DATA 2021

10th International Conference on Data Science, Technology and Applications

Final Program and Book of Abstracts

6 - 8 July, 2021

http://www.dataconference.org

SPONSORED BY

INSTICC

PAPERS AVAILABLE AT

SCITEPRESS DIGITAL LIBRARY

DATA 2021
Final Program and Book of Abstracts

10th International Conference on Data Science, Technology and Applications

Online Streaming
July 6 - 8, 2021

Sponsored by
INSTICC - Institute for Systems and Technologies of Information, Control and Communication

ACM In Cooperation
ACM SIGMIS - ACM Special Interest Group on Management Information Systems

Printed in Portugal
http://www.dataconference.org
data.secretariat@insticc.org

Table of Contents

Foreword 5
Important Information 7
General Information 8
Program Layout 9
Tuesday Sessions: July 6 17
Wednesday Sessions: July 7 21
Thursday Sessions: July 8 31
Author Index 39

Foreword

This book contains the abstracts and final program of the 10th International Conference on Data Science, Technologies and Applications (DATA 2021) which is sponsored by the Institute for Systems and Technologies of Information, Control and Communication (INSTICC), and held in cooperation with the ACM Special Interest Group on Management Information Systems (ACM SIGMIS). This year DATA was held as a web-based event due to the COVID-19 pandemic, from 6 - 8 July.

This conference brings together researchers, engineers and practitioners interested on databases, big data, data mining, data management, data security and other aspects of information systems and technology involving advanced applications of data.

The high quality of the DATA 2021 program is enhanced by the four keynote lectures, delivered by distinguished speakers who are renowned experts in their fields: Jan Recker (University of Cologne, Germany), Sandro Bimonte (INRAE, France), Hala Skaf-Molli (Nantes University, France), and Volker Markl (German Research Center for Artificial Intelligence (DFKI) and Technische Universität Berlin (TU Berlin), Germany).

DATA 2021 received 64 paper submissions from 27 countries of which 19% were accepted as full papers. In order to evaluate each submission, a double-blind paper review was performed by the Program Committee. All presented papers will be available at the SCITEPRESS Digital Library and will be submitted for indexation by Scopus, Google Scholar, DBLP, Semantic Scholar, Microsoft Academic, EI (Elsevier Index), and Web of Science (Clarivate). As in previous editions of the Conference, based on the reviewer's evaluations and the presentations, selected authors from the conference will be invited to submit extended versions of their papers for a book that will be published by Springer with the best papers of DATA 2021. This year, a short list of best papers will be invited for a post-conference special issue in the Springer Nature Computer Science Journal.

The program for this conference required the dedicated effort of many people. Firstly, we must thank the authors, whose research efforts are herewith recorded. Next, we thank the members of the Program Committee and the auxiliary reviewers for their diligent and professional reviewing. We would also like to deeply thank the invited speakers for their invaluable contribution and for taking the time to prepare their talks. Finally, we gratefully acknowledge the professional support of the conference secretariat and INSTICC team for all organizational processes, especially given the need to introduce online streaming, forum management, direct messaging facilitation and other web-based activities in order to make it possible for DATA 2021 authors to present their work and share ideas with colleagues in spite of the logistic difficulties caused by the current pandemic situation.

We wish you all an exciting conference and we look forward to having additional research results presented at the next edition of DATA.

Christoph Quix, Hochschule Niederrhein, University of Applied Sciences and Fraunhofer FIT, Germany
Slimane Hammoudi, ESEO, ERIS, France
Wil van der Aalst, RWTH Aachen University, Germany

Important Information

Event App
Download the Event App from the Play Store and App Store now, to have mobile access to the technical program and also to get notifications and reminders concerning your favorite sessions.

Create Your Own Schedule *
The option “My Program” gives you the possibility of creating a selection of the sessions that you plan to attend. This service also allows you to print-to-pdf all papers featured in your selection thus creating a pdf file per conference day.

Online Access to the Proceedings *
In the option “Proceedings and Final Program” you cannot only download the proceedings but also access the digital version of the book of abstracts with the final program.

Digital Access to the Receipt *
By clicking on the option “Delegate Home” and then “Registration Documents” it will enable you to access the final receipt which confirms the registration payment.

Keynotes Videos
The keynote lectures will also be available on video on the website after the event, as long as the appropriate authorization from the keynote is received, so you will be able to see them again or watch them should you have missed one.

Survey
Every year we conduct a survey to access the participants’ satisfaction with the conference and gather the suggestions. You will receive an e-mail after the event with the detailed information. Your contribution will be carefully analysed and a serious effort to react appropriately will be made.

* Please login to PRIMORIS (www.insticc.org/Primoris), select the role “Delegate” and the correct event.

If you have any doubt, we will be happy to help you at the Welcome Desk.

General Information

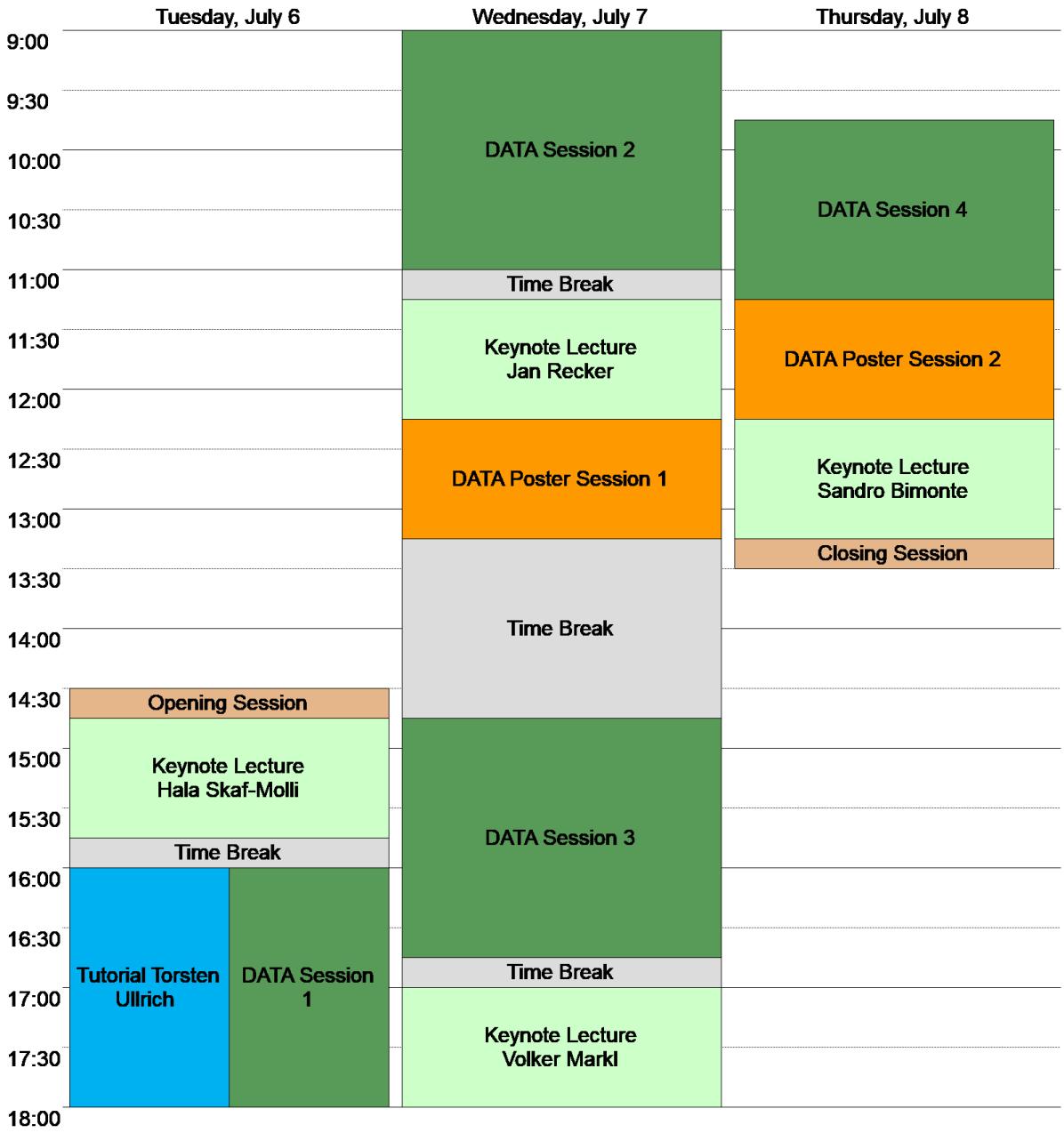
Welcome Desk
Tuesday, July 6 – Open from 14:15 to 18:00
Wednesday, July 7 – Open from 08:45 to 18:00
Thursday, July 8 – Open from 09:30 to 13:30

Opening Session
Tuesday, July 6, at 14:30 in the Plenary 1 room.

Closing Session
Thursday, July 8, at 13:15 in the Plenary 1 room.

Secretariat Contacts
DATA Secretariat
Address: Avenida de S. Francisco Xavier, Lote 7 Cv. C
2900-616 Setúbal, Portugal
Tel.: +351 265 520 185
Fax: +351 265 520 186
e-mail: data.secretariat@insticc.org
website: http://www.dataconference.org

Program Layout



Final Program and Book of Abstracts

Contents

Tuesday Sessions: July 6

Opening Session (14:30 - 14:45)
Room Plenary 1 . . . 19

Keynote Lecture (14:45 - 15:45)
Room Plenary 1 . . . 19
Querying Decentralized Knowledge Graphs, *by Hala Skaf-Molli* . . . 19

Session 1 (16:00 - 18:00)
Room 2: *Data Science* . . . 19
Complete Paper #31: textPrep: A Text Preprocessing Toolkit for Topic Modeling on Social Media Data, *by Rob Churchill and Lisa Singh* . . . 19
Complete Paper #60: Forecasting Stock Market Trends using Deep Learning on Financial and Textual Data, *by Georgios-Markos Chatziloizos, Dimitrios Gunopulos and Konstantinos Konstantinou* . . . 19
Complete Paper #14: Deep Learning for RF-based Drone Detection and Identification using Welch's Method, *by Mahmoud Almasri* . . . 19
Complete Paper #44: Data Driven Hybrid Approach for Health Monitoring and Fault Detection in Military Ground Vehicles, *by Indu Shukla, Antoinette Silas, Haley Dozier, Brandon Hansen and W. Bond* . . . 20
Complete Paper #50: Estimating Territory Risk Relativity for Auto Insurance Rate Regulation using Generalized Linear Mixed Models, *by Shengkun Xie, Chong Gan and Clare Chua-Chow* . . . 20

Tutorial (16:00 - 18:00)
Room 1 . . . 20
Introduction to the Data Analysis of Time Series, *by Torsten Ullrich* . . . 20

Wednesday Sessions: July 7

Session 2 (09:00 - 11:00)
Room 3: *Data Mining* . . . 23
Complete Paper #11: Similarity of Software Libraries: A Tag-based Classification Approach, *by Maximilian Auch, Maximilian Balluff, Peter Mandl and Christian Wolff* . . . 23
Complete Paper #20: A Comparative Study on Inflated and Dispersed Count Data, *by Monika Arora, Yash Kalyani and Shivam Shanker* . . . 23
Complete Paper #6: Data Mining for Animal Health to Improve Human Quality of Life: Insights from a University Veterinary Hospital, *by Oscar Tamburis, Elio Masciari, Christian Esposito and Gerardo Fatone* . . . 23
Complete Paper #19: A Survey of Social Emotion Prediction Methods, *by Abdullah Alsaedi, Phillip Brooker, Floriana Grasso and Stuart Thomason* . . . 23
Complete Paper #23: A Network based Approach for Reducing Variant Diversity in Production Planning and Control, *by Shailesh Tripathi, Sonja Strasser and Herbert Jodlbauer* . . . 23
Room 4: *Text Analytics* . . . 24
Complete Paper #61: A Graph-based Approach at Passage Level to Investigate the Cohesiveness of Documents, *by Ghulam Sarwar and Colm O'Riordan* . . . 24
Complete Paper #62: A Reference Process for Judging Reliability of Classification Results in Predictive Analytics, *by Simon Staudinger, Christoph Schuetz and Michael Schrefl* . . . 24
Complete Paper #39: Well-Being in Plastic Surgery: Deep Learning Reveals Patients' Evaluations, *by Joschka Kersting and Michaela Geierhos* . . . 24
Complete Paper #8: GRASP: Graph-based Mining of Scientific Papers, *by Navid Nobani, Mauro Pelucchi, Matteo Perico, Andrea Scrivanti and Alessandro Vaccarino* . . . 25
Complete Paper #13: A Comparison of Methods for the Evaluation of Text Summarization Techniques, *by Marcello Barbella, Michele Risi and Genoveffa Tortora* . . . 25

