BIG DATA AUF DEM WEG ZUR ANGEWANDTEN DATENWISSENSCHAFT

Big Data lebt vom Experimentieren mit Modellen. Das verlangt das Verständnis von Daten, Algorithmen und Anwendungsdomänen. Unser Ziel muss aber die Etablierung einer angewandten Datenwissenschaft zur sinnvollen Nutzung in der Praxis sein.

Von Prof. Dr. Reinhard Riedl

Big Data wird oft über seine charakteristischen Eigenschaften definiert: Sehr viele und sehr unterschiedliche Daten werden mit hoher Geschwindigkeit verarbeitet und dabei sind die Qualität und Relevanz der Daten unterschiedlich und ihre Gültigkeit von kurzer Dauer. Deshalb spricht man von 3-V-, 6-V-, N-V-Big-Data, wobei V für Volume, Variety, Velocity, Variability, Veracity, Validity, Visualization, Value etc. steht.

Wissen, Modelle und Daten – zwei Kehrseiten einer Medaille

Diese Zuschreibungen stimmen oft. Wesentlich für das Verständnis der Anwendung ist aber, dass bei Big Data das Modell zur Datenauswertung aus den Daten selbst abgeleitet wird. Das kann – je nachdem, wie man es macht – zu seriösen Resultaten oder verantwortungslosen Umtrieben führen. In die Kategorie seriös gehört, dass beim Arbeiten mit aus Daten abgeleiteten Modellen die wissenschaftlichen Qualitätsstandards gelockert werden, um den Spielraum für Experimente zu erhöhen. Man darf bei Big Data viele Dinge tun, ohne vorerst genau zu wissen, was man dabei tut, solange man im Nachhinein die Ergebnisse in Bezug auf ihre Gültigkeit und Relevanz überprüft. Denn die Qualität der Resultate wird nicht im Vorhinein beurteilt, sondern erst im Nachhinein. Diese Überprüfung ist aber für seriöse Anwendungen unerlässlich.

Zu den frivolen Umtrieben gehört es dagegen, dass viele Big-Data-Evangelisten behaupten, dass Big Data die traditionellen Wissenschaften überflüssig machen würde, weil in Zukunft Wissen durch Daten ersetzt werden könne. Nachzulesen sind solche Visionen von einer Null-Wissen-Datenwissenschaft in den einschlägigen Fachbüchern der Jahre 2012 bis 2014. Richtig ist das Gegenteil: Big Data ist deshalb so wertvoll, weil es ganz neue Valorisierungen von Wissen ermöglicht. Es passt sehr gut zum assoziativen Denken von uns Menschen und ist ideal geeignet für den Transfer von alten Erkenntnissen in neue Anwendungsbereiche. Damit schafft es „Economies of Scale" für menschliche Talente, weil diese viel breiter genutzt werden können.

Mittlerweile ist die Begeisterung für den Null-Wissen-Hype am Abflauen. Dazu hat das bisweilen spektakuläre Versagen von Big Data aufgrund ungenügender Modelle kräftig beigetragen, z. B. das Scheitern des Google-Grippeindikators. Auch aufgrund der Erfahrungen anderer Firmen mit unüberlegtem Einsatz von Big Data vermehrt sich das Bekenntnis zu einer seriösen Datenwissenschaft.

Datenwissenschaft und europäische Emanzipation

Ausgangspunkt einer angewandten Datenwissenschaft sind die Nutzenfokussierung und der Verzicht auf falsche Perfektion: Die beim Big Data angewandten Algorithmen sollen in der Summe möglichst großen Nutzen bringen, dafür soll aber NICHT für einen einzelnen Algorithmus der Anspruch bestehen, dass er immer korrekte Resultate liefert. Denn die Resultate müssen im Nachhinein sowieso genau überprüft werden. Das Ex-post-Reflektieren der Resultate ist genauso wichtig wie das Ex-ante-Wählen der richtigen Methoden. Somit steht Data-Science quasi in der Mitte zwischen richtiger Wissenschaft (die sich wesentlich über ex ante korrekte Methoden definiert) und der reinen Mathematik (die sich über den Ex-post-Beweis von Erkenntnissen definiert). Während – überspitzt formuliert – in der Wissenschaft die Methode alles und in der Mathematik die Methode nichts ist, spielen in der seriösen Datenwissenschaft die Methoden der Herleitung ebenso eine Rolle wie die nachträgliche Überprüfung der Resultate aus sich selbst heraus. Man nutzt quasi das Beste aus mehreren Welten.

