Engineering Trustworthy Data-Intensive Systems

Vorschlag für einen Profilbereich

Fachbereich für Informatik

DFG-Fachgebiete: Künstliche Intelligenz, Softwaretechnologie

Antragsteller: Patrick Delfmann, Jan Jürjens (Leitung), Ralf Lämmel, Andreas Mauthe, Viorica Sofronie-Stokkermans, Steffen Staab, Matthias Thimm, Claudia Wagner

1 Beschreibung des Forschungsthemas

Das Problem: Data Science ist ein interdisziplinäres Gebiet, das wissenschaftliche Methoden, Prozesse, Algorithmen und Systeme verwendet, um Wissen aus und Einsichten in Daten zu gewinnen, die in unterschiedlichen Formaten vorliegen. Kennzeichnend für das Gebiet ist einerseits die hohe Signifikanz, die es aufweist. Aufgrund der gewonnenen Resultate werden zahlreiche wichtige Entscheidungen getroffen, die den einzelnen oder die Gesellschaft als Ganzes betreffen: Diagnosen, Therapien, Kreditentscheidungen, Raumplanungen, etc. Andererseits ist Data Science charakterisiert durch die iterative und empirisch-heuristische Vorgehensweise, mittels derer Wissen extrahiert und Entscheidungen abgeleitet werden. Was typischerweise zu kurz kommt, ist eine systematische, ingenieurorientierte Vorgehensweise, die Aussagen über die Qualität der Datenanalyse erlaubt.

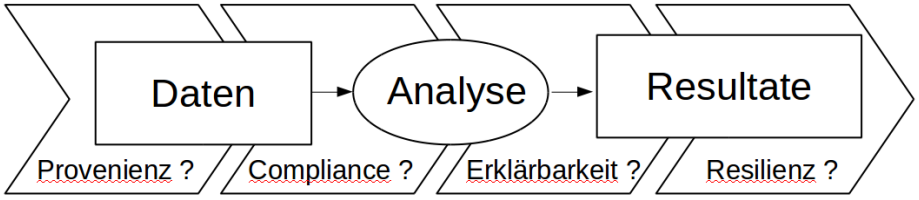
Ein prominentes Beispiel für einen Fehlschlag einer wissenschaftlichen Datenanalyse war (Reinhart und Rogoff, 2010). Aufgrund der abgeleiteten Einsicht, dass das Wirtschaftswachstum von Staaten ab einer Staatsverschuldung von 90% nachlässt, wurde international Politik gemacht. Im Jahr 2013 aber wurden solche Fehler in der Datenanalyse von Reinhart und Rogoff nachgewiesen, dass die Aussagen zurückgezogen werden mussten. Dass der Fehler überhaupt bemerkt wurde, war eher ein glücklicher Zufall.

Die Korrektheit von Softwaresystemen wird heute durch methodisches Vorgehen und formale Werkzeuge unterstützt oder gar bewiesen. Für Data Science und daraus entstehende datenintensive Software fehlt aber ein entsprechendes Inventar an Methoden, Prozessen, Algorithmen und Systemen, die zur Korrektheit beitragen. Insbesondere lässt sich auch nicht ohne weiteres "korrektes Verhalten" eines datenintensiven Systems beschreiben, denn das Ergebnis ist nicht vorherbestimmt und soll erst durch den Datenanalyseprozess gewonnen werden.

Das Ziel: Das Ziel dieses Profilbereiches ist es, methodische Vorgehensweisen und formale Werkzeuge zu erforschen, die die ingenieurmäßigen Entwicklung korrekter - oder zumindest vertrauenswürdiger - datenintensiver Software unterstützen. Dieses Ziel wird in verschiedenen Teilprojekten verfolgt (vgl. Abschnitt 1.1), in einer Data-Science-Entwicklungsumgebung wie Jupyter (http://jupyter.org; vgl. Abschnitt 1.2) implementiert, und in Anwendungsgebieten (vgl. Abschnitt 3) evaluiert werden.