Contents

Keynote Lecture (11:15 - 12:15)
Room Plenary 1 . . . 25
From Representation to Mediation: Modeling Information Systems in a Digital World, *by Jan Recker* . . . 25

Poster Session 1 (12:15 - 13:15)
Room Posters DATA . . . 25
Abstract #13: Real Estate Price Prediction with Artificial Intelligence Techniques, *by Sophia Zhou* . . . 25
Abstract #18: Knowledge Graph based Electrical Circuit Simulation and Component Selection, *by Rahman Syed, Johannes Bayer and Felix Thoma* . . . 26
Complete Paper #2: Determining How Different Factors Affect Police-Allegation's Sustainability in Chicago using Decision-Tree, *by Linxin Yang* . . . 26
Complete Paper #3: Archival and Museum Information as a Component of the Common Digital Space of Scientific Knowledge, *by N. Kalenov, I. Sobolevskaya and A. Sotnikov* . . . 26
Complete Paper #4: Building an Integrated Relational Database from Swiss Nutrition's (menuCH) and Multiple Swiss Health Datasets Acquired from 1992 to 2012 for Data Mining Purposes, *by Timo Lustenberger, Helena Jenzer and Farshideh Einsele* . . . 27
Complete Paper #9: Motif-based Classification using Enhanced Sub-Sequence-Based Dynamic Time Warping, *by Mohammed Alshehri, Frans Coenen and Keith Dures* . . . 27
Complete Paper #21: WFDU-net: A Workflow Notation for Sovereign Data Exchange, *by Heinrich Pettenpohl, Daniel Tebernum and Boris Otto* . . . 27
Complete Paper #33: Semantic Entanglement on Verb Negation, *by Yuto Kikuchi, Kazuo Hara and Ikumi Suzuki* . . . 27
Complete Paper #38: Using BPMN for ETL Conceptual Modelling: A Case Study, *by Bruno Oliveira, Óscar Oliveira and Orlando Belo* . . . 28

Session 3 (14:45 - 16:45)
Room 3: *Data Management and Quality* . . . 28
Complete Paper #7: DERM: A Reference Model for Data Engineering, *by Daniel Tebernum, Marcel Altendeitering and Falk Howar* . . . 28
Complete Paper #18: DQ-MeeRKat: Automating Data Quality Monitoring with a Reference-Data-Profile-Annotated Knowledge Graph, *by Lisa Ehrlinger, Alexander Gindlhumer, Lisa-Marie Huber and Wolfram Wöß* . . . 28
Complete Paper #42: Semantic Enrichment of Vital Sign Streams through Ontology-based Context Modeling using Linked Data Approach, *by Sachiko Lim, Rahim Rahmani and Paul Johannesson* . . . 28
Room 4: *Mobile Data and Data Integrity* . . . 29
Complete Paper #30: An Efficient Representation of Enriched Temporal Trajectories, *by Nieves Brisaboa, Antonio Fariña, Diego Otero-González and Tirso Rodeiro* . . . 29
Complete Paper #27: Database Recovery from Malicious Transactions: A Use of Provenance Information, *by Theppatorn Rhujittawiwat, John Ravan, Ahmed Saaudi, Shankar Banik and Csilla Farkas* . . . 29
Complete Paper #36: Invers Natural Number System to Maintain User-defined Sequence of Data Records, *by Seyfettin Öztürk* . . . 29

Keynote Lecture (17:00 - 18:00)
Room Plenary 1 . . . 29
Database Systems and Information Management: Trends and a Vision, *by Volker Markl* . . . 29

Thursday Sessions: July 8

Session 4 (09:45 - 11:15)
Room 3: *Data Science Applications* . . . 33
Complete Paper #51: Detecting Twitter Fake Accounts using Machine Learning and Data Reduction Techniques, *by Ahmad Homsi, Joyce Al Nemri, Nisma Naimat, Hamzeh Abdul Kareem, Mustafa Al-Fayoumi and Mohammad Abu Snober* . . . 33
Complete Paper #12: Biomedical Dataset Recommendation, *by Xu Wang, Frank van Harmelen and Zhisheng Huang* . . . 33
Complete Paper #59: Tailoring Taint Analysis for Database Applications in the K Framework, *by Md. Alam and Raju Halder* . . . 33
Complete Paper #48: Toward a Multimodal Multitask Model for Neurodegenerative Diseases Diagnosis and Progression Prediction, *by Sofia Lahrichi, Maryem Rhanoui, Mounia Mikram and Bouchra El Asri* . . . 33
Room 8: *Business Analytics* . . . 34
Complete Paper #49: A Study on the Effects of Response Time on Travel Package Attributes, *by Usha Ananthakumar and Sagun Pai* . . . 34
Complete Paper #54: A Longitudinal Model for Song Popularity Prediction, *by Ahmet Çimen and Enis Kayış* . . . 34
Complete Paper #63: A Company's Corporate Reputation through the Eyes of Employees Measured with Sentiment Analysis of Online Reviews, *by R. Loke and R. Lam-Lion* . . . 34

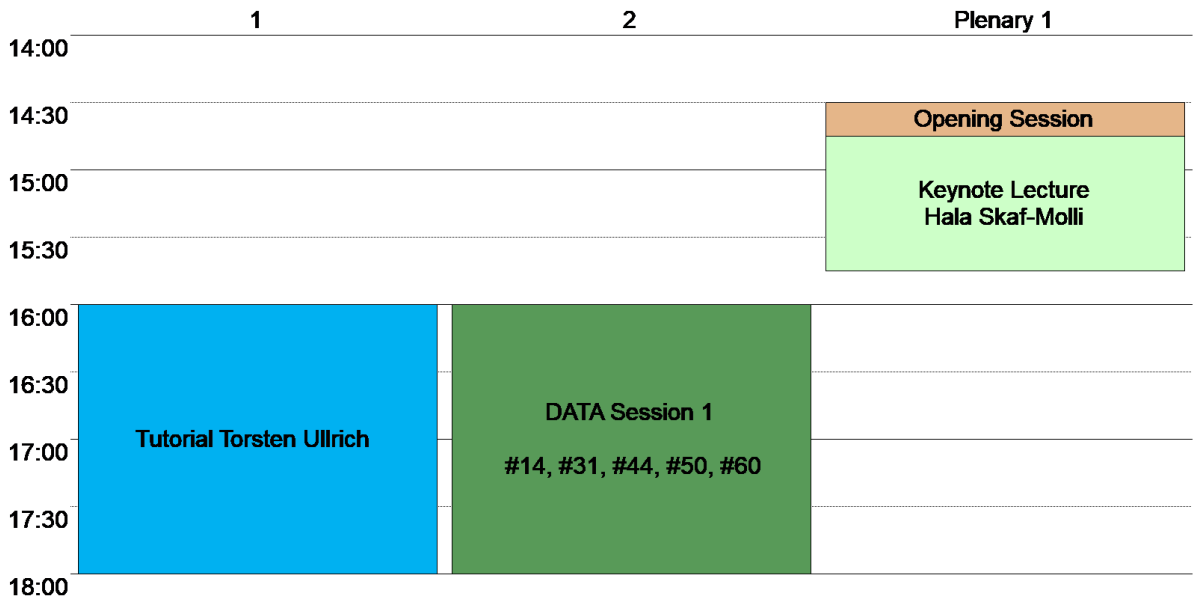
Poster Session 2 (11:15 - 12:15)
Room Posters DATA . . . 34
Abstract #19: Automatic Measurement of Corporate Reputation for Retail Companies from Online Public Data on the Web, *by Marselo Sitorus and Rob Loke* . . . 34
Complete Paper #28: Predicting Shopping Intent of e-Commerce Users using LSTM Recurrent Neural Networks, *by Konstantinos Diamantaras, Michail Salampasis, Alkiviadis Katsalis and Konstantinos Christantonis* . . . 35
Complete Paper #41: Applied Feature-oriented Project Life Cycle Classification, *by Oliver Böhme and Tobias Meisen* . 35
Complete Paper #45: Impact of Duplicating Small Training Data on GANs, *by Yuki Eizuka, Kazuo Hara and Ikumi Suzuki* 35
Complete Paper #46: Making Data Big for a Deep-learning Analysis: Aggregation of Public COVID-19 Datasets of Lung Computed Tomography Scans, *by Francesca Lizzi, Francesca Brero, Raffaella Cabini, Maria Fantacci, Stefano Piffer, Ian Postuma, Lisa Rinaldi and Alessandra Retico* . . . 36
Complete Paper #52: Knowledge Graph Analysis of Russian Trolls, *by Chih-yuan Li, Soon Chun and James Geller* . . 36
Complete Paper #53: Aspect Based Sentiment Analysis on Online Review Data to Predict Corporate Reputation, *by R. Loke and W. Reitter* . . . 36
Complete Paper #57: Evo-Path: Querying Data Evolution through Complex Changes, *by Theodora Galani, Yannis Stavrakas, George Papastefanatos and Yannis Vassiliou* . . . 37
Complete Paper #58: Enhanced AI On-the-Edge 3D Vision Accelerated Point Cloud Spatial Computing Solution, *by Gaurav Kumar Wankar and Shubham Vohra* . . . 37

Keynote Lecture (12:15 - 13:15)
Room Plenary 1 . . . 37
A Profile-aware Methodological Framework for Collaborative Multidimensional Modeling: Agro-biodiversity Case Study, *by Sandro Bimonte* . . . 37

Closing Session (13:15 - 13:30)
Room Plenary 1 . . . 37

Tuesday Sessions: July 6

Tuesday Sessions: July 6 Program Layout



Opening Session	DATA
14:30 - 14:45	Room Plenary 1

Keynote Lecture	DATA
14:45 - 15:45	Room Plenary 1

Querying Decentralized Knowledge Graphs

Hala Skaf-Molli
Nantes University, Nantes, France

Abstract: Following the principles of linked data, billions of RDF data have been produced and hundreds of interconnected knowledge graphs are available through public SPARQL endpoints. However, existing SPARQL servers suffer from availability and scalability issues. In this talk, I will present the latest research results that address these issues. I will present approaches that balance the cost of query processing between data providers and data consumers in order to reduce the pressure on data providers. I will conclude my presentation with open research questions.

Session 1A	DATA
16:00 - 18:00	Room 2
Data Science	

Complete Paper #31

textPrep: A Text Preprocessing Toolkit for Topic Modeling on Social Media Data

Rob Churchill and Lisa Singh
Georgetown University, U.S.A.

Keywords: Text Preprocessing, Topic Modeling, Data Science, Social Media, textPrep.

Abstract: With the rapid growth of social media in recent years, there has been considerable effort toward understanding the topics of online discussions. Unfortunately, state of the art topic models tend to perform poorly on this new form of data, due to their noisy and unstructured nature. There has been a lot of research focused on improving topic modeling algorithms, but very little focused on improving the quality of the data that goes into the algorithms. In this paper, we formalize the notion of preprocessing configurations and propose a standardized, modular toolkit and pipeline for performing preprocessing on social media texts for use in topic models. We perform topic modeling on three different social media data sets and in the process show the importance of preprocessing and the usefulness of our preprocessing pipeline when dealing with different social media data. We release our preprocessing toolkit code (textPrep) in a python package for others to use for advancing research on data mining and machine learning on social media text data.

Complete Paper #60

Forecasting Stock Market Trends using Deep Learning on Financial and Textual Data

Georgios-Markos Chatziloizos, Dimitrios Gunopulos and Konstantinos Konstantinou
Department of Informatics and Telecommunications, National and Kapodistrian University of Athens, Athens, Greece

Keywords: Technical Analysis, Sentiment Analysis, Machine Learning, Stock Market.

Abstract: Stock market research has increased significantly in recent years. Researchers from both economics and computer science backgrounds are applying novel machine learning techniques to the stock market. In this paper we combine some of the techniques used in both of these fields, namely Technical Analysis and Sentiment Analysis techniques, to show whether or not it is possible to successfully forecast the trend of the stock price and to what extent. Using the four tickers AAPL, GOOG, NVDA and S&P 500 Information Technology, we collected historical financial data and historical textual data and we used each type of data individually and in unison, to display in which case the results were more accurate and more profitable. We describe in detail how we analysed each type of data, how we used it to come up with our results.

