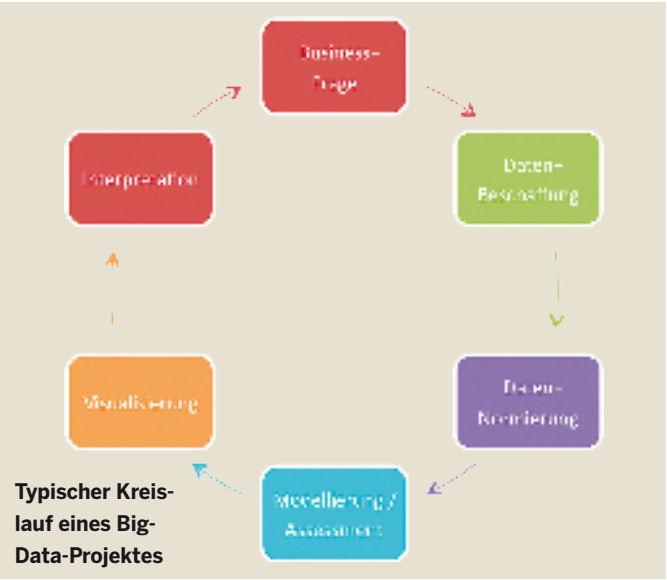
Big Data – mehr Schein als Sein? Mit dem Begriff Big Data wird nicht nur eine, sondern ein ganzes Portfolio neuer Technologien assoziiert. Physiker, Mathematiker, Informatiker, Soziologen, Ökonomen und Wissenschaftler aus anderen Bereichen setzen grosse Hoffnung in die Analyse grosser Datenmengen. Doch diese Analyse kann nur dann sinnvoll eingesetzt werden, wenn man organisatorisch vorbereitet ist.

VON MARCEL BLATTNER*

Nach der Definition von Gartner (2012) hat man ein Big-Data-Projekt, wenn folgende Voraussetzungen gegeben sind: Volume (Datenmenge), Velocity (Änderungsrate und Durchsatz), Variety (Heterogenität). Diese Definition ist aus verschiedenen Gründen problematisch. Das offensichtlichste Problem besteht darin, dass die Datenmenge immer relativ zu einer Infrastruktur betrachtet werden muss. Was für eine Unternehmung mit einer kleinen Infrastruktur ein Big-Data-Projekt ist, wird von einem Unternehmen mit einer grossen Infrastruktur nicht unter diesem Label geführt. Ein weit subtileres Problem zeigt sich darin, dass sich offenbar in der letzten Zeit die Meinung durchsetzt, dass nur dann sinnvolle Analysen und Modellierungen gemacht werden können, wenn man ein Big-Data-Projekt hat. In vielen Köpfen hat sich offenbar die Gleichung manifestiert: gewinnbringende Analyse von Daten = Big-Data-Projekt. In dieser Gleichung steckt vor allem eine Annahme, die man genauer betrachten muss.

Mehr Daten sind immer besser? Die Haltung, dass mehr Daten immer besser sind, geht davon aus, dass mehr Daten auch mehr Informationen enthalten. Dies muss überhaupt nicht der Fall sein. Es kann durchaus sein, dass ein wichtiges Signal in den Daten durch zusätzliche Anreicherung immer mehr «verrauscht» wird und somit nicht mehr entdeckt werden kann. Statistische Fallen wie der Yule-Simpson-Effekt können dazu führen, dass die Kombination von Variablen aus verschiedenen Datensätzen eine gegenteilige Signatur offenbart. Zum Beispiel: Positiv gemessene Trends in den einzelnen Variablen können sich bei Kombination in einen Negativtrend «transformieren». Erfolgversprechender hingegen wäre es, wenn man die Variablen einzeln innerhalb der jeweiligen Datensätze untersuchen würde. Eine weitere Schwierigkeit besteht darin, dass sich bei ausreichend grosser Datenmenge Zusammenhänge von Variablen zufälligerweise ergeben können. Je grösser die Datenmenge ist, desto höher ist die Wahrscheinlichkeit, dass dies passiert. Erkennt man diese zufälligen Zusammenhänge nicht, können falsche Schlüsse und damit falsche Resultate die Folgen sein. Mehr Daten sind also nicht immer besser. Anstatt sich auf die Datenmenge zu konzentrieren, wäre es besser, ein Augenmerk auf die Datenqualität zu legen. Aber wer kann jetzt etwas mit diesen Daten anfangen? Wer versteht es, aus diesen Daten Prognosemodelle abzuleiten?



