Schatzsuche im **Datendschungel**

Normalerweise formulieren Forscher eine Hypothese, ehe sie mit einem Experiment beginnen und Daten sammeln. **Pauli Miettinen** vom **Max-Planck-Institut für Informatik** in Saarbrücken stellt diesen wissenschaftlichen Grundsatz mit einem neuen Verfahren zur Datenanalyse auf den Kopf – dem Redescription Mining. Die Software kann vorhandene Datensätze analysieren und daraus nachträglich Hypothesen und unerwartete Korrelationen extrahieren, die Wissenschaftlern wiederum wichtige Anhaltspunkte für neue Fragestellungen liefern – zum Beispiel, wenn es darum geht, die politische Stimmung in der Bevölkerung einzufangen.

TEXT **TIM SCHRÖDER**

Über die Jahrzehnte haben Computer gelernt, Aufgaben zu erfüllen, die man ihnen vorgibt. Sie können komplexe Gleichungen lösen, das Wetter vorhersagen und inzwischen sogar mit einer menschlichen Stimme auf Fragen wie „Wo finde ich in der Nähe ein gutes und preiswertes chinesisches Restaurant?“ antworten. Pauli Miettinen vom Saarbrücker Max-Planck-Institut für Informatik aber geht noch einen Schritt weiter. Er hat Computern beigebracht, auf Fragen zu antworten, die der Mensch ihnen noch gar nicht gestellt hat – und so Zusammenhänge zu erkennen, auf die der Mensch allein gar nicht gekommen wäre.

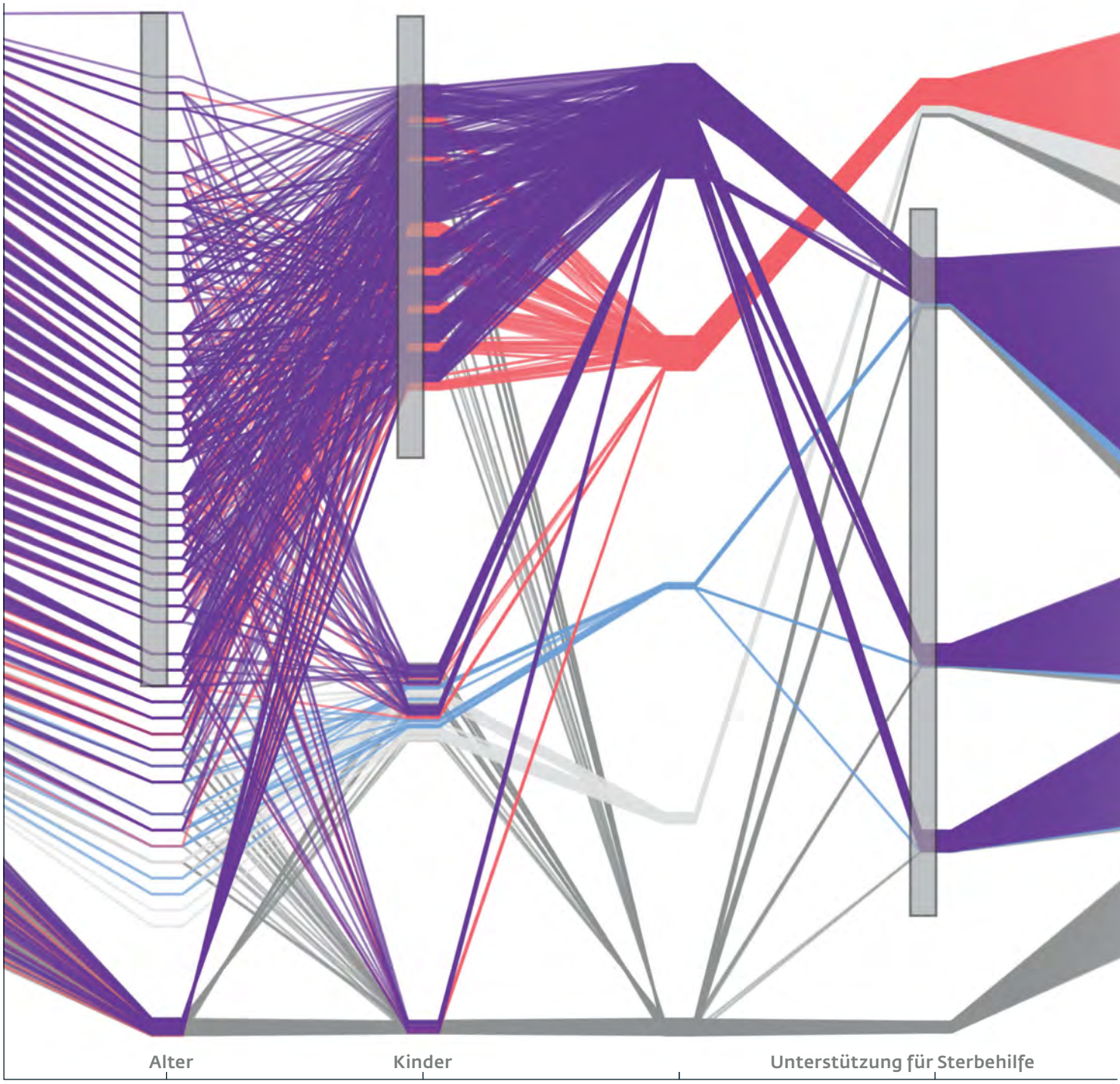
Pauli Miettinen ist damit dem Blick in die Kristallkugel schon recht nahe. Er selbst beschreibt seine Arbeit ein wenig nüchterner: „Im Grunde machen wir nichts anderes, als eine neue Hypothese aus vorhandenen Daten zu generieren.“ Das klingt bescheiden, ist aber nicht weniger als eine kleine Revolution des wissenschaftlichen Arbeitens. Denn seit Jahrhunderten gehen Forscher, gleich welcher Disziplin, immer nach demselben Muster vor. Erst stellen sie eine Hypothese auf wie etwa: „Der Mensch stammt vom Affen ab.“ Dann überprüfen sie diese Hypothese, indem sie beobachten und Daten sammeln.