Complete Paper #14

Deep Learning for RF-based Drone Detection and Identification using Welch's Method

Mahmoud Almasri
LABSTICC, UMR 6285 CNRS, ENSTA Bretagne, 2 rue F. Verny, 29806 Brest Cedex 9, France

Keywords: Artificial Intelligence, Deep Neural Network, Drone Identification and Classification, Welch.

Abstract: Radio Frequency (RF) combined with the deep learning methods promised a solution to detect the presence of the drones. Indeed, the classical techniques (i.e. radar, vision and acoustics, etc.) suffer several drawbacks such as difficult to detect the small drones, false alarm of flying birds or balloons, the influence of the wind on the performance, etc. For an effective drones's detection, two main stages should be established: Feature extraction and feature classification. The proposed approach in this paper is based on a novel feature extraction method and an optimized deep neural network (DNN). At first, we present a novel method based on Welch to extract meaningful features from the RF signal of drones. Later on, three optimized Deep Neural Network (DNN) models are considered to classify the extracted features. The first DNN model can be used to detect the presence of the drones and contains two classes. The second DNN help us to detect and recognize the type of the drone with 4 classes: A class for each drone and the last one for the RF background activities. In the third model, 10 classes have been considered: the presence of the drone, its type, and its flight mode (i.e. Stationary, Hovering, flying with or without video recording). Our proposed approach can achieve an average accuracy higher than 94% and it significantly improves the accuracy, up to 30%, compared to existing methods.

Complete Paper #44

Data Driven Hybrid Approach for Health Monitoring and Fault Detection in Military Ground Vehicles

Indu Shukla, Antoinette Silas, Haley Dozier, Brandon Hansen and W. Bond

US Army ERDC, Information Technology Laboratory, Vicksburg, MS, 39180, U.S.A.

Keywords: Long Short-Term Memory (LSTM), Vector Auto Regression (VAR), Prognostics and Health Management (PHM).

Abstract: This paper presents a data driven hybrid approach for Prognostics and Health Management (PHM) of military ground vehicles to mitigate a number of the unexpected failures, enabling intelligent decision-making for improved performance, safety, reliability, and maintainability. For military ground vehicles, the Controller Area Network (CAN) bus provides sensor data for collection and analysis. In this study we used collected operational time-series data for generating future operational time series data for military ground vehicles. Our sensor data share stochastic trends with more than one-time dependent variable to develop Vector AutoRegression (VAR) models suitable to forecast operational data. We have developed Long Short-Term Memory (LSTM) fault detection models which ingest VAR forecasted data to identify fault detection. Our experimental results show our hybrid approach provides promising fault diagnosis performance. Root mean squared error, mean absolute percentage error and mean absolute error have been used as the evaluation criteria.

Complete Paper #50

Estimating Territory Risk Relativity for Auto Insurance Rate Regulation using Generalized Linear Mixed Models

Shengkun Xie[1], Chong Gan[2] and Clare Chua-Chow[1]

[1] *Global Management Studies, Ted Rogers School of Management, Ryerson University, Toronto, Canada*

[2] *Department of Mathematics and Statistics, University of Guelph, Guelph, Canada*

Keywords: Generalized Linear Mixed Models, Rate-making, Insurance Rate Regulation, Business Data Analytics.

Abstract: Territory risk analysis has played an essential role in auto insurance rate regulation. It aims to obtain a set of regions to estimate their respective relativities to reflect the regional risk. Cluster as a latent variable has not yet been considered in modelling the regional risk of auto insurance. In this work, spatially constrained clustering is first applied to insurance loss data to form such regions. The generalized linear mixed model is then proposed to derive the risk relativities for obtained clusters and then for each basic rating unit. The results are compared to the ones from generalized linear models. The Forward Sortation Area (FSA) grouping to a specific region by spatially constrained clustering is to reduce the insurance rate heterogeneity caused by some smaller number of risk exposures. The spatially constrained clustering and risk relativity estimation help obtain a set of territory risk benchmarks, which can be used in rate filings within the regulation process. It also provides guidance for auto insurance companies on rate-making. The proposed methodologies could be helpful and applicable in many other fields, including business data analytic.

Tutorial
16:00 - 18:00 **Room 1**

Introduction to the Data Analysis of Time Series

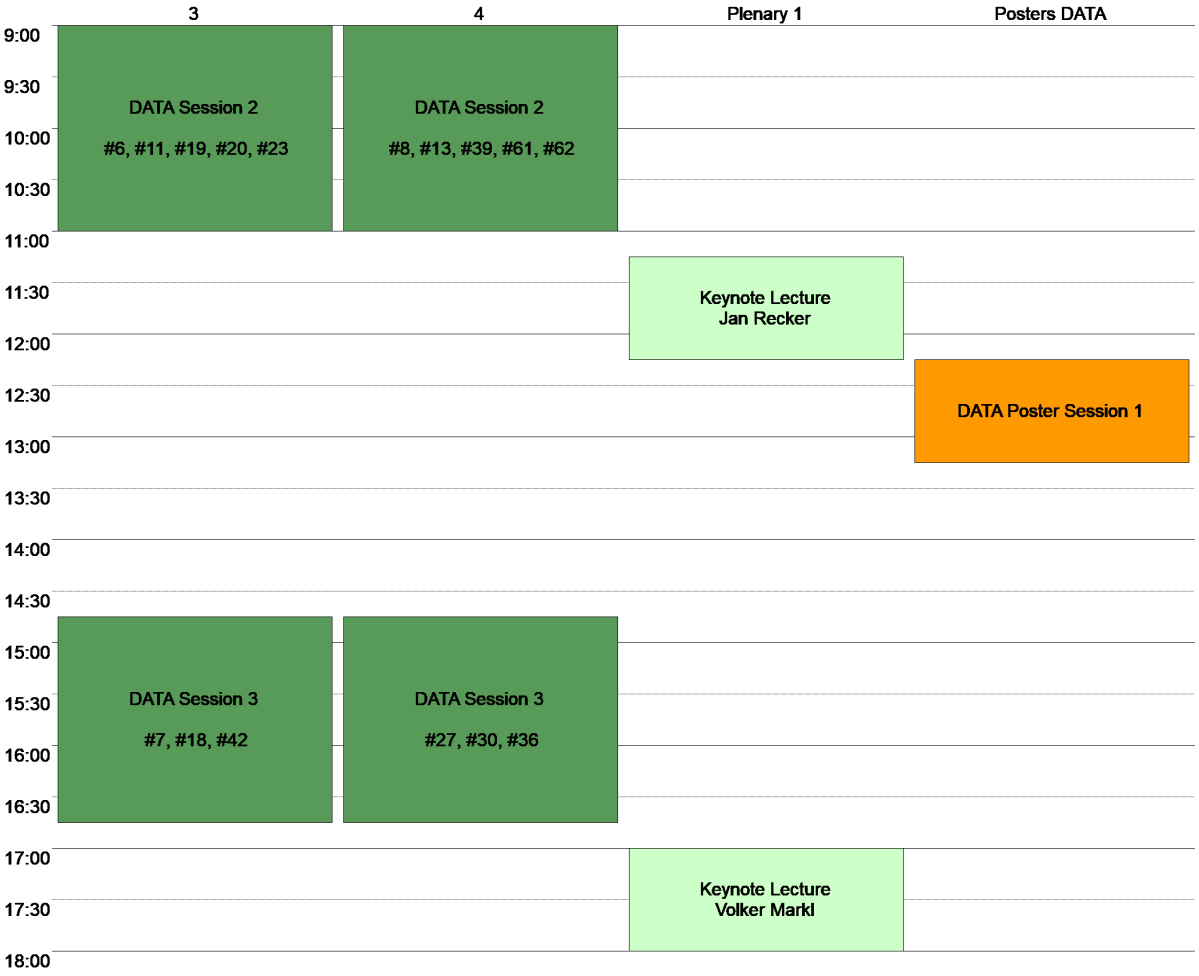
Torsten Ullrich

Fraunhofer Austria Research GmbH, Austria

Abstract: For time series, a variety of statistical methods exist to calculate future values and to make reliable forecasts. In this context, both the trend development and the extreme values (minima and maxima) are of interest. In this tutorial, statistical and machine learning methods for modeling will be presented and used for forecasting. A comparison of these methods shall give an overview and be a general decision support, which provides a recommendation for action for own data sets.

Wednesday Sessions: July 7

Wednesday Sessions: July 7 Program Layout



Session 2A **DATA**
09:00 - 11:00 **Room 3**
Data Mining

Complete Paper #11

Similarity of Software Libraries: A Tag-based Classification Approach

Maximilian Auch[1], Maximilian Balluff[1], Peter Mandl[1] and Christian Wolff[2]

[1] *University of Applied Sciences Munich, Lothstraße 34, 80335 Munich, Germany*

[2] *University of Regensburg, Universitätsstraße 31, 93053 Regensburg, Germany*

Keywords: Software Libraries, Classification, Tags, Similarity, Naíve Bayes, Logistic Regression, Random Forest, Neural Network.

Abstract: The number of software libraries has increased over time, so grouping them into classes according to their functionality simplifies repository management and analyses. With the large number of software libraries, the task of categorization requires automation. Using a crawled dataset based on Java software libraries from Apache Maven repositories as well as tags and categories from the indexing platform MvnRepository.com, we show how the data in this set is structured and point out an imbalance of classes. We introduce a class mapping relevant for the procedure, which maps the libraries from very specific, technical classes into more generic classes. Using this mapping, we investigate supervised machine learning techniques that classify software libraries from the dataset based on their available tags. We show that a tag-based approach to classify libraries with an accuracy of 97.46% can be achieved by using neural networks. Overall, we found techniques such as neural networks and naíve Bayes more suitable in this use case than a logistic regression or a random forest.

Complete Paper #20

A Comparative Study on Inflated and Dispersed Count Data

Monika Arora, Yash Kalyani and Shivam Shanker

Department of Mathematics, Indraprastha Institute of Information Technology, Delhi, India

Keywords: Zero Inflated Data Regression Models, Dispersion, Machine Learning, Predictive Modeling.

Abstract: The availability of zero inflated count data has led to the demonstration of various statistical models and machine learning algorithms to be applied in diverse fields such as healthcare, economics and travel. However, in real life there could be a count $k > 0$ that is inflated. There are only a few studies on $k-$ inflated count models. To the best of our knowledge, there is no article that demonstrates the machine learning algorithms on such data sets. We apply existing $k-$ inflated count models as well as machine learning algorithms on travel data to compare the prediction and fitness of the models and find the significant covariates. Our study shows that the $k-$ inflated models provide a good fit to the data, however, the predictions from machine learning algorithms are superior. This study can be extended further to include other artificial neural network approaches on a larger data set.

Complete Paper #6

Data Mining for Animal Health to Improve Human Quality of Life: Insights from a University Veterinary Hospital

Oscar Tamburis[1], Elio Masciari[2], Christian Esposito[3] and Gerardo Fatone[1]

[1] *Dept. of Veterinary Medicine and Animal Productions, Federico II University, Via Delpino 1, Naples, Italy*

[2] *Dept. of Computer Science and Electrical Engineering, Federico II University, Naples, Italy*

[3] *Dept. of Computer Science, University of Salerno, Fisciano (SA), Italy*

Keywords: Veterinary Medicine, Electronic Medical Record, Decision Tree Algorithm, One Health.

Abstract: The increasing importance of Veterinary Informatics is driving the implementation of integrated veterinary information management systems (VIMS) for the capture, storage, analysis and retrieval of animal data. In this paper, a decision tree algorithm was implemented, starting from the database of the University Veterinary Hospital at Federico II University of Naples, aiming at building a predictive model for an effective recognition of neoplastic diseases and zoonoses for cats and dogs focusing to Campania Region, in order to figure out, according to the One (Digital) Health perspective specifics, the connection between humans, animals, and surrounding environment.

Complete Paper #19

A Survey of Social Emotion Prediction Methods

Abdullah Alsaedi[1], Phillip Brooker[2], Floriana Grasso[1] and Stuart Thomason[1]

[1] *Department of Computer Science, University of Liverpool, U.K.*

[2] *Department of Sociology, Social Policy and Criminology, University of Liverpool, U.K.*

Keywords: Social Emotion Prediction Methods, Social Emotion, Reader's Emotion.

