10 Jahre Amt für Statistik Berlin-Brandenburg

Wozu noch amtliche Daten, wenn es Google gibt?

Festrede zur Jubiläumsfeier anlässlich des zehnjährigen Bestehens des Amtes für Statistik Berlin-Brandenburg

von **Katharina Schüller**

„Statistical Literacy bedeutet, zu verstehen, welcher Wert in Daten steckt und welcher Aufwand nötig ist, um daraus Wissen entstehen zu lassen."

Foto: Mike Auerbach

Liebe Gäste,
sehr geehrter Herr Frees,
sehr geehrte Frau Staatssekretärin Lange,
sehr geehrte Frau Dr. Michaelis-Merzbach,
sehr geehrter Herr Präsident Sarreither,
sehr geehrter Herr Wayand,
liebe weitere Gäste, Mitarbeiterinnen und Mitarbeiter des Amtes für Statistik Berlin-Brandenburg,

es ist mir eine große Freude und Ehre, die heutige Festrede halten zu dürfen.

Der amerikanische Zukunftsforscher John Naisbitt hat vor 25 Jahren gesagt: „Wir ertrinken in Informationen und hungern nach Wissen."

Seit zehn Jahren stillen Sie, liebe Vertreter des Amtes für Statistik Berlin-Brandenburg, unseren Wissenshunger durch weit mehr als verlässliche Zahlen. Von der Zahl zur Information, so lautet Ihre Devise. Und Sie liefern noch mehr als das – nämlich umfangreiches Wissen über Stadt und Land, Unternehmen und die Zivilgesellschaft in Form von zahlreichen Publikationen und Produkten. Daten, Informationen, Wissen mit hohem Qualitätsanspruch – darauf greifen nicht nur Statistiker wie ich dankbar zu.

Denn wir können davon ausgehen, dass bei Ihnen wie in einem Gourmet-Restaurant nur hochwertige Zutaten zum Einsatz kommen, dass diese professionell vorbereitet und bekömmlich zubereitet werden und dass sie schließlich in einem ansprechenden Kontext, entsprechend unseren eigenen hohen Ansprüchen, serviert werden.

Aber diese Qualität hat auch ihren Preis. Bei den Statistischen Ämtern gibt es eben kein billiges Fast Food, es gibt Bio-Hühnchen „sous vide" nach Art des Chefs und keine Chicken McNuggets – und wir müssen manchmal ziemlich lange darauf warten. Verlässliche Zahlen sind nicht unbedingt aktuell. Und so bekommen alternative Anbieter wie Google & Co. – in gewisser Weise so etwas wie der Daten-McDonalds – zunehmend Zulauf.

Wozu also noch amtliche Daten, wenn es Big Data gibt?

Machen Google und Facebook die Statistischen Ämter bald überflüssig? Weil Daten, das „Öl des 21. Jahrhunderts", inzwischen überall reichlich sprudeln – und noch dazu fast nichts mehr kosten?

Im Zeitalter der Digitalisierung scheinen Daten allgegenwärtig und endlos zu sein. Dennoch halte ich diese Daten zugleich für über- und unterbewertet. Diese Daten sind unterbewertet, weil unzählige datengenerierende Systeme gigantische Mengen von Bits und Bytes produzieren. So viele, dass man nur noch aus dem Vollen zu schöpfen braucht?

So einfach ist es nicht. Diese Daten sind gleichzeitig überbewertet, weil zwar alle Welt euphorisch von der „Macht der Daten" spricht. Aber kaum jemand ist sich im Klaren, wo die Grenzen dieser Macht liegen. Denn: Der Großteil der Daten ist weder verknüpft noch organisiert. Deswegen kann er nicht, zumindest nicht ohne Weiteres, in Wissen verwandelt werden.

Ich bin überzeugt: Damit wir zu einer zeitgemäßen Auffassung dessen kommen, wie wir die Nutzbarkeit und Relevanz von Daten im digitalen Zeitalter bewerten können, brauchen wir etwas ganz Altmodisches: Statistical Literacy.

Die Bedeutung von Statistical Literacy hat schon vor mehr als 100 Jahren der Schriftsteller Herbert George Wells in seinen politischen Schriften herausgestellt: „Wenn wir mündige Bürger in einer modernen technologischen Gesellschaft möchten, dann müssen wir ihnen drei Dinge beibringen: Lesen, Schreiben und statistisches Denken, das heißt den vernünftigen Umgang mit Risiken und Unsicherheiten."

Wells' bekanntestes Werk ist übrigens die „Zeitmaschine" …

Informationsdienstleister zu sein, heißt eben heute und zukünftig ganz gewiss nicht mehr, die alleinige Hoheit über die Daten zu besitzen. Dazu sprudeln die neuen Daten- und Informationsquellen längst viel zu stark. Und es widerspricht nicht zuletzt dem Anspruch an Transparenz unserer heutigen Zivilgesellschaft und der Forderung nach Open Data.

Vielleicht wird es auf Dauer nicht einmal gelingen, die offizielle Interpretationshoheit über die Daten zu behalten – falls das überhaupt gewollt sein kann. So etwas würde ja womöglich bedeuten: Open Data gibt es von den Ämtern nur in Kombination mit Interpretationshilfen. Damit nicht jeder selbsternannte Datenjournalist wilde Korrelationen berechnet und die Ergebnisse als Kausalitäten verkauft, die amtliche Statistiker dann wieder mühsam korrigieren müssen.

