

# WÖRTERBUCH AUS DATEN

Jeder und jede von uns erzeugt nahezu täglich Unmengen an Daten.  
Am Institut für Mathematik arbeitet START-Preisträgerin Karin Schnass mit ihrem Team daran, diese Daten auch optimal auszuwerten.

**D**ie Fotos vom letzten Urlaub in den Anden, zwei Staffeln „Breaking Bad“ als Netflix-Download, drei unfertige Text-Dokumente und eine ganze Reihe weiterer fertiger, gefühlte 5.000 E-Mails, alle Nirvana-Alben als MP3, und das ist nur der private Computer: Wir alle sammeln und erzeugen nahezu täglich Unmengen an Daten. Manche davon relevant, andere weniger; manche komplex, andere einfach strukturiert. Insgesamt gibt es Schätzungen,

*„Daten liegen praktisch nie so vor, wie man sich das theoretisch überlegt hat.“* Karin Schnass

dass vom weltweiten Datenvolumen – inzwischen zumindest einige Billionen Gigabyte – maximal ein Viertel auch nützlich ist und nur etwa ein Prozent tatsächlich analysiert werden kann.

Mit Methoden dieser Analyse beschäftigt sich die Mathematikerin Karin Schnass in ihrem vom österreichischen Wissenschaftsfonds FWF mit einem START-Preis geförderten Projekt: „Viele stellen sich die Analyse von Daten relativ einfach vor: Schlagworte sind dann Deep Learning, Machine Learning, das passiert alles automatisch und dann machen wir wunderbare Dinge damit. Leider ist es nicht so einfach: Daten sind nämlich leider nie wirklich sauber und liegen praktisch nie so vor, wie man sich das theoretisch überlegt hat. Nehmen wir als Beispiel die Auswertung von Fotos: Sie haben Bilder von Leuten, da ist dann einmal der Kopf schief, einmal ist die Person bei den Ohren, einmal bei der Stirn abgeschnitten. Aber man hätte gerne, dass der Kopf immer die gleiche Größe hat und zentriert am Bild ist – dabei ist aber die Tatsache, dass jemand einen sehr großen Kopf hat, auch ein Merkmal





**ZWEI BEISPIELE** für automatisch erstellte Dictionaries: Das Dictionary des Affen (4. Bild von links) enthält mehr Atome, weil das Bild komplexer ist. Das Dictionary der Paprikas hingegen kommt mit weniger Atomen aus (2. Bild von links).

und das gehört berücksichtigt. Bis man die Daten einmal soweit hat, dass das alles passt, hat man extrem viel Zeit investiert.“

### Dictionary Learning

Die Methode, die Schnass und ihr Team für die Auswertung von Daten – Bilder, Audio, aber theoretisch jede Art von Datenquelle – verwendet, heißt „Dictionary Learning“: Ein Bild wird in Stücke zerschnitten, diese Bildstücke bilden die Datenquelle. „Üblich ist dabei eine Größe von 8 x 8 Pixel, also 64 Pixel pro Stück. Was wir wollen, ist, dass alle denkbaren Formen, die im Bild vorkommen, auch in den Stücken repräsentiert sind“, erklärt Schnass. Diese Stücke werden dem Lern-Algorithmus übergeben, der am Ende ein Dictionary, ein Wörterbuch, der gewünschten Größe ausspuckt, aus dem sich das Ausgangsbild wieder zusammensetzen lässt. Wie viele einzelne Teile – in der Fachsprache „Atome“ – gebraucht werden, hängt natürlich von der Komplexität des Ausgangsbilds ab (siehe Bilder). „Es gibt eine Reihe von Bildern, an denen die Algorithmen weltweit getestet werden, unter anderem eines von Paprikas und eines von einem Affen. Der Affe ist aufgrund der feinen Fellstruktur schwieriger, da werden insgesamt mehr Atome gebraucht als bei den Paprikas, die aus vielen glatten Flächen bestehen“, sagt die Mathematikerin.