SINNVOLLE INFORMATIONEN AUS GROSSEN DATENMENGEN

Das Informatikwerkzeug, das Miettinen mit seinem Team entwickelt hat, stellt dieses Prinzip auf den Kopf. Es nutzt vorhandene Daten, analysiert diese und stellt ganz neue Bezüge her – die zum Teil verblüffend sind. Seine Methode ist so etwas wie der letzte Schrei in der Welt der Datenanalyse. Sie heißt Redescription Mining, was frei übersetzt in etwa „Alternativbeschreibung“ bedeutet. Soll heißen: Miettinen und seine Kollegen fahnden in bereits vorhandenen Daten nach neuen Korrelationen, nach neuen Aussagen, die in den Daten stecken – nach neuen Wegen, die Daten zu beschreiben. So helfen sie, Schätze im Datendschungel aufzuspüren.

Dabei, und auch das sind Stärken des Redescription Mining, lassen sich mit der Methode beliebige Arten von Daten analysieren. Und die Datenmenge, die verarbeitet werden kann, ist schier unbegrenzt. So hilft das Verfahren, aus den großen Datenmengen, die heute allerorten gesammelt werden, sinnvolle Informationen zu gewinnen.

Was die Methode kann, hat Pauli Miettinen mit seinen Kollegen anhand von Daten aus seiner Heimat Finnland gezeigt: Informationen über finnische Politiker, die 2011 und 2015 für einen Platz im Parlament kandidiert hatten. Der Forscher hat für seine Analyse zwei Datensätze miteinander verknüpft. Der erste Datensatz enthielt die öffentlich verfügbaren Daten über den sozialen Hintergrund der Politiker, das Alter, die



Grafik: Pauli Miettinen/MPI für Informatik

Eine Linie für jeden Politiker: Diese Grafik hat die Software Siren bei der Analyse der soziodemografischen Daten und der politischen Einstellungen, hier speziell zur Sterbehilfe, von Kandidaten für die finnische Parlamentswahl ergeben. Eine Erkenntnis daraus: Kandidaten über 34 und solche, die Kinder haben, lehnen Sterbehilfe eher ab.