Nein, wir Statistiker können und sollten uns nicht daran abarbeiten, statistischen Unsinn zu korrigieren, den andere in die Welt hinausblasen. Das wäre eine ähnlich unmögliche Aufgabe wie die Geschichte vom Huhn des Heiligen Philipp, die sich im Rom des 16. Jahrhunderts folgendermaßen zugetragen haben soll:

> Die Contessa Bianchi soll dem heiligen Philipp Neri gebeichtet haben, wie gern sie Fake News über andere Leute verbreitete. Zur Buße ließ er sie auf dem Markt ein Huhn kaufen und zu ihm bringen, es unterwegs aber sorgfältig rupfen. Als die Contessa ihm das gerupfte Huhn überreichte, gab er ihr die Aufgabe, die verstreuten Federn wieder einzusammeln. Das ist unmöglich, rief sie – der Wind hat sie ja schon über ganz Rom verteilt! Siehst du, antwortete Philipp, genauso wenig kannst du deine falschen Neuigkeiten wieder zurückholen.

Ein halbes Jahrtausend später ist schließlich „postfaktisch" das Wort des Jahres. Dabei müssen post- oder kontrafaktische statistische Aussagen noch nicht einmal böswillig erfundene Fake News sein, wie solche, denen wir das Brexit-Votum zu verdanken haben. Häufig entsteht statistischer Unsinn schlicht aus einem Mangel an statistischem Denken.

Und falls Sie zweifeln, dass dieser Mangel ein Problem sein könnte, lassen Sie mich Folgendes zitieren:

„Jeweils rund zwei Drittel der Deutschen haben nennenswerte Zweifel daran, dass die von amtlichen Stellen veröffentlichten Statistiken zur Zuwanderung, zur Einkommens- und Vermögensverteilung und zur Arbeitslosigkeit die Wirklichkeit einigermaßen korrekt widerspiegeln. Die vom Staat behaupteten wirtschaftlichen Realitäten werden von vielen Bürgern als postfaktische Beschreibungen wahrgenommen, die nicht den eigenen Erfahrungen entsprechen."

So kommentiert Jürgen Doeblin vom gleichnamigen Wirtschaftsforschungsinstitut seine jüngste Befragung. (Dass seine Stichprobe von 1 072 Befragten repräsentativ sei, begründet er übrigens durch den Vergleich ihrer Strukturmerkmale mit denjenigen der Bevölkerung, die er amtlichen Statistiken entnimmt.)

Statistisches Denken zu vermitteln, Statistical Literacy in die Welt zu tragen – da sehe ich eine große zukünftige Aufgabe der Statistischen Ämter. Eine Tagung wie das 24. Wissenschaftliche Kolloquium des Statistischen Bundesamtes, „Statistik verstehen – Orientierung in der Informationsgesellschaft", kann da nur der Anfang sein.

Was brauchen wir für die Orientierung in der Informationsgesellschaft?

Um diese Frage zu beantworten, brauchen wir ein klares Verständnis davon, was Wissen ist und was Daten sind. Daten sind Abstraktionen der realen Welt. Wenn man Daten bereinigt und verknüpft, entstehen aus ihnen Informationen. Analysierte und infolgedessen organisierte Informationen wiederum begründen Wissen. Und angewandtes, also sinnvoll interpretiertes und genutztes Wissen konstituiert schließlich Weisheit oder – wie es der französische Philosoph Michel Foucault nennt – Macht.

Digitalisierung schafft gigantische Datenfluten. Hierzu ein paar nicht-amtliche Zahlen: Das Wachstum neu entstandener Daten hat sich in gerade einmal fünf Jahren verzehnfacht, auf geschätzte knapp neun Zetabytes im Jahr 2015. Ein Ende ist nicht in Sicht – und über 90 % der Daten sind unstrukturiert. 85 % der Daten entstehen aus neuartigen Quellen wie beispielsweise Smartphones, Social Media und Sensoren.

Daten gab es allerdings schon immer und Datenanalyse ist auch keine neue Erfindung. Trotzdem gelten Daten als der Rohstoff des 21. Jahrhunderts. Wert, also Weisheit und Macht, aus Daten zu schöpfen, ist jedoch weit weniger einfach, als es den Anschein hat. Sie als Daten-Profis wissen das längst; die Wirtschaft lernt es gerade.

75 % aller Analytics-Projekte scheitern. 88 % der Daten in Unternehmen liegen brach und werden niemals ausgewertet. Amerikanischen Unternehmen entstehen jährlich 700 Mrd. $ Kosten aufgrund schlechter Datenqualität.

Wenn die Zähmung der Datenflut also schon in einem als derart innovativ geltenden Umfeld wie der amerikanischen Wirtschaft häufig zu scheitern droht, wie soll sie dann erst bei uns gelingen?

Wenn ich „bei uns" sage, dann möchte ich im Folgenden einige Beispiele aus der kommunalen Selbstverwaltung und der empirischen Stadtforschung aufgreifen. Da tut sich im Moment extrem viel, etwa bei Ihren Kolleginnen und Kollegen im Bundesinstitut für Bau-, Stadt- und Raumforschung.

Denn der Begriff „Smart City" ist gerade in aller Munde: Daten stehen in Städten und Stadtforschung in bisher ungeahntem Umfang zur Verfügung. Damit verändern sich Entscheidungsprozesse durch die Digitalisierung enorm. Umso wichtiger wird der Dialog zwischen Statistik, Stadtplanung und Verwaltung.

