Interview mit David Kriesel

In Daten stecken viel mehr Dinge, als zunächst erwartet ...

David Kriesel spricht mit OBJEKTspektrum über von ihm entdeckte Datenungewöhnlichkeit wie den Xerox-Kopierfehler, der schon mal aus einer 6 eine 8 macht, Bahn- und SpiegelMining.

David Kriesel, schön, dass du dir Zeit genommen hast. Am heutigen Eröffnungsabend der OOP-Konferenz in München hältst du einen Vortrag. Ich bin gespannt.

... zum Thema SpiegelMining. Ein Hobbyprojekt, in dem ich die digitalen Inhalte von Spiegel.de über einen Zeitraum von ein paar Jahren kontinuierlich runtergeladen habe.

Zum Spiegel kommen wir gleich zurück. Was ich tatsächlich als Erstes von dir gesehen habe, ist der Xerox-Vortrag auf dem Chaos Communication Congress (CCC) 2014. Bei dem hast du gezeigt, dass Xerox-Kopierer etwas Merkwürdiges machen.

Das stimmt. Das war in der Tat ein Bug von Xerox. Diese Kopierer vertauschen manchmal Charaktere, also Buchstaben und Zahlen. Und zwar so, dass man das zunächst nicht sieht, denn der neue Buchstabe, die neue Zahl wird perfekt in das Ursprungslayout eingebaut, sodass das Dokument auf den ersten Blick normal aussieht. Das passiert beim Scannen.

Das heißt, wenn ich beispielsweise eine Xerox-Kopie unseres Jahresabschlusses von unseren Steuerberatern unterschreibe, so könnte es sein, dass ich eine Bilanz unterschreibe, in der Buchstaben und Zahlen vertauscht wurden?

Genau, aber ich muss das technisch präzisieren: Das passiert beim Scannen in PDFs. Ich habe das Verhalten nicht nachvollziehen können, wenn man eine reine Kopie macht. Da PDFs aber gerne hin und her gemailt werden und man sie ewig aufheben und Ausdrucke davon machen kann, ist das vielleicht noch schlimmer. Die fehlerhaften PDFs gammeln dann irgendwo in Archiven rum.

„Millionen Views für einen Kopierer-Vortrag, das muss man sich mal vorstellen"

Das war natürlich damals der Lacher auf dem Congress. Bei dem Wort „Fotokopie" denke ich, ich mache nur ein Foto. Und warum ist das nicht so?

Es stellte sich heraus, dass es ein Bug in der Bildkompression war. Die Xerox-Kopierer zerlegen das Bild in einzelne Segmente, in die einzelnen Buchstaben, und speichern diese dann. Dabei werden die Symbole für Buchstaben nur jeweils einmal gespeichert. So wird aus einem großen Bild eine Speicherplatz-sparende Kette von Referenzen auf Buchstaben. Wenn das Bild dann aus den Buchstaben rekonstruiert wird, funktioniert das solange gut, wie die Ähnlichkeitsmetrik im Gerät akkurat ist. Und die arbeiten eigentlich immer unscharf. Wenn eine 6 für eine 6 gehalten wird, ist das gut. Wenn eine 6 für eine 8 gehalten wird und dann durch die 8 ersetzt wird, ist das schlecht. Und das war der Bug. Das Kompressionsformat heißt übrigens JBIG2.

Wie bist du denn darauf gekommen?

Zufall. Ich habe im Studium in Bonn als IT-Dienstleister gejobbt und eine der Firmen, für die ich gearbeitet habe, hat mich irgendwann angerufen und gesagt: „Komm mal vorbei, unser Kopierer wirft einfach falsche Zahlen aus". Ich habe natürlich gedacht, das wäre ein Scherz, bin dann aber vorbeigefahren. Es war kein Scherz. Und dann bin ich der Sache nachgegangen.

Das heißt, du warst studentischer Turnschuh-Admin.

Genau. Dann war ich ein Supportfall bei Xerox, der nicht beachtet wurde. Und irgendwann habe ich gedacht, jetzt ziehe ich da mal ein bisschen die Zügel an. Also ein Bug in so einer Größenordnung ..., da schwante mir, dass wir nicht die einzigen Betroffenen sind. Xerox ist ja ein Welt-Hersteller.

War deine erste große Veröffentlichung dann gleich dieser Vortrag auf dem 31. Chaos Communication Congress 2014?

Nein, das Ganze spielte sich 1,5 Jahre vorher ab. Zunächst habe ich mich mit Xerox auseinandergesetzt. Ich habe ein oder zwei Wochen nicht-öffentlich versucht, dem Problem beizukommen. Mit dem Distributor vor Ort, mit Xerox selbst. Dann passierte aber ewig nichts. Und schließlich habe ich eine Warnung auf meinem kleinen IT-Blog geschrieben, die dann um den Planeten ging. Diese Reaktion überraschte mich damals ...

Dann bist du sozusagen über Nacht ins Rampenlicht der Hackerszene getreten.

Sozusagen ja, aber das war ja kein Hack, sondern ein IT-Sicherheits-Thema, nur aus einem ungewohnten Angriffsvektor. Eineinhalb Jahre später habe ich dann beim CCC-Kongress den Kopierer-Vortrag gehalten.

Der dann sofort einer der viel zitierten Vorträge vom Kongress geworden ist.

Das stimmt. Er hat vielen Zuhörern Spaß gemacht. Ich gestaltete den Vortrag lustig, um auf spielerische Art Leuten zu zeigen, wie man sich Schritt für Schritt mit einem Gegner auseinandersetzt, der viel größer ist als man selbst, und dabei nach Möglichkeit überlebt. Das „Überleben" ist ja auch ein wichtiger Aspekt! Ich stand 2013 am Anfang meines Berufslebens und ich dachte, jetzt mache ich mich verklagbar auf soundso viele Milliarden Börsenwert. Aus heutiger Sicht klingt das übertrieben, aber vielleicht hast du auch beim allerersten Mal bei etwas Größerem ein bisschen Gänsehaut bekommen.

Hat das irgendwas bewirkt? Hat dein Vortrag ein Bewusstsein dafür erzeugt, dass Computer eben doch nicht perfekt sind?

Ich bekomme heute noch E-Mails wegen dieser einen Sache. Die werde ich vermutlich im Leben nicht mehr los. Insofern würde ich sagen, ich habe schon etwas bewirkt. Der Vortrag wird immer noch durchs Netz gereicht und hat mittlerweile Millionen Views. Millionen Views für einen Kopierer-Vortrag, das muss man sich mal vorstellen. Ich glaube, Xerox wird so ein Fehler auch nicht wieder passieren.

Wie reagiert eine so große Firma, wenn so ein Fehler öffentlich geworden ist?

Also erst mal ewig gar nicht. Dann hat mich zuerst die PR-Abteilung von Xerox-Deutschland angerufen. Die haben das zunächst als PR-Problem gesehen, bis irgendwann die ganze US-Presse eingestiegen ist. In den USA ist das medial, glaube ich, deutlich teurer als hier. Und dann hatte ich plötzlich deren Geschäftsspartenleiter an der Strippe, der das gerne aufklären wollte.

Hat Xerox versucht zu verhindern, dass du das weiter öffentlich machst?

Nein, gar nicht. Zu dem Zeitpunkt, als die mich angerufen haben, war das in fast jedem Massenmedium weltweit verbreitet. Bei uns war es im Spiegel, bei den Amerikanern war es in allen großen Medien. Der ganze asiatische Raum war auch dabei. Da sehe ich dann nur immer meinen Namen in für mich unlesbaren Zeichen. Da gab es nichts mehr zu verhindern. Insofern war die Stimmung entspannt.

Das war ja nicht die einzige Datenungewöhnlichkeit, der du auf die Schliche gekommen bist. Was ich zuletzt von dir gehört habe, war dein Vortrag beim letzten CCC zur Bahn-Pünktlichkeitsstatistik.

Ja genau, da habe ich ein paar Bahn-Daten runtergeladen und mal reingeguckt.

