methode mit zahlenbasierten Methoden vergleichbar ist. Die Anzahl der Eingangsmerkmale kann unter Beibehaltung der Klassifizierungsleistung erheblich reduziert werden.

Impact of Anonymization on Sentiment Analysis of Twitter Postings
Auswirkung der Anonymisierung auf die Stimmungsanalyse von Twitter-Posts

Der Prozess der strategischen Modellierung und der gesamte Bereich der Strategieplanung sind komplex und stellen die Entscheidungsträger vor große Herausforderungen. Eine Herausforderung dabei ist die Einbeziehung der Bürger in den Entscheidungsprozess. Dies kann über verschiedene Formen der E-Beteiligung erfolgen, wobei die aktive/passive Bürgergewinnung eine Möglichkeit darstellt, aktuelle Diskussionen zu Themen und Problemen, die für die Allgemeinheit relevant sind, zu steuern.

Ein besseres Verständnis der Gefühle gegenüber bestimmten Themen und das daraus resultierende Verhalten der Bürger kann öffentlichen Verwaltungen neue Einsichten bieten. Gleichzeitig ist es aber wichtiger denn je, die Privatsphäre der Bürger zu respektieren, rechtskonform zu handeln und damit das Vertrauen der Öffentlichkeit zu fördern. Während die Einführung der Anonymisierung zur Gewährleistung der Wahrung der Privatsphäre eine angemessene Lösung für die genannte Probleme darstellt, ist jedoch noch unklar, ob und inwieweit sich die Anonymisierung von Daten auf die aktuellen Datenanalysetechnologien auswirkt. Daher untersucht dieser Forschungsbeitrag die Auswirkungen der Anonymisierung auf die Stimmungs-analyse in sozialen Medien im Rahmen von Smart Governance.

Drei Anonymisierungsalgorithmen werden mit Twitter-Daten getestet und die Ergebnisse werden auf Veränderungen innerhalb der resultierenden Stimmung hin analysiert. Die Ergebnisse zeigen, dass die vorgeschlagenen Anonymisierungsansätze tatsächlich einen messbaren Einfluss auf die Stimmungsanalyse haben. Dies geht sogar so weit, dass Ergebnisse für die weitere Verwendung im Bereich der strategischen Modellierung möglicherweise problematisch sein können.

Data Analytics | Modelling

A Data-Driven Approach for Detecting Autism Spectrum Disorders
Ein datengesteuerter Ansatz zur Erkennung von Autismus-Spektrum-Störungen

Autismus-Spektrum-Störungen (ASS) sind eine Gruppe von Erkrankungen, die sich durch Beeinträchtigungen der wechselseitigen sozialen Interaktion und das Vorhandensein von eingeschränkten und sich wiederholenden Verhaltensweisen kennzeichnen. Die aktuellen Mechanismen zur Erkennung von ASS sind entweder subjektiv (umfragebasiert) oder konzentrieren sich nur auf die Reaktionen auf einen einzelnen Stimulus.

In dieser Arbeit entwickeln wir maschinelle Lernmethoden zur Vorhersage von ASS auf Grundlage von Daten aus Elektrokardiogrammen (EKG) und der elektrodermalen Aktivität (EDA), die während eines sogenannten *Sensory Challenge Protocol* (SCP) gesammelt wurden, mit dem die Reaktionen auf acht Stimuli von 25 Kindern mit ASS und 25 normal entwickelten Kindern zwischen 5 und 12 Jahren beobachtet wurden. Durch die Länge der Zeitsequenzen ist es schwierig, herkömmliche maschinelle Lernalgorithmen zur Analyse dieser Datentypen zu verwenden. Stattdessen haben wir Verarbeitungs-techniken für Merkmale entwickelt, die eine effiziente Analyse der Sequenzen ohne Effektivitätsverlust ermöglichen.

Die Ergebnisse unserer Analyse der aufgezeichneten Zeitsequenzen bestätigten unsere Hypothese, dass autistische Kinder besonders stark von bestimmten sensorischen Stimuli betroffen sind. Darüber hinaus erreichte unser gemeinsames ASS-Vorhersagemodell eine Genauigkeit von 93,33 %, d. h. 13,33 % besser als das Beste aus 8 verschiedenen Basismodellen, die wir getestet haben.