Bringt Licht ins Datendunkel: Pauli Miettinen und seine Mitarbeiter haben eine Software namens Siren (rechte Seite) entwickelt, um in Datensätzen Zusammenhänge aufzudecken, die bei der Datenerhebung noch nicht als Hypothese formuliert wurden.

Foto: Tom Pingel

Herkunft, den Bildungs- oder den Familienstand. Der zweite Datensatz enthielt Antworten auf Fragen, welche die Politiker für ein Webportal beantwortet hatten.

Solche Webportale sind schon seit einigen Jahren außerordentlich beliebt, in Deutschland ist unter anderem der Wahl-O-Mat angesagt. Politiker und Wähler antworten unabhängig voneinander auf dieselben Fragen. Das Webportal nennt dem Wähler dann jene Partei, mit der er die größten Übereinstimmungen aufweist. Miettinen speiste in die in seinem Team entwickelte Redescription-Mining-Software namens Siren die Informationen zum sozialen Hintergrund von 675 Politikern ein, außerdem deren Antworten auf 31 Fragen – etwa: „Sind Sie dafür, dass Sterbehilfe legalisiert wird?“

POLITIKERDATEN ALS TEST FÜRS REDESCRIPTION MINING

Pauli Miettinen ging es nicht darum aufzudecken, was welcher Politiker im Einzelnen denkt. Und dass er Politikerdaten verwendet hat, war eher ein Zufall und lag daran, dass er ganz einfach nach frei verfügbaren Daten über Menschen gesucht hatte, mit denen er Siren testen konnte. Die Politikerdaten sind frei verfügbar. Auf andere Personendaten hätte er aus Gründen des Datenschutzes gar nicht zugreifen können. Letztlich wollte er beweisen, dass es möglich ist, die Meinungen und Stimmungen in einer Gesellschaft anhand der Herkunft und der Aussagen der Menschen zu ermitteln.

„Unsere Datensätze sind weder riesig noch repräsentativ, doch machen sie das Prinzip klar“, sagt Miettinen. „Außerdem hat unsere Analyse gezeigt, dass ein Forscher ohne Softwarewerkzeug bereits bei einer solch überschaubaren Datenmenge überfordert wäre.“ Denn die Bezüge, welche die Software zwischen den beiden Datensätzen – in diesem Fall dem soziodemografischen Hintergrund und dem Antwortkatalog der Politiker – herstellt, sind zum Teil schwierig aufzuspüren. Zumindest, wenn eine Studie nicht von vornherein entsprechend ausgelegt wurde. So fand die Software unter anderem heraus, dass Personen zwischen 34 und 74 Jahren sowie Personen, die Kinder haben, Sterbehilfe eher ablehnen.

Solche Ergebnisse sind vor allem deshalb bemerkenswert, weil Siren sie

aus zwei Datensätzen gewonnen hat, die ursprünglich zu anderen Zwecken erhoben worden waren und eigentlich nichts miteinander zu tun haben. Im Fragenkatalog von 2015 wurde lediglich gefragt, ob man Sterbehilfe befürwortet oder nicht. Die Software aber stellt nun einen viel komplexeren Zusammenhang her, indem sie weitere Gemeinsamkeiten aufdeckt einerseits zwischen den Personen, die sich für Sterbehilfe aussprechen, und andererseits zwischen jenen, die dagegen sind. „Sie liefert im Nachhinein ganz neue Aussagen und generiert wertvolle Antworten auf Fragen, an die man damals noch gar nicht gedacht hatte“, sagt Miettinen.