Nicht zuletzt ist das Interesse der Privatwirtschaft, technologische Lösungen und Daten an unsere Städte zu verkaufen, so groß wie nie zuvor. Deshalb müssen wir unseren konstruktiven wie kritischen Blick für die Probleme schärfen.

Weil es oft um hohe Investitionssummen geht, spielt die Datenqualität eine ganz besondere Rolle. Etwa bei der Frage:

Wie plant man eigentlich eine neue Schule?

Zu einem meiner ersten Aufträge kam ich im Jahr 2005 ganz nach dem Motto „Glaube keiner Statistik, die du nicht selbst gefälscht hast". Mit diesen Worten zweifelte eine Elternvertreterin eine Schülerzahlprognose an, die von der Hochschule erstellt worden war, an der ich lehrte.

Kommunen, wie damals der Landkreis Erding, geben Schulentwicklungsgutachten in Auftrag, damit es in Zukunft weder zu wenige noch zu viele Schulen gibt. Ob eine Schule neu gebaut, erweitert oder geschlossen werden soll oder wie viele Lehrkräfte in fünf oder zehn Jahren benötigt werden – Entscheidungen hierzu stützen sich stets auf eine Prognose der Schülerzahlen. (Dass es oft genug darum geht, mit Hilfe der scheinbar objektiven Statistik eine politische Entscheidung zu vermeiden, steht auf einem ganz anderen Blatt.)

Eine Prognose ist aber nur so gut wie die zugrundeliegenden Daten. Typischerweise sind das vergangenheitsbezogene Selbstauskünfte von Schulen sowie Daten der Statistischen Ämter.

Man kann daraus eine Menge ablesen, gerade wenn man auf Sonderauszählungen zugreift – da haben wir hervorragende Erfahrungen gemacht, nicht nur was die Datenqualität angeht, sondern auch die Freundlichkeit und Hilfsbereitschaft der Mitarbeiter. Das Preis-Leistungs-Verhältnis und die Serviceorientierung Ihrer Kolleginnen und Kollegen haben mich immer wieder positiv überrascht.

Bloß: Eine Prognose auf Basis der Vergangenheit ist wie Autofahren mit Blick in den Rückspiegel. Das funktioniert nur gut, so lange die Straße frei und gerade ist. Die zunehmende Krümmung der Bildungswege schafft jedoch viele Unsicherheiten. Mit Vergangenheitsdaten können wir diese Unsicherheiten kaum in den Griff bekommen. Wir behelfen uns dann mit Szenario-Rechnungen und mit Annahmen darüber, wie mögliche Zukünfte aussehen könnten.

Vielleicht könnte man aber mit Verfahren wie Prognosemärkten zukünftige Absichten von Eltern, Schülern, Lehrern oder Ausbildungsbetrieben in verschiedenen Szenarien genauer erfassen. Vielleicht könnte man dadurch Daten und Wissen generieren, in der Hoffnung auf bessere Entscheidungen. So wie Unternehmen das machen, wenn sie Kunden mittels Crowdsourcing neue Produkte entwickeln lassen.

Doch dabei geht es nicht mehr nur um die Frage, welche Daten am Ende besser sind: die objektiven, standardisierten, aber nicht besonders aktuellen Zahlen der Statistischen Ämter oder die Absichtserklärungen der Beteiligten, die zwar sehr zeitnah und feinräumig erhebbar sind, aber weder repräsentativ noch in sonst irgendeiner Weise valide sein müssen?

Sondern es führt uns zu der viel fundamentaleren Frage, welches methodische Umdenken Big Data in der Planung hervorrufen kann: Heute ziehen wir dafür Ursache-Wirkungs-Modelle heran, die auf standardisierten Daten aufbauen. Kinder werden von Frauen im gebärfähigen Alter geboren, treten in die Schule ein und in andere Schulen über, Familien ziehen hinzu oder weg usw.

An die Stelle solcher Modelle tritt die Auswertung heterogener und teilweise gigantisch umfangreicher Datenmengen mit Algorithmen. Entscheidend dabei ist die Erkenntnis: Mehr Daten heißt nicht zwangsweise mehr Wissen.

Denn trotzdem bleibt der zukünftige Schulbedarf mit Unsicherheit behaftet. Selbst wenn alle Beteiligten nach bestem Wissen am Prognosemarkt teilnehmen, können sich relevante Bedingungen ändern. Das zeigt, dass auch Big Data nicht zwangsweise einen Blick in die Zukunft erlaubt.

Daten schaffen zwar Wissen – aber wie genau geht das?

Ich stelle die Frage mal anders. Warum ist ein Diamantring von Tiffany so wertvoll? Warum ist er viel wertvoller als dieser (mein) Ring? Weil Tiffany eine Strategie der Wertschöpfung verfolgt, die man ganz analog auf Daten anwenden kann.

In der ersten Stufe geht es darum, hochwertige Rohstoffe zu gewinnen: Erz und Rohdiamanten. Bei der Wissensschöpfung heißt das: Wie können wir Daten von Bedeutung erhalten? Hier unterscheiden sich mein Ring und der von Tiffany noch nicht sonderlich.

