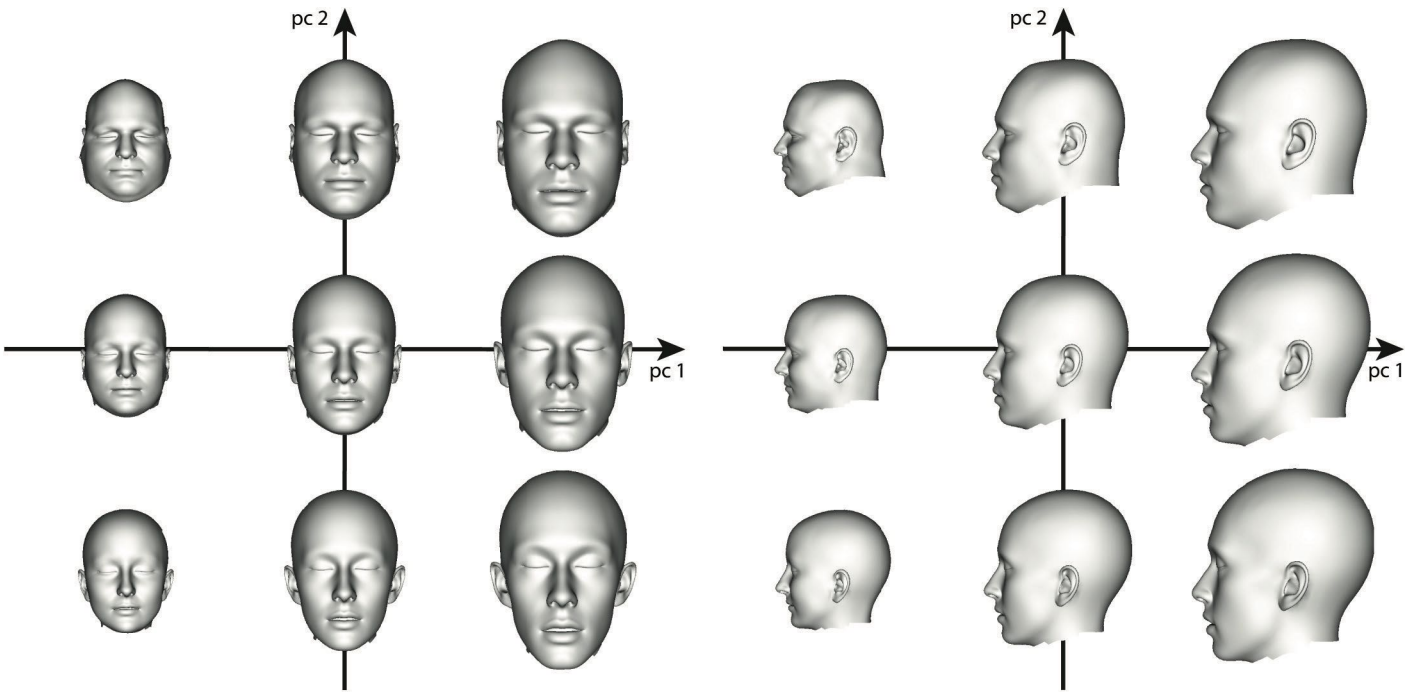
Organ der Gesellschaft für Informatik e.V.
und mit ihr assoziierter Organisationen

Band 42 • Heft 5 • Oktober 2019

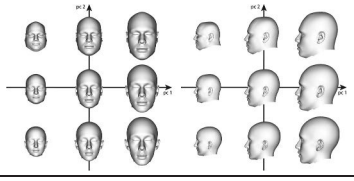
Informatik Spektrum

Privacy in Process Analytics



Springer

Informatik Spektrum
Band 42 | Heft 5 | Oktober 2019
Organ der Gesellschaft für Informatik e.V. und mit ihr assoziierter Organisationen



322

Automatisierte Rekonstruktion von Gesichtern

Springer

Informatik Spektrum

Organ der Gesellschaft für Informatik e.V. und mit ihr assoziierter Organisationen

Hauptaufgabe dieser Zeitschrift ist die Weiterbildung aller Informatiker durch Veröffentlichung aktueller, praktisch verwertbarer Informationen über technische und wissenschaftliche Fortschritte aus allen Bereichen der Informatik und ihrer Anwendungen. Dies soll erreicht werden durch Veröffentlichung von Übersichtsartikeln und einführenden Darstellungen sowie Berichten über Projekte und Fallstudien, die zukünftige Trends aufzeigen.

Es sollen damit unter anderem jene Leser angesprochen werden, die sich in neue Sachgebiete der Informatik einarbeiten, sich weiterbilden, sich einen Überblick verschaffen wollen, denen aber das Studium der Originalliteratur zu zeitraubend oder die Beschaffung solcher Veröffentlichungen nicht möglich ist. Damit kommt als Leser nicht nur der ausgebildete Informatikspezialist in Betracht, sondern vor allem der Praktiker, der aus seiner Tagesarbeit heraus Anschluss an die wissenschaftliche Entwicklung der Informatik sucht, aber auch der Studierende an einer Fachhochschule oder Universität, der sich Einblick in Aufgaben und Probleme der Praxis verschaffen möchte.

Durch Auswahl der Autoren und der Themen sowie durch Einflussnahme auf Inhalt und Darstellung – die Beiträge werden von mehreren Herausgebern referiert – soll erreicht werden, dass möglichst jeder Beitrag dem größten Teil der Leser verständlich und lesenswert erscheint. So soll diese Zeitschrift das gesamte Spektrum der Informatik umfassen, aber nicht in getrennte Sparten mit verschiedenen Leserkreisen zerfallen. Da die Informatik eine sich auch weiterhin stark entwickelnde anwendungsorientierte Wissenschaft ist, die ihre eigenen wissenschaftlichen und theoretischen Grundlagen zu einem großen Teil selbst entwickeln muss, will die Zeitschrift sich an den Problemen der Praxis orientieren, ohne die Aufgabe zu vergessen, ein solides wissenschaftliches Fundament zu erarbeiten. Zur Anwendungsorientierung gehört auch die Beschäftigung mit den Problemen der Auswirkung der Informatikanwendungen auf den Einzelnen, den Staat und die Gesellschaft sowie mit Fragen der Informatik-Berufe einschließlich der Ausbildungsrichtlinien und der Bedarfsschätzungen.

Vertrieb, Abonnement, Versand

Papierausgabe: ISSN 0170-6012
elektronische Ausgabe: ISSN 1432-122X
Erscheinungsweise: zweimonatlich

Den Bezugspreis können Sie beim Customer Service erfragen: customerservice@springernature.com. Die Lieferung der Zeitschrift läuft weiter, wenn sie nicht bis zum 30.9. eines Jahres abbestellt wird. Mitglieder der Gesellschaft für Informatik und der Schweizer Informatiker Gesellschaft erhalten die Zeitschrift im Rahmen ihrer Mitgliedschaft.

Bestellungen oder Rückfragen nimmt jede Buchhandlung oder der Verlag entgegen.
SpringerNature, Kundenservice Zeitschriften,
Tiergartenstr. 15, 69121 Heidelberg, Germany
Tel. +49-6221-345-0, Fax: +49-6221-345-4229,
e-mail: customerservice@springernature.com
Geschäftszeiten: Montag bis Freitag 8–20 h.

Bei **Adressänderungen** muss neben dem Titel der Zeitschrift die neue und die alte Adresse angegeben werden. Adressänderungen sollten mindestens 6 Wochen vor Gültigkeit gemeldet werden. **Hinweis gemäß §4 Abs. 3 der Postdienst-Datenschutzverordnung:** Bei Anschriftenänderung des Beziehers kann die Deutsche Post AG dem Verlag die neue Anschrift auch dann mitteilen, wenn kein Nachsendeauftrag gestellt ist. Hiergegen kann der Bezieher innerhalb von 14 Tagen nach Erscheinen dieses Heftes bei unserer Abonnementsbetreuung widersprechen.

Elektronische Version

springerlink.com

Hinweise für Autoren

http://springer.com/journal/00287

Impressum

Verlag:
Springer, Tiergartenstraße 17,
69121 Heidelberg

Redaktion:
Peter Pagel, Vanessa Keinert
Tel.: +49 611 787 8329
e-mail: Peter.Pagel@springer.com

Herstellung:
Philipp Kammerer,
e-mail: Philipp.Kammerer@springer.com

Redaktion GI-Mitteilungen:
Cornelia Winter
Gesellschaft für Informatik e.V. (GI)
Wissenschaftszentrum,
Ahrstraße 45, D-53175 Bonn,
Tel.: +49 228-302-145, Fax: +49 228-302-167,
Internet: http://www. gi.de,
e-mail: gs@gi.de

Wissenschaftliche Kommunikation:
Anzeigen: Eva Hanenberg
Abraham-Lincoln-Straße 46
65189 Wiesbaden
Tel.: +49 (0)611/78 78-226
Fax: +49 (0)611/78 78-430
eva.hanenberg@springer.com

Satz:
le-tex publishing services GmbH, Leipzig

Druck:
Printforce,
The Netherlands

springer.com

Hannes Federrath, Präsident der GI, Universität Hamburg

Privacy in Process Analytics

Wer große Konzerthäuser, noch größere Flughäfen oder einfach nur komplexe Software bauen soll, muss die Prozesse verstehen und beherrschen, die dort später ablaufen werden. Ich kann mich gut erinnern, wie in der Stadt, in der ich einst wohnte, ein nagelneues Kaufhaus für die Kaufhauskette A gebaut wurde. Während der Bauzeit übernahm Kaufhauskette B die Kaufhauskette A und entschied sich, das Kaufhaus künftig unter dem eigenen Namen B betreiben zu wollen. Hierfür wurde das halbe noch im Rohbau befindliche Kaufhaus wieder abgerissen und noch einmal neu gebaut. „Weil die Logistikprozesse sonst nicht funktionieren", war die Begründung.

Software ist manchmal auch wie einstürzende Neubauten. Wäre es nicht gut, wenn wir schon vorher alles wüssten? – Zumindest wäre es doch schön, wenn wir aus den Erfahrungen vergangener und aktueller Software-Zeiten lernen könnten und aktuelle oder künftige Systeme prozessoptimiert bauen oder verbessern könnten. Der Schlüssel hierzu ist die systematische Analyse von prozessrelevanten Ereignisdaten. Verglichen mit dem Versuch einer formalen Beschreibung kann Process Analytics zumindest im operativen Bereich sogar bessere oder wenigstens pragmatische Ergebnisse liefern.

Die Gesellschaft für Informatik wird nicht müde darauf hinzuweisen, dass der Mensch im Mittelpunkt der Gestaltung von Prozessen, Systemen und Software stehen soll. Menschen produzieren prozessrelevante Ereignisdaten und sie haben Rechte. Nach Artikel 8 der EU-Grundrechtecharta hat jede Person das Recht auf Schutz der sie betreffenden personenbezogenen Daten. Diese Daten dürfen nur nach Treu und Glauben für festgelegte Zwecke und mit Einwilligung der betroffenen Person oder auf einer sonstigen gesetzlich geregelten legitimen Grundlage verarbeitet werden. Dieses Grundrecht auf Privatheit haben auch die Prozessanalysten zu respektieren – und genau darum geht es in diesem Themenheft.

Den Anstoß für dieses Themenheft haben einige Juniorfellows der GI gegeben. Agnes Koschmider, Martin Degeling, Matthias Weidlich haben in ihrer Funktion als Gastherausgeber dankenswerterweise den von den Juniorfellows zugespielten Ball angenommen. Allerdings ist dieses Themenheft etwas anders geworden als üblich: Mit Kurzbeiträgen, Interviews und Extended Abstracts möchten die Gastherausgeber das Thema in seiner Breite vorstellen sowie zu Diskussionen und zum Nachdenken anregen.

Thorsten Strufe, Karin Schuler und Wil van der Aalst erzählen in drei Interviews, warum ihnen Privacy wichtig ist und wo sie die Herausforderungen beim datenschutzfreundlichen Process Mining sehen. In mehreren Extended Abstracts wird das Thema Privacy in Process Analytics aus unterschiedlichen Perspektiven beleuchtet. Wie Sie schon erahnen können, spielen beim Thema Process Analytics selbstverständlich auch Techniken der künstlichen Intelligenz eine Rolle. Der Beitrag zum „Privacy-Preserving Federated Learning" schlägt somit eine Brücke zu der derzeit sehr aktuellen und spannenden Problematik der datenschutzfreundlichen Datenakquise und Datennutzung zum Anlernen von KI-Systemen.

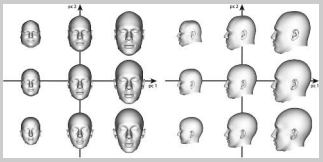
Man muss nicht gleich dystopische Phantasien bemühen, um die Problematik des respektlosen Umfangs mit unseren persönlichen Daten anzuprangern. Wenn etwa in Marc-Uwe Klings „QualityLand" die Lieferdrohne von TheShop, dem weltweit beliebtesten Versandhändler, Produkte bringt, noch bevor man sie bestellt hat, ja, noch bevor man weiß, dass man sie brauchen wird, dann war vielleicht Process Analytics (mehr oder weniger) erfolgreich, aber die Privacy ist tot. Übrigens, die Kaufhausketten A und B gibt es noch, aber die sind zwischenzeitlich längst wieder in neuer Eigentümerschaft.

Hannes Federrath
Präsident der GI

https://doi.org/10.1007/s00287-019-01217-7

i

Zum Titelbild



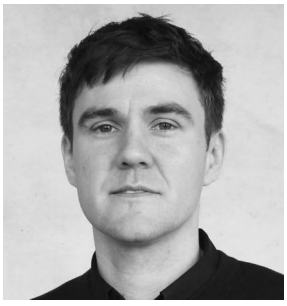
Automatisierte Rekonstruktion von Gesichtern

Die Grafik stammt aus dem Paper:
A method for automatic forensic facial reconstruction based on dense statistics of soft tissue thickness
Thomas Gietzen, Robert Brylka, Jascha Achenbach, Katja zum Hebel, Elmar Schömer, Mario Botsch, Ulrich Schwanecke, Ralf Schulze
PLoS ONE, 14(1), 2019.
https://www.researchgate.net/publication/330569236_A_method_for_automatic_forensic_facial_reconstruction_based_on_dense_statistics_of_soft_tissue_thickness

In dem Paper wird eine Methode zur automatisierten Rekonstruktion eines menschlichen Gesichts auf Basis eines gegebenen Schädels vorgestellt. Die Methode basiert auf drei statistischen Modellen. Ein volumetrisches Schädelmodell, das die Variationen verschiedener Schädel kodiert, ein Oberflächenkopfmodell, das die Kopfvariationen kodiert, und einer dichten Statistik der Weichgewebeverteilung des Gesichts, welche die Verbindung zwischen Schädel und Kopfoberfläche herstellt. Diese Modelle sind jeweils Hauptkomponenten-Modelle (Principal Component Modelle), was es erlaubt Variationen in den Schädeln bzw. Köpfen mit wenigen Parametern zu beschreiben. Die Abbildung zeigt unterschiedlichen Köpfe, die entstehen, wenn die Gewichtung der ersten beiden Hauptkomponenten des Kopfmodells geändert wird (pc1 und pc2 stehen für Principal Component 1 und 2). Die Arbeit ist eine Kooperationsarbeit zwischen der Arbeitsgruppe von Prof. Dr. Ulrich Schwanecke (cvmr.info) an der Hochschule RheinMain und Arbeitsgruppen von Prof. Dr. Elmar Schömer an der Johannes Gutenberg-Universität Mainz, Prof. Dr. Ralf Schulze an der Universitätsmedizin Mainz und Prof. Dr. Mario Botsch an der Universität Bielefeld.

Agnes Koschmider
Christian-Albrechts-Universität zu Kiel
ak@informatik.uni-kiel.de

Martin Degeling
Ruhr-Universität Bochum



Matthias Weidlich
Humboldt-Universität zu Berlin

Process Analytics over IoT-based Event Streams with Privacy Guarantees

Every day large amounts of event streams are created in the Internet of Things (IoT). A lot of these events include or refer to personal data, which need to be protected. Awareness of privacy issues in IoT-based event streams has risen, particularly since the General Data Protection Regulation (GDPR) was put into force. Among other things, it requires organizations to consider privacy throughout the whole data lifecycle, from collection and processing to deletion. Systems that exploit IoT-based event streams, however, require fundamentally new practices for privacy-by-design or privacy-by-architecture development that are different from those for conventional information systems in terms of user-centered access control and identity management. In the past, the focus of privacy considerations in system design was often on the protection of data that directly relate to a person, i.e. salary, email address, or customer number. With IoT-based event streams, privacy-aware system design faces plenty of new challenges. Such streams continuously produce an infinite number of heterogonous events that are highly dependent on each other and can occur concurrently, alternatively, or independently. Stream processing systems aim to identify meaningful knowledge and process-related information from event streams in real-time, thereby creating threats to personal privacy. This calls for techniques for data analytics over IoT-based event streams with provable data protection through privacy guarantees.

Analytics in terms of process mining enables valuable insights into the compliance and performance of process execution using IoT-based event streams. However, process mining over IoT event data can also uncover plenty of personal activities and behavior patterns that require protection. Even process mining over encrypted event data might uncover sensitive information. Moreover, privacy requirements have to be chosen in relation to the threat the data pose and are, therefore, situation and application specific. Eventually, it will, therefore, be necessary to provide models and methods to balance privacy requirements and analysis capabilities based on IoT event data.

This special issue sets out to discuss concerns related to privacy in process analytics over IoT-based event streams. Three interviews with leading experts illustrate the challenges in the field. The reports Process Mining and Privacy in Smart Manufacturing, Process Mining for Learning User Privacy in D2D and IoT Networks, *and* Post-Quantum Cryptography and its Application to the IoT *provide an overview of how some of the challenges encountered could be approached in industry or research settings. Several extended abstracts then outline recent scientific advances on privacy-awareness in process analytics.*

We hope that you will enjoy reading this special issue, and that it will stimulate further work in the field,

Agnes Koschmider, Martin Degeling, Matthias Weidlich

https://doi.org/10.1007/s00287-019-01214-w

Interview mit Thorsten Strufe zu Herausforderungen von Datenschutz und Datensicherheit für das Internet der Dinge

Agnes Koschmider

Herr Strufe, an der TU Dresden haben Sie den Lehrstuhl für Datenschutz und Datensicherheit. Was sind Ihre Forschungsschwerpunkte?

Derzeit arbeiten wir parallel an drei Bereichen. Der erste Bereich ist die Entwicklung technischer Methoden, Verfahren und Protokolle, welche die Ziele des Datenschutzes in und durch verteilte Systeme ermöglichen (Privacy Enhancing Technologies [PET]). Hier entwickeln wir Verfahren für die Netzwerkanonymisierung(AN.ON), Analysen, um diese formal zu untersuchen, und Darknets (Freenet, I2P) sowie Methoden zur Reduzierung von Inferenzangriffen.

Parallel forschen wir auch an Verfahren für die Netzwerksicherheit, um Netze und Applikationen für das Edge/Cloud-Computing zu sichern.

Der dritte Bereich sind datenschutzfreundliche Algorithmen und Verfahren für Social Media. Hier untersuchen wir, welche Gefahren von welcher Nutzung ausgehen und welche Gefahren für welche Daten bestehen. Wir möchten Methoden und Verfahren bereitstellen, die die Privatsphäre des Benutzers schützen und neue Systeme, bei denen sie nicht beeinträchtigt wird. Dabei untersuchen wir am Rande auch User Interfaces von Systemen und deren Auswirkungen auf den Datenschutz. Beispielsweise untersuchen wir, wie Systeme überhaupt bedient werden und wie Fehlbedingungen durch Benutzer erkannt werden können. So untersuchen wir in einem interdisziplinären Verbund, wie Datenschutz und Nudging zusammenpassen. Die Analysen helfen uns bei der Entwicklung von beispielsweise Algorithmen zur Erkennung von Social Bots.

Häufig werden die Begriffe Datenschutz und Datensicherheit vermischt. Wie grenzen Sie die beiden Begriffe ab?

Ganz klassisch kann man die beiden Bereiche wie folgt abgrenzen. Datensicherheit beschäftigt sich mit den Daten selbst und den datenverarbeitenden Systemen, um Schutz vor Fremdzugriff, Veränderung oder Sabotage zu gewährleisten. Hierzu werden kryptografische oder rechtliche Ansätze verwendet. Beim Datenschutz hingegen gehen wir davon aus, dass wir Menschen davor schützen müssen, dass Daten gegen sie verwendet werden. Auch bedarf es sowohl technischer als auch juristischer Mechanismen, sodass die Sammlung dieser Daten nicht zum Nachteil der Bürger gereichen kann.

Meine Wahrnehmung ist, dass in Deutschland im Bereich Privacy überwiegend zu Usable Privacy geforscht wird und weniger zu Privacy Engineering wie beispielsweise Differential Privacy. Wie ist Ihre Wahrnehmung?

Grundsätzlich deckt sich bei mir auch dieser Eindruck. Wenn man aber aus der Sicht der Netzwerksicherheit (z. B. der Anonymisierung von Systemen) schaut, dann findet man in Deutschland Forschung zur anonymen Kommunikation, etwa unsere Weiterentwicklung von AN.ON gemeinsam mit Hamburg und Regensburg). Im Bereich

https://doi.org/10.1007/s00287-019-01210-0

Agnes Koschmider
Kiel, Deutschland
E-Mail: ak@informatik.uni-kiel.de

der Kryptografie existieren zahlreiche Arbeiten, die sich beispielsweise mit Protokollen für Private Information Retrieval oder Privacy Preserving Machine Learning beschäftigten (z. B. in Darmstadt). Es gibt natürlich außerhalb der Informatik auch Forschung, die sich aus juristischer Sicht intensiv mit dem Thema beschäftigt, hier wird insbesondere der gesellschaftliche Wert von Privatheit untersucht, z. B. natürlich in Kassel, aber auch in Freiburg und Karlsruhe. Inzwischen haben wir aber gelernt, dass wir die Benutzer einbinden müssen, weswegen es natürlich einen großen Fokus auf Usable Privacy gibt. Aktuell sind wir an einem Punkt angekommen, an dem wir feststellen, dass die Technik nicht ohne die Juristen auskommt und die Juristen nicht ohne die Technik. Man kann noch so sicherere Systeme bauen, wenn die Benutzer sie nicht haben wollen, sondern sich den bunteren Angeboten großer Firmen zuwenden, die tief, vielleicht auch illegal in die Privatsphäre eingreifen, dann hilft dieser Schutz natürlich nichts. Deswegen brauchen wir die Juristen und Privacy Engineering braucht auch Usable Privacy.

Nach zahlreichen Datenklauvorfällen herrscht in der Bevölkerung eine große Verunsicherung. Was müsste passieren, damit das Vertrauen in Systeme steigt, die personenbezogene bzw. sensible Daten speichern und verarbeiten?

Derzeit sind wir in der Situation, dass der Benutzer pauschal benachteiligt wird und keine realistische Chance hat, sich zu wehren. Vielen Benutzern ist diese Gefahr sicher bewusst, sodass man einerseits eine gewisse Resignation, andererseits auch einen Rückzug aus unterschiedlichen technischen Systemen beobachten kann. Die Datenschutz-Grundverordnung (DSGVO) ist sicher ein guter Schritt in die richtige Richtung – zumindest hat sie die Verpflichtung zum Datenschutz auch seitens der Wirtschaft und Institutionen ins Bewusstsein gerufen. Einiges ist da vielleicht noch schwierig umzusetzen: Wir Informatiker denken ja üblicherweise in Worst-case-Szenarien und wenn man dies konsequent bei etwa der Datenschutzfolgeabschätzung durchzieht, dann würde das wahrscheinlich zu einem Verbot fast aller Datenverarbeitungsprozesse oder der Notwendigkeit der Einholung von sehr umfassenden Einwilligungen nach sich ziehen. Da müssen wir wahrscheinlich lernen, vom Worst-case-Denken auf die Einschätzung absehbarer Risiken umzuschwenken – um eine Situation zu vermeiden, in der die Firmen schlicht wieder alles tun und behaupten, die DSGVO ließe sich ja eh nicht einhalten. Gleichzeitig müssen wir dafür sorgen, dass es datenschutzfreundliche Konkurrenzlösungen gibt – einmal, damit diese tatsächlich eingesetzt werden und so Datenschutz betrieben wird, aber auch, um der Argumentation, das ginge ja eh alles nicht und wir bräuchten halt den Datenreichtum, den Wind aus den Segeln zu nehmen. Dies wird insbesondere im Kontext des Internet of Things (IoT) sehr wichtig, weil hier ja eine immer umfassendere Beobachtung und potenzielle Überwachung der Benutzer einzieht. Da wird es natürlich auch schwierig, weil diese Geräte häufig keine Interfaces haben, über die man die Benutzer sinnvoll aufklären oder über die der Benutzer sinnvoll Schutzeinstellungen konfigurieren könnte. Einige dankenswerte Ansätze, z. B. in der Richtung von rein lokal agierenden Sprachassistenten und natürlich auch PET gibt es ja bereits. Schließlich, denke ich, müssen auch die Datenschutzbehörden, Verbraucherzentralen und Interessengruppen stärker im Interesse der individuellen Privatheit eingreifen. Natürlich dürfen wir nicht nur publikumswirksam Google verklagen, wir müssen auch bei den europäischen Firmen genauer hinschauen und sie zur Einhaltung der Grundrechte und der Abkehr von unfairen Praktiken zwingen. Ich glaube, nur dann können wir mittelfristig echtes Vertrauen in diese Systeme wiedergewinnen.

Im Kontext von intelligenten und vernetzten Systemen, wie man sie im Smart Home oder autonomen Fahrzeug findet, spielen Datenschutz und Datensicherheit eine sehr große Rolle. Welche konkreten Herausforderungen stellen sich für den Datenschutz und die Datensicherheit?

Hier gibt es eine ganze Reihe von Herausforderungen. In einer solchen Umgebung gibt es schon mal einen Paradigmenwechsel: von der aktiven Benutzung und passiven Beobachtung zur unbewussten Benutzung bei parallel sehr aktiver Beobachtung. Während es im Internet mehr darum geht zu untersuchen, was wer wann preisgibt, beobachten smarte Systeme den Menschen, was der Person unter Umständen gar nicht klar ist. Siri oder Echo sind da exzellente Beispiele, aber auch die Videokameras an Spielkonsolen oder SmartTV. Wie soll man da jemandem klar machen, dass Daten über ihn gesammelt werden und welche Auswirkun-

gen die Datensammlung für ihn hat? Eine große Herausforderung stellt sich sicherlich auch für die Netzwerksicherheit und Haftung. Verbraucher kaufen oft die günstigsten Systeme, für die es schnell keine Patches mehr gibt. Wie wir die Netze trotzdem sichern können und wer da in welchen Fällen wie zu haften hat, müssen wir sicherlich bald klären. Was sicherlich auch sehr spannend ist, ist der Bereich Schutz für Datenströme. In der Vergangenheit haben wir uns viele Gedanken darüber gemacht, wie man einzelne Daten (z. B. das Gehalt) von Personen schützen kann. Im IoT-Kontext untersuchen wir Datenströme. Problematisch hier ist, dass wir normalerweise davon ausgehen, dass die einzelnen Datenpunkte, die wir in der Vergangenheit betrachtet haben, statistisch unabhängig voneinander sind. Diese Annahme ist für Datenströme natürlich völlig unrealistisch. Denn Elemente in Datenströmen sind natürlich nicht statistisch voneinander unabhängig, sondern zwischen ihnen bestehen intensive Abhängigkeiten. Hier muss man sich Gedanken machen, was Privacy bedeutet (d. h. was kann ich schützen). Was bedeutet Differential Privacy oder wie funktionieren beweisbare sichere Ansätze zum Erreichen von Differential Privacy in diesem Kontext überhaupt? Das bleibt ganz sicher eine spannende Forschungsfrage – und Privacy im Allgemeinen wird sicher zu einer immer wichtigeren Herausforderung.

Vielen Dank für das Gespräch.

Das Interview wurde am 5. April 2019 von Agnes Koschmider geführt.

What spreadsheets are to numbers, process mining is to events

Matthias Weidlich

Wil van der Aalst is a full professor at RWTH Aachen University where he is leading the Process and Data Science (PADS) group. He is also part-time affiliated with Technische Universiteit Eindhoven (TU/e). Until December 2017, he was the scientific director of the Data Science Center Eindhoven (DSC/e) and led the Architecture of Information Systems group at TU/e. His research interests include process mining, Petri nets, business process management, workflow management, process modeling, and process analysis. Wil van der Aalst has published over 220 journal papers, 20 books (as author or editor), 510 refereed conference/workshop publications, and 75 book chapters. Many of his papers are highly cited and his ideas have influenced researchers, software developers, and standardization committees working on process support. He is also an elected member of the Royal Netherlands Academy of Arts and Sciences, the Royal Holland Society of Sciences and Humanities, and the Academy of Europe. In 2018, he was awarded a Humboldt Professorship.