Für wissenschaftliche Arbeiten können die von Siren ausgespuckten Korrelationen sehr interessant sein. Vor allem deshalb, weil die Software viele „und“/„oder“-Verknüpfungen präsentiert, die viele andere Datenanalyse-Programme in dieser Komplexität nicht liefern. Wissenschaftler können mit Siren ganz neue Hypothesen aufstellen – zum Beispiel: „Menschen im mittleren Lebensalter lehnen Sterbehilfe ab.“ Solche Aspekte können wiederum eine Anregung für zukünftige wissenschaftliche Studien oder Umfragen sein. Siren steht Forschern aller Disziplinen zur Verfügung und ist über den Link *siren.mpi-inf.mpg.de* kostenlos herunterzuladen.

Wissenschaftler können ihre Daten so einfach wie bei einem Statistikprogramm in die Software einspielen. Siren ermittelt dann innerhalb weniger Minuten eine Vielzahl von Korrelationen. „Natürlich sind manche Korrelationen trivial oder unsinnig“, sagt Pauli Miettinen. Eine Aussage wie: „Menschen über 60 interessieren sich weniger für Kinderkrippenplätze“ wäre zum Beispiel wenig überraschend.

Wie ein anderes Experiment Miettinens zeigt, ist Siren aber immer wieder für eine Überraschung gut. In diesem Fall fütterte er die Software zusammen mit Biologen mit Informationen zur Verbreitung der Säugetiere Europas. Der eine Datensatz enthielt 54 000 Einzelnachweise von Säugetieren mit Ortsangaben, der zweite die Klimadaten der verschiedenen Orte und Regionen – etwa Höchst- und Tiefsttemperaturen sowie die Niederschlagswerte. Auch diese Datensätze waren ursprünglich unabhängig voneinander erhoben worden, stammten aus verschiedenen Quellen und hatten eigentlich nichts miteinander zu tun. „Dieses Beispiel macht deutlich, mit welch großen Datenmengen man es oft zu tun hat, wenn man zwei Datensätze verknüpft“, sagt Miettinen.