Abstract: Emotions are an important factor that affects our communication. Considerable research has been done to detect and classify emotion in text. However, most deal with emotion from the writer's perspective. Social emotion is the emotion of the reader when exposed to the text. With the increased use of social media, many works are performed for social emotion prediction. In this paper, we attempt to provide a survey of social emotion prediction methods. To the best of our knowledge, this is the first work to survey the literature of social emotion, review methods, and used techniques, compare the methods, and highlight their limitations.

Complete Paper #23

A Network based Approach for Reducing Variant Diversity in Production Planning and Control

Shailesh Tripathi, Sonja Strasser and Herbert Jodlbauer

University of Applied Sciences Upper Austria, Austria

Keywords: Discrete Event Simulation, Network Analysis, Bipartite Graph, Community Detection, Production Planning and Control, Data Mining.

Abstract: This paper presents a network-based procedure for selecting representative materials using routings of materials as features and applies this procedure to a sheet metal processing case study which is used for parameterizing discrete event simulation models for PPC control. The discrete event simulation model (simgen) is a generic and scalable model that is commonly used to deal with optimization problems in production planning and control, such as manufacturing resource planning. The preparatory steps of discrete event simulations for production planning and control are data preprocessing, parameterization, and experimental design. Given the complexity of the manufacturing environment, discrete event simulation models must incorporate appropriate model details for parameterization and a practical approach to experimental design to ensure efficient execution of simulation models in a reasonable time. The parameterization for discrete event simulation is not trivial; it requires optimizing parameter settings for different materials dependent on routing, bill of materials complexity, and other production process-related features. For a suitable parameterization that completes the execution of discrete event simulation in an expected time, we must reduce variant diversity to an optimized level that removes redundant materials and reflects the validity of the overall production scenario. We employ a network based approach by constructing a bipartite graph and Jaccard-index measure with an overlap threshold to group similar materials using routing features and identify representative materials and manufacturing subnetworks, thus reducing the complexity of products and manufacturing routes.

Session 2B **DATA**
09:00 - 11:00 **Room 4**
Text Analytics

Complete Paper #61

A Graph-based Approach at Passage Level to Investigate the Cohesiveness of Documents

Ghulam Sarwar and Colm O'Riordan
Department of Information Technology, National University of Ireland, Galway, Ireland

Keywords: Passage-based Document Retrieval, Passage Similarity Graph, Document Cohesion, Inter-passage Similarity, Weighted Graph, Query Difficulty, Re-ranking.

Abstract: Approaches involving the representation of documents as a series of passages have been used in the past to improve the performance of ad-hoc retrieval systems. In this paper, we represent the top returned passages as a graph with each passage corresponding to a vertex. We connected the vertices (passages) that belongs to the same document to form a graph. The underlying intuition behind this approach is to identify some measure of the cohesiveness of the documents. We introduce a graph-based approach at the passage level to calculate the cohesion score of each document. The scores for both relevant and non-relevant documents are compared, and we illustrate that the cohesion score differs for relevant and non-relevant. Moreover, we also re-ranked the documents by applying the cohesion score with a document similarity score to inspect its impact on the system's performance.

Complete Paper #62

A Reference Process for Judging Reliability of Classification Results in Predictive Analytics

Simon Staudinger, Christoph Schuetz and Michael Schrefl
Institute of Business Informatics, Data and Knowledge Engineering, Johannes Kepler University Linz, Austria

Keywords: Business Intelligence, Business Analytics, Decision Support Systems, Data Mining, CRISP-DM.

Abstract: Organizations employ data mining to discover patterns in historic data. The models that are learned from the data allow analysts to make predictions about future events of interest. Different global measures, e.g., accuracy, sensitivity, and specificity, are employed to evaluate a predictive model. In order to properly assess the reliability of an individual prediction for a specific input case, global measures may not suffice. In this paper, we propose a reference process for the development of predictive analytics applications that allow analysts to better judge the reliability of individual classification results. The proposed reference process is aligned with the CRISP-DM stages and complements each stage with a number of tasks required for reliability checking. We further explain two generic approaches that assist analysts with the assessment of reliability of individual predictions, namely perturbation and local quality measures.

Complete Paper #39

Well-Being in Plastic Surgery: Deep Learning Reveals Patients' Evaluations

Joschka Kersting[1] and Michaela Geierhos[2]
[1] *Paderborn University, Warburger Str. 100, Paderborn, Germany*
[2] *Bundeswehr University Munich, Research Institute CODE, Carl-Wery-Straße 22, Munich, Germany*

Keywords: Aspect-based Sentiment Analysis, Information Extraction, Deep Learning, Transformer Applications.

Abstract: This study deals with aspect-based sentiment analysis, the correlation of extracted aspects and their sentiment polarities with metadata. There are millions of review texts on the Internet that cannot be analyzed and thus people cannot benefit from the contained information. While most research so far has focused on explicit aspects from product or service data (e.g., hotels), we extract and classify implicit and explicit aspect phrases from German-language physician review texts. We annotated aspect phrases that indicate ratings about the doctor's practice, such as waiting time or general perceived well-being conveyed by all staff members of a practice. We also apply a sentiment polarity classifier. While we compare several traditional and transformer networks, we apply the best model, the multilingual XLM-RoBERTa, to a dedicated German-language dataset dealing with plastic surgeons. We choose plastic surgery as sample domain because it is especially sensitive with its relation to a person's self-image and felt acceptance. In addition to standard evaluation measures such as Precision, Recall, and F1-Score, we correlate our results with metadata from physician review websites, such as a physician's gender. We figure out several correlations and present methods for analyzing unstructured review texts to enable service improvements in healthcare.

Complete Paper #8

GRASP: Graph-based Mining of Scientific Papers

Navid Nobani[1,2], Mauro Pelucchi[3], Matteo Perico[4], Andrea Scrivanti[3] and Alessandro Vaccarino[3]

[1] *Dept. of Informatics, Systems & Communication, University of Milan-Bicocca, Milan, Italy*

[2] *Digital Attitude, Milan, Italy*

[3] *CRISP Research Center, University of Milan-Bicocca, Milan, Italy*

[4] *Oròbix, Bergamo, Italy*

Keywords: Graph Networks, Scientific Documents, Information Retrieval, Literature Review.

Abstract: Over the past two decades, academia has witnessed numerous tools and search engines which facilitate the retrieval procedure in the literature review process and aid researchers to review the literature with more ease and accuracy. These tools mostly work based on a simple textual input which supposedly encapsulates the primary keywords in the desired research areas. Such tools mainly suffer from the following shortcomings: (i) they rely on textual search queries that are expected to reflect all the desired keywords and concepts, and (ii) shallow results which makes following a paper through time via citations a cumbersome task. In this paper, we introduce GRASP, a search engine that retrieves scientific papers starting from a sub-graph query provided by the user, offering (i) a list of time papers based on the query and (ii) a graph with papers and authors as vertices and edges being cited and published-by. GRASPhas been created using a Neo4j graph database, based on DBLP and AMiner corpora provided by their API. Acting performance evaluation by asking ten computer science experts, we demonstrate how GRASPcan efficiently retrieve and rank the most related papers based on the user's input.

Complete Paper #13

A Comparison of Methods for the Evaluation of Text Summarization Techniques

Marcello Barbella, Michele Risi and Genoveffa Tortora

Department of Computer Science, University of Salerno, Fisciano (SA), Italy

Keywords: Automatic Text Summarization Algorithms, Extractive, Abstractive, ROUGE Metric, Bert.

Abstract: Automatic Text Summarization techniques aim to extract key information from one or more input texts automatically, producing summaries and preserving the meaning of content. These techniques are divided into two main families: Extractive and Abstractive, which differ for their operating mode. The former picks up sentences directly from the document text, whilst the latter produces a summary by interpreting the text and rephrases sentences by incorporating information. Therefore, there is the need to evaluate and verify how close a summary is to original text. The research question is: how to evaluate the quality of the summaries produced by these techniques? Different metrics and scores have been proposed in the literature (e.g., ROUGE) for the evaluation of text summarization. Thus, the main purpose of this paper is to deeply estimate the behaviour of the ROUGE metric. In particular, we performed a first experiment to compare the metric efficiency for the evaluation of the Abstractive versus Extractive Text Summarization algorithms while, in a second one, we compared the obtained score for two different summary approaches: the simple execution of a summarization algorithm versus the multiple execution of different algorithms on the same text. The conclusions lead to the following interesting results: ROUGE does not achieve excellent results, because it has similar performance on both the Abstractive and Extractive algorithms; multiple execution works better than single one most of the time.

Keynote Lecture **DATA**
11:15 - 12:15 **Room Plenary 1**

From Representation to Mediation: Modeling Information Systems in a Digital World

Jan Recker

University of Hamburg, Germany

Abstract: The role of information systems is changing in an increasingly digitalized world. Does this situation mean that established conceptual modeling practices relevant to the analysis and design of systems must change as well? In this talk, I will answer this question with a definite and affirmative "yes". I will review the traditional assumptions around the conceptual modeling of information systems and demonstrate how advances in digital technology increasingly challenge these assumptions. I will then present a new framework for conceptual modeling that is consistent with the emerging requirements of a digital world. The framework draws attention to the role of conceptual models as mediators between physical and digital realities. It identifies new research questions about grammars, methods, scripts, agents, and contexts that are situated in intertwined physical and digital realities. I will discuss several implications for conceptual modeling scholarship for systems analysis and design that relate to the necessity of developing new methods and grammars for conceptual modeling, broadening the methodological array of conceptual modeling scholarship, and considering new dependent variables.

Poster Session 1 **DATA**
12:15 - 13:15 **Room Posters DATA**

Abstract #13

Real Estate Price Prediction with Artificial Intelligence Techniques

Sophia Zhou

Mass Academy of Math and Science at Worcester Polytechnic Institute, 85 Prescott Street, Worcester, MA 01605, U.S.A.

Keywords: N/A

Abstract: For investors, businesses, consumers, and governments, an accurate assessment of future housing prices is crucial to critical decisions in resource allocation, policy formation, and investment strategies. Previous studies are contradictory about macroeconomic determinants of housing price and largely focused on one or two areas using point prediction. This study aims to develop data-driven models to accurately predict future housing market trends in different markets.

This work studied five different metropolitan areas representing different market trends and compared three time lagging situations: no lag, 6-month lag, and 12-month lag. Linear regression (LR), random forest (RF), and artificial neural network (ANN) were employed to model the real estate price using datasets with S&P/Case-Shiller home price index and 12 demographic and macroeconomic features, such as gross domestic product (GDP), resident population, personal income, etc. in five metropolitan areas: Boston, Dallas, New York, Chicago, and San Francisco. The data from March, 2005 to December 2018 were collected from the Federal Reserve Bank, FBI, and Freddie Mac. In the

original data, some factors are monthly, some quarterly, and some yearly. Thus, two methods to compensate missing values, backfill or interpolation, were compared. The models were evaluated by accuracy, mean absolute error, and root mean square error.
The LR and ANN models outperformed the RF model due to RF's inherent limitations. Both ANN and LR methods generated predictive models with high accuracy (>95%). It was found that personal income, GDP, population, and measures of debt consistently appeared as most important factors. It also showed that technique to compensate missing values in the dataset and implementation of time lag can have significant influence in the model performance and require further investigation. The best performing models varied for each area, but the backfilled 12-month lag LR models and the interpolated no lag ANN models showed best stable performance overall, with accuracies >95% for each city. This study reveals the influence of input variables in different markets. It also provides evidence to support future studies to identify the optimal time lag and data imputing methods for establishing accurate predictive models.

Abstract #18

Knowledge Graph based Electrical Circuit Simulation and Component Selection

Rahman Syed[1], Johannes Bayer[1] and Felix Thoma[2]

[1] *Deutsches Forschungszentrum für Künstliche Intelligenz GmbH, Trippstadter Str. 122, Kaiserslautern, Germany*

[2] *DFKI, Germany*

Abstract: Electrical circuits can be considered graph structures with components (e.g. resistors, capacitors or inductors) as nodes and wiring as edges. For simulation and hardware implementation purposes, these nodes must be equipped with several attributes like electrical properties and referenced against real-world product libraries.
The system presented in this work takes an RDF representation of a netlist and uses Ngspice to calculate circuit parameters. Additional parameters can be specified using formulas represented using RDF. The parameters and signals calculated for devices are then used as constraints to shortlist candidates from the device knowledge graph and shortlisted candidates can then be optimised for cost if marginal costs of procurement for devices is known. Signal and device characteristics matching criteria and unit standardization formulas are also stored as RDF triples to simplify the addition of new device types, circuit characteristics and matching criteria.
The operating point simulation of a circuit outputs the voltage at nodes and current flows in branches. These values and parameters are substituted in formulae to arrive at values such as power consumption for components. Formulae are specified in RDF and the system checks which of the specified formulae for a component can be applied given the set of known parameters. These values and parameters are added to an enhanced RDF representation of the circuit, which can then be used to shortlist devices.
Product information for components such as resistance, capacitance, power output and prices has been collected from web stores to build a knowledge graph of different device types. Multiple physical devices from various manufacturers and vendors, with differing parameters and physical characteristics can match component requirements known at this stage. Our system achieves this with the help of filters for each known parameter for a circuit component to list the most suitable device matches.
Component level shortlist of matching devices from the knowledge graph also provide the designer with pricing information extracted from vendor sites. While detailed costing for the end hardware implementation and cost optimization is still not achievable because of complicated pricing rules and ordering costs, the designer is given an overview of the potential options along with the pricing per piece and the minimum order size.
In further work, the system is envisioned to support the designer with automated constraint checking and device recommendations when circuits are altered.