Dann muss das Gold geschmolzen und geschmiedet und der Diamant muss geschliffen werden. In der Analogie lautet die Frage: Wie können wir durch Bereinigung und Aggregation aus Daten Informationen erzeugen? Dabei entsteht schon deutlich mehr Wert.

Drittens werden Gold und Stein zu einem Ring zusammengefügt: Wie können wir entsprechend Wissen aus Informationen gewinnen, indem wir diese verknüpfen und analysieren? Ab hier wird es wirklich interessant.

Der größte Teil der Wertschöpfung geschieht nämlich auf der vierten und letzten Stufe: Wie sollen wir das Wissen interpretieren und auf dieser Basis handeln? Richtig wertvoll macht den Ring erst der Stempel, den Tiffany am Ende hinzufügt – und das ist nichts anderes als eine Interpretation und Handlungsanweisung: Kauf' mich, verschenk' mich, dann sagt sie „ja".

Heute ertrinken wir in Daten. Die wenigsten holen aus dem Rohstoff das ganze Potenzial heraus, auch das gilt analog zum Ring-Beispiel – und im Übrigen auch für das Gourmet-Restaurant. Damit es gelingt, neues Wissen zu schaffen, müssen wir begreifen, wo Digitalisierung Wissen produziert und wo bloß Daten – und dass (ich sage es nochmal, weil es so wichtig ist) ein Mehr an Daten nicht zwangsläufig ein Mehr an Wissen bedeutet. Big Data oder Open Data an sich sind noch kein Wert.

Statistical Literacy bedeutet, zu verstehen, welcher Wert in Daten steckt und welcher Aufwand nötig ist, um daraus Wissen entstehen zu lassen. Nicht immer lohnt dieser Aufwand. Manchmal verbieten rechtliche Hürden die Verknüpfung sensibler Datenquellen, manchmal liegen die benötigten finanziellen, technischen und personellen Ressourcen jenseits des Möglichen, manchmal enthalten die Daten auch schlicht zu wenig relevante Informationen, um eine Fragestellung hinreichend zu beantworten.

Bloß – wann sind neue Daten relevant und wann sind sie nutzbar?

Die großen Herausforderungen liegen darin, die Chancen und die Risiken neuer Datenquellen realistisch abzuschätzen. Dabei spielt eine ganze Reihe von Kriterien eine Rolle. Einerseits geht es um technische Fragen der Implementierung einer Wissensschöpfungskette. Darunter fallen die Datenintegration, die Qualitätssicherung, die Standardisierung und Analyse sowie die konkrete Umsetzung im Tagesgeschäft.

Aber genauso wenig dürfen wir die Auswirkungen auf die interne Organisation eines Unternehmens oder einer Stadt ignorieren. Rechtliche und organisatorische Rahmenbedingungen müssen betrachtet und auf den Prüfstand gestellt werden: Wer ist „Herr der Daten", welcher Bedarf an Kooperation und Koordination ergibt sich, wer haftet für Fehlentscheidungen?

Und schließlich können Beziehungen zwischen Verwaltung und Zivilgesellschaft oder zwischen einem Unternehmen und seiner Außenwelt betroffen sein. Hier geht es um informationelle Selbstbestimmung, die nicht jede legal mögliche Datennutzung legitimiert, wie auch um digitale Glaubwürdigkeit: Was macht eine privatwirtschaftliche oder öffentliche Organisation mit den Daten ihrer Bürger, Kunden, Lieferanten und Partner und welche Machtverschiebungen können sich daraus ergeben?

Ich möchte auf diese Fragen im Folgenden näher eingehen, weil sie uns verstehen helfen, wie relevant qualitativ hochwertige Daten und Informationen der Statistischen Ämter auch in Zukunft sein werden. Nichtsdestotrotz beschäftigt sich auch die öffentliche Hand mit der Frage, wie man neue Datenquellen zur Planung, Steuerung und Entscheidungsfindung nutzen kann.

Die Digitalisierung ist auf dem Vormarsch, neue Datenanbieter drängen aufs Spielfeld. Das können wir nicht wegdiskutieren. Aber wir können uns überlegen, was das bedeutet und wie wir diese Daten möglichst sinnvoll integrieren können.

Datenintegration heißt: den Bohrturm bauen. Wie können wir das „Öl des 21. Jahrhunderts" fördern?

Naheliegend ist, dass ein erster Blick den technischen Standards und Schnittstellen gelten muss. Hier müssen Städte und Kommunen ermitteln, wo sie ihre Infrastruktur anpassen müssen und welcher Aufwand nötig ist, um Big Data in die bestehende Systemlandschaft zu integrieren.

Weiterhin stellt sich die Frage, welche Lizenzkosten oder -einschränkungen bei der Software anfallen. Womöglich gibt es passende Open-Source-Lösungen, die aber nur scheinbar kostenlos sind. Sie sind eben oft mit einer höheren Einarbeitungsdauer verbunden, weil eine einheitliche Dokumentation fehlt. Das wirkt sich auch darauf aus, wie hoch der spätere Aufwand für Wartung und Support sein wird. Auch Google Analytics kostet nur scheinbar nichts. Wir bezahlen dafür mit unseren Daten.

Besonders wichtig ist schließlich die Prüfung, wie zuverlässig das System im Ernstfall ist. Wenn ausgerechnet im Katastrophenfall die Internet-Verbindung ausfällt, darf eine Stadt sich nicht allein aufs Crowdmapping via Facebook verlassen.