1.1 Teilprojekte

Die folgende Abbildung verdeutlicht den Kern-Lebenszyklus bei der Analyse von Daten. Das geplante Projekt beschäftigt sich mit den wissenschaftlichen Fragestellungen in Hinblick auf die Vertrauenswürdigkeit dieses Prozesses und seiner Ergebnisse, die sich vor, zwischen und nach den einzelnen Phasen des Prozesses stellen:



Vertrauenswürdige Daten-intensive Software muss demnach die folgenden Aspekte des Datenanalyse-Prozesses berücksichtigen:

- **Datenprovenienz:** Woher kommen die Daten?
- **Compliance:** Welche Einschränkungen müssen bei der Datenanalyse und der Verwendung ihrer Ergebnisse berücksichtigt werden?
- **Erklärbarkeit:** Wie kommt die Software zu ihren Ergebnissen?
- **Resilienz:** Wie wird dabei eine Manipulation verhindert?

Diese Aspekte werden in den folgenden Teilprojekten untersucht.

1.2.1 Datenprovenienz [Ralf Lämmel / Steffen Staab]

Der Begriff der Datenprovenienz bzw. des verwandten Bereiches der Datenlineage erfasst grundsätzlich die Verwaltung der Ursprünge von Datenpartikeln in zusammengesetzten Datenartefakten (z.B. Tabellen in einem Data Warehouse) und eine Beschreibung der Zusammenhänge - also wie die Datenpartikel aus Quellen entsprechend in den vorliegenden Artefakt überführt wurden. Datenprovenienz ist eine grundlegende Komponente, um überhaupt in einem nichttrivialen System Compliance, Erklärbarkeit und Resilienz im Sinne der folgenden Teilprojekte zu realisieren (vgl. Abschnitte 1.2.2 - 1.2.4).

Der Stand der Kunst in der Datenprovenienz ist vorrangig mit homogenen Systemen beschäftigt, z.B. komplett SQL-basierte Data Warehouses oder Daten- bzw. Informationsflussanalyse für Programmiersprachen. Das Teilprojekt fokussiert auf die Realität von heterogenen und komplexen Systemen mit verschiedenen Komponenten, verschiedenen Datenrepräsentationen (etwa JSON versus relationale Datenbank versus RDF/Ontologien versus Programmdaten) und verschiedenen Abhängigkeiten (etwa Mapping, Streaming, Pipelines, Ad-hoc Programmlogik).

Das Teilprojekt behandelt Herausforderungen wie die folgenden: a) die Verwendung von so ausreichend differenzierten Datentypen, dass Unterschiede (etwa Zeitzonen oder Verfallsgrenzen oder Genauigkeits- oder Auflösungsdetails) beachtet werden; b) die zuverlässige Verschmelzung verschiedener Datenmodelle und Typsysteme so dass keine Verluste an den Übergängen auftreten bzw. diese Verluste etwa durch Wahrscheinlichkeitsverteilungen quantifiziert werden; c) die zuverlässige Komposition von Komponenten so dass keine Fehlinterpretationen der weitergereichten und kompositionierten Daten auftreten (etwa im Fall von Verzögerungen beim Loggen bzw. Streamen).

Die folgenden Arbeiten nennen ausgewählte Grundlagen und Anwendungen von Datenprovenienz sowie erste relevante Veröffentlichungen der am Teilprojekt Beteiligten. Im Teilprojekt erfolgt eine Konzentration auf Herausforderungen im Überlappungsbereich Ontologien, Programmiersprachen, Datenbanken, und Softwareengineering.

Eigene Vorarbeiten: (Leinberger et al., 2017, 2014; Ringelstein et al., 2011; Schenk et al., 2011)

Sonstige relevante Literatur: (Chen 2016; Gehani&Tariq, 2012; Han et al. 2018; Moreau et al., 2015; Sáenz-Adán et al. 2018)

1.2.2 Compliance [Patrick Delfmann / Jan Jürjens / Andreas Mauthe]

Themengegenstand: Unter Compliance wird die Konformität von Software mit geltenden Regulierungen verstanden (Becker et al. 2016). Im weitesten Sinne umfassen solche Regulierungen die Korrektheit der Software allgemein, im engeren Sinne bezeichnet Compliance die Konformität der Software mit juristischen und/oder unternehmensinternen, betriebswirtschaftlichen Regelwerken. Datensouveränität bezeichnet dabei insbesondere die Fähigkeit eines Datengebers, zu kontrollieren, wer die eigenen Daten erhält, und für welche Zwecke sie verwendet werden. Dies betrifft sowohl industrielle als auch personenbezogene Daten. Im letzteren Fall gibt es dabei einen engen Bezug zum Thema Datenschutz; so fordert z. B. das GDPR: "Werden personenbezogene Daten […] erhoben, so teilt der Verantwortliche […] Folgendes mit: das Bestehen einer automatisierten Entscheidungsfindung […] und […] aussagekräftige Informationen über die involvierte Logik […]."[1]