Optimal Regression Tree Models through Mixed Integer Programming
Optimale Regressionsbaummodelle durch gemischt-ganzzahlige Optimierung

Die Regressionsanalyse kann zur Vorhersage von Ausgabevariablen bei einem Satz bekannter unabhängiger Variablen eingesetzt werden. Durch Regression wird eine Funktion, die die Beziehung zwischen den Variablen erfasst, an die Daten angepasst. Regressionsbaummodelle sind in der Literatur beliebt, da sie schnell berechnet werden können und einfach zu interpretieren sind.

Das Erstellen komplexer Baumstrukturen kann jedoch zu einer Überanpassung der Lerndaten und damit zu einem schlechten Vorhersagemodell führen. Diese Arbeit stellt einen Regressionsbaumalgorithmus, der anhand der mathematischen Optimierung Daten optimal in zwei Unterbereiche gliedert – sogenannte Knoten –, sowie einen statistischen Test zur Beurteilung der Qualität der Aufteilung vor. Eine Reihe von öffentlich zugänglichen Literaturbeispielen wurde verwendet, um die Leistung der Methode mit anderen, in der Literatur verfügbaren Methoden zu vergleichen.

A Spatial Data Analysis Approach for Public Policy Simulation in Thermal Energy Transition Scenarios

Ein Ansatz zur Analyse von Geodaten für die Simulation der öffentlichen Strategie in thermischen Energiewendeszenarien

Der Beitrag erläutert einen Ansatz zur Simulation der Auswirkungen öffentlicher Strategien auf die thermischen Energiewendeszenarien in städtischen Gemeinden. Der Beitrag beschreibt die zugrundeliegenden Methoden zur Berechnung des Heizenergiebedarfs von Gebäuden und die Gründe für potenzielle Zonen für thermische Energiesysteme.

Um die Auswirkungen öffentlicher Strategien auf die Gemeinden zu simulieren, entwickelten die Autoren ein räumliches, agentenbasiertes Modell, bei dem die Gebäude die Hauptobjekte sind, die sich basierend auf einer Reihe von technischen und soziodemografischen Parametern ändern können.

Um ein räumliches, agentenbasiertes Modell mit Daten zu füllen, muss eine Reihe von Open-Source- und kommerziell verfügbaren Datensätzen räumlich analysiert und zusammengeführt werden. Die ersten Ergebnisse der räumlichen, agentenbasierten Modellierung zeigen, dass öffentliche Strategien für die thermische Energiewende entsprechend simuliert werden können.

Data Analytics | Comprehensibility

A Probabilistic Approach to Web Waterfall Charts
Wahrscheinlichkeitsansatz für webbasierte Wasserfalldiagramme

Ziel dieses Beitrags ist es, einen effizienten und zuverlässigen Modellierungsansatz für probabilistische Wasserfalldiagramme zu erarbeiten, der die Zeitabläufe von webbasierten Ressourcen veranschaulicht und sich dabei besonders auf dessen Anpassung an große Datenmengen konzentriert. Eine Umsetzung mit realen Daten wird diskutiert und anhand von Beispielen veranschaulicht. Die Methode basiert auf der nichtparametrischen Dichteschätzung und wir besprechen einige subtile Aspekte, wie verrauschte Eingaben und singuläre Daten. Wir untersuchen des Weiteren Optimierungstechniken für die numerische Integration, die im Rahmen der Modellierung entsteht.

Facilitating Public Access to Legal Information: A Conceptual Model for Developing an Agile Data-driven Decision Support System
Erleichterung des öffentlichen Zugriffs auf rechtliche Informationen: Ein konzeptionelles Modell zur Entwicklung eines flexiblen, datengesteuerten Entscheidungsunterstützungssystems