Geschafft. Die Daten sind drin – aber wie brauchbar sind sie? Qualitätssicherung heißt: das Rohöl säubern und sortieren.

Neue Datenquellen sollten Städte und Kommunen nur nutzen, wenn sie korrekte Informationen generieren. Das klingt trivial, ist es in der Praxis aber nicht. Man braucht dazu allgemein akzeptierte Regeln, um diese Gültigkeit zu prüfen und sicherzustellen.

Wer von Ihnen schon mal in einer Kreistags- oder Ausschusssitzung saß, der weiß, wie häufig der „gesunde Menschenverstand", Erfahrungswissen oder schlicht Machtspielchen solche Regeln substituieren sollen. Im Schulentwicklungs-Fall kennt man dann Eltern, die ihr Kind niemals auf die Schule im Nachbarlandkreis schicken würden, und überhaupt, also kann die Prognose gar nicht stimmen …

Selbst wenn die Daten richtig sind, heißt das längst noch nicht, dass sie relevant sind. Schätzungen zufolge sind bei Großkatastrophen nur etwa 8 % der Tweets und Social-Media-Einträge relevant, das heißt, trotz entsprechender Hashtags enthielten 92 % keine Informationen, die für Betroffene oder Helfer von Bedeutung waren.

Nicht unterschätzen darf man auch das Problem der Repräsentativität. Verzerrungen sind eine ernstzunehmende Gefahr, etwa weil nicht alle Bevölkerungsgruppen gleichermaßen an einer digitalen Form der Bürgerbeteiligung teilnehmen. Selten liefern neue Datenquellen also alle Informationen, die man braucht. Die Frage lautet dann sofort: Welche zusätzlichen Datenquellen sind notwendig und welcher Aufwand entsteht, um Lücken zu schließen und alle Informationen zu verknüpfen?

Oft wird argumentiert, dass neue Datenquellen aktuellere, präzisere oder feinräumigere Informationen ermöglichen. Dies kann durchaus zu Ressourceneinsparungen führen. Andererseits müssen die hoch granularen Daten oft eben doch wieder sinnvoll aggregiert werden. Und das kostet Ressourcen.

Kommen wir zur Standardisierung und Analyse, das heißt: den Rohstoff zum Treibstoff machen.

Was messen die Daten und Informationen eigentlich? Diese Frage zielt einerseits auf die Skalierung der möglichen Messung ab. Ein wichtiger Punkt ist auch, wie die Erhebungsmerkmale charakterisiert sind, etwa im Vergleich zu den Demographischen Standards der amtlichen Statistik.

Wenn in einer Online-Befragung von „Studenten" die Rede ist, dann meint dieser Begriff nicht unbedingt Studenten im Sinne des Statistischen Bundesamtes oder Studenten im Sinne der Arbeitsagentur. Die sachliche Standardisierung der Messung, beispielsweise eine Ableitung von Branchenzuordnungen, kann sehr aufwendig sein.

Gleiches gilt für die räumliche Standardisierung. Sollen die Daten räumlich verortet werden, entstehen oft Unsicherheiten im Raumbezug. Nicht jede räumliche Information kann man direkt auf einer Karte abbilden. Wenn auf Facebook von der „Brücke hinter dem Aldi" die Rede ist, lässt sich daraus kaum eine Georeferenzierung ableiten.

Andere räumliche Einteilungen, die beispielsweise die Industrie- und Handelskammern in ihren Datenbanken zur Prognose des Fachkräftebedarfs nutzen, decken sich nicht unbedingt mit Gemeindegrenzen.

Ganz analoge Überlegungen müssen wir schließlich für die zeitliche Standardisierung anstellen. Die wenigsten neuen Datenquellen orientieren sich an den Stichtagen der amtlichen Statistik. Aber wie gut lassen sich die Daten dann zeitlich verorten und verknüpfen?

Städte kommen deshalb zumindest bei ihren ersten Gehversuchen, neue Datenquellen zu analysieren – egal, ob Small Data oder Big Data –, kaum um einen hohen manuellen Aufwand herum. Sie müssen Indikatoren bilden, Zusammenhangsstrukturen analysieren, geeignete Formen der Visualisierung ausprobieren und anpassen.

Werden Datenströme kontinuierlich integriert, etwa in den zunehmend populären City-Dashboards, dann braucht es dazu automatisierte Prozesse. Um zu beurteilen, ob ein solches Dashboard für die Stadtverwaltung und die Bürger relevant ist, muss man sich erst einmal überlegen, welche Daten überhaupt visualisiert werden sollen, was diese Daten messen (und was nicht), wie präzise sie sind, wie sie laufend eingebunden und analysiert werden können und welcher Aufwand damit verbunden ist. Offensichtlich spielt also schon für eine recht naheliegende Anwendung – wie sie ja in Unternehmen vielfach verbreitet ist – eine Vielzahl von Kriterien eine wichtige Rolle.

Auf einer Metaebene gilt es zu klären, was ein solches Dashboard eigentlich abbildet: Sind es Daten? Sind es Informationen? Ist es Wissen? Und wie sollen Städte und Kommunen in Abhängigkeit von diesem Wissen handeln, etwa wenn virtuelle Warnleuchten auf Grenzwerte der Indikatoren reagieren und einen kritischen Zustand der Verkehrssituation oder der Luftverschmutzung anzeigen? Für solche Aufgaben der Prognose und Steuerung braucht es Data Analytics und Modelle.