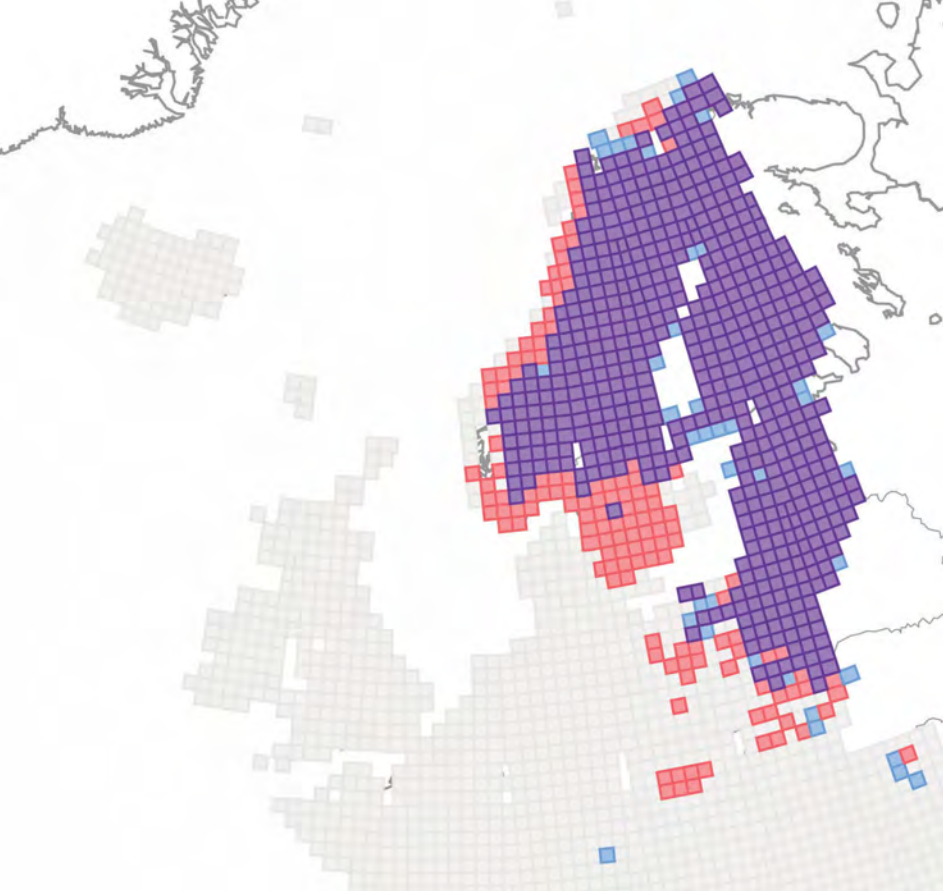
SIREN DEFINIERT REGELN UND AUSNAHMEN

Eigentlich sollte die Studie klären, inwieweit sich die Säugetierpopulationen in Europa mit der Erderwärmung verschieben könnten. Doch Siren lieferte unabhängig davon einige Korrelationen, die für Biologen aufschlussreich waren. Etwa zu den Lebensräumen von Elchen. Wie die Software herausfand, kommen Elche vor allem in Gebieten vor, in denen die maximale Temperatur im Februar zwischen minus zehn und null Grad Celsius liegt und im Juli zwischen zwölf und 25 Grad. Zudem beträgt der Niederschlag im August dort zwischen 57 und 136 Millimetern. Von dieser Regel gibt es allerdings auch Ausnahmen, die Siren gleich mitlieferte: So leben Elche auch an Norwegens Küste, wo im August mehr Regen fällt. Und in Österreich gibt es eine kleine Elchpopulation in einem Gebiet mit deutlich höheren Februartemperaturen.

Dank Siren verstehen Biologen die klimatischen Bedingungen, die für die Verbreitungsgebiete der Elche und anderer Säugetiere gelten, besser – obwohl es ihnen in ihrer Studie darum anfangs gar nicht ging. Sie müssen allerdings noch die Regeln definieren und entscheiden, wie sie etwa mit der österreichischen Elchpopulation umgehen: „Biologen können die Bedingungen so definieren, dass auch solche Habitate eingeschlossen werden, oder aber sie betrachten Situationen wie jene in Österreich als Anomalie“, sagt Miettinen.

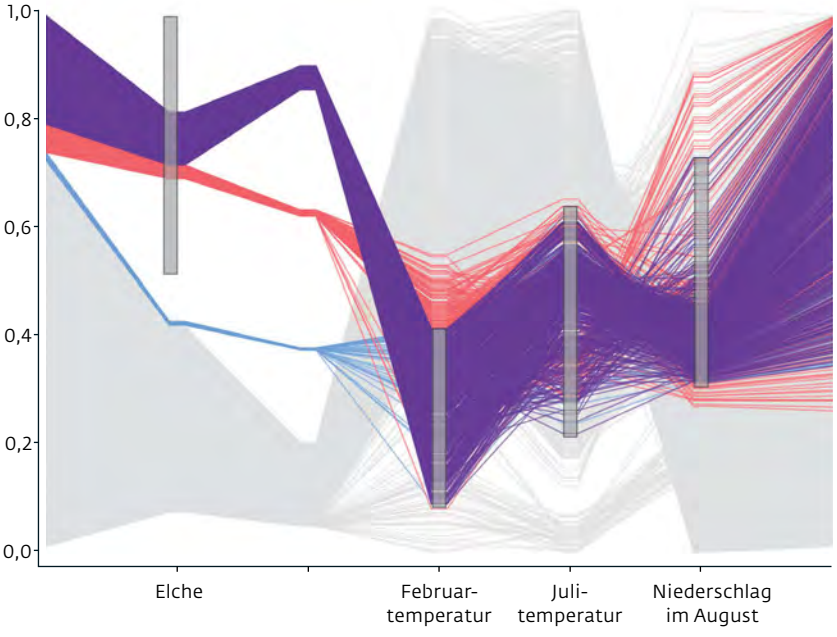
Softwarewerkzeuge wie Siren sind bisher selten, weil die Disziplin Redescription Mining noch relativ jung ist. Informatiker nutzen diese Methode erst seit etwa zehn Jahren. Zudem gibt es weltweit nur einige wenige Arbeitsgruppen, die sich damit befassen. Und das, obwohl Siren ausgesprochen vielseitig ist. Das Programm stellt nicht nur Korrelationen zwischen zwei unterschiedlichen Datensätzen her, sondern kann auch in einem einzigen Datenpool Bezüge finden. Eine Software so zu programmieren, dass sie so große Mengen an „und“/„oder“-Verknüpfungen oder Verneinungen wie „Wenn x zutrifft, ist y ausgeschlossen“ verarbeiten kann, sei eine Herausforderung, sagt Miettinen. >

