

Talente finden und fördern

Segen oder Fluch? Big Data und die Auswahl von Mitarbeitern

Kann die Personalauswahl mithilfe von Big Data und ausgefeilten Algorithmen verbessert werden? Und was taugen die bisherigen Ansätze? Eine Bestandsaufnahme.

Glaukt man Anbietern von HR-Software, wird die Personalauswahl künftig nicht nur einfacher und schneller, sondern auch besser. Schließlich ist das Verbesserungspotenzial enorm. Denn noch immer setzen viele Personalmanager bei der Bewerberauswahl auf ihre Erfahrung und ihr Bauchgefühl (zusammenfassend: Kanning, 2015a), orientieren sich an Tippfehlern oder Lücken im Lebenslauf, führen unstrukturierte Einstellungsinterviews und unterliegen bei ihrer Urteilsbildung zahlreichen systemischen Fehlern. So werden beispielsweise gut aussehende Menschen oft überschätzt. Hier könnte der Computer ein Stück weit Abhilfe schaffen und Entscheidungsprozesse zumindest teilweise objektivieren. Dass hieraus am Ende eine bessere Auswahlentscheidung resultiert, setzt allerdings voraus, dass die erhobenen Daten auch tatsächlich aussagekräftig im Hinblick auf die zu besetzende Stelle sind.

Valide Informationen: Mangelware

Bislang haben sich vor allem zwei Methoden etabliert: computergestützte Testverfahren sowie Online-Bewerbungsformulare. Bei *computergestützten Testverfahren* kann es sich um Leistungstests oder Fragebogen zur Selbsteinschätzung handeln. *Online-Bewerbungsformulare* ermöglichen eine regelgeleitete Vorauswahl. Die Bewerber füllen einen kurzen Fragebogen aus, in dem nur valide und für die Auswahlentscheidung tatsächlich relevante Informationen erfasst werden (zum Beispiel Abiturnote, Studienfach, Examensnote, Auslandserfahrung). Die Vorauswahl trifft dann der Computer nach zuvor definierten Entscheidungsregeln. Studien, die sich damit beschäftigen, inwieweit deutsche Unternehmen diese Prinzipien umsetzen, kommen zu ernüchternden Ergebnissen. So werden Bewerbungsformulare vor allem dazu eingesetzt, die Unterlagen zu digitalisieren. Nach wie vor werden immer noch kaum valide Informationen (wie Anschreiben oder Selbstcharakterisierungen) angefordert und subjektiv durch Mitarbeiter der Personalabteilung bewertet (Kanning, 2015b; Kanning & Colpan, 2016).

Big Data in der Personalauswahl

In den vergangenen Jahren sind verschiedene Ansätze entstanden, die sich unter dem Schlagwort „Big Data“ zusammenfassen lassen. Big Data bezeichnet zunächst nur eine besonders große Ansammlung von Daten, die so gewaltig ist, dass man sie als Mensch nicht mehr ökonomisch sinnvoll oder überhaupt nicht mehr verarbeiten kann. Es geht also darum, mithilfe von Algorithmen eine mathematisch optimale Aggregation der Daten zu erreichen. Ein bestimmtes Validitätskriterium – in unserem Fall die berufliche Leistung an einem Arbeitsplatz – soll durch eine optimale Kombination der Bewerberdaten möglichst gut vorhergesagt werden. Die Suche nach dem richtigen Algorithmus ist dabei nichts anderes als die Suche nach einer mathematischen Formel, in die einzelne Daten, zum Beispiel gewichtet oder einander kompensierend, einfließen. Im Idealfall sollte die Prognose der beruflichen Leistung so besser gelingen als auf herkömmlichem Weg. Die Qualität der späteren Personalauswahl-Entscheidung hängt dabei von mehreren Faktoren ab:

- **Beziehung der Daten zur Person des Bewerbers**
Die erhobenen Daten müssen in einem möglichst unverfälschten und zeitlich stabilen Zusammenhang zu Eigenschaften des Bewerbers stehen. Die Information „Herr Meier trägt am 15. Januar einen dunklen Pullover“ wäre irrelevant.
- **Validität der Einzelinformationen** Die Informationen, die später in die Berechnung einfließen, müssen in einem Zusammenhang zur beruflichen Leistung stehen. Das Ausmaß der Validität der Einzelinformati-

Prof. Dr. Uwe Peter Kanning,
Diplom-Psychologe,
Professor für Wirtschaftspsychologie
an der Hochschule Osnabrück,
u.kanning@hs-osnabrueck.de



Talente finden und fördern

onen ist dabei wichtiger als ihre bloße Menge. Zumindest deutet die bisherige Forschung darauf hin: Ein einfacher Intelligenztest kann die berufliche Leistung deutlich besser vorhersagen als ein klassisches Vorstellungsgespräch, obwohl Letzteres vielfältigere Daten erfasst.

- **Relative Überschneidungsfreiheit der Daten** Würden alle validen Einzelinformationen genau denselben Aspekt der beruflichen Leistung vorhersagen, so wären sie völlig redundant. Statt 20 Informationen würde dann lediglich eine einzige ausreichen. Je weniger stark die validen Einzelinformationen einander überschneiden, desto größer ist die Wahrscheinlichkeit, dass sie gemeinsam eine bessere Prognose der beruflichen Leistung ermöglichen.

- **Definition der beruflichen Leistung** Im Kern geht es in der Personalauswahl darum, die spätere Leistung zu prognostizieren. Es ist daher wichtig, dass die berufliche Leistung am Arbeitsplatz bei der Entwicklung der Algorithmen auch möglichst gut erfasst wird. Ein Vorgesetztenurteil wäre ein einfacher und auch üblicher Weg. Besser wäre es jedoch, wenn weitere Indikatoren guter Leistung in die Definition einfließen würden (zum Beispiel Kennzahlen zur Produktivität, Arbeitszufriedenheit, Kundenzufriedenheit). Je umfassender die Operationalisierung des Leistungskriteriums ausfällt, desto praktisch wertvoller wird dann später die Aussage des Auswahlverfahrens sein.

- **Qualität der Algorithmen** Der ganze Ansatz steht und fällt mit der Frage, inwieweit es den Entwicklern tatsächlich gelingt, eine gute Aggregation der Einzeldaten zu finden. Dies ist keine banale Frage. Studien wie die der Informatikerin Aylin Caliskan haben beispielsweise gezeigt, dass Algorithmen manchmal dieselben Stereotype reproduzieren, wie sie auch in den Köpfen der Menschen existieren.

Computergestützte Analyse von Internetdaten

Die größte öffentliche Aufmerksamkeit findet Big Data, wenn es um die Analyse von Daten geht, die jeder von uns im Internet hinterlässt. Inzwischen beschäftigen sich mehrere Studien mit der Frage, inwieweit aus den Daten sozialer Netzwerke Rückschlüsse auf grundlegende Persönlichkeitsmerkmale möglich sind. Hierzu bittet man Probanden zunächst, einen Persönlichkeitsfragebogen auszufüllen, und erfasst anschließend Internetdaten auf zwei unterschiedlichen Wegen. Entweder geben die Probanden in einem weiteren Fragebogen Auskunft über ihre Internetdaten, wie etwa die Anzahl der Freunde bei Facebook, oder sie erklären sich damit einverstanden, dass die Forscher entsprechende Daten selbst erfassen. Zudem unterscheiden sich die Studien dahingehend, wie viele Internetinformationen sie berücksichtigen. Anschließend werden die Internetdaten zu den Ergebnissen des Persönlichkeitsfragebogens in Beziehung gesetzt. Die Tabelle unten gibt einen Überblick über die Befunde. Zwei Dinge fallen hierbei auf: Zum einen steigt die Aussagekraft an, wenn gleichzeitig mehrere Informationen berücksichtigt werden. Zum anderen sind die Zusammenhänge insgesamt gering ausgeprägt.