Damit geht es konkret an die Umsetzung: Wir wollen die PS auf die Straße bringen.

Mit Big Data und Data Analytics eröffnen sich Wege, an räumlich und zeitlich hoch aufgelöste Informationen zu gelangen und diese in Steuerungswissen zu verwandeln. Besonders interessant ist dabei die Prognose, auch „Predictive Analytics" genannt. Sie erfordert jedoch neue Kompetenzen bei den Datenanalysten.

Bisher wertet die öffentliche Verwaltung mithilfe der deskriptiven Statistik üblicherweise Vergangenheitsdaten aus; selten nutzt sie Szenarioanalysen und Trendmodelle (es gibt Ausnahmen, aber die Privatwirtschaft schreitet da sehr viel forscher voran). Sie spekuliert eben nicht – das ist auch gut so.

Prognosen auf Basis großer Datenmengen erfordern aber nicht nur neue Verfahren des Data Mining und Machine Learning, sondern häufig auch spezielle Datenstrukturen. Daten müssen womöglich auf der Ebene des einzelnen Bürgers oder der einzelnen Interaktion abgebildet werden. Es kann ziemlich aufwendig sein, solche Datensichten zu erzeugen.

Mit einem Prognosemodell an sich ist es dann auch nicht getan. Eine Prognose muss man auch evaluieren und eventuell korrigieren. Deswegen braucht man Methoden, um sicherzustellen, dass die Analyseergebnisse und die daraus abgeleiteten Entscheidungen und Entscheidungsregeln korrekt sind.

Besonders spannend ist dabei, was passieren soll, wenn sich Widersprüche ergeben, etwa wenn eine Prognose der Schülerzahlen zu ganz anderen Standortentscheidungen führen würde als eine Elternbefragung mittels einer App. Es kann nötig sein, Experten aus verschiedensten Disziplinen einbeziehen, beispielsweise GIS-Experten, Soziologen, Computerlinguisten, Statistiker, aber auch Fachexperten aus den einzelnen Referaten.

Das führt uns nun zu den organisatorischen Fragen, oder: Wie kommen wir von der Probefahrt zum dauerhaften Fahrbetrieb?

Es gibt immer mehr Daten, und der Druck, auf Basis dieser Daten zu entscheiden, steigt. Datenbasiertes Entscheiden führt aber zu tiefgreifenden Veränderungen. Häufig werden Machtstrukturen in Frage gestellt und bisherige Entscheidungsregeln entpuppen sich als Mythen.

Analytics schafft Wert, aber darf das Tagesgeschäft nicht stören. Die Beteiligten müssen darum überzeugt werden, ihre Daten zu teilen und damit Wissen und Macht abzugeben, damit mehr Wert für alle entstehen kann.

„This persuasion task is probably more difficult than any technological issues that might come up", schreiben Stephen Goldsmith und Susan Crawford in ihrem sehr lesenswerten Buch „The Responsive City".

Viele Menschen betrachten solche neuen Ideen kritisch, weil sie in Bezug auf Datensicherheit und Datenschutz unsicher sind. Städte und Kommunen müssen genau klären, welche rechtlichen Rahmenbedingungen sie beachten müssen und welcher Aufwand sie erwartet, um sicherzustellen, dass diese eingehalten werden. Gewisse Ideen sind mit der aktuellen Rechtslage in Deutschland nicht zu vereinbaren, weil manche Datenquellen eben qua Gesetz nicht miteinander verknüpft werden dürfen.

Auch das Thema der Haftung darf man nicht vernachlässigen: Haftungsfälle können beispielsweise auftreten, wenn Fehlentscheidungen passieren, wegen mangelhafter Datenqualität oder falscher Modelle. Wer haftet in solchen Fällen?

Das spielt nicht nur bei selbstfahrenden Autos eine Rolle, wenn sie die Umweltdaten nicht richtig verarbeiten und dann in Unfälle verwickelt werden. Genauso relevant ist es bei Fehlprognosen in der kommunalen Planung. Daten und Datenanalyse entbinden niemanden von der Verantwortung für seine Entscheidungen auf der Basis dieser Daten.

Nicht nur deshalb bedarf es einer kontinuierlichen Schulung der Beteiligten. Und es schließen sich weitere Fragen an: Welche neuen Bedarfe entstehen an die Kommunikation und Kooperation verschiedener kommunaler Einrichtungen untereinander und mit verwaltungsfremden Akteuren, etwa im Zuge von Open-Governance-Bestrebungen und übergreifenden Informationsportalen? Braucht es vielleicht sogar Umstrukturierungen?

Wie können die verschiedenen Akteure voneinander lernen? Wie lassen sich Verwaltungsprozesse anpassen? Und welchen Aufwand bedeutet es für das Projektmanagement, für interne sowie externe Kommunikation, die Akzeptanz der neuen Daten wie auch der mit ihrer Hilfe getroffenen Entscheidungen sicherzustellen – insbesondere, wenn Ergebnisse der datenbasierten Entscheidungsfindung der bisherigen Praxis widersprechen?

Können wir uns dann einfach auf die Algorithmen berufen?

Algorithmen sind heute das, was im 20. Jahrhundert chemische Formeln, Bau- und Produktionspläne waren. Sie bestimmen, wie genau in einem spezifischen Anwendungsfall Wert aus Daten extrahiert werden kann.