Wil, everybody talks about data science. Yet, your chair is about process and data science. Why do we need a process perspective?

I think the need for data science is very clear and, indeed, everybody talks about it. At the same time, there has also been a long tradition of analysing processes, in operations research, in business process management, and other fields where people would typically ignore the data. The unique feature of my chair is that we try to combine process expertise with data science. So, to answer the question on why we need a process perspective: Processes are very complicated and notions such as concurrency are hard to grasp. Also, many processes happen, but you cannot touch them. This makes it a very interesting and important topic and people are struggling to connect the worlds of processes and data.

Speaking of this context, the general field of data mining is much older and more established than process mining. However, whenever data is generated, there is some behaviour that governs the creation of data. In many scenarios in which data mining is used these days, there should therefore be some potential for process mining. Would you agree and, if so, what is needed to leverage the potential of process mining more widely?

Let me first expand a bit on a general tendency to confuse things with different terms. The terms AI and machine learning get a lot of attention these days. There are many interesting successes in these fields, but there is also the problem that things quickly get confused, and there is a general ignorance of people not working in these domains related to terminology. If people talk about AI or machine learning, they often refer to deep neural networks and image classification. Everybody would agree that these techniques are very successful. Yet, they cover only a tiny part of the huge data science space. Today, public media and politicians talk about AI and it is unclear whether they refer to deep

https://doi.org/10.1007/s00287-019-01198-7

Matthias Weidlich
Humbold-Universität zu Berlin
E-Mail: matthias.weidlich@hu-berlin.de

neural networks, the whole spectrum of data science, or something else. This is not so much a matter of a discussion on the terminology. What counts is whether methods and tools provide certain capabilities. If I look at a topic such as discovering end-to-end process models from data, there are no classical data mining tools, no classical machine learning approaches, no AI approaches that would return such a process model. Discovering an end-to-end process model is a unique capability that is not provided by these methods and tools. In the end, it is very important that we talk about such concrete capabilities.

Coming back to your question, one could argue that process mining is a comparatively young discipline. Yet, given the relative age of the field, I am quite happy with the recent developments. Also, there are many things that people think are established, but their adoption in industry is not as big as we would like it to be, think of techniques for simulation, traditional data mining, or operations research. In comparison, the relative performance and adoption of process mining is quite ahead already at this time, and it is rapidly growing.

This is a good point as establishing a research field concerns various dimensions, such as industry uptake and academic teaching. Do you see an opportunity in the sense that the successful uptake of process mining in industry helps to further establish the field in academia and in particular in teaching?

That is very much the case. Just look at the academic programs of Celonis and Fluxicon and check the universities where these tools are used in courses. That is quite impressive. Also, my MOOC has now been taken by over 120,000 people. All these things are helping to increase the visibility. The success of process mining in industry also impacts what is being taught. I notice that for students, it is completely clear that this is something that is real, interesting, and challenging.

You mentioned the buzz on AI and machine learning. Is that something that is problematic or beneficial for the field of process mining?

The buzz certainly has a positive side effect. It creates opportunities for data science programs, data science research, etc. If you look specifically at process mining, however, the overall effect could be negative, because the buzz creates expectations which, I think, are fairly unrealistic. And that is going to lead to disappointment. Process mining is already a reality helping organizations to save millions of euros. Therefore, the association to AI and machine learning can be negative once we have passed the peak of inflated expectations.

… so, the field may be hit by the next AI winter?

It is very clear that the next AI winter is coming. Many companies spend a lot of time and money on AI-like initiatives, without being very concrete. There will be some backlash related to this, which will make it more difficult to do process mining projects, as people will think that all this is the same.

Speaking of expectations, there have been debates on whether data science shall be considered as an independent field, or whether it shall always be grounded in a domain. What is your take on the importance of the application domain for process mining?

I think that data science is an independent discipline. The debate reminds me of when I started my bachelor in computer science in 1984. This was two years after the creation of a computer science program in Eindhoven. At the time, there was discussion whether computer science is a real science (and not just applied mathematics). I think nobody will question that today and the same will happen with data science. It grew out of computer science and statistics, but there are now many techniques that are not so much related anymore to the original concepts in statistics and computer science.

Process mining is generic and not specific for a particular application domain. Process mining is like a spreadsheet. With a spreadsheet you do anything with numbers, with process mining you do anything with events. And if you take that metaphor, you realize there is no special process mining for health care, logistics, etc. Of course, different application domains lead to different challenges, for example in terms of scalability or how data is preprocessed. Overall, however, I think of process mining as something that is very generic.

Let's talk a bit more about process mining. Your first papers on this topic have been published around 20 years ago. Two decades are a long time and with initiatives such as the international conference on process mining and successful industry adoption,

the field has come a long way. Looking back, what do you consider most surprising in how the field has developed?

What is surprising is how broad the field has developed. It started with discovering control-flow models, but today there is also conformance checking, performance analysis, predictive analytics, and much more. It is also surprising that many of the core problems have not been cracked yet. If you look at my papers between the years 2000 and 2006, we defined many algorithms to tackle problems that are still a challenge today. For example, how do you evaluate a discovered process model? Despite many peoples' claims, this is an unsolved problem. Also, if you look at process model discovery, for me it is still amazing that a human, first analysing the data then doing some modelling, can still come up with a better model through various iterations than algorithms. So, the field has been much broader and, in a way, also much deeper than I expected. For many challenges, there has been a lot of progress, but at the same time, there are still many open problems.

So, back then, you expected that this case would have been closed within a decade?

(laughing) Well, as a researcher, I'm very happy with all the open problems.

When thinking about the field becoming broader, one may also consider successful exchange with other communities, thinking of work on predictive monitoring and links that have been established to operations research and natural language processing. What are the communities where such exchange is not yet happening to the desired extent?

The process mining community, as it emerged from the BPM community, is relatively open to adopting ideas from other fields. For example, if you have a control flow model and would like to make predictions, people adopt the ideas of deep neural networks or traditional data mining. If it is about finding performance bottlenecks, people know that concepts of queueing theory can be exploited. So, the rather small process mining community has been looking at various other fields. What is a bit disappointing is that not many people in these other fields pick up on the challenges that were presented. This may be related to communities, to personalities, also the fact that people simply do not understand and see the problems. It is an issue that process mining is often considered as a specific data mining technique, not acknowledging its unique challenges. For the future, I would hope that people from, for example, the data mining community would embrace the posed problems. Similarly, looking into the field of stochastic processes, there are visionary people like Avishai Mandelbaum, who is one of the few that see that this field needs to embrace the data that is available today. Those are just two examples of other disciplines that could contribute to process mining and I would hope that they actually see and embrace these challenges.

When thinking about research communities with which exchange may be beneficial, people working on models for privacy and confidentiality come to mind. Process mining data is often related to clients, to process workers, to stakeholders, to customers – there are simply a lot of people involved in processes. Recently, initiatives like the General Data Protection Regulation received a lot of attention. How will these trends affect process mining?

In the Netherlands, I led a general initiative on Responsible Data Science (RDS). On the one hand, I could see how important it is to bring together different communities. At the same time, I also saw the difficulties. You mentioned the GDPR and I have listened to many talks by legal experts in this field. It was always very difficult for me to see the connection to challenges that we could actually undertake. However, if these people with a legal background would think about privacy-preserving data mining techniques, they would also consider it to be very foreign. So, to bring these different people together is very important for RDS, but a big challenge.

As computer scientists, I suspect we love technical problem formulations, for which we can then define notions, algorithms, proofs, approximations, etc. May the issue be that the challenges related to responsible data science are not primarily on the technical level?

Indeed there are different challenges. However, it is our responsibility, as say data scientists, to try to develop techniques and tools that facilitate people in handling the negative side effects of the use of data. And these challenges are technical challenges. If you use the metaphor of driving a car as using data science techniques, then we would like to prevent pollution and people dying in traffic. It is clear that safety and the environment are strongly influenced

by how people use cars. If they always drive 250km/h, there is a lot of pollution and they are more likely to kill people. At the same time, the car itself can be made cleaner and safer. Therefore, if you use that metaphor, then you see that you need to do both. The bottom line for me personally is that one should not only try to solve these challenges in terms of legislation, but also provide people with the tools to make analytics safer.

The GI recently published ethical guidelines for computer scientists and engineers, whether they work in academia or industry. Clearly, these guidelines are abstract. Yet, they provide some general principles to relate to. What is your take on such measures?

There is an important educational aspect. In any data science education, these issues shall be addressed. For example, I give the lecture "Introduction to Data Science" and talk two weeks about these issues. While this is very important, it is not the only thing that should be done. We need to think creatively of technology that will protect people. You may say that people can use a hammer in all kinds of horrible ways, so let us get rid of the hammer. Partly, it is the decision of people to have laws to decide on what people are allowed to do with the hammer. However, for certain purposes, you can think of a safer device with less potential for negative side effects, which protects people.

In your curriculum as part of the responsible data science initiative, you advocate the goals of fairness, confidentiality, accuracy, and transparency, for data science in general, and process mining specifically. These goals are easy to grasp on the abstract level. Personally, however, I find it non-trivial to understand the specific implications for an analysis technique or application. Would you agree and what can we do about that?

I agree, so we should think of how we can achieve these goals for specific examples. In process mining, event data are often super sensitive. You only need a few events that are correlated to directly link some process execution to a person. Also, if people make decisions about processes, informed by process mining techniques, then this may have negative consequences for customers and employees working in the process.

Now, think about confidentiality. Knowing that event data are very sensitive, we work on techniques, where the analyst does not get the raw data, but only encrypted data which prevents any reconstruction of the original events. The analyst is still able to analyse the process without going back to the individual events. That is a bit like Apple advertising the fact that they are not storing the routes that somebody takes, but they break long routes into smaller parts. If you have a long route, it would be very easy to identify specific people, whereas you may get almost the same quality of analytics by just using route fragments. You could do similar things with event data, where you break these longer correlations and still create process models showing bottlenecks and deviations, etc., without being able to link these insights to individual events. That would be one example of a concrete technology to ensure confidentiality. Such techniques are currently not available in commercial tools, though, where you can always drill down to the lowest level of detail to see any information you want.

Another example is the notion of fairness. So if you use data science techniques, it's very easy to make decisions that are unfair or simply wrong. For example, if you look at commercial tools for conformance checking, you will see a list of deviations. For every deviation, it is indicated what kind of effect it has, like the process running five days longer if a specific activity was skipped. At the click of a button, you can then drill down and see which person worked on specific activities. The tools make it very easy to look at a particular process outcome, a bottleneck, a deviation, and link it back to characteristics of the cases where the undesirable behaviour takes place. This potentially leads to completely wrong conclusions. There may be strong correlations between a bottleneck in your process and particular people. However, it could very well be that the involved people are simply overloaded with too much work. Here, we need techniques to avoid making unfair conclusions.

I would also argue that these questions become even more important when we extend process mining from traditional event logs to sensed data. Would you think that this creates a danger that, if we as a community do not have an answer to these challenges, further advances in data analytics will simply not be appreciated anymore?

This is something related to the data science field as a whole. Earlier, we talked about AI winters and, in the past, those have been characterized by the fact that people were disappointed that analytics did not really work. There will be a winter in this sense again, as we already see people become sceptical about how organizations use data. Also, handling data in a responsible way will increasingly become a competitive advantage. As a company, you may want to provide evidence that you are handling data in a clean way, even though the latter is still to be defined. This is also an opportunity here in Europe, compared to other countries where people care much less about these issues.

These differences create tensions in particular for international companies that face different expectations by people, who may be more or less willing to adopt certain types of analytics and share certain data.

That is why we need notions like confidentiality and fairness. They need to become a parameter of the algorithms, such that a company can adjust them to the situation at hand. There is a clear trade-off, for example, between the accuracy of analysis results and confidentiality. You can anonymise and anonymise, until you cannot conclude anything anymore. What I envision is that more and more algorithms will have a slider functionality to capture this trade-off. Of course, there are further challenges. For example, in process mining based on SAP, people are surprised what we can all find and the results are sometimes considered highly sensitive. What they do not realize, though, is that any system administrator could have typed in a query and could have gotten exactly the same view. Beyond the algorithmic challenges, there is therefore a need to raise the awareness for these issues.

Wil, I take that as a call for action. Thank you very much for sharing these insights!

Thank you, I enjoyed talking about topics that are close to my heart.

I'm sure we will hear more about those topics from you in the years to come. Thanks again.

The interview has been conducted by Matthias Weidlich.

Was bedeutet Process Mining für Datenschutz und Mitbestimmung im Unternehmen?

Martin Degeling

Beim Process Mining werden verschiedene Daten, die im Unternehmen entstehen, zusammengeführt und ausgewertet. Das können technische Systemdaten sein, aber auch solche, die direkt oder indirekt einer Person zuordenbar sind. Das wirft Fragen rund um Datenschutz und Verhaltenskontrolle von Beschäftigten auf.

Richtig, aber man muss hier zwischen zwei Rechtsgebieten unterscheiden. Es geht um Datenschutz auf der einen und Mitbestimmung auf der anderen Seite, auch wenn es zwischen beiden eine große Verzahnung und Abhängigkeiten gibt. Beim Datenschutz ist die erste Frage die gestellt werden muss: Hab ich es überhaupt mit personenbezogenen Daten zu tun? Das betrifft direkt personenbezogene Daten, etwa wenn ich feststelle, dass eine Maschine mehrfach am Tag überhitzt, und dann nachschaue, wer damit gearbeitet hat. Es gilt aber auch für die Analyse mit Big-Data-Methoden, bei denen man vorhandenen Daten in ein anderes System überführt und dann dort möglicherweise keine personenbezogenen, sondern rein statistische Auswertungen durchführt – auch dann muss man sich Gedanken machen, welche Datenschutzanforderungen zu beachten sind.

Auf der anderen Seite – und das ist, grundsätzlich ein anderes Rechtsgebiet, das sich vorrangig an den Systemen orientiert und nicht an den Daten – gilt es zu schauen: ist ein System geeignet, das Verhalten von Beschäftigten zu kontrollieren? In dem Fall wird die Mitbestimmung ausgelöst. Das heißt, das System darf nicht eingeführt werden, ohne dass der zuständige Betriebs- oder Personalrat zugestimmt hat. Mitbestimmung und Datenschutz sind die zwei Säulen des Beschäftigtenschutzes.

Wo liegen die Unterschiede?

Natürlich spielt Datenschutz im Rahmen der Mitbestimmung bei der Gestaltung von Systemen eine Rolle. Genauso wie eine Betriebsvereinbarung Vorgaben zur sachgerechten Nutzung und zur Berechtigungsgestaltung enthält, müssen auch Datenschutzanforderungen berücksichtigt werden.

Aber, ganz unabhängig von Mitbestimmung, zum Beispiel auch, wenn es gar keinen Betriebsrat gibt, sind Arbeitgeber verpflichtet, datenschutzkonform zu handeln. Wenn also ein System die Beschäftigten nach Strich und Faden kontrolliert und überwacht, es aber keinen betrieblichen Partner gibt, der das Mitbestimmungsrecht wahrnimmt, müssen die geltenden Datenschutzvorgaben, wie beispielsweise das Verbot mit Erlaubnisvorbehalt eingehalten werden.

Konkret heißt das für das Process Mining: Ich muss mir als Betreiber darüber klar werden, was für mich die Rechtsgrundlage ist. Bei vielen derartigen Systemen dürfte es schwerfallen zu argumentieren, dass sie zur Durchführung des Beschäftigungsverhältnisses notwendig sind. Denn der normale Betrieb, wie Personalbeschäftigung, -bezahlung und -verwaltung sowie die Einsatzplanung konnte auch bisher ohne Process-Mining durchgeführt werden. Außerdem handelt es sich beim Process-Mining ja häufig um eine Systemart, die eher auf

https://doi.org/10.1007/s00287-019-01197-8

Martin Degeling
ist Mitarbeiter am Horst Görtz Institut für IT Sicherheit an der Ruhr-Universität Bochum.
Er forscht zur Anwendungbarkeit von Privacy-by-Design in der Software Entwicklung und Datenschutz im Internet.
E-Mail: martin.degeling@ruhr-uni-bochum.de

der Metaebene liegt: nicht selbst produktiv, sondern andere Systeme analysierend und zur Unterstützung bei der Optimierung betrieblicher Prozesse. Auf der Datenschutzebene und aus Sicht des Unternehmens kann man hier zur Identifizierung einer Rechtsgrundlage eigentlich nur mit eigenen Interessen argumentieren. Denn es liegt im berechtigten Interesse eines Unternehmens, die eigenen Arbeitsweisen kontinuierlich zu optimieren, um wettbewerbsfähig zu bleiben.

Das heißt, dass man nicht versucht, mit der Durchführung des Beschäftigungsverhältnisses zu argumentieren, oder gar Einwilligungen einzuholen. Beides verspricht datenschutzrechtlich keinen Erfolg. Das Eigeninteresse muss dann allerdings gut begründet werden: ich muss nachvollziehbar darlegen, dass ich genauso vorgehen muss, wie ich es vorhabe. Die entscheidende Frage ist dann, ob es gelingen kann, die Verarbeitung, im Sinne der betroffenen Beschäftigten so datenschutzfreundlich zu gestalten, dass eine datenschutzrechtlich saubere Abwägung zwischen deren Rechten und meinen eigenen Interessen zu meinen Gunsten ausgeht.

In einer solchen Abwägung muss abzulesen sein, welche Einschränkungen der Persönlichkeitsrechte der Beschäftigten zu erwarten sind, um welche es sich handelt, und wie schwerwiegend diese jeweils sind. Und auf der anderen Seite muss dargelegt werden, wie groß mein Interesse an der Durchführung der Verarbeitung ist und worauf es sich stützt.

Warum ist die Einwilligung keine sinnvolle Rechtsgrundlage?

Generell ist die Einwilligung eine potenzielle Rechtsgrundlage für Verarbeitung personenbezogener Daten, aber sie ist an bestimmte Voraussetzungen geknüpft. Eine wesentliche ist: Sie muss freiwillig sein. Die zweite: sie muss jederzeit zurücknehmbar sein und es dürfen keine negativen Konsequenzen für den Betroffenen daraus erwachsen. Das ist im Arbeitsverhältnis in den seltensten Fällen möglich. Vor allem, weil es kein Kräftegleichgewicht zwischen Arbeitgeber und Beschäftigten gibt. Wenn ich als Arbeitnehmer von meinem Chef gefragt werde: „Du bist doch bestimmt einverstanden, dass ich das und das mit deinen Daten mache!“ habe ich schnell das Gefühl: Wenn ich jetzt „Nein“ sage, dann werde ich Nachteile haben. Da kann von einer freiwilligen Einwilligung keine Rede sein.

Es gibt nur sehr wenige Situationen im Arbeitsleben, in denen wirksam eine Einwilligung von Beschäftigten durch den Arbeitgeber eingeholt werden kann. Und mein weiterer Zweifel beruht auf Praktikabilität: ich muss – egal, ob im Beschäftigungsverhältnis oder sonst wo – bei Einwilligung immer dafür sorgen, dass ich in der Lage bin, einen Widerspruch angemessen zu behandeln. Das heißt, wenn jemand mir heute eine Einwilligung gibt, seine Daten dann verarbeitet werden, und die Person morgen kommt und die Einwilligung zurückzieht, dann muss ich bzw. mein System in der Lage sein, dessen Daten herauszusuchen und die Verarbeitung einzustellen. Und wenn das eine größere Menge von Personen macht, dann müssen meine Verarbeitungen trotzdem noch sinnvoll möglich sein. Wenn etwa plötzlich 50 % meiner Beschäftigten der Verarbeitung widersprechen, dann werden Ergebnisse meiner Berechnungen vermutlich nicht mehr repräsentativ sein. Deswegen glaube ich nicht, dass in einem solchen Zusammenhang eine Einwilligung wirklich eine sinnvolle Rechtsgrundlage ist.

Wenn der gangbare Weg ist, über das berechtigte Interesse zu argumentieren: Wie muss dann eine Abwägung aussehen?

Das Ziel aus Sicht derer, die personenbezogene Daten verarbeiten wollen, ist es, die eigenen Interessen nicht künstlich, aber berechtigt und gut nachvollziehbar, hoch einzuschätzen und auf der anderen Seite ist genauso nachvollziehbar darzulegen, dass die Interessen der Betroffenen nicht überwiegen. Dazu müssen ausreichende Schutzmaßnahmen ergriffen werden, damit die Einschränkungen der Persönlichkeitsrechte der Betroffenen möglichst gering bleiben.

Darunter fallen die üblichen Maßnahmen, wie beispielsweise die frühzeitige Anonymisierung oder zumindest Pseudonymisierung. Und echte Anonymisierung bedeutet wohlgemerkt die Entfernung eines Personenbezugs, nicht nur das einfache Löschen von direkt identifizierenden Merkmalen, wie z. B. des Namens.

Zu den Schutzmaßnahmen gehört aber auch die ordentliche Strukturierung und Konzeptionierung der Sicherheit des Gesamtsystems. Die einfache Zusicherung, man wolle die Zweckbindung einhalten, ist nicht ausreichend. Sondern auch durch technisch sichere Gestaltung des Systems muss dafür gesorgt werden, dass diese auch eingehalten werden kann.

Etwa durch ein sachgerechtes Berechtigungssystem. Es ist aber auch zu beachten, wo die Daten verarbeitet werden. Passiert das vor Ort oder werden sie in einen Cloud-Dienst überführt, dessen Server im nicht-europäischen Ausland stehen? Dies würde zu einer schwierigeren Abwägung der eigenen Interessen mit denen der Beschäftigten führen, weil Beschäftigte ein sehr hohes Interesse daran haben, dass ihre Daten den Geltungsbereich der Datenschutzgrundverordnung nicht verlassen.

Wichtig, und auch „gern" vergessen, wird ein ordentliches Löschkonzept. Das heißt, es muss klar sein, wann welche Daten nicht mehr erforderlich sind und wann sie gelöscht werden. Das betrifft nicht nur den Datenpool, in dem Daten aus verschiedenen Systemen für das Process Mining zusammengeführt werden, sondern auch die Quelldatenbanken selbst. Es ist also festzulegen, wann welche Daten nicht mehr erforderlich sind und wann sie gelöscht werden. Gelöscht werden müssen außerdem die Ergebnisse von Analysen, wenn diese personenbezogene Daten enthalten. Und auch für Ergebnisse und Auswertungen, die papierhaft weiterverarbeitet werden (z. B. in Infos an das Management) sind Vernichtungsregeln festzulegen, wenn sie personenbezogene Daten enthalten.

Dazu müssen auch organisatorische Prozesse betrachtet werden. Alle Löschfristen und deren Begründung müssen sachgerecht dokumentiert werden.

Am Ende sollten durch die ergriffenen Sicherheits- und technisch-organisatorischen Schutzmaßnahmen die Risiken für die Beschäftigten so überschaubar und gering sein, dass man im Rahmen der Abwägung berechtigterweise von einem überwiegenden Eigeninteresse ausgehen kann.

Das ist aber nur die eine Seite.

Genau, da kommen wir zurück auf die andere Säule. Auch wenn eine Lösung datenschutzkonform ist, müssen trotzdem noch Mitbestimmungsrechte eingehalten werden, wenn das System zur Verhaltenskontrolle geeignet ist. Natürlich können all die erwähnten Schutzmaßnahmen auch Inhalt der Betriebsvereinbarung sein. Im Einzelfall kann eine solche Betriebsvereinbarung dann sogar selbst die Rechtsgrundlage sein. Dann gehen Datenschutz und Mitbestimmungsrechte Hand in Hand.

Ok, kommen wir nochmal zurück zur Frage der Clouddienste. Häufig sind diese ja auch attraktiv, weil man diese auch für einen ersten Test nutzen kann, um kurzfristig die Auswertung der Daten zu starten, ohne eigene Infrastruktur aufzubauen. Wie ist das mit dem Datenschutz vereinbar?

Sowohl für Mitbestimmung als auch für den Datenschutz, gibt es keinen Testbetrieb, der in irgendeiner Weise schwächere Anforderungen hätte als ein Regelbetrieb. Und daher muss ich auch für einen Testbetrieb sowohl datenschutzrechtliche als auch mitbestimmungsbezogene Überlegungen anstellen. Denn in dem Moment, in dem ich das erste personenbezogene Datum in die Cloud übertrage, habe ich eine datenschutzrelevante und mitbestimmungsrelevante Verarbeitung vorgenommen.

Im Bereich der Mitbestimmung kommt es häufig vor, dass man sich auf Pilotvereinbarungen einigt, in denen festgelegt wird, was das Ziel des Tests ist und die Begrenzungen in Bezug auf Laufzeit, Umfang der Verarbeitung, Zugriffsrechte und Zwecke enthalten.

Wenn sich anschließend abzeichnet, dass ein System dauerhaft genutzt werden soll, kann man dazu die endgültige Betriebsvereinbarung erarbeiten. Auf der Datenschutzebene ist dieses stufenweise Vorgehen aber nicht möglich. Wenn Clouddienste genutzt werden, muss immer ein Auftragsverarbeitungsvertrag abgeschlossen werden, der den Anforderungen der Grundverordnung entspricht, unabhängig davon, ob die Nutzung testweise oder dauerhaft erfolgt.

Beim Process Mining werden Daten ja häufig für einen anderen Zweck genutzt als ursprünglich vorgesehen. Was gibt es da zu beachten?

Die Zweckänderung von Daten erfordert letztlich immer, dass man nochmal von vorne mit der Prüfung der Rechtsgrundlage anfängt. Man hat Daten für einen bestimmten Zweck rechtmäßig erhoben und muss, wenn man Sie für einen anderen Zweck nutzen will, auch hier wieder Erforderlichkeit, Rechtmäßigkeit und so weiter nachweisen. Wenn man dann zum Schluss kommt, dass man sie auch für den neuen Zweck erheben dürfte und es gibt noch keine Betriebsvereinbarung die genau diesen neuen Zweck anschließt, muss man diesen mindestens in der alten ergänzen.

Sie haben bereits kurz die Problematik von Pseudonymisierung und Anonymisierung angesprochen. Können Sie hier noch etwas in die Tiefe gehen?

Wir können heute mit Sicherheit sagen, dass z. B. das einfache Löschen einer Namensspalte aus einer Tabelle in großen Datenbeständen keine Anonymität herstellt. Es gibt einfach zu viele Möglichkeiten, auch über Querverbindungen und Zusatzwissen noch Aussagen darüber treffen zu können, um wen es sich da wohl handelt. Anonymität herzustellen stellt in der Informatik eine anspruchsvolle Aufgabe dar. Aber das bedeutet nicht, dass es unmöglich ist. Bei vielen existierenden Systemen ist mein Eindruck eher, dass die Umsetzung von Anonymität versäumt wurde, weil es methodisch und inhaltlich nicht trivial und damit eben auch teuer für die Hersteller ist.

Im wissenschaftlichen Bereich ist die Diskussion relativ weit gediehen, aber ich bin mir nicht sicher wie weit das in Systemen umgesetzt ist. Ähnliche Probleme gab es viele Jahre im Bereich der Löschkonzepte. Auch vor 10 Jahren war schon bekannt, dass Unternehmenssoftware Löschkonzepte vorsehen muss, aber damals hat sich ein sehr bekannter, deutscher Hersteller einer verbreiteten Unternehmenssoftware noch auf einer Konferenz so geäußert: „Wenn ein Kunde kommt, der uns eine Million in die Hand drückt, weil ihm das wichtig ist, dann kümmern wir uns ums Löschen. Solange machen wir das nicht." Hier hat die Diskussion um die Datenschutzgrundverordnung gezeigt, dass, wenn es Kunden an den Geldbeutel zu gehen droht, bestimmte Funktionalität stärker nachgefragt und dann auch umgesetzt wird.

Welche Möglichkeiten für datenschutzfreundliches Process Mining gibt es noch?

Eine zentrale Frage ist ja: Brauche ich eine Zuweisung zu irgendeiner eindeutigen personenbezogenen Kennung, damit ich überhaupt eine allgemeine Entwicklung im Unternehmen wahrnehmen kann? Muss ich unbedingt wissen, dass: ein bestimmter Datensatz in Zusammenhang mit Person X entstanden ist und darüber hinaus personenbezogen alle Datensätze in Zusammenhang mit Person X in Beziehung setzen können? Oder bliebe die Erkenntnis die gleiche, auch wenn ich die Datensätze ohne Personenbezug beliebig durcheinander würfeln würde? Eine weitere Frage: welche Angabe stellt an diesem jeweiligen Datensatz der Personenbezug her? Recht häufig kann man den durchaus vermeiden, ohne den Erkenntnisgewinn zu schmälern. Nehmen wir das Beispiel Versandhandel. Eine Mitarbeiterin geht mit einem Handscanner durch das Lager und hat vom System eine Liste mit Waren bekommen, die sie zusammenstellt. Sie scannt jeden Artikel und das datenschutzrechtliche Problem entsteht erst in dem Moment, in dem die Zuordnung von Handscanner zur Person möglich wird. Das Unternehmen interessiert sich nun vielleicht dafür, wie die Prozesse funktionieren und sucht Optimierungspotenziale. Dafür muss es aber eigentlich nicht wissen, wer den Handscanner benutzt hat. Statt die Daten technisch vor der Auswertung zu anonymisieren, könnte man auch Prozesse geschickt gestalten. So zum Beispiel, indem man im beschriebenen Fall einfach ein rollierendes System einsetzt, bei dem der Handscanner jeden Tag von jemand anderem benutzt wird. So lassen sich die Prozesse über die Handscanner-Nummer beobachten, ohne zu wissen, welcher konkrete Beschäftigte da jeweils diesen Handscanner in der Hand gehabt hat.

Natürlich ist das eine sehr spezifische Lösung, aber häufig lassen sich auch organisatorische Lösungen finden, die technisch weniger aufwändig sind.

Abschließend die Frage nach der Umsetzung. Welche Rollen können Datenschutzbeauftragte spielen?

Datenschutzbeauftragten haben den Datenschutz für die Beschäftigten genauso zu beachten und zu fordern wie für die Kundendaten oder für Partnerdaten. Das gilt ganz unabhängig von Mitbestimmungsrechten. Die Rolle der Datenschutzbeauftragten unterscheidet sich auch nicht, wenn es keinen Betriebsrat gibt. Gibt es jedoch einen Betriebsrat und sind Betriebsvereinbarungen in Kraft, die auch datenschutzrelevante Regelungen enthalten, dann muss der Datenschutzbeauftragte auf deren Einhaltung hinwirken – genauso wie auf die Einhaltung aller anderen Datenschutzvorschriften.

Vielen Dank für das Gespräch.

Karin Schuler
freiberufliche Beraterin für Datenschutz, IT-Sicherheit und Mitbestimmung (www.schuler-ds.de)
Gründungsmitglied des Netzwerks Datenschutzexpertise (www.netzwerk-datenschutzexpertise.de)
Mitglied der Fachgruppe PET der Gesellschaft für Informatik e. V.
E-Mail: buero@schuler-ds.de

Das Interview wurde am 12. April 2019 von Martin Degeling geführt.

Process Mining and Privacy in Smart Manufacturing

Felix Mannhardt
Sobah Abbas Petersen
Manuel Oliveira

Introduction

Operators in industrial manufacturing environments are under pressure to cope with the ever-increasing flexibility and complexity of their work. In Industry 4.0, the use of smart technologies and sensors is seen as the future of manufacturing [14]. Tomorrow's factories [6] will leverage on new and emerging technologies and digital solutions to enhance collaboration, not only among the workers but also between the human workers and technology – Human in the Loop (HITL). Technologies such as exoskeletons and other wearables promise collaboration networks between humans and technology, leading to improved health and safety at the workplace [10] and reduced physiological load. Similarly, other types of technologies, such as augmented reality (AR), support workers with cognitive load minimization and performance improvement [3]. Many of these technologies offer personalized services to help industry workers through enhanced data collection, which in turn poses potential risks for privacy [4]. Moreover, with the increase in the availability of data, new opportunities arise to (1) use such recorded data in combination with machine learning techniques to deliver timely and relevant assistance to the operator, and (2) employ it retrospectively for work process optimization through data analytics techniques such as process mining [8]. Despite these clear opportunities, challenges are raised concerning the perceived threats of the usage of the data that outweigh the perceived benefits [12]. Thus, it is paramount to strongly consider privacy concerns when designing a system for a smart manufacturing environment from the onset rather than as an afterthought. This experience report summarizes our work on how to deal with process mining and the associated privacy challenges in a human-centered smart manufacturing environment as it is developed in the EU H2020 project HUMAN Manufacturing [7]. First, we present a brief summary of process mining in manufacturing and describe how it is used in the HUMAN project. Then, we describe the proposed HUMAN Trust and Privacy Framework [13] that is used in the project to raise awareness of privacy issue and guide developer in the design of privacy aspects in the system.

Process mining in manufacturing

Process mining promises to provide a data-driven view on the actual execution of any kind of process in which the execution of discrete activities is logged [1]. Typically, the execution of a process instance results in a sequence of events being recorded. In general, such an execution trace, also denoted log trace, contains at least the timestamps of activity executions and names or identifiers of the executed activity. Each log trace groups together the activities performed in one instance of a recurring process. An event log contains a set of several such log traces from which process mining methods discover an end-to-end process model (process discovery) or diagnose deviations from existing process models (conformance checking). Thus, in comparison to

https://doi.org/10.1007/s00287-019-01199-6

Felix Mannhardt · Sobah Abbas Petersen · Manuel Oliveira
Technology Management, SINTEF Digital,
Trondheim, Norway
E-Mail: {felix.mannhardt, sobah.petersen,
manuel.oliveira}@sintef.no

data mining methods, process mining adds the notion of process instances and activity execution sequences that are captured in log traces and analyzes end-to-end processes. A comprehensive survey on process discovery methods is presented in [2], and an introduction to conformance checking is given in [5].

In the context of smart manufacturing, several processes are worth analyzing using process mining [12]. Typical candidates for activities are, e. g., the individual assembly tasks performed by operators and logistical activities around the supply with parts and materials. However, data sources are not limited to the execution of activities. Manual tasks that are often hidden from databases or logs may be automatically recognized by means of activity recognition based on sensor data [9] and [11]. Moreover, it is possible to fuse sensor data from wearable device and machines (e. g., connected to the Internet of Things [IoT]) together with the execution of work activities [8] and overlay sensor data with a discovered process model. Finally, machine learning can be employed to detect undesired situations such as mental or physical stress encountered by operators. Such events can also be used as input for process mining.

The HUMAN project

The goal of the HUMAN project is to digitally enhance the operator on the shop floor to support them in their work, assisting them in mitigating any productivity losses resulting from both physical and cognitive fatigue, whilst contributing to their greater well-being. Towards this goal, the cognitive system envisioned in HUMAN captures physiological data from the operator (through wearable sensors), is aware of the production context (e. g., tasks, workplace) in which the operator is embedded and uses data analytics on historical data to provide timely and contextualized support for operators. A few examples of services provided through the HUMAN system are described below:

- The Knowledge in Time (KIT) service provides support to the actions of operators on the shop floor using an augmented reality (AR) system that helps the operator when needed, i. e., it provides in-time cognitive support.
- The EXOS service is coupled to a light-weight exoskeleton that an operator wears to distribute the physical stress on their body. The HUMAN solution monitors the different physiological signals, such as heart rate and galvanic skin response, to determine the onset of fatigue and, thereby, activate the exoskeleton to support the operator.
- The Workflow Optimization Service (WOS) is a HUMAN service that utilizes virtual as well as augmented reality in order to simulate real manufacturing environments and processes, assess

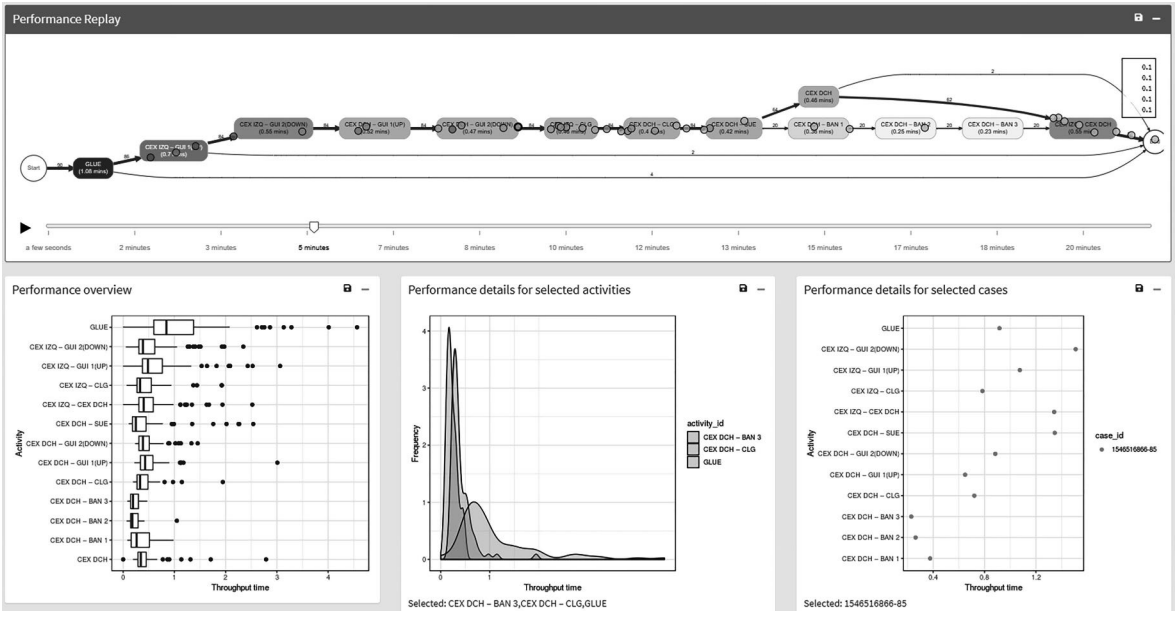


Fig. 1 Process mining widget of the SII service showing data from a furniture assembly process

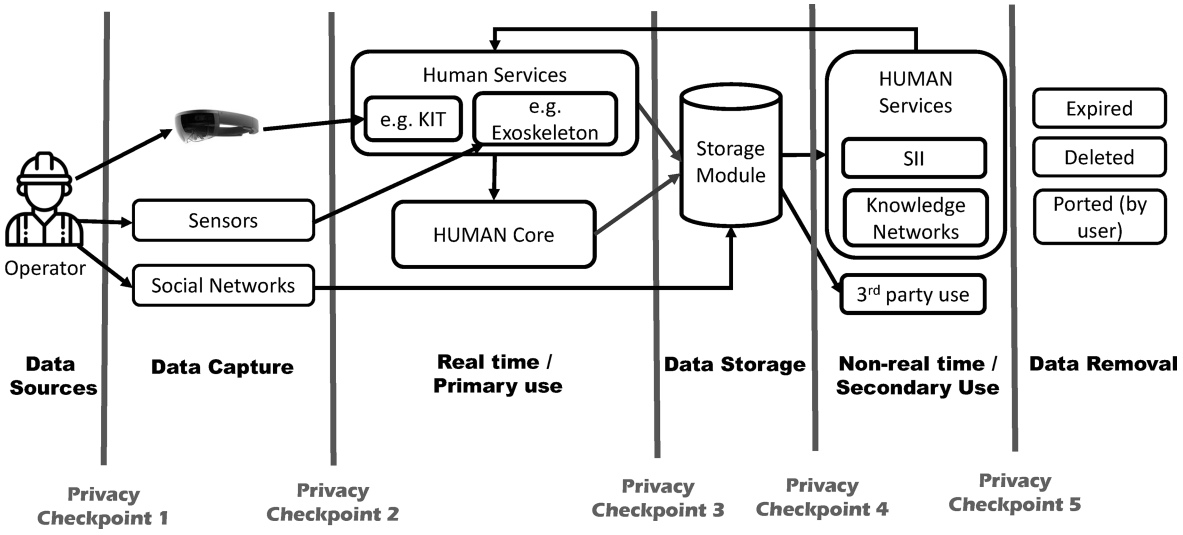


Fig. 2 The HUMAN trust and privacy framework for smart manufacturing environments applied to the HUMAN project [13]

them, modify the virtual workplace, and check for the effectiveness of the modifications it terms of ergonomics, usability, and efficiency.
- The Shopfloor Insight Intelligence (SII) service provides process mining capabilities based on historical data. It relies on the data captured from other services and combines data from several sources into a coherent event log on which process mining methods are applied to reveal what happened on the shop floor. Based on aggregated data as well as individual executions of work processes, the SII aims to reveal and help identifying root causes for recurring physical and cognitive stress that cannot be alleviated automatically by the available intervention measures of the HUMAN system.

As an example, a widget from the SII service is shown in Fig. 1. The widget was used to analyze an assembly process for bathroom furniture based on data collected from the KIT service, which provided AR support to the worker.

The HUMAN trust and privacy framework

Despite the opportunities to improve both the work environment and the way of working through analyzing work processes, which may also benefit the operators, the collection, storage, and processing of such personalized data comes with many challenges. In fact, there may be many justified reasons for operators to oppose the processing of their data. Operators may fear that collected data may compromise their career progression or may not be suitably protected against access from adversaries. It is not only privacy regulations, such as the omnipresent EU GDPR (Europe's General Data Protection Regulation), that impose compliance challenges for organizations that aim to employ such kind of data analytics, but also the trust challenges in the relationship between employer and employee. Clearly, this leads to both technological and organizational challenges in using data for analysis with process mining as outlined in [12].

To address this challenge, we create a trust and privacy framework for the smart manufacturing environment. In Fig. 2, an application of the this trust and privacy framework [13] to the situation in the HUMAN project is shown. In the envisioned application scenario of the framework, there is a strong emphasis on trust beyond the system, and in the organization and among individuals. The framework tracks the life cycle of collected data through the stages from its source (data source) and recording (data capture) across the uses (primary use and secondary use) in multiple interconnected services for various purposes until its removal (data removal). The framework includes five privacy checkpoints that are crossed when transitioning between stages of the data element's life cycle. For each privacy checkpoint, the framework closely examines the implications of

data crossing the stages (e. g., being stored after primary real-time use) in terms of trust and privacy and provide general guidelines for developers. In summary, the proposed framework aims to fill a gap in the role of privacy for smart manufacturing in the context of Industry 4.0 by integrating and facilitating the understanding of the role of trust and privacy in complex smart manufacturing systems, which provide contextualized and situationally aware services combined with data analytical services.

Conclusion

This experience report summarizes our experience with trust and privacy aspects in smart manufacturing environments such as the one built in the HUMAN project. We highlight the use of process mining as a possible method to analyze data collected from manual work on the shop floor and briefly describe the HUMAN trust and privacy framework that is used in the project to raise awareness of trust and privacy aspects and to help developers considering trust and privacy when designing systems such as the one built in HUMAN.

Acknowledgements

This research is funded by the EU's H2020 research and innovation program, grant agreement no. 723737 (HUMAN). We thank all participants of the trust and privacy workshops conducted within the context of the HUMAN project.

References

1. van der Aalst WMP (2016) Process Mining – Data Science in Action, 2nd ed. Springer, Berlin Heidelberg
2. Augusto A, Conforti R, Dumas M, La Rosa M, Maggi FM, Marrella A, Mecella M, Soo A (2019) Automated discovery of process models from event logs: review and benchmark. IEEE T Knowl Data Eng 31(4):686–705
3. Blattgerste J et al (2017) Comparing Conventional and Augmented Reality Instructions for Manual Assembly Tasks. PETRA '17. Rhodes, Greece
4. Carayon P (2006) Human factors of complex sociotechnical systems. Appl Ergon 37(4):525–535
5. Carmona J, van Dongen B, Solti A, Weidlich M (2018) Conformance Checking. Springer, Cham
6. EFFRA European Factories of the Future Research Association (2016) Factories 4.0 and Beyond: Recommendations for the Work Programme 18-19-20 of the FoF PPP under Horizon 2020
7. HUMAN Manufacturing (2019) http://humanmanufacturing.eu/, last access: 27.8.2019
8. Janiesch C, Koschmider A, Mecella M, Weber B, Burattin A, Di Ciccio C, Gal A, Kannengiesser U, Mannhardt F, Mendling J, Oberweis A, Reichert M, Rinderle-Ma S, Song WZ, Su J, Torres V, Weidlich M, Weske M, Zhang L (2017) The Internet-of-things meets business process management: mutual benefits and challenges. Arvix Preprint: arXiv:1709.03628 [cs.CY]
9. Knoch S, Ponpathirkoottam S, Fettke P, Loos P (2018) Technology-enhanced process elicitation of worker activities in manufacturing. In: Business Process Management Workshops. Springer, Cham, pp 273–284
10. de Looze MP, Bosch T, Krause F, Stadler KS, O'Sullivan LW (2015) Exoskeletons for industrial application and their potential effects on physical work load. Ergonomics 59(5):671–681
11. Mannhardt F, Bovo R, Oliveira MF, Julier S (2018) A taxonomy for combining activity recognition and process discovery in industrial environments. In: IDEAL 2018. Springer, Cham, pp 84–93
12. Mannhardt F, Petersen SA, Oliveira MF (2018) Privacy challenges for process mining in human-centered industrial environments. In: Intelligent Environments (IE). IEEE Xplore: 64–71
13. Mannhardt F, Petersen SA, Oliveira MF (2019) A trust and privacy framework for smart manufacturing environments. J Ambient Intell Smart Env 11(3):201–219
14. Thoben KD, Weisner SA, Wuest T (2017) Industrie 4.0 and smart manufacturing – A review of research issues and application examples. Int J Autom Technol 11(1):4–16

Process Mining and User Privacy in D2D and IoT Networks

Muhammad Usman · Marwa Qaraqe
Muhammad Rizwan Asghar
Imran Shafique Ansari

As the communication industry is moving towards the 5th generation (5G) of cellular networks and beyond, the traffic it carries is also becoming a mixture of higher and lower data rate rate traffic originating from cellular users and Internet-of-Things (IoT) networks, respectively. The main industries where IoT has found its applications include, but are not limited to, automotive, manufacturing, supply chain, agriculture, healthcare, and energy. All the aforementioned industries have their specific quality of service (QoS) requirements in terms of bandwidth, latency, and storage regarding the transmission of the data they generate. For instance, vehicle-to-everything (V2X) communication in the automotive industry requires ultra-low latency without any need for higher bandwidth; however, the IoT networks employed in agriculture generally require lower latency and lower bandwidth communication.

In order to accommodate such a diverse requirement in 5G networks, many solutions have been proposed to tackle the problem from different angles. For instance, network functions virtualization (NFV) and software-defined networking (SDN) have been proposed as possible solutions, designed to decouple network services from the hardware they are executed upon. These technologies are also termed virtualization of network resources, which are meant to accommodate diverse QoS requirements onto a single physical network. Although these solutions are promising, there is a long way to go towards end-to-end virtualization of cellular networks.

In the literature, some researchers [1, 2] propose employing device-to-device (D2D) communication in the cellular access network to accommodate many QoS requirements of IoTs. For the uplink, D2D nodes can act as communication hubs to collect slower data from an IoT network and transfer it to the Cloud over cellular communication for storage and processing. For instance, a D2D node can collect data from various appliances in a smart home and transfer it to a remote cloud server via cellular or WiFi network. Similarly, for the downlink, D2D nodes can act as a caching device for many IoT applications to reduce latency [2]. In addition, D2D communication can also be used as an enabler of many proximity-based applications, such as peer-to-peer communication, proximity-based social networking, and proximity-based advertisement broadcasting.

All the aforementioned applications generate huge amounts of data, which a cellular network

https://doi.org/10.1007/s00287-019-01212-y

Muhammad Usman · Marwa Qaraqe
Information and Computing Technology,
College of Science and Engineering,
Hamad Bin Khalifa University (HBKU),
Education City, 34110 Doha, Qatar
E-Mail: {musman, mqaraqe}@hbku.edu.qa

Muhammad Rizwan Asghar
School of Computer Science, The University of Auckland,
1142 Auckland, New Zealand
E-Mail: r.asghar@auckland.ac.nz

Imran Shafique Ansari
School of Engineering, University of Glasgow,
G12 8QQ, UK
E-Mail: imran.ansari@glasgow.ac.uk

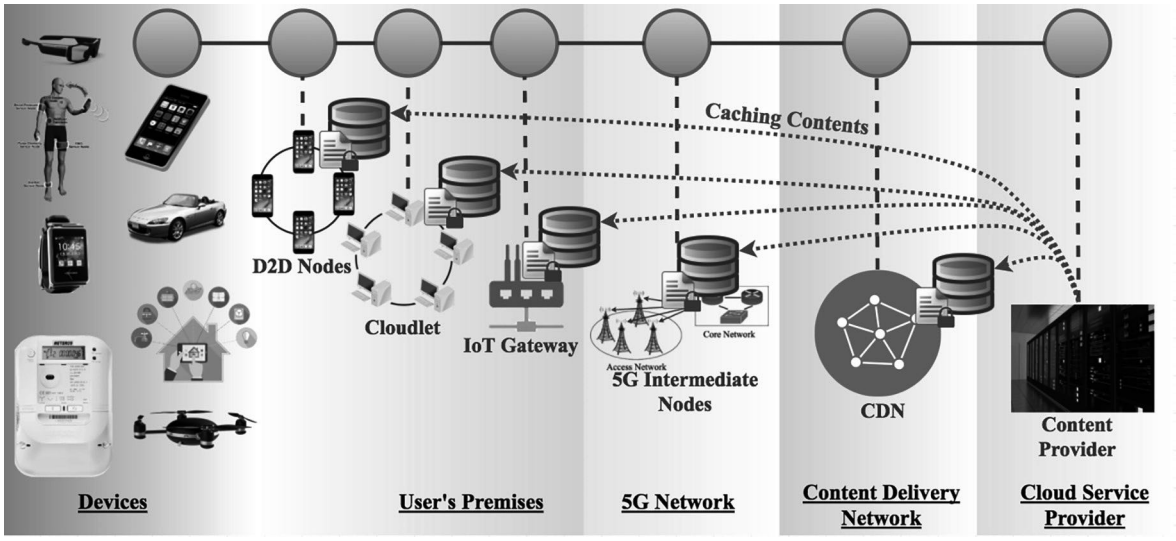


Fig. 1 Different caching options for a content provider: The contents can be cached at any feasible option between the content provider and the end user depending on the QoS requirement of the user and availability of resources at different caching servers [2]

and/or D2D node may utilize to prioritize traffic to ensure QoS for different IoT applications. This data may contain some information that may be personal to someone, who may not want to reveal it. For instance, an advertising company may log all the events while a customer visits a particular shop. It also logs the time and date when the customer was near the shop. It can even log the duration a customer stayed near the shop. If the shop is on the way to the office, the events are logged on a daily basis. Carrying out process mining on such events could expose sensitive information that may reveal the personal habits of a person (e. g., going to the office late orbeing absent on a particular day).

Additionally, as D2D communication extends the caching option right at the user's proximity, the cellular network needs to know exactly what a user is looking for at a particular time in order to cache the content in D2D networks. This becomes true for all the caching options presented in Fig. 1. For instance, depending on the requirements, a content provider may place the content across a CDN (content delivery network), a 5G intermediate node, an IoT gateway, or cloudlet or D2D nodes. However, placing the content within cellular network premises (5G intermediate node to D2D nodes) will allow the cellular network to log all the events, i. e., data accesses. Moreover, this type of content caching enables numerous applications that try to save bandwidth across 5G cellular networks. For instance, if two users are accessing the same contents over the Internet, and they are in the immediate proximity, the cellular network will only provide contents to the single user with the better channel conditions (say user 1), and the other user (say user 2) will be served by the first one. It is worth mentioning that user 2 will not know that she is accessing the content from user 1, and, similarly, user 1 will be unaware of her delivery of content to user 2.

It is worth mentioning that a lot of personal information can be exposed to process miners in all these scenarios. Process mining is a technique that takes as an input the event logs and records of the sequence of steps and discovers a process of the model to expose personal information. However, in the light of the above discussion regarding data generation and processing in IoT-based D2D networks, we believe that process mining can be employed to reveal personal information of customers using the services of a particular organization, such as a cellular network or an IoT services provider.

In previous work [2], we proposed a convergent encryption-based solution to ensure security in caching environments. However, we stress that similar approaches can be employed to ensure anonymity in IoT networks that

use D2D either as a caching server or as data distribution hub.

Acknowledgements

Funding. Open access funding provided by the Qatar National Library.

References

1. Usman M, Asghar MR, Ansari IS, Granelli F (2017) Towards bootstrapping trust in D2D using PGP and reputation mechanism. In: 2017 IEEE International Conference on Communications (ICC). IEEE, pp 1–6
2. Usman M, Asghar MR, Ansari IS, Granelli F, Abbasi QH, Qaraqe K (2018) A marketplace for efficient and secure caching for IoT applications in 5G networks. In: 2018 IEEE Wireless Communications and Networking Conference (WCNC). IEEE, pp 1–6

Post-Quantum Cryptography and its Application to the IoT

(Extended Abstract)

Juliane Krämer

Cryptography for the IoT

In our everyday lives, many applications have to be secured with cryptography. The need for cryptography rises drastically especially through the Internet of Things (IoT). Billions of heterogeneous devices are interconnected. If IoT applications are not secured, not only is our privacy threatened, and financial loss may be expected, but it may even cause danger to life and limb. Thinking of autonomous cars, for instance, multiple collisions can be easily imagined. Also cars with a human driver are at risk, since modern cars are connected to the Internet and use several digital tools that can be attacked, as was shown in the Fiat-Chrysler hack. Hackers remotely hijacked a Jeep's digital system, enabling them to cause, among other things, unintended acceleration or turning the vehicle's steering wheel at any speed.

Another example of a system that has to be secured to protect our lives is telemedicine. Unprotected transmission of medical data can allow an attacker to modify the transferred data so that a seriously ill patient appears to be in good health to the treating physician, so that the patient does not receive the necessary medical treatment. Using telemedicine without sufficient security can also threaten our privacy, for instance when the transmission is tapped and someone who should not know about them learns about our health data. Hence, communication channels and IoT-centric communication protocols have to be secured.

Further privacy risks can stem from the usage of Smart Home solutions such as a smart TV. When the camera and the microphone are not sufficiently protected, an attacker can access them and thereby observe and eavesdrop on what happens in front of the TV and in the room, respectively.

Thinking only for some minutes about potential risks that can occur through the usage of unsecured IoT applications leads to many scenarios where our privacy or our life and limb are threatened, and where financial loss may be expected. Hence, all IoT devices and applications have to be secured, i. e., all functionalities and transmissions of data have to be protected such that no unauthorized person can read the data and no one can unnoticedly modify the transmitted data.

Public-key cryptography

These two properties of secured systems – that no unauthorized person can read the data and that no one can modify the transmitted data unnoticedly – are two of the main goals of cryptography. They are called confidentiality and integrity. The third important goal that can be achieved with cryptography is authenticity. Authenticity means that the sender of data cannot be forged, i. e., it is confirmed that the data come from the stated sender. Authenticity is, for instance, important for software updates where we trust that the operating system's producer is the sender, if this is stated in the update. If this were not ensured, it would be easy for any attacker to install malicious code on our devices.

The fact that we benefit not only from confidentiality in our modern, highly-connected IT systems, but also from integrity and authenticity, stems from the development of public-key cryptography (PKC) in the mid-1970s. Contrary to secret-key cryptogra-

https://doi.org/10.1007/s00287-019-01200-2

Juliane Krämer
TU Darmstadt, Darmstadt
E-Mail: jkraemer@cdc.informatik.tu-darmstadt.de

phy, where both sender and receiver use the same key, in PKC each participant of a network has a pair of keys, i. e., a public key and a private key. While the public key can be published (for instance, on a website), the private key has to be kept secret. When someone wants to send encrypted data to a receiver (thus, to modify them so that no one can understand the modified data), he uses the public key of the receiver for the encryption and sends the ciphertext, i. e., the encrypted data, over a potentially insecure network. Only the receiver with his private key can then decrypt the ciphertext and gain the plaintext, i. e., the original data. Similarly, it is possible to protect data with a digital signature that yields both integrity of data and authenticity of the sender. Both for public-key encryption and digital signatures it is paramount that it is computationally infeasible for an attacker to compute the private key from the public key. In currently used public-key schemes, such as RSA or elliptic curve cryptography, this is ensured by basing the security of the schemes on two mathematical problems which are assumed to be computationally hard, meaning that it is assumed that they cannot be solved efficiently, i. e., in polynomial time. These problems are the problem of factoring a large number into its prime factors and the problem of computing the discrete logarithm.

Post-quantum cryptography

Since the mid-1990s it has been known that with a quantum computer, both the factorization problem and the discrete logarithm problem can be solved efficiently [2]. Hence, it has been known for 25 years that we need replacements for currently used public-key cryptography as soon as large quantum computers exist. These replacements have to provide the same functionalities as classical PKC but rely on mathematical problems which resist attacks with quantum computers. Cryptography that is based on mathematical problems which are assumed to resist attacks with quantum computers is called post-quantum cryptography (PQC). Several types of PQC exist. The five most important types are lattice-based, hash-based, isogeny-based, code-based, and multivariate cryptography, which all have certain advantages, but also disadvantages. Hence, post-quantum cryptography is a very active field of research. Although no one knows when exactly large quantum computers will exist, many estimations range between 10 and 20 years from now (e. g. [3]). Hence, the era in which we have to use PQC is coming up – and data with very sensitive information, in fact, should be encrypted with PQC already today, since otherwise the encrypted data can be stored and decrypted once large quantum computers exist. One avenue to promote using PQC is the standardization of post-quantum algorithms, since the standardization of cryptography makes the adoption of new schemes easier for practitioners. The best known PQC standardization effort is taking place at NIST, the US-American National Institute of Standards and Technology [1].