Doch es gibt noch ein sehr viel grundlegenderes Missverständnis. In der Personalauswahl geht es im Kern nicht darum,

Persönlichkeitsmerkmale zu erfassen, sondern die berufliche Leistung zu prognostizieren. Dies kann unter anderem vermittelt über Persönlichkeitsmerkmale laufen, sofern die Merkmale tatsächlich in der Lage sind, beruflichen Erfolg zu prognostizieren. Werfen wir in diesem Zusammenhang einen Blick in eine klassische Metaanalyse (Barrick, Mount & Judge, 2001), so fällt auf, dass die Prognosegüte grundlegender Persönlichkeitsmerkmale über verschiedene Berufe hinweg sehr gering ausfällt. Den höchsten Wert erzielt die Gewissenhaftigkeit mit etwa 5,8 Prozent. Verbinden wir beide Ergebnisse miteinander, so ergibt sich ein sehr ernüchterndes Fazit (siehe Tabelle unten). Über Big Data lässt sich die Gewissenhaftigkeit bestenfalls zu 12,3 Prozent erklären. Die Gewissenhaftigkeit selbst ist aber nur in der Lage, die berufliche Leistung zu 5,8 Prozent vorherzusagen. Dies wiederum bedeutet, dass sich die berufliche Leistung über Big Data im Schnitt zu weniger als einem Prozent vorhersagen lässt. Dies ist kaum besser als ein Münzwurf, der bei null Prozent liegt.

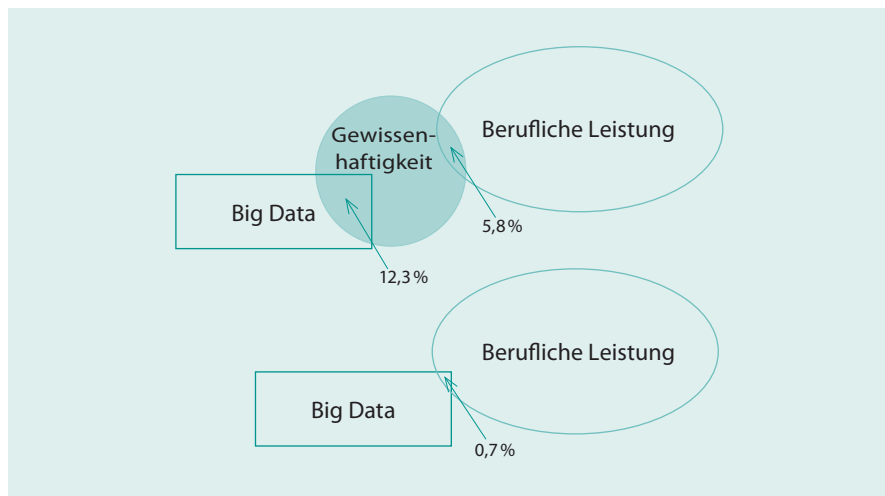
Der Denkfehler oder Trick einiger Anbieter von Software zur Personalauswahl liegt darin, dass sie nicht das eigentlich entscheidende Erfolgskriterium der beruflichen Leistung untersuchen, sondern Persönlichkeitsmerkmale. Damit produzieren sie Kennzahlen, die den tatsächlichen Nutzen ihrer Software positiv verzerrt darstellen. Den meisten Kunden

Zusammenhang Internetdaten und Ausprägung von Persönlichkeitsmerkmalen in Prozent

Studie	Persönlichkeitsmerkmal				
	Neurotizismus	Extraversion	Offenheit	Gewissenhaftigkeit	Verträglichkeit
Einzelne Kriterien aus dem Internet					
Amichai-Hamburger et al. (2010)	0	4,4–5,8	0–3,2	0–1,2	0
Schwartz et al. (2013)	0–2,3	0–1,7	0–1,7	0–1,7	0–3,6
Roulin (2014)	–	1,2	0	0	0
Marshall et al. (2015)	0–3,2	0–2	0–8,4	0–5,3	0
Kombination von mehreren Kriterien aus dem Internet					
Schwartz et al. (2013)	9,6	14,4	17,6	12,3	9,6