Typischer Kreislauf eines Big-Data-Projektes

Data Scientist. Da eine Aggregation von Daten noch keine sinnvolle Interpretation zulässt, braucht es Spezialisten, welche die Daten analysieren, die Resultate interpretieren und bewerten und schliesslich so aufbereiten, dass eine Entscheidungsgrundlage für eine nachfolgende Aktion vorliegt. Diese Prozesskette ist äusserst komplex und kann nicht von einem Individuum alleine gemeistert werden. Es gibt immer mehr Unternehmen, welche Vakanzen mit dem Label «Data Scientist» publizieren. Liest man solche Stelleninserate, wird einem schnell klar, dass eine Einzelperson die nötigen Skills auf einem hohen Niveau für die ganze Prozesskette unmöglich mitbringen kann. Dazu braucht es ein Team, das aus Akteuren von unterschiedlichen Gebieten zusammengesetzt ist – technisches Personal sowie Analysten mit mathematischem Hintergrund und Entscheidungsträger aus strategisch wichtigen Abteilungen wie Marketing, Verkauf etc.

Strategisch relevante Fragestellungen. Ein Big-Data-Projekt beginnt nicht mit der Technologie, sondern mit strategisch relevanten Fragen, die den Dreh- und Angelpunkt des gesamten Projektes darstellen. Zum Beispiel kann sich eine HR-Abteilung die Frage stellen, welche Faktoren dafür verantwortlich sind, dass in gewissen Bereichen hohe Personalfluktuationen

vorherrschen. Für die Ausarbeitung der Fragestellung muss viel Zeit eingeplant werden. Es sind ebendiese Fragen, welche die Weichen stellen, um das weitere Vorgehen in die richtigen Bahnen zu lenken (siehe Abbildung).

Sind die relevanten Fragen einmal fixiert, sollte man sich Gedanken über die nötigen Daten machen und darüber, woher man diese bekommt. Eigene Daten werden dabei oft angereichert durch Daten aus den sozialen Medien, Open-Government-Daten oder bereits statistisch aufbereitete Daten, z.B. vom Bundesamt für Statistik. Die Datenmenge und das Format (strukturiert und/oder unstrukturiert) geben ein Stück weit vor, welche Speichertechnologien zum Zuge kommen und ob z.B. eine Cluster-Infrastruktur wie Hadoop wirklich nötig ist.

Sind diese Daten einmal gespeichert, müssen sie in eine Form gebracht werden, damit eine Maschine (ein Algorithmus) etwas damit anfangen kann. Variablen müssen transformiert, skaliert werden. Ausserdem wird man oft mit sehr vielen Variablen konfrontiert (mehrere Hundert Variablen sind keine Ausnahme). Damit trotzdem sinnvolle Analysen gemacht werden können, wendet man in einem vorgelagerten Schritt eine Technik an, welche «dimensionality reduction» genannt wird. Damit werden die wichtigen Variablen kombiniert und in einem neuen (reduzierten) Datensatz dargestellt.

Modelle und Algorithmen abwägen. Nachdem die Daten aufbereitet sind, werden diese den Algorithmen zugeführt. Man wählt eine Menge an Algorithmen aus, welche für die entsprechende Aufgabe (Frage) geeignet sind. Zum Beispiel möchte man herausfinden, wie wahrscheinlich es ist, dass ein Mitarbeiter in naher Zukunft seine Stelle kündigt. Einerseits hat man die ausgewählten Variablen und zusätzlich ist man im Besitz der «ground truth», welche Mitarbeiter und wann diese in der Vergangenheit gekündigt haben. Deshalb wäre bei dieser Fragestellung ein Algorithmus auszuwählen, der zur Gruppe «supervised learning» gehört. Der Algorithmus wird dann auf den historischen Daten trainiert, um eine Prognose für Mitarbeiter zu erstellen, welche noch nicht gekündigt haben. Dabei spielt der Tradeoff zwischen der Komplexität eines Algorithmus und erzieltem Resultat (richtige Voraussagen) eine zentrale Rolle. Es hat keinen Sinn, wenn man einen hochkomplexen Algorithmus einsetzt, welcher gegenüber einer einfacheren Lösung nur marginale Verbesserungen liefert und gleichzeitig mehrere Magnituden mehr an Rechenzeit braucht. Mit der Auswahl und Fütterung der Algorithmen ist es aber nicht getan. Man muss verschiedene Modelle und Algorithmen gegeneinander abwägen. Dabei muss ein besonderes Augenmerk auf die Generalisierbarkeit der Algorithmen gelegt werden, denn es nützt nichts, wenn eine Methode nur auf den bestehenden Daten funktioniert und bei neuen Datensätzen gleicher Herkunft keine sinnvollen Resultate mehr produzieren kann. Man nennt dieses Abwägen in der Fachsprache «bias-variance tradeoff». Das ganze Model Assessment ist eine äusserst heikle Angelegenheit. Es kann nicht genug betont werden, dass diese Prozesse (noch) nicht alleine von einer Maschine bewältigt werden können. Die Evaluierung muss durch erfahrene Spezialisten gemacht werden.