Post-quantum cryptography for the IoT

When it comes to the Internet of Things, however, the schemes that are currently in the process of standardization by NIST might be inappropriate. These schemes were designed for desktop and server environments. IoT applications, in contrast, use resource-constrained devices which are limited in terms of, e. g., power, processing time, and memory consumption. Hence, IoT applications impose particular requirements on cryptographic schemes and existing post-quantum algorithms have to be adapted such that they meet these requirements. Since, in general, there is a tradeoff between security and efficiency, the security of the modified schemes will decrease. There may still be enough security, since many IoT applications do not need high-level security. However, there is also a tradeoff between performance and resource requirements, and the performance of the modified schemes may not be acceptable once they meet the resource requirements of IoT devices. Hence, it might be necessary to develop post-quantum algorithms tailored for resource-constrained devices, i. e., dedicated light-weight post-quantum algorithms for the IoT.

Unfortunately, even the use of classical cryptography is an enormous problem in the context of IoT; too many IoT applications do not use any cryptography at all and do not provide any security. Hence, on the way to a post-quantum secure Internet of Things, there is still a lot to be done.

References

1. NIST CSRC (2017) Post-quantum cryptography standardization – Post-quantum cryptography. https://csrc.nist.gov/projects/post-quantum-cryptography/post-quantum-cryptography-standardization
2. Shor (1994) Algorithms for quantum computation: discrete logarithms and factoring. FOCS 1994:124–134
3. de Touzalin M, Heijman C, Murray C (2016) Quantum Manifesto – A New Era of Technology

We Value Your Privacy ... Now Take Some Cookies[1]

Measuring the GDPR's Impact on Web Privacy

Martin Degeling · Christine Utz
Christopher Lentzsch · Henry Hosseini
Florian Schaub · Thorsten Holz

Problem description

On May 25, 2018, the General Data Protection Regulation (GDPR) went into effect in the European Union. The GDPR is supposed to set high and consistent standards for the processing of personal data within the European Union and whenever personal data of people residing in Europe is involved.

As a result, the GDPR affects millions of web services from around the world, which are available in Europe. In addition to potentially changing how they process personal data, companies have to disclose transparently how they handle personal data, the legal bases for their data processing, and need to offer their users mechanisms for individual consent, data access, data deletion, and data portability. Even outside Europe, online services had to prepare for the GDPR because it not only applies to companies in Europe but any company that offers its service in Europe. As a result, the GDPR is expected to have a major impact on companies across the world.

Previous work found that about 70 to 80 % of websites in the US have privacy policies [4, 5]. However, the analysis of privacy policies has been focused on English policies, performing in-depth studies on their content [1, 3, 6, 7]. Cookie consent notices have just recently seen research attention with respect to their usability [2], but their use and implementations have not been studied in detail.

We describe an empirical study to measure the actual impact of the GDPR on a representative set of websites. We monitored this rare event by analyzing the 500 most-visited websites, according to Alexa country rankings, in each of the 28 member states of the EU over the course of 6 months. In total, this resulted in a set of 6759 websites available in 24 different languages.

We used a combination of automated and manual methods and compared the privacy policies of these websites before and after the GDPR effective date and, together with historic data, retrieved 112 041 privacy policies.

Results

We conduct an empirical, longitudinal study of privacy policies and cookie consent notices of 6759 websites representing the 500 most popular websites in each of the 28 member states of the EU.

We performed monthly scans to measure changes in adoption rates. Between January and the end of May, we observed an average rise of websites providing privacy policies by 4.9 % and cookie consent notices by 16 %.

While prior studies primarily focused on English-language privacy policies, we analyze privacy policies in 24 different languages We use natural language processing techniques to iden-

[1] This paper was presented as Degeling, Martin, Christine Utz, Christopher Lentzsch, Henry Hosseini, Florian Schaub, and Thorsten Holz. "We Value Your Privacy . . . Now Take Some Cookies: Measuring the GDPR's Impact on Web Privacy." In *Proceedings of the NDSS 2019*. San Diego, CA: Internet Society, 2019. https://doi.org/10.14722/ndss.2019.23378.

https://doi.org/10.1007/s00287-019-01201-1

Martin Degeling · Christine Utz · Christopher Lentzsch
Henry Hosseini · Thorsten Holz
Ruhr-Universität Bochum, Bochum, Germany
E-Mail: {martin.degeling, christine.utz, christopher.lentzsch, henry.hosseini, thorsten.holz}@rub.de

Florian Schaub
University of Michigan, Ann Arbor, MI, USA
E-Mail: fschaub@umich.edu

tify how privacy policies' content has changed and whether the GDPR's new transparency requirements are reflected in the texts. We find that not too many websites make use of GDPR terminology, but for those that do, the amount of information about users' rights and the legal basis of processing have increased.

We compare the use of cookies and third-party libraries in our set of websites between January and June 2018 to determine whether the GDPR's transparency and consent requirements affected the prevalence of web tracking. While neither was significantly impacted, 147 sites stopped using tracking libraries, and 37 chose to ask for explicit consent before activating them.

We categorize observed cookie consent notices based on their options for interaction. In our data set, we found many distinct implementations of cookie consent notices. We analyze these libraries for key features required to implement the GDPR notion of "informed consent" and identify technical obstacles to achieving this goal.

References

1. Harkous H, Fawaz K, Lebret R et al (2018) Polisis: automated analysis and presentation of privacy policies using deep learning. arXiv:180202561 [cs]
2. Kulyk O, Hilt A, Gerber N, Volkamer M (2018) "This Website Uses Cookies": Users' Perceptions and Reactions to the Cookie Disclaimer. In: 3rd European Workshop on Usable Security. Internet Society, London, England. https://doi.org/10.14722/eurousec.2018.23012
3. Libert T (2018) An Automated Approach to Auditing Disclosure of Third-Party Data Collection in Website Privacy Policies. In: Proceedings of the 2018 World Wide Web Conference, pp 207–216. International World Wide Web Conferences Steering Committee, Republic and Switzerland. https://doi.org/10.1145/3178876.3186087
4. Liu C, Arnett KP (2002) Raising a red flag on global WWW privacy policies. J Comp Inf Syst 43:117–127. https://doi.org/10.1080/08874417.2002.11647076
5. Nokhbeh Zaeem R, Barber KS (2017) A study of web privacy policies across industries. J Inf Priv Secur 1–17. https://doi.org/10.1080/15536548.2017.1394064
6. Tesfay WB, Hofmann P, Nakamura T et al (2018) PrivacyGuide: Towards an implementation of the EU GDPR on internet privacy policy evaluation. In: Proceedings of the Fourth ACM International Workshop on Security and Privacy Analytics. ACM, New York, pp 15–21
7. Wilson S, Schaub F, Dara A et al (2016) The Creation and Analysis of a Website Privacy Policy Corpus. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, Berlin, Germany. https://doi.org/10.18653/v1/P16-1126

User Centered and Privacy-Driven Process Mining System Design[1]

(Extended Abstract)

Judith Michael · Agnes Koschmider
Felix Mannhardt · Nathalie Baracaldo
Bernhard Rumpe

Problem description

Process mining uses event data recorded by information systems to reveal the actual execution of business processes in organizations. Since most activities in modern organizations are supported by technology, each process execution leaves behind a digital trace (i. e., event log) indicating the occurrence and timing of activities in the databases of the company. This way, the discovered process model from event logs can expose performance information, bottlenecks, workarounds, and much more. Events and traces may contain sensitive information pertaining to the data provider and are accessible to data controller(s) and data consumer(s), who demand protection. Due to GDPR, organizations are obliged to consider privacy throughout the complete development process, which also applies to the design of process mining systems. To a certain degree, process mining methods already abstract from such privacy related details by deriving a process model that reveals only the sequences of activity execution observed. However, often occurrence frequencies, performance information, and decision rules are discovered in addition to the basic control-flow of the process, which may leak additional information from the event log. By discovering several process models and slightly varying the filtering condition it is possible to identify workers. Obviously, privacy preservation should be taken into account for process mining.

Use scenario: Privacy and IoT manufacturing tasks

In the past, process mining has been hampered by the fact that processes are often incomplete or erroneous. With the Internet of Things (IoT) producing a large amount of data stored, more data become available for analysis, possibly resolving issues of incompleteness [2]. IoT is a domain with a high demand for privacy and security considerations. The large amount of data that is tracked and analyzed with, e. g., learning (AI) software, can originate from internet-enabled machines, working modules labeled with QR-code, and workers equipped with wearables such as smart watches, interacting as autonomous agents forming a complex system. In the context of IoT, GDPR relates to the privacy compliance of a large number of attributes such as GPS location, working time, and salary. From this data, the working practices and performance of workers can be inferred, which may be considered very sensitive information [3].

[1] This paper is in press as Michael, Koschmider, Mannhardt, Baracaldo, Rumpe: User Centered and Privacy-Driven Process Mining System Design for IoT. 31st International Conference on Advanced Information Systems Engineering (CAiSE) Forum, 2019.

https://doi.org/10.1007/s00287-019-01202-0

Judith Michael · Bernhard Rumpe
RWTH Aachen University, Software Engineering, Aachen, Germany
E-Mail: {michael, rumpe}@se-rwth.de

Agnes Koschmider
Karlsruhe Institute of Technology, AIFB, Karlsruhe, Germany
E-Mail: agnes.koschmider@kit.edu

Felix Mannhardt
SINTEF Digital, Trondheim, Norway
E-Mail: Felix.Mannhardt@sintef.no

Nathalie Baracaldo
IBM Almaden Research Center, San Jose, USA
E-Mail: baracald@us.ibm.com

Results

To ensure user-centered privacy for process mining in the IoT context, we suggest a system design relying on privacy policies. First, a context meta-model [4] is adopted to the IoT use case, which is used as schema for data storage. Next, the context meta-model is enriched with privacy concepts and process mining concerns captured in a privacy preserving meta-model. Lastly, an architectural model is designed allowing to monitor the compliance of policies. The user-centered privacy-driven system relies on the eXtensible Access Control Markup Language (XACML), which we adapted for our purpose. The language allows evaluating access requests of data consumers according to the rules defined in policies (between, e. g., the data provider and data controller) with the notions of the policy enforcement point (PEP), policy decision point (PDP), policy information point (PIP), and the policy administration point (PAP). Besides these common components, our system architecture consists of an information portal, a data collection engine, and an obligation engine. The objective of an information portal is to provide a user-friendly representation of stored data, data access attempts, and the management of policies, and to foresee privacy preservation strategies for each stage. The objective of the data collection engine is to collect data from heterogeneous data sources and to link them to attributes and persons. The obligation engine is responsible for keeping track of obligation triggers. This system architecture allows us to (a) define and manage privacy policies, (b) determine more accurately who can do what with which data, (c) monitor compliance, and (d) preview which privacy mechanisms are foreseen in which stages of process mining. Also, the system design supports the eight privacy design strategies [1], which means that the system design ensures a higher level of privacy protection from the perspective of users than existing approaches.

References

1. Hoepman JH (2014) Privacy design strategies. In: Cuppens-Boulahia N, Cuppens F, Jajodia S, Abou El Kalam A, Sans T (eds) ICT systems security and privacy protection. Springer, Berlin Heidelberg, pp 446–459
2. Janiesch C, Koschmider A, Mecella M, Weber B, Burattin A, Ciccio CD, Gal A, Kannengiesser U, Mannhardt F, Mendling J, Oberweis A, Reichert M, Rinderle-Ma S, Song W, Su J, Torres V, Weidlich M, Weske M, Zhang L (2017) The internet-of-things meets business process management: Mutual benefits and challenges. CoRR abs/1709.03628
3. Mannhardt F, Bovo R, Oliveira MF, Julier S (2018) A taxonomy for combining activity recognition and process discovery in industrial environments. In: Intelligent data engineering and automated learning. LNCS. Springer, Berlin, Heidelberg, pp 84–93
4. Michael J, Steinberger C (2017) Context modeling for active assistance. In: ER Forum and the ER Demo Track, pp 221–234

Privacy-preserving Process Mining: Differential[1]

Privacy for Event Logs (Extended Abstract)

Felix Mannhardt · Agnes Koschmider
Nathalie Baracaldo · Matthias Weidlich
Judith Michael

Problem description

The GDPR defines personal data as "any information relating to an identified or identifiable natural person" (referred to as data provider) [1]. Privacy protection goes further than security and regulates the authorized access to data based on a lawful basis (e. g., it may be based on consent or on legal requirements such as auditing) and organizational measures that should build trust between the individual (i. e., the data provider), the entity that processes and stores the data (referred to as the data controller), and entities who use or bought the data (referred to as the data consumer). Process mining uses event data recorded by information systems to reveal the actual execution of business processes in an organization. Since most activities in modern organizations are supported by technology, each process execution leaves behind a digital trace indicating the occurrence and timing of activities in the databases of the company. Process mining takes event logs, records of the sequence of steps, and discovers a de-facto model of the process that can expose performance information, bottlenecks, workarounds, and much more. This way, events and traces may contain sensitive information pertaining to the data provider and being accessible to the data controller(s) and the data consumer(s). To a certain degree process mining methods already abstract from such privacy-related details by deriving a process model that reveals only the sequences of activity execution observed. However, often occurrence frequencies, performance information, and decision rules are discovered in addition to the basic control-flow of the process, which may leak additional information from the event log. Furthermore, process mining is often an iterative process in which multiple process models for different subsets of the event log, filtered according to conditions of interest, are discovered and compared. By discovering several process models and slightly varying the filtering conditions it is possible to identify workers. Obviously, privacy preservation should be taken into account for process mining.

Starting from common assumptions on the event logs used in process mining, we study potential privacy leakages and means of protection from them. We show how to exploit the notion of privacy checkpoints to identify potential issues in the design of system that shall preserve privacy in process mining. Based on this analysis, we propose an approach for protection against secondary use of

[1] This paper is in press as Mannhardt, Koschmider, Baracaldo, Weidlich, Michael: Privacy-preserving Process Mining: Differential Privacy for Event Logs. Business & Information System Engineering. Special Issue Data Sovereignty and Data Space Ecosystems, 5/2019.

https://doi.org/10.1007/s00287-019-01207-9

Felix Mannhardt
SINTEF Digital,
Trondheim, Norway
E-Mail: felix.mannhardt@sintef.no

Agnes Koschmider
Group Process Analytics, Kiel University,
Kiel, Germany
E-Mail: ak@informatik.uni-kiel.de

Nathalie Baracaldo
IBM Almaden Research Center,
San Jose, USA
E-Mail: baracald@us.ibm.com

Matthias Weidlich
Humboldt-Universität zu Berlin,
12489 Berlin, Germany
E-Mail: matthias.weidlich@hu-berlin.de

Judith Michael
RWTH Aachen University, Software Engineering,
Aachen, Germany
E-Mail: michael@se-rwth.de

event logs. To this end, we lift the well-established notion of differential privacy to the specific model of event logs. We instantiate this notion for process discovery methods, i. e., algorithms that aim at the construction of a process model from an event log. The general feasibility of our approach is demonstrated by its application to two publicly available real-life events logs.

Privacy issues for process mining of healthcare processes

Hospital information systems have a high demand for privacy and security considerations, since electronic health records need privacy protection. While we use a hospital use case to motivate our approach, there are a many similar situations in which organizations have centralized control over an event log and want to protect the privacy of individuals for which cases are processed. To understand at which stage of data passes a protection model for event log privacy is required, we apply the privacy checkpoint diagram from [3]. According to this privacy checkpoint diagram, data passes six stages within healthcare processes. (1) the data source: given our use case, the sources of data originate from the medical stage, the administrative stage, and patients. We refer to this data as personal data. (2) Data capture: data from these data sources is captured when devices and systems log tasks of the medical stage, the administrative stage, and patients, when recognizing the identity or requesting actions. Since this stage tracks who does what, when, and where with data, anonymization techniques should be used here for protection against disclosure of sensitive events. (3) Primary use: the hospital determines the purposes for which and the means by which the captured data is processed. For instance, the captured data can be used to support the work of medical and administrative stages for the diagnosis or treatment of patients. (4) Data storage: personal data and events of medical and administrative stages, and patients are stored in a database or event logs. The data might be processed by data mining approaches aiming to address performance indicators such as the number of pancreas operations, the length of waiting lists, or the success rate of surgeons. (4) Data (re)use: at this stage, data from event logs is used for process mining aiming to determine the main paths that are followed by patients or medical stage in the process. Such analysis demands privacy techniques to protect personally identifiable records in event logs. Personal data might also be retrieved from third-party sources such as public databases or other hospitals, which obviously triggers a GDPR requirement (i. e., demonstration that the data was retrieved in compliance with GDPR regulations). Compliance is a central concern in the context of hospital processes [4]. At this stage, data from several sources is required, which increases the number of leakages. (5) Data removal: raw data is permanently deleted. We consider privacy checkpoint 4 and stage data (re)use as points of privacy leakage for process mining.

Protection model

The strongest privacy model available to date which provides provable privacy guarantees is differential privacy [2]. Therefore, the protection model presented in this paper relies on differential privacy. Differential privacy establishes a theoretical limit on the influence of a single row on a dataset (e. g., an individual's data), thus limiting an attacker's ability to infer such membership. Typically, noise is added proportionally to the sensitivity of the output. Sensitivity measures the maximum change of the output due to the inclusion of a single data instance. Given the fact that in the use case of a hospital, access to patient and hospital records exists, we assume a centralized privacy approach to protect the data (re)use of event data for process mining using differential privacy. Please note that this centralized approach to handling privacy would also be possible in many other scenarios with a centralized data management, e. g., in public administration. Our protection model works as follows. The environment is divided into a trusted environment, in which data is processed to provide the primary services of the hospital (primary use) in accordance with the consent of patients and stage (data sources). Additionally, the captured sensitive data is stored as an event log in a protected data storage for later analysis with process mining methods. Up to the data storage stage we rely on organizational and technological measures (e. g., access control, encryption) to fully protect the privacy of stakeholders. The main idea of the envisioned protection model is to guarantee differential privacy for the data providers. We introduce a privacy engine, which acts as the single point of access for process mining algorithms. All data required by the algorithms needs to be queried according to a set of restricted query operations. This privacy

engine resides in the trusted environment and introduces noise to each query result such that differential privacy guarantees are maintained at all times.

References

1. D'Acquisto G, Domingo-Ferrer J, Kikiras P, Torra V, de Montjoye YA, Bourka A (2015) Privacy by design in big data: An overview of privacy enhancing technologies in the era of big data analytics
2. Dwork C (2008) Differential privacy: A survey of results. In: International Conference on Theory and Applications of Models of Computation. Springer, pp 1–19
3. Mannhardt F, Petersen S, de Oliveira MFD (2018) Privacy challenges for process mining in human-centered industrial environments. In: 14th International Conference on Intelligent Environments (IE). IEEE, Xplore, pp 64–71
4. Mans RS, van der Aalst WMP, Vanwersch RJB, Moleman AJ (2013) Process mining in healthcare: Data challenges when answering frequently posed questions. In: Process Support and Knowledge Representation in Health Care. Springer, Berlin Heidelberg, pp 140–153

PRETSA: Event Log Sanitization for Privacy-aware Process Discovery[1]

(Extended Abstract)

Stephan A. Fahrenkrog-Petersen
Han van der Aa · Matthias Weidlich

Information systems record data in the form of event logs, while executing business processes. Event logs can, therefore, be used for data-driven analysis of business processes. In recent years, various such analysis techniques have been proposed under the umbrella of process mining [1]. For instance, techniques for process discovery construct a model of a business process from an event log (see [2]), which can then be enriched with performance information for quantitative process analysis.

However, logs potentially contain sensitive information about individual employees involved in process execution. Due to legal frameworks such as the General Data Protection Regulation (GDPR), organizations are obliged to ensure a certain level of privacy and to protect the personal data of individuals [3]. In many scenarios, obfuscation of the event log is not sufficient to achieve such data protection. Rather, one has to rely on explicit approaches for data sanitization, which provide privacy guarantees through data transformation mechanisms. Data sanitization is typically lossy, meaning that the utility of the data for some analysis task is hampered. Therefore, it is necessary to develop techniques that achieve privacy, but preserve as much utility as possible for the analysis task at hand.

This work introduces a data sanitization technique suited for event logs used to discover performance-annotated process models. Specifically, we consider a trace linking attack on an event log with pseudonymized employee information that correlates events of the log with background knowledge on possible activity assignments during process execution. For this setting, we present PRETSA (PREfix-Tree based event log SAnitisation for t-closeness), a sanitization technique that guarantees privacy in terms of k-anonymity and t-closeness for the transformed event log. It thereby avoids disclosure of employee identities, their membership in the event log, and their characterization based on sensitive attributes, such as performance information. PRETSA takes up ideas on achieving k-anonymity for sequential data [4]. In essence, PRETSA constructs a prefix tree representation of an event log that is annotated with frequencies and attribute values. This tree is then step-wise transformed; subtrees are merged and relocated until the required privacy guarantees have been obtained. The resulting log transformations are comparatively fine granular. As a consequence, the log's utility for discovery of a performance-annotated process model is largely preserved.

Experiments with three real-world data sets demonstrate that sanitization with PRETSA yields

[1] This paper is in press as: Stephan A. Fahrenkrog-Petersen, Han van der Aa, Matthias Weidlich: PRETSA: Event Log Sanitization for Privacy-aware Process Discovery. 1st International Conference on Process Mining (ICPM), June 24–26, 2019, Aachen, Germany.

https://doi.org/10.1007/s00287-019-01203-z

Stephan A. Fahrenkrog-Petersen · Han van der Aa
Matthias Weidlich
Humboldt-Universität zu Berlin, Institut für Informatik, Berlin
E-Mail: {stephan.fahrenkrog-petersen, han.van.der.aa, matthias.weidlich}@hu-berlin.de

event logs of higher utility compared to methods that exploit frequency-based filtering, while providing the same privacy guarantees. In some cases, frequency-based filtering is not even able to provide any event log that fulfills the requested privacy guarantees. PRETSA, in turn, was always able to provide a sanitized event log with high utility. The latter is reflected in the amount of preserved process variants, the fitness of the resulting process model with the event log, and the deviation in the generated performance annotations compared to the annotations generated based on the original event log.

References

1. Van der Aalst WMP (2016) Process Mining – Data Science in Action. Springer, Berlin Heidelberg
2. Augusto A, Conforti R, Dumas M, La Rosa M, Maggi FM, Marrella A, Mecella M, Soo A (2019) Automated discovery of process models from event logs: Review and benchmark. IEEE T Knowl Data Eng 31(4):686–705
3. Mannhardt F, Petersen SA, Oliveira MF (2018) Privacy challenges for process mining in human-centered industrial environments. In: 2018 14th International Conference on Intelligent Environments (IE). IEEE, pp 64–71
4. Monreale A, Pedreschi D, Pensa RG, Pinelli F (2014) Anonymity preserving sequential pattern mining. Artif Intell Law 22(2):141–173

Business Process Privacy Analysis in Pleak[1]

(Extended Abstract)

**Aivo Toots · Reedik Tuuling
Maksym Yerokhin · Marlon Dumas
Luciano García-Bañuelos · Peeter Laud
Raimundas Matulevičius
Alisa Pankova · Martin Pettai
Pille Pullonen · Jake Tom**

Gap

The Business Process Model and Notation (BPMN) is a standard notation for capturing enterprise-wide and interorganizational business processes. BPMN models are commonly used to analyze the compliance of business processes with respect to regulations [2] as well as their performance with respect to efficiency and quality criteria [3].

A less exploited application of BPMN models is to analyze compliance with respect to privacy regulations, in great part because of BPMN's inability to capture privacy requirements and privacy-enhanced technologies (PETs), such as secure multiparty computation, differential privacy, and homomorphic encryption. Currently, there is no approach available to analyze the way private data is used (and potentially leaked) along entire business processes in the presence of PETs.

Approach

To address this gap, we have designed an extension of BPMN, namely privacy-enhanced BPMN (PE-BPMN), that allows users to capture private data objects, privacy policies, and PETs alongside the tasks, decisions, and other elements of a business process model [4].

To support the design and analysis of PE-BPMN models, we have developed a tool, namely Pleak [5], that allows users to identify privacy leakages in PE-BPMN models, to quantify the extent of these leakages, to determine which PETs can be used to eliminate or reduce these leakages, and to analyze the trade-offs that these PETs bring along, for example in terms of loss of accuracy of the disclosed data.

Pleak supports the analysis of PE-BPMN models at three levels. The top level (Boolean analysis) tells us whether or not a given output of a process may reveal information about a given input. The middle level (qualitative analysis), goes further by indicating which attributes of (or functions over) a given input data object are potentially leaked by each output, and under what conditions this leakage may occur. The lower level (quantitative analysis) quantifies to what extent a given output leaks information about an input, either in terms of a sensitivity (differential privacy) measure or in terms of the guessing advantage that an attacker gains by having the output. The Boolean level is based on the high-level flow of the data and privacy-enhancing technologies that are deployed to avoid leakages. In the qualitative and quantitative levels, Pleak relies on the specification of the actual computations carried out by the steps in the business process, as well as the definitions of data structures. Quantitative analyses output both the

https://doi.org/10.1007/s00287-019-01204-y

Aivo Toots · Reedik Tuuling · Peeter Laud · Alisa Pankova
Martin Pettai · Pille Pullonen
Cybernetica, Tartu, Estonia
E-Mail: {aivo.toots, reedik.tuuling, peeter.laud, alisa.pankova, martin.pettai, pille.pullonen}@cyber.ee

Maksym Yerokhin · Marlon Dumas · Luciano García-Bañuelos
Raimundas Matulevičius · Jake Tom
University of Tartu, Tartu, Estonia
E-Mail: {maksym.yerokhin@ut.ee, marlon.dumas, luciano.garcia-banuelos, raimundas.matulevicius, jake.tom}@ut.ee

[1] This paper is in press as Aivo Toots, Reedik Tuuling, Maksym Yerokhin, Marlon Dumas, Luciano García-Bañuelos, Peeter Laud, Raimundas Matulevičius, Alisa Pankova, Martin Pettai, Pille Pullonen, Jake Tom: Business Process Privacy Analysis in Pleak. International Conference on Fundamental Approaches to Software Engineering (FASE), 2019.

expected privacy risk and the necessary measures to overcome the leakage. In addition, they highlight the trade-off of a private process versus an accurate process. Furthermore, Pleak can also make use of the actual data used in the process (or data samples) if available.

With respect to previous research proposals on privacy analysis of business processes [1], Pleak stands out for its ability to analyze processes not only in terms of yes-no questions (e. g., is private data unduly disclosed to a given party?) but also in quantitative terms, allowing analysts to address questions such as: How much private information about a given individual can a participant in a process infer from the data disclosed to them? Or, how much noise should be added to the data disclosed to a party so that this party cannot infer some private information within certain bounds?

Pleak has been validated via three case studies in the fields of emergency response, international aid distribution, and IoT-based building occupancy monitoring. The case studies have demonstrated the usefulness of combining qualitative and quantitative analysis to understand the scope and impact of privacy leakages.

An online version of Pleak is available for evaluation purposes at https://pleak.io/. The source code is available at https://github.com/pleak-tools.

References

1. Accorsi R, Lehmann A, Lohmann N (2015) Information Leak Detection in Business Process Models: Theory, application, and tool support. Inf Syst 47: 244–257
2. Governatori G, Milosevic Z, Sadiq S (2006) Compliance checking between business processes and business contracts. In: Proceedings of the 10th International IEEE Enterprise Distributed Object Computing Conference EDOC'06, pp 221–232
3. Van Looy A, Shafagatova A (2016) Business process performance measurement: A structured literature review of indicators, measures and metrics. SpringerPlus 5(1):1797
4. Pullonen P, Tom J, Matulevičius R, Toots A (2019) Privacy-enhanced BPMN: Enabling data privacy analysis in business processes models. Softw Syst Model. https://doi.org/10.1007/s10270-019-00718-z
5. Toots A, Tuuling R, Yerokhin M, Dumas M, García-Bañuelos L, Laud P, Matuleviičius R, Pankova A, Pettai M, Pullonen P, Tom J (2019) Business Process Privacy Analysis in Pleak. CoRR abs/1902.05052 (2019). https://arxiv.org/abs/1902.05052

A Hybrid Approach to Privacy-Preserving Federated Learning

(Extended Abstract)

Stacey Truex · Nathalie Baracaldo
Ali Anwar · Thomas Steinke
Heiko Ludwig · Rui Zhang · Yi Zhou

Data Privacy versus Machine Learning

Machine learning (ML) models' predictive accuracy has increased significantly in the last decade, and ML use has extended into everyday life applications from business to healthcare. Training ML models often requires access to quality data generated by multiple individuals and organizations. In a traditional ML process data are collected in one repository for processing. In an era of increased discussion of privacy this is often not a viable approach for a number of reasons: Some data – of a personal or financial nature – are inherently sensitive to individuals and organizations. In other cases, disclosure of data might lead to reputational damage: organizations may be reluctant to share data on security incidents, for example. Most importantly, regulation such as the General Data Protection Regulation (GDPR) and Health Insurance Portability and Accountability Act (HIPAA) impose clear limitations on what data can be shared. Finally, even if a third party is trusted in general, security incidents can expose private data unwillingly [5]. The best way to protect privacy is to keep data with the owner.

These privacy concerns often inhibit the training of beneficial ML models from larger data sets: banks could benefit from pooling data for fraud prediction. Hospitals could improve radiology assistance based on patient data from multiple institutions. We need new approaches that allow the training of ML models without having to share data.

The case for privacy-preserving federated learning

Privacy enabling techniques such as federated learning (FL) [4] and differential privacy (DP) [3] help address this problem. In an FL training process, each of the parties involved in learning maintains their data locally, trains a local model, and exchanges model parameters with other parties, typically through an aggregator. The information exchanged depends on the specific model type: gradients for neural networks and counts for decision tree. Through FL, an ML model can be trained without transferring or sharing data with a central entity.

However, this is not enough to prevent information disclosure. It has been shown that it is possible to infer private training data based on messages exchanged during the training process or querying the final model [7]. To prevent these inference attacks, FL has been combined with local differential privacy (local-DP) [1] in recent work to ensure inference attacks do not leak information about training data. In such a schema, each party adds differentially private noise to the model information sent to the aggregator to achieve a desired privacy guarantee (measured by epsilon). This is done independently by each party. Although this method ensures the privacy of data in a trained model, it also reduces substantially the predictive performance of the model due to the large amount of noise injected by each participating entity. This is not viable in many domains. To apply FL meaningfully to privacy-sensitive domains we

https://doi.org/10.1007/s00287-019-01205-x

Stacey Truex
Georgia Tech,
Atlanta, GA, USA
E-Mail: staceytruex@gatech.edu

Nathalie Baracaldo · Ali Anwar · Thomas Steinke
Heiko Ludwig · Rui Zhang · Yi Zhou
IBM Research – Almaden
San Jose, CA, USA
E-Mail: baracald@us.ibm.com

need ways to reduce the DP noise introduced by each party without scarifying the DP guarantee (epsilon).

Reducing noise

Hybrid-One is a federated learning framework designed to produce ML models with high predictive value while maintaining the desired differential privacy guarantee, epsilon, offered by state-of-the-art techniques. To achieve these results, Hybrid-One combines the use of DP with multiparty computation (MPC). Since messages exchanged during training time are protected using MPC techniques and disguise from which party data is coming from, each of the involved parties can use less noise than it would in a traditional scheme of local DP. Specifically, the noise introduced is reduced by a factor of n, where n is the number of parties involved in the learning process. Additionally, the learning processes in Hybrid-One can be configured to address potential collusion threats where participants who may not fully trust all parties remain from colluding to infer data of other participants. In this case, the total noise is reduced by the number of non-colluding parties with respect to local DP approaches. To achieve these results, Hybrid-One utilizes a threshold-based additive homomorphic approach [2] based on the Pallier cryptosystem [6]. This allows us to provide end-to-end privacy guarantees with respect to the participants as well as any attackers of the model itself.

We validated Hybrid-One with experimental results on three popular and significantly different ML algorithms: decision trees, support vector machines, and convolutional neural networks. Our experiments also demonstrate that our approach outperforms state of the art solutions in accuracy and customizability, and allows collaborations where parties do not have large amount of data. We believe Hybrid-One provides a good framework to allow collaborative learning without sharing data and protecting participating parties from inference attacks.

References

1. Blum A, Dwork C, McSherry F, Nissim K (2005) Practical privacy: the SuLQ framework. In: Proceedings of the Twenty-Fourth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems. ACM, pp 128–138
2. Damgård I, Jurik M (2001) A generalisation, a simplification and some applications of Paillier's probabilistic public-key system. In: Proceedings of the 4th International Workshop on Practice and Theory in Public Key Cryptography: Public Key Cryptography, PKC '01. Springer, London, pp 119–136
3. Dwork C (2008) Differential privacy: A survey of results. In: International Conference on Theory and Applications of Models of Computation. Springer, pp 1–19
4. McMahan HB, Moore E, Ramage D, Hampson S (2016) Communication-efficient learning of deep networks from decentralized data. arXiv preprint, arXiv:1602.05629
5. NBC News (2018) Yahoo to pay $50 million, offer credit monitoring for massive security breach. https://www.nbcnews.com/tech/tech-news/yahoo-pay-50m-offer-credit-monitoring-massive-security-breach-n923531
6. Paillier P (1999) Public-key cryptosystems based on composite degree residuosity classes. In: International Conference on the Theory and Applications of Cryptographic Techniques. Springer, pp 223–238
7. Shokri R, Stronati M, Song C, Shmatikov V (2017) Membership inference attacks against machine learning models. In: Security and Privacy (SP), IEEE Symposium. IEEE, pp 3–18

PrivApprox: Privacy-Preserving Stream Analytics[1]

(Extended Abstract)

Do Le Quoc · Martin Beck
Pramod Bhatotia · Ruichuan Chen
Christof Fetzer · Thorsten Strufe

Problem description

Many online services continuously collect users' private data for real-time analytics. In the current ecosystem of data analytics, the analysts usually have direct access to users' private data and must be trusted not to abuse it. A pragmatic ecosystem has two desirable, but contradictory design requirements. Users seek stronger privacy, while analysts strive for high-utility analytics in real time. Much of this data arrives as a data stream and in huge volumes, requiring real-time stream processing based on distributed systems.

New computing paradigms try to address these concerns. On the one hand, *privacy-preserving analytics* protects the user data by local storage of information and integration of noise, while on the other hand, *approximate computation* enables low-latency, high-throughput stream analytics. However, the respective research fields address the issue separately. State-of-the-art privacy-preserving systems typically operate on single-shot batch queries and are, therefore, not applicable to real-time stream processing. Approximate computation proposals build on the underlying assumption of a centralized dataset. Thus, both solutions are not compatible for trivial integration.

Usage scenario:

Today, sources of privacy-relevant data streams are pervasive and can be found almost everywhere. IoT devices, automotives, mobile phones, wearables, or workstations, all stream privacy-related information. This includes, but is not limited to, environmental measurements, health data, location for navigation, augmented reality, or social contacts, as well as statistical information about the usage of devices. We concentrate our analysis on location information, more specifically on the distances traveled using taxis based on the *NYC taxi ride* dataset and power consumption of households represented by the *Household Electricity Consumption* dataset.

Inference upon this information can easily reveal working practices, habits, personal preferences, social connections, medical issues, and other highly personal information. It is therefore of utmost importance to preserve the privacy of the users involved.

Results

To guarantee privacy of users within a stream analytic ecosystem, we suggest a system that combines the results of the computing paradigms mentioned. Indeed, we make the observation that privacy-preserving analytics and approximate computation are complementary. Both paradigms strive for an

[1] This paper is published as Quoc, Beck, Bhatotia, Chen, Fetzer, Strufe: PrivApprox: Privacy-Preserving Stream Analytics. USENIX Annual Technical Conference (ATC), 2017.

https://doi.org/10.1007/s00287-019-01206-w

Do Le Quoc · Christof Fetzer
TU Dresden, Systems Engineering,
Dresden, Germany
E-Mail: do.le_quoc@tu-dresden.de

Martin Beck · Thorsten Strufe
TU Dresden, Privacy and Security,
Dresden, Germany

Pramod Bhatotia
The University of Edinburgh,
Computing Systems Architecture,
Edinburgh, UK

Ruichuan Chen
Nokia Bell Labs,
Stuttgart, Germany

approximate instead of the exact output, but they differ in their means and goals for approximation. Privacy-preserving analytics adds explicit noise to the aggregate query output to protect user privacy, whereas approximate computation relies on a representative sampling of the entire dataset to compute over only a subset of data items to enable low-latency/efficient analytics. Therefore, we marry these two existing paradigms to leverage the benefits of both. The high-level idea is that adding noise and sampling for approximate computing both increase privacy. Therefore, the overall privacy level and approximation for efficient analysis can be adjusted independently, but nevertheless strongly support each other.

As a result, the system presented achieves the provably secure privacy notion of zero-knowledge privacy, which is strictly stronger than the well-known privacy notion of differential privacy. Despite providing higher privacy guarantees than the available related work, our proposal offers lower latencies and similarly high throughput.

Parallele Dateisysteme

Michael Kuhn

Einleitung

Die weltweit produzierte Datenmenge verdoppelt sich momentan ungefähr alle zwei Jahre. Diese exponentiell wachsende Datenflut muss gespeichert, analysiert und weiterverarbeitet werden. Traditionelle Dateisysteme können die wachsenden Anforderungen an Speicherkapazität und -geschwindigkeit nicht erfüllen, weswegen parallele Dateisysteme bereits in vielen datenintensiven Bereichen wie z. B. der Hochenergiephysik oder den Klimawissenschaften Verwendung finden.

Das Hochleistungsrechnen hat sich dabei als ein nützliches und mittlerweile unverzichtbares Werkzeug für viele Wissenschaftsdisziplinen etabliert. Durch die dadurch möglichen Analysen und Simulationen kann der wissenschaftliche Erkenntnisgewinn in vielen Bereichen deutlich gesteigert werden. In den vergangenen Jahrzehnten konnte die Rechenleistung der in der TOP500-Liste vertretenen Hochleistungsrechner durchschnittlich alle 14 bis 15 Monate verdoppelt werden [6]. Obwohl seit 2015 ein Abflachen der Steigerungsraten zu beobachten ist, führt dieser weiterhin exponentielle Anstieg zu einer rasanten Zunahme der produzierten Daten.

Dateisystemgrundlagen

Dateisysteme ermöglichen einen komfortablen Zugriff auf Speichergeräte. Sie stellen eine Abstraktionsschicht zwischen Anwendungen und der tatsächlichen Speicherhardware dar, sodass sich Anwendungsentwickler nicht mit den Eigenheiten der darunterliegenden Speicherhardware auseinandersetzen müssen. Darüber hinaus stellen Dateisysteme üblicherweise standardisierte Zugriffsschnittstellen bereit, die die Portabilität der Ein-/Ausgabe (E/A) ermöglichen.

Die Speicherhardware setzt sich heutzutage typischerweise aus Festplatten und Solid-State-Drives zusammen, die unterschiedliche Leistungscharakeristika und Kosten aufweisen. Solche Speichergeräte geben keine Organisationsstruktur vor und erlauben einen block- oder byteweisen Zugriff auf den bereitgestellten Speicher.

Die beiden grundlegenden Datenstrukturen, die sich in fast allen Dateisystemen wiederfinden, sind Dateien und Verzeichnisse. Dateien enthalten die eigentlichen Daten, während Verzeichnisse zur Strukturierung des Dateisystemnamensraumes dienen. So können Verzeichnisse üblicherweise sowohl Dateien als auch weitere Verzeichnisse enthalten, wodurch ein hierarchischer Namensraum entsteht. Der Zugriff auf Dateien und Verzeichnisse findet normalerweise über Namen statt, wobei ein voll qualifizierter Name auch Pfad genannt wird.

Dateien und Verzeichnisse stellen hierbei nur das Minimum an Dateisystemfunktionalität dar. Viele Dateisysteme bieten zusätzliche Funktionalitäten wie z. B. Named Pipes [4]. Durch die immer weiter steigenden Datenmengen integrieren moderne Dateisysteme außerdem zunehmend Funktionen zur Volumenverwaltung, Kompression und Sicherstellung der Datenintegrität [7].

Obwohl sich die Funktionsweise von Dateisystemen auf unterschiedlichen Betriebssystemen

https://doi.org/10.1007/s00287-019-01209-7

Michael Kuhn
Universität Hamburg, Hamburg
E-Mail: michael.kuhn@informatik.uni-hamburg.de

nicht grundlegend unterscheidet, werden sich die Erklärungen in diesem Artikel auf Linux und seine durch POSIX (Portable Operating System Interface) standardisierten Schnittstellen beschränken. Insbesondere im Hochleistungsrechnen stellt Linux einen De-facto-Standard dar und so finden sich in der aktuellen TOP500-Liste ausschließlich Hochleistungsrechner mit einem auf Linux basierenden Betriebssystem. Die E/A-Schnittstelle des POSIX-Standards wurde primär für lokale Dateisysteme entwickelt. Eine erste formale Spezifikation erfolgte im Rahmen von POSIX.1 im Jahr 1988, wobei diese für asynchrone und synchrone E/A im Jahr 1993 durch POSIX.1b erweitert wurde. Zusätzlich zur Syntax der E/A-Schnittstellen wird auch deren Semantik, d. h. deren Verhalten, durch POSIX spezifiziert. So wird beispielsweise festgelegt, dass Leseoperation nach dem Zurückkehren einer Schreiboperation sofort die neuen Daten zurückliefern müssen, was Auswirkungen auf das Cachingverhalten haben kann [5]. Fast alle unter Linux verfügbaren Dateisysteme sind POSIX-konform, wodurch eine hohe Portabilität gewährleistet werden kann. Insbesondere bieten auch die meisten parallelen Dateisysteme eine POSIX-konforme Schnittstelle an, sodass Anwendungen ohne größere Anpassungen auch auf Hochleistungsrechnern lauffähig sind.

Ein weiteres wichtiges Konzept sind die Metadaten. Dateien und Verzeichnisse bestehen jeweils sowohl aus Daten als auch aus Metadaten. Im Fall einer Datei bezeichnen die Daten den eigentlichen Inhalt der Datei, während bei einem Verzeichnis die Verzeichniseinträge gemeint sind. Die Metadaten beschreiben in beiden Fällen weitergehende Informationen wie z. B. Zugriffsberechtigungen oder Eigentümer (siehe Abb. 1). Im Fall von POSIX-konformen Dateisystemen werden die Metadaten als sogenannte Inodes verwaltet. Während die Größe der Daten stark variieren kann, nehmen die Metadaten üblicherweise nur sehr wenig Platz ein. Im Fall des traditionellen Linux-Dateisystems ext4 haben Inodes eine Standardgröße von 256 Byte. Insbesondere enthalten die Inodes auch Verweise auf die eigentlichen Daten. Das darunterliegende Speichergerät wird meist in Blöcke gleicher Größe aufgeteilt, die dann im Inode referenziert werden. So kann der Benutzer komfortabel über Namen auf Dateien und Verzeichnisse zugreifen, während sich das Dateisystem intern um die Allokation und Verwaltung des notwendigen Speicherplatzes kümmert.

Feldgröße	Inhalt
2 Byte	Berechtigungen
2 Byte	Benutzer-ID (Untere Bits)
4 Byte	Dateigröße (Untere Bits)
4 Byte	Zugriffszeit (Sekunden)
⋮	⋮
4 Byte	Versionsnummer (Obere Bits)
100 Byte	Freier Speicher

Abb. 1 ext4-Inode [1]: Inodes sind in feste Felder und einen freien Bereich am Ende aufgeteilt. Aus Gründen der Rückwärtskompatibilität sind unter anderem die Felder für die IDs, die Größe und die Zeitstempel in jeweils zwei Felder aufgeteilt. Der freie Bereich kann für erweiterte Attribute oder spätere Erweiterungen der Inode-Datenstruktur genutzt werden.

Parallele Dateisysteme

Um die Anforderungen moderner datenintensiver Anwendungen erfüllen zu können, bieten parallele Dateisysteme sowohl einen effizienten parallelen Zugriff von mehreren Anwendungen sowie eine Verteilung der Daten über mehrere Speichergeräte und Dateisystemserver hinweg. Einerseits erlaubt es der parallele Zugriff verteilten Anwendungen gleichzeitig mit einem gemeinsamen Datensatz zu arbeiten, andererseits können durch die Verteilung der Daten die Kapazität und der Durchsatz vieler Speichergeräte genutzt werden (siehe Abb. 2). Dadurch sind aktuell massiv parallele Dateisysteme mit Speicherkapazitäten im dreistelligen Petabytebereich und Durchsätzen im einstelligen Terabytebereich realisierbar. Durch die vom Dateisystem bereitgestellte Abstraktion kann die Verteilung der Daten vollstän-

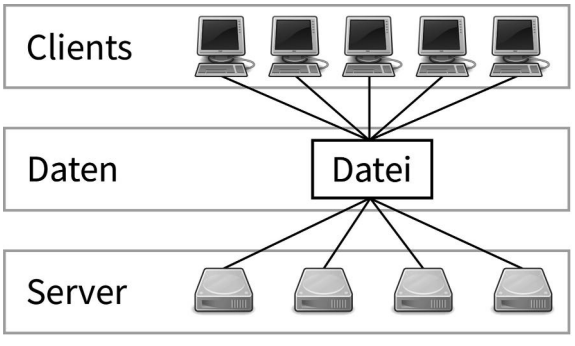


Abb. 2 Paralleler Zugriff und Datenverteilung: Mehrere Anwendungen können gleichzeitig auf eine gemeinsame Datei zugreifen, die wiederum über mehrere Speichergeräte und Dateisystemserver verteilt wird.

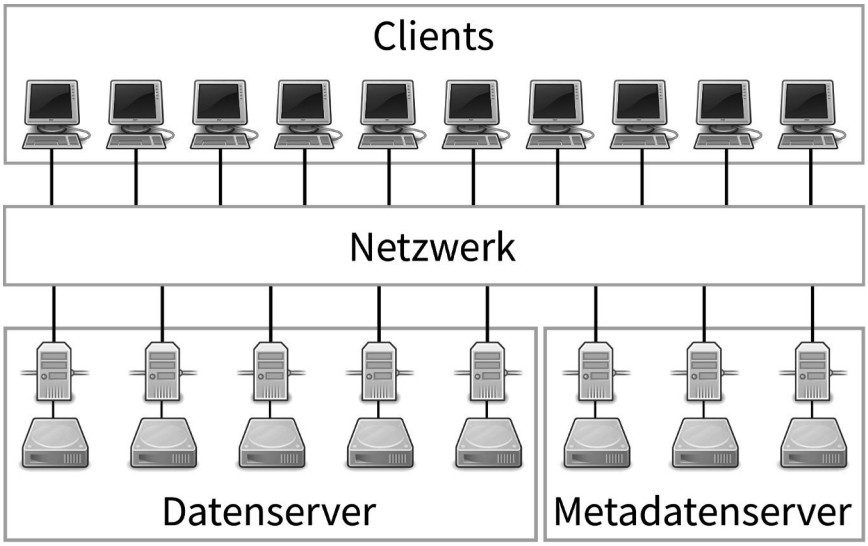


Abb. 3 Paralleles Dateisystem: Clients kommunizieren über ein Netzwerk mit den Daten- und Metadatenservern. Auf den Clients ist das Dateisystem häufig als POSIX-konformes Kernelmodul in das Betriebssystem eingebunden.

dig transparent für die Anwender geschehen. Drei der am weitesten verbreiteten parallelen Dateisysteme sind Lustre von DDN, Spectrum Scale (ehemals GPFS) von IBM und BeeGFS von ThinkParQ und Fraunhofer. Da im Bereich paralleler Dateisysteme keine einheitliche Nomenklatur existiert, werden diese häufig auch als verteilte Dateisysteme oder seltener Cluster-Dateisysteme bezeichnet.

Ein paralleles Dateisystem wird üblicherweise in Clients und Server getrennt (siehe Abb. 3). Dies erlaubt einerseits eine Spezialisierung auf die jeweilige Funktionalität und schließt andererseits eine gegenseitige Beeinflussung aus. Beispielsweise können die Clients so mit vielen Kernen für Berechnungsaufgaben ausgestattet werden, während die Server entsprechende Speichergeräte und ausreichend Arbeitsspeicher für das Caching enthalten. Darüber hinaus werden die Server häufig in Daten- und Metadatenserver aufgeteilt, da auf Daten meist große Zugriffe erfolgen, wohingegen Metadatenzugriffe häufig klein und zufällig verteilt sind. So können für die Datenserver optimalerweise Festplatten mit höherer Kapazität eingesetzt werden, während für die Metadatenserver kleinere Festplatten mit höherer Umdrehungszahl oder Solid-State-Drives genutzt werden.

Durch diese Aufteilung müssen die Clients über das Netzwerk mit den Servern kommunizieren, was zusätzliche Latenz verursacht. Eine typische Architektur verlagert den Großteil der Dateisystemlogik in die Clients, sodass diese selbstständig entscheiden können, welche Server kontaktiert werden müssen. In diesem Fall müssen die Server nicht untereinander kommunizieren und können als relativ einfache Datenspeicher fungieren.

Die eigentliche Verteilung der Daten und Metadaten wird dabei von Verteilungsfunktionen übernommen. Die häufigste Verteilungsfunktion für Daten ist ein simpler Round-Robin-Ansatz, der die Daten in gleichgroße Blöcke zerlegt und in Form von Streifen über die Datenserver verteilt (siehe Abb. 4). Etwaige Ungleichgewichte bei der Verteilung der Daten können aufgrund der zufällig gewählten Startparameter und der großen Dateigrößen für gewöhnlich vernachlässigt wer-

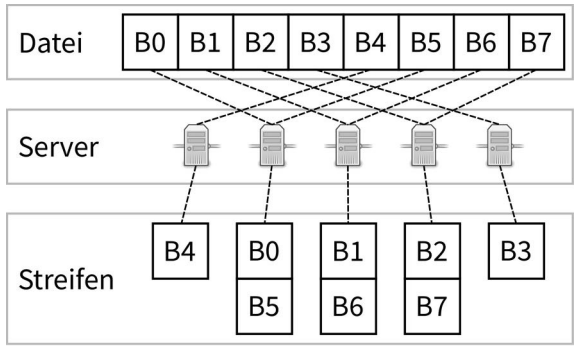


Abb. 4 Round-Robin-Verteilungsfunktion: Die Datei wird in acht gleichgroße Blöcke zerlegt und über fünf Datenserver verteilt. Der Startserver wird zufällig gewählt, um eine gleichmäßige Lastverteilung mit mehreren Dateien zu gewährleisten. Sobald der letzte Datenserver erreicht ist, wird mit dem ersten fortgefahren.

den. Da Metadaten häufig zu klein für eine solche Verteilung sind, werden die Metadaten einer einzelnen Datei oder eines einzelnen Verzeichnisses durch einen einzigen Metadatenserver verwaltet. Eine Ausnahme bilden sehr große Verzeichnisse, die zur Leistungssteigerung über mehrere Metadatenserver verteilt werden können. Zur Bestimmung des zuständigen Metadatenservers wird dabei häufig auf Hashing zurückgegriffen, wobei kryptographische Hashfunktionen den Vorteil einer gleichmäßigen Verteilung haben. So kann der zuständige Metadatenserver beispielsweise durch das Hashen des Dateinamens oder des vollen Pfads bestimmt werden.

Die grundlegende Funktionsweise soll an einem einfachen Beispiel erläutert werden: Ein Client will 42 Byte an Offset 4242 in die bereits existierende Datei `/meine/datei` schreiben. Dafür sind mehrere Schritte notwendig:

1. Zuerst muss das Wurzelverzeichnis (`/`) gelesen werden. Der Ort des Wurzelverzeichnisses ist daher meist statisch festgelegt, sodass keine weiteren Informationen für die Suche notwendig sind. Der Client kommuniziert mit dem zuständigen Metadatenserver und ruft die Verzeichnisinformationen ab. Daraufhin werden die Zugriffsrechte überprüft und bei Bedarf ein Fehler zurückgeben. Darf der Client auf das Wurzelverzeichnis zugreifen, werden die Einträge gelesen und nach dem Eintrag `meine` durchsucht.
2. Im nächsten Schritt muss das darauf folgende Verzeichnis (`meine`) gelesen werden. Der Ort und somit der zuständige Metadatenserver kann aus dem Verzeichniseintrag im Wurzelverzeichnis bestimmt werden, sodass nun der vorherige Schritt für diese Pfadkomponente wiederholt wird.
3. Als Nächstes müssen die Metadaten der eigentlichen Datei (`datei`) abgerufen werden. Der zuständige Metadatenserver kann aus dem im vorherigen Schritt gesuchten Verzeichniseintrag bestimmt werden. Daraufhin kommuniziert der Client mit dem Metadatenserver und ruft alle notwendigen Informationen ab. Es werden wieder die Zugriffsrechte geprüft und bei Bedarf ein Fehler zurückgeben.
4. Abschließend müssen die Daten geschrieben werden. Die im vorherigen Schritt abgerufenen Metadaten enthalten auch Informationen über die Verteilung der Daten. Auf deren Basis wird der zuständige Datenserver für das Offset 4242 bestimmt, auf dem wiederum die Schreiboperation durchgeführt wird. Sollte die Datei gleichzeitig von einem anderen Client geöffnet worden sein, müssen außerdem Sperren für den betroffenen Bereich angefordert werden, um das durch POSIX festgelegte Verhalten sicherzustellen [3].

Ausblick

Auf Grundlage der vorgestellten Architektur lassen sich parallele Dateisysteme in Betrieb nehmen, die einige Tausend Server enthalten, auf die wiederum mehrere Zehntausend Clients zugreifen. Trotz ihrer verteilten und skalierbaren Architektur stoßen parallele Dateisysteme allerdings immer häufiger an die Grenzen ihrer Leistungsfähigkeit. Während die mittleren Zugriffszeiten handelsüblicher Festplatten im zweistelligen Millisekundenbereich liegen, erreichen Solid-State-Drives Latenzen im zwei- bis dreistelligen Mikrosekundenbereich. In der Entwicklung befindliche Speichertechnologien wie beispielsweise NVRAM werden noch geringere Zugriffszeiten aufweisen. Um die Leistungsfähigkeit dieser neuen Speichergeräte ausnutzen zu können, dürfen die durch die Softwareschichten des Dateisystems verursachten Latenzen nicht überhand nehmen. Einen signifikanten Anteil am Dateisystem-Overhead hat dabei der Betriebssystemkernel, da zur Ausführung von E/A-Operationen mit dem Kernel kommuniziert werden muss. Die notwendigen Modus- und Kontextwechsel können je nach durchgeführter Operation und Hardwarearchitektur bis zu einigen Mikrosekunden dauern. Ein möglicher Ansatz ist die Umgehung des Kernels, wie dies bereits für diverse Netzwerktechnologien wie z. B. InfiniBand umgesetzt wird. Die strengen Kohärenz- und Konsistenzanforderungen des POSIX-Standards stellen ein weiteres Problem für die Skalierbarkeit paralleler Dateisysteme dar. Um diesen Anforderungen gerecht zu werden, entstehen aktuell neue Speichersystemkonzepte wie z. B. der Distributed Application Object Storage (DAOS), der vollständig im Benutzermodus läuft und eine transaktionsbasierte E/A-Schnittstelle bietet [2].

In Zeiten der rechen- und datenintensiven Forschung ist es notwendig, die Bemühungen um leistungsfähige und skalierbare Speichersysteme zu intensivieren. Auch grundlegende Dateisystemkonzepte müssen überdacht und gegebenenfalls verbessert werden, um den steigenden Anforderungen gerecht werden zu können.

Literatur

1. djwong. Ext4 Disk Layout. https://ext4.wiki.kernel.org/index.php/Ext4_Disk_Layout, last access: 30.8.2019
2. Liu J, Koziol Q, Butler GF, Fortner N, Chaarawi M, Tang H, Byna S, Lockwood GK, Cheema R, Kallback-Rose KA, Hazen D, Prabhat (2018) Evaluation of HPC application I/O on object storage systems. In: 3rd IEEE/ACM International Workshop on Parallel Data Storage & Data Intensive Scalable Computing Systems, PDSW-DISCS@SC 2018. Dallas, TX, USA, November 12, 2018, pp 24–34
3. Moore M, Farrell P, Cernohous B (2018) Lustre lockahead: Early experience and performance using optimized locking. Concurr Comp-Pract E 30(1):1–14
4. The Linux man-pages project. fifo(7). http://man7.org/linux/man-pages/man7/fifo.7.html, last access: 30.8.2019
5. The Linux man-pages project. write(2). http://man7.org/linux/man-pages/man2/write.2.html, last access: 30.8.2019
6. TOP500.org. Performance Development. https://www.top500.org/statistics/perfdevel/, last access: 30.8.2019
7. Zhang Y, Rajimwale A, Arpaci-Dusseau AC, RH Arpaci-Dusseau RH (2010) End-to-end data integrity for file systems: A ZFS case study. In: 8th USENIX Conference on File and Storage Technologies, San Jose, CA, USA, February 23–26, 2010, pp 29–42

FORUM

Gewissensbits – wie würden Sie urteilen?

Christina B. Class,
Ernst-Abbe-Hochschule Jena
E-Mail: cclass@acm.org,
Stefan Ullrich,
Weizenbaum-Institut für die vernetzte Gesellschaft, Berlin

Mit den ethischen Leitlinien der GI haben wir es uns zur Aufgabe gemacht, Diskurse zu ethischen Problemen der Informatik zu initiieren und zu fördern. Mitglieder der Fachgruppe „Informatik und Ethik" der GI stellen jeweils ein hypothetisches, aber realistisches Fallbeispiel vor, das zur Diskussion anregen soll. Die Fälle können jeweils von Interessierten im Blog der Fachgruppe auf der GI-Website https://gewissensbits.gi.de kommentiert und diskutiert werden.

Fallbeispiel: Statistische Irrungen

Alex hat vor etwas mehr als einem Jahr seine Masterarbeit im Bereich Künstliche Intelligenz und Gesichtserkennung abgeschlossen. Sein adaptiertes selbstlernendes Verfahren konnte die bisherigen Ergebnisse der Echtzeit-Gesichtserkennung deutlich verbessern. Als er seine Abschlussarbeit auf einer Tagung vor einem Jahr vorgestellt hat, inklusive Proof-of-Concept auf der Bühne, wurde er vom Leiter der KI Forschungs- und Entwicklungsabteilung der EmbraceTheFuture GmbH angesprochen. Diese wurde vor drei Jahren gegründet, mit Schwerpunkt auf der Entwicklung angepasster Softwaresysteme, insbesondere im Bereich der intelligenten Systeme sowie Sicherheitssysteme. Nach einem kurzen Urlaub nach Ende seines Studiums nimmt Alex eine Stelle bei EmbraceTheFuture GmbH an.

In einem kleinen Team arbeitet er zurzeit im Auftrag der Bundespolizei an Gesichtserkennungssoftware für ein neues Sicherheitssystem namens „QuickPicScan" an Flughäfen. In Echtzeit werden die Gesichter der Passagiere an der Sicherheitskontrolle mit Fahndungsbildern verglichen, um verdächtige Personen zur Seite zu nehmen und extra zu kontrollieren. Die Behörden erhoffen sich so, gesuchte Personen bei Flügen innerhalb des Schengenraums identifizieren zu können, da dort keine umfassenden Passkontrollen vorgenommen werden. Zudem soll der Durchsatz der kontrollierten Personen gesteigert werden.

Das System wurde mit Millionen von Bildern umfassend trainiert. Bilder von gesuchten Personen, Fahndungsbilder, sind in einer Datenbank gespeichert, auf die bei einem Bildabgleich zugegriffen wird. Dadurch kann das System leicht an aktuell gesuchte Personen angepasst werden.

Am Flughafen werden Bilder aller Passagiere in niedriger Qualität aufgenommen, sobald die Sicherheitsschleuse durchschritten wird. Wenn die Software anschlägt, wird der übliche „Metall gefunden"-Alarm ausgelöst. In der darauffolgenden Durchsuchung wird das Gesicht in höherer Auflösung unter besseren

https://doi.org/10.1007/s00287-019-01213-x

Lichtverhältnissen fotografiert und erneut mit den Bilddaten verglichen. Erst wenn dieser zweite Test auch positiv ist, erfolgt eine tiefergehende Kontrolle im Nebenraum mit Abgleich der Personalien. Das Ergebnis des zweiten Tests wird an einem Kontrollterminal angezeigt. Die Fotos der Passagiere werden nicht gespeichert, ein eigenes Team stellt sicher, dass die aufgenommenen Bilder wirklich aus dem Hauptspeicher gelöscht werden und nicht von außen ausgelesen werden können.

QuickPicScan wurde in Simulationen sowie in einem eigens dafür gebauten Studio mit einer nachgebauten Sicherheitskontrolle und Schauspielern umfangreich getestet. Basierend auf den Tests geht das Team von einer False Negative Rate von 1 % aus, d. h. von 100 gesuchten Personen wird eine nicht gefunden. Die False Positive Rate – Personen, die zu Unrecht als verdächtig eingestuft werden – ist kleiner als 0,1 %. Sabine, die Marketingleiterin, ist von dem Ergebnis begeistert. Nur 0,1 % Fehlerrate für unschuldige Personen, das sei ein Riesenerfolg!

Um das System unter realen Bedingungen zu testen, wird es in Abstimmung mit der Bundespolizei während zweier Sommermonate in einem kleineren Flughafen mit einem Passagieraufkommen von ca. 400.000 Passagieren pro Jahr getestet. Das Kontrollterminal wird von einem Angestellten des Auftraggebers überwacht. Von 370 Darstellern wurden „Fahndungsbilder" in unterschiedlich guter Qualität und verschiedenen Positionen aufgenommen und ins System gespeist. Während der zwei Testmonate gehen die Darsteller zu vorher zufällig festgelegten Zeiten insgesamt 1.500 Mal durch die Sicherheitskontrolle. Sie geben sich nach Durchgang der Person am Kontrollterminal zu erkennen, damit das System getestet werden kann.

Aufgrund der Ferienzeit werden in den zwei Testmonaten 163.847 Passagiere kontrolliert. Bei 183 Passagieren leuchtet die Lampe fälschlicherweise auf. Bei 8 der 1.500 Sicherheitskontrollen der Darsteller wurde die Übereinstimmung nicht erkannt. Der Gesamtprojektleiter Viktor ist begeistert. Zwar war die False Positive Rate mit 0,11 % etwas schlechter als ursprünglich erhofft, die False Negative Rate mit 0,53 % aber deutlich besser als angenommen. Mit diesen Zahlen und der Fehlerrate von 0,11 % geht EmbraceTheFuture GmbH an die Presse. Die Bundespolizei kündigt den baldigen Einsatz in einem Terminal eines großen Flughafens an.

Am Abend trifft Alex seine alte Schulfreundin Vera, die zufällig in der Stadt ist. Sie arbeitet als Geschichts- und Mathematiklehrerin. Nachdem sie sich über das neueste aus ihrem Alltag und Liebesleben aufs Laufende gebracht haben, berichtet Alex Vera begeistert von dem Projekt und erzählt von der Pressekonferenz. Vera reagiert ziemlich kritisch, automatische Gesichtserkennung behagt ihr irgendwie gar nicht. Darüber hatten sie schon während Alex' Master häufiger diskutiert. Alex berichtet begeistert von den geringen Fehlerraten, der erhöhten Sicherheit und der Möglichkeit, untergetauchte Personen zu identifizieren. Vera schaut ihn skeptisch an. Sie findet die Fehlerrate überhaupt nicht gering. 0,11 % – bei einem großen Flughafen sind das doch Dutzende Personen, die für weitere Kontrollen beiseite genommen werden. Das findet sie gar nicht lustig. Auch fragt sie sich, wie viele Personen, von denen es Fahndungsfotos gibt, tatsächlich mit dem Flugzeug fliegen. Alex will darüber nicht wirklich was hören und beginnt, ihr den Algorithmus, den er im Rahmen der Masterarbeit weiterentwickelt hat, genauer zu erläutern…

Einige Monate später ist das System im AirportCityTerminal fertig installiert, Beamte wurden geschult und die Presse meldet den erfolgreichen Start. Wenige Tage später fliegt Alex vom AirportCityTerminal ab und freut sich schon darauf, an QuickPicScan vorbeizugehen und sich in dem Gefühl zu sonnen, dass er einen Beitrag zu erhöhter Sicherheit leisten konnte. Doch als er in die Sicherheitsschleuse getreten ist, piepst der Metall-Alarm. Er wird gebeten, die Arme auszustrecken, die Füße abwechselnd auf einen Hocker zu stellen und zu guter Letzt geradeaus zu schauen. Er schielt nach rechts auf den Monitor der Sicherheitsbeamten und sieht, wie die kleine Kontrollleuchte am QuickPicScan-Terminal leuchtet. Hoffentlich geht das schnell, es wird knapp mit seinem Flug. Da er kein Gepäck eingecheckt hat, würden sie nicht auf ihn warten.

Er wird in einen separaten Raum geführt wo man ihn bittet, seine Papiere bereitzuhalten. Ein Beamter steht ihm gegenüber. Alex will ihm seinen Personalausweis reichen, dieser meint jedoch, dass die zuständige Kollegin gleich kommen würde, sie müsse noch jemand anderen überprüfen. Alex wird langsam ungeduldig. Er bittet darum, dass seine Identität überprüft wird. Nein, das ginge nicht, der postierte Sicherheitsbeamte habe noch keine Einweisung für das neue System bekommen. Erst acht Minuten später taucht die verantwortliche Beamtin auf.

Nach der Identitätsfeststellung ist klar, dass es sich bei Alex nicht um eine gesuchte Person handelt. Sein Gepäck wird dennoch minutiös untersucht. „Ist Vorschrift", sagt die Beamtin knapp. Alex wird unruhig, den Flieger wird er wohl verpassen.

Plötzlich kommt ihm das Gespräch mit Vera wieder in den Sinn. „Passiert das öfter?“, fragt er mit gespielter Freundlichkeit. „Ach, ein paar Dutzend sind es schon am Tag“, sagt die Beamtin, als sie ihn wieder zurück ins Terminal geleitet.

Fragen

- Alex wurde fälschlicherweise vom System als „Verdächtiger“ identifiziert und hat in Folge seinen Flug verpasst. Dies bezeichnet man als *false positive*. In welchen Fällen muss hingenommen werden, dass es *false positive* gibt? Welche Folgen sind für die Betroffenen hinnehmbar? Wie müssten Entschädigungen geregelt werden?
- Auch Menschen können Fehleinschätzungen vornehmen. In einer ähnlich gelagerten Situation wie in der geschilderten könnte Alex auch von einem Sicherheitsbeamten zur Seite genommen werden, um genauer kontrolliert zu werden. Gibt es hier einen prinzipiellen Unterschied?
- Menschen haben Vorurteile. Es ist bekannt, dass ausländisch aussehende Männer zum Beispiel häufiger kontrolliert werden. Welche Chancen bestehen, solche Diskriminierungen durch Menschen mithilfe von Softwaresystemen zu verringern?
- Selbstlernende Algorithmen benötigen Trainingsdaten. Die Ergebnisse der Algorithmen hängen damit stark von den Trainingsdaten ab. Dies kann zu im Algorithmus manifestierter Diskriminierung führen. Denkbar ist auch, dass z. B. Gesichter einer bestimmten Personengruppe weniger genau erkannt werden, wenn weniger Trainingsdaten zur Verfügung stehen. Dies kann sich auf Hautfarbe, Alter, Geschlecht, Vorhandensein eines Barts etc. beziehen. In einem System wie dem beschriebenen könnte dies dazu führen, dass Personen mit bestimmten äußerlichen Merkmalen schneller beiseite genommen werden, um sie zu kontrollieren. Wie kann man Trainingsdaten sinnvoll wählen, um diskriminierende Systeme nach Möglichkeit zu verhindern? Wie kann man Systeme mit Blick auf solche Diskriminierungen testen?
- Gibt es einen konzeptionellen Unterschied zwischen im System manifestierter Diskriminierung und Diskriminierung durch Menschen? Welche ist einfacher zu identifizieren?
- Menschen tendieren dazu, Antworten, die von einer Software gegeben wird, schnell zu vertrauen und Verantwortung abzugeben. Macht dies Diskriminierung durch technische Systeme besonders gefährlich? Welche Möglichkeiten der Sensibilisierung gibt es? Sollte, und wenn ja in welcher Form, eine Sensibilisierung in den Schulen erlernt werden? Ist sie Teil notwendiger digitaler Kompetenzen für die Zukunft?
- Zahlen für die *false positive* und *false negative* Rate werden oft in Prozent angegeben. Fehlerraten von unter 1 % klingen zuerst mal gar nicht so schlecht. Oftmals fällt es Menschen schwer, sich vorzustellen, wie viele Personen in realen Anwendungen davon betroffen wären, welche Folgen dies haben könnte und was das bedeutet. Oft werden auch beide Zahlen nebeneinander gestellt, ohne das Verhältnis zwischen Positives (in unserem Fall die Personen, die per Fahndungsbild gesucht werden) und Negatives (in unserem Fall alle anderen Passagiere) abzubilden. Oft ist dieses Verhältnis sehr unausgeglichen. Beim beschrieben Testlauf sollten 1.500 Personen (Positives) von 163.847 Passagieren identifiziert werden, also ein Verhältnis von ca. 1:100. Ist ein solcher Vergleich irreführend? Dürfen solche Zahlen in Produktbeschreibungen bzw. Marketingbroschüren genutzt werden? Handeln die Verantwortlichen von EmbraceTheFuture GmbH unethisch, wenn Sie an die Presse gehen? Gibt es andere Fehlermaße? Wie kann man Fehlerraten realistisch darstellen, sodass Systeme realistisch eingeschätzt werden?

Die Fachgruppe ist unter http://www.fg-ie.gi.de/ erreichbar. Unser Buch „Gewissensbisse – Ethische Probleme der Informatik. Biometrie – Datenschutz – geistiges Eigentum“ ist im Oktober 2009 im Transkript-Verlag erschienen. Ein neues Werk ist in Arbeit.

FORUM

Distributed Ledger und Governance

Ursula Sury
Luzern, Schweiz

Bei der Distributed-Ledger-Technologie (DLT) werden gleichgestellte oder gleiche Kopien von Daten von unterschiedlichen Parteien unterhalten und weiterentwickelt.

Als Basis dafür wird heute häufig die Blockchaintechnologie gewählt, aber nicht nur. Der Unterschied zwischen der Blockchaintechnologie und anderen DLT-Anwendungen liegen insbesondere in den Möglichkeiten, die die (Informations-)Verkettung (Chain) bei der Blockchain mit sich bringt.

Die Möglichkeit, über offene und transparente Datensätze Interaktionen zu unterstützen, könnte große Veränderungen in der Wirtschaft und im Staat mit sich bringen.

Die Transparenz und redundante Haltung identischer Datensätze macht Fälschungen jeglicher Art schwieriger. Das ist ein großer Pluspunkt.

Je nach Art der Interaktion ist es auch möglich, traditionelle Vermittler, wie Banken, Treuhänder, Notare etc., auszulassen und somit kostengünstiger und schneller zu arbeiten.

Die Anforderung der Bewirtschaftung gleicher Inhalte durch verschiedene Parteien erfordert aber spezifische Regelungen, wie festgelegt werden soll, welche Informationen denn in den verteilten Datenbanken aufgenommen werden sollen.

Es stellen sich insbesondere folgende Fragen, die bei Governanceinstrumenten berücksichtigt werden müssen:

- Sind die Informationen überhaupt relevant für das konkrete Business
- Wer steht konkret hinter den Informationen? In einer offenen Blockchain ist dies zum Beispiel nicht nachvollziehbar. Habe ich verlässliche Ansprechpartner?
- Dürfen die Informationen überhaupt offengelegt werden? Je nach Design der DLT ist ein unterschiedlicher Adressatenkreis möglich. Schranken der Offenlegung können sich insbesondere aus dem Datenschutz, dem Urheberrecht, dem Wettbewerbsrecht oder gesetzlichen und vertraglichen Geheimhaltungen ergeben.
- Warum sind wir sicher, dass die Informationen stimmen? Wer garantiert dafür?
- Was dürfen zugriffsberechtigte Parteien mit den Informationen machen? An welche Regeln müssen sie sich halten und warum sind wir sicher, dass sie das auch machen? Gibt es entsprechende Kontrollen?
- Wie werden die Governanceinstrumente und die Regeln weiterentwickelt? Gibt es dafür Quoren?

Es gilt dabei zu unterscheiden zwischen Off-Chain-Governance, die außerhalb und rund um die Blockchain vorgegeben wird, und On-Chain-Governance, die direkt durch und auf dem Blockchain-Protokoll vorgeben ist.

Dazu gibt es auch Branchenvorgaben, wie zum Beispiel diejenige der Crypto Valley Association in Zug.

Über die spezifische Organisation der Blockchain, mit einerseits den Erfindern/Initiatoren oder auch Foundern genannt und anderer-

https://doi.org/10.1007/s00287-019-01194-x

seits den Softwareentwicklern oder Minern und in dritter Linie dann den Nodes-Betreibern, wird Aufbau und Betrieb einer Blockchain auch als demokratisches Konstrukt verstanden. Die Softwareentwickler wären dann die Legislative, was die nicht wollen, entwickeln sie nicht, diese Gesetze wird es nie geben. Die Judikative wären dann die Node-Betreiber, sie entscheiden, ob sie eine Softwareentwicklung überhaupt akzeptieren, verwenden. Und die Initiatoren wären dann die Exekutive, die die Blockchain eben grundsätzlich betreibt/ausführt.

Das Abbilden und Betreiben von Inhalten und Prozessen in DLT impliziert und generiert grundsätzlich andere Möglichkeiten, Risiken und Verbindlichkeiten. Dies ergibt sich aus dem parallelen, d. h. gleichzeitig für alle zugänglichen und einsehbaren System und den darin enthaltenen Daten.

Diese Transparenz und Koordination, die diese Technologie mit sich bringt, erzwingt eine andere Vertrags- und Zusammenarbeitskultur als in klassischen Austauschverhältnissen. Die geheime Umsetzung von Eigeninteressen ist auf jeden Fall eingeschränkter möglich.

Diese Transparenz und Koordination bedingt gut durchdachte Verträge für die Konsensbildung und eine großes Commitment zu Vertragstreue auch in nachgelagerten oder vorgelagerten Prozessen.

Bei der Programmierung von Smarten Contracts muss zudem genau der Inhalt durchdacht werden, da Änderungen nur mit großem Aufwand wieder möglich sind.

Nicht zu vergessen ist dabei jeweils, ob noch zusätzlich Hardware und Software mit spezifischen Verträgen für die Funktionsweise des Systems abgeschlossen werden müssen.

DLT sind auch ein geeignetes Vehikel, um Businessprozesse und die dort verwendeten/gehandelten Assets zu dematerialisieren. Die Herausforderung in dieser neuen Welt von Tokens ist nebst der Frage, wie diese rechtlich zu qualifizieren sind, wie eine vertrauensvolle Verbindung zwischen physischer und digitaler Welt abgebildet und rechtlich begründet werden kann.

Fazit

DLT impliziert die Möglichkeit, umfassende Businesssysteme zu gestalten. Hauptmerkmale sind Transparenz und Impact fürs ganze System, bei Nicht-Compliance. Für die beteiligten Akteure ergeben sich dadurch andere und neue Verbindlichkeiten, welche in neuen Arten von Governancesystemen umgesetzt werden. Die Umsetzung kann mit Smarten Contracts direkt auf der Chain oder Offchain erfolgen.

Ursula Sury ist selbständige Rechtsanwältin in Luzern, Zug und Zürich (CH) und Vizedirektorin an der Hochschule Luzern – Informatik. Sie ist zudem Dozentin für Informatikrecht, Datenschutzrecht und Digitalisierungsrecht.

FORUM

Einsichten eines Informatikers von geringem Verstande

Au, Toren schafft Autorenschaft

Alfred Alhelm, Belfried Behelm, Siegfried Schelm und Reinhard Wilhelm
Saarland Informatics Campus, Saarbrücken

Publizierte Ergebnisse und Zitate darauf, mehr als das kümmerliche Gehalt, sind der Lohn des Wissenschaftlers. Die Länge der Publikationsliste und die Zahl der Zitate sind ein unverzichtbarer Bestandteil bei der Vorstellung eines wackeren Forschers. „Dies ist meine Kollege John Doe. Er hat 345 Publikationen in angesehenen Organen veröffentlicht, welche 23 145-mal zitiert worden sind." ist eine typische Vorstellung eines Kollegen in einschlägigen Kreisen. Ist ja schließlich auch wichtiger als zu wissen, ob der Kollege menschlich ein Vollpfosten, sozial ein Blindgänger oder politisch ein ausgewiesener Verschwörungstheoretiker ist. Nicht ganz unwichtig für die Bewertung der Publikationslage ist natürlich, wie, genauer von wem, die publizierten Ergebnisse erzielt wurden. Wissenschaftler in den Geistes- und Sozialwissenschaften sind meist heroische Einzelkämpfer. Da stellt sich die Frage der Autorenschaft eigentlich nicht.

In natur- und ingenieurwissenschaftlichen Disziplinen dagegen werden die meisten Ergebnisse von Teams erzielt und anschließend auch publiziert. In der Informatik erinnern die Längen der Autorenlisten an Tennis: Einzel, Doppel, auch mal Mixed, höchstens ein Daviscup-Team. In den Laborwissenschaften sind Autorenlisten von der Größe von Fußballmannschaften (mit Ersatzspielern) eher die Regel als die Ausnahme. Da finden sich fairerweise auch der Laborant und die studentische Hilfskraft auf der Liste, die am Wochenende die Zellkulturen gefüttert oder die Labormäuse gestreichelt haben. Die letzte Position in der Autorenliste nimmt in der Regel der Chef des Teams, der Professor, Institutsdirektor oder Laborleiter in der Industrie ein. Da er auf jeder der Publikationen zu stehen pflegt, die sein Labor verlassen, kommen auf Dauer sehr eindrucksvolle Publikationslisten zustande. Die Rechtfertigung für seine Anwesenheit auf der Autorenliste besteht meist darin, dass einer der tatsächlich beitragenden Autoren an seiner geöffneten Bürotür vorbeigekommen ist und vom Fluidum des Chefs erfasst wurde. Dieses Fluidum pflegt ungeheure Inspirationen frei zu setzen. Sollten die beitragenden Autoren diesem Fluidum nicht ausgesetzt gewesen sein, bleibt noch die Motivation von prekär beschäftigten Wasserträgern des Wissenschaftsbetriebs, die Autorenliste durch die Hinzufügung des Chefnamens zu veredeln. Mit der letzten Position in der Autorenliste sind gewisse Privilegien verbunden. So verschont sie den Besetzer dieser Position in der Regel von der Verpflichtung, wis-

https://doi.org/10.1007/s00287-019-01196-9

sen zu müssen, was der Inhalt der Publikation ist. Das kommt dann recht händig, wenn sich der Inhalt der Publikation als dubios herausstellt, also getürkte experimentelle Ergebnisse oder verfälschende Interpretationen enthält, „Bin zutiefst enttäuscht! Habe mich auf meine Leute verlassen! Werde in Zukunft …".

Nachdem die letzte Position der Publikationsliste in den experimentellen Wissenschaften hiermit endgültig geklärt sein dürfte, ist die Frage der Reihenfolge der weiteren Autoren noch offen. In den Naturwissenschaften richtet sich die Reihenfolge nach der Größe der Beteiligung an den publizierten Ergebnissen. Nennen wir das die *Beitragsordnung*. Der Erstautor ist der Hauptbeteiligte, oft ein Doktorand. Denn Doktoranden sind das Öl im Motor der Forschung. Ohne sie käme der Betrieb zu einem qualmenden Halt, sozusagen Kolbenfresser allenthalben.

In der Informatik gibt es neben der Beitragsordnung auch die *alphabetische Reihenfolge*. Da wird ein Autor weniger dafür belohnt, dass er führend am Erzielen der Ergebnisse beteiligt war als dafür, dass sein Name in der lexikalischen Ordnung vorne steht.

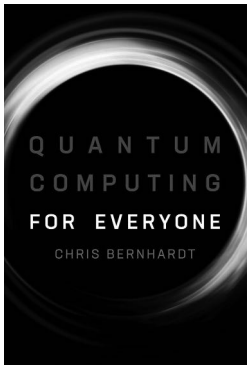
Die meisten Informatiker sind seltsam unentschieden, ob sie jetzt lieber die Beitragsordnung oder die alphabetische Ordnung bevorzugen, natürlich mit Ausnahme der Besitzer alphabetisch privilegierter Anfangsbuchstaben. Bei einer alphabetisch sortierten Autorenliste können sich die relativen Beiträge natürlich auch genau in dieser Reihenfolge verhalten; zum Beispiel könnte die Doktorandin Abel die ganze Arbeit gemacht, ihr Betreuer Babel ihr viele Anregungen gegeben, Kollege Cnabel sie auf einige einschlägige Publikationen hingewiesen, einige weitere, lexikografisch unter ferner liefen einzusortierende Kollegen den beiden über die Schulter gesehen und die Fortschritte kommentiert haben, während Professor Zabel die Verantwortung getragen hatte. Andererseits könnte die Reihenfolge dem Alphabet geschuldet sein. Die Frage ist ohne Kontextwissen nicht entscheidbar. Die Autorenreihenfolge Professor Abel vor Doktorandin Babel, Postdoc Cnabel wird mit einem gewissen Maß an gutem Willen vermutlich so interpretiert werden, dass Professor Abel die Arbeit gemacht hat und in großzügiger Weise die von ihm angeleitete Doktorandin Babel, den Postdoktoranden Cnabel und ein paar weitere Lehrstuhlangehörige auf die alphabetisch sortierte Autorenliste gesetzt hat. Sie wirft eventuell Rätsel auf, wenn man die gesamten Publikationen der Arbeitsgruppe von Professor Zabel betrachtet und ein System hinter den Autorenreihenfolgen zu entdecken glaubt. Aber Unentscheidbarkeit ist ein Problem, mit dem Informatiker und Logiker zu leben gelernt haben.

Dies ist ein Auszug aus meinem demnächst nicht erscheinenden Roman *Künstliche Intelligenz und natürliche Dummheit* oder *Natürliche Intelligenz und künstliche Dummheit*, weiß noch nicht.

FORUM

Rezension

Quantum Computing for Everyone



Chris Bernhardt:
Quantum Computing for Everyone,
MIT Press, Cambridge 2019
ISBN 978-0-262-03925-3

„Quantum Computing is a beautiful fusion of quantum physics with computer science"
(Chris Bernhardt)

„Quantum Computing" ist ein Begriff, der immer häufiger in unserer Computer-Fachliteratur auftaucht. Es ist keine einfache Technologie. Die klassischen Informatiker reagieren darauf entweder mit (1) *Furcht* (da kommt etwas Schlimmes auf uns zu), (2) *Gleichgültigkeit* (das geht noch lange) oder (3) *verhaltenem Interesse* (das müssen wir verstehen lernen).

Das Quantum Computing wird zweifellos unsere Informations-/Communications-Technologiewelt massiv verändern – nur der Zeitpunkt ist noch nicht so klar. Es gibt große Chancen, aber auch signifikante Gefahren lauern dahinter – z. B. die Gefährdung unserer heutigen, weitverbreiteten Kryptoalgorithmen wie RSA.

Über Quantum Computing existiert bereits eine riesige Literaturbasis (Googlesuche „Quantum Computing" ergibt 221.000.000 Ergebnisse [Stand 13.9.2019]). Wieso also *diese* Buchbesprechung und Buchempfehlung?

Zuerst die Motivation: Jeder Informatiker sollte heute die Grundlagen, Implikationen, die möglichen Anwendungen, und die Gefahren des Quantum Computing kennen und verstehen. Dabei stellen sich zwei Fragen: (1) Wie tief soll das Verständnis gehen? Und (2) Wie kommt er mit vertretbarem Zeitaufwand zu diesem Wissen?

Eine gute Antwort ist dieses Buch: Es beginnt mit den *Physikalischen Grundexperimenten*, welche die Quantentheorie begründeten (Kapitel 1). Im 2. Kapitel werden die notwendigen mathematischen Voraussetzungen (lineare Algebra: Vektoren, Matrizen, Tensorprodukt) eingeführt – ein richtiges Verständnis des Quantum Computing ohne das (minimale) mathematische Modell ist nicht möglich, also muss man genügend Zeit in die – sehr gut präsentierte – Auffrischung investieren! Der Autor benutzt die von Paul Dirac eingeführte Notation: |v> für Kolonnenvektoren und <w| für Zeilenvektoren, etwas gewöhnungsbedürftig, aber elegant. Dieser Teil erfordert etwas Durchhaltevermögen, zahlt sich aber im Rest des Buches aus.

Kapitel 3 führt die ersten zwei fundamentalen Konzepte des Quantum Computing ein: *Spin* (Eigendrehimpuls) und *Qubit* (Quantenbit). Mit den vorher gelernten mathematischen Konzepten werden die Mechanik und die Beziehungen zwischen Qbits beschrieben. Das Kapitel bietet immer wieder physikalische Beispiele, um die Mathematik zu illustrieren.

In Kapitel 4 wird das dritte fundamentale Konzept des Quantum Computing eingeführt: *Entanglement* (Verschränkung). Das Entanglement bedeutet, dass die Eigenschaften (z. B.

https://doi.org/10.1007/s00287-019-01229-3

Spin) von zwei oder mehreren Teilchen (Photonen oder Elektronen) voneinander abhängig („verknüpft“) sind, auch wenn die Teilchen durch große Distanzen getrennt sind. Messung einer Eigenschaft eines Teilchens wirkt sich sofort auf das andere Teilchen aus – möglicherweise über große Distanz. Entanglement ist eines der großen unverstandenen Phänomene der Quantenphysik und die Ursache für viele Untersuchungen sowie Veröffentlichungen und bildet die Grundlage für die Quantenphänomene. Hier wird Entanglement mathematisch über das Tensorprodukt $\otimes$ von Vektoren elegant beschrieben.

Das 5. Kapitel ist für das Verständnis nicht notwendig, aber wissenschaftsgeschichtlich spannend. Der Autor führt das Gedankengut der *Bell'schen Ungleichung* ein – eine der wichtigsten Erkenntnisse des 20. Jahrhunderts. Die Ursache ist eine tiefgehende Auseinandersetzung der Befürworter der neuen Quantentheorie mit „Fernwirkungen“: Speziell Albert Einstein und Erwin Schrödinger konnten dieses neue Gedankengut nie akzeptieren. Albert Einstein, Boris Podolsky und Nathan Rosen griffen diese neue Theorie 1935 in einem brillanten Paper mit dem Vorwurf der Unvollständigkeit an (Original zum Download: https://journals.aps.org/pr/abstract/10.1103/PhysRev.47.777). Der Physiker John Stewart Bell erkannte als erster, dass die Messung einer großen Anzahl von Qubits zu verschiedenen Wahrscheinlichkeiten führt, je nachdem, ob man die Berechnung mit dem klassischen Modell oder mit dem Quantenmodell durchführt. Die Bell'sche Ungleichung definiert damit einen experimentellen Ansatz zum Entscheid „klassische Theorie $\Leftrightarrow$ Quantentheorie“. Der Physiker Alain Aspect erfand 1980 einen Test, um die Bell'sche Ungleichung zu prüfen, der in sensationeller Weise klar die Richtigkeit der Theorie der verschränkten Qubits (= Quantentheorie) bewies! Da das Aspects-Experiment verschiedene Mängel aufwies, wurden weitere Experimente von anderen Physikern durchgeführt – bis heute ist jeder Zweifel an der Korrektheit der Qubit-Verschränkung ausgeräumt.

Kapitel 6 führt zuerst die klassische Boole'sche Logik und die logischen Schaltungen ein (für Informatiker Schulwissen). Neu werden in diesem Kapitel die „reversible gates“ (reversible logische Gatter) erklärt. Ein reversibles Gate erlaubt den eindeutigen Rückschluss auf die Inputs aus den Outputs. Ein Konzept, das später bei den Quantum Gates wichtig wird.

Kapitel 7 ist heavy: Hier wird das Verfahren zur Verarbeitung von Qubits in Quantum Gates mit der Mathematik aus Kapitel 1 und 4 voll durchgespielt. Man versteht, wie Entanglement entsteht, kommuniziert und schließlich aufgelöst wird. Die Anfälligkeit von physikalischen Qubits auf Zerstörung durch Umwelteffekte wird durch fehlerkorrigierende Codierung aufgefangen – eine wichtige Technologie für reelle Quantensysteme.

Das 8. Kapitel erklärt einige Quantumalgorithmen. Wichtig ist, dass klassische (von-Neumann-)Algorithmen *nicht* auf Quantumcomputer übertragen werden können. Die Verwendung von Quantumcomputern verlangt nach ganz neuen Algorithmen, welche die Eigenschaften der Qubits – die sehr hohe Parallelität – ausnutzen. Zurzeit existieren relativ wenige Algorithmen, die sinnvolle Anwendungen erlauben. Der Autor beschreibt – allerdings nur sehr summarisch – einige der heute bekannten Algorithmen: (1) David Deutschs Algorithmus zur Identifikation einer Funktion aus einer Menge von Funktionen mittels minimaler Anzahl von Funktionsauswertungen, (2) Deutsch-Jozsa-Algorithmus als Generalisierung des Deutsch-Algorithmus auf Funktionen von n Variablen, (3) Simons Algorithmus sucht nach einer Funktion zur Generierung von Bitstrings, ausgehend von einem unbekannten, geheimen Bitstring. Wieder geht es um die Anzahl notwendiger Evaluationen, bis die Funktion und der Bitstring gefunden sind, (4) Shors Algorithmus zur schnellen Primzahlfaktorisierung von großen Zahlen. Hier bringt ein Quantumalgorithmus unglaubliche Zeitgewinne und gefährdet die moderne Kryptografie.

Das letzte Kapitel geht kurz auf die möglichen Auswirkungen der Quantencomputer ein: Dieses Kapitel ist sehr kurz und illustriert die große Gefahr des Shor'schen Faktorisierungsalgorithmus für unsere gesamte Finanzinfrastruktur. Der zweite Algorithmus stammt von Grover und erlaubt die effiziente Suche in riesigen Datenmengen (Big Data). Quantumsimulationen, speziell im Gebiet der Chemie (Computational Chemistry) versprechen wichtige Resultate, sobald Quantencomputer mit genügender Anzahl von Qubits verfügbar sind.

Das Buch geht bewusst nicht auf die *Implementation* von Qubits ein: Diese Technologie ist im Fluss und es gibt genügend aktuelle Literatur. Allerdings ist interessant, dass eine spezifische *Programmiersprache* für Quantumrechner bereits besteht: „Qiskit“ (www.qiskit.org).

Der Buchtitel „Quantum Computing for Everyone“ ist nicht ganz zutreffend. Der Leser muss Grund-

kenntnisse in linearer Algebra und Elementarteilchenphysik besitzen. Zudem muss er bereit sein, genügend Zeit mit Bleistift und Papier zu investieren, um die Herleitungen zu verstehen. Die Freude am zunehmenden Verständnis ist der Lohn dafür!

Das Buch ist wunderbar aufgebaut und führt den Leser zielsicher durch das notwendige Wissen bis zu einigen absehbaren, durchschlagenden Anwendungen. Die Darstellung ist klar und es werden nur die wirklich notwendigen Konzepte eingeführt. Am Schluss des Buches – das doch einiges an Durchhaltevermögen verlangt – verfügt der Leser über ein solides Verständnis des Quantum Computing. Dies erlaubt ihm, die zahlreich erscheinenden Publikationen in den nächsten Jahren zu verstehen und zu beurteilen. Anstelle einer Anzahl oberflächlicher Veröffentlichungen zu lesen (jede Menge vorhanden) investiert man diese Zeit wesentlich besser in diese 189 Seiten! Das Buch ist sehr wertvoll, auch wenn man nicht jedes Kapitel im Detail versteht.

Wir sind heute (2020) an einem Punkt angelangt, an dem jeder Informatiker Quantum Computing und seine absehbaren Folgen verstehen sollte.

„*It will become accepted that there is a more fundamental level of computing – and the most elemental level of computing involves qubits, entanglement, and superpositions*“
(Chris Bernhardt)

Frank J. Furrer,
Dresden 2019

Gesellschaft für Informatik e.V. (GI), Wissenschaftszentrum, Ahrstraße 45, 53175 Bonn, Tel.: 0228/302-145, Fax 0228/302-167,
e-mail: gs@gi.de, Server: http://www.gi.de
Geschäftsführung: Cornelia Winter, e-mail: cornelia.winter@gi.de, Tel.: -145, Daniel Krupka, e-mail: daniel.krupka@gi.de, Tel.: 030-7261566-15
Sekretariat: Kerstin Schimmel, e-mail: kerstin.schimmel@gi.de, Tel.: -145
Mitgliederwerbung: Ludger Porada, e-mail: ludger.porada@gi.de, Tel.: -146
Finanzen/Buchhaltung: Elena Kerkmann, e-mail: elena.kerkmann@gi.de, Tel.: -153, Swetlana Ruppel, e-mail: swetlana.ruppel@gi.de, Tel.: -152
Mitgliederverwaltung: Tanja Diegeler, e-mail: tanja.diegeler@gi.de, Tel.: -149, Andrea Mertins, e-mail andrea.mertins@gi.de, Tel.: -151
ITK: Viktor Schröder, e-mail: viktor.schroeder@gi.de, Tel.: -156, Karl-Heinz Künkel, e-mail: karl-heinz.kuenkel@gi.de, Tel.: -143

Aus Vorstand und Präsidium

Beschlüsse aus der Präsidiumssitzung vom 27./28. Juni 2019 in Bonn:

- Das Präsidium verabschiedet den vorgelegten Entwurf zum Referenzrahmen Informatik mit drei Enthaltungen.
- Das Präsidium stimmt der Änderung der Wahlordnung zu.
- Das Präsidium benennt folgende Personen für die Kandidatenfindungskommission Präsidium für die Wahl 2020:
  - David Richter
  - Detlef Lippert
  - Dr. Ursula Köhler
  - Prof. Dr. Daniela Nicklas (Sprecherin)
  - Bernhard C. Witt
- Das Präsidium benennt folgende Personen für den Wahlausschuss für die Wahl 2020:
  - Prof. Dr. Alexander Rossnagel (Vorsitz)
  - Prof. Dr. Rüdiger Grimm (Stellvertretung)
  - Prof. Dr. Gerd Stumme
  - Karl-Heinz-Künkel
  - Viktor Schröder
  - Cornelia Winter
- Folgende Personen werden bestätigt bzw. benannt:
  - Achim Rettberg als GI-Vertreter in TC10 der IFIP (Nachfolge Rammig)
  - Prof. Dr. Oliver Thomas als Sprecher des FB WI (Nachfolge Strecker)
  - Prof. Dr. Ali Sunyaev als Stellvertretender Sprecher des FB WI (Nachfolge Thomas)
  - David Richter als Sprecher des FB RVI (Nachfolge Wimmer)
  - Tanja Krins als stellvertretende Sprecherin des FB RVI
  - Prof. Dr. Roland Vollmar als GI-Vertretung in Schloss Dagstuhl
  - Robert Heinrich als Sprecher der Regionalgruppe Karlsruhe
  - Steffen Schilke als Sprecher der Regionalgruppe Rhein-Main

Zur finanziellen Lage der GI

1. August 2019

Zusammenfassung

Das Jahr 2018 wurde mit einem Überschuss in Höhe von ca. 121 T€ abgeschlossen. Dieses positive Ergebnis setzt die positive Entwicklung aus 2018 fort. Ursprünglich war der Plan für 2018, einen kleinen Überschuss von ca. 35 T€ zu erzielen. Dies bedeutet für die GI eine weitere positive Entwicklung nach fünf Verlustjahren vor 2017. Auch dieser Überschuss ist das Ergebnis verschiedener Maßnahmen. Neben Kostensenkungen und Ausbau des Projektbereiches ist die Erhöhung der Mitgliedsbeiträge vor zwei Jahren für diese Verbesserung der Einnahmenseite ausschlaggebend.

Die Mitgliederzahl ist 2018 wiederum leicht gesunken, die Maßnahmen der vergangenen Jahre zur Mitgliedergewinnung zeigten bisher nur geringe Wirkung. Es zeichnet sich jedoch ein erfreulicher Trend in 2019 ab bei der Gewinnung von Mitgliedern unter den Studierenden. Hier zeigt der Beschluss, diese Mitgliedschaften kostenfrei für die Studierenden zu ermöglichen, eine sehr positive Wirkung. Dies muss anspornen, die neuen Mitglieder langfristig zu halten. Daher gilt nach wie vor, dass trotz der im Moment immer noch positiven finanziellen Lage der GI auch weiterhin sparsames Haushalten und Maßnahmen bzw. Investitionen zur Steigerung der Attraktivität der GI sowie zur Gewinnung neuer Mitglieder durchgeführt werden müssen.

Jahresabschluss 2018

Im Haushaltsjahr 2018 haben die Gesamterträge von 3.405 T€ im Vorjahr um 213 T€ auf 3.618 T€ zugenommen. Gleichzeitig sind die Aufwendungen von 3.103 T€ im Vorjahr um 394 T€ auf 3.497 T€ gestiegen. Insgesamt ergibt sich somit für 2018 ein Überschuss von 121 T€.

Die Zunahme der Erträge im Vergleich zu 2017 ergibt sich einerseits aus dem moderateren Sinken des Beitragsaufkommens und wesentlich höheren Einnahmen bei den Zuschüssen (Projekte) andererseits. Weitere positive finanzielle Effekte entstanden durch Einsparungen in der Mitgliederverwaltung und im Betrieb.

Wie bereits in den Jahren seit 2012 erfolgte auch 2018 keine Gewinnausschüttung durch die DLGI (nach einer letztmaligen Ausschüttung in Höhe von 203 T€ in 2011).

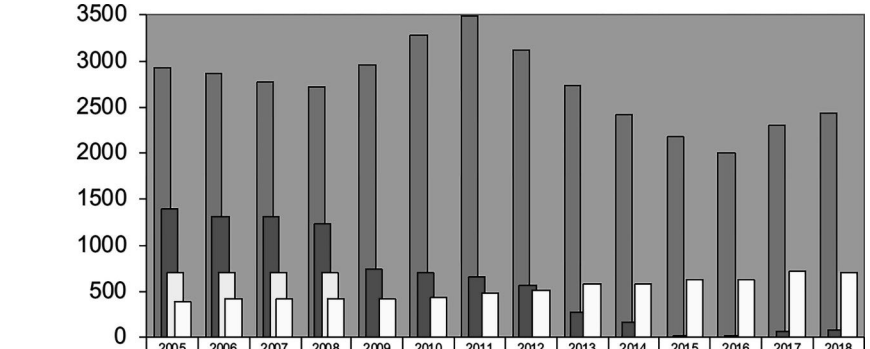
Auf der Aufwandsseite haben 2018 die Aufwendungen für Tagungen (116 T€) und die Personalkosten

https://doi.org/10.1007/s00287-019-01195-w

(248 T€) zugenommen. Die Entwicklung bei den Kosten von Tagungen ist im Rahmen der großen Tagungen zu sehen, die nicht in einem jährlichen Zyklus stattfinden, sowie die Jahrestagung der GI, die in 2018 außer der Reihe in Berlin – bedingt durch den Ausfall eines lokalen Veranstalters – direkt durch die Berliner Geschäftsstelle organisiert wurde. Die Zunahme der Kosten in den Projekten ist durch den personellen Zuwachs in diesen Projekten bedingt. Ähnliches gilt für die Bundeswettbewerbe im BWINF, die sehr erfolgreich sind und daher auch höhere Kosten (146 T€) haben. Zurückgegangen sind die Aufwendungen für Publikationen (–156 T€) und Abschreibungen (–21 T€).

Das per Saldo im Jahr 2018 erzielte leicht negative Tagungsergebnis in Höhe von ca. 24 T€ ist i. W. bedingt durch die erwähnte Ausrichtung der Jahrestagung in Berlin. Die Veranstaltung wurde jedoch als ganz wesentlich für die generelle Wahrnehmung der GI und damit auch als Erfolg erachtet. Das Präsidium würde eine Wiederholung der GI-Jahrestagung in diesem Format und in Berlin in regelmäßigen Abständen begrüßen. Neben dem traditionellen Format der Jahrestagung wäre das Format in Berlin ein attraktiver Baustein, um die Bemühungen zu verstärken, in vielerlei Beziehung attraktive Tagungen für unsere Mitglieder auszurichten.

Die Mitgliederzahl ist 2018 von 15.519 auf 15.122 persönliche Mitglieder gesunken, dabei sind 237 beitragsfreie Mitgliedschaften mitgezählt. Die Zahl der korporativen Mitglieder ist um 5 auf 223 zurückgegangen. Die Steigerung der Attraktivität der GI, die Gewinnung von Neumitgliedern sowie die Mitgliederbindung bleiben daher auch in Zukunft zentrale Herausforderungen. Wie bereits oben erwähnt, ist die kostenfreie Mitgliedschaft momentan ein Erfolgsmodell. Jedoch muss durch besondere Anstrengungen dafür gesorgt werden, dass dieser Trend anhält und die neuen Mitglieder langfristig dabeibleiben.



	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018
Ges.vermögen	2931	2868	2770	2723	2960	3281	3484	3127	2741	2413	2178	2007	2309	2430
zweckgeb. Rücklagen	1394	1312	1310	1230	730	707	657	557	281	161	16	16	66	86
davon Immobilienrückl.	700	700	700	700	0	0	0	0	0	0	0	0	0	0
Rücklagen Fachglied.	390	410	419	414	414	430	475	510	581	577	623	630	720	713

Abb. 1 Vermögensentwicklung 2005–2018

Vermögensentwicklung

Das Vereinsvermögen der GI ist durch den Gewinn 2018 gegenüber dem Vorjahr von 2.309 T€ auf 2.430 T€ gestiegen. Die Entwicklung des GI-Vermögens in den Jahren 2005 bis 2018 wird in Abb. 1 dargestellt.

Im Vereinsvermögen enthalten sind interne Rücklagen der Fachgliederungen, die 2018 nur leicht gesunken sind. Es muss darauf geachtet werden, dass aus den Rücklagen auch Entnahmen für zeitnahe Aktivitäten der Fachgliederungen erfolgen.

Die zweckgebundenen Rücklagen wurden leicht erhöht (auf 86 T€), um erwartete Arbeiten im Bereich der IT abzudecken, und enthalten u. a. auch in 2018 unverändert 16 T€ als Rücklage für den inzwischen eingestellten ICSI-Beirat. Dieser Betrag soll für GI-Nachwuchsarbeit verwendet werden.

Rechnungsprüfung 2019

Die Jahresrechnung und die zugrundeliegende Buchführung der Gesellschaft für Informatik für das Geschäftsjahr vom 1. Januar 2018 bis zum 31. Dezember 2018 wurden von Freudenhammer Maas & Partner mbB im Hinblick auf die Einhaltung der Rechenschaftslegungsgrundsätze für Vereine sowie satzungsmäßiger Regelungen geprüft und am 18. April 2019 bescheinigt.

Die Prüfung durch die von der Mitgliederversammlung bestellten internen Rechnungsprüfer Prof. Michael Meier und Prof. Ali Sunyaev fand am 24. Mai 2019 in der GI-Geschäftsstelle in Bonn statt. In ihrem Bericht bestätigen die Rechnungsprüfer die Ordnungsmäßigkeit der Haushaltsabwicklung sowie der kaufmännischen Buchführung für das Haushaltsjahr 2018.

Wirtschaftsplan 2019

Im Jahr 2019 ergibt sich eine gegenüber den Planungen für 2019 eine leicht günstigere Situation bei den Einnahmen und Ausgaben für die Mitgliederverwaltung. Unter Berücksichtigung der zu erwartenden Mitgliederzahlen und Verringerung der Ausgaben für die Verwaltung der Mitglieder ist gegenüber der ursprünglichen Planung in diesem Bereich ein leichtes Plus von etwa 8 T€ zu erwarten.

Bei den weiteren Ausgaben haben sich jedoch zusätzliche, bisher nicht erwartete Mehrausgaben ergeben. Die Situation ist dadurch gekennzeichnet, dass einerseits weiterhin mit einem niedrigen Zinsniveau und

Wirtschaftsplan 2019

01.08.19

1) Budget Zentral

Kostenstelle	Einnahmen				Ausgaben				Saldo
	Beiträge T€	Tagungen T€	Sonstige T€	Summe T€	Mitgl.-bezg. T€	Tagungen T€	Sonstige T€	Summe T€	T€
1. Mitglieder	1.373,9	-	-	1.373,9	261,5	-	-	261,5	1.112,4
2. Gliederungen									
2.1 Fachbereiche	0,0	0,0	0,0	0,0	0,0	0,0	42,0	42,0	-42,0
2.2 Regionalgruppen	12,0	0,0	0,0	12,0	0,0	0,0	36,0	36,0	-24,0
2.3 Beiräte u. Anwendergruppen	0,0	0,0	0,0	0,0	0,0	0,0	33,0	33,0	-33,0
3. Leitungsorgane									
3.1 Präsidium	-	0,0	0,0	0,0	-	0,0	34,0	34,0	-34,0
3.2 Vorstand	-	0,0	0,0	0,0	-	0,0	9,0	9,0	-9,0
3.3 Kuratorium	-	0,0	0,0	0,0	-	0,0	0,0	0,0	0,0
3.4 GS Bonn, Berlin	-	-	45,0	45,0	-	-	905,0	905,0	-860,0
4. Div. Fachaktivitäten									
4.1 Projekte	-	-	758,0	758,0	-	-	873,0	873,0	-115,0
4.2 Öffentlichkeitsarbeit	-	-	0,0	0,0	-	-	20,0	20,0	-20,0
4.3 Zeitschriften und and.	119,8	-	0,0	119,8	115,3	-	0,0	115,3	4,5
4.4 Sonstige	-	0,0	2,0	2,0	-	0,0	26,0	26,0	-24,0
Summe	**1.505,7**	**0,0**	**805,0**	**2.310,7**	**376,8**	**0,0**	**1.978,0**	**2.354,8**	**-44,1**

2) Budget Fachgliederungen *)										**IST Rücklage per 1.1.2019**	**Geplante Rücklage per 31.12.19**
Fachbereich GInf	4,3	0,0	2,0	6,3	3,6	1,0	2,1	6,7	-0,4	38,3	37,9
Fachbereich KI	30,5	0,0	0,0	30,5	16,5	10,0	4,0	30,5	0,0	64,6	64,6
Fachbereich SWT	16,0	0,0	0,6	16,6	15,1	0,0	1,5	16,6	0,0	65,7	65,7
Fachbereich MCI	21,5	0,0	3,5	25,0	37,5	0,0	9,5	47,0	-22,0	217,7	195,7
Fachbereich DBIS	19,3	0,0	0,0	19,3	15,2	0,0	1,1	16,3	3,0	7,6	10,6
Fachbereich TI	4,7	0,0	0,0	4,7	1,9	1,0	4,8	7,7	-3,0	82,4	79,4
Fachbereich BS, KuVS	2,3	1,3	0,0	3,6	0,0	0,0	5,1	5,1	-1,5	38,9	37,4
Fachbereich ILW	0,0	0,0	0,0	0,0	0,0	0,0	1,0	1,0	-1,0	7,5	6,5
Fachbereich WI	9,9	0,0	0,0	9,9	6,4	0,0	3,5	9,9	0,0	46,3	46,3
Fachbereich RVI	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
Fachbereich IAD	9,7	0,0	0,0	9,7	0,0	0,0	9,7	9,7	0,0	50,1	50,1
Fachbereich IUG	5,5	0,0	0,0	5,5	3,7	0,0	1,8	5,5	0,0	7,2	7,2
Fachbereich GDV	0,0	0,0	0,0	0,0	0,0	0,0	1,0	1,0	-1,0	15,1	14,1
Fachbereich SICHERHEIT	0,0	0,0	0,0	0,0	0,0	2,0	3,0	5,0	-5,0	58,5	53,5
Summe 2	**123,7**	**1,3**	**6,1**	**131,1**	**99,9**	**14,0**	**48,1**	**162,0**	**-30,9**	**699,9**	**669,0**

3) Budget Gesamt (Summen 1 + 2)

	1.629,4	**1,3**	**811,1**	**2.441,8**	**476,7**	**14,0**	**2.026,1**	**2.516,8**	**-75,0**

*) jeweils kumuliert über alle Untergliederungen

Abb. 2 Angepasster Wirtschaftsplan 2019

keinen Einnahmen von der DLGI zu rechnen ist, andererseits weitere bisher nicht geplante Kosten bei den Projekten anfallen werden. Es konnten Einsparungen erzielt werden, sodass die Personalkosten insgesamt etwas geringer als geplant gestiegen sind und auch eine halbe Stelle nicht wie vorgesehen aus dem GI – Haushalt direkt, sondern durch ein Projekt finanziert werden konnte. Andererseits fallen durch die Anmietung neuer Arbeitsräume in der Geschäftsstelle in Berlin und mehrere kleinere Vorhaben weitere Kosten an.

In den Fachbereichen wird gegenüber dem ursprünglichen Plan etwas weniger ausgegeben, so dass in der Summe für das Haushaltsjahr 2019 ein kleiner verkraftbarer Verlust von ca. 75 T€ zu erwarten ist (Details siehe Abb. 2).

Haushaltsentwurf 2020

Bereits 2016 haben Präsidium und Vorstand beschlossen, das Informatik Spektrum den Mitgliedern ab 2018 im Rahmen der Mitgliedschaft ausschließlich digital zur Verfügung zu stellen. Dies soll weiter fortgeführt werden, da diese Vertriebsform neben ein paar kritischen Stimmen doch eher allgemein auf Zustimmung stößt und weiterhin gegenüber dem Papier-basierten Versand erhebliche Kosteneinsparungen ermöglicht.

Ein weiterer wichtiger Punkt für die Erstellung des Haushaltsentwurfs ist der Antrag des Vorstands,

Haushaltsentwurf 2020

01.08.19

1) Budget Zentral

Kostenstelle	Einnahmen				Ausgaben				Saldo
	Beiträge T€	Tagungen T€	Sonstige T€	Summe T€	Mitgl.-bezg. T€	Tagungen T€	Sonstige T€	Summe T€	T€
1. Mitglieder	1.359,6	-	-	1.359,6	247,9	-	-	247,9	1.111,7
2. Gliederungen									
2.1 Fachbereiche	0,0	0,0	0,0	0,0	0,0	0,0	42,0	42,0	-42,0
2.2 Regionalgruppen	12,0	0,0	0,0	12,0	0,0	0,0	36,0	36,0	-24,0
2.3 Beiräte u. Anwendergruppen	0,0	0,0	0,0	0,0	0,0	0,0	33,0	33,0	-33,0
3. Leitungsorgane									
3.1 Präsidium	-	0,0	0,0	0,0	-	0,0	35,5	35,5	-35,5
3.2 Vorstand	-	0,0	0,0	0,0	-	0,0	9,0	9,0	-9,0
3.3 Kuratorium	-	0,0	0,0	0,0	-	0,0	0,0	0,0	0,0
3.4 GS Bonn, Berlin	-	-	45,0	45,0	-	-	960,0	960,0	-915,0
4. Div. Fachaktivitäten									
4.1 Projekte	-	-	778,0	778,0	-	-	860,0	860,0	-82,0
4.2 Öffentlichkeitsarbeit	-	-	0,0	0,0	-	-	20,0	20,0	-20,0
4.3 Zeitschriften und and.	119,8	-	0,0	119,8	115,3	-	0,0	115,3	4,5
4.4 Sonstige	-	0,0	2,0	2,0	-	0,0	26,0	26,0	-24,0
Summe	**1.491,4**	**0,0**	**825,0**	**2.316,4**	**363,2**	**0,0**	**2.021,5**	**2.384,7**	**-68,3**

2) Budget Fachgliederungen *)										**Geplante Rücklage per 1.1.2020**	**Geplante Rücklage per 31.12.19**
Fachbereich GInf	4,3	0,0	2,0	6,3	3,6	1,0	2,1	6,7	-0,4	37,9	37,5
Fachbereich KI	30,5	0,0	0,0	30,5	16,5	10,0	4,0	30,5	0,0	64,6	64,6
Fachbereich SWT	16,0	0,0	0,6	16,6	15,1	0,0	1,5	16,6	0,0	65,7	65,7
Fachbereich MCI	21,5	0,0	3,5	25,0	37,5	0,0	9,5	47,0	-22,0	195,7	173,7
Fachbereich DBIS	19,3	0,0	0,0	19,3	15,2	0,0	1,1	16,3	3,0	10,6	13,6
Fachbereich TI	4,7	0,0	0,0	4,7	1,9	1,0	6,8	9,7	-5,0	79,4	74,4
Fachbereich BS, KuVS	2,3	0,0	0,0	2,3	0,0	0,0	2,6	2,6	-0,3	37,4	37,1
Fachbereich ILW	0,0	0,0	0,0	0,0	0,0	0,0	1,0	1,0	-1,0	6,5	5,5
Fachbereich WI	9,9	0,0	0,0	9,9	6,4	0,0	3,5	9,9	0,0	46,3	46,3
Fachbereich RVI	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
Fachbereich IAD	9,7	0,0	0,0	9,7	0,0	0,0	9,7	9,7	0,0	50,1	50,1
Fachbereich IUG	5,5	0,0	0,0	5,5	3,7	0,0	1,8	5,5	0,0	7,2	7,2
Fachbereich GDV	0,0	0,0	0,0	0,0	0,0	0,0	1,0	1,0	-1,0	14,1	13,1
Fachbereich SICHERHEIT	0,0	0,0	0,0	0,0	0,0	2,0	3,0	5,0	-5,0	53,5	48,5
Summe 2	**123,7**	**0,0**	**6,1**	**129,8**	**99,9**	**14,0**	**47,6**	**161,5**	**-31,7**	**669,0**	**637,3**

3) Budget Gesamt (Summen 1 + 2)

1.615,1	**0,0**	**831,1**	**2.446,2**	**463,1**	**14,0**	**2.069,1**	**2.546,2**	**-100,0**	

*) jeweils kumuliert über alle Untergliederungen

Abb. 3 Haushaltsentwurf 2020

die Mitgliedsgebühren um etwa 3 % zu erhöhen. Bei der vorherigen Erhöhung der Mitgliedsbeträge war auch beschlossen worden, eine stetige Anpassung der Mitgliedsbeiträge orientiert beispielsweise an die Inflationsrate vorzunehmen. Laut Statistischem Bundesamt lag die Inflationsrate[1] in 2017 bei 1,5 % und 2018 bei 1,8 %. Eine Erhöhung um 3 % bleibt damit unter der kombinierten Inflationsrate über diese beiden Jahre. Auf der Sitzung des Präsidiums wurde diese Erhöhung beschlossen und wird der Ordentlichen Mitgliederversammlung auf der INFORMATIK 2019 in Kassel zur Bestätigung vorgelegt.