Foto: MPI für Informatik



Oben Siren analysierte, ob sich die Lebensräume europäischer Säugetiere mit den klimatischen Bedingungen in dem jeweiligen Gebiet erklären lassen. Die violetten und roten Felder zeigen, wo Elche leben. In den violetten Bereichen stimmen die Klimaverhältnisse dabei mit den Erwartungen der Biologen überein: maximale Temperaturen im Februar zwischen minus zehn und null Grad Celsius, im Juli zwischen zwölf und 25 Grad und im August Niederschlagsmengen zwischen 57 und 136 Millimetern. In den roten Bereichen leben Elche, obwohl diese Kriterien nicht erfüllt sind. Vor allem das Vorkommen in einem Teil Österreichs mit deutlich höheren Februartemperaturen überraschte die Biologen. In den blauen Bereichen leben keine Elche, obwohl das Klima passt.

Unten Dieselben Zusammenhänge gibt die Grafik unten für die einzelnen Habitate wieder, die jeweils durch eine Linie repräsentiert werden. Ein Wert über 0,5 an der Marke Elche bedeutet, dass es die Tiere dort gibt, darunter nicht. Auch die durchschnittlichen Temperaturen und Niederschlagsmengen im Februar, Juli bzw. August sind relativen Werten zugeordnet. Die grauen Balken definieren hier jeweils die Erwartungen von Biologen. Zwischen der Angabe, ob es Elche gibt oder nicht, und den Werten für die Februartemperatur werden die Linien der einzelnen Habitate nach Farben bei willkürlichen Werten gebündelt. Wo die Linien den linken und rechten Rand der Grafik schneiden, hat keine Bedeutung.



„Es ist ziemlich schwierig, das in Algorithmen umzusetzen."

Die Arbeitsweise von Redescription-Mining-Programmen aber lässt sich relativ einfach erklären. Die Programme suchen nach Ähnlichkeiten zwischen den Objekten einer Menge – solche Ähnlichkeiten können gleiche Antworten der Politiker auf bestimmte Fragen, derselbe Bildungs- oder Familienstand oder dasselbe Alter sein. Zwischen allen diesen Aspekten stellt die Software Korrelationen her. Zuerst wählt sie simple, sogenannte schwache Korrelationen aus – so ordnet sie etwa Personen danach, ob sie Sterbehilfe ablehnen oder befürworten.

Diese einfachen Verknüpfungen werden dann in einem zweiten Schritt um präzisere Verknüpfungen ergänzt – beispielsweise um die Frage, ob Personen, die Sterbehilfe ablehnen, Kinder haben. In einem nächsten Schritt kann die Software dann das Alter berücksichtigen. Schritt für Schritt fügt die Software beliebige weitere Verknüpfungen hinzu und findet so die Objekte, die die größte Ähnlichkeit haben. Daraus wird dann die allgemeingültige Hypothese beziehungsweise Korrelation generiert.

Beim Redescription Mining testet das Programm zugleich, wie wahrscheinlich oder zutreffend eine gefundene Korrelation ist. In der Sprache der Informatiker klingt das so: Die Software

Grafiken: Pauli Miettinen/MPI für Informatik (2)

Pauli Miettinen, Sanjar Karaev und Saskia Metzler (von links) diskutieren, wie sie das Data Mining zukünftig weiterentwickeln können.

maximiert den „Jaccard-Koeffizienten" – einen Wert, an dem sich die Ähnlichkeit zwischen zwei sogenannten Support-Sets, etwa finnischen Politikern mit bestimmten Eigenschaften, misst.