Visualisierung – und neue Fragen. Sind das optimale Modell und der zugehörige Algorithmus einmal ausgewählt, müssen die Resultate visualisiert und interpretiert werden. Dieser Schritt ist enorm wichtig, da er die Entscheidungsgrundlage liefern wird, wie eine Strategie allenfalls angepasst werden muss. Zum Beispiel kann eine HR-Abteilung durch die Analysen ihre Strategie bei Einstellungsgesprächen ändern. Ein eindrückliches Beispiel dazu liefert Xerox. Eine hohe Fluktuationsrate im Bereich Callcenter veranlasste Xerox, die relevanten Daten genauer zu untersuchen. Es stellte sich heraus, dass z.B. die Erfahrung als Callcenter-Mitarbeiter vor der Anstellung praktisch kein Indikator war, ob ein Mitarbeiter früher oder später kündigen wird. Hingegen spielten spezifische Persönlichkeitsmerkmale eine sehr starke Rolle. Nach dieser Erkenntnis änderte Xerox die Assessment-Strategie bei der Einstellung von Callcenter-Mitarbeitern. Dies führte zu signifikant weniger Fluktuationen und damit zu erheblichen Kosteneinsparungen.

Die Ergebnisse solcher Analysen «triggern» ihrerseits wieder neue Fragen oder zwingen einen dazu, die bestehenden Fragen zu ändern oder ganz zu verwerfen.

Schätze heben. Um den geschilderten Kreislauf Erfolg versprechend in eine Organisation zu integrieren, muss man vor allem auf zwei Dinge achten. Erstens ist es eminent wichtig, dass ein heterogen zusammengestelltes Team eine gemeinsame Sprache ausbilden kann. Dies braucht Zeit und verlangt von jedem Teammitglied ein hohes Mass an Reflexion. Zweitens muss der geschilderte Zyklus möglichst schnell durchlaufen werden. Es ist normal, dass die ersten Ansätze, erste Fragestellungen und Analysemethoden nicht zum gewünschten Ziel führen. Es ist also das Credo gefragt: «develop explorative, fast, and fail early».

Big Data ist eine Opportunität für alle Unternehmen, die gewillt sind, ihren Daten eine Chance zu geben. Wer sich mit seinen Daten ernsthaft auseinandersetzt, wird schnell einsehen, dass es wirklich «Schätze» gibt, die nur darauf warten, gehoben zu werden. Ob nun das «Big» gerechtfertigt ist oder nicht, ist hierfür absolut irrelevant.

LITERATUR:

- Hook, Ernest B., and Regal, Ronald R. «Conceptus viability, malformation, and suspect mutagens or teratogens in humans. The Yule–Simpson paradox and implications for inferences of causality in studies of mutagenicity or teratogenicity limited to human livebirths.» Teratology 43.1 (1991): 53–59.
- Mayer-Schönberger, Viktor, and Cukier, Kenneth. Big Data. Computer Press, 2014.
- Provost, Foster, and Fawcett, Tom. Data Science for Business: What you need to know about data mining and data-analytic thinking. «O'Reilly Media, Inc.», 2013.
- HR and Big Data: It's a Union With Limitless Possibilities, http://www.tlnt.com/2013/01/21/hr-and-big-data-its-a-union-with-limitless-possibilities/

*** Dr. Marcel Blattner** hat Physik und Mathematik an der Universität Zürich studiert und ist Head Research des Laboratory for Web Science an der Fernfachhochschule Schweiz. Ab 2015 wird er bei Tamedia Digital mithelfen, ein Data-Science-Team aufzubauen.