# Treasure Hunt **in the Data Jungle**

Researchers normally formulate a hypothesis before beginning an experiment and collecting data. **Pauli Miettinen** from the **Max Planck Institute for Informatics** in Saarbrücken is turning this scientific principle on its head with a new procedure for analyzing data – redescription mining. The software can analyze existing datasets and retrospectively extract hypotheses and unexpected correlations. These, in turn, give scientists important clues for asking new questions – for example, when the task is to capture the political mood among the population.

TEXT **TIM SCHRÖDER**

Over the decades, computers have learned to complete specified tasks. They can solve complex equations, predict the weather, and now even reply in a human voice to such questions as "Where can I find a good, inexpensive Chinese restaurant near here?" However, Pauli Miettinen from the Max Planck Institute for Informatics in Saarbrücken has taken things one step further. He has taught computers to answer questions that nobody has asked them yet – and in this way to discern connections that humans wouldn't have noticed on their own.

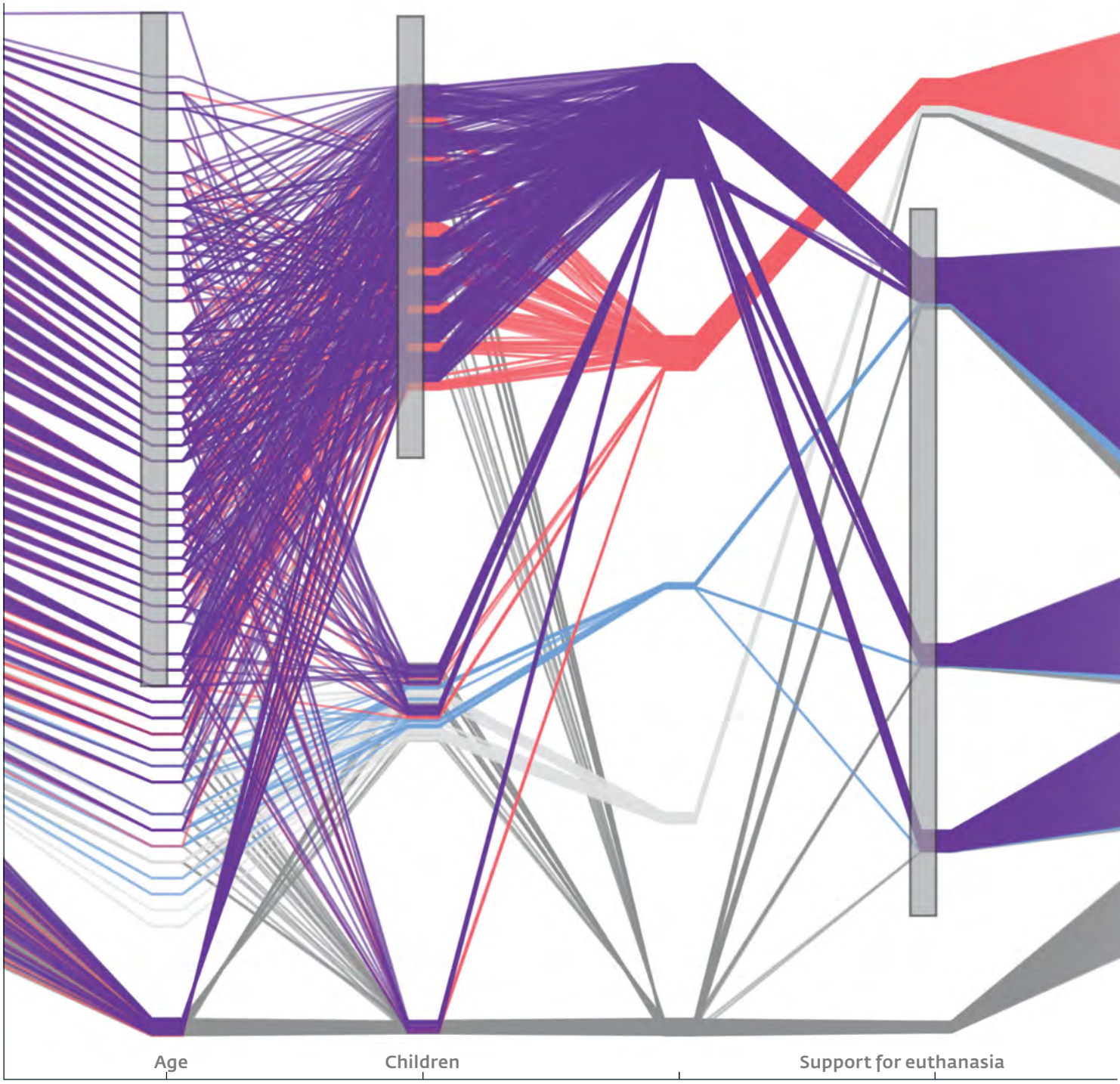
With that, Pauli Miettinen is pretty close to looking into a crystal ball. He himself describes his work a little more soberly: "Basically, all we're doing is generating a new hypothesis from existing data." That sounds modest, but it's nothing less than a minor revolution in the ways of scientific work. For centuries, researchers have always proceeded in accordance with the same template, regardless of the discipline. First they posit a hypothesis such as: "Man is descended from the apes." Then they test this hypothesis through observation and by collecting data.