Kanning

Talente finden und fördern



Zusammenhänge zwischen Internetdaten, Persönlichkeit und Berufserfolg

wird dies überhaupt nicht auffallen. Würden sie komplett auf Big Data verzichten und ihren Bewerbern stattdessen einen (gegebenenfalls computergestützten) Fragebogen vorlegen, der die Gewissenhaftigkeit direkt erfasst, so würde sich die Aussagekraft ihrer Prognose gegenüber der Big-Data-Analyse um ein Vielfaches verbessern. Der Einsatz eines solchen Fragebogens wäre vermutlich auch deutlich günstiger. Es bliebe jedoch das Problem der insgesamt geringen Prognosegüte. Es müssten in jedem Fall zusätzlich deutlich validere Methoden zum Einsatz kommen, um eine gute Stellenbesetzung realisieren zu können.

Beispiel: Computergestützte Sprachanalyse

Seit einigen Jahren wird in Deutschland ein Verfahren angeboten, dessen Anbieter verspricht, Bewerber mithilfe einer computergestützten Sprachanalyse beurteilen zu können. Dazu müssen sie in einem Telefoninterview etwa 15 Minuten lang Fragen beantworten, die keinen direkten Bezug zur ausgeschriebenen Stelle haben. Sie erzählen beispielsweise, wie bei ihnen ein typischer Sonntag abläuft und wie sie Weihnachten verbracht haben. Für den Anbieter geht es lediglich darum, irgendeine zeitlich hinreichend lange Sprachprobe zu erzeugen, egal, worum es inhaltlich geht. Denn in der Sprachpro-

be soll sich die Persönlichkeit eines Menschen spiegeln.

Anschließend wird die Sprachprobe mithilfe einer Software und geheimer Algorithmen analysiert und ein Gutachten erstellt, das bei so manchem den Anschein großer Wissenschaftlichkeit erwecken dürfte. Da ist von Basislinguistik, prosodischen Parametern, Perturbation, deiktischer Verteilung und Harmonizität die Rede. Zudem werden viele Zahlenwerte präsentiert, etwa zur Tonlage in Hertz, zum Stimmumfang in Halbtonschritten oder zur durchschnittlichen Wortlänge. Am Ende steht eine Diagnostik der Persönlichkeit. Man erfährt etwas über Neugier, Dominanz, Bescheidenheit, Autonomie, Kontaktfreude, Risikobereitschaft, die Big Five (Neurotizismus, Extraversion, Offenheit, Gewissenhaftigkeit, Verträglichkeit), die populären Kommunikationsstile nach Schulz von Thun, das Ausmaß der Beanspruchung und vieles mehr. Die Bandbreite ist so groß, dass sie bei fachkundigen Lesern schon wieder Skepsis hervorrufen müsste. Die Darstellung der Ausprägung der einzelnen Persönlichkeitsmerkmale ist kunterbunt. Da gibt es fünfstufige und zehnstufige Skalen. Manche sind unipolar, andere wiederum bipolar. Zudem wird eine Art Normierung vorgenommen, indem die Werte der einzelnen Person mit dem Mittel einer Referenzstichprobe verglichen werden. Übliche Normwerte, wie

sie international Standard sind (zum Beispiel Prozentrangnormen, Standardwertnormen), findet man nicht. Und am Ende wird noch über das Gesamtprofil eine prozentuale Übereinstimmung mit dem Referenzprofil berechnet.

Die Sprachanalyse bringt alles mit, was sich Laien von einem Test erhoffen. Mit nur wenigen Minuten Aufwand erhält man auf wundersame Weise ein vielfältiges Persönlichkeitsprofil. Der Computer sorgt dafür, dass alles mit rechten Dingen zugeht. Zudem soll der Bewerber die Ergebnisse nicht bewusst zu seinem Vorteil manipulieren können. Es erstaunt nicht, dass ein solcher „Wundertest“ auch bei den Medien auf großes Interesse stößt. Von „Spiegel online“ über „FAZ“ und „Die Zeit“ bis zum ZDF „heute journal“ gibt es unzählige, völlig kritikfreie Berichte. Dabei sind die Schwächen des Ansatzes fundamental:

- Die physikalischen Eigenschaften der Sprache variieren mit den Inhalten, über die gesprochen wird. Insofern ist es nicht unerheblich, welche Fragen gestellt werden und ob zum Beispiel den Probanden in der Referenzstichprobe dieselben Fragen gestellt wurden. Offenbar variieren die Fragen aber von Bewerber zu Bewerber (= mangelnde Durchführungsobjektivität). Bewerber, die kurze Antworten geben, erhalten zusätzliche Fragen, damit man auf eine Mindestredezeit kommt.
- Die physikalischen Eigenschaften der Sprache variieren mit den aktuellen Emotionen des Sprechers. Für viele Bewerber dürfte das Interview eine Situation der Anspannung sein, die zu einem nicht-repräsentativen Spracheindruck führt (= mangelnde Reliabilität).
- Dadurch wird die Vergleichbarkeit zur Normierungsstichprobe reduziert, da nicht davon auszugehen ist, dass hier eine ähnliche Anspannung vorlag.
- Die Erfassung der physikalischen Eigenschaften der Sprache wird durch Rahmenbedingungen beeinflusst, die zu

Talente finden und fördern

einer verzerrten Erhebung und damit zu einer falschen Datenbasis führt (= mangelnde Durchführungsobjektivität). Hierzu zählen zum Beispiel das verwendete Telefon, die Qualität der Übertragung per Handy oder die Uhrzeit, die Einfluss auf die Müdigkeit und damit auf die Stimmlage nehmen kann.

- Die Sprache kann willentlich beeinflusst werden. Fast jeder weiß, wie er mit seiner Stimme besonders warmherzig, charmant oder selbstsicher wirken kann.
- Die zeitliche Stabilität der Befunde: Angesichts der zuvor genannten Probleme ist zu erwarten, dass die Ergebnisse der Untersuchung zu verschiedenen Zeitpunkten unterschiedlich ausfallen (= mangelnde Retest-Reliabilität).
- Die Menge der abgeleiteten Persönlichkeitsmerkmale lässt erwarten, dass untereinander starke Korrelationen vorliegen (= mangelnde faktorielle Validität). Offenbar wird hier der Versuch unternommen, möglichst viele Dimensionen abzubilden, um eine breite Vermarktung zu gewährleisten. Die Abhängigkeit der vielen Merkmale wird folgerichtig nicht thematisiert, obwohl dies in einem seriösen Test selbstverständlich wäre.
- Es liegen keine Befunde vor, die dokumentieren, ob – und wenn ja mit welchem Ergebnis – Validierungsstudien durchgeführt wurden. Besonders wichtig wären Hinweise zur Höhe der Korrelationen. Reden wir beispielsweise über einen Zusammenhang zur Gewissenhaftigkeit von zwei Prozent oder von 50 Prozent? Ersteres dürfte wahrscheinlicher sein als Letzteres (= fragwürdige innere kriterienbezogene Validität).
- Selbst wenn sich nennenswerte Zusammenhänge zu Persönlichkeitsmerkmalen finden ließen, fehlen Studien, die einen Zusammenhang zu Kriterien beruflicher Leistung aufzeigen (= fragwürdige äußere kriterienbezogene Validität). Hier finden wir dasselbe Problem wie bei der Analyse von Internetdaten.
- Zwar gibt es für Unternehmen, die die Sprachanalyse einsetzen, die Möglich-

keit eines direkten Vergleichs zwischen dem Sprachprofil eines Bewerbers und dem seiner erfolgreichen Mitarbeiter. Doch dieser Vergleich hilft nicht weiter, da unbekannt ist, ob die gemessenen Eigenschaften der erfolgreichen Mitarbeiter in einem kausalen Zusammenhang zu ihrer Leistung stehen. So könnten erfolgreiche Mitarbeiter zum Beispiel Schuhgröße 43 haben. Dennoch wäre es nicht sinnvoll, Bewerber mit derselben Schuhgröße zu bevorzugen, denn wir haben es hier mit einer illusorischen Korrelation zu tun.

- Es liegen keine Informationen zu den verwendeten Algorithmen vor. Man erwartet von den Kunden, dass sie dem Anbieter einfach glauben.
- Es ist keine unabhängige Forschung zu den Angaben des Anbieters bekannt.

Dennoch setzen einige Unternehmen bei ihrer Personaldiagnostik auf diese Software. Wer sich dabei an klassische Pseudowissenschaften wie etwa Grafologie oder Physiognomie erinnert fühlt, liegt nicht ganz falsch. Auch hier gibt es das Versprechen, man würde einen direkten und unverfälschten Zugang zur Persönlichkeit eines Menschen schaffen, indem verschlüsselte Botschaften in der Handschrift oder im Gesicht des Betroffenen dechiffriert werden. Dabei berufen sich die Anbieter gern auf ein jahrhundertealtes Erfahrungswissen. Bei der Sprachanalyse ist es der geheimnisvolle Algorithmus.

Fazit

Computertechnologie hat durchaus das Potenzial, die Praxis der Personalauswahl zu verbessern. Allerdings drängen derzeit Technologien auf den Markt, die beim Laien mit Begriffen wie „Big Data“ und „Algorithmen“ den Anschein hochkomplexer Wissenschaftlichkeit erzeugen, dabei aber auf eine Blackbox setzen und ihre Algorithmen gegenüber den Kunden nicht offenlegen. Unter dem Deckmantel des Schutzes ihrer Produkte verhindern sie unabhängige Forschung und eine

kritische Prüfung der Qualität. Der Kunde kauft die sprichwörtliche Katze im Sack.

Weitere Literatur

Amichai-Hamburger, Y. & Vinitzky, G. (2010). Social Network Use and personality. *Computers in Human Behavior*, 26 (6), 1289–1295.

Barrick, M. R., Mount, M. K. & Judge, T. A. (2001). Personality and Performance at the Beginning of the New Millennium: What Do We Know and Where Do We Go Next? *International Journal of Selection and Assessment*, 9, 9–30.

Caliskan, A., Bryson, J. J. & Narayanan, A. (2017). Semantics Derived Automatically from Language Corpora Contain Human-Like Biases. *Science*, 356 (6334), 183–186.

Kanning, U. P. (2015a). *Personalauswahl zwischen Anspruch und Wirklichkeit – Eine wirtschaftspsychologische Analyse*. Berlin: Springer.

Kanning, U. P. (2015b). E-Recruitment – Chancen und Realität der Bewerbervorauswahl per Internet. *Personalführung*, 5, 61–65.

Kanning, U. P. & Colpan, C. (2016). Alter Wein in neuen Schläuchen: Einsatz von Online-Bewerbungsformularen in deutschen Unternehmen. *Human Resources Manager*, 6, 76–77.

Marshall, T. C., Lefringhausen, K. & Ferenczi, N. (2015). The Big Five, Self-Esteem, and Narcissism as Predictors of the Topics People Write About in Facebook Status Updates. *Personality and Individual Differences*, 85, 35–40.

Roulin, N. (2014). The Influence of Employer's Use of Social Networking Websites in Selection, Online Self-Promotion, and Personality on the Likelihood of Faux Pas Postings. *International Journal of Selection and Assessment*, 22, 80–87.

Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Ramones, S. M., Agrawal, M., Shah, A., Kosinski, M., Stillwell, D., Seligman, M. E. P. & Ungar, L. H. (2013). Personality, Gender, and Age in the Language of Social Media: The Open-Vocabulary Approach. *PLoS ONE*, 8 (9).