Du hast, glaube ich, fast das ganze Jahr Bahn-Daten runtergezogen mit einer kleinen Lücke drin und hast analysiert und geguckt, ob man daraus etwas lernen kann.

Genau, so funktionieren solche Hobbyprojekte von mir.

„Ausgefallene Züge fließen nicht in die Pünktlichkeitsstatistik ein"

Die Bahn hat dir die Daten dann auf Nachfrage zur Verfügung gestellt, die Schnittstelle ist nicht geschützt und öffentlich …

Genau. Ich habe die Erlaubnis erhalten, die heruntergeladenen Daten auch öffentlich zu verwerten. Da war ich mir anfangs nicht sicher. Vermutlich hätte ich das einfach ohne Nachfrage machen können, aber bei Sachen in dieser Größenordnung gehe ich lieber auf Nummer sicher. Ich habe eine E-Mail an die bei der Bahn angegebene Adresse geschickt. Daraufhin wurde mir der Datenzugriff einfach genehmigt.

Dann hast du mit Akribie und Leidenschaft angefangen, diese Daten zu analysieren, und hast Informationen und Zusammenhänge über die tatsächliche Pünktlichkeit der Deutschen Bahn herausgefunden.

Die Stichproben, mit denen ich arbeitete, haben erst mal ergeben, dass die Bahn-Daten in Bezug auf die Pünktlichkeit korrekt zu sein scheinen. Ich fand keine Anhaltspunkte dafür, dass sie Verspätungsminuten bei Zügen abziehen oder Ähnliches. Was mich allerdings dann ein bisschen überraschte, ist, dass ausgefallene Züge nicht in die Pünktlichkeitsstatistik einfließen. Von da kann man dann dem Kaninchenloch folgen, immer weiter,

David Kriesel
Alter: 36 Jahre
Spezialist für Data Science und Machine Learning
Mission: Menschen für Datenspuren begeistern
Meine Hoffnung für die Zukunft:
„In Zukunft führen wir Diskurse weniger anekdotisch, sondern auf Basis von Daten"

immer weiter, und sich fragen, was bringt dir das denn, wenn du ausgefallene Züge rausstreichst?

Kann man denn einen ausgefallenen Zug überhaupt in eine Pünktlichkeitsstatistik mit einbeziehen? Der hat ja dann eine unendliche Verspätung.

Ja, du kannst ihm keine Pünktlichkeits-, keine Verspätungsminute mehr zuweisen, aber die Deutsche Bahn hat ein binäres Kriterium für Pünktlichkeit. Die sagen „war der Zug pünktlich: ja/nein". Wenn er mehr als sechs Minuten zu spät war, heißt es: nein, sonst: ja. Ich würde halt sagen, wer nicht kam, ist: nein. Aber im Moment gilt, wenn ein Zug nicht kommt, dann fällt er aus der Pünktlichkeitsstatistik raus.

Wie veränderte sich die Pünktlichkeitsstatistik, wenn die ausgefallenen Züge berücksichtigt werden?

Zunächst einmal erfüllt die Bahn auch so schon ihren eigenen, Anfang 2019 angesagten KPI-Wunsch von 76,5 Prozent Pünktlichkeit nicht. Das haben sie um locker 1 Prozent unterboten. Wenn man die Ausfälle mit einberechnet, liegen sie sogar nur bei 72 Prozent. Jedes Prozent weniger macht richtig, richtig viel aus. Und wenn du deinen Anschluss nicht mehr bekommst, verschlechtert sich die Reisebequemlichkeit. Man muss aber auch sagen:

Bonner Altstadt: Kirschblütentraum, Foto: David Kriesel

Es ist nicht ganz klar, ob und wenn ja welche Ersatzzüge in den Daten drinstehen. Bis heute gibt es allerdings keinen Protest der Bahn.

Bei deinem Bahn-Vortrag auf dem CCC-Kongress 2019 ist auch eine Menge Verständnis für die Bahn mitgeschwungen, auch wenn es bei deinen Darstellungen oft Lacher gab ...

Das mache ich aber eigentlich immer. Ich stelle mir vor, wie es wäre, selbst in so einer Situation zu sein. Sei es die Bahn, bei der das Deployment nun mal nicht mit drei Skripten regelbar ist. Das geht eben nicht in drei Minuten. Oder wie viele Interviews Xerox zu seinen Kopierern gegeben hat. Ich glaube, die Stimmung war bei meinen drei CCC-Vorträgen vergleichbar und auch der Humor darin. Die Bahn hat sogar Applaus bekommen. Mein Eindruck war, dass die Bahn bei all ihren Fehlern beim Publikum (und auch bei mir) durchaus beliebt ist.

Christa Koenen hatten wir auch in der Interviewreihe. Als DB-CIO betonte sie, dass es eine Herkulesaufgabe ist, so einen Technologiekonzern zu innovieren ...

... mit Equipment wortwörtlich aus Kaiser-Wilhelm-Zeiten.

Wie hatte sich denn der Bahn-Konzern verhalten, als dieser Vortrag rausgekommen ist?

Ich bekam inoffiziell eine Menge positives Feedback. Offiziell hat die Bahn sich dazu nicht geäußert. Sie haben das allerdings intern über ihren Newsletter an alle Angestellten weitergeschickt. Nun habe ich ein paar Einladungen der Bahn erhalten und fahre im Sommer mal hin, sofern Corona mich lässt. Mit der Bahn übrigens.

Sehr viel Zuschauer haben auf dem CCC-Kongress deinen BahnMining-Vortrag gesehen, oder?

Ja, das hat mich sehr beeindruckt. Wir hatten einen Saal für ca. 5.000 Zuschauer. Dann haben sie sich live per Handy noch mal eine Tribüne Stehplätze nachgenehmigen lassen und es gab eine Live-Übertragung. Waren also sehr viele ...

Das heißt, so ungefähr jeder dritte Teilnehmer des CCC hat den Vortrag live gesehen?

Ja, live im oder am Saal. Der Livestream ist am Schluss auch noch zusammengebrochen, den haben wir überlastet. Deshalb sind wahrscheinlich noch einige auf YouTube ausgewichen. Da war ich schon etwas überwältigt.

„Ich möchte Leuten nahebringen, wie man Daten sammelt"

Und dann gibt es deine statistische Analyse der SpiegelOnline-Veröffentlichungen von 2016 ...

Ich habe immer eine Hintergrundmessage bei meinen Vorträgen. Ich möchte Leuten nahebringen, wie man Daten sammelt, wie man das auch zuhause tun kann und wie aus der Analyse ein systematisches Bild entstehen kann. Das habe ich zuerst mit dem Spiegel und mit der Bahn so gemacht, aber auch mit vielen anderen Sachen. Nicht über alle halte ich Vorträge. Beim Spiegel habe ich alle paar Minuten die Seite gepollt und dann alles abgespeichert. Dazu habe ich Skripte geschrieben, die dann diese Datenmengen parsen. Ich habe relativ leistungsstarke Desktop-Rechner zuhause, mit denen das geht. So entsteht ein feiner Datensatz, den man auswerten kann.

Für wie lange hast du das gemacht?

Seit 2014, und ich glaube, ich habe vergessen, das abzustellen (schmunzelt). Die haben beim Spiegel aber jetzt gerade ihr Interface geändert. Jetzt muss ich gucken, ob das noch funktioniert. Aber was ich garantieren kann, ist, dass ich die Daten bis Ende 2018 habe. Den Datensatz danach habe ich bislang nicht angeguckt.

„Es ist möglich, ein soziales Netz der Spiegel-Onliner-Redakteure zu erstellen"

Und Daten bedeutet, du hast die Webseiten geparst?

Genau. Der Spiegel hat eine Indexseite, da stehen immer alle Artikel-URLs drauf und daraus komme ich auf die Artikel selbst. Die lade ich dann runter und speichere den kompletten Quellcode ohne Bilder. Da kommen schon ein paar Gigabyte zusammen über die Zeit.

Und dann kannst du dir angucken, wann die arbeiten, wann der Spiegel veröffentlicht und wann die Redakteure Pause machen?

Exakt. Du kannst gucken, welche Sparte oder welcher Autor schreibt am längsten, wer am kürzesten. Wer veröffentlicht wann am liebsten. Wer arbeitet mit wem zusammen. Es ist möglich, ein soziales Netz der Spiegel-Onliner-Redakteure zu erstellen. Man kann sogar nachvollziehen, wer mit wem zeitgleich in Urlaub ist und Mutmaßungen darüber anstellen, wo es möglicherweise Affären gibt. Jeder hatte schon mal Kollegen, die immer gleichzeitig im Urlaub waren, und das lag nicht an den Schulferien. So kann man Leuten sehr gut in einem Vortrag beibringen, was Data Science ist. In Daten stecken viel mehr Dinge, als man es zunächst erwartet. Ich nehme möglichst Datensätze, die man ohne Weiteres einfach verstehen kann. Wenn da überraschende Sachen bei rauskommen, dann erzeugt das bei der Hörerschaft Faszination. Und das macht mir Spaß.

Hat der Spiegel inzwischen etwas über dich veröffentlicht? Hast du dich selbst in deiner Suche wiedergefunden?

Es gab, wie schon gesagt, im Rahmen der Xerox-Saga mehrere Artikel. Dann haben sie Vorträge vom CCC empfohlen und meinen einfach mit. Und jetzt, im Rahmen der Bahn-Sache, haben sie wieder etwas geschrieben und ein Interview gemacht.

Hast du es mal erlebt, dass du aufgrund einer Datenanalyse ein System reparieren konntest, wo der Fehler vorher unklar war?

Ja, oft. Nur ein Beispiel: Ich hatte mal ein Projekt für Flottenleitsysteme für Busse im Nahverkehr. Beim Umstieg auf das System meines damaligen Auftraggebers fielen deutliche Umsatzverluste in einer Region auf. Schlussendlich kam über Anomalien in den Daten heraus, dass einige Busfahrer Geld unterschlugen.

Wäre Corona ein interessantes nächstes Thema? Und falls ja, warum [Anm. d. Red: Diese Frage kam erst während der Schlussredaktion hinzu]?

Zu dem Thema mache ich mir sehr viele Gedanken. Nach 11 Jahren kontinuierlichem Aufschwung und 70 Jahren ohne Krieg ging es uns so gut wie niemals zuvor. Und nun haben wir wieder eine ernsthafte Krise, die von außen über uns hereinbricht und geeignet ist, verkrustete Strukturen aufzubrechen. Ich hoffe, dass wir Diskurse in Zukunft weniger anekdotisch führen, sondern mehr auf Basis von Daten. Und genau das erlebe ich jetzt. Leute gieren nach Infektionsdaten und studieren sie. In den letzten Wochen sind vermutlich vom normalen Bürger mehr Plots mit Exponentialverläufen angeguckt worden als im Jahrzehnt davor. Toll!
Auf der anderen Seite werden Scheinprobleme, die von Politik, Medien und Gesellschaft mangels tatsächlicher Krisen herbeigeredet wurden, vermutlich jetzt einfach unter den Tisch fallen. Neulich habe ich einen interessanten Text gelesen, in dem prophezeit wurde, dass die Leute nach Corona in neugewonnener Eintracht beisammensitzen und sagen: „Political Correctness, worum ging es da noch mal ...?" Vielleicht kehren wir also zu weniger Spaltung und mehr Daten in der Diskussion zurück. Ob die Coronapandemie ein Auswertungsthema für mich ist, habe ich noch nicht entschieden. Es gibt bereits sehr viele gute Sachen, und ich möchte da nur beitragen, wenn ich einen ernsthaften Mehrwert bieten kann. Wenn Corona ein Thema für mich werden sollte, kann der Leser dies unter http://www.dkriesel.com/corona finden.

David, vielen Dank für das Gespräch.

Das Interview führte ...

Dr. Johannes Mainusch
(johannes.mainusch@kommitment.works)
Berater für Unternehmen, die Bedarf im Bereich IT, Architektur und agiles Management haben. Dr. Mainusch ist seit 2012 Mitglied der OBJEKTspektrum-Redaktion.