Stand der Wissenschaft: Klassischerweise wird im Rahmen des Business Process Compliance Management mithilfe von Anfragesprachen (model query languages, s. z.B. (Ahmadian et al. 2017; Bürger et al. 2018; Ramadan et al. 2017)) oder Logfileanalysen (process mining) untersucht, ob eine Software Abläufe zulässt, die in dieser Form aufgrund des Regelwerks unzulässig sind (z.B. existieren im Kreditwesen Regulie-

[1] https://www.informatik-aktuell.de/entwicklung/methoden/kuenstliche-intelligenz-und-erklaerbarkeit.html

rungen, die bestimmte Abarbeitungsreihenfolgen/Eskalationen vorschreiben, und die bei Nichteinhaltung juristische Konsequenzen für das jeweilige Kreditinstitut nach sich ziehen können). Unzulässige Abläufe können so erkannt und eliminiert werden. Anwendungsstudien haben unlängst ein erhebliches Risikovermeidungspotenzial einer solchen Vorgehensweise offenbart. Korrespondierende Untersuchungen im Bereich datenintensiver Software existieren bisher nicht, obwohl zu erwarten ist, dass sie erhebliches Anwendungspotenzial eröffnen.

Ziele: Ziel des Teilprojekts *Compliance* ist es, die Erkenntnisse, die im Bereich des klassischen Compliance Management erzielt wurden, auf datenintensive Software und die Anforderung der Datensouveränität sowie die Einhaltung von Datensicherheits- und Datenschutzanforderungen zu übertragen. So ist z.B. zu vermeiden, dass eine Datenanalyse aufgrund unzulässiger Annahmen zu verzerrten Ergebnissen kommt (vgl. Amazon-Recruiting-Software, die weibliche Bewerber benachteiligte sowie in den USA eingesetzte Algorithmen, die die Rückfallwahrscheinlichkeit von Straftätern berechnen und Afroamerikaner benachteiligen). Es wird angenommen, dass derartige Compliance-Verletzungen sich - ähnlich wie Compliance-Verletzungen im klassischen Sinne - durch typische Muster sowohl in der Nutzung der Datenquellen als auch in den Abläufen manifestieren. Zur Aufdeckung solcher Muster sollen sowohl die modellhaften Abbildungen des Softwarecodes als auch die von Software generierten Logs untersucht werden. Dabei sollen sowohl graph- und logikbasierte Pattern Matching-Verfahren als auch ML-Algorithmen der grammatischen Inferenz zum Einsatz kommen (vgl. z. B. (Breuker et al. 2016)). Die Ergebnisse dieses Teilprojekts werden u. a. im Kontext einer Anwendung auf den Industrial Data Space evaluiert (s. https://www.internationaldataspaces.org).

Eigene Vorarbeiten: (Ahmadian et al. 2017; Becker et al. 2016; Breuker et al. 2016; Bürger et al. 2018; Ramadan et al. 2017)

Sonstige relevante Literatur: (Polyvyanyy et al. 2017; van der Aalst 2016; Di Francescomarino 2018)

