01 Backup is key

Da sie vergessen hatten den Schlüssel mitzunehmen, mussten sie ein weiteres Jahr physischer Strapazen erleiden.

Ein Forscherteam arbeitete über ein Jahr lang an einer großen Studie im Bereich Physik. Im Laufe des Jahres wurden in langen und zeitaufwändigen Prozessen viele Terabyte an Videodaten von Mikron-Mustern erzeugt. Die Videos wurden systematisch beschriftet und die experimentellen Informationen, die für die Interpretation der Daten in anschließende Analysen notwendig sind, in einer Excel-Datei (Schlüsseldatei) gespeichert. Während die Videodateien in einem langwierigen Prozess in mehreren Instanzen regelmäßig gesichert wurden, war die Sicherung der Schlüsseldatei weit weniger ausgereift.

Als der verantwortliche Doktorand eines Tages seinen Rechner aufräumte, löschte er versehentlich die Datei. Wiederherstellungsversuche schlugen fehl und die Arbeitsgruppe wurde um rund 1 Jahr zurückgeworfen. Ohne den experimentellen Kontext zu den Videodateien war die gesamte Arbeit sinnlos, im wörtlichen Sinne.

Die Story zeigt, dass ein Datenmanagementplan (DMP) von Anfang an sinnvoll ist. Er identifiziert sowohl die zu sichernden Datenmengen als auch ihre Relevanz im Forschungsprozess. Der DMP verhindert somit (hoffentlich), dass zentrale Dateien bei der Sicherung übersehen werden, weil sie im Vergleich zu anderen ein geringes Volumen aufweisen und ihre Sicherung damit vergleichsweise trivial wirkt.

Quellen:

- persönliche Kommunikation

02 Seltsam - Preise haben Geburtstag

Sie konnte ihre Kostenanalyse einfach nicht beenden, da ihr immer wieder Geburtstage dazwischenkamen.

Beim Versuch eine Kostenanalyse zu erstellen, war die zuständige Mitarbeiterin überrascht als in ihrer Tabelle immer wieder abwegige Ergebnisse am Ende ihrer Berechnungen standen. Nach kurzer Recherche war klar, dass beim Import der CSV-Datei für die Nutzungsstatistik elektronischer Medien in Excel aus den Preisen automatisch Datumsangaben gemacht wurden. Nutzungsstatistiken für elektronische Medien werden typischerweise als CSV-Dateien ausgeliefert. Werden diese in den Grundeinstellungen in eine Excel-Datei importiert, werden Preise teilweise in Datumsangaben umgewandelt, obwohl dies nicht gewünscht ist. Das verhindert dann in der Folge eine sinnvolle Interpretation der Nutzungsdaten insgesamt. So lassen sich z.B. keine durchschnittlichen Nutzungskosten mehr berechnen, wenn ein Teil der Preise nicht ausgelesen werden kann.

Das Beispiel zeigt, dass beim Import von Daten in Tabellenkalkulationsprogramme sorgfältig auf die korrekte Formatierung der Zellen geachtet werden muss, da es sonst zu automatischen Formatierungsänderungen kommen kann.

Quellen:

- persönliche Kommunikation

03 Verlorenes Spielzeug

Nur durch die Geburt eines unschuldigen Kindes konnte die Beziehung von Woody und Cowgirl Jessi auf Film festgehalten werden.

Das Unternehmen Pixar konnte nur knapp verhindern einen großen Teil der Daten für den Film Toy Story 2 zu verlieren. Jemand nutzte versehentlich das Kommando "rm *" (für Nicht-Geeks: "remove all" - entferne alles) und löschte damit sämtliche Dateien des Projektes. Unglücklicherweise funktionierte das automatische Backup nicht wie es sollte und die letzte Sicherung lag 2 Monate zurück. Durch einen glücklichen Zufall gab es eine halb-private Arbeitskopie der technischen Direktorin. Diese hatte ein Baby zuhause und hatte den ganzen Film auf ihren Computer transferiert, um im Homeoffice arbeiten zu können. Nach einer sehr vorsichtigen Fahrt zu ihrem Haus und zurück zu Pixar war klar, dass die Daten größtenteils gerettet werden konnten.

Das Beispiel zeigt, dass selbst bei State-of-the-Art Backups Datenverluste auftreten können, wenn mehrere unglückliche Zufälle zusammenkommen. Eine gute Grundlage für das sichere Speichern von Daten bietet die 3-2-1 Regel. Danach sollen Daten an drei verschiedenen Orten auf mindestens 2 verschiedenen Speichermedien gesichert werden, wobei einer der Speicher an einem externen Ort sein sollte. Zudem sollten regelmäßig Tests auf die Effektivität des Backups vorgenommen werden. Praktisch wird das umgesetzt, indem zu zufälligen Zeitpunkten zentrale Dateien aus dem Backup mit den Originalen verglichen werden.

Quellen:

- https://kottke.org/12/05/how-pixar-almost-deleted-toy-story-2
- https://www.techdirt.com/articles/20120514/17243918918/how-toy-story-2-almost-got-deleted-except-that-one-person-made-home-backup.shtml
- https://de.wikipedia.org/wiki/Toy_Story_2

04 The Sound of Silence

Als er nach langer Zeit in seinen Raum zurückkehrte, war keine Musik mehr zu hören.

2019 gab die Plattform MySpace bekannt, dass sie bei der Migration der Daten auf einen neuen Server einen Großteil aller Musik-, Bild- und Videodaten verloren hatte, die zwischen 2003 und 2015 hochgeladen wurde. Nach Angaben von MySpace gab es für die verlorenen Daten kein Backup. Die Onlineplattform war zu Beginn vor allem für Musiker gedacht, um dort ihre Werke zu präsentieren und umfasste mehr als 50 Millionen Musikstücke. Darunter auch frühe Arbeiten von Künstlern, die auf MySpace ihre Karriere gestartet hatten. Insbesondere für ältere Beiträge gab es oft keine lokalen Kopien bei den Nutzern, da sie sich auf den Speicher in der Cloud verlassen hatten. Somit sind die Werke möglicherweise für immer verloren.

Das Beispiel zeigt sehr gut, dass es immer sinnvoll ist Dateien an mehr als einem Ort zu verwahren. Eine gute Grundlage für das sichere Speichern von Daten bietet die 3-2-1 Regel. Danach sollen Daten an drei verschiedenen Orten auf mindestens zwei verschiedenen Speichermedien gesichert werden, wobei einer der Speicher an einem externen Ort sein sollte.

Quellen:

- https://www.heise.de/newsticker/meldung/Datenverlust-Myspace-verliert-riesiges-Musikarchiv-4338737.html
- https://taz.de/Datenverlust-bei-MySpace/!5581108/

05 Aus allen Wolken gefallen

Nachdem seine Kollegen zusammen in die Wolken gestarrt hatten, blicken sie ihn finster an.

Eine Gruppe von Studierenden arbeitete am Ende des Semesters an einem Abschlussbericht für eine Lehr-veranstaltung. Das Dokument wurde gemeinsam in einem iCloud Drive bearbeitet und stand kurz vor der Fertigstellung. Einer der Studierenden löschte den Bericht in dem Glauben, dass es sich dabei um einen alten Entwurf handelte. Zu dieser Zeit gab es keine Funktion zur Wiederherstellung von Daten und die Gruppe musste den Bericht von Grund auf neu schreiben. Es ist nicht überraschend, dass die restlichen Mitglieder der Gruppe in dieser Zeit nicht unbedingt gut auf ihren Kommilitonen zu sprechen waren.

Das Beispiel zeigt sehr gut, dass die Speicherung an nur einem Ort nicht ausreicht, um eine sichere Aufbewahrung der Daten zu gewährleisten. Generell sollten Daten nach der 3-2-1-Regel gesichert werden. Dementsprechend sollte es mindestens 3 Kopien der Daten auf 2 verschiedenen Speichermedien geben, wobei eine davon an einem externen Ort verwahrt werden sollte. Außerdem sollte man bei der gemeinsamen Arbeit in einem zentralen Speicherort besonders umsichtig beim Entfernen von Dateien sein.

Quellen:

- persönliche Kommunikation

06 Gen-Datierung

Die Metaanalysen legen nahe, dass der 2. September für die Funktion von Zellen eine entscheidende Rolle spielt.

Werden Dateien in Tabellenkalkulationssoftware (wie z.B. Microsoft Excel) unter Verwendung der Standardeinstellungen importiert oder eingetragen, werden Einträge in Zellen teilweise automatisch neu formatiert.

Bereits 2004 wurde in einer Studie festgestellt, dass dieser Fehler auch häufig in wissenschaftlichen Veröffentlichungen in Tabellen mit Genbezeichnungen zu finden ist. Die Namen werden dabei entweder in ein Datum oder in Gleitkommazahlen umgewandelt. Diese Änderung der Formatierung ist irreversibel: die ursprüngliche Information zu den betreffenden Genen geht vollständig verloren.

Eine neuere Studie aus dem Jahr 2016 zeigte, dass dieses Problem immer noch aktuell ist und bisher keine standardisierten Lösungen für das Problem bestehen. Etwa 20 % der untersuchten Artikel aus anerkannten Journalen zum Thema Genomik enthielten Fehler in den Namen der Gene in Tabellen. Ein Beispiel ist das SEPT2 (Septin 2) Gen, das eine wichtige Rolle für die Funktion des Zellskeletts spielt und in Tabellen schnell zum 2. September geändert wird.

Da die Daten solcher Studien für die wissenschaftliche Gemeinschaft eine wichtige Ressource darstellen und häufig wiederverwendet werden, ist der Informationsverlust sehr problematisch. Ein Sprecher von Microsoft kommentierte die Ergebnisse der Studie mit dem Hinweis: „Excel is able to display data and text in many different ways. Default settings are intended to work in most-day-to-day scenarios". Dementsprechend ist die Aufzeichnung wissenschaftlicher Daten, wie z.B. Gen-Namen, keine alltägliche Aufgabe für viele Tabellenkalkulationsprogramme und muss entsprechend aufmerksam durchgeführt werden.

Das Beispiel zeigt, dass es bei der Verwendung von Tabellenkalkulationsprogrammen wichtig ist auf die entsprechende Formatierung der Zellen zu achten und die richtige Übertragung der Inhalte sorgfältig zu überprüfen.

Quellen:

- https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-5-80
- https://genomebiology.biomedcentral.com/articles/10.1186/s13059-016-1044-7
- https://www.bbc.com/news/technology-37176926

07 Auf dem falschen Fuß erwischt

Als das Schiff in den Fluten versinkt, betrachten die Zimmermänner sorgenvoll ihre Füße.

Die schwedische Galeone Vasa stellte ein Prestigeprojekt des Schwedischen Königs Gustav II. Adolf dar und zählte zu den größten Kriegsschiffen ihrer Zeit. Am 10. August 1628 versank sie bei ihrer Jungfernfahrt nach nur etwa einem Kilometer auf See. In einer Untersuchung zur Ursache und einem anschließenden Prozess konnte zwar festgestellt werden, dass das Schiff über eine unzureichende Stabilität und eine zu geringe Breite verfügte, aber es konnte kein Hauptschuldiger verurteilt werden.

Archäologen haben mittlerweile herausgefunden, dass die Verwendung unterschiedlicher Längenmaße zu dem Unglück beigetragen hat. Zwar wurde bei der Planung des Schiffes alles in Fuß angegeben, aber diese Angabe war zu jener Zeit nicht standardisiert. So verwendete eine Gruppe von Zimmermännern Maßbänder in „Schwedischen Fuß“ während die andere „Amsterdam Fuß“ benutzte. Die beiden Maße unterscheiden sich zwar nur um einen Zoll, aber auf die Gesamtlänge von 69m ergeben sich dabei deutliche Ungenauigkeiten.

Das Beispiel zeigt sehr gut, wie wichtig die Verwendung von gemeinsamen und gut definierten Standards für die erfolgreiche Durchführung von Projekten sowie die Vergleichbarkeit und Nachvollziehbarkeit ist.

Quellen:

- https://www.pri.org/stories/2012-02-23/new-clues-emerge-centuries-old-swedish-shipwreck
- https://www.instmc.org/Festival/Why-Measurement-Matters
- https://de.wikipedia.org/wiki/Vasa_(Schiff)

08 Babylonische Verwirrung

Haben Sie einen japanischen Namen? Fürchten Sie um Ihre Reputation.

Am 1. Januar 2020 rollte eine kleine lexikalische Revolution durch Japan. Eine neue Verordnung bestimmt, dass offizielle Dokumente die Reihenfolge der Namen des japanischen Volkes umkehren sollten, wenn sie im lateinischen Alphabet gerendert werden. Bisher wurden beispielsweise in englischen Dokumenten japanische Namen mit dem Vornamen zuerst geschrieben und somit die westliche Praxis verwendet. Von nun an wird der Familienname an erster Stelle stehen und, um jede Mehrdeutigkeit zu verbannen, vollständig kapitalisiert. Damit sind die japanischen Namenskonventionen dann für japanische und fremde Zeichen identisch, zum Preis eines Bruchs mit der bis dahin gültigen Transkriptionspraxis. Die Folge im akademischen Leben könnte eine höhere Zahl von übersehenen Zitaten und damit ein relativer Reputationsverlust japanischer Autoren sein bis sich die Anbieter der entsprechenden Metriken und, noch wichtiger, die Zitierenden an die neue Konvention gewöhnt haben.

Das Beispiel zeigt, dass es immer günstig ist über einen persistenten Identifikator (PID) zu verfügen, um derartige Probleme zu umgehen. Für Personen gibt es hier zum Beispiel die unabhängig vergebene ORCID oder die über Thompson Reuters vergebene ResearcherID. Werden diese PIDs zitiert, dann erfolgt unabhängig von geänderten Namenskonventionen eine korrekte Zuordnung des Autors zu den Nennungen seiner Artikel. Für den Zitierenden ergibt sich der Vorteil, dass er die Beiträge von Autoren trotz Namens- oder Konventionsänderungen im Blick behalten kann. Für die Autoren selbst ist es ein klares Plus eine direktere Kontrolle über die korrekte Zuordnung von Artikeln und damit über die wissenschaftliche Reputation zu haben.

Quellen:

- https://www.economist.com/asia/2020/01/02/why-japanese-names-have-flipped
- https://www.forschungsdaten.info/themen/bewahren-und-nachnutzen/persistente-identifikatoren/
- https://orcid.org/about
- http://www.researcherid.com/?returnCode=ROUTER.Unauthorized&Init=Yes&SrcApp=CR#rid-for-researchers

09 Die vergessene Fantasie?

Trotz 20 Jahren Warten auf das Spiel, konnten nicht alle Fantasien erfüllt werden.

Die von SquareEnix (damals noch Squaresoft) entwickelten Playstation 1 RPG-Spiele Final Fantasy VII, VIII und IX gelten als die „Goldene Ära" der Final Fantasy Geschichte und noch bis heute müssen sich viele neue Spiele an diesen messen. Dennoch hat sich SquareEnix lange Zeit darüber ausgeschwiegen, warum gerade der achte Teil der Serie nie ordentlich für PC oder eine aktuelle Konsolengeneration aufgesetzt wurde – obwohl dies mit fast jeden anderen Teil der Serie gemacht wurde.

Die Antwort darauf war leicht und kam durch mehrere Interviews zum Vorschein: Man hatte sowohl die Rohdaten (u.a. Hintergrundbilder, Musik, 3D-Modelle) als auch den fertigen Quellcode des Spieles einfach nicht archiviert. In den 90er Jahren gab es weniger AAA-Titel (Großprojekte, die über mehrere Jahre liefen) und Spiele wurden in geringeren Abständen produziert. Dies führte wie bei SquareEnix dazu, dass für neue Projekte immer wieder Platz geschaffen werden musste und man alten Projektdaten wenig Aufmerksamkeit schenkte. So kam es, dass sämtliche Projektdaten von Final Fantasy VIII verloren gegangen sind und erst 2019 in Zusammenarbeit mit anderen Firmen eine neue Version des Spiels veröffentlicht werden konnte.

Obwohl es jetzt viele „Remastered" Versionen der damaligen Final Fantasy Spiele gibt, die meist neu nachprogrammiert wurden, ist das Fehlen der Rohdaten für die Entwickler und die Spielergemeinschaft immer noch ein Problem. So wurden z.B. die Hintergrundgrafiken der Spiele in einer hohen Auflösung erstellt, aber nur die komprimierten Versionen für die damaligen Konsolen aufbewahrt, die nach heutiger HD-Darstellung der Bildschirme nicht mehr zeitgemäß sind.

An diesem Beispiel kann man gut erkennen, wie wichtig es ist sowohl die Rohdaten als auch die eigentlichen Projektdaten aufzubewahren. Die Veröffentlichung oder der Abschluss eines Projekts, sollte nicht gleichzeitig das Ende der entstandenen Daten bedeuten. Erst durch das richtige Organisieren der Roh- und Ergebnisdaten können Folgeprojekte die Daten weiterverwenden und die Arbeit wird mehr gewürdigt.

Quellen:

- https://www.vg247.com/2018/09/14/isnt-ps4-xbox-switch-port-final-fantasy-8-preservation-may-answer/
- https://www.vg247.com/2019/01/09/final-fantasy-7-hd-remaster-remake-upscale/
- https://ffviiiremastered.square-enix-games.com/

10 Hallo? Bist du noch da?

Es hätte Routine sein sollen und nun konnte niemand mehr mit den anderen reden.

2009 war T-Mobile mit über 40 Millionen Kunden der größte Handy-Netz-Anbieter in Deutschland. Dennoch ist am 21.04. gegen 16 Uhr dem Anbieter eine Panne geschehen, die bis dato als größte in die Geschichte eingehen sollte. Mit einem Schlag konnte sich kein Kunde mehr mit dem Netz verbinden. Weder Anrufe konnten verbunden noch SMS verschickt werden. Grund dafür war ein gleichzeitiger Ausfall aller drei Home Location Register (zu Deutsch: Verzeichnis des Heimatortes). Diese drei Server stellen zusammen eine verteilte Datenbank dar und sind ein zentraler Bestandteil eines jeden Mobilfunknetzes. Im Normalfall könnte das Netz weiterhin aufrecht erhalten bleiben, solange nur einer der drei Server noch aktiv ist. Doch wie kam es dazu, dass alle drei Server plötzlich abgestürzt sind?

Die Antwort wurde wenige Tage später über die Presse öffentlich gemacht. Auf allen drei Servern wurde gleichzeitig ein fehlerhaftes Software-Update aufgespielt. Deswegen konnten sich die Server nicht gegenseitig stützen, da jeder mit dem gleichen Problem zu kämpfen hatte. Erst gegen 20 Uhr desselben Tages, konnte man das Software-Update bereinigen und einen Großteil des Netzes wieder zum Laufen bekommen.

Da neue Software manchmal unerwartet reagieren kann, sollte sie niemals auf alle kritischen Punkte eines Systems gleichzeitig aufgespielt werden. Zu empfehlen ist ein schrittweises Vorgehen, idealerweise mit einem Probelauf in einer Testumgebung.

Quellen:

- https://www.thelocal.de/20090422/18791
- https://www.news-on-tour.de/13160/mobilfunk-t-mobile-netz-wieder-neu-hoch-gefahren-softwarefehler-im-sogenannten-home-location-register-hlr-gefunden/
- https://www.computerwoche.de/a/groesste-panne-im-t-mobile-handynetz-wird-untersucht,1893599
- https://www.deutschlandfunk.de/kein-netz-nach-vier.676.de.html?dram:article_id=26367

11 Atlantischer Lazarus

Sein Mitgefühl wurde positiv wahrgenommen, trotzdem kam es den Forschenden falsch vor.

Eine Forschergruppe wollte die Funktion ihres Gerätes für funktionelle Magnetresonanztomographie (fMRI) testen und suchte dafür nach Objekten mit viel Kontrast und verschiedenen Texturen. Nachdem ein Kürbis und ein totes Huhn ihre Erwartungen nicht vollständig erfüllen konnten, testeten sie einen toten atlantischen Lachs. Diesem wurden anschließend Bilder von sozialen Situationen gezeigt und seine Reaktionen aufgezeichnet.

Die Daten sollten für eine Lehrveranstaltung verwendet werden, um die Auswertung von fMRI-Daten und mögliche Fehlerquellen an einem absurden Beispiel zu demonstrieren. Die Forschenden waren doch sehr überrascht als sie bei dem toten Lachs plötzlich eine Reaktion auf die gezeigten Bilder in Gehirn und Wirbelsäule feststellten. Ein wichtiger Schritt bei der Datenauswertung war die Korrektur für mögliche falsch-positive Ergebnisse. Ohne diese Korrektur würden die gemessenen Werte fälschlicherweise als signifikante Veränderungen in der Gehirnaktivität des toten Tieres interpretiert. Zum Zeitpunkt der Veröffentlichung der Studie basierten viele Publikationen zu fMRI-Daten auf Analysen ohne die entsprechenden Korrekturen und die Ergebnisse sorgten daher für viel Aufsehen.

Das Beispiel zeigt, dass bei Experimenten nicht nur die Kalibrierung von Messinstrumenten eine entscheidende Rolle spielt, sondern auch die korrekte Auswertung der gewonnen Daten einschließlich angemessener Kontrollen und Korrekturen.

Quellen:

- https://teenspecies.github.io/pdfs/NeuralCorrelates.pdf
- https://blogs.scientificamerican.com/scicurious-brain/ignobel-prize-in-neuroscience-the-dead-salmon-study/

12 Familienbande

Hätte sie die Geburtstage nicht in Ordnung bringen wollen, hätten sich die Geschwister nicht aufgeregt.

Eine Studie im Jahr 2003 wollte die Kontaktdaten zufällig ausgewählter Familienmitglieder der Hauptbefragten, zu denen auch Geschwister zählten, erheben. Die Kontaktdaten der Geschwister wurden in einer Excel-Tabelle aufbewahrt, um später Adressaufkleber für den Fragebogenversand zu fertigen. Der Plan war, die ausgewählten Geschwister darüber zu informieren, dass ihr Bruder oder ihre Schwester mit dem Geburtsdatum TT/MM/JJJJ den Forschenden die Kontaktdaten überlassen hatte.

Einer der Forschenden wollte die Daten kurz vor Druck der Adressaufkleber nach Geburtsdatum sortieren. Das war zumindest die Absicht. Tatsächlich wurde in der Tabelle ausschließlich die Spalte mit den Geburtsdaten sortiert. In der Folge bekamen tausende Personen einen Brief mit der Nachricht, dass sie von einem Bruder oder einer Schwester mit einem völlig falschen Geburtsdatum für die Studie registriert wurden. Dies führte zu sehr vielen wütenden und aufgeregten Anrufen und E-Mails. Viele der so kontaktierten Personen dachten, sie hätten ein Geschwisterkind, von dem sie bisher nichts wussten, oder dass ihr Vater vielleicht ein heimliches Doppelleben mit einer anderen Familie führe. Die Forschenden mussten anschließend die Daten manuell aus den originalen Papier-Fragebögen rekonstruieren und diese zusammen mit einem Entschuldigungsbrief erneut versenden.

Diese Geschichte zeigt, dass Master-Dateien niemals überschrieben werden sollten und dass beim Umgang mit persönlichen Daten besondere Vorsicht geboten ist.

Quellen:

- https://www.lcrdm.nl/horror-family-stress

13 Haushaltsangelegenheiten

Weil ein Hausmädchen unachtsam war, musste Mills hilfsbereiter Freund Carlyle die Französische Revolution zweimal nacherleben.

John Stuart Mill hatte zwar mit seinem Verleger einen Vertrag für eine Geschichte über die Französische Revolution unterzeichnet, war aber in andere Projekte verwickelt und nicht in der Lage, die Vertragsbedingungen zu erfüllen. Er schlug vor, dass stattdessen sein Freund Thomas Carlyle das Buch schreiben sollte. Mill schickte seinem Freund sogar eine Bibliothek mit Büchern und anderen Materialien über die Revolution. Carlyle arbeitete er sein einziges vollständiges Manuskript an Mill. Während es in Mills Obhut war, wurde das Manuskript zerstört. Mill zufolge geschah dies durch ein unvorsichtiges Hausmädchen, dass es für Abfall hielt und als Feueranzünder benutzte. Carlyle schrieb dann das gesamte Manuskript um und bezeichnete sein Buch als „direkt und flammend aus dem Herzen“. Carlyles übliche Arbeitsweise die Notizen zu einem Kapitel zu vernichten, nachdem er den Text geschrieben hatte, erschwerte die Arbeit an der zweiten Version des Buches maßgeblich. Er hatte freiwillig auf das verzichtet, was zu seiner Zeit einem Backup am nächsten kam.

Diese Geschichte zeigt zwei Dinge: Zum einen sind Daten-management und seine Missgeschicke kein neues Thema, und zum anderen ist es, unabhängig vom Medium in dem ein Manuskript verfasst wird, wichtig eine Kopie zu haben, falls etwas mit dem Original passiert.

Quellen:

- https://en.wikipedia.org/wiki/The_French_Revolution:_A_History

14 Desorientiert

Sie räumte nicht nur den Keller auf, sondern riss das ganze Haus ab.

Tracy Teal war eine Studentin, die Computerlinguistik im Rahmen eines Master-Abschlusses in Biologie an der University of California in Los Angeles studierte. Sie hatte Monate damit verbracht, Simulationssoftware zu entwickeln und zu betreiben. Danach war sie endlich bereit mit ihrer Analyse zu beginnen. Der erste Schritt vor der Analyse war es, alle wichtigen Daten zu organisieren und alle unnötigen zu löschen. Für den Löschvorgang benutzte sie den typischen Routinebefehl "rm -rf *", der alle Daten im aktuellen Verzeichnis und in den Unterverzeichnissen löscht. Das Problem war nur: Sie war gar nicht im Verzeichnis, wo die entsorgbaren Daten lagen, sondern im Stammverzeichnis ihres Projekts. Da durch diesen Befehl in ihrem Unix-Systemen, die Dateien nicht erst in den Papierkorb geschoben werden, wie unter Windows oder Macintosh, wurden alle Projektdaten mit einem Schlag gelöscht.

Tracy hatte Glück, denn ein automatisiertes Backup rettete ihre Arbeit. Dafür musste sie den IT-Helpdesk ihrer Abteilung freundlich anfragen, ob dieser ihre Dateien wiederherstellen könne. Jetzt ist Tracy Teal Executive Director bei The Carpentries, einer Non-Profit Organisation, die Forschenden weltweit grundlegende Kenntnisse in Codierung und Datenwissenschaft vermittelt. Dennoch denkt Tracy beschämt an ihre damalige Situation zurück, weil sie vor dem Datenunfall selbst für den IT-Helpdesk gearbeitet hatte. Für sie war es "wie der Rettungsschwimmer, der gerettet werden muss".

Diese Geschichte zeigt, dass auch erfahrene Forschende im Umgang mit Daten Fehler machen können. Bei wichtigen Daten sollte immer ein Versionierung- bzw. Backupsystem verwendet werden, damit aufwendig erhobene Daten nicht durch ein Missgeschick verloren gehen.

Quellen:

- https://www.nature.com/articles/d41586-019-01040-w

15 Keine gute Tat bleibt ungesühnt

Hätte er es liegen gelassen und nicht eingesteckt, wäre er der Attacke entgangen.

Wissenschaftler der Universität Illinois führten einen Versuch durch, um herauszufinden, inwieweit Menschen unbekannte USB-Sticks, die sie auf der Straße finden, an ihren Computer anschließen.

Die Ergebnisse der Studie zeigten, dass die Mehrheit die Sticks innerhalb kurzer Zeit von der Straße aufgehoben hatte und in vielen Fällen auch auf die Inhalte zugegriffen wurde. Eine Datei auf den Sticks erlaubte es den Wissenschaftlern in Erfahrung zu bringen, wann die USB-Sticks ausgelesen wurden und auf welche Dateien zugegriffen wurde. Obwohl auch Neugier eine Rolle spielte, wollte ein Großteil der Leute, die den USB-Stick mitgenommen hatten, die Identität des Besitzers ermitteln, um den Stick zurückzugeben. Die Autoren der Studie stellten außerdem fest, dass nur wenige der Versuchspersonen Vorsichtsmaßnahmen beim Öffnen der Inhalte ergriffen hatten. Es ist nicht schwer sich vorzustellen, dass ähnliche Methoden wie bei diesem harmlosen Angriff zu Forschungszwecken, auch leicht von wirklichen Kriminellen angewendet werden können.

Das Beispiel zeigt gut, wie wichtig der richtige Umgang mit fremden Speichermedien ist, um Angriffe auf den eigenen Computer zu vermeiden. Andererseits werden verlorene Datenträger oft von den Findern durchsucht, wodurch sensible Daten für nicht autorisierte Personen zugänglich werden. Daher sollten externe Speichermedien (z.B. USB-Sticks und Festplatten) immer mit einem Passwort geschützt und/oder verschlüsselt werden.

Quellen:

- https://elie.net/publication/users-really-do-plug-in-usb-drives-they-find/

16 Nur schauen, nicht anfassen

Hätte sie den Spruch "Nur schauen, nicht anfassen!" ernster genommen, wäre ihr die zusätzliche Arbeit erspart geblieben.

Eine Forscherin hatte die Daten eines Experimentes auf ihrem Computer gespeichert. Eines Tages öffnete sie die originalen Rohdaten in Microsoft Excel, wobei in einigen Spalten die Formatierung automatisch geändert wurde. Dadurch wurden die ursprünglichen Werte unwiederbringlich umformatiert und waren nicht länger auswertbar. Da keine Kopie der Rohdaten existierte, musste das Experiment wiederholt werden.

Das Beispiel zeigt, dass bei der langfristigen Sicherung von Rohdaten besondere Aufmerksamkeit erforderlich ist. Diese sollten mit entsprechenden Backups gesichert werden, wobei der Zugriff generell auf eine Leseberechtigung beschränkt werden sollte. Dadurch erfolgen Auswertungen in separaten Dateien und die ursprünglichen Rohdaten können nicht verändert werden.

Quellen:

- https://www.nature.com/articles/d41586-019-01040-w

17 Intervention von oben

Ein gefallener Präsident zerstörte seinen Laptop.

Das Büro eines Ingenieurs wurde von einem kleinen Erdbeben durchgeschüttelt, was in Kalifornien wahrlich keine große Überraschung ist. Dabei fiel das Bildnis des früheren US-Präsidenten und einstigen Kunden, Gerald Ford, von der Wand auf den Laptop und zerstörte den Bildschirm. Nach diesem Vorfall machte sich der Ingenieur deutlich mehr Gedanken darüber, was mit seinen Geräten und Daten passieren könnte, sollte sich ein solcher Vorfall wiederholen.

Die Geschichte zeigt, dass die Vorbereitung auf (Natur-)Katastrophen Teil eines effizienten Forschungsdatenmanagements sein sollte. Dies kann man unter anderem durch die Nutzung eines gut funktionierenden Backup-Systems erreichen, bei dem die wichtigen Daten auf verschiedenen Speichermedien an mehreren Orten abgelegt werden.

Quellen:

- https://www.nature.com/articles/d41586-019-01040-w

18 Schlechtes Recycling

Die Uni spart Ressourcen und er hat den Zitate-Salat.

Ein Forscher fing an der Universität an und startete enthusiastisch seine Karriere. Er bekam einen neuen und eine neue E-Mail-Adresse. Schnell machte er sich an die Arbeit und bald hatte er seine erste wissenschaftliche Publikation veröffentlicht! Um zu sehen, ob seine Veröffentlichung schon in Suchmaschinen aufgelistet wird, prüfte er mit Google Scholar. Tatsächlich hatte die Webseite seine Publikation aufgenommen, allerdings war dies nicht die einzige Arbeit, die er dort zu seiner "neuen" E-Mail-Adresse aufgelistet fand. Was war passiert?

Die E-Mail-Adresse wurde vorher von einem anderen Wissenschaftler mit gleichem Vor- und Zunamen genutzt, der mittlerweile den Standort gewechselt hatte. Das universitäre Rechenzentrum hatte die Adresse "recycelt", den alten E-Mail-Verkehr gelöscht und die Adresse wieder freigegeben. Es hatte nicht bedacht, dass noch Publikationen und andere Services mit dieser Adresse verbunden waren. Für den neuen Forscher war es leicht sich eine neue Adresse geben zu lassen, aber doch recht schwer seine falsch angelegten Zitationen auf Google Scholar zu korrigieren. Andererseits hatte der Forscher die Möglichkeit auf alle registrierten Services mit der recycelten E-Mail-Adresse zuzugreifen, was für den vorherigen Nutzer ein hohes Sicherheitsrisiko darstellte.

Die Geschichte zeigt die Probleme der Wiederwendung von E-Mail-Adressen. Bei wichtigen Registrierungen sollte auch für Personen ein persistenter Identifikator (z.B. ORCID, ResearcherID) verwendet werden, der auch bei Umzug oder Namensänderung erhalten bleibt. Zudem sollten eigene Registrierungen dokumentiert und bei Institutionswechsel auf die neue E-Mail-Adresse umgestellt werden, um eigene Sicherheitsrisiken zu minimieren.

Quellen:

- https://www.lcrdm.nl/horror-address-recycled

19 Antiquitäten

Da sie sich den Beginn Ihrer wissenschaftlichen Karriere anschauen wollte, verbrachte sie viel Zeit auf dem Flohmarkt.

Am Anfang ihrer Karriere sicherte eine Forscherin Daten auf Disketten, dem damals gängigen lokalen Speichermedium. Danach wurden die Daten nicht mehr aufgerufen oder auf modernere Datenträger (CDs, DVDs, USB-Sticks, oder externe Festplatten) migriert. Die Datenträger sind vorhanden und eindeutig beschriftet. Aber selbst wenn die Daten nach all der Zeit noch in Takt und vollständig sind, die Hardware zum Auslesen der Datenträger ist schlicht nicht mehr vorhanden. Um ihre alten Daten zu sichten, müsste Leslie also großes Glück haben und ein noch funktionsfähiges Gerät finden: auf dem Flohmarkt, in einem Abstellraum an der Uni oder im schlimmsten Fall im Museum.

Um dieses Problem zu vermeiden, sollten Speichermedien vergangener Projekte dokumentiert und die Verfügbarkeit der komplementären Hardware kontinuierlich beobachtet werden. Spätestens wenn Hersteller ankündigen, Anschlüsse oder Laufwerke in neueren Generationen ihrer Geräte nicht mehr zu verbauen, sollten die Daten migriert werden, um die Bitstream-Preservation sicher zu stellen. Im Vorfeld werden derartige Probleme unwahrscheinlicher, wenn Daten nicht ausschließlich auf 1 Medium gespeichert werden, sondern möglichst nach der 3-2-1 Regel auf mehreren Medientypen.

Quellen:

- https://www.nature.com/articles/d41586-019-01040-w

20 Auf die inneren Werte kommt es an

Nachdem sie über die ermittelte Identität informiert wurde, war sie den Tränen nahe.

Eine Doktorandin nutzte für einen Versuch eine DNA-Probe, die sie von einem Kollegen bekommen hatte. Da sie ihrem Kollegen vertraute, überprüfte sie die Identität der Probe nicht weiter. Im Verlauf der nächsten Monate arbeitete sie intensiv damit, aber die Versuche führten zu keinen sinnvollen Ergebnissen. Die Probe wurde schließlich sequenziert, um zu bestätigen, dass es sich dabei um die richtige Probe handelte. Dabei stellte sich heraus, dass sie aufgrund eines Fehlers in der Beschriftung vertauscht wurde. Somit war die monatelange Arbeit umsonst.

Das Beispiel zeigt gut, dass beim Umgang mit physischen Proben auf die korrekte und deutliche Kennzeichnung geachtet werden muss. Sollte es Unklarheiten bezüglich der Identität von Proben geben, sollten diese überprüft werden bevor die Proben für weitere Versuche verwendet werden. Außerdem sollten die Proben und die dazugehörigen Informationen in einem physischen und/oder digitalen Katalog dokumentiert werden.

Quellen:

- https://www.lcrdm.nl/horror-wrong-sample

21 Startschwierigkeiten

Erst als er vollständig wiederhergestellt war, konnte die Reise in die Berge beginnen.

Ein Forscher machte sich für eine Felduntersuchung auf den Weg in das Pamir-Gebirge nach Tadschikistan. Als er dort ankam stellte er fest, dass der Laptop, auf dem er alle Materialien für die Felduntersuchung gespeichert hatte, nicht hochgefahren werden konnte.

Da er die Materialien nicht auf einem zusätzlichen Speichermedium oder in ausgedruckter Form mitgenommen hatte, und aufgrund der schlechten Internetverbindung auch nicht auf seinen Onlinespeicher zugreifen konnte, war seine gesamte weitere Feldarbeit gefährdet. Glücklicherweise konnte er den einzigen Vertragspartner seines Laptopherstellers im Land ausfindig machen und seinen Computer reparieren lassen.

Die Geschichte zeigt, wie wichtig Backups insbesondere bei der Planung von Feldarbeiten sind. Da hier eventuell keine einfachen Möglichkeiten für Reparaturen und Neuanschaffungen vorhanden sind, müssen mögliche Probleme schon vor dem Start mit eingeplant werden. Onlinespeicher sind als Backup nur geeignet, wenn eine dauerhafte Versorgung mit einer guten Internetverbindung gewährleistet ist.

Quellen:

- https://www.lcrdm.nl/horror-crash-on-field-trip

22 Wenn die Kommandozeile Schwierigkeiten macht

Als sie versuchte die Fehler zu bereinigen, gingen die Probleme erst richtig los.

Während des ersten Jahres ihrer Doktorarbeit im Bereich computergestützte Biologie versuchte eine junge Forscherin sich mit der Arbeit mittels Befehlszeile vertraut zu machen. Der Supercomputer, auf welchem alle Analysen liefen, gab für jede Analyse zwei Dateien aus – eine standardisierte Output-Datei mit der Namensendung ".o" und eine standardisierte Error-Datei mit der Endung ".e".

Bei dem Versuch all die Error-Dateien, die sich angesammelt hatten, zu löschen, tippte sie den Befehl "rm *e*" in die Befehlszeile und vergaß dabei den alles entscheidenden Punkt. In Folge dieses kleinen Fehlers wurden alle Dateien mit einem "e" im Dateinamen gelöscht. Dies galt auch für alle Evolutions-Bäume, an denen sie monatelang gearbeitet hatte und von denen einige mehrere Wochen auf dem Supercomputer in Anspruch genommen hatten. Das war jedoch leider nicht das Ende der Probleme. Zu ihrem Entsetzen musste sie feststellen, dass sie viele der verlorenen Dateien nicht gesichert hatte. So verursachte ein kleiner Tipp-Fehler eine große Verzögerung in der Fertigstellung der Doktorarbeit.

Die Moral der Geschichte ist, dass man seine Arbeit täglich sichern sollte. Und wer eine Datenbereinigung durchführt, sollte sich ganz sicher sein, dass er die richtigen Dateien löscht.

Quellen:

- https://www.lifehacker.com.au/2016/06/file-error-your-nightmare-data-loss-stories/

23 Geteilte Arbeit ist halbe Arbeit?

Nachdem er seine Dateien geteilt hatte, war sein Team nicht mehr gut auf ihn zu sprechen.

Ein Forscher arbeitete in einem 6-köpfigen Team, um eine Enzyklopädie im Umfang von 800.000 Wörtern zu schreiben. Das Team hatte nach längerer Zeit entschieden, dass alle Arbeiten auf einem gemeinsamen Arbeitsbereich geteilt werden sollen. Hierfür lag schon ein Server vor, der von einigen Mitgliedern kooperativ verwendet wurde für die Projektdaten. Also verschob er das Resultat seiner Arbeit auf den gemeinsamen Speicher und plötzlich waren die anderen sauer.

Das Problem bestand darin, dass sowohl die eigenen Dateien des Forschers, als auch die bereits vorhandenen Dateien der anderen Teammitglieder generisch benannt wurden. Der Verzicht auf eine explizite, zweckgebundene Namenskonvention führte dazu, dass beim Kopieren der lokalen Dateien auf den gemeinsamen Speicher viele Dateien überschrieben wurden. Der Forscher dachte sich nicht viel dabei, da man die alten Dateien sicher durch ein Backup wiederherstellen könnte – nur leider lag dies mehr als 1 Monat zurück und das aktuelle Backup hat durch schlechtes Timing nur die neuen Dateien gesichert.

Die Geschichte verdeutlicht zwei wesentliche Punkte des FDM: Zum einen sollten Organisationsstrukturen und Namenskonventionen festgelegt werden. So können z.B. Datum, Thema und Autor/-in Teil des Dateinamens sein. Zum anderen sollte Wert auf ein gründliches Versionierungs- und Backup-System gelegt werden.

Quellen:

- persönliche Kommunikation

24 Kalendarische Fragen

1908: Russen kommen zu spät zum Schuss.

Im Jahr 1908 dauerten die Olympischen Spiele ganze 6 Monate an. Dennoch war die russische Mannschaft bei den Schießwettbewerben am 11. Juli nicht anwesend. Das Problem lag darin, dass die russischen Athleten den Julianischen anstelle des Gregorianischen Kalenders benutzten, wie er im Rest Europas üblich ist (dies änderte sich erst mit der russischen Revolution 1917). Als die russische Delegation endlich eintraf, waren daher die ersten anderthalb Monate der Spiele bereits vorüber. Sie konnten noch aktiv an den Spielen teilnehmen und einige Medaillen gewinnen. Dies war aber nur ein schwacher Trost für die Schützen.

Die Olympischen Spiele 1908 waren jedoch nicht nur wegen dieser kalendarischen Verwirrung bemerkenswert. Auch die Dokumentation der Spiele ließ viel Raum für Verbesserungen. Es gab keinen umfassenden und schlüssigen Bericht. So ist es beispielsweise immer noch fraglich, ob die Türkei überhaupt bei den Spielen vertreten war!

Diese Geschichte zeigt, dass die Festlegung einheitlicher Konventionen und Standards zentral ist - sei es für die Austragung von internationalen Wettbewerben oder in der Wissenschaft. Ohne die klare Benennung der jeweils verwendeten Maße und Einheiten werden Daten von Dritten gegebenenfalls nicht richtig interpretiert oder repliziert, weil Memos oder Laborbucheinträge nicht richtig verstanden werden.

Quellen:

- https://www.rbth.com/history/331074-russia-late-for-olympics-1908

25 Feuer und Flamme

Wäre der Tag etwas wärmer gewesen, wäre ihm viel Ärger erspart geblieben.

Im Wohnzimmer eines Forschers stand ein Karton mit Originaldokumenten, die wichtige Daten mit Personenbezug enthielten. Die Babysitterin seiner Tochter suchte an einem besonders kalten Tag nach Material zum Anheizen des Kamins und nahm die Papiere als Feueranzünder, da sie davon ausging, dass es sich bei den beschriebenen Blättern nur um Schmierpapier handelte. Ein Großteil der Dokumente wurde dadurch unwiederbringlich vernichtet.

Die Geschichte zeigt, dass Dokumente mit persönlichen Daten immer nur in verschlossenen Möbeln und in verschlossenen Räumen gelagert werden sollten. Dies gewährleitest nicht nur die Wahrung des Datenschutzes, sondern auch den Erhalt der Unterlagen. Während das Anlegen von digitalen Kopien von Unterlagen in diesem Beispiel den Ausgang verbessert hätte, ist dies nicht immer für alle Dokumente möglich. Gerade bei unwiederbringlichen Materialien ist somit besondere Sorgfalt geboten.

Quellen:

- persönliche Kommunikation

26 Never change a running system?

Da sie die alten Sprachen nicht mehr beherrschten, konnte kein Geld verteilt werden.

Anfang 2020 brach global die Krankheit COVID-19 (verursacht durch das Coronavirus SARS-CoV-2) aus, was aus Quarantänegründen zur Schließung vieler Geschäfte und Betriebe führte. Die Folge war vor allem in den USA eine große Zahl an Arbeitslosen, die dringend Geld für die nächste Mietzahlung, Lebensmittel oder andere Ausgaben benötigten. Die Regierung beschloss daraufhin ein Entlastungspaket für jeden, der sich arbeitslos meldete. Doch warum gelangte das Geld trotzdem nicht an die Leute?

Grund dafür war die Überlastung von kritischen Systemen, auf denen noch COBOL läuft. COBOL ist eine Programmiersprache, welche Ende der 1950er-Jahre entwickelt wurde, um kaufmännische Anwendungen zu steuern. Sie gilt als veraltet und wird in der Ausbildung von Programmierern nicht mehr unterrichtet. Darum gab es kein Personal, welches sich um die Systeme kümmern konnte, als diese zusammenbrachen. Leider laufen im Wirtschaftssektor noch bis heute viele Anwendungen mit der veralteten Programmiersprache. Um das Problem zu lösen, suchte die Trump-Administration nun verzweifelt nach 'pensionierten' COBOL-Programmierern.

Dieses Beispiel zeigt, dass auch wenn ein System (vermeintlich) gut läuft, Bestehendes hinterfragt werden sollte, da sich Anforderungen gerade im IT-Bereich schnell ändern. In den Computerwissenschaften sind Innovation und Entwicklung wichtig. So können zum Beispiel Daten irgendwann nicht mehr abgerufen werden oder liegen in Formaten vor, deren Bearbeitung immer schwerer zu gewährleisten ist.

Quellen:

- https://youtu.be/PpV_5-tCS-c?t=310
- https://www.theverge.com/2020/4/14/21219561/coronavirus-pandemic-unemployment-systems-cobol-legacy-software-infrastructure
- https://fortune.com/2020/04/15/how-to-get-unemployment-benefits-coronavirus-extra-600-dollars/
- https://www.datacenter-insider.de/cobol-eine-programmiersprache-wird-uns-alle-ueberleben-a-865219/
- https://www.linkedin.com/pulse/never-change-running-system-warum-diese-weisheit-im-zeitalter-welsch-1e/

27 Auf ganzer Linie verlabelt

Obwohl er die Linie im Inventar fand, war seine ganze Arbeit letztendlich umsonst.

Während seiner Promotion brauchte ein Wissenschaftler eine bestimmte Zelllinie für seine Forschung. Er fand diese im Inventar, welches der ganze Fachbereich gemeinsam nutzte, und begann die Zellen zu kultivieren. Dann startete er ein teures Massenspektrometer-Experiment mit dieser Zelllinie, nur um später heraus zu finden, dass diese nicht korrekt gelabelt worden war. Da sein Promotionsvertrag kurz vor dem Ende stand, war unglücklicherweise keine Zeit mehr, um das Problem zu lösen.

Dies hätte sich durch das korrekte Labeln der Probe und das Verifizieren der Zelllinie vor dem Beginn des Experiments vermeiden lassen. Das Management von Proben, einschließlich Verifikation, korrektem Labeling, Dokumentation in einer analogen oder digitalen Inventarliste, ist ein essenzieller Teil des Forschungsdatenmanagements. Wie am Beispiel der Geschichte zu sehen, ist es von großer Wichtigkeit, dass alle diese Schritte korrekt durchgeführt und dokumentiert werden.

Quellen:

- https://www.lcrdm.nl/horror-mislabelling

28 Teure Mäuse

Transportstress machte das geplante Experiment nicht nur teuer, sondern undurchführbar.

Eine Doktorandin in Deutschland bestellte Labormäuse einer bestimmten Knockout-Reihe direkt bei einer Firma in den USA, die das entsprechende Patent hielt. Bei Knockout-Mäusen werden mittels genetischer Manipulation gezielt Gene deaktiviert, um die durch sie regulierten biologischen Mechanismen zu untersuchen. Außerdem eignen sich derartige Tiere als Modell für menschliche Erkrankungen und pharmakologische Fragestellungen. Die Bestellung in den USA führte dazu, dass die Transportkosten der Mäuse (mehrere Tausend Euro für Tiertransport im Flugzeug) deren Wert (ca. 3,50 Euro je Maus) um ein Vielfaches überstiegen. Problematisch war zudem, dass die Tiere für das geplante Experiment durch den langen und stressigen Transport unbrauchbar waren. Eine spätere Bestellung über einen europäischen Lizenznehmer für das Patent war zwar problemlos möglich, aber die Aktion kostete die Doktorandin wertvolle Zeit und ihr Institut unnötig Geld. Wahrscheinlich handelte es sich aus Sicht der meisten Mitarbeiter/innen im Labor um „Allgemeinwissen“, dass nicht nur die Patenthalter die richtigen Mäuse liefern können, sondern auch Lizenznehmer in Europa. Es war so selbstverständlich, dass es keinen Gesprächsanlass gab, und die Doktorandin hätte diese Information aktiv abfragen müssen.

Die Bestellunterlagen wurden zwar daraufhin geprüft, dass der richtige Typ Maus bestellt wurde. Die eher banale Frage woher die Mäuse kamen, hatte der Projektleiter wahrscheinlich nicht auf dem Schirm und bestätigte die Bestellung, ohne den US-Lieferanten in Frage zu stellen.

Diese Geschichte zeigt, wie wichtig es ist, implizites Wissen zu verschriftlichen. Eine Liste mit Auswahlkriterien und möglichen Bezugsquellen für Versuchstiere verschiedener Genlinien hätte diesen Fehler verhindert. Der Doktorandin wäre sicher aufgefallen, dass die US-Firma nur einer von mehreren in Frage kommenden Lieferanten ist. Ein systematisches Vorgehen bei der Auswahl ist nicht nur für Versuchstiere, sondern auch für Software, Hardware, Messgeräte und Verbrauchsmaterialien sinnvoll.

Quellen:

- persönliche Kommunikation

29 Die Einstellung ist alles

Kurz vor der Abgabe stimmte die Einstellung der Mitarbeiter nicht.

Eine Forschergruppe arbeitete gemeinsam an einem Projekt, bei dem sie kurz vor der Deadline standen. Damit jeder Zugriff auf die vielen Projektdateien hat, wurden alle Daten auf einen Microsoft SharePoint Server abgelegt. Hier sieht man den aktuellen Stand der Dokumente und kann kollaborativ an ihnen arbeiten. In der finalen Phase sollte der Bericht für das Projekt fertig gestellt und alle alten und unnötigen Dateien gelöscht werden. Ein Mitarbeiter löschte jedoch den Bericht, weil er dachte, es wäre ein alter Entwurf. Dies sollte jedoch kein Problem sein, da SharePoint Dateien im Hintergrund versioniert und man die Datei leicht wiederherstellen kann. Warum lief es aber doch nicht wie geplant?

Es waren die Einstellungen. Auch wenn SharePoint eins von vielen Systemen ist, welches Projektdateien versionieren und wiederherstellen kann, so ist dies in den Werkseinstellungen der Software nicht vorgegeben, da dies natürlich auch mehr Speicherplatz kostet. Erst über verschiedene Optionenmenüs können Versionierung von Dateien und Abstände für Backups festgelegt wird. Das hatte das Team jedoch nicht getan und so war der Bericht für immer verloren und musste in kürzester Zeit komplett neu geschrieben werden.

Diese Geschichte zeigt, dass das Funktionieren verschiedener Backup-, Cloud- und anderer Softwarelösungen für Projektdaten nicht einfach vorausgesetzt werden sollte. Es ist wichtig sich vorher zu informieren, wie die Sicherung der Daten bei den Anwendungen funktioniert und am besten diese auch zu testen. Viele Funktionen müssen erst freigeschaltet oder an die Bedürfnisse des Projekts angepasst werden.

Quellen:

- persönliche Kommunikation

30 Versunkene Schätze

Sandy's stürmische Art war nicht hilfreich bei dem Versuch, die kleinen Plagegeister besser zu verstehen.

Eine Neurobiologin an der Rockefeller Universität in New York, hatte im Keller ihres Privathauses einen Server für die Speicherung ihrer Forschungsdaten eingerichtet. In Folge von Hurricane Sandy im Jahr 2012 wurde ihr Keller überschwemmt und beinahe wären ihre gesamten Daten zu einem Mosquito-Genom Projekt verloren gegangen.

Die Geschichte zeigt, dass unerwartete Ereignisse und Katastrophen wie Wirbelstürme oder Brände die Datenspeicher an einem Standort schnell gefährden können. Dabei sind private Wohngebäude meist schlechter gegen solche Ereignisse gesichert als spezialisierte Gebäude (z. B. Rechenzentren) und Daten sollten wenn möglich nicht in den privaten Räumlichkeiten gelagert werden. Generell sollten entsprechend der 3-2-1 Regeln immer 3 Kopien der Daten auf 2 unterschiedlichen Speichermedien existieren. Eine der Kopien sollte dabei an einem externen Standort aufbewahrt werden, damit bei größeren Katastrophen an einem Standort nicht alle Kopien der Daten zerstört werden.

Quellen:

- https://www.nature.com/articles/d41586-019-01040-w

31 Unbeschriebenes Blatt

Die Daten existierten, konnten aber trotz größter Bemühungen nicht nachgenutzt werden.

Zu Beginn seiner Promotion wurde einem jungen Wissenschaftler mitgeteilt, er solle an unveröffentlichten Daten arbeiten, die drei Jahre zuvor erhoben wurden. Er erhielt mehrere Ordner voller Daten. Darin enthalten waren Dateien mit identischem Namen, aber unterschiedlichem Inhalt, Skripte von denen niemand mehr wusste, was sie tun oder warum sie existieren und Tabellen mit unklaren Spaltenbezeichnungen. Noch dazu war teilweise unbekannt, welche Geräte und/oder Einstellungen genau für die Datenerhebung verwendet wurden. Da die Daten mehrere Jahre alt waren, konnten weder intensive Gespräche mit den Herstellern der identifizierten Geräte noch mit den damaligen Forschern die Nachnutzbarkeit der Daten ermöglichen. Am Ende konnten die Daten einfach nicht mehr verwendet werden.

Dies zeigt, wie essenziell das Beschreiben und Dokumentieren von Datensammlungen und Analyseprozessen ist. Auch wenn Datendokumentation Zeit braucht, ist es noch zeitintensiver schlecht dokumentierte, jahrealte Daten aufzubereiten. Obwohl viele Forschende denken, dass sie ihre Daten kennen, ist es sehr wahrscheinlich, dass die meisten von ihnen einen Großteil der Details innerhalb weniger Jahre vergessen. Daher sollte die Datendokumentation immer so umfangreich, detailliert, präzise und für Dritte leicht verständlich sein wie möglich.

Quellen:

- https://www.lcrdm.nl/horror-lack-of-documentation

32 Es sind die kleinen Dinge

Wäre sein Name leichter zu buchstabieren gewesen, wäre seine Leistung stärker wahrgenommen worden.

Kleine Fehler in der Schreibweise von Namen können großen Einfluss auf die Karriere von Forschenden haben. Insbesondere Namen, die zum Beispiel Sonderzeichen enthalten, können schnell in verschiedenen Schreibweisen erscheinen. Diese uneinheitliche Schreibweise kann zu einer systematisch geringeren Erfassung von entsprechenden Zitationen führen.

Die Forscherin Terje Tüür-Fröhlich zeigte dies in ihrer Arbeit unter anderem am Beispiel des bekannten Soziologen Pierre Bourdieu, für dessen Namen sie in den Wissenschaftsdatenbanken 85 Mutationen feststellen konnte. Die fehlerhafte Aufzeichnung der Zitationen ist insbesondere zu Beginn einer wissenschaftlichen Karriere problematisch, da die Zitierhäufigkeit oft als Maß für die wissenschaftliche Leistung verwendet wird und unter Anderem bei der Vergabe von Fördermitteln oder Stellen eine Rolle spielt.

Eine Möglichkeit, um dieser Problematik entgegenzutreten, ist die Verwendung von persistenten Identifikatoren (PID) für Personen. Hierdurch wird der Einfluss von Fehlern in der Schreibweise von Namen minimiert. Beispiele für PIDs für Forschende sind die unabhängig vergebene ORCID oder die über Thompson Reuters vergebene ResearcherID. Mittlerweile sind PIDs für Personen recht weit verbreitet und müssen bei einigen Verlagen bei der Einreichung von Artikeln mit angegeben werden.

Quellen:

- https://www.heise.de/tp/features/Auch-Pierre-Bourdieu-ist-ein-Indexierungsopfer-3727711.html
- https://lisa.gerda-henkel-stiftung.de/fehler_in_zitationsdatenbanken_sind_nicht_zufaellig_verteilt?nav_id=7314

33 Arbeitnehmer mit Tarnumhang

Er schien der Personalabteilung einfach immer wieder durch die Finger zu rutschen.

Obwohl die Personalabteilung mehrfach einen Eintrag für Steve Null anlegte, verschwand dieser wiederholt aus der Datenbank. Das System nahm "Null" wörtlich und deutete den Eintrag als fehlendes Datum. Für die Datenbank existierte Steve nicht, ja er schrie seine Nichtexistenz durch seinen Namen laut heraus. Bevor es die Anfrage nach Steve verarbeitete, kontrollierte das System zunächst, ob überhaupt Daten eingegeben waren. Moderne Systeme verhindern so, dass die häufig versehentlich ohne Inhalt abgeschickten Anfragen das System unnötig belasten. Leider ist der Nebeneffekt dieses "search_term!= NULL", dass Menschen mit dem Namen Null in derartigen Systemen nicht gefunden werden können, obwohl der entsprechende Eintrag existiert. Die Suche wird einfach zu früh abgebrochen.

Die Geschichte zeigt, dass es sinnvoll ist die Grenzen eines verwendeten Datenbank-Sytems genau anzusehen [1]. Gibt es innerhalb des Systems Regeln, die dazu führen, dass bestimmte Einträge systematisch nicht gefunden oder aber auf eine Art und Weise interpretiert werden, die nicht intendiert ist? Hier hilft ein Blick in die (hoffentlich existierende) Dokumentation der Software bzw. bei übernommenen Daten in die verwendeten Konventionen (z.B. für fehlende Werte). Außerdem sollte eine entsprechende Dokumentation für alle selbst erstellten Daten angelegt werden, damit zukünftig noch nachvollzogen werden kann, was gestern selbstverständlich war.

[1] Sie sind Programmierer und halten das Problem für einen alten Hut? Suchen Sie mal nach bug report FLEX-33644 für den XMLEncoder in Apache Flex.

Quellen:

- Matt Parker (2020): Humble Pi - When Math Goes Wrong in the Real World, S. 259.

34 Falscher Alarm

Dem Feuer waren sie nicht anheim gefallen – aber weg waren die Daten doch.

Der Chefingenieur einer Data Recovery Firma hatte einmal auf Grund eines Flächenbrandes sein sämtliches Hab und Gut verloren. Umso ironischer war es, was einem seiner Kunden passierte. Der Kunde hatte 96 Laufwerkschränke direkt unter einer Sprinkleranlage aufgestellt. Eines Tages löste wahrscheinlich ein Fehlalarm die Sprinkleranlage aus und flutete die Schränke mit Wasser. Viele der Daten waren unwiederbringlich verloren, da kein Update vorgenommen wurde.

Die Geschichte zeigt anschaulich, dass Daten immer nach der 3-2-1-Regel gelagert werden sollten (mindestens 3 Kopien auf mindestens 2 verschiedenen Speichermedien und 1 der Kopien an einem anderen Ort). Zudem ist ein regelmäßiges und verlässliches Backup-System essenziell.

Quellen:

- https://www.nature.com/articles/d41586-019-01040-w

35 In der feindlichen Basis

Als der Agent die Basis infiltrierte, reagierte man mit Hardwareentsorgung.

Im Jahr 2008 entschied das Militär der USA alle entfernbaren USB-Speichergeräte aus den Militärbasen zu entsorgen und keine USB-Geräte mehr einzusetzen. Was führte zu diesem Entschluss?

Als beim Außeneinsatz im mittleren Osten ein USB-Stick entdeckt und analysiert werden sollte, stellte sich später heraus, dass dieser Malware eines ausländischen Geheimdienstes enthielt. Die Daten mit dem bösartigen Code gerieten damit unbemerkt in das innere Netzwerk des US Militärs. Ironischerweise hieß das bösartige Malwareprogramm sogar "Agent.btz". Als das Programm und die Sicherheitslücke entdeckt wurde, entschied das Pentagon, alle entfernbaren Geräte, die über USB-Schnittstelle laufen, sofort zu entsorgen. Bis heute stellt dieses Geschehen eines der größten Sicherheitseinbrüche in der Geschichte des US-Militärs dar.

Prinzipiell stellen Geräte, die einen USB-Anschluss verwenden ein Sicherheitsrisiko dar, da sie durch ihren universellen Zweck schnell Kontrolle über ein System erlangen (z.B. indem sie sich als Tastatur ausgeben) oder einfach nur gefährliche Daten einschleusen können. Insbesondere für Forschungs-einrichtungen und Unternehmen ist es wichtig niemals fremde USB-Geräte einfach an kritische Systeme anzuschließen. Es sollte immer ein zusätzlicher Hardwareadapter, virtuelle Arbeitsumgebung oder zumindest ein Virenscanner als Softwarelösung eingeschaltet sein, um Schadsoftware auf einem USB-Gerät frühzeitig zu erkennen.

Quellen:

- https://www.computerworld.com/article/2514879/infected-usb-drive-blamed-for--08-military-cyber-breach.html
- https://www.vice.com/en_us/article/7xy5ky/the-american-military-sucks-at-cybersecurity

36 Lost in Translation

Da sie sich nicht an gemeinsame Normen hielten, kam er irgendwann vom rechten Weg ab.

Der Mars Climate Orbiter (MCO) war Teil eines NASA-Programms zum besseren Verständnis des Mars. Am 11. Dezember 1998 startete der MCO an Bord einer Rakete zu seiner Mission. Ziel war es, dass der MCO den Mars in einem kreisförmigen Orbit umrunden und Messungen zu Atmosphäre und Klima des Planeten aufzeichnen sollte. Allerdings kam es bei dem geplanten Manöver zu einem Fehler, der MCO kam deutlich näher an den Mars heran als geplant und die Sonde ging verloren.

Die Ursache des Navigationsfehlers war die Verwendung unterschiedlicher Maßeinheiten durch die verschiedenen beteiligten Projektpartner. Während das Navigationsteam das metrische System verwendete, nutzte das amerikanische Unternehmen Lockheed Martin Astronautics, das mit der Herstellung der Sonde beauftragt wurde, das Angloamerikanische Maßsystem. Die Umrechnung der unterschiedlichen Maßeinheiten (z.B. Pfund-Sekunden und Newton-Sekunden) wurde nicht immer korrekt berücksichtigt und resultierte in fehlerhaften Kurskorrekturen.

Tatsächlich hatte die NASA in ihren Spezifikationen klar dargelegt, dass das metrische System verwendet werden sollte. Der Sondenhersteller missachtete allerdings die Vorgaben und verursachte so den Verlust der Sonde.

Die Geschichte zeigt, dass die Verwendung gemeinsamer Standards bei Projekten mit internationalen Partnern (z.B. USA) besonders zu beachten ist, wenn deren Maßsysteme nicht auf dem Internationalen Einheitssystem (SI-Einheiten) basieren. Zudem sind gut definierte Standards Voraussetzung für die Vergleichbarkeit und Nachvollziehbarkeit von Projekten.

Quellen:

- https://www.simscale.com/blog/2017/12/nasa-mars-climate-orbiter-metric/
- https://de.wikipedia.org/wiki/Mars_Climate_Orbiter#Verlust

37 Null Island

Statistisch gesehen, war die Umgebung des Polizeireviers der gefährlichste Ort von allen.

Auf der online verfügbaren Verbrechensübersicht des Los Angeles Police Departments konnte man sehen, dass zwischen Oktober 2008 und März 2009 über 1380 Einträge aus der Umgebung des Polizeireviers stammten. Dies machte fast 4% aller aufgezeichneten Verbrechen der Stadt in diesem Zeitraum aus. Erst als die Los Angeles Times sich deswegen beschwerte, weil sie ihren Sitz ebenfalls in dem Viertel hat, ist dem Polizeirevier der Fehler im System aufgefallen. Doch was ist passiert?

Alle Polizeiberichte wurden händisch verfasst und meist automatisch in die Datenbank eingespeist. Wurde der Ort des Verbrechens nicht erkannt, so wurde als Default-Wert einfach der Standort des Polizeireviers eingetragen, was zur Verfälschung der Kriminalstatistik führte. Das Polizeirevier hatte den Fehler dahingehend bereinigt, dass es die fehlenden Ortsangaben mit "Null" (Angabe für fehlenden Wert in der Informatik) korrigiert hat. Null-Angaben können aber Teile von Datensätze unbrauchbar machen, wenn sie für bestimmte Visualisierungen oder Berechnungen benötigt werden. Man spricht deshalb auch von "Null Island - where bad data goes to die".

Die Geschichte zeigt, wie wichtig es ist Attribute von Tabellen und Datenbanken richtig zu bestimmen. besonders wenn diese auch fehlende Werte haben können. Setzt man bei fehlenden Werten einen für Maschinen als logisch lesbaren Wert (wie z.B. „Null“ als Kommentartext oder als Ortsangabe (0.0,0.0)) ein, so werden die Daten fehlinterpretiert und können Ergebnisse verfälschen. Die Dokumentation solcher Ersatzwerte ist daher essenziell.

Quellen:

- "When Good Data Turns Bad" aus dem Buch "Humble Pi: A Comedy of Maths Errors", Seite 253

38 In die "Scheiße" gegriffen

Ein weniger spezielles Futter hätte den Kriechtierzensus repräsentativer gemacht.

Weil sie einfach und quasi überall zur Verfügung stehen, haben Biologen bisher häufig menschliche Fäkalien verwendet, um Bestandsaufnahmen von kotfressenden Insekten zu machen. Da einige Spezies dieses Futter deutlich attraktiver finden als andere, könnte es durch dieses Verfahren zu Verzerrungen in der Erfassung der örtlichen Biodiversität gekommen sein, wie Studien der Oxforder Zoologin Elizabeth Raine zeigen. Alternative Verfahren werden derzeit evaluiert. Der bisherige Standard hat sich also als suboptimal erwiesen und eine konzertierte Entwicklung eines neuen Standards ist der nächste logische Schritt.

Die Geschichte zeigt, dass tradierte Vorgehensweisen in ihren Auswirkungen auf die Ergebnisse hinterfragt werden sollten und dass es wichtig sein kann, präzisere Verfahren für die Erfassung der tatsächlich gewünschten Messgröße zu entwickeln. Gleichzeitig wird deutlich, dass die Verwendung von Standards einheitliche Fehlerkorrekturen und eine konsistente Einführung neuer Verfahren erleichtert.

Quellen:

- https://www.economist.com/science-and-technology/2020/01/09/dung-beetles-prefer-human-faeces-to-those-of-wild-animals

39 Wie gewonnen so zerronnen

Wäre der Besuch etwas später gekommen, hätte er keine Punkte verloren.

Ein Forscher führte langwierige Messungen zu den Eigenschaften von Plasma in einem Reaktor zur plasmaunterstützten chemischen Gasphasenabscheidung durch. Er hatte die Daten der aktuellen Messung noch nicht gespeichert als ein Theoretiker aus dem Projekt in das Labor kam, um mit ihm über seine Modelle zu sprechen. Der Besucher drückte eine Taste auf der Tastatur und löschte dabei versehentlich die Daten der Messung. Glücklicherweise handelte es sich dabei nur um einen einzelnen Messpunkt, denn die Daten waren nicht wiederherzustellen.

Die Geschichte zeigt, dass neue Daten so schnell wie möglich gespeichert und entsprechende Backups erstellt werden sollten. Außerdem sollten während der Aufzeichnung und Sicherung von Daten Ablenkungen vermieden werden, um Fehler zu vermeiden.

Quellen:

- https://www.lcrdm.nl/horror-erroneous-keyboard-key

40 Das schickt sich nicht

Hätten sie nicht umsortiert, wäre die Auslieferung deutlich problemloser gelaufen.

Der Verein Deutscher Bibliothekarinnen und Bibliothekare brauchte 2019 drei Anläufe, um sein Jahrbuch an alle Mitglieder auszuliefern. Die Zuordnung von Namen und Adressen in Excel wurde korrumpiert, da nur eine Spalte in der Excel-Tabelle umsortiert und so die ursprünglichen Zeilen inhaltlich durchbrochen wurden. Die Auslieferung der Bücher erzeugte durch diesen Fehler viele Rückläufer. Dass die Zustellung über die Weihnachtsfeiertage stattfand, erschwerte die Kommunikation mit den Vereinsmitgliedern zusätzlich. Über die Feiertage waren sowohl die Vereinsverantwortlichen als auch die Mitglieder eher mit anderen Angelegenheiten beschäftigt. Die Mitglieder wurden zwar darüber informiert, dass sie die zugesandten Bücher annehmen sollten, auch wenn sie nicht die richtigen Adressaten waren. Dennoch konnten nicht alle Jahrbücher korrekt ausgeliefert werden. Im zweiten Anlauf trat ein Programmierfehler im Nachbestellungsformular auf. Diejenigen, die keine Mitgliedsnummer angegeben hatten, konnten nicht bedient werden und mussten ihren Band erneut nachbestellen. Insgesamt dürfte die Häufung von Versandproblemen für erhebliche Kosten für den Verein und Irritation bei den Vereinsmitgliedern gesorgt haben.

Die Geschichte zeigt deutlich, dass beim Postversand besondere Vorsicht geboten ist. Adressdaten sollten auf jeden Fall schreibgeschützt gespeichert, mehrfach gesichert und vor dem Versand auf ihre Unversehrtheit hin kontrolliert werden. Software, die im „Kunden“-Kontakt genutzt wird, sollte vor ihrem Einsatz gründlich auf die Funktionalität getestet werden.

Quellen:

- persönliche Kommunikation

41 Das Ende ist nah

Wie Excel einmal fast Wiki-Leaks sabotierte.

Als Wiki-Leaks-Gründer Julian Assange im Jahr 2010 eine Datei mit 92.000 geleakten Feldberichten aus Afghanistan an Journalist/-innen von The Guardian und The New York Times übergab, endeten die Aufzeichnungen abrupt im April 2009, obwohl auch für den Rest des Jahres Daten hätten vorhanden sein sollen. Was war passiert?

Die Journalist/-innen hatten die Datei in Excel geöffnet. Zum damaligen Zeitpunkt war Excel auf eine maximale Größe von 65.536 Zeilen beschränkt und die große Menge an Daten sprengte das Fassungsvermögen der Tabelle. Mit dem Öffnen in Excel wurden so alle Daten nach Zeile 65.536 abgeschnitten.

Obwohl die maximale Zeilenanzahl seitdem auf 1.048.576 erhöht wurde, zeigt diese Geschichte trotzdem eindrücklich, dass Excel kein adäquater Ersatz für ein professionelles Datenbanksystem ist. Dies gilt vor allem für Forschungsvorhaben, die mit einem hohen Aufkommen an Daten rechnen. Hier ist es wichtig, sich schon vor der Antragstellung Gedanken über Alternativen und deren mögliche Kosten zu machen.

Quellen:

- "When Good Data Turns Bad" aus dem Buch "Humble Pi: A Comedy of Maths Errors", Seite 244-245

42 Zu Schön, um wahr zu sein

Die Materialien waren zu schön um wahr zu sein, was zu einer großen Blamage führte.

Beginnend im Jahr 1998 erschienen in kurzer Folge einige beachtliche Artikel von Mitarbeiter/-innen der Bell Laboratories zur Entdeckung neuer kohlenstoff-basierter Materialien. Das Problem war jedoch, dass andere Materialwissenschaftler/-innen die Ergebnisse nicht replizieren konnten.

Trotz des starken Interesses dauerte es weitere 3 Jahre bis anderen Forschenden auffiel, dass sich die Zahlen in vielen der Veröffentlichungen auffällig glichen und einige Grafiken einfach zu "schön" waren, um tatsächlich existierende Systeme abzubilden. Ironischerweise war ein junger deutscher Wissenschaftler namens Jan Hendrik Schön Ko-Autor aller zweifelhaften Veröffentlichungen und an den Arbeiten zu diesen beteiligt. Ein unabhängiges Expertenkomitee wurde eingesetzt und kam zu dem schockierenden Ergebnis, dass in mindestens 16 von 25 Fällen, die den Veröffentlichungen zugrundeliegenden Daten niemals existiert hatten. Schöns Erklärungen, dass er die Primärdaten aus Platzmangel gelöscht und verwendete Speichermedien nicht mehr funktionierten bzw. weggeworfen wurden, erschien dem Ausschuss mehr als zweifelhaft.

Eine solide Verpflichtung auf Open Data hätte den Schwindel sehr viel schneller auffliegen lassen bzw. von vornherein unmöglich gemacht. Der Fall führte zudem an den Bell Laboratories zur Einführung neuer Richtlinien zur Datenhaltung, zur Verantwortlichkeit von Ko-Autor/-innen sowie zum Review von Primärdaten vor der Publikation.

Quellen:

- On Being a Scientist. A Guide to Responsible Conduct in Research: Third Edition (2009), https://doi.org/10.17226/12192

43 Katzenjammer

Wären sie auf den Hund gekommen, wäre das Home-Office wohl angenehmer gewesen.

Corona-bedingt musste ein Wissenschaftler im Home-Office arbeiten. Dies war auch recht unproblematisch, da er alle Arbeitsschritte und Dokumente gut von Zuhause aus organisieren und bearbeiten konnte. Dennoch hatte er mit einer Sache nicht gerechnet, die ihm für die nächsten Tage Probleme bescheren sollte.

Zuhause hatte er Katzen und diese hatten bisher wenig Interesse an all den Kabeln gezeigt, die zu der technischen Ausstattung des Heimbüros gehörten. Dies sollte sich aber eines Nachts ändern. Als der Wissenschaftler an einem Morgen aufstand, war plötzlich das Stromkabel des Arbeitslaptops durchgebissen. Zum Glück hatte er noch einen Ersatzlaptop und da die Arbeitsdaten kontinuierlich über eine Cloud der Universität abgesichert wurden, konnte er auch seinen virtuellen Arbeitsplatz dort schnell wieder einrichten und all seine Termine einhalten.

Diese Geschichte zeigt, dass man getreu nach dem Sprichwort von Murphys Gesetz "Alles was schief gehen kann, wird auch schief gehen." für alle Problemfälle gerüstet sein sollte. Auch wenn die Forschungsdaten auf dem eigenen Laptop sicher scheinen, sollte immer ein Backup existieren, um die Daten bei einem Problem wieder herstellbar zu machen. Neben Ersatzhardware empfehlen sich besonders Cloud-Lösungen, die Dateien auf verschiedenen Geräten schnell synchronisieren können.

Quellen:

- persönliche Kommunikation

44 Sinnlos Herumgefragt

Weniger Vertrauen in professionelle Prozesse und sie hätte sich Arbeit erspart.

Eine Forscherin ärgerte sich gewaltig als ihr auffiel, dass das kommerzielle Umfragetool ihre letzte Version der Umfrage nicht gespeichert hatte. Sowohl der uniinterne Support als auch der Kundendienst des Unternehmens konnten allerdings keine befriedigende Erklärung für den Vorfall liefern. Die nicht gespeicherte Umfrageversion musste neu erstellt werden, was dank des noch vorhandenen lokalen Backups der Forscherin mit vertretbarem Aufwand möglich war.

Auch wenn kommerzielle Produkte ein integriertes Backup anbieten, ist es notwendig dessen Funktionalität in regelmäßigen Abständen zu testen. Ein eigenes automatisiertes Backup kann zusätzlich die Redundanz des Systems erhöhen und die Wahrscheinlichkeit von Datenverlusten senken. Die Verbindung zwischen Online-Tool und Server sollte nicht unterbrochen werden, um die Datenübertragung zu gewährleisten. Häufige Ursachen hierfür sind Fehler der Anwendung (z.B. Session time out) oder Probleme mit der Internetverbindung. Eine Alternative kann die Erstellung auf dem lokalen Rechner und anschließend das Hochladen bzw. Eintragen der Inhalte im Tool sein.

Quellen:

- persönliche Kommunikation

45 Mit Intelligenz vermessen

Sie rutschte mit Exzellenz durchs Raster.

Eine Expertin für Künstliche Intelligenz (KI) bewarb sich auf eine Position in diesem Bereich. Weil in ihren Bewerbungsunterlagen die geforderten Praktika nicht auftauchten und die KI die vorhandene Berufserfahrung im Ausland für die Stelle nicht erwartete, hätte die KI-Expertin beinahe keine Einladung zum Vorstellungsgespräch bekommen. Tatsächlich war sie nicht aussortiert worden, weil sie einen ungewöhnlichen und weiblichen Vornamen hatte. Außerdem fragte sie beim Unternehmen nach, was sowohl ihr Engagement als auch ihr Verständnis für die Funktionsweise künstlicher Intelligenz bewies. Ihre Vermutung, dass ihr Lebenslauf nicht dem Muster "erfolgreicher" Lebensläufe im Unternehmen entsprach, war nicht weit von der Wahrheit entfernt. Es gab zwar mit der Anforderung nach Praktika ein "objektives" Kriterium, aber der Abgleich erfolgte nur daraufhin, ob das Kriterium vorlag oder nicht. Gleiche oder höherwertige Alternativen wurden von der Software einfach nicht anerkannt.

Da künstliche Intelligenz musterbasierte Entscheidungen trifft, sind für einen erfolgreichen Einsatz zwei Punkte essentiell: Es muss ein geeigneter Trainingsdatensatz vorliegen und die Aussagen müssen mit Testdaten überprüft werden, um unerwünschte Effekte bei der späteren Anwendung möglichst frühzeitig zu erkennen und zu beheben.

Quellen:

- https://www.spiegel.de/wissenschaft/technik/ki-forscherin-ueber-algorithmen-sind-wir-menschen-wirklich-so-simpel-a-00000000-0002-0001-0000-000172493030

46 Nicht nur in Stein gemeißelt

Dank seiner Vorgänger konnte er das Rätsel lösen.

Während einer Ägyptenexpedition Napoleons entdecke der französische Offizier Piere François Xaver Bouchard im Jahre 1799 den Stein von Rosette im Nildelta. Er ist das Bruchstück einer höheren Stele, auf der ein Dekret in drei unterschiedlichen Sprachen (altgriechisch, demotische Schrift und Hieroglyphen) eingemeißelt ist. Unmittelbar nach der Entdeckung fertigten französische Wissenschaftler vor Ort zahlreiche Kopien der Inschriften an. Nach der Niederlage gegen die Briten, fiel auch der Stein von Rosette in britischen Besitz und der Forscher Thomas Young begann sich mit den Texten zu beschäftigen. Glücklicherweise hatten die Franzosen aber Kopien, sodass es 1822 Jean-François Champillion gelang mithilfe der ihm bekannten altgriechischen Sprache die demotische Schrift und die Hieroglyphen zu entschlüsseln. Nach der Veröffentlichung seiner Entdeckung erfolgte auch die Entzifferung weiterer Hieroglyphen und der Grundstein der modernen Ägyptologie wurde gelegt.

Die Geschichte zeigt, dass Dank der Abschriften gleich zwei Gedanken des Forschungs-datenmanagements umgesetzt wurden. Zum einen gab es Kopien für den Fall des Verlusts des Originals (Backup) und zum anderen erhielten auch weitere Wissenschaftler/-innen Zugriff auf die Texte und konnten daran forschen (Open Data).

Quellen:

- https://www.youtube.com/watch?v=TDnuTzAyCss
- https://de.wikipedia.org/wiki/Stein_von_Rosette

47 Die tönernen Füße der Schuldenbremse

Eine konventionellere Gewichtung hätte ihrem Ruf weniger geschadet.

Die Forschenden Kenneth Rogoff und Carmen Reinhard von der Harvard University postulierten 2010, dass eine Überschreitung der Staatsverschuldung um 90% der Wirtschaftsleistung eine negative Auswirkung auf das Wirtschaftswachstum eines Staates hat. Diese Grundannahme ist sowohl die Basis der deutschen Schuldenbremse als auch der Sparauflagen der Euro-Rettungspolitik. Der Versuch das Ergebnis anhand der zugrunde liegenden Daten zu replizieren gelang allerdings nicht. Tatsächlich wurden in die Studie aus dem Jahr 2010 Daten für bestimmte Jahre nicht aufgenommen, einige Fälle ungewöhnlich stark gewichtet und mehrere Länder versehentlich nicht berücksichtigt.

Das der fehlgeschlagenen Replikation folgende Medienecho schadete dem Ruf der beiden Ökonomen erheblich. In der Replikationsstudie als auch in einer Nachfolgestudie der Harvard-Forschenden blieb der grundlegende Zusammenhang zwar bestehen, allerdings war die Verminderung der Wachstumsraten weniger gravierend als ursprünglich berechnet.

Dieses Beispiel zeigt, dass einzelne Entscheidungen im Verlauf der Datenanalyse gut dokumentiert und in den zugehörigen Publikationen genannt werden müssen, um den Verdacht von Datenmanipulation zugunsten besonders spektakulärer oder signifikanter Ergebnisse gar nicht erst aufkommen zu lassen. Auch die Präregistrierung von Studien bei einem entsprechenden Journal ist ein guter Schritt um Forschungsfrage, Forschungsdesign und –umsetzung unabhängig begutachten zu lassen.

Quellen:

- https://www.spiegel.de/wirtschaft/panne-mit-excel-tabelle-rogoff-und-reinhart-haben-sich-verrechnet-a-894893.html
- https://www.spiegel.de/wirtschaft/soziales/excel-panne-von-kenneth-rogoff-das-war-ein-massaker-a-929248.html

48 Der kleine Unterschied

Auf dieser magnetischen Spur kam das System ins Schlingern.

Im Frühjahr 2020 kam heraus, dass Hardwareverkäufer wie Western Digital nicht gekennzeichnete HDD (Hard Disk Drive) Festplatten im SMR-Format (Shingled Magnetic Recording) an Stelle von herkömmlichen Festplatten im CMR-Format (Conventional Magnetic Recording) in verschiedenen Backup-Systemen verkauft hatten. Bei dieser Unterart von HDD-Festplatten wird zwar Platz gespart, indem sich die Magnetspuren (ähnlich wie bei Dachschindeln) überlappen, anstatt nur nebeneinander zu liegen, allerdings ist dafür die Lese- und Schreibgeschwindigkeit geringer. Dies führte dazu, dass Backup-Systeme einfach ausgefallen sind, da sie nicht nach Zeitplan ihre Routinen zum Sichern der Daten durchführen konnten.

Die Geschichte zeigt, dass der Einsatz von falscher oder veralteter Hardware zu Fehlern in der Arbeit führen kann. Auch wenn in diesem Beispiel nicht die Forschenden daran Schuld waren, so sollte trotzdem darauf geachtet werden, dass beim Anstieg immer größerer Datenmengen (Stichwort: Big Data) auch die richtigen Voraussetzungen für das Arbeiten mit der entsprechenden Software vorliegen. Die Hardware sollte dafür immer getestet werden, bevor sie in den routinemäßigen Einsatz kommt.

Quellen:

- https://arstechnica.com/gadgets/2020/04/caveat-emptor-smr-disks-are-being-submarined-into-unexpected-channels/
- https://de.wikipedia.org/wiki/Shingled_Magnetic_Recording

49 Clone Wars

Mehr Offenheit bezüglich der Abstammung hätte viel Aufwand und Geld gespart.

Zu Beginn der 2000er Jahre galt die Stammzellenforschung als große Hoffnung für die Entwicklung neuer Therapien. Und so war es geradezu eine Sensation als ein relativ unbekanntes koreanisches Labor unter der Leitung von Woo-Suk Hwang in zwei Veröffentlichungen im prestigeträchtigen Journal Science 2004 und 2005 bekannt gab, ganze 11 Stammzellenlinien aus geklonten menschlichen Embryonen gewonnen zu haben. Zunächst schien der Weg in ein neues Zeitalter der Stamm-zellenforschung geebnet und weltweit versuchten Forschende die Hwang-Methode zu replizieren. Leider ohne Erfolg.

Ein gutes Jahr nach Erscheinen des zweiten Papers fielen Unstimmigkeiten und auffällige Ähnlichkeiten in den Abbildungen auf, die in den Artikeln enthalten waren. Eine eingesetzte Kommission sichtete und analysierte die Primärdaten, was sie dazu veranlasste Tests an den DNA-Proben durchzuführen. Es stellte sich heraus, dass keine der Stammzellenlinien von geklonten Embryonen stammten und sämtliche Angaben und Darstellungen in den Veröffentlichungen erfunden waren. Bis dahin waren bereits Millionen von Fördergeldern rum um die Welt geflossen, um die Ergebnisse zu replizieren - völlig umsonst.

Hätte die Zeitschrift, in welchen die Artikel veröffentlicht wurden, darauf bestanden, dass Hwang und seine Ko-Autor/-innen die Primärdaten als Supplement liefern, wäre der Betrugsversuch höchstwahrscheinlich sofort aufgeflogen. Das Bestehen auf offenen Forschungsdaten hätte hier Transparenz schaffen und die Verschwendung von Fördergeldern verhindern können. Im Zuge dieses Falles änderten viele große Zeitschriften ihre Regularien und auch Fördergeber begannen ein größeres Augenmerk auf offen zugängliche Forschungsdaten zu legen.

Quellen:

- On Being a Scientist. A Guide to Responsible Conduct in Research: Third Edition (2009), https://doi.org/10.17226/12192

50 Erfolgversprechende Entwicklerpersönlichkeit

Als Kapitänin des Schachclubs hatte sie schlechte Karten bei der Auswahl.

Amazon entwickelte ab 2014 ein Computerprogramm, das Lebensläufe von Bewerbern evaluieren sollte. Ziel war es, die Suche nach geeigneten Kandidaten für eine zu besetzende Stelle zu automatisieren. Das Programm nutzte künstliche Intelligenz und bewertete die Eignung der Bewerbung mit ein bis fünf Sternen. Allerdings stellte das Unternehmen 2015 fest, dass das Programm bei Bewerbungen für Softwareentwicklung oder andere technische Positionen keine geschlechterneutrale Auswahl traf. Das Online-Rekrutierungsprogramm mochte schlichtweg keine Frauen.

Die Ursache dafür fand sich im Training des Computermodels, welches mit Lebensläufen von Bewerbern der letzten 10 Jahre gespeist wurde. Die meisten Bewerbungen kamen von Männern, was die männliche Dominanz in der Hightech-Industrie reflektierte. So brachte sich das System selbst bei, dass männliche Bewerber zu bevorzugen sind und stufte Bewerber herab, wenn im Lebenslauf das Wort „Frau“ wie bspw. „Kapitänin des Frauenschachclubs“ vorkam. Nach Bekanntwerden des Falls behauptete Amazon, dass das Programm nie von Personalverantwortlichen eingesetzt wurde. Insider sagten allerdings, dass das auf künstlicher Intelligenz basierende Ranking genutzt wurde, wenn auch nicht ausschließlich.

Das Beispiel zeigt, dass die Datenqualität beim maschinellen Lernen von entscheidender Bedeutung ist. Nach dem Motto „Garbage In, Garbage Out“ kann ein Algorithmus nur so gut sein, wie der Datensatz, der ihm vom Menschen zum Training zu Verfügung gestellt wird.

Quellen:

- https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G