Complete Paper #2

Determining How Different Factors Affect Police-Allegation's Sustainability in Chicago using Decision-Tree

Linxin Yang

Korhal Internet, Columbus, Ohio, U.S.A.

Keywords: Decision Tree, Criminal Justice, Data Analysis.

Abstract: The Citizen Police Data Project (CPDP) is a database of allegations made against the Chicago Police Department. Reports made against officers are rarely sustained, which results in the perception of little officer accountability and contributes to widespread distrust of law enforcement. Using a decision tree model on the CPDP database, this work explores how the following factors: officer years of employment, complainant type, investigation agency, and allegation severity level, affects the outcome of an allegation work together to increase or decrease the sustainability of allegations made against CPD between 2008 to 2018. The results found that when a CPD employee reports an allegation, it has higher chances to be sustained. However, for allegations reported by civilians, a third-party agency increases the likelihood of allegation sustainability.

Complete Paper #3

Archival and Museum Information as a Component of the Common Digital Space of Scientific Knowledge

N. Kalenov, I. Sobolevskaya and A. Sotnikov

Joint Supercomputer Center of the Russian Academy of Sciences, Branch of Federal State Institution "Scientific Research Institute for System Analysis of the Russian Academy of Sciences" (JSCC RAS, Branch of SRISA), 119334, Moscow, Leninsky Av., 32 a, Russia

Keywords: Scientific Archive, Science Museum, Knowledge Space, Scientific Heritage, Network Technologies, Digital Libraries, Digital Information Resources, Metadata.

Abstract: The Common Digital Space of Scientific Knowledge (CDSSK), in its modern interpretation, is a fundamentally new information environment that accumulates knowledge from various fields of science and is the basis for solving a wide range of problems: from artificial intelligence to the science popularization. One of the prototypes of the CDSSK model is the digital library "Scientific Heritage of Russia" (DL SHR), within which methods and means of integrating heterogeneous digital information (including archival and museum information) related to Russian scientific achievements are being developed. For several years, the Archive of the Russian Academy of Sciences (ARAN) and the V. I. Vernadsky State Geological Museum Russian Academy of Sciences (GGM RAS) participated in the development and in the DL SHR development and filling the DL SHR with digital content of the DL SHR. The paper discusses the metadata profiles adopted for displaying archival and museum objects in the CDSSK, Provides examples of search and visualization.

Complete Paper #4

Building an Integrated Relational Database from Swiss Nutrition's (menuCH) and Multiple Swiss Health Datasets Acquired from 1992 to 2012 for Data Mining Purposes

Timo Lustenberger[1], Helena Jenzer[2] and Farshideh Einsele[1]

[1] *Section of Business Information, Bern University of Applied Sciences, Switzerland*

[2] *Hospital of Psychiatry, University of Zurich, Switzerland*

Keywords: Health Informatics, Data Mining, Nutritional and Health Databases, Nutritional and Chronical Databases, Modelling and Managing Large Data Systems, Data Management for Analytics, Large Scale Databases, Database Architecture and Performance.

Abstract: Objective: The objective of the study was to integrate a large database from Swiss nutrition national survey (menu-CH) with 5 extensive databases derived from 5 consecutive Swiss health national surveys from 1992 to 2012 for data mining purposes. Each database has additionally a demographic base data. An integrated Swiss database is built to later discover critical food consumption patterns linked with lifestyle diseases known to be strongly tied with food consumption and compare the derived rules with the rules resulted with a previous study which used a significantly smaller database. Design: Swiss nutrition national survey (menu-CH) with approx. 2000 respondents from two different surveys, one by Phone and the other by questionnaire along with Swiss health national surveys from 1992 to 2012 with over than 100000 respondents were preprocessed, cleaned, transformed and finally integrated to a unique relational database. Results: The result of this study is an integrated relational database from the Swiss nutritional and 20 years of Swiss health data.

Complete Paper #9

Motif-based Classification using Enhanced Sub-Sequence-Based Dynamic Time Warping

Mohammed Alshehri[1,2], Frans Coenen[1] and Keith Dures[1]

[1] *Department of Computer Science, University of Liverpool, Liverpool, U.K.*

[2] *Department of Computer Science, King Khalid University, Abha, Saudi Arabia*

Keywords: Time Series Analysis, Dynamic Time Warping, K-Nearest Neighbour Classification, Sub-Sequence-Based DTW, Matrix Profile, Motifs.

Abstract: In time series analysis, Dynamic Time Warping (DTW) coupled with k Nearest Neighbour classification, where k = 1, is the most commonly used classification model. Even though DTW has a quadratic complexity, it outperforms other similarity measurements in terms of accuracy, hence its popularity. This paper presents two motif-based mechanisms directed at speeding up the DTW process in such a way that accuracy is not adversely affected: (i) the Differential Sub-Sequence Motifs (DSSM) mechanism and (ii) the Matrix Profile Sub-Sequence Motifs (MPSSM) mechanism. Both mechanisms are fully described and evaluated. The evaluation indicates that both DSSM and MPSSM can speed up the DTW process while producing a better, or at least comparable accuracy, in 90% of cases.

Complete Paper #21

WFDU-net: A Workflow Notation for Sovereign Data Exchange

Heinrich Pettenpohl[1], Daniel Tebernum[1] and Boris Otto[1,2]

[1] *Data Business, Fraunhofer Institute for Software and Systems Engineering, Dortmund, Germany*

[2] *Industrial Information Management, TU Dortmund University, Dortmund, Germany*

Keywords: Petri Net, Data Sovereignty, WFD-net, Usage Control, CTL*.

Abstract: Data is the main driver of the digital economy. Accordingly, companies are interested in maintaining technical control over the usage of their data at any given time. The International Data Spaces initiative addresses exactly this aspect of data sovereignty with usage control enforcement. In this paper, we introduce the so-called Workflow with Data and Usage control network (WFDU-net) model. The data consumer can visually define his or her workflow using the WFDU-net model and annotate the data operations and context. With model checking we validate that the WFDU-net follows the usage policies defined by the data owner. Afterwards, the compliant WFDU-net can be executed by exporting the WFDU-net in a Petri Net Markup Language (PNML). We evaluated our approach by using our example WFDU-net in a data analytics use case.

Complete Paper #33

Semantic Entanglement on Verb Negation

Yuto Kikuchi[1], Kazuo Hara[1] and Ikumi Suzuki[2]

[1] *Yamagata University, 1-4-12 Kojirakawa-machi, Yamagata City, 990-8560, Japan*

[2] *Nagasaki University, 1-14 Bunkyo, Nagasaki City, 852-8521, Japan*

Keywords: Word2vec, Word Vector, Word Analogy Task.

Abstract: The word2vec, developed by Mikolov et al. in 2013, is an epoch-creating method that embeds words into a vector space to capture their fine-grained meaning. However, the reliability of word2vec is inconsistent. To evaluate the reliability of word vectors, we perform Mikolov's word analogy task, where $word_A$, $word_B$, and $word_C$ are provided. Under the condition that $word_B$ exhibits a particular relation with $word_A$, the task involves searching the vocabulary and returning the most relevant word for $word_C$ for the same relation. We conduct an experiment to return negative words for verbs using word2vec for 100 typical Japanese verbs and investigate the effect of context (i.e., surrounding words) on correct or incorrect responses. It is shown that the task fails when the sense of verbs and negative relation are entangled because the semantic calculation of verb negation does not hold.

Complete Paper #38

Using BPMN for ETL Conceptual Modelling: A Case Study

Bruno Oliveira[1], Óscar Oliveira[1] and Orlando Belo[2]

[1] *CIICESI, School of Management and Technology, Porto Polytechnic, Rua do Curral, Felgueiras, Portugal*

[2] *ALGORITMI R&D Centre, University of Minho, Braga, Portugal*

Keywords: Data Warehousing, ETL, Conceptual Modelling, BPMN.

Abstract: One of the most important parts of a Data Warehousing System is the Extract-Transform-Load (ETL) component. It is responsible for extracting, transforming, conciliating, and loading data for supporting decision-making requirements. Usually, due to the complexity of managing heterogeneous data, this component is responsible for consuming most of the resources required for implementing a Data Warehousing System, representing a critical component that compromises the adequacy of the system. Despite their importance, the ETL development method is essentially ad-hoc, which does not always follow or embodies the best practices. With the emergence of Big Data and associated tools, script-based ETL became, even more, a common approach. In the last years, BPMN – Business Process Model and Notation – have been proposed and used to support ETL conceptual models. Still, as an expressive language, it provides different approaches for representing the same requirements. In this paper, we explore the use of BPMN for ETL conceptual modelling, analyzing existing approaches, and proposing a set of guidelines for using this notation in a more consistent way.

Session 3A **DATA**
14:45 - 16:45 **Room 3**
Data Management and Quality

Complete Paper #7

DERM: A Reference Model for Data Engineering

Daniel Tebernum[1], Marcel Altendeitering[1] and Falk Howar[2]

[1] *Data Business, Fraunhofer ISST, Emil-Figge-Strasse 91, 44227 Dortmund, Germany*

[2] *Chair for Software Engineering, TU Dortmund University, Otto-Hahn-Strasse 12, 44227 Dortmund, Germany*

Keywords: Reference Model, SLR, Data Lifecycle, Data Engineering, Research Map.

Abstract: Data forms an essential organizational asset and is a potential source for competitive advantages. To exploit these advantages, the engineering of data-intensive applications is becoming increasingly important. Yet, the professional development of such applications is still in its infancy and a practical engineering approach is necessary to reach the next maturity level. Therefore, resources and frameworks that bridge the gaps between theory and practice are required. In this study, we developed a data engineering reference model (DERM), which outlines the important building-blocks for handling data along the data lifecycle. For the creation of the model, we conducted a systematic literature review on data lifecycles to find commonalities between these models and derive an abstract meta-model. We successfully validated our model by matching it with established data engineering topics. Using the model derived six research gaps that need further attention for establishing a practically-grounded engineering process. Our model will furthermore contribute to a more profound development process within organizations and create a common ground for communication.

Complete Paper #18

DQ-MeeRKat: Automating Data Quality Monitoring with a Reference-Data-Profile-Annotated Knowledge Graph

Lisa Ehrlinger[1,2], Alexander Gindlhumer[1], Lisa-Marie Huber[1] and Wolfram Wöß[1]

[1] *Johannes Kepler University Linz, Altenberger Straße 69, 4040 Linz, Austria*

[2] *Software Competence Center Hagenberg GmbH, Softwarepark 32a, 4232 Hagenberg, Austria*

Keywords: Data Quality Monitoring, Data Profiling, Automation, Knowledge Graphs, Heterogeneous Data, Sensor Data.

Abstract: High data quality (e.g., completeness, accuracy, non-redundancy) is essential to ensure the trustworthiness of AI applications. In such applications, huge amounts of data is integrated from different heterogeneous sources and complete, global domain knowledge is often not available. This scenario has a number of negative effects, in particular, it is difficult to monitor data quality centrally and manual data curation is not feasible. To overcome these problems, we developed DQ-MeeRKat, a data quality tool that implements a new method to automate data quality monitoring. DQ-MeeRKat uses a knowledge graph to represent a global, homogenized view of local data sources. This knowledge graph is annotated with reference data profiles, which serve as quasi-gold-standard to automatically verify the quality of modified data. We evaluated DQ-MeeRKat on six real-world data streams with qualitative feedback from the data owners. In contrast to existing data quality tools, DQ-MeeRKat does not require domain experts to define rules, but can be fully automated.