1.2.3 Erklärbarkeit [Viorica Sofronie-Stokkermans / Matthias Thimm / Claudia Wagner]

Ein wesentlicher Aspekt, der zur Vertrauenswürdigkeit datenintensiver Systeme beitragen kann, ist ihre Erklärbarkeit, d.h. deren Fähigkeit, getroffene Entscheidungen zu erklären, aber auch, allgemeiner, die korrekte Ausführung solcher Systeme nachzuweisen oder zu beweisen. Obwohl in datenintensiven Systemen "korrektes Ausführen" oft nicht klar definiert ist, gibt es Aspekte der "Korrektheit", die nachgewiesen können, für (i) Daten, (ii) Algorithmen, die Daten verarbeiten, und (iii) Entscheidungen, die auf solche Algorithmen basieren. Die *Daten* können - z.B. aufgrund von Modellierungsfehlern - falsche Informationen enthalten. Falls Inkonsistenz nachgewiesen werden kann, sind Erklärungen (Beweise, unerfüllbare Teilmengen) für das Auffinden der Fehler hilfreich. Darüber hinaus kann eine quantitative Analyse der Inkonsistenz (durch Methoden der Inkonsistenzmessung) verwendet werden, um die Schwere der Fehler zu bewerten und die Ermittlung der wahrscheinlichsten Schuldfaktoren zu unterstützen. Auch werden manchmal Inkonsistenzen falsch erkannt, weil die Wahrheit der Fakten nur mit einem bestimmten Maß an Vertrauen garantiert wird; in solchen Situationen ist es wichtig, die Fakten bzw. Schlussfolgerungen zu analysieren, um zu verstehen, inwieweit man sich auf die Ergebnisse verlassen kann. Ein weiteres Problem, das berücksichtigt werden muss, ist die Repräsentativität von Daten auf deren Basis Algorithmen Entscheidungen treffen. Minderheiten sind häufig unterrepräsentiert in Trainingskorpora. Dies kann zu asymmetrischen Fehlraten für unterschiedliche Gruppen führen. Die *Algorithmen,* die Daten verarbeiten, können ebenfalls Fehler enthalten, deshalb sind Korrektheitsnachweise oder Erklärungen der Probleme ebenfalls wichtig. Transparente datenintensive Systeme, insbesondere auch solche, die auf Methoden des Maschinellen Lernens aufsetzen, sollten auf Anfrage erläutern können, warum bestimmte *Entscheidungen* getroffen wurden.[2]

Stand der Wissenschaft: Viele der oben genannten Aspekte wurden in den letzten Jahren untersucht. Erklärungen (in Form von Beweisen, Modellen oder unerfüllbaren Kernen) oder Interpolanten werden heutzutage von vielen automatisierten Beweisern für viele Klassen logischer Theorien zur Verfügung gestellt, siehe z.B. (Sofronie-Stokkermans 2018, Barrett et al. 2009). Verschiedene Verfahren zur Bewertung von Inkonsistenzmaßen werden beispielsweise in (Thimm 2018) diskutiert. Die Arbeiten (Damm et al. 2015; Sofronie-Stokkermans 2013) schlagen Methoden zur Überprüfung von Sicherheitseigenschaften in reaktiven und hybriden Systemen vor, die für unsichere Systeme Gegenbeispiele bereitstellen, die erklären, wo die Probleme liegen, oder Constraints auf Parameter generieren, die die Sicherheit gewährleisten. Das Gebiet der formalen Argumentation (Baroni et al. 2018) bietet Formalismen zur Modellierung von Argu-

[2] Dies wird auch auf politischer Ebene gefordert; vgl.:
https://www.zeit.de/digital/internet/2017-10/google-facebook-algorithmen-regulierung-bundestag-gutachten

mentationsprozessen durch Interaktionen von Argumenten und Gegenargumenten, die eine natürliche Repräsentation für Erklärungen sind. Diese Formalismen führen eine dialektische Analyse durch, bei der alle Argumente dazu beitragen, zu entscheiden, ob eine bestimmte Aussage wahr ist. Diese Analyse kann dem Benutzer gezeigt werden, um zu erklären, warum eine bestimmte Entscheidung getroffen wurde. Die Arbeit (Thimm and Kersting 2017) diskutiert erste Anwendungen dieser Formalismen auf Methoden des maschinellen Lernens. Die Arbeit (Hannak et al 2017) zeigt, dass algorithmische Benachteiligung von Minderheiten in Freelancer-Plattformen ein Problem darstellt, während die Arbeit (Wagner et al 2017, Karimi et al 2018) die Bedingungen analysiert die zu algorithmischen Benachteiligungen führen können.

Ziel: Ziel des Teilprojekts ist es, Methoden zu entwickeln, um Erklärungen für die Korrektheit und Repräsentativität der Daten sowie Korrektheit und Fairness der Algorithmen, die diese Daten bearbeiten, bereitzustellen, und um Entscheidungen, die auf datenintensiver Software basieren, zu erklären. Die Herausforderung besteht darin, die Methoden des automatisierten Schließens und der Softwareanalyse zu erweitern, damit sie für datenintensive Systeme – ggf. mit heterogenen Daten - geeignet sind. Die Überprüfung der Algorithmen, die Daten verarbeiten, kann problematisch sein, wenn kein formaler Begriff von „korrektem Verhalten“ vorliegt. Wir planen jedoch, Situationen zu identifizieren, in denen klassische Methoden der Softwareverifizierung verwendet werden können, um Erklärungen zu finden - z.B. Korrektheitsnachweise oder Beschreibung von Situationen, in denen Probleme auftreten, oder Einschränkungen („Constraints“) auf Parametern (welche entweder in den Algorithmen oder in den Daten vorkommen), die zusätzliche Bedingungen beschreiben, welche Korrektheit und Fairness gewährleisten. Des Weiteren sollten die Ergebnisse zur formalen Argumentation zur Modellierung von Erklärungen verwendet werden.