Damit ergeben sich folgende Eckpunkte für die Haushaltsplanung. Auf der Einnahmenseite wird für den Haushaltsentwurf von einem Rückgang der vollzahlenden GI-Mitglieder um etwa 500 und weiterhin von einem Verzicht auf Beiträge für Studierende ausgegangen. Durch die gleichzeitig moderate Erhöhung der Beiträge ergibt sich ein leichter Rückgang dieser Einnahmen (ca. 14 T€). Angenommen wird weiter eine Steigerung der Personalkosten in den Geschäftsstellen durch Gehaltssteigerungen um 3 %, möglicherweise die Kosten für eine ½ Mitarbeiterstelle (Öffentlichkeitsarbeit) in der Geschäftsstelle

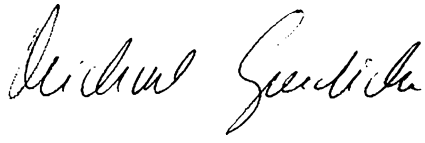
[1] Siehe https://de.statista.com/statistik/daten/studie/1046/umfrage/inflationsrate-veraenderung-des-verbraucherpreisindexes-zum-vorjahr/, abgerufen am 1.8.2019.

in Berlin sowie erhöhte Mietkosten für die Räumlichkeiten in Berlin. Außerdem sind Ausgaben für die Weiterentwicklung und Pflege der Mitgliederverwaltungssoftware aus der dafür vorgesehenen Rücklage vorgesehen. Bei den Projekten ergeben sich kleinere Änderungen durch einen neuen Vertrag BWInf (–5 T€), Digitalisierung der FiBu (–10 T€). Bei den Projekten wird angenommen, dass sie im Wesentlichen die diesbezüglichen Gemeinkosten decken. Die anderen Positionen auf Ein- und Ausgabenseite sind weitgehend aus dem laufenden Jahr übernommen worden.

Das Präsidium hat in seiner Sitzung am 28. Juni 2019 in Bonn dem in Abb. 3 dargestellten Haushaltsentwurf 2019 zugestimmt. Er wird der Mitgliederversammlung 2019 zur Genehmigung vorgelegt. Als Ergebnis für das Jahr 2020 wird ein Verlust in Höhe von 100 T€ erwartet. Dabei ist ein geplanter Abbau von Vermögen der Fachgliederungen in Höhe von knapp 32 T€ zur Finanzierung von Vereinsaktivitäten bereits berücksichtigt.

Abschließen möchte ich diesen Bericht mit einem herzlichen Dank an die in der Geschäftsstelle für die Finanzen zuständigen Mitarbeiterinnen für ihre Unterstützung, die Bewältigung immer neuer Herausforderungen und die stets sehr angenehme Zusammenarbeit.

Essen, im August 2019



Prof. Dr. Michael Goedicke
Universität Duisburg-Essen, Institut für Informatik und Wirtschaftsinformatik und
paluno, The Ruhr Institute for Software Technology

Presse- und Öffentlichkeitsarbeit der GI

GI-Umfrage: Wenn KI ein Geschlecht hätte, wäre sie männlich (17.7.2019)

Die Gesellschaft für Informatik e. V. (GI) hat im Wissenschaftsjahr 2019 – Künstliche Intelligenz des Bundesministeriums für Bildung und Forschung in einer repräsentativen Allensbach-Umfrage gefragt, ob die deutsche Bevölkerung KI-Maschinen eher als weiblich oder männlich wahrnimmt.

Sie schlagen uns Restaurants vor, leiten uns durch den Verkehr und beantworten uns alltägliche Fragen: Digitale Sprachassistenten mit Künstlicher Intelligenz (KI) finden sich heute bereits in Autos, Handys oder Uhren. Auffallend ist: In den meisten Fällen sind die assistierenden Computer wie Siri, Alexa oder Google Assistant per Werkseinstellung mit weiblichen Stimmen ausgestattet. Um zu erfahren, ob die KI-Systeme allgemein als eher männlich oder eher weiblich wahrgenommen werden, hat die Gesellschaft für Informatik im Rahmen des Projektes „#KI50 – Künstliche Intelligenz – gestern, heute, morgen“ im Wissenschaftsjahr 2019 das Allensbach-Institut mit einer repräsentativen Bevölkerungsumfrage beauftragt. Mit #KI50 will die GI in Anlehnung an ihr 50-jähriges Bestehen dazu anregen, über die deutsche KI-Geschichte zu reflektieren, einen Blick nach vorne zu werfen und das Thema einer breiten Öffentlichkeit besser zugänglich zu machen.

KI wird sechsmal häufiger als männlich wahrgenommen

Zwar weist eine Mehrheit der Deutschen KI noch kein Geschlecht zu, rund ein Fünftel der Befragten nehmen KI-Maschinen aber als „eher männlich“ (19,3 %) wahr. Das sind fast sechsmal so viele Menschen, wie die, die KI als „eher weiblich“ (3,5 %) einordnen. Ein Ungleichgewicht, das sich durch die gesamte Bevölkerung zieht, egal ob alt oder jung, Mann oder Frau, Ost oder West. Lediglich eine Ausnahme zeigen die Daten: Bei Science-Fiction-Kennern (1) liegt der Wert derjenigen, die KI als „eher weiblich“ einordnen mit 7,2 % doppelt so hoch, wie in der Gesamtbevölkerung.

Die kompletten Pressemitteilungen der GI finden Sie unter https://gi.de/aktuelles/presse/.

Aus den GI-Gliederungen

Neue Leitung der Fachgruppe „Internet und Gesellschaft“

Die Fachgruppe Internet und Gesellschaft hat Ende Mai ein neues Sprecherinnen-Team gewählt: Das Präsidiumsmitglied Agata Królikowksi, Referentin beim IT-DLZ Bayern, ist stellvertretende Specherin. Dr. Andrea Knaut, Leiterin des mobilen Bildungsprojekts Turing-Bus der GI und der OKF, ist Sprecherin. Bei Interesse an Mitarbeit in der Fachgruppe wenden Sie sich bitte an info@fg-internet.gi.de.

Treffen der Fachgruppe Requirements Engineering

Das jährliche Treffen der Fachgruppe RE findet am 28./29.11.2019 am Institut für Informatik der Universität Heidelberg unter dem Leitthema „Innovatives RE für die Herausforderungen der Zukunft“ statt. Neben zahlreichen interessanten Vorträgen zu aktuellen Entwicklungen des RE gibt es wieder die Möglichkeit zu

intensivem Austausch und Diskussionen. Weitere Informationen zum Programm und zur Anmeldung siehe: https://fg-re.gi.de/veranstaltung/treffen-2019/.

Personalia

Oliver Thomas neuer Sprecher des Fachbereichs Wirtschaftsinformatik

Seit dem 1. Mai 2019 hat der GI-Fachbereich Wirtschaftsinformatik einen neuen Sprecher. Als Nachfolger von Stefan Strecker hat Oliver Thomas den Vorsitz übernommen. Als Stellvertreter wurde Ali Sunyaev gewählt. Nach einstimmiger Wahl stellte Prof. Thomas seine Ziele für seine Amtszeit vor, in der er unter anderem die Marke „WI" stärken und somit für die Öffentlichkeit präsenter machen wird.

„Ich freue mich über das entgegengebrachte Vertrauen und auf die bevorstehenden Aufgaben, um die Marke „WI" zukunftsweisend zu positionieren." sagte Prof. Thomas im Anschluss an die Wahl. Der Fachbereich Wirtschaftsinformatik ist einer der großen Fachbereiche in der GI und hat das Ziel, die Themen der Wirtschaftsinformatik innerhalb der GI und darüber hinaus weiterzuentwickeln und zu vertreten, bspw. durch die Organisation von Konferenzen, Herausgabe von wissenschaftlichen Zeitschriften und engem Austausch mit Politik und Wirtschaft. Als Schnittstellendisziplin zwischen Informatik und Betriebswirtschaft steht die Anwendungsorientierung und somit die praktische Einsetzbarkeit bspw. in Unternehmen im Fokus. „Ein Ziel muss es daher sein, die Kooperationen auf unterschiedlichen Ebenen voranzutreiben. Insbesondere mit Unternehmen, aber auch in der Ausbildung von Nachwuchskräften in Schule und Universität", sagte Thomas.

Abb. 4 Bild (v.l.n.r.): Prof. Dr. Stefan Strecker (FernUniversität Hagen), Prof. Dr. Jan Marco Leimeister (Universität Kassel), Prof. Dr. Oliver Thomas (Universität Osnabrück) und Prof. Dr. Ali Sunyaev (Karlsruher Institut für Technologie). Bildquelle: Fachgebiet Informationsmanagement und Wirtschaftsinformatik (IMWI), Universität Osnabrück

Förderpreis des Fachbereichs „Informatik in den Lebenswissenschaften" verliehen

Der diesjährige Förderpreis des Fachbereichs Informatik in den Lebenswissenschaften (ILW) für die beste Masterarbeit geht an Johanna Schwarz aus Marburg für ihre Arbeit zum Thema „Projecting Machine Learning Scores to Well-Calibrated Probability Estimates". Aus den hervorragenden Bewerbungen um den ILW Förderpreis wurde darüber hinaus die Masterarbeit von Aline Sindel aus Erlangen mit dem 2. Preis ausgezeichnet. Frau Sindel erhält den Preis für ihre Masterarbeit zum Thema „Learning-based Image Super-Resolution for 3-D Magnetic Resonance Imaging".

(Richard Lenz, Falk Schreiber (FB ILW))

GI-Junior-Fellows 2019 gekürt

Die Gesellschaft für Informatik e. V. (GI) ernennt Prof. Dr. Ziawasch Abedjan (TU Berlin), Prof. Dr. Viktor Leis (Uni Jena), Dr. Sandra Schulz (HU Berlin) und Dr. Felix Gessert (Baqend) zu Junior Fellows 2019. Mit dem Junior-Fellowship will die Gesellschaft für Informatik herausragende Informatik-Talente aus Wissenschaft und Praxis dazu ermutigen, sich für die weitere Entwicklung der Informatik zu engagieren. Die Junior-Fellows erhalten hierzu fachliche, ideelle und finanzielle Unterstützung der GI, um eigenverantwortlich Ideen zur Gestaltung der Informatik in Gesellschaft und Wissenschaft umzusetzen. Prof. Hannes Federrath, Präsident der Gesellschaft für Informatik: „Trotz ihres jungen Alters haben sich alle vier Persönlichkeiten bereits durch ihre hervorragenden Leistungen und ihr Engagement für die Informatik einen Namen gemacht. Mit dem Junior-Fellowship wollen wir diese

vorbildhaften Talente weiter fördern. Ich freue mich schon jetzt auf die gemeinsame Arbeit und die vielen neuen Impulse, die sie innerhalb und außerhalb unserer Fachgesellschaft setzen werden."

Prof. Dr. Ziawasch Abedjan ist Juniorprofessor an der TU Berlin und leitet dort das „Fachgebiet Big Data Management". Die Forschung seiner Gruppe befasst sich mit der Entwicklung von Methoden zur Automatisierung zeitintensiven Datenvorbereitungsschritten, wie die der Datenintegration und Datenreinigung, um Data-Science-Anwendungen zu beschleunigen. Als Junior-Fellow möchte er die GI bei der Erprobung und Dissemination der Idee von Data Literacy und der Ausarbeitung von Empfehlungen für die universitäre Ausbildung von Data Scientists unterstützen.

Prof. Dr. Viktor Leis ist Professor für Datenbanken und Informationssysteme an der Fakultät für Mathematik und Informatik der Friedrich-Schiller-Universität Jena. Sein Forschungsgebiet ist die Entwicklung von Datenbanksystemen zur effizienten Speicherung und Verarbeitung von großen Datenmengen auf moderner Hardware. Er setzt sich für die Verbesserung der Softwareentwicklungsausbildung an Universitäten und die Stärkung der Startup-Kultur in Deutschland ein

Dr. Sandra Schulz forscht an der Humboldt-Universität zu Berlin zu aktuellen Themen der Didaktik der Informatik und ist darüber hinaus an einer Berliner Schule tätig. Sie untersucht Problemlöseprozesse beim Umgang mit Physical-Computing-Geräten in der Informatik sowie fächerübergreifende MINT-Problemlöseprozesse. Um die informatische Bildung an der Basis – v. a. in der Schule – zu fördern, hat sie die Schülergesellschaft Informatik an der HU Berlin ins Leben gerufen sowie weitere Kooperationen von Universitäten und Schulen initiiert.

Dr. Felix Gessert ist Gründer und CEO von Baqend, einer Ausgründung der Universität Hamburg mit mittlerweile 15 Mitarbeitern und Mitarbeiterinnen, die eine neue Technologie für Ladezeiten-Optimierung von E-Commerce Webseiten entwickelt. In seiner Promotion an der hat Felix Gessert die zugrundeliegenden Caching-Algorithmen erforscht. Für seine Vision eines Webs ohne Ladezeiten möchte er Wissenschaft und Praxis enger zusammenbringen.

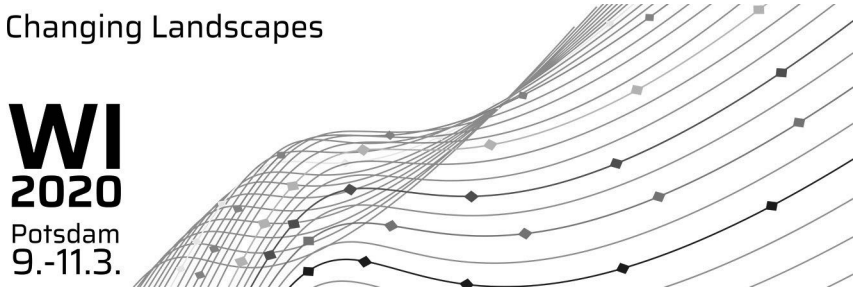


Abb. 5

Tagungsankündigungen

Wirtschaftsinformatik 2020 in Potsdam

Die aktuellen Entwicklungen, Chancen und Herausforderungen der Digitalisierung werden auf der internationalen Tagung der Wirtschaftsinformatik (WI2020) vom 9.–11.3.2020 in Potsdam unter dem Motto „Changing Landscapes – Shaping Digital Transformation and its Impact" diskutiert. Ergänzt wird das wissenschaftliche Programm durch praxisorientierte Formate, Nachwuchskräfte finden hier Kontakte für ihre weitere berufliche Laufbahn. Das Programm und Termine sind unter www.wi2020.de zu finden.

Bundeswettbewerb Informatik

Bronze bei der zentraleuropäischen Informatikolympiade CEOI 2019

Das deutsche Team reist, wie schon im vergangenen Jahr, mit einer starken Bronzemedaille heim von der zentraleuropäischen Informatikolympiade (CEOI). In diesem Jahr fand die CEOI im slowakischen Bratislava statt. Gegen das besonders anspruchsvolle Teilnehmerfeld bei der CEOI ist gewöhnlich nur schwer anzukommen – das macht die Bronzemedaille, die Lennart Ferlemann aus dem westfälischen Bad Oeynhausen gewann, besonders wertvoll. Lennart Ferlemann hatte bereits bei der Ostseeolympiade (BOI) im letzten Jahr teilgenommen und verzeichnete in diesem Jahr einen deutlichen Leistungssprung. Auch die anderen Mitglieder des deutschen Olympiateams haben sich gut geschlagen und scheiterten zum Teil denkbar knapp nur an der Implementation. Das deutsche Team ist gut aufgestellt – mit erfahrenen Veteranen wie Erik Sünderhauf und Neulingen wie Vincent de Bakker – und hat in diesem Jahr bei der BOI und nun auch bei der CEOI Edelmetall gewonnen.

Abb. 6 Erik Sünderhauf, Luis Bahners, Lennart Fehrlemann, Vincent de Bakker (von links nach rechts). Quelle: BWINF

So kann sich das Team mit Zuversicht auf die Reise nach Aserbaidschan machen. Dort, in der Hauptstadt Baku nämlich, findet vom 4. bis zum 11. August die Internationale Informatikolympiade statt, das Hauptevent im Olympia-Kalender der deutschen Mannschaft. Die zentraleuropäische Informatikolympiade 2020 findet vom 29. Juni bis zum 5. Juli 2020 in Nagykanizsa in Ungarn statt, und wird hoffentlich wieder mit starker, deutscher Beteiligung laufen.

GI-Veranstaltungskalender

23.10.2019 – Stuttgart
Kotlin – 7 Languages in 7 Months
JUGS
https://www.jugs.org/va2019/10-23.html

23.10.–24.10.2019 – Nürnberg
SOPHIST DAYS 2019
SD19
https://www.sophist.de/sophist-days/sophist-days-2019/?L=0

24.10.–25.10.2019 – Lörrach
Neue Vorgehensmodelle in Projekten – Führung, Kulturen und Infrastrukturen im Wandel
PVM2019
https://pvm-tagung.de

28.10.–31.10.2019 – Berlin
The 30th IEEE International Symposium on Software Reliability Engineering
ISSRE 2019
http://2019.issre.net

29.10.–30.10.2019 – Potsdam
Designing Digital Transformation – 50 Jahre Internet
https://hpi.de/20-jahre-hasso-plattner-institut/festprogramm/designing-digital-transformation-50-jahre-internet.html

04.11.–06.11.2019 – Würzburg
10th Symposium on Software Performance 2019
SSP2019
https://www.performance-symposium.org/2019/

04.11.–07.11.2019 – Salvador/Brazil
ER 2019 – 38th International Conference on Conceptual Modeling
ER 2019
http://www.inf.ufrgs.br/er2019/

07.11.–08.11.2019 – Berlin
10. Fraunhofer FUSECO Forum
FFF 2019
https://www.fokus.fraunhofer.de/ngni/events/fuseco_forum_2019

13.11.–14.11.2019 – Aachen
CDO Aachen 2019 – Convention on Digital Opportunities
http://cdo-aachen.de/

21.11.–22.11.2019 – Boppard
40. Fachtagung Echtzeit „Autonome Systeme – 50 Jahre PEARL“
Echtzeit 2019
https://www.real-time.de/echtzeit.html

22.11.2019 – Frankfurt am Main
SECMGT-Workshop
https://fg-secmgt.gi.de/

22.11.–24.11.2019 – Magdeburg
1. Convention „KI & Wir*“ zu Künstlicher Intelligenz & Gender
www.ki-convention.com

27.11.2019 – Stuttgart
Elixir – 7 Languages in 7 Months
JUGS
https://www.jugs.org/va2019/11-27.html

28.11.2019 – Stuttgart
Ringvorlesung: Forum Software und Automatisierung
https://www.ias.uni-stuttgart.de/lehre/vorlesungen/

03.12.–05.12.2019 – Pretoria/ South Africa
Intelligent Systems Design and Application
FRAI-ISDA'2019
http://www.mirlabs.org/isda19/

08.12.–11.12.2019 – National Harbor/ USA
Winter Simulation Conference 2019
WSC 2019
http://www.wintersim.org

09.12.–12.12.2019 – Frankfurt am Main
IT-Tage 2019
ITT
http://www.it-tage.org

18.02.–22.02.2020 – Furtwangen
meccanica feminale 2020
https://scientifica.de/bildungsangebote/meccanica-feminale/meccanica-feminale-call-for-lectures/

07.03.2020 – Rostock
Landestagung der Informatiklehrer/innen MV
LT-MV2020
https://gi-ibmv.de/

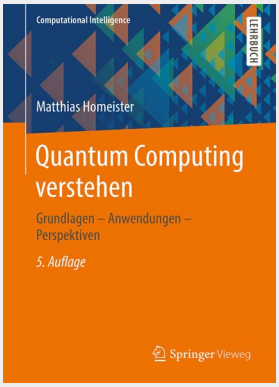
17.03.–20.03.2020 – Göttingen
Jahrestagung des Fachbereichs Sicherheit Schutz und Zuverlässigkeit
GI Sicherheit 2020
https://www.uni-goettingen.de/de/603140.html

08.06.–10.06.2020 – Berlin
Languages & the Media 2020
13th International Conference on Language Transfer in Audiovisual Media
LM2020
https://www.languages-media.com/

06.09.–09.09.2020 – Magdeburg
Tagung Mensch und Computer 2020
MuC2020
https://muc2020.mensch-und-computer.de

29.09.–01.10.2020 – Karlsruhe
INFORMATIK 2020: Jahrestagung der Gesellschaft für Informatik
INFORMATIK 2020
gs@gi.de

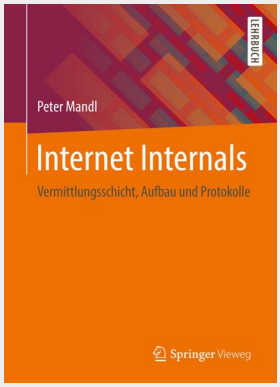
Lehrbuch Informatik



M. Homeister
Quantum Computing verstehen
Grundlagen – Anwendungen – Perspektiven
5., aktualisierte u. erw. Aufl. 2018, XI, 328 S. 83 Abb., 15 Abb. in Farbe. Brosch.
€ (D) 34,99 | € (A) 35,97 | *sFr 39,00
ISBN 978-3-658-22883-5
€ 26,99 | *sFr 31,00
ISBN 978-3-658-22884-2 (eBook)

- Quantum Computing begreifen – ohne spezielle Vorkenntnisse
- Erklärt konkrete Anwendungen und aktuelle Forschung: von Grovers Suchalgorithmus bis Quantenkryptographie
- Alle wichtigen Begriffe werden umfassend erklärt

Nach der technologischen Revolution, die die Erfindung des Computers ausgelöst hat, steht mit der Zusammenführung von Computing und Quantenmechanik die nächste bevor: Quantum Computing. Anschaulich und auf Beispiele gestützt führt dieses Buch in die Grundlagen des Quantum Computing ein. Was ein Quantencomputer ist und was er kann wird anhand von Algorithmen erläutert, also anhand von konkreten Rechenverfahren. Dieser an den Grundlagen orientierte Zugang befähigt Leserinnen und Leser aktuelle und auch künftige Entwicklungen einzuordnen. Um zu verstehen, wie Quantencomputer rechnen, erklärt der Autor, der als Professor für Informatik an der TH Brandenburg lehrt, zunächst die einfachen quantenmechanischen Prinzipien und stellt diese so anwendungsorientiert wie nur möglich dar.



P. Mandl
Internet Internals
Vermittlungsschicht, Aufbau und Protokolle
2019, XII, 197 S. 102 Abb. Brosch.
€ (D) 32,99 | € (A) 33,92 | *sFr 36,50
ISBN 978-3-658-23535-2
€ 24,99 | *sFr 29,00
ISBN 978-3-658-23536-9 (eBook)

- Überblick über den Aufbau des Internets
- Erklärung von Protokollen und Algorithmen der Internet-Vermittlungsschicht
- Anschauliche Erläuterung des Zusammenwirkens der Internet-Mechanismen

Lernen Sie in diesem Buch alles über Vermittlungsschicht und Aufbau des Internets und die Übertragung von Informationen Sie haben sich schon immer gefragt, wie das World Wide Web aufgebaut ist? Sie suchen eine Einführung in die „Internet Internals"? Dieses Buch über die Vermittlungsschicht des Internets gibt Ihnen Antworten. Es führt Sie in die Grundlagen der Übertragung von Daten im Netz ein und vermittelt Ihnen dabei einen Überblick über den Aufbau des Internets. Die große Stärke dieses Buchs über die Vermittlungsschicht des Internets liegt darin, dass es komplexe Sachverhalte wie Protokolle und Algorithmen im Detail beschreibt, dabei jedoch stets verständlich und übersichtlich bleibt. Die Inhalte von „Internet Internals" Autor Peter Mandl gewährt dem Leser einen möglichst tiefen wie breiten Einblick in die Kommunikationsstrukturen des World Wide Webs.

Ihre Vorteile in unserem Online Shop:

Über 280.000 Titel aus allen Fachgebieten | eBooks sind auf allen Endgeräten nutzbar | Kostenloser Versand für Printbücher weltweit

€ (D) sind gebundene Ladenpreise in Deutschland und enthalten 7 % für Printprodukte bzw. 19 % MwSt. für elektronische Produkte. € (A) sind gebundene Ladenpreise in Österreich und enthalten 10 % für Printprodukte bzw. 20% MwSt. für elektronische Produkte. Die mit * gekennzeichneten Preise sind unverbindliche Preisempfehlungen und enthalten die landesübliche MwSt. Preisänderungen und Irrtümer vorbehalten.

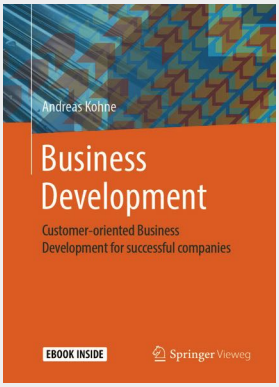
Wirtschaftsinformatik

T. Barton, C. Müller, C. Seel (Hrsg.)
Digitalisierung in Unternehmen
Von den theoretischen Ansätzen zur praktischen Umsetzung
2018, XX, 305 S. 91 Abb. Book + eBook. Geb.
€ (D) 39,99 | € (A) 40,93 | *sFr 43,50
ISBN 978-3-658-22772-2
€ 29,99 | *sFr 34,50
ISBN 978-3-658-22773-9 (eBook)

- Zeigt Ansätze und Szenarien, um Digitalisierungsprojekte erfolgreich umzusetzen
- Geht auf Fragestellungen aus der unternehmerischen Praxis ein
- Auch für Studium und Lehre geeignet

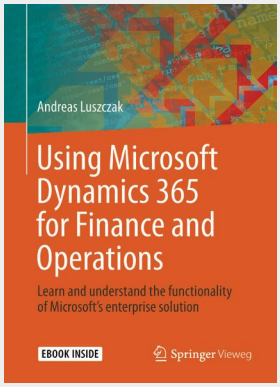
S. Robra-Bissantz, C. Lattemann (Hrsg.)
Digital Customer Experience
Mit digitalen Diensten Kunden gewinnen und halten
2019, XXVI, 298 S. 56 Abb., 36 Abb. in Farbe. Book + eBook. Geb.
€ (D) 44,99 | € (A) 46,06 | *sFr 48,50
ISBN 978-3-658-22541-4
€ 34,99 | *sFr 38,50
ISBN 978-3-658-22542-1 (eBook)

- Aktuelle Forschung zu Digital Customer Experience
- User Experience Konzepte für den stationären Einzelhandel
- Zeigt soziale und kooperative Lösungsansätze



A. Kohne
Business Development
Customer-oriented Business Development for successful companies
2019, IX, 110 p. 12 illus., 8 illus. in color. Book + eBook. Geb.
€ (D) 49,22 | € (A) 50,39 | *sFr 53,50
ISBN 978-3-658-24725-6
€ 39,26 | *sFr 42,50
ISBN 978-3-658-24726-3 (eBook)

- Easy to implement business development process
- General approach that can be used in any business
- Relevant businsess development case study included



A. Luszczak
Using Microsoft Dynamics 365 for Finance and Operations
Learn and understand the functionality of Microsoft's enterprise solution
2019, XIII, 475 p. 1 illus. Book + eBook. Geb.
€ (D) 49,22 | € (A) 50,39 | *sFr 53,50
ISBN 978-3-658-24106-3
€ 39,26 | *sFr 42,50
ISBN 978-3-658-24107-0 (eBook)

- Easily learning Microsoft Dynamics 365 for Finance and Operations through hands-on examples
- Including a simple but comprehensive case study
- Knowledge to handle all basic business processes
- Exercises make it a good choice for self-study

Ihre Vorteile in unserem Online Shop:

Über 280.000 Titel aus allen Fachgebieten | eBooks sind auf allen Endgeräten nutzbar | Kostenloser Versand für Printbücher weltweit

€ (D) sind gebundene Ladenpreise in Deutschland und enthalten 7 % für Printprodukte bzw. 19 % MwSt. für elektronische Produkte. € (A) sind gebundene Ladenpreise in Österreich und enthalten 10 % für Printprodukte bzw. 20% MwSt. für elektronische Produkte. Die mit * gekennzeichneten Preise sind unverbindliche Preisempfehlungen und enthalten die landesübliche MwSt. Preisänderungen und Irrtümer vorbehalten.