Complete Paper #42

Semantic Enrichment of Vital Sign Streams through Ontology-based Context Modeling using Linked Data Approach

Sachiko Lim, Rahim Rahmani and Paul Johannesson

Department of Computer and Systems Science, Stockholm University, Kista, Sweden

Keywords: Internet of Things (IoT), Semantic Enrichment, Ontology, Linked Data, Patient Health Monitoring, Patient Management, Vital Sign, Healthcare, Infectious Disease Outbreak.

Abstract: The Internet of Things (IoT) creates an ecosystem that connects people and objects through the internet. IoT-enabled healthcare has revolutionized healthcare delivery by moving toward a more pervasive, patient-centered, and preventive care model. In the ongoing COVID-19 pandemic, it has also shown a great potential for effective remote patient health monitoring and management, which leads to preventing straining the healthcare system. Nevertheless, due to the heterogeneity of data sources and technologies, IoT-enabled healthcare systems often operate in vertical silos, hampering interoperability across different systems. Consequently, such sensory data are rarely shared nor integrated, which can undermine the full potential of IoT-enabled healthcare. Applying semantic technologies to IoT is a promising approach for fulfilling heterogeneity, contextualization, and situation-awareness requirements for real-time healthcare solutions. However, the enrichment of sensor streams has been under-explored in the

existing literature. There is also a need for an ontology that enables effective patient health monitoring and management during infectious disease outbreaks. This study, therefore, aims to extend the existing ontology to allow patient health monitoring for the prevention, early detection, and mitigation of patient deterioration. We evaluated the extended ontology using competency questions and illustrated a proof-of-concept of ontology-based semantic representation of vital sign streams.

Session 3B **DATA**
14:45 - 16:45 **Room 4**
Mobile Data and Data Integrity

Complete Paper #30

An Efficient Representation of Enriched Temporal Trajectories

Nieves Brisaboa, Antonio Fariña, Diego Otero-González and Tirso Rodeiro

Universidade da Coruña, CITIC, Fac. Informática, Database Lab. Elviña, 15071, A Coruña, Spain

Keywords: Compression, Data Management, Trajectories, Correlated Sequences.

Abstract: We present a novel representation of enriched trajectories of a mobile workforce management system. In this system, employees are tracked during their working day and both their routes and the tasks performed at each time instant are recorded. Our proposal tackles the representation of this information paying special attention to the space footprint without neglecting query time. We performed experiments using real and synthetic datasets where we show the compression effectiveness as well as the efficiency at query time. Our results showed that our proposal yields promising results in terms of the space needed to represent both users' locations and activities while performing access queries to the original data within microseconds.

Complete Paper #27

Database Recovery from Malicious Transactions: A Use of Provenance Information

Theppatorn Rhujittawiwat[1], John Ravan[1], Ahmed Saaudi[1], Shankar Banik[2] and Csilla Farkas[1]

[1] *Computer Science & Engineering Dept., University of South Carolina, Columbia, SC, U.S.A.*

[2] *Dept. of Mathematics and Computer Science, The Citadel, The Military College of South Carolina, Charleston, SC, U.S.A.*

Keywords: Database, Malicious Transaction, Security, Dependency Graph, Data Provenance.

Abstract: In this paper, we propose a solution to recover a database from the effects of malicious transactions. The traditional approach for recovery is to execute all non-malicious transactions from a consistent rollback point. However, this approach is inefficient. First, the database will be unavailable until the restoration is finished. Second, all non-malicious transactions that committed after the rollback state need to be re-executed. The intuition for our approach is to re-execute partial transactions, i.e., only the operations that were affected by the malicious transactions. We develop algorithms to reduce the downtime of the database during recovery process. We show that our solution is 1.) Complete, i.e., all the effects of the malicious transactions are removed, 2.) Sound, i.e., all the effects of non-malicious transactions are preserved, and 3.) Minimal, i.e., only affected data items are modified. We also show that our algorithms preserve conflict serializability of the transaction execution history.

Complete Paper #36

Invers Natural Number System to Maintain User-defined Sequence of Data Records

Seyfettin Öztürk

Nokia Solutions and Networks GmbH & Co. KG, Nürnberg, Germany

Keywords: Database Indexing, Storing User-defined Row Sequence, Avoiding Reordering, Increasing Performance, Decreasing Computing Time and Energy Consumption, Reducing Internet Data Traffic.

Abstract: The objective of this paper is to present a method to insert, edit, and delete database records without affecting the sequence of existing data. Typically, databases comprise integer data fields, in this paper named sequence number, meant to determine the user-defined sequence of data records. Inserting new data records or editing the sequence number of data records might cause a resequencing of the existing data records. This resequencing can be avoided by using a numbering system that decreases the value of a number when a digit is added to its end. Such a numbering system allows to insert an infinite quantity of additional sequence numbers between two sequence numbers even if their difference is 1.

Keynote Lecture **DATA**
17:00 - 18:00 **Room Plenary 1**

Database Systems and Information Management: Trends and a Vision

Volker Markl

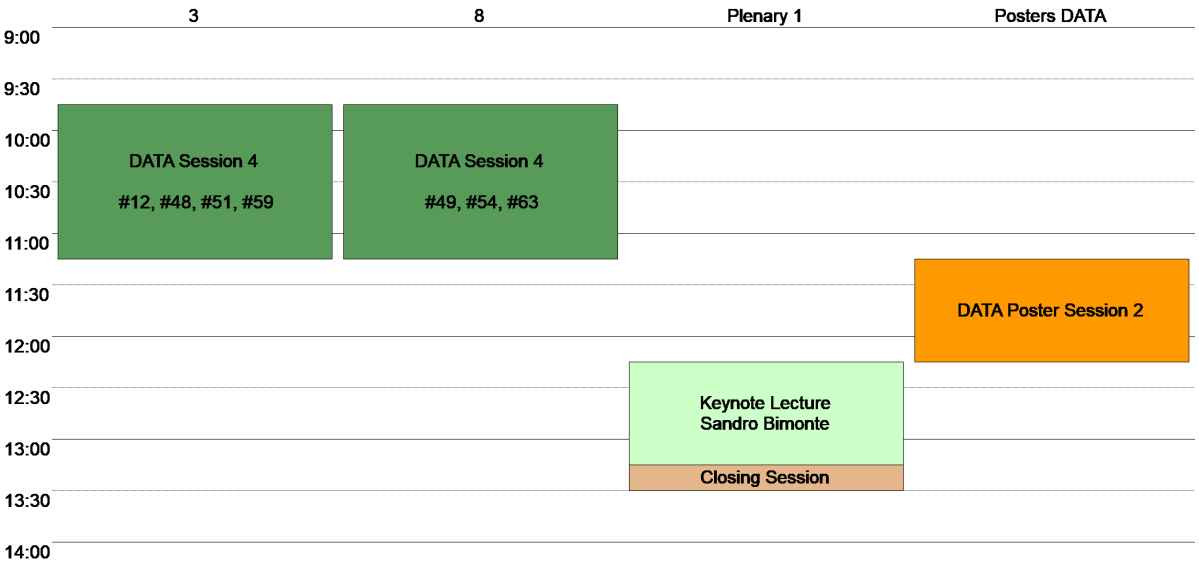
German Research Center for Artificial Intelligence (DFKI) and Technische Universität Berlin (TU Berlin), Germany

Abstract: The global database research community has greatly impacted the functionality and performance of data storage and processing systems along the dimensions that define "big data", i.e., volume, velocity, variety, and veracity. Locally, over the past five years, we have also been working on varying fronts. Among our contributions are: (1) establishing a vision for a database-inspired big data analytics system, which unifies the best of database and distributed systems technologies, and augments it with concepts drawn from compilers (e.g., iterations) and data stream processing, as well as (2) forming a community of researchers and institutions to create the Stratosphere platform to realize our vision. One major result from these activities was Apache Flink, an open-source big data analytics platform and its thriving global community of developers and production users. Although much progress has been made, when looking at the overall big data stack, a major challenge for database research community still remains. That is, how to maintain the ease-of-use despite the increasing heterogeneity and complexity of data analytics, involving specialized engines for various aspects of an end-to-end data analytics pipeline, including, among others, graph-based, linear algebra-based, and relational-based algorithms, and the underlying, increasingly heterogeneous hardware and computing infrastructure. At TU Berlin, DFKI, and the Berlin Institute for Foundations of Learning and Data (BIFOLD) we currently aim to advance research in this field via the Nebula Stream and Agora projects. Our goal is to remedy some of the heterogeneity challenges that hamper developer productivity and limit the use of data science technologies to just the privileged few, who are coveted experts. In this talk, we will outline how state-of-the-art SPEs have to change to exploit the new capabilities of the IoT

and showcase how we tackle IoT challenges in our own system, NebulaStream. We will also present our vision for Agora, an asset ecosystem that provides the technical infrastructure for offering and using data and algorithms, as well as physical infrastructure components.

Thursday Sessions: July 8

Thursday Sessions: July 8 Program Layout



Session 4A **DATA**
09:45 - 11:15 **Room 3**
Data Science Applications

Complete Paper #51

Detecting Twitter Fake Accounts using Machine Learning and Data Reduction Techniques

Ahmad Homsi, Joyce Al Nemri, Nisma Naimat, Hamzeh Abdul Kareem, Mustafa Al-Fayoumi and Mohammad Abu Snober

Department of Computer Science, Princess Sumaya University for Technology, Khalil Al-Saket Street, Amman, Jordan

Keywords: Twitter, ML, Detecting Fake Accounts, Spearman's Correlation, PCA, J48, Random Forest, KNN, Naive Bayes.

Abstract: Internet Communities are affluent in Fake Accounts. Fake accounts are used to spread spam, give false reviews for products, publish fake news, and even interfere in political campaigns. In business, fake accounts could do massive damage like waste money, damage reputation, legal problems, and many other things. The number of fake accounts is increasing dramatically by the enormous growth of the online social network; thus, such accounts must be detected. In recent years, researchers have been trying to develop and enhance machine learning (ML) algorithms to detect fake accounts efficiently and effectively. This paper applies four Machine Learning algorithms (J48, Random Forest, Naive Bayes, and KNN) and two reduction techniques (PCA, and Correlation) on a MIB Twitter Dataset. Our results provide a detailed comparison among those algorithms. We prove that combining Correlation along with the Random Forest algorithm gave better results of about 98.6%.

Complete Paper #12

Biomedical Dataset Recommendation

Xu Wang, Frank van Harmelen and Zhisheng Huang

Vrije University Amsterdam, De Boelelaan 1105, 1081 HV Amsterdam, The Netherlands

Keywords: Dataset Recommendation, Scientific Datasets.

Abstract: Dataset search is a special application of information retrieval, which aims to help scientists with finding the datasets they want. Current dataset search engines are query-driven, which implies that the results are limited by the ability of the user to formulate the appropriate query. In this paper we aim to solve this limitation by framing dataset search as a recommendation task: given a dataset by the user, the search engine recommends similar datasets. We solve this dataset recommendation task using a similarity approach. We provide a simple benchmark task to evaluate different approaches for this dataset recommendation task. We also evaluate the recommendation task with several similarity approaches in the biomedical domain. We benchmark 8 different similarity metrics between datasets, including both ontology-based techniques and techniques from machine learning. Our results show that the task of recommending scientific datasets based on meta-data as it occurs in realistic dataset collections is a hard task. None of the ontology-based methods manage to perform well on this task, and are outscored by the majority of the machine-learning methods. Of these ML methods only one of the approaches performs reasonably well, and even then only reaches 70% accuracy.

Complete Paper #59

Tailoring Taint Analysis for Database Applications in the K Framework

Md. Alam[1,2] and Raju Halder[1]

[1] *Indian Institute of Technology Patna, India*

[2] *Università Ca' Foscari Venezia, Italy*

Keywords: Database Applications, Taint Analysis, K Framework, Security.

Abstract: Maintaining the integrity of underlying databases of any information systems is one of the challenges. This could be either due to coding flaws or due to improper flow of information from source to sink in the associated database applications. Compromising this may lead to either disclosure of sensitive information to the attackers or illegitimately modification of private data stored in the databases. Taint analysis is a widely used program analysis technique that aims at averting malicious inputs from corrupting data values in critical computations of programs. In this paper, we propose K-DBTaint, a rewriting logic-based executable semantics for taint analysis of database applications in the K framework. We specify the semantics for a subset of SQL statements along with host imperative program statements. Our K semantics can be seen as a sound approximation of program semantics in the corresponding security type domain. With respect to the existing methods, K-DBTaint supports context- and flow-sensitive analysis, reduces false alarms, and provides a scalable solution. Experimental evaluation on several PL/SQL benchmark codes demonstrates encouraging results as an improvement in the precision of the analysis.