Algorithmen entstehen, bildlich gesprochen, in Daten-Innovationslaboren und werden später in Datenfabriken für die tägliche Produktion von Planungs-, Steuerungs- und Entscheidungsergebnissen eingesetzt:
Welche Ampeln sollen wann auf Grün schalten, damit der Verkehr in der Stadt möglichst störungsfrei fließen kann? Wenn Bürger über ein Mängelportal Straßenschäden melden, welche sollen dann mit hoher Priorität repariert werden, welche mit nachgelagerter? Wann werden verstärkt U-Bahnen und Sicherheitskräfte eingesetzt, damit die Besucherströme zum Oktoberfest optimal gelenkt werden? Wie lassen sich Lebenszyklen von Quartieren in der Planung so abbilden, dass Angebot an und Nachfrage nach öffentlichen Einrichtungen nachhaltig und dynamisch aufeinander abgestimmt sind?

Damit man solche Fragestellungen identifizieren und mithilfe von Daten beantworten kann, braucht es nicht nur kompetente Mitarbeiterinnen und Mitarbeiter, sondern auch geeignete Organisationsmodelle.

Damit aus dem Daten-Öl Mobilität für alle wird, braucht es aber noch verlässliche Verkehrsregeln: Es geht um informationelle Selbstbestimmung und digitale Glaubwürdigkeit.

Der digitale Wandel hat erhebliche Auswirkungen auf die Zivilgesellschaft, auch und gerade wenn Städte neue Datenquellen nutzen wollen. Für amtliche Daten haben wir Gesetze und Regeln; nicht jedes Unternehmen ist zwar glücklich über die Statistikgesetze, aber zumindest ist klar, wer welche Daten zur Verfügung stellen muss. Im Idealfall ist sogar klar, warum wir alle diese Daten brauchen.

„Digitale Glaubwürdigkeit" ist deswegen unabdingbare Voraussetzung dafür, dass Städte und Kommunen neue Daten erheben und auf ihrer Basis Entscheidungen treffen können. Ohne Glaubwürdigkeit wird es keine Bereitschaft zur digitalen Interaktion zwischen Zivilgesellschaft und behördlicher Organisation geben. Anders gesagt, „Vertrauen ist der Anfang von Allem".

Denken wir nur an den Aufschrei, als die Supermarktkette „Target" potenziell schwangere Kundinnen selektiert hat, um ihnen gezielt Werbung zuzusenden. Oder an das jüngste Patent von Mastercard: „Ein System, ein Verfahren und ein computerlesbares Speichermedium, das so konfiguriert ist, dass es die physische Größe der Zahlungsbegünstigten auf der Grundlage von Zahlungsvorgängen analysiert und es einem Transportanbieter erlaubt, die physische

Größe der Zahlungsbegünstigten bei der Zuteilung eines Sitzplatzes zu berücksichtigen." – Da werden aus den Einkäufen eines Kunden also Größe und Gewicht errechnet und an eine Fluglinie weitergegeben, die ihn dann womöglich nicht mitfliegen lässt. Stellen wir uns das mal für die Berliner U-Bahn vor – undenkbar!

„Achtung vor informationeller Selbstbestimmung" muss die zugrundeliegende Leitlinie sein: Es gibt kein allgemeines Recht des Staates auf die Daten der Bürger. Wer Daten generiert, soll auch das Recht auf Einsicht und Verwendung dieser Daten haben, da ein Recht auf Eigentum an den eigenen Daten besteht.

Vielleicht gibt es einen spezifischen Anspruch der Öffentlichkeit auf spezifische Daten des Individuums, ähnlich wie der Staat einen gewissen Anspruch auf das Einkommen der Bürger in Form von Steuern und Gebühren erhebt. Diesen Gedanken lohnt es sich weiterzudenken, gerade wenn wir Daten als werthaltiges Gut betrachten.

Konkret könnten wir folgende fünf Überlegungen anstellen:

Erstens stellt sich die Frage nach dem Anspruch auf und dem Ausgleich für Daten.

Auf welche Daten kann die öffentliche Hand aus berechtigtem Grund Anspruch erheben und für welche Daten muss sie eine entsprechende Gegenleistung erbringen? Was können wir aus Statistikgesetzen lernen, können neue Technologien hier vielleicht sogar entlasten, weil sie die Datenbereitstellung vereinfachen? Welche positiven Effekte ergeben sich aus der Partizipation, weil Bürger erleben, dass ihre Daten wichtig sind und zu Veränderungen führen?

Zweitens geht es um die Balance zwischen Transparenz und Diskretion.

Wem wird welcher Einblick in die Daten gewährt und gegebenenfalls zu welchem Preis? Wie viel Aufwand muss die öffentliche Hand selbst tragen, wie viel muss der Bürger zahlen, wie geht man mit weiteren Anspruchsgruppen um? Gibt es berechtigte Gründe, Einblicke zu verweigern, weil das Gemeinwohl und die Interessen des Individuums in Konflikt geraten? Welche positiven Effekte entstehen, weil Bürger sich besser informiert fühlen?

Drittens sollten wir über Aufbewahrungs- und Verfallsfristen nachdenken.