> „Man darf bei Big Data viele Dinge tun, ohne vorerst genau zu wissen, was man dabei tut, solange man im Nachhinein die Ergebnisse in Bezug auf ihre Gültigkeit und Relevanz überprüft."

Daraus folgt, dass Datenwissenschaft auf drei Pfeilern ruhen sollte: dem Pfeiler, gebildet aus Methoden zur Erzielung von Resultaten, dem Pfeiler, gebildet aus praktischen Prinzipien zur Validierung von Resultaten in den jeweiligen Anwendungsbereichen, und dem Pfeiler, gebildet aus rechtlicher und ethischer Hinterfragung des Vorgehens. Die Schweizer Regierung hat im Frühsommer dieses Jahres ein nationales Forschungsprogramm Big Data beschlossen, das diese Auslegeordnung im Kern abbildet: Ein Modul beschäftigt sich mit Rechen- und Informationstechnologie, ein Modul mit Anwendungen und ein Modul mit gesellschaftlichen, regulatorischen und bildungsbezogenen Herausforderungen. In Letzterem geht es insbesondere auch um Rechts- und Ethikfragen.

Ich bin Mitglied des Steuerungsausschusses dieses Forschungsprogramms und muss mich immer wieder kritischen Fragen stellen, ob denn das Thema Big Data ein nationales Forschungsprogramm – noch dazu eines der bestdotierten in der Geschichte des Schweizer Nationalfonds – verdiene, wo doch die Amerikaner bereits alle Fragen beantwortet hätten. Ich meine, dass das Outsourcen wissenschaftlicher Exzellenz zu den Amerikanern in einem so zukunftskritischen Bereich wie der angewandten Datenwissenschaft für die Schweiz ein No-Go ist. Und dass es auch für Österreich keine sinnvolle Option wäre. Denn es geht darum, von Big Data frühzeitig und umfassend zu profitieren – und nicht etwa mit zehn Jahren Verspätung – sowie außerdem darum, die Standards für die Anwendung von Big Data mitzubestimmen.

> „Big Data ist deshalb so wertvoll, weil es ganz neue Valorisierungen von Wissen ermöglicht."

Big Data, ein weites Land

Wobei es nicht DAS Big Data gibt, sondern sehr unterschiedliche Methoden und Ziele, die damit verfolgt werden. In diesem weiten Land dreht sich vieles darum, aus den Daten über viele ein maßgeschneidertes Handeln für die eine oder den einen abzuleiten. Aber es geht auch darum, Muster und Zusammenhänge besser zu erkennen, weil man die Phänomene ganzheitlich betrachtet und auf vermeintlich effiziente Fokussierungen verzichtet.

Eine der einfachsten Formen von Big Data ist die Methode der invertierten Files, die den Algorithmen in modernen Suchmaschinen zugrunde liegt. Statt mathematisch zu erklären, wie das funktioniert, will ich es am Beispiel von Foundation Medicine erläutern, bei der man sowohl eine Verbreiterung der Perspektive als auch eine skalpellhafte Präzisierung der Therapie anstrebt. Bei Foundation Medicine werden für relevante genetische Veränderungen, die in einer Tumorzelle auftreten können, alle verfügbaren Informationen zusammengestellt. Dafür werden unter anderem auch Daten aus Patientendossiers genutzt, von Patienten, deren Tumore ebenfalls diese genetischen Veränderungen aufweisen. Wenn nun ein Patient behandelt werden soll, werden alle relevanten genetischen Veränderungen in seinem Tumor identifiziert und für jede dieser Veränderungen wird dem behandelnden Arzt die Informationssammlung zu dieser Veränderung inklusive Therapieoptionen zur Verfügung gestellt. Das ermöglicht eine für die spezielle Art des Tumors gezielte Therapie, die spezifische Studienergebnisse, die bisherigen Therapieerfahrungen und das Wissen um das Individuum Mensch berücksichtigt.

Das Beispiel zeigt, dass schon einfaches Big Data großen Nutzen bringen kann. Doch die angewandte Datenwissenschaft kann noch viel mehr. Wir sollten uns auch in Österreich proaktiv mit ihr beschäftigen!

Herzlichst
Ihr Reinhard Riedl