Das europäische Rechtssystem ist vielschichtig und komplex. Seit seiner Einführung wurden große Mengen an Rechtsdokumenten erstellt. Dies hat erhebliche Auswirkungen auf die europäische Gesellschaft, deren verschiedene verfassungsgebende Organe regelmäßigen Zugriff auf präzise und zeitnahe rechtliche Informationen benötigen, aber häufig mit einem grundlegenden Verständnis der Rechtssprache zu kämpfen haben. Das Projekt, auf das sich dieser Beitrag konzentriert, schlägt die Entwicklung einer Reihe von nutzerzentrierten Diensten vor, die die Bereitstellung und Visualisierung von Rechtsinformationen in Echtzeit für Bürger, Unternehmen und Verwaltungen auf Grundlage einer Plattform gewährleistet, die von semantisch kommentierten Big Open Legal Data (BOLD) unterstützt wird. Ziel dieses Forschungsbeitrags ist es, durch die Entwicklung eines konzeptionellen Modells kritisch zu untersuchen, wie die aktuelle Nutzeraktivität mit den Komponenten der vorgeschlagenen Projektplattform interagiert. Aufgrund des Model Driven Design (MDD) wird die vorgeschlagene Projektarchitektur beschrieben. Sie wird durch die Anwendung des Agent Oriented Modelling (AOM) auf Grundlage von UML-(Unified Modelling Language)-Nutzeraktivitätsdiagrammen ergänzt, um sowohl die Nutzeranforderungen der vorgeschlagenen Plattform zu entwickeln als auch die Abhängigkeiten aufzuzeigen, die zwischen den verschiedenen Komponenten des vorgeschlagenen Systems bestehen.

Do we have a Data Culture?
Gibt es eine Datenkultur?

Heutzutage ist die Einführung einer „Datenkultur“ oder der „datengesteuerte“ Betrieb für viele Führungskräfte ein wünschenswertes Ziel. Was bedeutet es jedoch, wenn ein Unternehmen behauptet, eine Datenkultur zu haben? Es gibt dafür keine klare Definition. Dieser Beitrag zielt darauf ab, das Verständnis einer Datenkultur in Unternehmen zu verbessern, indem die aktuelle Verwendung des Begriffs besprochen wird. Er zeigt, dass Datenkultur eine Art Organisationskultur ist.

Eine besondere Form der Datenkultur ist die datengesteuerte Kultur. Wir kommen zu dem Schluss, dass sich eine datengesteuerte Kultur durch die Befolgung bestimmter Werte, Verhaltensweisen und Normen kennzeichnet, die eine effektive Datenanalyse ermöglichen. Neben diesen Werten, Verhaltensweisen und Normen erläutert dieser Beitrag die professionellen Rollen, die für eine datengesteuerte Kultur notwendig sind. Wir schließen die wichtige Rolle des Dateneigners ein, der die Datenkultur durch die Datenlenkung erst zu einer datengesteuerten Kultur macht. Schließlich schlagen wir eine Definition der datengesteuerten Kultur vor, die sich auf das Streben nach einer datenbasierten Entscheidungsfindung und einem ständig verbesserten Prozess der Datenanalyse konzentriert.

Dieser Beitrag unterstützt Teams und Organisationen jeder Größe, die ihre – nicht notwendigerweise großen – Datenanalysefähigkeiten verbessern möchten, indem wir auf häufig vernachlässigte, nicht-technische Anforderungen aufmerksam machen: Datenlenkung und eine geeignete Unternehmenskultur.

Short Papers

Neural Machine Translation from Natural Language into SQL with state-of-the-art Deep Learning methods
Neuronale maschinelle Übersetzung natürlicher Sprache in SQL mit modernsten Deep-Learning-Methoden

Einen Text lesen, wichtige Aussagen erkennen, zusammenfassen, Verbindungen herstellen und andere Aufgaben, die Verständnis und Kontext erfordern, sind einfach für Menschen, aber das Trainieren eines Computers in diesen Aufgaben ist eine Herausforderung. Die jüngsten Fortschritte im Bereich Deep Learning ermöglichen es, Texte tatsächlich zu interpretieren und leistungsstarke Ergebnisse bei der natürlichen Sprachverarbeitung zu erzielen. Die Interaktion mit relationalen Datenbanken über natürliche Sprache ermöglicht es Benutzern unterschiedlichster Hintergründe, große Datenmengen auf benutzerfreundliche Weise abzufragen und zu analysieren. Dieser Beitrag fasst die wichtigsten Herausforderungen und unterschiedlichen Ansätze im Zusammenhang mit Natural Language Interfaces to Databases (NLIDB) zusammen. Ein von Google entwickeltes, hochmodernes Übersetzungsmodell – Transformer – wird zur Übersetzung natürlicher Sprachabfragen in strukturierte Abfragen verwendet, um die Interaktion zwischen Benutzern und relationalen Datenbanksystemen zu vereinfachen.

