Big Data und Data-Science-Ansätze in der öffentlichen Verwaltung

Ines Mergel

Universität Konstanz

Big Data und Data-Science-Ansätze finden Einzug in die öffentliche Verwaltung. Dieses Kapitel bietet zunächst eine Definition von Big Data in der öffentlichen Verwaltung an und leitet die unterschiedlichen Datenquellen für historische, Echtzeit- und prädiktive Big-Data-Analysen ab. Danach werden Beispiele für organisationale Einheiten in der öffentlichen Verwaltung erläutert, die Big-Data-Analysen durchführen. Anhand der folgenden drei ausgewählten Beispiele wird das Potenzial von Big Data aufgezeigt: USGS »Did you feel it?«-Twitter-Karten, prädiktive Analysen in Finanzbehörden und Vorhersagen von Grippewellen mit Hilfe von Google Flu Trends. Aus diesen und weiteren Beispielen werden dann die Herausforderungen für die Verwendung von Big Data und Data-Science-Ansätzen in der öffentlichen Verwaltung erläutert sowie offene Forschungsfragen für die Verwaltungswissenschaft abgeleitet.

1. Einleitung

Big Data entsteht dann, wenn zum Beispiel Millionen von Datenpunkten durch Online-Interaktionen auf Social-Media-Plattformen kreiert werden und mit anderen Datensätzen verbunden werden. Zusätzlich entstehen durch die Sharing Economy, die Nutzung von Internet-Plattformen riesige Mengen von Daten, die kontinuierlich und in nahezu Echtzeit ausgewertet werden können.

Die meisten Publikationen, die den Begriff verwenden, berichten positiv über die Chancen, die große Datenmengen zu bieten scheinen. So veröffentlichte der Economist kürzlich einen Artikel, in dem der ökonomische Wert von Big Data herausgehoben wurde: »Daten sind für dieses Jahrhundert das, was Öl für das letzte Jahrhundert war: ein Motor für Wachstum und Wandel«[1]. Forbes gibt an, dass Big Data das sind, was intelligente Städte antreibt: die Kontrolle über die Menge der Daten und damit das Erkennen von Bedürfnisse der Bürger in Echtzeit.[2] Andere Autoren verwerfen den Begriff Big Data als ein Konstrukt, das von der Industrie entwickelt wurde, um neue Geschäftsfelder zu erschließen, jedoch oftmals zu irreführenden Anwendungen von Algorithmen führt, die nicht für riesige Datenmengen entwickelt wurden.[3]

Im Folgenden werden deshalb zunächst der Begriff Big Data geklärt, dessen unterschiedliche Anwendungen in der öffentlichen Verwaltung dargestellt und die bestehenden Herausforderungen aufgezeigt.

2. Was ist Big Data?

Big Data wird derzeit als Oberbegriff verwendet, um verschiedene Arten von Daten und Aspekte datenintensiver Ansätze zu beschreiben.[4] Deshalb ist es sinnvoll, die verwendeten Datenbegriffe zunächst zu definieren und miteinander zu vergleichen.

Administrativ gesammelte Daten entstehen durch Datensammlungen, die entweder durch reguläre Verwaltungsakte erzeugt werden, wenn die Bürger eine Dienstleistung erhalten, oder wenn andere

1 n. a., 2017

2 Newman, 2016

3 Nijhus, 2017

4 Mergel, Rethemeyer et al., 2016; Mergel, Rethemeyer et al., 2016a

Formen von Transaktionen durchgeführt werden. Diese Datensammlungen sind zumeist systematisch strukturiert und werden von offiziellen Behörden oder Firmen gesammelt. Dadurch sind es oftmals hochwertige vorstrukturierte Datensätze, in denen genau definierte persönliche Daten, wie Alter, Geschlecht, Einkommen usw., enthalten sind. Sie entstehen in der öffentlichen Verwaltung, wenn Register erstellt werden oder wenn Volkszählungsdaten in bestimmten Zeitabständen erhoben werden. Administrative Daten werden in Form von Berichten der Öffentlichkeit in aggregierter Form (mit Zeitverzögerung) zur Verfügung gestellt. Ein Teil dieser administrativ gesammelten Daten wird dann auf offenen Datenplattformen oder anderen Regierungswebsites veröffentlicht und wird damit zu *Open Data*. Sie können wiederverwendet werden, da sie oftmals in maschinenlesbaren Formen zur Verfügung gestellt werden, und eignen sich damit für die Weiterverarbeitung in Form von Visualisierungen oder anderen Interpretationsformen. Im Unterschied zu anderen Arten von Datensätzen wurde über die Sammlung, Bereinigung, Kombination und andere Analyseschritte vorab entschieden und die Daten stehen der Öffentlichkeit oder anderen Behörden oftmals nur in aggregierten Formaten zur Verfügung.

Benutzer- oder bürgergenerierte Daten sind Daten, die außerhalb der öffentlichen Verwaltung von Bürgern erstellt werden, die sowohl mit Online-Inhalten oder auch miteinander interagieren, um einen Wert für sich selbst zu schaffen. Beispiele hierfür sind: Amazon Mechanical Turks Click Worker, Online-Kreditvergabeseiten, Crowdsourcing-Plattformen, wie ThreadLess, aber auch Social-Media-Feeds, wie Twitter, Facebook, YouTube, Weblogs, Clickstreams, Online-Suchdaten, oder Daten aus Online-Verkaufstransaktionen (wie *Amazon Sales).* Abhängig von den Benutzereinstellungen der Webseiten, aber auch dem individuellen Nutzerverhalten sind diese Daten entweder öffentlicher oder privater Natur.

Automatisch generierte Daten sind Daten von menschlichen und physikalischen Sensoren, die z.B. an Gebäuden angebracht sind und den Personenverkehr aufnehmen, in Form von Dashcams an Fahrzeugen Aufzeichnungen machen, durch Handysignalen entstehen oder aber mit Hilfe von polizeilichen Bodycams gesammelt werden. Jede Bewegung wird automatisch erfasst und es entstehen umfassende kontinuierlich erweiterbare Datensätze, die sich auf ganze Populationen anstatt nur auf gezielte Stichproben beziehen. Dadurch können beispielsweise alle Interaktionen von Nutzern einer App innerhalb einer Stadt, Landes oder sogar weltweit mit Hilfe ihrer Handydaten nachvollzogen werden. Als Resultat entstehen umfassende Datensätze von ganzen Populationen, die automatisch und oft ohne Wissen der Benutzer und Bürger generiert werden. Die Daten werden angereichert durch zusätzliche Metadaten, wie z. B. Geolokalisierung, die durch Wetter-Applikationen auch dann gespeichert wird, wenn der Nutzer nicht aktiv nach seinen lokalen Wetterbedingungen sucht.

Big Data umfasst somit große, komplexe, unstrukturierte Datenmengen, die zu groß sind, um herkömmliche Tools zur Erfassung und Analyse zu verwenden.[5] Dabei handelt es sich nicht um eine einzige Datenbank, sondern Daten, die aus den folgenden Quellen gesammelt werden: Unstrukturierte Internetquellen, wie z. B. Social-Media-Interaktionen, Handy-Apps, Videos, geteilte Bilder, menschliche und physische Sensoren oder Online-Suchverhalten, Online-Verkäufe von Internet-Shops, oder auch Telefonverbindungen in Mobilfunknetzen.

[5] Cox & Ellsworth, 1997

3. Big Data-Analyseformen

Eine *historische Analyse* mit Hilfe von administrativ designten und gesammelten Daten enthält historische Daten, die gesammelt, bereinigt und dann zeitverzögert analysiert werden. Über diese Datensätze wurden in der öffentlichen Verwaltung Entscheidungen getroffen, d.h. es sind keine Rohdaten mehr, sondern Daten, die bereits bereinigt wurden, und die Analyse bezieht sich auf vergangene Entwicklungen. Dazu gehören z.B. alle Verwaltungsakte, Transaktionen mit Dritten oder auch Census-Daten. Historische Analysen koennen sowohl zu Trendanalysen genutzt werden, aber auch um den zukünftigen Ressourcenbedarf des öffentlichen Sektors oder einer einzelnen Behörde zu ermitteln.

Unstrukturierte Interaktionen von Bürgern und Organisationen im Internet über soziale Medien, z. B. Verbreitung von Meinungen, *Fake News*, abgeleitetes Abstimmungsverhalten aufgrund von Such- und Lesepräferenzen, Verbindungen auf Social-Networking-Websites, individuelle strukturelle Positionen (wer mit wem wie verbunden ist) oder Inhalte von Beziehungen (Stimmung, politische Meinungen etc.) können dazu genutzt werden, um *Echtzeitanalysen* zu wirtschaftlichen Online-Aktivitäten durchzuführen. Diese Analysen sind für (fast) Echtzeit-Einblicke in die aktuellen Präferenzen und Verhaltensweisen von Nutzern geeignet und für sogenanntes Nowcasting - Vorhersagen der Gegenwart – verwendbar.[6]

In Kombination können administrativ gestaltete Datensätze zusammen mit unstrukturierten, automatisch generierten und kontinuierlich einfließenden Daten verwendet werden, um genauere und

[6] Banbura, Giannone et al., 2013

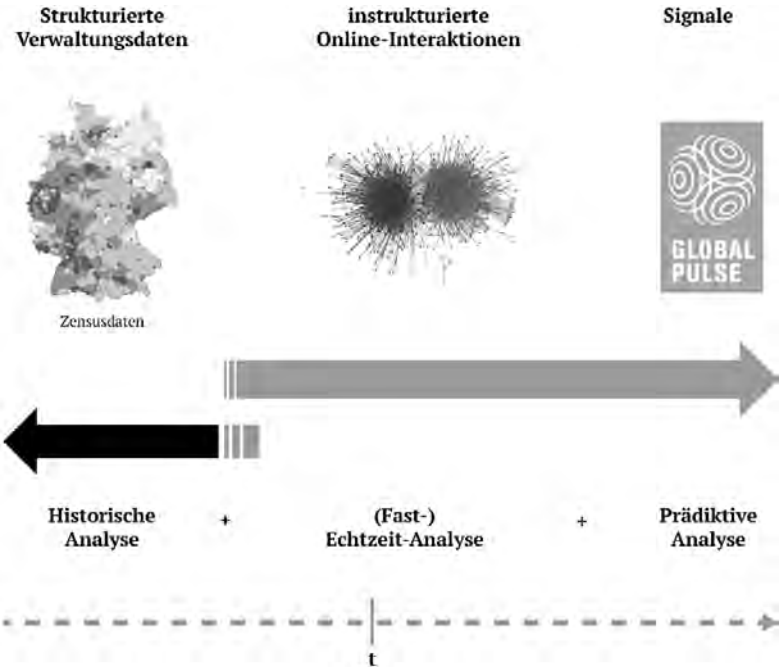

Abbildung 1: Big-Data-Analysen

schnellere Vorhersagen über Verhaltensänderungen und -präferenzen zu treffen.[7] Diese Daten werden als Signale genutzt, mit denen sich zum Beispiel die Ausbreitung von Krankheiten, menschliche Mobilität und wirtschaftliche Entwicklungen modellieren lassen. Historische Daten werden herangezogen, um Muster zu identifizieren und mathemaische Modelle zu erstellen, wodurch Trends erkennbar werden können. Diese prädikativen Modelle (können auf aktuelle Daten angewendet werden und) erlauben

Wahrscheinlichkeitsaussagen über zukünftige Entwicklungen. Damit kann die öffentliche Verwaltung beispielsweis Flüchtlings-

[7] George, Haas et al., 2014

ströme voraussagen, sie werden z.B. durch die United Nations in ihrem Global Pulse verwendet. Diese Form der Datenanalyse wird als predictive analytics – also *prädiktiven Analysen* - zur Ableitung von Vorhersagen verwendet und hilft dem öffentlichen Sektor zukünftige Bedürfnisse und Ressourceneinsätze zu planen.

Abbildung 1 zeigt, welche unterschiedlichen Formen von Big-Data-Analysen mit Hilfe der Datentypen vorgenommen werden können. Administrativ designte Datensätze eignen sich vor allem für historische Analysen, unstrukturierte Internetinteraktionen und -transaktionen für Echtzeitanalysen, und Signale für prädiktive Analyse.

4. Anwendungsbeispiele: Big Data in der öffentlichen Verwaltung

Big-Data-Analysen oder Data-Science-Vorgehensweisen sind mittlerweile in der öffentlichen Verwaltung angekommen. Data Scientists können jedoch nicht wie bisher nur traditionelle Statistiker oder Programmierer sein. Vielmehr müssen Verwaltungsfachkräfte ihr bisheriges inhaltliches Spezialwissen mit um Datenanalysefaehigkeiten erweitern.

In der öffentlichen Verwaltung gibt es aktuell unterschiedliche Organisationseinheiten, die Big Data-Analyseverfahren einsetzen:

- *Thematische Data-Science-Teams* in spezifischen Ministerien oder Behörden, wie z.B. im Finanzministerium, die beispielsweise Vorhersagen von Mehrwertsteuerbetrug und Mehrwertsteuer-Karussellbetrug durchführen können, indem sie Netzwerkanalysen mit Risikoindikatoren verbinden, um so prädiktive Analysen durchzuführen.
- *Social & Behavioral Insights Teams*, wie z. B. das vom britischen Cabinet Office eingesetzte Team, das sich mit der Messung der

Regierungsleistung beschäftigt, um aus diesen Einsichten die öffentliche Leistungserbringung effektiver gestalten zu können.

– *Digital Operation Center,* z. B. des Roten Kreuzes, verarbeiten eine Kombination aus Daten von menschlichen und physischen Sensoren in den unterschiedlichen Phasen des Katastrophenmanagements und nutzen dafür Nowcasting-Methoden.

– *Citizen Data Scientists* sind Laien und Nicht-Experten, die von der öffentlichen Verwaltung z. B. durch Hackathons eingebunden werden, um mithilfe von Predictive-Modelling-Methoden bei der Auswertung von Big Data zu helfen.

Abbildung 2: Ausgewählte Big-Data- und Data-Science-Organisationseinheiten im öffentlichen Sektor

Es gibt mittlerweile viele verschiedene Beispiele für Big-Data-Analysen in benachbarten Forschungsdisziplinen:

– *Politikwissenschaft*: Vorhersage des Wählerverhaltens basierend auf Twitter-Interaktionen.[8]

[8] Mislove, Lehmann et al., 2011, Gayo-Avello, 2012

– *Public Health*: Die Auswertung von Google-Flu-Trends-Algorithmen zur Vorhersage von Grippeausbrüchen ist fehlgeschlagen, weil der Algorithmus den Kontext nicht berücksichtigt hat und Google zugab, dass ihre Algorithmen ungeeignet sind.[9] Google analysiert darüber hinaus die Krankheitssymptome von Millionen von Nutzern und sagt damit den potenziellen Krankheitsverlauf voraus.[10]

– *Notfall- und Disastermanagement*: Zusammenführung offizieller wissenschaftlicher Sensordaten zur Abschätzung der Auswirkungen von Erdbeben mit Twitter-Daten der Bürger, die die tatsächlichen Auswirkungen in ihrer Region wahrnehmen und berichten. Als Beispiel dient hier die Analyse der Twitter-Daten und deren Vergleich mit seismologischen Messungen während eines Erdbebens durch den U.S. Geological Service.[11]

Im Folgenden werden drei Beispiele von Big-Data-Analysen im Detail besprochen, die auch für Verwaltungswissenschaftler und Verwaltungspraktiker von hoher Relevanz sein können.

4.1. Beispiel 1: »Did you Feel It?«-Twitter-Karten

Der United States Geological Service (USGS) war in den U.S.A. eine der ersten Verwaltungseinheiten, die große, nicht-wissenschaftlich erhobene, öffentlich zugängliche Datenmengen genutzt hat:[12] Auf

[9] Google.org, 2008, Lazer, Kennedy et al., 2014, Raghupathi & Raghupathi, 2014

[10] Rajkomar, Oren et al., 2018

[11] USGS, 2015.

[12] Robbins, Simonsen et al., 2008; USGS, 2015.

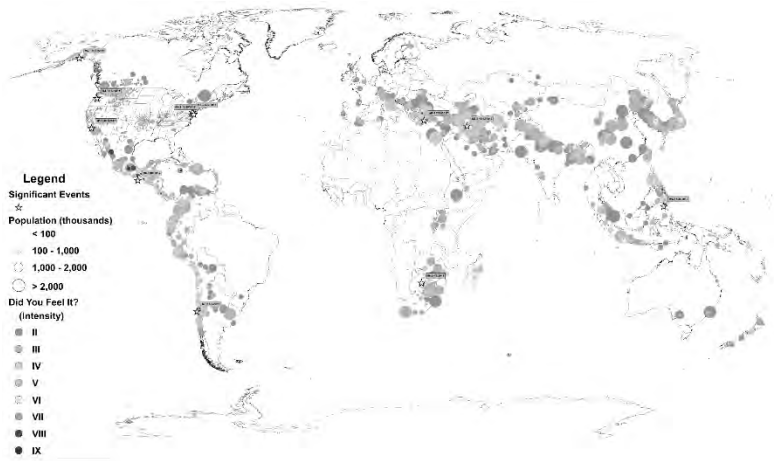

Abbildung 3: USGS »Did you feel it?«-Karte (https://earthquake.usgs.gov/data/dyfi)

den »Did you feel it?«-Twitter-Karten kartografierte die Behörde die mit Hilfe von geo-markierten Tweets von Bürgern gesammelten Updates, die während eines Erdbebens die gespürten Auswirkungen auf Twitter teilten und damit auch ermöglichten, Schäden während der Katastrophe zu dokumentieren. Auf den Twitter-Karten werden dann die gefühlten Auswirkungen mit wissenschaftlich gesammelten seismographischen Daten des Erdbebens kombiniert. Die Twitter-Nutzern sortieren in selbstreflektierenden kollektiven Prozessen, sogenannte »social milling«-Prozesse die Fake News aus den Fakten aus, um daraus die tatsächlichen Schäden aufzudecken und ein realistisches Lagebild zu erschließen.

Herkömmliche Seismometer benötigen normalerweise 2 bis 20 Minuten um mit einer Schnelligkeit von 3–5 km pro Sekunde seismische Daten zu verbreitet und einen Alarm auszulösen. Twitter-Daten werden dagegen über Glasfaserkabel mit 200.000 km pro Sekunde verbreitet. Die öffentliche Verwaltung verwendet diese Big-Data-Analysen, um während einer Katastrophenlage bessere und

schnellere Entscheidungen zu treffen und Ressourcen gezielter einsetzen zu können. Die folgende Grafik zeigt die Ausbreitung von Tweets, die entsprechend der gefühlten Heftigkeit der Erdbebenauswirkung farblich kodiert sind und kombiniert werden mit den Messungen der tatsächlichen geologischen Tätigkeiten:

4.2. Beispiel 2: Der Einsatz von prädiktiven Analysen in Steuer- und Finanzbehörden

Traditionell werden Daten der öffentlichen Verwaltung in Datenbanken gespeichert, die tief in Regierungsbehörden versteckt sind und nicht über Behördensilos hinweg mit anderen Abteilungen geteilt werden. Die hochstrukturierten Daten von Steuerzahlern oder Steuerberatern, die in vordefinierten Zeitabständen eingereicht werden, erlauben meist vor allem eine chronologische Analyse historischer Daten, leisten jedoch selten einen Beitrag, um prädiktive Analysen oder Echtzeitanalysen durchzuführen.

Mit Hilfe von prädiktiven Analysen nutzen Finanzministerien der OECD-Länder Daten von Rechnungen, Kontoauszügen, Zollerklärungen, Lieferantenrechnungen und Bankbelegen von Unternehmen und anderen im Internet generierten Daten von Online-Verkaufsplattformen ermöglichen.[13] Diese Daten werden mit Risikoindikatoren der einzelnen Wirtschaftsakteure einer sozialen Netzwerkanalyse zugeführt, sodass in nahezu Echtzeit Einblicke gewonnen werden können, wie das folgende Zitat eines Data Scientists in einer Finanzbehörde nahelegt:

> *»Wir senden unsere Informationen über riskante Objekte an die Revisionsabteilung. Wir suchen dann nach Hintergrundinformationen, um herauszufinden, ob es ein Risiko gibt oder*

[13] OECD – Advanced analytics for better tax administration: http://s.fhg.de/as4

nicht, oder ob die Indikatoren normal sind. Wir setzen uns mit der Person in Verbindung, um zu verstehen, warum es Anomalien gibt. Bleibt das Risiko bestehen, dann beginnen wir mit einigen Aktionen. Die Revisionsabteilung leitet dann eine Untersuchung ein: In 95 Prozent der Fälle bekommen wir das Geld in zwei Tagen zurück. Nicht nach einem Vierteljahr oder einem ganzen Jahr.«[14]

Die Analysen führen damit zu Einsichten über das Kundenverhalten und deren Präferenzen, tragen dem Datenaustausch zwischen den Finanzbehörden und anderen Behörden für Monitoring-Zwecke und Performance-Analysen bei. Sie führen oftmals zu einer Änderungen der Risikobewertung von wirtschaftlichen Akteuren.[15] Als Resultat konnte beispielsweise Estland die Erhebungsquote seiner Mehrwertsteuer auf 98 Prozent steigern, was einer Steigerung um 12 Prozent durch die Anwendung von prädiktiven Analyseverfahren entspricht.[16]

Andere Anwendungen im öffentliche Sektor beziehen sich auf die Kriminalitäts- und Korruptionsprävention, Rechnungs- und Konformitätsprüfungen in nahezu Echtzeit zur Erhöhung der Mehrwertsteuererhebung. Zukünftig können Big Data-Analysen auch zur Verbesserung der Entscheidungsfindung genutzt werden, z.B. Folgenabschätzung auf das Sozialsystem hervorzusagen, falls große Arbeitgeber in einer bestimmten Region schließen sollten. Damit entwickeln sich reaktive öffentliche Verwaltungen hin zu proaktiv handelnden Einheiten.

[14] Interview geführt 2017 mit Verantwortlichem für Big Data-Analyse im österreichisches Finanzministerium.

[15] Mergel, 2017

[16] Interview geführt 2017 mit Verantwortlichem für Big Data-Analyse im estnischen Finanzministerium.

Der öffentliche Sektor steht zurzeit vor der Herausforderung Big Data-Analysen in die Fachverfahren der Verwaltung zu integrieren, in die Fachprozesse miteinzubeziehen und nicht als unabhängige Data Science-Abteilungen ohne Fachverständnis aufzustellen. Der Leiter einer Data-Science-Gruppe in einem europäischen Land sagt: »*Wir müssen uns auf eine kontextspezifische Fallauswahl zubewegen, statt auf rein mathematische oder methodengetriebene Ansätze.*« Ein anderer fügt hinzu: »*Die Verantwortlichkeiten und administrativen Prozesse müssen angepasst werden, um die Datenstrategien widerzuspiegeln*.« So entsteht in der öffentlichen Verwaltung die Notwendigkeit für organisatorische Änderungen und kulturelle Veränderungen, wie zum Beispiel die Notwendigkeit für: »*Kulturellen Wandel hin zur Integration von Rechenmethoden und IT mit tiefem Verwaltungswissen*.«

4.3. Google Flu Trends

Google Flu Trends (GFT) wurde 2008 entwickelt, um Grippepandemien aus Google-Keyword-Suchen früher vorherzusagen als die offiziellen Arzt- und Seuchenberichte veröffentlicht werden.[17] Die anfängliche Datenerhebung wurde so konzipiert, dass große Datenmengen (50 Millionen Google-Suchanfragen) mit kleinen Datenmengen (ca. 1.500 offiziellen Meldungen von Ärzten, die Grippefälle an die Seuchenbehörden melden) kombiniert wurden. Im Jahr 2013 berichtete allerdings das Magazin Nature, dass GFT weit höhere Wahrscheinlichkeiten einer Grippeepidemie vorhersagte als durch die tatsächlichen an das Center for Disease Control and Prevention (CDC) eingereichten offiziellen Ärzteberichte.[18] Programmierer bei Google hatten die Entscheidung getroffen, bestimmte Suchbegriffe

[17] Google.org, 2008

[18] Lazer, Kennedy et al., 2014

und Korrelationen auszusortieren. Das Ergebnis: Google Flu Trends überschätzte den tatsächlichen Ausbruch der Grippe und die prädiktiven Analysen führten zu falschen Ergebnissen.

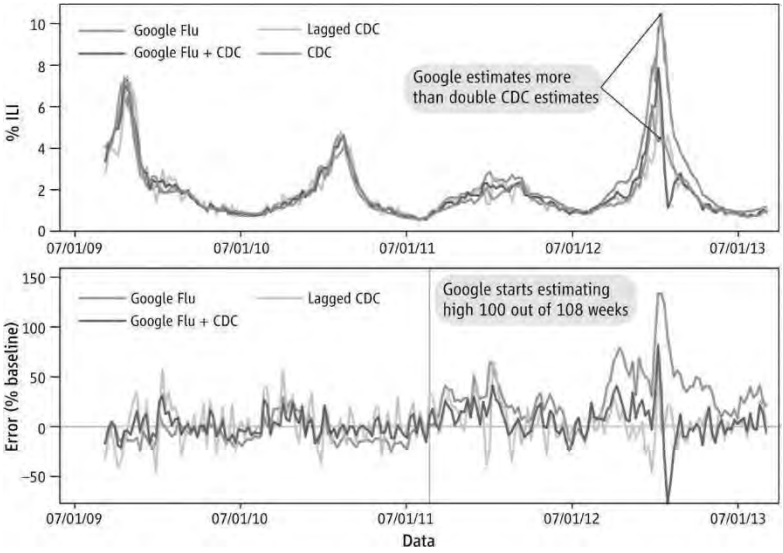

Abbildung 4: Google Flu Trends[19]

Da Google Flu Trend oft als das Paradebeispiel für die prädiktive Wirkungskraft von Big-Data-Analysen genutzt wird, um die Vorhersagekraft von großen Daten zu zeigen, hat Google seine Analysen auf falschen Algorithmen bezogenen Analysen zurückziehen müssen. Google Flu Trends wurde archiviert und Google arbeitet jetzt mit Forschern zusammen, um die Ergebnisse der Algorithmen zu verbessern.

[19] Lazer, Kennedy et al., 2014

5. Herausforderungen für die Verwaltung

Angesichts der allgegenwärtigen Verfügbarkeit von Big Data und des Versprechens, mit riesigen Datenmengen neue Erkenntnisse zu gewinnen, gibt es für die öffentliche Verwaltung und die Politik viele Herausforderungen.

Nutzung des sogenannten »digital exhaust«: Die schiere Masse des Sammel- und Reinigungsaufwandes von riesigen Datenmengen überfordert oftmals die traditionellen Statistikämter der öffentlichen Verwaltung. Die relativ leicht zugänglichen Daten über Bürger und die zusätzlich von ihnen in sozialen Medien produzierten Big Data führen oftmals zu der Versuchung, auf die Daten von Gesamtpopulationen zuzugreifen. Jedoch wird dabei häufig außer Acht gelassen, dass diese großen Datenmengen für technische und kommerzielle Zwecke gesammelt werden, wobei Verknüpfungen und Datenkonstrukte verwendet wurden, die möglicherweise nicht zuverlässig und für die Zwecke der öffentlichen Verwaltung ungeeignet sind.

Garbage in, garbage out? Auf die vor allem auf sozialen Medien entstehenden Big Data werden Algorithmen verwendet, die für wesentlich kleinere Datensätze entwickelt wurden. Aus diesem Grund funktionieren Algorithmen möglicherweise nicht wie erwartet, unabhängig davon, mit wie vielen Daten sie »gefüttert« werden.[20] Sie erlauben deshalb oftmals nur eine eingeschränkte Suche nach Mustern und dem Durchschnitt,[21] jedoch nicht nach Anomalien. Deshalb ist es notwendig Algorithmen anzupassen und zu testen da ansonsten das Dilemma von Goodhards Gesetz verstärkt wird: Oftmals kon-

[20] Boyd & Crawford, 2011

[21] Lazer, Pentland et al., 2009

zentrieren sich Data Scientists zurzeit auf das mathematisch mögliche. Erst durch das Hinzufügen von fachlichem Kontext können die Daten interpretiert werden und möglicherweise erkannt werden, ob Algorithmen korrekt sind oder revidiert werden müssen.[22]

Hathaway-Effekt: Algorithmen brauchen menschliche Interpretationen und Eingriffsmöglichkeiten. Jedes Mal, wenn die Schauspielerin Anne Hathaway in den Nachrichten genannt wird, z. B. in Form von Filmkritiken, Oscar-Verleihungen oder Garderobenunglücken, steigen die Börsenkurse von Warren Buffetts Holdinggesellschaft Berkshire-Hathaway.[23] Daher könnte man annehmen, dass es eine Kausalität zwischen dem Anstieg des Aktienkurses und Anne Hathaways Erwähnungen in den Medien gibt. Was jedoch viel wahrscheinlicher ist, ist, dass automatisierte, computergestützte Trading-Programme den Eintrag »Hathaway« aufgreifen und auf die Börse übertragen. Die Programmierung kann nicht zwischen der Schauspielerin Anne Hathaway und Berkshire-Hathaway-Aktien unterscheiden und führt zu Verzerrungen der Börsenkurse.

Streetlight- & Beobachtungseffekte:[24] Durch die oftmals bestehende Beschränkung auf die Analyse von bestimmten, leicht zugänglichen Social-Media-Datensätzen, wie z. B. von Twitter, werden Populationen vor-ausgewählt, die nicht unbedingt repräsentativ für die Bevölkerung sind[25]. Beispielsweise nutzen nur 17 Prozent der Internetnutzer auch Twitter, wobei viele der Twitterkonten inaktiv sind. Trotzdem werden aus den Äußerungen von Twitternutzern Vorhersagen für politische Wahlergebnisse oder Trends in der öffentlichen Meinung getroffen. Dieser sogenannte *Drunkard's Search* bzw.

[22] Chrystal & Mizen, 2003

[23] Mirvish, 2017

[24] Kaplan, 1964

[25] Freedman, 2010

Streetlight Effect bedeutet: Wir suchen dort, wo Daten leicht zugänglich sind, aber nicht da, wo Erkenntnisse über die Gesamtpopulation gewonnen werden können.

Gefahren durch den öffentlichen Charakter der Datensätze: Online Interaktionen erlauben direkte Rückschlüsse auf persönliche identifikationsmerkmale einzelner Bürger. Der kürzlich bekannt gewordene Cambridge-Analytica-Fall hat verdeutlicht, dass Marketing- und Datenanalysefirmen die sozialen Graphen aller Facebooknutzer erheben und sogar Rückschlüsse auf Bürger zulassen, die kein eigenes Facebook-Konto besitzen. Es gibt zurzeit jedoch wenig Handhabe gegen US-Firmen – auch wenn die EU und Deutschland versuchen mit lokalen Vorschriften die Nutzung der Daten einzuschränken.

6. Schlussfolgerungen und Aufgaben für Verwaltung und Regierung

Für die Nutzung von Big Data und Data Analytics in der öffentlichen Verwaltung besteht sowohl aus verwaltungsinterner kultureller wie auch aus rechtlicher Sicht weiterhin großer Handlungsbedarf. Die Erstellung von Big Data-Datensätzen und die Nutzung mit Hilfe von Data-Science-Ansätzen ist immer von bestimmten menschlichen Annahmen und Entscheidungen getrieben. Das heißt, dass die Art und Weise, wie besonders Internetunternehmen Datensätze generieren und die öffentliche Verwaltung oder Forscher sie nutzen ist niemals neutral und ohne systematische menschliche Fehler einzustufen, sondern immer mit bestimmten Entscheidungen verbunden. Die öffentliche Verwaltung und der Gesetzgeber müssen die Regulierung nichtöffentlicher und auch öffentlich zugänglicher Daten, deren Wiederverwendung oder gemeinsame Nutzung über Sektorengrenzen und Websites hinweg begutachten und gegebenenfalls

einschränken, sodass Persönlichkeitsrechte der Bürger gewahrt werden.

Es bleiben viele Forschungsfragen im Zusammenhang zur Nutzung von Big Data und Data Science-Ansätzen in der öffentlichen Verwaltung offen, wie z. B.: Was sind die ethischen Überlegungen, die sich aus der Wiederverwendung großer Datenmengen als Teil der Sharing & Data Economy ergeben? Wie soll vor allem der Privatsektor zur Rechenschaft gezogen werden, wenn Datensätze erstellt werden, die persönliche Informationen über das politische Verhalten von Bürgern nachvollziehbar machen? Welchen Einfluss haben Big Data-Analysen auf die Demokratie und Repräsentation der Bevölkerung? Wie können Data Scientists bei der Nutzung von Big Data-Analysen auch die Randgruppen, die nicht auf populären Social-Media-Plattformen vertreten sind oder gar nicht online interagieren, mit einbeziehen?

Des Weiteren sollten Big-Data-Analysen auch zur Messung und zum Vergleich von Verwaltungsleistungen genutzt werden: Wie können wir große Datenmengen nutzen, um vergleichende Einblicke in Verwaltungshandeln zu erhalten? Wie kann die Analyse großer Datenmengen dazu beitragen, dass Behördenvorgänge agiler, anpassungsfähiger und reaktionsschneller werden?[26] Es bleibt weiter wichtig, menschliche Einsichten und analytische Fähigkeiten in der öffentlichen Verwaltung auszubauen: Wie können also analytisch orientierte Data Scientists menschliches Wissen (Bauchgefühl, Gefühle, Veränderungen, Einstellungen) mit einbeziehen und analysieren?

[26] Mergel, 2016

Quellen

Banbura, M., Giannone, D., Modugno, M. & Reichlin, L. (2013). Now-casting and the real-time data flow. In: *Handbook of economic forecasting* 2 (Part A), S. 195-237

Boyd, D. & Crawford, K. (2011). Six provocations for big data. In: *A decade in internet time. Symposium on the dynamics of the internet and society.* Oxford Internet Institute. 21, Oxford

Chetty, R., Hendren, N., Kline, P. & Saez, E. (2014). Where is the land of opportunity? The geography of intergenerational mobility in the United States. In: *The Quarterly Journal of Economics* 129(4), S. 1553-1623

Chrystal, K. A. & Mizen, P. D. (2003). Goodhart's law: its origins, meaning and implications for monetary policy. In: *Central banking, monetary theory and practice: Essays in honour of Charles Goodhart* 1 (S. 221-243)

Cox, M. & Ellsworth, D. (1997). Application-controlled demand paging for out-of-core visualization. In: *Proceedings of the 8th conference on Visualization '97,* IEEE Computer Society Press

Gayo-Avello, D. (2012). »I Wanted to Predict Elections with Twitter and all I got was this Lousy Paper« - A Balanced Survey on Election Prediction using Twitter Data. arXiv preprint arXiv:1204.6441

George, G., Haas, M. R. & Pentland, A. (2014). Big data and management. In: *Academy of Management Journal* 57(2), S. 321-326

Google.org (2008). Google Flu Trends. http://s.fhg.de/sB4

Kaplan, A. (1964). The Conduct of Inquiry. Chandler, San Francisco, CA

Lazer, D., Kennedy, R., King, G. & Vespignani, A. (2014). Big Data: The Parable of Google Flu: Traps in Big Data Analysis. In: *Science* 343(6176), S. 1203-1205

Lazer, D.,Pentland, A. S., Adamic, L, Aral, S.,Barabasi, A. L., Brewer, D., Christakis, N., Contractor, N., Fowler, J., Gutmann, M., Jebara, T., King, G., Macy, M., Roy, D. & Alstyne, M. V. (2009). Life in the network: the coming age of computational social science. In: *Science* 323(5915)

Mergel, I. (2016). Agile Innovation Management in Government: A Research Agenda. In: *Government Information Quarterly* 33(3), S. 516-523

Mergel, I. (2017). Korruptionsbekämpfung in Echtzeit. In: Behörden Spiegel. http://s.fhg.de/kQ6

Mergel, I., Rethemeyer, R. K. & Isett, K. R. (2016). Big data in public affairs. In: *Public Administration Review* 76(6), S. 928-937

Mergel, I., Rethemeyer, R. K. & Isett, K. R. (2016a). What does Big Data mean to public affairs research? Understanding the methodological and analytical challenges. LSE Impact Blog

Mervis, J. (2014). How two economists got direct access to IRS tax records. In: *Science-Insider*

Mirvish, D. (2017). The Hathaway Effect: How Anne Gives Warren Buffett a Rise. In: *Huffington Post*

Mislove, A., Lehmann, S., Ahn, Y.-Y., Onnela, J.-P. & Rosenquist, J. N. (2011). *Understanding the Demographics of Twitter Users*

n. a. (2017) The world's most valuable resource is no longer oil, but data. In: *The Economist*

Newman, D. (2016) Big Data And The Future Of Smart Cities. Forbes

Nijhus, M. (2017) How to call Bullshit on Big Data: A Practical Guide. In: The New Yorker.

Raghupathi, W. & Raghupathi, V. (2014). Big data analytics in healthcare: promise and potential. In: *Health Information Science and Systems* 2(1), S. 3

Rajkomar, A., Oren, E., Chen, K., Dai, A. M., Hajaj, N. Liu, P. J., Liu, X., Sun, M., Sundberg, P. & Yee, H. (2018). Scalable and accurate deep learning for electronic health records. arXiv preprint arXiv:1801.07860.

Robbins, M. D., Simonsen, B. & Feldman, B. (2008). Citizens and Resource Allocation: Improving Decision Making with Interactive Web-Based Citizen Participation. In: *Public Administration Review* 68(3), S. 564 - 575

USGS (2015). Did you feel it? Internet Impact Map. http://s.fhg.de/3xC

Über die Autorin

Ines Mergel

Dr. Ines Mergel ist Universitätsprofessorin für *Public Administration* an der Universität Konstanz, wo sie im Fachbereich Politik- und Verwaltungswissenschaften zu Themen der Digitalisierung und Digitalen Transformation des öffentlichen Sektors forscht und lehrt. Nach ihrem Diplom in Wirtschaftswissenschaften, Universität Kassel, hat sie 2005 ihren Doktor in Informationsmanagement an der Universität St. Gallen (Schweiz) abgeschlossen und von 2002 - 2008 an der *Harvard University, Kennedy School of Government* am *National Center for Digital Government* und dem *Center for Networked Governance* geforscht. Danach war sie von 2008 - 2016 *Assistant* und *Associate Professor* mit *Tenure* an der *Maxwell School of Citizenship and Public Affairs, Syracuse University (USA)* tätig.