MEHRERE ERKLÄRUNGEN FÜR EINEN DATENBESTAND

Gerhard Weikum, Direktor am Max-Planck-Institut für Informatik und Leiter der Abteilung Databases and Information Systems hält das Redescription Mining für ein „extrem nützliches Werkzeug" bei der Analyse großer Datenmengen. Beim Datamining geht es generell darum, in großen, mehrdimensionalen Datenbeständen interessante Muster zu finden. „Ein Analyst, der daraus Erkenntnisse ziehen will, braucht aber oft auch eine Erklärung oder kompakte Charakterisierung eines Musters", sagt Weikum. „An dieser Stelle ist Redescription Mining extrem nützlich, weil es nicht nur eine Erklärung für einen Datenbestand, sondern mehrere Erklärungen liefert."

Weikum nennt ein Beispiel: Ein Computerprogramm könnte in einem Personendatenbestand etwa ein Muster erkennen, das Personen umfasst, die bei einer Hightech-Firma arbeiten, jeden Tag lange Pendelstrecken zurücklegen und ein hohes Jahreseinkommen zwischen 100 000 und 300 000 Dollar haben. Redescription Mining würde aus den Daten eine alternative Beschreibung dieser Personengruppe generieren können, die so aussehen könnte: IT-Experten, die einen Universitätsabschluss in einem technischen Fach haben, aus Asien stammen und in einem US-amerikanischen Ballungsraum arbeiten.

Selbst wenn der Begriff Redescription Mining für Nichtinformatiker ungewohnt und abstrakt klingen mag, regt Pauli Miettinen Forscher anderer Disziplinen an, die Software zu nutzen. Sie sei einfach zu bedienen und für ganz verschiedene Fragestellungen nutzbar. Zudem eigne sie sich sowohl für sogenannte bestätigende als auch für explorative Analysen. Diese unterscheiden sich darin, dass eine Analyse entweder mit oder ohne Arbeitshypothese startet.

Ein Beispiel für eine bestätigende Analyse war die Studie über die Säugetierpopulationen, bei der erwartet wurde, dass der Klimawandel die Verbreitung verändern wird. Bei einer explorativen Analyse stürzt sich die Software hingegen ganz unvoreingenommen auf einen Datensatz. Insofern ist die explorative Analyse mit Redescription Mining geradezu eine Überraschungskiste, die alte Hypothesen stürzen oder auch neue hervorzaubern kann.

In der Regel nutzen die Anwender Siren allein. In schwierigen Fällen aber gibt Pauli Miettinen Unterstützung – etwa wenn unklar ist, ob die Daten überhaupt geeignet sind, um eine Hypothese zu überprüfen. Siren kann so manche wissenschaftliche Fragestellung in neuem Licht erscheinen lassen – und erinnert ein wenig an die Maschine aus dem Roman *Per Anhalter durch die Galaxis*, die einige Millionen Jahre rechnet, um auf die Frage nach dem Sinn des Lebens die Zahl 42 auszuspucken. Die ist freilich relativ nichtssagend. Den ratlosen Menschen rät die Maschine, sich auf die Suche nach der richtigen Frage zu machen, für die die Antwort „42" einen Sinn ergibt. Hätten sie Siren gehabt, hätten sie die richtige Frage vielleicht gefunden. ◂

AUF DEN PUNKT GEBRACHT

- **Mit einer Software namens Siren generieren Forscher des Max-Planck-Instituts für Informatik aus vorhandenen Daten neue Hypothesen. Diese Methode der Datenanalyse heißt Redescription Mining.**
- **Mit dem Programm Siren analysierten die Forscher unter anderem Zusammenhänge zwischen dem soziodemografischen Hintergrund und politischen Haltungen von Kandidaten für die finnischen Parlamentswahlen und die klimatischen Bedingungen der Habitate europäischer Landsäugetiere, speziell von Elchen.**
- **Die Software steht Forschern aller Disziplinen zur Verfügung und lässt sich über den Link *siren.mpi-inf.mpg.de* kostenlos herunterladen.**

Foto: Tom Pingel