## MEANINGFUL INFORMATION FROM LARGE VOLUMES OF DATA

The data analysis tool that Miettinen and his team developed turns this principle on its head. It uses existing data, analyzes it and makes entirely new connections – some of which are astonishing. The method used is pretty much the cutting edge in the world of data analysis. It's called redescription mining, which, freely interpreted, means something like "alternative description." In other words, Miettinen and his colleagues search for new correlations in existing data, for new statements contained in the data – for new ways of describing the data. In this way, they are helping to track down treasure in the data jungle.

Any kind of data can be analyzed with the method, and that, too, is one of the strengths of redescription mining – and the volume of data that can be processed is just about unlimited. For example, the procedure can help extract meaningful information from the large volumes of data that are collected everywhere today.

Pauli Miettinen and his colleagues showed what the method can do using data from his home country of Finland: information on Finnish politicians who stood as candidates for a seat in parliament in 2011 and 2015. For his analysis, the researcher combined two datasets: the first contained publicly available data on the politicians' social background, their age, origin, education and marital status. The second dataset contained replies to questions the politicians had answered for a web service. >




Graphic: Pauli Miettinen/MPI for Informatics

One line for every politician: This chart was produced by the Siren software in analyzing the sociodemographic data and political attitudes, in this case specifically on euthanasia, of candidates in the Finnish parliamentary elections. One finding: candidates over 34 and those with children are more likely to reject euthanasia.

Illuminating the data jungle: Pauli Miettinen and his staff developed software by the name of Siren (right-hand page) in order to identify associations in datasets that had not yet been formulated as a hypothesis at the time the data was collected.

Photo: Tom Pingel

Such web services have been extraordinarily popular for some years – the German Wahl-O-Mat website is fashionable, to name but one. The idea is that politicians and voters answer the same questions independently of each other, and the website reveals to the voter which party or candidate they have the greatest degree of agreement with. Miettinen fed the information on the social background of 675 politicians into Siren, the redescription mining software his team developed, as well as their answers to 31 questions, such as: "Are you in favor of legalizing euthanasia?"

## POLITICIANS' DATA AS A TEST OF REDESCRIPTION MINING

For Pauli Miettinen, it wasn't about discovering the details of what each politician thinks. And the fact that he used data from politicians was more a matter of chance and owed to the fact that he was simply looking for freely available data about people with which he could test Siren. Politicians' data is freely available. He wouldn't have been able to access other personal data for reasons of data protection. Ultimately, he wanted to prove that it's possible to determine the opinions and moods in a society based on where people come from and the statements they make.

"Our datasets are neither huge nor representative, but they reveal the principle clearly," says Miettinen. "Our analysis also showed that researchers without a software tool would be out of their depth even with a manageable volume of data such as this." Because the association that the software establishes between the two datasets – in this case, the sociodemographic background and the politicians' lists of answers – are sometimes hard to track down. At least if the study hasn't been correspondingly designed from the outset. For instance, the software found out, among other things, that people between 34 and 74 and people with children tend to reject euthanasia.