Eigene Vorarbeiten: (Sofronie-Stokkermanns 2013, 2018; Damm et al. 2015; Thimm&Kersting 2017; Thimm 2018; Hannak et al. 2017; Wagner et al. 2017; Karimi et al. 2018)

Sonstige relevante Literatur: (Baroni et al. 2018), (Barrett et al. 2009)

1.2.4 Resilienz [Andreas Mauthe / Steffen Staab / Jan Jürjens]

Themengegenstand: Resilienz sozio-technischer Systeme wird als Fähigkeit definiert, in Anbetracht interner und externer Herausforderungen (z.B. Attacken und gezielten Manipulationen, aber auch Überlastsituationen und Fehlkonfigurationen) einen akzeptablen Service, vorhersagbares Verhalten und zuverlässige (bzw. vertrauenswürdige) Ergebnisse zu liefern. Um dies zu erreichen, sind zum einem Schutzvorkehrungen struktureller Art (z.B. im Algorithmusdesign), sowie aktive Maßnahmen zur Entdeckung der Herausforderungen und ihrer Behebung (bzw. Gegensteuerung) notwendig. Im Bereich vertrauenswürdiger Daten-intensiver Software bedeutet dies zum einen, Analyseverfahren so zu gestalten, dass die zugrundeliegenden Machine-Learning- und Data-Mining-Algorithmen möglichst robust gegenüber Manipulationen sind (sogenanntes “Resilience-by-Design”). Dies kann z.B. den Einsatz paralleler Alternativverfahren einschließen. Zum anderen sollten die Datenbasis, Dateneinund -ausgabe auf Abweichungen und Anomalien untersucht werden. Eine Möglichkeit ist z.B. die Analyse von Metadaten, die die Datenherkunft (z. B. Mac-Adressen von Sensoren) erkennen lassen, um eine Man-in-the-Middle-Attacke bei der Dateneingabe zu erkennen, oder der Analyse von fortlaufenden und historischen Ausgabedaten, um Extreme und Abweichungen herauszufinden und anzuzeigen. Bei der Überprüfung der Ausgabedaten sollten auch zusätzliche situative Daten berücksichtigt werden. Dies können, je nach Analysegegenstand, Umgebungsdaten sein (z.B. meterologische Daten), Sozial- und Wirtschaftsdaten, oder auch Daten über Nutzerverhalten. Die (Meta-)Datenuntersuchung kombiniert Daten und Analyseverfahren von mehreren Systemebenen, um ein möglichst vollständiges Bild über den Zustand der Datenherkunft und Datenanalyse zu erhalten.

Stand der Wissenschaft: Resilienz-Forschung findet in unterschiedlichen Wissenschaftsbereichen statt, innerhalb der Informatik liegt der Fokus auf Resilienz in Verteilten Systemen und datenintensiven Systemen. Im Bereich der verteilten und vernetzten Systeme, die mit diesem Teilprojekt im Zusammenhang stehen, gibt es drei wesentliche Gebiete, (i) systemübergreifende Resilienz und die Erforschung von Konzepten, die es erlauben, Resilienz-Ziele und Schutzmaßnahmen über Systemgrenzen zu koordinieren (d.h. Multi-level Resilience (Sterbenz et al 2014; Ariffin et al 2017)), unterschiedliche Algorithmen zur Anomalieerkennung (Shirazi et al 2016; Chen et al. 2017) und die Berücksichtigung situativer Daten im Resilienzprozess (Marnerides et al 2017). Bis vor kurzem fokussierte die Forschung primär auf Daten, die den Systemzustand beschreiben, und nicht, wie in diesem Teilprojekt, auf der Resilienz der Datenquellen, Algorithmen und Ergebnisse.