Complete Paper #48

Toward a Multimodal Multitask Model for Neurodegenerative Diseases Diagnosis and Progression Prediction

Sofia Lahrichi, Maryem Rhanoui, Mounia Mikram and Bouchra El Asri

IMS Team, ADMIR Laboratory, Rabat IT Center, ENSIAS, Mohammed V University, Rabat, Morocco

Keywords: Alzheimer's Disease, Multimodal Multitask Learning, Machine Learning, Deep Learning, Progression Detection, Time Series.

Abstract: Recent studies on modelling the progression of Alzheimer's disease use a single modality for their predictions while ignoring the time dimension. However, the nature of patient data is heterogeneous and time dependent which requires models that value these factors in order to achieve a reliable diagnosis, as well as making it possible to track and detect changes in the progression of patients' condition at an early stage. This article overviews various categories of models used for Alzheimer's disease prediction with their respective learning methods, by establishing a comparative study of early prediction and detection Alzheimer's disease progression. Finally, a robust and precise detection model is proposed.

Session 4B **DATA**
09:45 - 11:15 **Room 8**
Business Analytics

Complete Paper #49

A Study on the Effects of Response Time on Travel Package Attributes

Usha Ananthakumar and Sagun Pai
Indian Institute of Technology Bombay, Mumbai, India

Keywords: Consumer Behavior, Conjoint Analysis, Demographic Profiling, Tourism Preferences, Willingness to Pay.

Abstract: The rapid growth of online surveys in the past decade has raised questions about the effects of response time on the results. The focus of our current study is to discuss the impact of response time on various travel package attributes, thereby understanding consumer cognitive process. This study makes use of a recently conducted conjoint analysis experiment on travel package preferences in order to gain insights into the impact of response time on attribute importance and willingness to pay (WTP). Accordingly, the respondents are grouped as fast and slow depending on their response time and their differences in conjoint attribute importance estimates are investigated. The study also examines the changes in consumer willingness to pay for the two groups. Additionally, the distinctions in socioeconomic characteristics between the fast and slow respondents are also analyzed. The results and conclusions obtained from this research will help tour operators to scrutinize the time taken by consumers and thereby deploy appropriate marketing strategy based on the respective importance values and WTP trends.

Complete Paper #54

A Longitudinal Model for Song Popularity Prediction

Ahmet Çimen and Enis Kayış
Department of Industrial Engineering, Ozyegin University, Istanbul, Turkey

Keywords: Music Analytics, Time-varying Coefficients, Mathematical Programming.

Abstract: Usage of new generation music streaming platforms such as Spotify and Apple Music has increased rapidly in the last years. Automatic prediction of a song's popularity is valuable for these firms which in turn translates into higher customer satisfaction. In this study, we develop and compare several statistical models to predict song popularity by using acoustic and artist-related features. We compare results from two countries to understand whether there are any cultural differences for popular songs. To compare the results, we use weekly charts and songs' acoustic features as data sources. In addition to acoustic features, we add acoustic similarity, genre, local popularity, song recentness features into the dataset. We applied Flexible Least Squares (FLS) method to estimate song streams and observe time-varying regression coefficients using a quadratic program. FLS method predicts the number of weekly streams of a song using the acoustic features and the additional features in the dataset while keeping weekly model differences as small as possible. Results show that the significant changes in the regression coefficients may reflect the changes in the music tastes of the countries.

Complete Paper #63

A Company's Corporate Reputation through the Eyes of Employees Measured with Sentiment Analysis of Online Reviews

R. Loke and R. Lam-Lion
Centre for Market Insights, Amsterdam University of Applied Sciences, Amsterdam, The Netherlands

Keywords: Sentiment Analysis, Corporate Reputation, Natural Language Processing, Semantic Search, Scraping.

Abstract: Corporate reputation can be defined as the overall assessment of a company's performance over time (Kircova & Esen, 2018). Organizations with a positive corporate reputation create a competitive advantage and are more likely to influence customer's behaviors and attitudes (Kircova, 2018). Measuring corporate reputation from online data is an increasingly important area in business studies because the amount of opinions and comments is increasingly growing on the internet and has become very accessible to strangers (Shayaa, 2018). Traditionally, corporate reputation is measured with well-known approaches such as surveys, qualitative interviews, and sample groups (Smith, 2010). Researchers like Fombrun, Fonzy and Newburry (2015) developed instruments to measure corporate reputation and predictivily modeled its impact on stakeholder outcomes. So far, however, there has been little attention in the literature on sophisticated measurement techniques for corporate reputation that can be applied to online reviews from the public web. This paper applies sentiment analysis in combination with semantic search as a suitable technique to explore how employees perceive organizations. By using our toolbox, organizations can adapt to market changes and cater to stakeholders' needs. Also, it can be used to raise awareness for organizations that are unaware of negative reviews online.

Poster Session 2 **DATA**
11:15 - 12:15 **Room Posters DATA**

Abstract #19

Automatic Measurement of Corporate Reputation for Retail Companies from Online Public Data on the Web

Marselo Sitorus and Rob Loke
Centre for Market Insights, Amsterdam University of Applied Sciences, Amsterdam, The Netherlands

Keywords: Aspect Based Sentiment Analysis (ABSA), Unsupervised Learning, Retail Industry, Corporate Reputation, Web Scraping, Online Reviews.

Abstract: Retail industry consists of the establishment of selling consumer goods (i.e. technology, pharmaceuticals, food and beverages, apparels and accessories, home improvement etc.) and services (i.e. specialty and movies) to customers through multiple channels of distribution including both the traditional brick-and-mortar and online retailing. Managing corporate reputation of retail companies is crucial as it has many advantages, for instance, it has been proven to impact generated revenues (Wang et al., 2016). But, in order to be able to manage corporate reputation, one has to be able to measure it, or, nowadays even better, listen to relevant social signals that are out there on the public web. One of the most extensive and widely used frameworks for measuring corporate reputation is through conducting elaborated surveys with respective stakeholders (Fombrun et al., 2015). This approach

is valuable but deemed to be laborious and resource-heavy and will not allow to generate automatic alerts and quick and live insights that are extremely needed in this era of internet. For these purposes a social listening approach is needed that can be tailored to online data such as consumer reviews as the main data source. Online review datasets are a form of electronic Word-of-Mouth (WOM) that, when a data source is picked that is relevant to retail, commonly contain relevant information about customers' perceptions regarding products (Pookulangara, 2011) and that are massively available.

The algorithm that we have built in our application provides retailers with reputation scores for all variables that are deemed to be relevant to retail in the model of Fombrun et al. (2015). Examples of such variables for products and services are high quality, good value, stands behind, and meets customer needs. We propose a new set of subvariables with which these variables can be operationalized for retail in particular. Scores are being calculated using proportions of positive opinion pairs such as <fast, delivery> or <rude, staff> that have been designed per variable. With these important insights extracted, companies can act accordingly and proceed to improve their corporate reputation. It is important to emphasize that, once the design is complete and implemented, all processing can be performed completely automatic and unsupervised.

The application makes use of a state of the art aspect-based sentiment analysis (ABSA) framework because of ABSA's ability to generate sentiment scores for all relevant variables and aspects. Since most online data is in open form and we deliberately want to avoid labelling any data by human experts, the unsupervised aspectator algorithm has been picked. It employs a lexicon to calculate sentiment scores and uses syntactic dependency paths to discover candidate aspects (Bancken et al., 2014).

We have applied our approach to a large number of online review datasets that we sampled from a list of 50 top global retailers according to National Retail Federation (2020), including both offline and online operation, and that we scraped from trustpilot, a public website that is well-known to retailers.

The algorithm has carefully been evaluated by manually annotating a randomly sampled subset of the datasets for validation purposes by two independent annotators. The Kappa's score on this subset was 80%.

Complete Paper #28

Predicting Shopping Intent of e-Commerce Users using LSTM Recurrent Neural Networks

Konstantinos Diamantaras, Michail Salampasis, Alkiviadis Katsalis and Konstantinos Christantonis

Intelligent Systems Laboratory, Department of Information and Electronic Engineering, International Hellenic University, Sindos, Thessaloniki, Greece

Keywords: Purchase Intent, e-Commerce, LSTM-RNN, Web Usage Mining.

Abstract: An e-commerce web site is effective if it turns visitors into buyers achieving a high conversion rate. To this realm, it is useful to predict each user's purchase intent and understand their navigation behavior. Such predictions may be utilized to improve web design and to personalize shopper's experience, hopefully leading to increased conversion rates. Additionally, if such predictions can be done in real-time, during the ongoing navigation of an e-commerce user, the e-commerce application can take proactive stimuli actions to offer incentives with a view to increase the probability that a user will finally make a purchase. This paper presents a method for predicting in real-time the shopping intent of e-commerce users using LSTM recurrent neural networks. We test several variants of our method in a dataset created from the processing of Web server logs of an industry e-commerce web application, dividing user sessions in three different classes: browsing, cart abandonment, purchase. The best classifier achieves a predictive accuracy of almost 98%. This result is competitive with other state-of-the-art methods, which affirms that accurate and scalable purchasing intention prediction for e-commerce, using only session-based data, is feasible without any intense feature engineering.

Complete Paper #41

Applied Feature-oriented Project Life Cycle Classification

Oliver Böhme and Tobias Meisen

Chair for Technologies and Management of Digital Transformation, Bergische Universität Wuppertal, Rainer-Gruenter-Str. 21, Wuppertal, Germany

Keywords: Machine Learning, Classification, Prediction, Deep Neural Networks, MLP, LSTM, Multivariate, Automotive, R&D, Projects Progressions, Project Life Cycle, Comparative Analysis.

Abstract: The increasing complexity in automotive product development is forcing traditional manufacturers to fundamentally rethink. As a result, many companies are already investing in the development of methods to increase the controllability of their development processes. The use of data-driven approaches is a promising way to provide an early prediction of potential problems in the course of a project by learning from the past. In vehicle development, projects can be divided into two basic categories: new vehicle launches and model enhancement projects. The course of projects according to the above-mentioned categories can be based on different influencing factors. To verify this hypothesis and to determine the extent of the differences in the data, we carry out a data-driven classification of the project category. In contrast to the recognition of other time-dependent data (e.g., univariate sensor data courses), we use multivariate project information from the automotive industry. With this paper, which is of an application nature, we prove that a multivariate classification of automotive projects can be realized based on the underlying project's progression.

Complete Paper #45

Impact of Duplicating Small Training Data on GANs

Yuki Eizuka[1], Kazuo Hara[1] and Ikumi Suzuki[2]

[1] *Yamagata University, 1-4-12 Kojirakawa-machi, Yamagata City, 990-8560, Japan*

[2] *Nagasaki University, 1-14 Bunkyo, Nagasaki City, 852-8521, Japan*

Keywords: Generative Adversarial Networks, Small Training Data, Emoticons.

Abstract: Emoticons such as (^_^) are face-shaped symbol sequences that are used to express emotions in text. However, the number of emoticons is miniscule. To increase the number of emoticons, we created emoticons using SeqGANs, which are generative adversarial networks for generating sequences. However, the small number of emoticons means that few emoticons can be used as training data for SeqGANs. This is concerning because as SeqGANs underfit small training data, generating emoticons using SeqGANs is difficult. To address this problem, we duplicate the training data. We observed that emoticons can be generated when the duplication magnification is of an appropriate value. However, as a trade-off, it was also observed that SeqGANs overfit the training data, i.e., they produce emoticons that are exactly the same as the training data.

Complete Paper #46

Making Data Big for a Deep-learning Analysis: Aggregation of Public COVID-19 Datasets of Lung Computed Tomography Scans

Francesca Lizzi[1,2], Francesca Brero[3,4], Raffaella Cabini[3,5], Maria Fantacci[2,6], Stefano Piffer[7,8], Ian Postuma[3], Lisa Rinaldi[3,4] and Alessandra Retico[2]

[1] *Scuola Normale Superiore, Pisa, Italy*

[2] *National Institute for Nuclear Physics (INFN), Pisa Division, Pisa, Italy*

[3] *INFN, Pavia Division, Pavia, Italy*

[4] *Department of Physics, University of Pavia, Pavia, Italy*

[5] *Department of Mathematics, University of Pavia, Pavia, Italy*

[6] *Department of Physics, University of Pisa, Pisa, Italy*

[7] *Department of Biomedical Experimental Clinical Science "M. Serio", University of Florence, Florence, Italy*

[8] *INFN, Florence Division, Florence, Italy*

Keywords: COVID-19, Lung CT, U-net, Data Aggregation, Image Segmentation.