Smart recommendation system to simplify projecting for an HMI/SCADA platform
Intelligentes Empfehlungssystem zur Vereinfachung der Projektierung für eine HMI/SCADA-Plattform

Die Modellierung und Verbindung von Maschinen und Hardware von Produktionsanlagen in HMI/SCADA-Softwareplattformen gilt als zeitaufwändig und erfordert Fachkenntnisse. Ein intelligentes Empfehlungssystem könnte die Projektierung unterstützen und vereinfachen. In diesem Beitrag werden überwachte Lernmethoden erörtert, um dieses Problem zu lösen. Datenmerkmale, Herausforderungen bei der Modellierung und zwei mögliche Modellierungsansätze – 1-aus-n-Code und probabilistische Themenmodellierung – werden besprochen.

Adversarial Networks — A Technology for Image Augmentation
Gegensätzliche Netzwerke – Eine Technologie zur Datenanreicherung

Eine wichtige Anwendung der Datenanreicherung ist die Unterstützung des hochmodernen maschinellen Lernens, um fehlende Werte zu ergänzen und mehr Daten aus einem bestimmten Datensatz zu generieren. Neben Methoden wie Transformation oder Patch-Extraktion können auch gegensätzliche Netzwerke genutzt werden, um die Wahrscheinlichkeitsdichtefunktion der ursprünglichen Daten zu erlernen. Mit sogenannten Generative Adversarial Networks (GANs) können neue Daten aus Rauschen generiert werden, indem ein Generator und ein Diskriminator eingesetzt werden, die in einem Nullsummenspiel versuchen, ein Nash-Gleichgewicht zu finden. Mit diesem Generator kann dann Rauschen in Erweiterungen der ursprünglichen Daten umgewandelt werden. Dieser kurze Beitrag erläutert die Verwendung von GANs, um gefälschte Bilder von Gesichtern zu erzeugen, und enthält Tipps zur Verbesserung des immer noch schwierigen Trainings von GANs.

Using supervised learning to predict the reliability of a welding process
Der Einsatz von überwachtem Lernen zur Vorhersage der Zuverlässigkeit eines Schweißprozesses

In diesem Beitrag wird überwachtes Lernen zur Vorhersage der Zuverlässigkeit von Herstellungsprozessen im industriellen Umfeld verwendet. Zur Illustration wurden die Lebensdauerdaten einer speziellen Vorrichtung aus Blech gesammelt. Es ist bekannt, dass das Schweißen der entscheidende Schritt in der Produktion ist. Um die Qualität der Schweißfläche zu prüfen, wurden mit jedem Gerät End-of-Life-Tests durchgeführt. Zur statistischen Auswertung stehen nicht nur die erfasste Lebensdauer, sondern auch Daten zur Verfügung, die das Gerät vor und nach dem Schweißprozess beschreiben, sowie Messkurven während des Schweißens, z. B. Verlauf über die Zeit. In der Regel werden die Weibull- und Log-Normalverteilung zur Modellierung der Lebensdauer verwendet. Auch in unserem Fall gelten beide als mögliche Verteilungen.

Obwohl beide Verteilungen für die Daten geeignet sind, wird die Log-Normalverteilung verwendet, da der KS-Test und der Bayes'sche Faktor etwas bessere Ergebnisse zeigen. Zur Modellierung der Lebensdauer in Abhängigkeit der Schweißparameter wird ein multivariables, lineares Regressionsmodell verwendet. Um die signifikanten Kovariablen zu finden, wird eine Mischung aus Vorwärtsauswahl und Rückwärtselimination verwendet. Mit dem T-Test wird die Wichtigkeit jeder Kovariable bestimmt, während der angepasste Determinationskoeffizient als globales Anpassungskriterium verwendet wird.

Nachdem das Modell, das die beste Anpassung bietet, bestimmt wurde, wird die Vorhersagekraft mit einer eingeschränkten Kreuzvalidierung und der Residuenquadratsumme bewertet. Die Ergebnisse zeigen, dass die Lebensdauer anhand der Schweißeinstellungen vorhergesagt werden kann. Für die Lebensdauerprognose liefert das Modell genaue Ergebnisse, wenn die Interpolation verwendet wird. Eine Extrapolation über den Bereich der verfügbaren Daten hinaus zeigt jedoch die Grenzen eines rein datengesteuerten Modells auf.