Wie lange dürfen personenbezogene Daten gespeichert werden, und dürfen sie danach in anonymisierter Form weiter genutzt werden? Wie lange müssen umgekehrt Daten gespeichert werden, damit Auskunftsrechte gewahrt werden können, und wer trägt dafür die Kosten? Welche positiven Effekte entstehen, weil Behörden besser vernetzt sind und mehr Bürgerservice möglich wird?

Wer ist viertens zuständig für die Lösung von Konflikten?

Wie können wir die Interessen des Bürgers sicherstellen, wenn es dafür noch keine allgemeinen Regeln gibt? Braucht es Ombudspersonen? Wie können wir die Interessen insbesondere derjenigen berücksichtigen, die sich mangels Wissen oder Ressourcen nicht aktiv informieren oder klagen, wenn ihre Rechte verletzt werden? Welche positiven Effekte entstehen, weil die öffentliche Hand sich aktiv um die Rechte der Bürger kümmert – anders als Google, Facebook & Co.?

Was ist fünftens und letztens eigentlich mit der Ressourcenschonung?

Einerseits bezieht sich Ressourcenschonung auf die Frage, ob Digitalisierung von Prozessen per se zu Einsparungen gegenüber bisherigen analogen Verfahren führt. Andererseits bezieht sich Ressourcenschonung auf den Aspekt der Datensparsamkeit. Sollte es generell zulässig sein, Daten zu erheben, auch wenn sie im konkreten Anwendungsfall nicht benötigt, aber später genutzt werden können, wenn dadurch Ressourcen (also letztlich Steuergelder) gespart werden können?

Neue Daten: ja, aber nicht um jeden Preis.

Die öffentliche Hand – Städte, Kommunen, Behörden – steht vor demselben Problem wie die Wirtschaft. Auf der einen Seite muss sie eine Vielfalt von Datenquellen zusammenführen, um aus Big Data Smart Data oder vielleicht sogar Open Data zu erzeugen – dabei muss sie unter anderem technische, datenschutzrechtliche und lizenzrechtliche Aufgaben bewältigen.

Auf der anderen Seite steht der Wunsch nach besserer Planung und Steuerung, nach einer effizienteren Durchführung von Verwaltungsprozessen und vielleicht sogar nach neuen Datenprodukten, die man weiteren Akteuren zur Verfügung stellen kann.

Eins ist klar: Es ist nicht damit getan, Data Scientists einzukaufen und möglichst viele Daten aus möglichst vielen innovativen Quellen in eine Cloud zu packen.

Die größte Herausforderung ist eben auch keine technische, sondern eine organisationspolitische. Es ist die Herausforderung, die vielen verschiedenen Akteure dazu zu bewegen, dass sie ihre Daten teilen, um mehr Wert für alle zu schaffen und diesen Wert auch sichtbar zu machen.

Dazu gehört aber auch das Verständnis, was neue Daten, insbesondere Big Data und Data Analytics leisten können – und was nicht. Dazu gehört es zu begreifen, wie unschätzbar wertvoll etablierte Prozesse zur Bereinigung und Qualitätssicherung von Daten sind – auch wenn sie viel Zeit und Geld kosten.

Und sich zu vergegenwärtigen, welche Bedeutung darin liegt, dass solche Prozesse transparent nachvollziehbar sind: Wir wissen, was in amtlichen Daten steckt und was nicht. Wir wissen, wie wir sie interpretieren dürfen und wie nicht – und im Zweifelsfall

gibt es kompetente Ansprechpartner in den Statistischen Ämtern, die wir fragen können.

All dies müssen wir noch lernen, wenn es um Big Data geht. Meine Hoffnung ist, dass Sie, liebe Anwesende, hier einen wertvollen Beitrag leisten können.

Denn Big Data funktioniert nicht ohne Statistical Literacy.

Am Ende wird trotz aller Euphorie längst nicht alles exakt prognostizierbar sein. Unsicherheiten bleiben selbst mit der leistungsfähigsten Hard- und Software, den größten Datenmengen und den komplexesten Algorithmen. Entscheidungen können durch Daten unterstützt werden, aber sie lassen sich in den seltensten Fällen vollständig berechnen.

Deswegen müssen wir auch zukünftig Mut zu Entscheidungen haben, wenn Daten keine eindeutige Antwort geben. Entscheidungskompetenz ist nicht die Fähigkeit, eine Illusion von „sicherem Wissen" zu schaffen – jedenfalls nicht im gewohnten Sinn. Vielmehr könnte eine neue Definition lauten: „Sicheres Wissen" entsteht aus „unsicherem Wissen" und dem „Wissen über das Ausmaß der Unsicherheit".

Entscheidungskompetenz ist die Fähigkeit, mit Unsicherheit umzugehen. Und nichts anderes ist Statistical Literacy. Davon brauchen wir zukünftig eher mehr als weniger. Und dafür stehen Sie – hoffentlich noch viele weitere Jahrzehnte lang.

Katharina Schüller gründete 2003 das Unternehmen STAT-UP Statistical Consulting & Data Science in München, das mit Niederlassungen in Madrid und London europaweit für Unternehmen, Forschungsinstitute und die öffentliche Hand tätig ist. Der Öffentlichkeit bekannt ist sie durch regelmäßige Radio- und Fernsehbeiträge sowie Fach- und populärwissenschaftliche Publikationen. Sie ist Lehrbeauftragte an verschiedenen Hochschulen und als Expertin für Digitalisierung und Data Analytics zudem Mitglied des Beirats der Deutschen Bank und des Beirats von Burda Forward.