Anwendung findet die Methode allerdings weniger in der Kompression von Daten – dafür gibt es effizientere Verfahren – als in der Wiederherstellung von beschädigten Portionen. „Stellen Sie sich ein zerknülltes Foto vor: Die Risse, die beim Zerknüllen entstehen, sind Leerstellen, die wir aber gerne gefüllt hätten. Nun kommt uns zugute, dass Bilder ja nie komplett zufällig zusammengesetzt sind, sondern immer etwas zeigen – wir können also unseren Algorithmus sogar

über das zerstörte Foto laufen lassen und erhalten unbeschädigte Atome.“ Mit diesen unbeschädigten Atomen können die Forscherinnen und Forscher dann beschädigte Teile des Bilds wiederherstellen.

„Daten, egal welcher Form, kann man sich als Punkte in hochdimensionalen Räumen vorstellen. Aber es ist nicht so, dass diese Räume voll sind, also dass ein Datenpunkt tatsächlich ein ganz beliebiger Punkt im Raum sein kann – meistens sind diese Punkte in irgendwelchen Strukturen angeordnet, etwa auf einer Linie“, sagt Karin Schnass. Diese Eigenschaft von Daten machen sich die Forscherinnen und Forscher zunutze: „Wenn ich jetzt nicht die gesamten Informationen habe, kann ich dennoch von vorhandenen Informationen auf das Gesamtbild schließen. Ich habe zum Beispiel ein Computertomografie-Gerät, will aber Patienten nicht einer hohen Strahlenbelastung aussetzen und viel Energie verbrauchen, al-

so mache ich weniger Messungen. Dann habe ich zwar weniger Informationen, aber ich weiß, dass die Daten auf einer bekannten Struktur liegen. Ich verwende also bekannte Zusatzinformationen, um Daten zu rekonstruieren und von weniger Punkten auf das Gesamtbild zu schließen. Bei der Magnetresonanztomografie wird das bereits verwendet: Man spart sich Messungen, weil man weiß, dass die Daten eine niedrigdimensionale Struktur haben, und das verwendet man, um die Anzahl der Messungen zu reduzieren. Auch ein Dictionary ist eine Art einer niedrigdimensionalen Struktur.“

Denkbar ist ein Einsatz auch in Szenarien, wo Energie nur begrenzt zur Verfügung steht: Wenn etwa ein Mars-Rover nur die nötigsten Messungen vornimmt, hält die Batterie länger. Der Aufwand, die gesendeten Daten dann auf der Erde umzurechnen, ist größer, hier spielt der Energieverbrauch aber auch eine geringere Rolle. „Das gleiche Phänomen haben wir übrigens auch bei handelsüblichen Kameras, wenn in JPG gespeichert wird: Die Kamera macht derzeit das Foto und rechnet dann selbst auf das komprimierte JPG-Format runter. Das benötigt natürlich auch Energie. Ein anderer Zugang wäre, wenn man wirklich viele Bilder speichern will und die Batterie lang halten soll, auch hier mit der Kamera nur die nötigsten Messungen für ein Foto zu machen und den Rest erst später am Computer rechnen zu lassen. Aber das ist derzeit noch Zukunftsmusik.“

Karin Schnass und ihr Team arbeiten im Moment an der Verbesserung ihrer eigenen Dictionary-Learning-Algorithmen und an möglichen Einsatzgebieten – so ist es etwa alles andere als trivial, automatisch festzustellen, wie viele Atome im Dictionary tatsächlich gebraucht werden, vor allem, wenn die Ausgangsdaten schon Fehler aufweisen. Das Projekt ist derzeit bis 2021 angesetzt. *sh*

**KARIN SCHNASS** (\*1980 in Klosterneuburg) erwarb 2004 ihr Diplom in Mathematik an der Universität Wien, bevor sie im Jahr 2009 an der ETH Lausanne in der Schweiz promovierte. Nach Geburt ihres ersten Kindes und einer Karenz 2009 forschte die Mathematikerin als Postdoc am Johann Radon Institute for Computational and Applied Mathematics (RICAM) in Linz. Nach Geburt ihres zweiten Kindes 2011 und einer Karenz arbeitete sie als Erwin-Schrödinger-Stipendiatin des Wissenschaftsfonds FWF an der Università degli studi di Sassari in Italien, bevor sie im Oktober 2014 an die Universität Innsbruck wechselte. Für ihr Projekt „Optimierung, Modelle und Algorithmen für Dictionary Learning“ erhielt Karin Schnass 2014 einen START-Preis, mit bis zu 1,2 Millionen Euro die höchstdotierte Auszeichnung für junge Wissenschaftlerinnen und Wissenschaftler in Österreich.