Such results are remarkable above all because Siren extracted them from two datasets that were originally collected for different purposes and actually have nothing to do with each other. In the 2015 list of questions, it was asked merely whether the respondent is in favor of euthanasia or not. The


software, however, establishes a much more complex connection by discovering further things in common, on the one hand between people who are in favor of euthanasia, and on the other, between those who reject it. "It delivers wholly new statements retrospectively and generates valuable answers to questions that no one had thought of at the time," says Miettinen.

The correlations identified by Siren can be very interesting for scientific work. Above all because the software presents many "AND"/"OR" links that many other data analysis programs can't identify with this degree of complexity. Scientists can use Siren to formulate completely new hypotheses – for example: "Middle-aged people reject euthanasia." Such aspects can, in turn, stimulate future scientific studies or surveys. Siren is available to researchers of all disciplines and can be downloaded free of charge from siren.mpi-inf.mpg.de.

Scientists can feed their data into the software as easily as with a statistics program. Siren then uncovers numerous correlations in a matter of minutes. "Of course, some correlations are trivial or meaningless," says Pauli Miettinen. A statement such as "People over 60 are less interested in available spots in daycare centers," for example, would hardly be surprising.

Time and again, however, Siren comes up with surprises, as another experiment of Miettinen's shows. In this case, he worked with biologists to feed the software with information on the distribution of Europe's mammals. One dataset contained 54,000 individual records of mammals with location details, and the second, the climate data of different locations and regions – for instance, the maximum and minimum temperatures and rainfall figures. These datasets, too, had originally been collected independently of each other, came from different sources and actually had nothing to do with each other. "This example underscores the sheer volume of data you often have to deal with when you link two datasets," says Miettinen.

## SIREN DEFINES RULES AND EXCEPTIONS

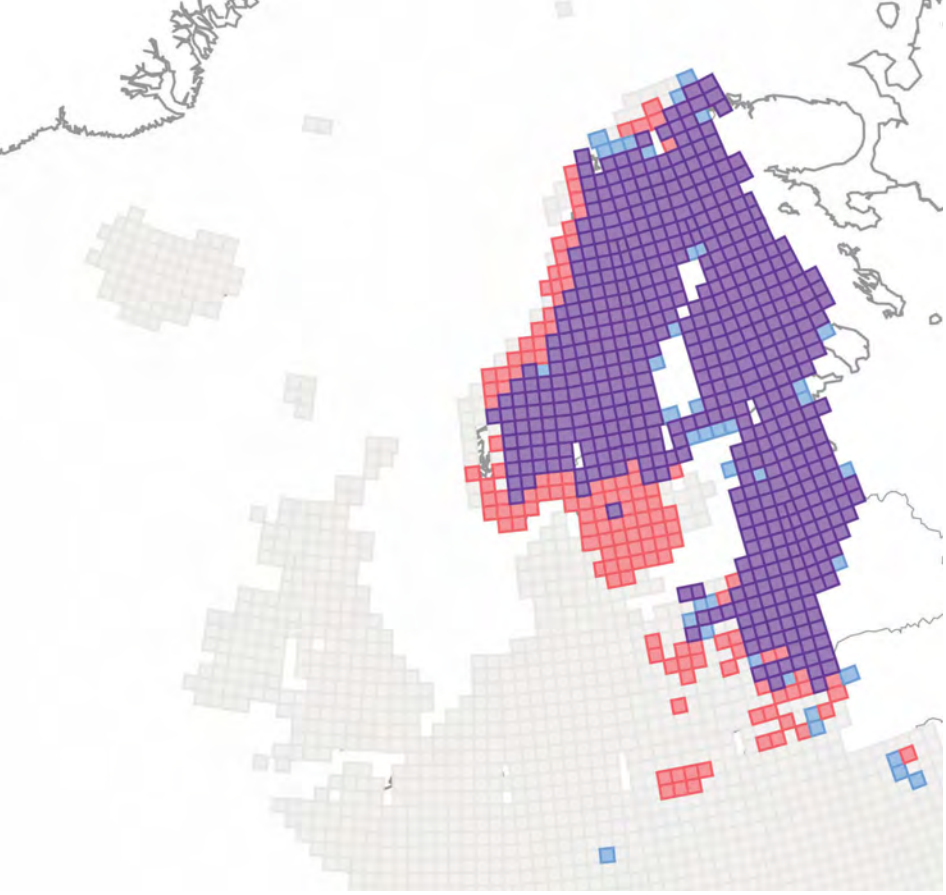
The study was actually supposed to clarify to what extent mammal populations in Europe might move in response to global warming. But Siren identified some unconnected correlations that were revealing for biologists – on the habitats of moose, for example. As the software discovered, moose are found primarily in regions in which the maximum temperature in February lies between -10 and 0 degrees Celsius, and in July between 12 and 25 degrees Celsius. In addition, the rainfall in August in these regions is between 57 and 136 millimeters. However, there are some exceptions to this rule, which Siren also identified: for example, moose also live on the coast of Norway, where there is more rainfall in August. And there is a small population of moose in Austria in a region with significantly higher temperatures in February.

Thanks to Siren, biologists can gain a better understanding of the climatic conditions that apply to the distribution of moose and other mammals – although this wasn't the original purpose of the study. However, they still have to define the rules and decide how to treat the Austrian moose population, for example. "Biologists can define the conditions in such a way that those habitats are also included, or they can view situations such as the one in Austria as an anomaly," says Miettinen.

Software tools such as Siren are rare to date, as the discipline of redescription mining is still relatively young – computer scientists have only been using this method for around 10 years. There are also only a few groups in the world working on this subject, even though Siren is decidedly versatile. Not only can the program establish correlations between two different datasets, but it can also find associations in a single data pool. Programming software to enable it to process large volumes of "AND"/"OR" links or negations such as "If x is true, y is impossible," is a challenge, says Miettinen. "It's fairly difficult to translate that into algorithms."

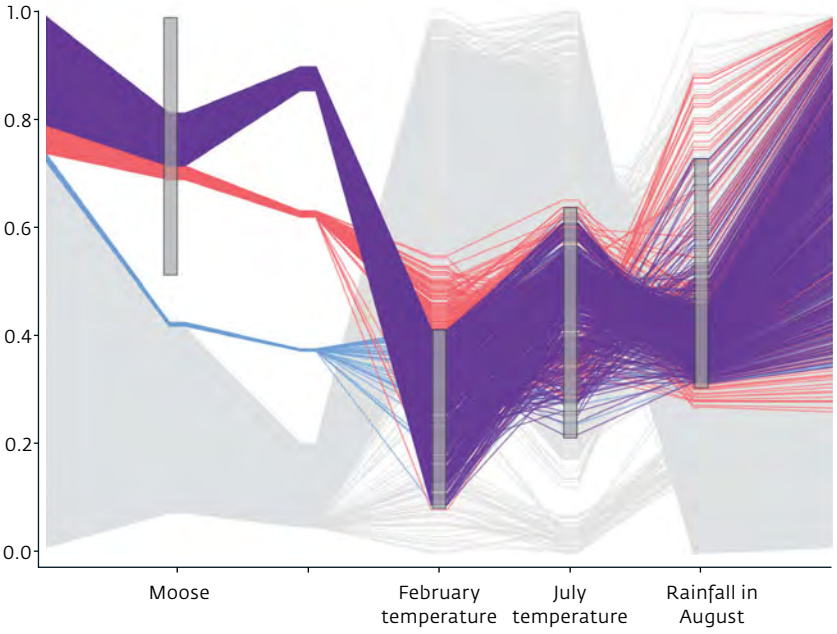
However, it is relatively easy to explain how redescription mining programs work. They search for similarities between the objects in a dataset – such similarities can be the same answers given by politicians to certain questions, the same level of education or marital status, or the same age. The

Photo: MPI for Informatics



***Above*** Siren analyzed whether the habitats of European mammals can be explained by the climatic conditions in the respective region. The purple and red fields show where moose live. The climatic conditions in the purple areas match the expectations of the biologists: maximum temperatures in February between -10 and 0 degrees Celsius, in July between 12 and 25 degrees Celsius and rainfall in August between 57 and 136 millimeters. Moose also live in the red areas even though these don't meet the criteria. The biggest surprise to the biologists was their presence in a region of Austria with significantly higher temperatures. There are no moose in the blue regions even though the climate fits.

***Below*** This chart shows the same associations for the individual habitats, each represented by a line. A value of over 0.5 for moose signifies that the species occurs there, and a value below 0.5, that it doesn't. The average temperatures and volumes of rainfall in February, July and August have also been assigned relative values. The gray bars define the models Siren built in each case. The lines for the individual sites are grouped depending on whether moose are present and on what values are encountered for temperature and precipitation, and are colored accordingly. It doesn't matter where the lines intersect with the left and right edge of the chart.




software establishes correlations between all these aspects. First it selects simple, so-called weak correlations – for example, it classifies people according to whether they are in favor of euthanasia or reject it.

These simple associations are then complemented by more precise associations in the second step – for example, by the question of whether people who reject euthanasia have children. In the next step, the software takes age into account. Step by step, the software adds any number of additional links, and in this way identifies the objects that have the greatest similarity. These results are then used to generate the universal hypothesis or correlation.

## SEVERAL EXPLANATIONS FOR ONE DATASET

With redescription mining, the program simultaneously tests how probable or accurate any discovered correlation is likely to be. As a computer scientist would put it: the software maximizes the "Jaccard coefficient" – a value by which the similarity between two so-called support sets can be measured, such as Finnish politicians with certain characteristics.

Gerhard Weikum, Director at the Max Planck Institute for Informatics

Graphics: Pauli Miettinen/MPI for Informatics (2)

Pauli Miettinen, Sanjar Karaev and Saskia Metzler (from left) discuss how they will be able to refine data mining in the future.

and Head of the Databases and Information Systems Department, regards redescription mining as an extremely useful tool when it comes to analyzing large volumes of data. The purpose of data mining is generally to find interesting patterns in large, multidimensional databases. "An analyst wanting to draw conclusions from it often also needs an explanation or compact characterization of a pattern," says Weikum. "Redescription mining is extremely useful in such cases because it supplies not just one explanation for a database but several."

Weikum gives an example: A computer program could recognize a pattern in a database comprising people who work for a high-tech company, have a long commute every day and earn a high annual salary of between 100,000 and 300,000 dollars. Redescription mining would be able to generate an alternative description of this group from the data that might look as follows: IT experts who have a university degree in a technical field, come from Asia and work in a US metropolitan area.

Even if the term redescription mining sounds unfamiliar and abstract to non-computer scientists, Pauli Miettinen encourages researchers from other disciplines to use the software. It's easy to operate, he says, and can be used for very different questions. In addition, it's suitable for both so-called confirmatory and exploratory analyses, he says. These differ in that an analysis starts either with or without a working hypothesis.

An example of a confirmatory analysis was the study of mammal populations where it was expected that climate change will change their distribution. In an exploratory analysis, on the other hand, the software tackles a dataset with no preconceptions. In that regard, an exploratory analysis with redescription mining is essentially a surprise package that can overturn old hypotheses or conjure up new ones.

Users generally work with Siren on their own. In difficult cases, however, Pauli Miettinen adds support – for instance if it is unclear whether the data is fundamentally suitable for reviewing a hypothesis. In this way, Siren can show many scientific questions in a new light – and it's a little reminiscent of the machine from the book *The Hitchhiker's Guide to the Galaxy*, which calculates for several million years only to spit out the number 42 in response to the question of the meaning of life. That is, of course, relatively meaningless. The machine advises the baffled person to embark on a search for the right question for which the answer "42" makes sense. If they had had Siren, they might have found the right question. ◂

## TO THE POINT

- **Researchers at the Max Planck Institute for Informatics use a software known as Siren to generate new hypotheses from existing data. This method of data analysis is called redescription mining.**
- **Using Siren, the researchers analyzed, among other things, the connections between the sociodemographic background and the political attitudes of candidates in the Finnish parliamentary elections, as well as the climatic conditions prevailing in the habitats of European land mammals, specifically of moose.**
- **The software is available to researchers of all disciplines and can be downloaded free of charge from siren.mpi-inf.mpg.de.**

Photo: Tom Pingel