Ziele: Ziel des Teilprojekts ist es, sowohl strukturelle als auch aktive Resilienz zu erforschen. Strukturelle Resilienz ist z.B. im Bereich datenintensiver Software-Systeme und Algorithmen notwendig. Dies schließt z.B. die Erforschung von Parallel- und Alternativverfahren zur Datenanalyse in Rahmen eines Data-

Analysis-Framework ein, das es erlaubt, je nach Anwendungsfall und Anforderungen geeignete Algorithmen dynamisch auszuwählen, und neue und verbesserte Algorithmen aufzunehmen. Dabei muss die Nachvollziehbarkeit des Einsatzes bestimmter Algorithmen gewährleistet sein und Datensicherheits- und Datenschutzanforderungen in der Datenanalyse unterstützt werden (vgl. z.B. (Ramadan et al 2018)).

Im Bereich der Analyse wird das Teilprojekt sich mit der proaktiven Analyse möglicher Attacken beschäftigen, speziellen mit der Erforschung von Datenmanipulationen, die zu Fehlklassifikation oder -regression führen können. Des Weiteren sollen Verfahren zur Anomalie-Erkennung erforscht werden, bei denen Eingabedaten und ihrer Herkunft basierend auf System- und Meta-Daten untersucht werden, um Manipulation zu erkennen und Gegenmaßnahmen zu ergreifen. Dies schließt die gemeinsame Analyse von Daten aus verschiedenen Systemebenen (sogenannte "Multi-Level Resilience Analysis") mit ein. Im Rahmen der Datenausgabeanalyse werden Analyseverfahren erforscht, die aufgrund von zusätzlichen (z.B. situativen) und historischen Daten die Analyseergebnisse auf Abweichungen und Plausibilität prüfen. Im Bereich der Dateneingabe besteht ein enger Zusammenhang zum Teilprojekt Datenprovenienz; sowohl Resilience-by-Design als auch die Ausgabeanalyse tragen zur Nachvollziehbarkeit und Verlässlichkeit bei.

Eigene Vorarbeiten: (Ramadan et al 2018; Ariffin et al 2017; Shirazi et al 2016; Marnerides et al 2017)

Sonstige relevante Literatur: (Sterbenz et al 2014; Moreau et al 2015; Goodfellow 2014; Chen et al 2017)

1.2 Vorgehensweise

Die vier betrachteten Teilaspekte werden in gemeinsamen Arbeiten exploriert. Als gemeinsame Grundlage wählen wir eine Programmier-/Datenanalyseumgebung (z.B. Jupyter, R oder Weka) und betrachten anfallende Probleme anhand dieser Umgebung. Startpunkt für die Entwicklung von Datenanalyseverfahren ist die CRISP-DM Methodologie, die allerdings auch angepasst bzw. verfeinert werden muss, um die iterativen und entdeckungsorientierten Entwicklung datenintensiver Systemen noch besser zu unterstützen.

Ziel ist es insgesamt, nicht nur einzelne Methoden zu entwickeln, sondern diese in einem Gesamtkonzept und umfassenden Prototypen realisieren und evaluieren zu können. Wir erwarten Synergien zwischen den Teilprojekten bei der Verwendung von Werkzeugen (z.B. Parsern und Transformatoren), formalen Repräsentationssprachen, Schlussfolgerungsmethoden und Beweistechniken.

In Hinblick auf ein integratives Vorgehen innerhalb des Profilbereiches, aber auch die Verwertung der Forschungsergebnisse durch andere ist geplant, die Ergebnisse im Rahmen eines Gesamtansatzes bereitzustellen, der durch die untenstehende Abbildung verdeutlicht wird. Als Rechnerinfrastruktur kann dabei auf die an den Instituten vorhandene Cloud zurückgegriffen werden. An Daten können sowohl Forschungsdaten betrachtet werden (auch durch eine existierende Beteiligung der Antragsteller an der European Open Science Cloud, s. http://eoscpilot.eu), als auch öffentliche Daten oder die privaten Daten der Transferpartner (s. Abs. 3.1.3), für deren Analyse Knowledge Graphs eingesetzt werden. Durch die Resultate der vier Teilprojekte können den Unternehmen dann vertrauenswürdige Datenanalysealgorithmen integriert in interaktiven Notebooks sowie in Form von Programmierabstraktionen angeboten werden. Die Transferpartner treten dabei sowohl als Datengeber, als auch Nutzer der Algorithmen auf. Durch Einbezugnahme des Industrial Data Space (s. http://internationaldataspaces.org), an dem Antragsteller aktuell beteiligt sind, wird in dem Zusammenhang auch der vertrauenswürdige Datenaustausch zwischen Unternehmen unterstützt.