Abstract: Lung Computed Tomography (CT) is an imaging technique useful to assess the severity of COVID-19 infection in symptomatic patients and to monitor its evolution over time. Lung CT can be analysed with the support of deep learning methods for both aforementioned tasks. We have developed a U-net based algorithm to segment the COVID-19 lesions. Unfortunately, public datasets populated with a huge amount of labelled CT scans of patients affected by COVID-19 are not available. In this work, we first review all the currently available public datasets of COVID-19 CT scans, presenting an extensive description of their characteristics. Then, we describe the design of the U-net we developed for the automated identification of COVID-19 lung lesions. Finally, we discuss the results obtained by using the different publicly available datasets. In particular, we trained the U-net on the dataset made available within the COVID-19 Lung CT Lesion Segmentation Challenge 2020, and we tested it on data from the MosMed and the COVID-19-CT-Seg datasets to explore the transferability of the model and to assess whether the image annotation process affects the detection performances. We evaluated the performance of the system in lesion segmentation in terms of the Dice index, which measures the overlap between the ground truth and the predicted masks. The proposed U-net segmentation model reaches a Dice index equal to 0.67, 0.42 and 0.58 on the independent validation sets of the COVID-19 Lung CT Lesion Segmentation Challenge 2020, on the MosMed and on the COVID-19-CT-Seg datasets, respectively. This work focusing on lesion segmentation constitutes a preliminary work for a more accurate analysis of COVID-19 lesions, based for example on the extraction and analysis of radiomic features.

Complete Paper #52

Knowledge Graph Analysis of Russian Trolls

Chih-yuan Li[1], Soon Chun[2] and James Geller[1]

[1] *Department of Computer Science, New Jersey Institute of Technology, Newark, NJ 07102, U.S.A.*

[2] *City University of New York, College of Staten Island, New York City, NY 10314, U.S.A.*

Keywords: Relationship Analysis of Troll Tweets, Entity Extraction, Triple Extraction, Sentiment Analysis.

Abstract: Social media, such as Twitter, have been exploited by trolls to manipulate political discourse and spread disinformation during the 2016 US Presidential Election. Trolls are users of social media accounts created with intentions to influence the public opinion by posting or reposting messages containing misleading or inflammatory information with malicious intentions. There has been previous research that focused on troll detection using Machine Learning approaches, and troll understanding using visualizations, such as word clouds. In this paper, we focus on the content analysis of troll tweets to identify the major entities mentioned and the relationships among these entities, to understand the events and statements mentioned in Russian Troll tweets coming from the Internet Research Agency (IRA), a troll factory allegedly financed by the Russian government. We applied several NLP techniques to develop Knowledge Graphs to understand the relationships of entities, often mentioned by dispersed trolls, and thus hard to uncover. This integrated KG helped to understand the substance of Russian Trolls' influence in the election. We identified three clusters of troll tweet content: one consisted of information supporting Donald Trump, the second for exposing and attacking Hillary Clinton and her family, and the third for spreading other inflammatory content. We present the observed sentiment polarization using sentiment analysis for each cluster and derive the concern index for each cluster, which shows a measurable difference between the presidential candidates that seems to have been reflected in the election results.

Complete Paper #53

Aspect Based Sentiment Analysis on Online Review Data to Predict Corporate Reputation

R. Loke and W. Reitter

Centre for Market Insights, Amsterdam University of Applied Sciences, Amsterdam, The Netherlands

Keywords: Aspect Based Sentiment Analysis (ABSA), Machine Learning, Natural Language Processing, Scraping.

Abstract: Corporate reputation is an intangible resource that is closely tied to an organization's success but measuring it and to derive actions that can improve the reputations can be a long and expensive journey for an organization. In the available literature, corporate reputation is primarily measured through surveys, which can be time and cost intensive. This paper uses online reviews on the web as the source for a machine-learning driven aspect-based sentiment analysis that can enable organizations to evaluate their corporate reputation on a fine-grained level. The analysis is done unsupervised without organizations needing to manually label datasets. Using the insights generated through the analysis, on one hand, organizations can save costs and time to measure corporate reputation, and, on the other hand, it provides an in-depth analysis that splits the overall reputation into multiple aspects, with which organizations can identify weaknesses and in turn improve their corporate reputation. Therefore, this research is relevant for organizations aiming to understand and improve their corporate reputation to achieve success, for example, in form of financial performance, or for organizations that help and consult other organizations on their journeys to increased success. Our approach is validated, evaluated and illustrated with Trustpilot review data.

Complete Paper #57

Evo-Path: Querying Data Evolution through Complex Changes

Theodora Galani[1], Yannis Stavrakas[2], George Papastefanatos[2] and Yannis Vassiliou[1]

[1] *School of Electrical and Computer Engineering, NTUA, Iroon Polytechniou 9, Athens, Greece*

[2] *RC ATHENA, Artemidos 6 & Epidavrou, Marousi, Greece*

Keywords: Querying Data Evolution, Change Modelling, XPath.

Abstract: Evo-graph is a model for data evolution that captures data versions and treats changes as first-class citizens. A change in evo-graph can be compound, comprising disparate changes, and is associated with the data items it affects. In previous work, we specified how an evo-graph can be reduced to a snapshot holding under a specific time instance, we presented an XML representation of evo-graph called evoXML, we defined how evo-graph is constructed as the current snapshot evolves, as well as presented and evaluated the C2D framework that implements these concepts using XML technologies. In this paper, we formally define evo-path, an XPath extension for querying the data history and change structure in a uniform way over evo-graph. We specify the evo-path syntax, semantics and implementation, and present several query categories.

Complete Paper #58

Enhanced AI On-the-Edge 3D Vision Accelerated Point Cloud Spatial Computing Solution

Gaurav Kumar Wankar and Shubham Vohra

Neal Analytics Services Private Limited, Pune, Maharashtra, 411014, India

Keywords: Industry - 4.0, Industry - 5.0, Digital Transformation, Business Transformation, Disruptive Technologies, 3D Vision, Product Market, Immersive Business Solutions, Artificial Intelligence, Deep Learning, Voxelization, PointNet, Pointpillars, Point Cloud, NVIDIA Jetson Tx2, Azure Kinect DK, GPU, Edge Applications, Edge AI, AI On-the-Edge, Transformative Experiences, Smart Everything Revolution.

Abstract: With the emergence of Industry - 5.0, the 3D Vision Product Market is growing rapidly. Leveraging Disruptive Technologies, we are exploring Artificial Intelligence driven Advanced 3D Vision immersive Business Solutions with transformative experiences leveraging Deep Learning accelerated with Voxelization, PointPillars and PointNet approaches for classification of Point Clouds enhancing the feature extraction to be more accurate bringing our work and data to life. NVIDIA Jetson Tx2 targeted at power constrained AI on-the-edge applications maintains awareness of its surroundings by visualizing in 3D space leveraging Azure Kinect DK depth sensing instead of 2D space thereby improving the performance in Edge AI computing device. Leveraging state of the art technologies converging AI and Mixed Reality we further encourage the readers to explore the possibilities of Next Generation services bringing accurate and immersive real-world information allowing decision-making based on Digital Reality driving Digital Transformation.

Keynote Lecture **DATA**
12:15 - 13:15 **Room Plenary 1**

A Profile-aware Methodological Framework for Collaborative Multidimensional Modeling: Agro-biodiversity Case Study

Sandro Bimonte

INRAE, France

Abstract: Multidimensional modeling, i.e., the design of cube schemata, has a key role in data warehouse (DW) projects, in self-service business intelligence, and in general to let users analyze data via the OLAP paradigm. Though an effective involvement of users in multidimensional modeling is crucial in these projects, not much has been said about how to establish a fruitful collaboration in projects involving numerous users with different skills, reputations, and degrees of authority. This issue is especially relevant in citizen science projects, where several volunteers can contribute their requirements despite not being formally-trained experts in the application domain. To fill this gap, we propose a framework for collaborative multidimensional modeling that can adapt itself to the pro-files and skills of the actors involved. We first classify users depending on their authoritativeness, skills, and engagement in the project. Then, following this classification, we identify four possible methodological scenarios and propose a profile-aware methodology supported by two sets of quality attributes. Finally, we describe a Group Decision Support System that implements our methodological framework and present some experiments carried out on a real case study.

Closing Session **DATA**
13:15 - 13:30 **Room Plenary 1**

Author Index

Abdul Kareem, H. 33
Abu Snober, M. 33
Al Nemri, J. 33
Al-Fayoumi, M. 33
Alam, M. 33
Almasri, M. 19
Alsaedi, A. 23
Alshehri, M. 27
Altendeitering, M. 28
Ananthakumar, U. 34
Arora, M. 23
Auch, M. 23

Balluff, M. 23
Banik, S. 29
Barbella, M. 25
Bayer, J. 26
Belo, O. 28
Böhme, O. 35
Bond, W. 20
Brero, F. 36
Brisaboa, N. 29
Brooker, P. 23

Cabini, R. 36
Chatziloizos, G. 19
Christantonis, K. 35
Chua-Chow, C. 20
Chun, S. 36
Churchill, R. 19
Çimen, A. 34
Coenen, F. 27

Diamantaras, K. 35
Dozier, H. 20
Dures, K. 27

Ehrlinger, L. 28
Einsele, F. 27
Eizuka, Y. 35
El Asri, B. 33
Esposito, C. 23

Fantacci, M. 36
Fariña, A. 29
Farkas, C. 29
Fatone, G. 23

Galani, T. 37
Gan, C. 20
Geierhos, M. 24
Geller, J. 36
Gindlhumer, A. 28
Grasso, F. 23
Gunopulos, D. 19

Halder, R. 33
Hansen, B. 20
Hara, K. 27, 35
Homsi, A. 33
Howar, F. 28
Huang, Z. 33
Huber, L. 28

Jenzer, H. 27
Jodlbauer, H. 23
Johannesson, P. 28

Kalenov, N. 26
Kalyani, Y. 23
Katsalis, A. 35
Kayış, E. 34
Kersting, J. 24
Kikuchi, Y. 27
Konstantinou, K. 19

Lahrichi, S. 33
Lam-Lion, R. 34
Li, C. 36
Lim, S. 28
Lizzi, F. 36
Loke, R. 34, 36
Lustenberger, T. 27

Mandl, P. 23
Masciari, E. 23
Meisen, T. 35
Mikram, M. 33

Naimat, N. 33
Nobani, N. 25

O'Riordan, C. 24
Oliveira, B. 28
Oliveira, Ó. 28
Otero-González, D. 29
Otto, B. 27
Öztürk, S. 29

Pai, S. 34
Papastefanatos, G. 37
Pelucchi, M. 25
Perico, M. 25
Pettenpohl, H. 27
Piffer, S. 36
Postuma, I. 36
Rahmani, R. 28
Ravan, J. 29
Reitter, W. 36
Retico, A. 36
Rhanoui, M. 33
Rhujittawiwat, T. 29
Rinaldi, L. 36
Risi, M. 25
Rodeiro, T. 29

Saaudi, A. 29
Salampasis, M. 35
Sarwar, G. 24
Schrefl, M. 24
Schuetz, C. 24
Scrivanti, A. 25
Shanker, S. 23
Shukla, I. 20
Silas, A. 20
Singh, L. 19
Sitorus, M. 34
Sobolevskaya, I. 26
Sotnikov, A. 26
Staudinger, S. 24
Stavrakas, Y. 37
Strasser, S. 23
Suzuki, I. 27, 35
Syed, R. 26

Tamburis, O. 23
Tebernum, D. 27, 28
Thoma, F. 26
Thomason, S. 23
Tortora, G. 25
Tripathi, S. 23

Ullrich, T. 20

Vaccarino, A. 25
van Harmelen, F. 33
Vassiliou, Y. 37
Vohra, S. 37

Wang, X. 33
Wankar, G. 37
Wöß, W. 28
Wolff, C. 23

Xie, S. 20

Yang, L. 26

Zhou, S. 25

Final Program and Book of Abstracts of DATA 2021

10th International Conference on Data Science, Technology and Applications

http://www.dataconference.org

LOGISTICS PARTNERS:

EVENT MANAGEMENT SYSTEM:

PROCEEDINGS WILL BE SUBMITTED FOR INDEXATION BY: