



Hochschule für Angewandte Wissenschaften Hamburg
Hamburg University of Applied Sciences

(Explorative) Datenvisualisierung

Quirine Philippen

Ausarbeitung zur Ringvorlesung InnovationCity 2030

Sommersemester 2015

Quirine Philipsen

(Explorative) Datenvisualisierung

Ausarbeitung eingereicht im Rahmen Ringvorlesung
InnovationCity 2030

im Studiengang Next Media
am Department Informatik
der Fakultät Technik und Informatik
der Hochschule für Angewandte Wissenschaften Hamburg

Betreuender Prüfer: Prof. Dr. Kai von Luck

Abgegeben am 23. August 2015

Quirine Philippen

Thema der Ausarbeitung

(Explorative) Datenvisualisierung

Stichworte

Explorative Datenvisualisierung, Informationsvisualisierung, Datenvisualisierung, Wissensvisualisierung, Datenjournalismus, Big Data, Datamining, automatisierte Datenerfassung

Kurzzusammenfassung

Die Ausarbeitung beschäftigt sich mit den Chancen, durch unterschiedliche visuelle Darstellung von großen Datenmengen aus Datamining-Verfahren, Zusammenhänge und Strukturen erkennen zu können. Dabei ist die (explorative) Datenvisualisierung eine Möglichkeit der Datenanalyse.

Inhaltsverzeichnis

1	Einleitung	5
1.1	Motivation	5
1.2	Aufbau der Arbeit	6
2	Definitionen	7
3	Explorative Visualisierung	8
3.1	Historie.....	8
3.2	Allgemeine Vorgehensweise im Visualisierungsprozess.....	10
3.3	Datentypen	10
3.4	Visualisierungs-Methoden	11
3.4.1	Herkömmliche statistische Verfahren	11
3.4.2	Beispiele moderner Visualisierungstechniken großer Datensätze	12
3.5	Weiterführende Informationen	16
4	Fazit/Ausblick	17
5	Literaturhinweis	
6	Abbildungen	

1 Einleitung

„Exploratory data analysis can never be the whole Story, but nothing else can serve as the foundation stone – as the first step“ John Turkey, 1977

1.1 Motivation

Menschen und Maschinen erzeugen heute täglich 2,5 ExaByte Daten [IBM, 2014]. Die persistente Speicherung von Massendaten, die unter dem allgemeinen Begriff Big Data verstanden wird, ist dank immer leistungsfähigerer Computer heute kein Problem mehr. Online- und Offline-Einkäufe, Webprotokolle aus dem Browser, Einträge in sozialen Medien, finanzielle Transaktionen, Versicherungs-, Verkehrs-, Wetter- und Sensordaten sind nur einige Beispiele, bei denen Daten erhoben und gespeichert werden. Jede Maschine und jeder Mensch hinterlässt bewusst oder unbewusst, bemerkt oder unbemerkt, on- oder offline täglich Millionen von Daten.

Aus der Studie „EMC Digital Universe“ des Marktanalytisten IDC von 2014 geht hervor, dass im Jahre 2020 das weltweite Datenvolumen 44 Billionen Gigabyte beträgt. Vor allem das „Internet der Dinge“ trägt dazu bei, dass sich das „Digitale Universum“ künftig alle zwei Jahre verdoppelt [vgl. IDC 2014].

Nun stellt sich das Problem, wie Big Data auszuwerten und zu analysieren ist, damit die gesammelten Daten zu brauchbaren Informationen werden, um Schlussfolgerungen und Voraussagungen treffen zu können. Laut ICM Studie waren „2013 nur 22 Prozent aller Informationen im Digitalen Universum nutzbare Daten – von diesen wurden wiederum nur fünf Prozent tatsächlich ausgewertet. 2020 werden bereits mehr als 35 Prozent aller Daten nutzbar sein.“ [ebd.]

Dazu berichtet Sabine Bendiek [ebd.], Geschäftsführerin Emc Deutschland GmbH über den Transformationsdruck der Unternehmen:

„... Sie müssen agil und effizient werden und zudem Spezialisten ausbilden, die den Umgang mit den neuen Analyse-Tools beherrschen.“

Welche Technik ermöglicht aber die optimale Auswertung der Massendaten?

Genau dieser Aufgabe stellt sich das Datamining. Durch die Anwendung statistischer Methoden, Techniken und Algorithmen auf große Datenbestände sollen selbständig neue Muster und Zusammenhänge entdeckt und ausgewertet werden. [Cleve, Lämmel, 2014, S. 2-3]

Aber oft reicht das alleinige Auffinden von Zusammenhängen in Datenstrukturen nicht aus, denn auch auf die richtige Aufbereitungsmethode kommt es an.

In dieser Hausarbeit beschäftige ich mich besonders mit der visuellen Darstellung explorativer Datensätze der Datenanalyse. Durch die Visualisierung werden komplexe Datenstrukturen aus Datamining-Verfahren in Zusammenhang gebracht, um daraus bessere Analysen und Hypothesen zu ermöglichen. Man kann daher die Visualisierung als eigene Datamining-Technik in der Datenanalyse bezeichnen. [ebd., S.14]

Ziel der Ausarbeitung ist es, den Vorteil der Visualisierung von Datensätzen herauszustellen sowie einen Überblick über die Vorgehensweisen, wie die Klassifizierung von Datensätzen und die unterschiedlichen Methoden der Visualisierung zu geben.

1.2 Aufbau der Arbeit

Zunächst werden Begriffe definiert, die Voraussetzungen für das Verständnis der Ausarbeitung sind. Im Hauptteil wird weiter auf die Geschichte und die Bedeutung der Visualisierung von Daten eingegangen. Es folgt die Klassifizierung von Daten, Darstellungsmethoden und Anwendungsbeispiele sowie abschließend eine Selbsteinschätzung über das Forschungsthema.

2 Definitionen

(Explorative) Datenvisualisierung vereint mehrere wissenschaftliche Forschungsgebiete, die Themengebiete und Schwerpunkte aus Statistik, Datamining, Datenanalyse, Datenjournalismus sowie computergestützte Techniken zur Visualisierung vereinen.

Es werden zunächst einige Begriffe definiert, die für das Verständnis der Ausarbeitung wichtig sind.

Explorative Datenanalyse

Die explorative Datenanalyse bezeichnet statistische Verfahren zur Aufdeckung von Datenstrukturen, Abhängigkeiten und Abweichungen einer vorhandenen Grundstruktur [vgl. Gabler Wirtschaftslexikon, 2015]. Sie umfasst häufig grafische Verfahrensweisen und dient dazu als Teil des Datamining-Prozesses, Daten zunächst zu explorieren, d.h. sie zu erkunden, um darin durch Visualisierungsmethoden besser die enthaltenen Muster und Strukturen zu erkennen, Schlussfolgerungen zu ziehen sowie mit den Daten interagieren zu können. Der Begriff explorative Datenvisualisierung bezeichnet die interaktive Visualisierung der Ergebnisse aus der explorativen Datenanalyse (EDA) oder auch explorativen Statistik.

Big Data

Big Data bezeichnet große Mengen an Daten, die gespeichert, verarbeitet und ausgewertet werden [vgl. Gabler Wirtschaftslexikon, 2015].

Diese Daten stammen aus unterschiedlichsten Bereichen wie beispielsweise Internet, Mobilfunk, Finanzindustrie, Energiewirtschaft, Gesundheitswesen und Verkehr. Täglich erzeugen die Menschen weltweit u.a. über soziale Netzwerke, Kredit- und Kundenkarten sowie Überwachungskameras diese riesigen Datenmengen. Laut Bitkom [vgl. 2015] sagte Google-Chef Eric Schmidt bereits vor vier Jahren: „Wir generieren heute alle zwei Tage so viele Daten wie vom Anbeginn der menschlichen Zivilisation bis 2003.“

Data Mining

„Data Mining ist eine Datenfilterung, bei der die gespeicherten Daten auf themenspezifische Zusammenhänge und Gesetzmäßigkeiten hin untersucht werden. Durch Extraktion, Filterung und Aggregation der gespeicherten Daten können produktions-, vertriebs- oder marketingrelevante Erkenntnisse gewonnen und umgesetzt werden. Die durch das Data Mining extrahierten Daten sind nur durch die neuen Verknüpfungen erkennbar.“ [IT Wissen, 2015]

Diagramm

Als Diagramm bezeichnet man die allgemeinste Form der graphischen Darstellung von Daten. Es ist eine graphische Repräsentation von Informationen mit Hilfe graphischer Elemente wie Punkte, Linien, Kurven und geometrischen Formen mit Darstellungsattributen wie Farbe und Textur [vgl. Schumann, Müller 2000, S.126].

Datenjournalismus

Datenjournalismus generiert Geschichten aus großen Datensätzen [vgl. Jakubetz, 2013]. Als Informationen dienen alle Daten und Zahlen, die aus großen Datenmengen in Zusammenhang gebracht werden können.

3 Explorative Visualisierung

3.1 Historie

Die Grundidee, durch eine geeignete Darstellung von Daten Schlussfolgerungen ziehen zu können, ist nicht neu. Das zeigt das Beispiel der Cholera-Map von 1854 in der Abbildung 1. Diese Karte gilt als eine der ersten nachgewiesenen räumlichen Analysen [Wikipedia, 2015]. Der Arzt John Snow zeichnete auf einer London-Karte alle Adressen der an Cholera Verstorbenen auf. Durch die Häufung der Fälle an bestimmten Adressen erkannte er, dass die Epidemie einer verseuchten Wasserpumpe zu verdanken war. Er widerlegte damit auch die Annahme der Verbreitung durch Luftübertragung.



Abbildung 1
[Wikipedia, 2015]:
Cholera-Map.
John Snow
visualisierte die
Choleraepidemie an
der Broad Street
Pump in London,
1854. Der Graph
zeigt Anhäufungen
von Todesfällen im
Umkreis um die
Pumpe.

Ein weiteres geschichtliches Beispiel zeigt auch der Statistiker und Grafik Designer Edward R. Tuftes auf [Tuftes, 1997]. In seiner Überzeugung wäre der Challenger-Absturz 1986 zu verhindern gewesen, wenn die Daten auf expressive Weise präsentiert worden wären. Die Anzahl defekter Dichtungsringe stieg bei gewissen Temperaturdifferenzen stark an, was Tuftes in seiner grafischen Darstellung verständlich verdeutlichte.

Daten-Visualisierungen können im Vergleich zu einfachen Tabellen deutlich aussagekräftiger sein. Das zeigt Edward Tufte auch in Abbildung 2 und 3 [1993, S. 13-14].

I		II		III		IV	
x	y	x	y	x	y	x	y
10,00	8,04	10,00	9,14	10,00	7,46	8,00	6,58
8,00	6,95	8,00	8,14	8,00	6,77	8,00	5,76
13,00	7,58	13,00	8,74	13,00	12,74	8,00	7,71
9,00	8,81	9,00	8,77	9,00	7,11	8,00	8,84
11,00	8,33	11,00	9,26	11,00	7,81	8,00	8,47
14,00	9,96	14,00	8,10	14,00	8,84	8,00	7,04
6,00	7,24	6,00	6,13	6,00	6,08	8,00	5,25
4,00	4,26	4,00	3,13	4,00	5,39	19,00	12,50
12,00	10,84	12,00	9,13	12,00	8,15	8,00	5,56
7,00	4,82	7,00	7,26	7,00	6,42	8,00	7,91
5,00	5,69	5,00	4,76	5,00	5,73	8,00	6,89

Abbildung 2 [Tufte, 1993]: Datentabelle vs. Grafik 1.
 Vier Tabellen zeigen Werte, die über eine große Anzahl identischer statistischer Maße und Eigenschaften verfügen.

N=11
 Mittelwert X = 9.0
 Mittelwert Y = 7.5

Regressionsgerade: $Y = 0.5 X + 3$

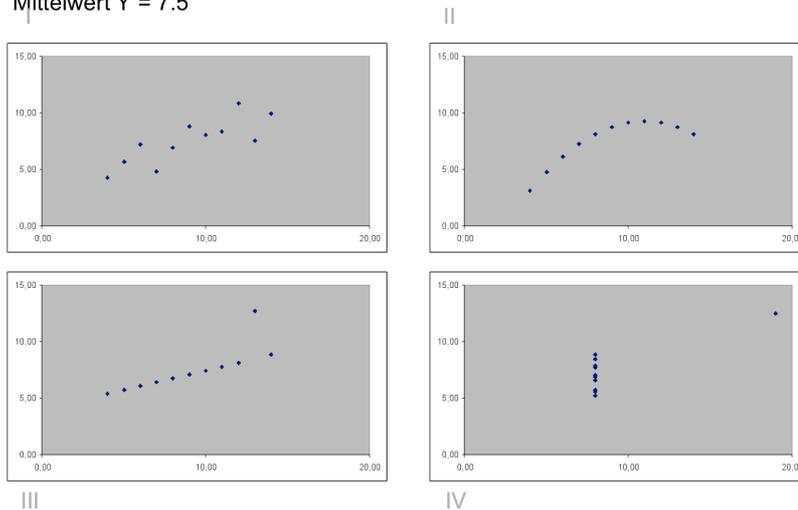


Abbildung 3 [Tufte, 1993; Anscombe, 1973]: Datentabelle vs. Grafik 2.
 Die Graphen zeigen, dass die Tabellenwerte aus Abbildung 2 sehr unterschiedlich sind.

Es ist deutlich zu sehen, dass die Zahlen in der Tabelle auf dem ersten Blick sehr ähnlich ausschauen (Abb. 2). Anhand der tabellarischen Darstellung ist es fast unmöglich, richtige Schlussfolgerungen treffen zu können. Erst die Visualisierung zeigt die deutlichen Unterschiede, die tabellarisch nicht erkennbar waren (Abb. 3).

Die Visualisierung hat gegenüber Zahlentabellen und textlicher Beschreibung folglich nicht nur den Vorteil der besseren und klareren Darstellung sondern auch, dass eine deutlich größere Menge an Informationen wahrgenommen werden kann, die vom Menschen sofort interpretierbar ist. Zudem lassen sich laut Prof. Daniel. A. Keim [vgl. Keim, 2002, S. 1] in textlicher Form deutlich weniger Zeilen auf dem Bildschirm ausgeben, als es eigentlich für die zunehmend wachsenden und zu analysierenden Datenmengen erforderlich wäre.

Aus diesem Grund spricht er weiter auch über den Hauptvorteil des Einbindens menschlicher Fähigkeiten wie Kreativität, Flexibilität und Allgemeinverständnis im Vergleich zu vollautomatischen Verfahren aus der Statistik. Diese werden für ein effektives Data Mining im Datenexplorationsprozess mit den Speicherkapazitäten und Rechenleistungen von Computersystemen kombiniert und lassen sich auch von Nicht-Spezialisten durchführen [vgl. ebd.].

3.2 Allgemeine Vorgehensweise im Visualisierungsprozess

„Exploratory data analysis is detective work.“ John W. Turkey, 1977

Beim visuellen Datenexplorationsprozess werden die Rohdaten aufbereitet, vervollständigt oder reduziert, damit ein Überblick über die relevanten und geglätteten Daten möglich ist. Im zweiten Schritt werden die Daten genauer mit Hilfe von Zoom und Selektionstechniken untersucht, um sie im letzten Schritt detailliert zu analysieren [vgl. Keim, 2002, S. 2]. Diese Gliederung wird von Ben Shneiderman [Shneiderman, 1996] auch Information Seeking Mantra genannt.

Um eine geeignete Visualisierungsmöglichkeit herauszufinden ist es vorteilhaft, die Datensätze nach Kriterien des Datentyps, der Visualisierungstechnik und Interaktions- und Verzerrungstechnik zu unterscheiden und einzuordnen [op. cit., S. 3].

3.3 Datentypen

Daten unterscheiden sich in der Anzahl der Merkmale, auch Dimensionen genannt, und ihrer Struktur. Dazu stellen Jürgen Cleve und Uwe Lämmel fest, dass Daten aus Zahlen-, Ordnungs- und Rechenmerkmalen bestehen können [2014, S. 39-40]. Das stellt unterschiedliche Herausforderungen an das Datamining, da man beispielsweise mit Farbenmerkmalen nicht rechnen kann. Daraus ergeben sich verschieden darzustellende Datentypen wie ein-, zwei und multidimensionale Daten, Textdaten, Graphen (Beziehungen zwischen Daten) sowie Algorithmen und Software. Nach Daniel A. Keim können die verschiedenen Datentypen wie folgt eingeordnet werden [2002, S. 5-8]:

Eindimensionale Daten

Eindimensionale Daten haben ein kontinuierliches Attribut. Oft handelt es sich um zeitabhängige Daten. Ein Beispiel für ihre Darstellung sind Aktien- oder Klimakurven.

Zweidimensionale Daten

Zweidimensionale Daten haben zwei Attribute. Als klassisches Beispiel gelten geographische x-y-Koordinaten wie auf Landkarten.

Multidimensionale Daten

Multidimensionale Daten haben zwischen drei und unendlich viele Attribute. Sie sind zunehmend komplex und undurchsichtig, so dass die einfache 2D- und 3D-Darstellung schwierig wird und neue Darstellungstechniken sinnvoll sind.

Graphen und Hierarchien:

Graphen visualisieren Beziehungen zwischen Daten (Abb. 7). Beispiele dafür können Telefonverbindungen oder Stromnetze sein. Eine hierarchische Anordnung wäre wie beim Prinzip eines Stammbaumes.

Text und Hypertext:

Text ist unstrukturiert und schwer darstellbar, da es keine festen Größeneinheiten gibt. Daher verwendet man Beschreibungsfaktoren, wie etwa das Zählen bestimmter Worte (Abbildung 11).

Algorithmen und Software

Da Software-Projekte sehr komplex und umfangreich sind, kann die Darstellung der Struktur und aller involvierten Komponenten helfen, die Arbeit zu vereinfachen, schneller zu verstehen und den Überblick zu behalten.

3.4 Visualisierungs-Methoden

3.4.1 Herkömmliche statistische Visualisierungs-Verfahren

Für eine statistische Auswertung von Daten gibt es zunächst alte grundlegende graphische Methoden. Darstellungs-Beispiele dafür sind:

- **Histogramm**
(Darstellung von Häufigkeitsverteilungen, vorzugsweise mit gleichgroßen Klassen, Beispiel Klimadiagramm [vgl. Schumann, Müller, 2000, S. 135].)
- **Boxplot**
(Fünf-Punkte-Zusammenfassung für einen schnellen, direkten Vergleich, bestehend aus Median, dem unteren und oberen Quantil und den beiden Extremwerten [vgl. Gabler Wirtschaftslexikon, 2015].)
- **Quantil-Quantil-Diagramm**
(Die Quantile zweier statistischer Variablen werden gegeneinander abgetragen, um ihre Verteilungen zu vergleichen [vgl. Wikipedia, 2015].)
- **Streudiagramm**
(Darstellung von Einzelpunkten, um aus der Punkteverteilung Rückschlüsse auf die Beziehungen zwischen zwei Größen ziehen zu können [vgl. Wirtschaftslexikon 24, 2015].)

Diese Verfahren sind für kleinere Datenmengen geeignet und jede der klassischen statistischen Visualisierung stößt mit steigender Dimension an ihre Grenzen [Cleve, Lämmel, 2014, S. 243]. Eine Möglichkeit wäre daher, die Dimensionen der zu betrachtenden Datensätze entweder zu reduzieren oder neue Visualisierungstechniken anzuwenden, die aufgrund der rasant fortschreitenden Rechenleistung moderner Computersysteme und neuster Softwarelösungen einfacher und vielfältiger werden.

So profitiert der Betrachter durch unterschiedliche Interaktionstechniken, kann die Explorationsziele anpassen und verändern (Abbildungen 7-8). Nach Keim [2002, S. 11-15] können zur effektiven Erkundung der Datensätze Teilmengen individuell herausgefiltert und fokussiert werden. Durch interaktives Zooming werden Daten unterschiedlich repräsentiert und interaktive Verzerrungen ermöglichen die visuelle Darstellung von Ausschnitten im Kontext zur Überblickdarstellung. Die „Linking- und Brushing-Methode“ kombiniert die einzelnen Methoden, indem durch Einfärben und Verknüpfungen Abhängigkeiten und Korrelationen in den Daten erkennbar werden [vgl. ebd., S.15].

3.4.2 Beispiele moderner Visualisierungstechniken großer Datensätze

Als weitverbreitete Data Mining-Tools gelten zur Zeit Rapidminer, SAS, Orange, Microsoft Excel 2013, IBM SPSS und Weka. Aber aufgrund der unzähligen Möglichkeiten, die modernste Software und Computersysteme bieten, können im Rahmen dieser Ausarbeitung nur einige Beispiele für mehrdimensionale und große Datensätze nachfolgend vorgestellt werden.

Parallel-Koordinaten

Die Dimensionen werden in parallelen Achsen dargestellt. Ihre Anordnung ist ausschlaggebend für die Struktur wie in Abbildung 4 und 5.

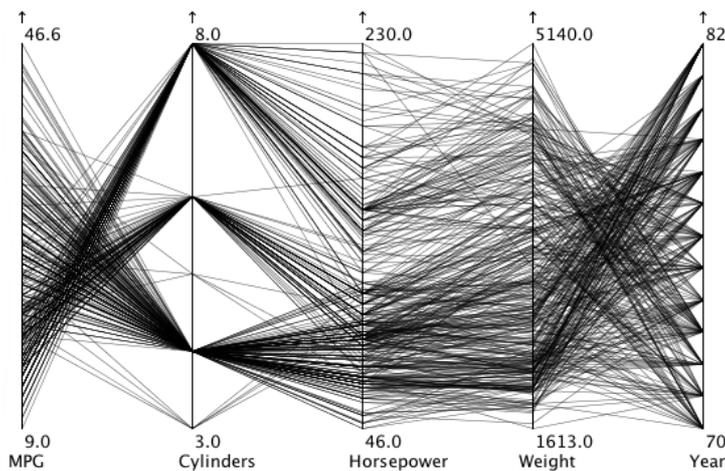


Abbildung 4
[Kosara, 2010]:
Parallel-Koordinaten
zu Automodellen 1.
 Die Grafik zeigt verschiedene Automodelle mit Anzahl der Zylinder, PS, Gewicht und Jahreszahl.

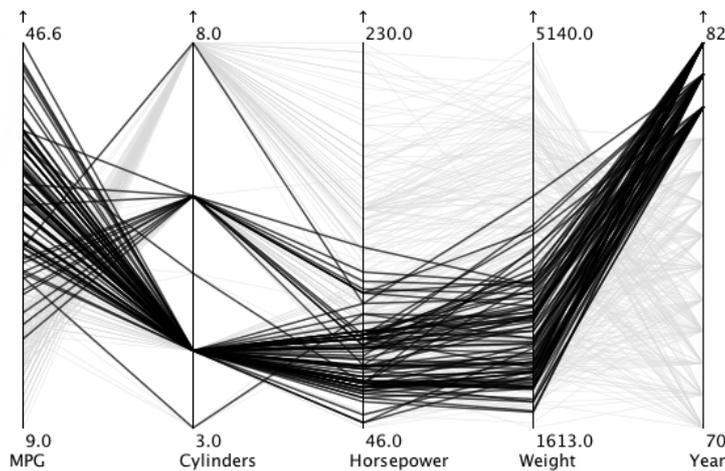
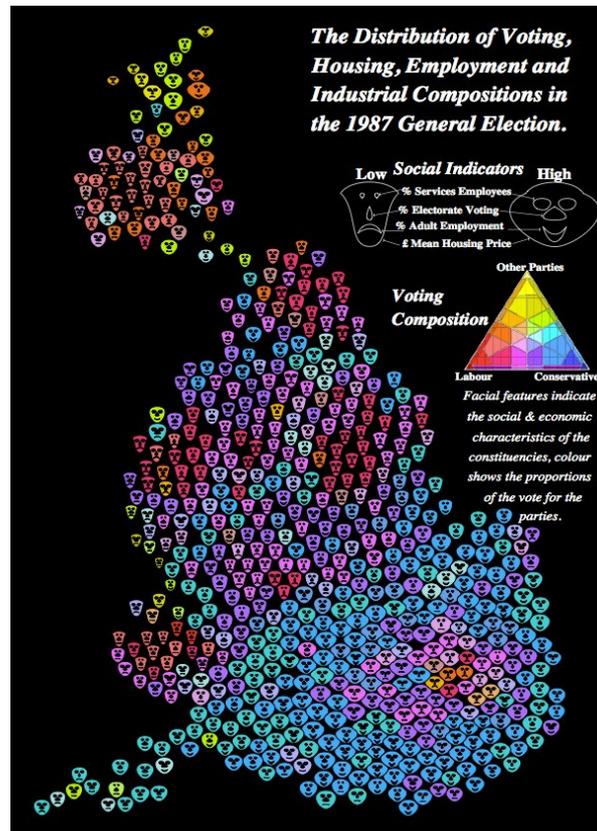


Abbildung 5
[Kosara, 2010]:
Parallel-Koordinaten
zu Automodellen 2.
 Die Teilmenge des Datensatzes, ausgewählt nach den Jahren 1980-1982. Daraus lassen sich neue Verhältnisse der Attribute ablesen.

Icon-basierte Technik

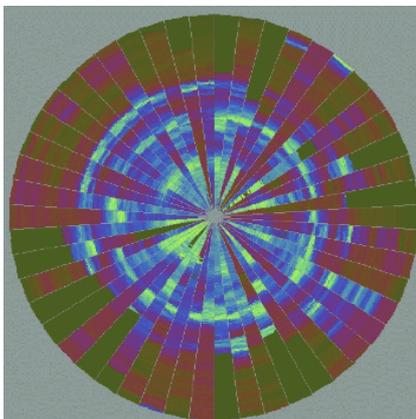
Bei der Icon-basierten Technik werden die einzelnen Eigenschaften anhand graphischer Primitive dargestellt, so dass Muster und Texturen entstehen. Ein bekanntes Beispiel sind die Chernoff-Gesichter. 12 Merkmale werden in einem Gesicht kombiniert und über Form und Größe von Kopf, Nase, Mund und Augen kodiert [vgl. Chernoff, 1973]. Die grundlegende Idee dabei ist, dass es dem Menschen leicht fällt, auch kleine Unterschiede in menschlichen Gesichtern zu erkennen [vgl. Schumann, Müller, 2000 S. 192-194]. Der Nachteil ist, dass die Darstellung schnell unübersichtlich werden kann.

**Abbildung 6 [Dorling, 1991]:
Die Wahlkreise Großbritanniens 1987.**
Das Kartogramm zeigt Gesichter mit vier Merkmalen für soziale Indikatoren. Die Farben stehen für die Parteien. Die industriellen Zentren zeigen schmale, traurige Gesichter, resultierend aus niedrigen Mieten und geringer Beschäftigungszahl. Um London herum sind die Gesichter rund und freundlich, es wird konservativer gewählt.



Pixel-Visualisierungen

Jeder Datenwert wird genau auf einem Pixel abgebildet. Diese sehr kompakte Darstellung eignet sich besonders für große Datensätze, beispielsweise Aktiendaten eines Jahrzehntes, es können jedoch schlecht Detailangaben gemacht werden [op. cit., S. 199].



**Abbildung 7 [Keim, 2002]:
Die Circle Segments Technik.**

Die Visualisierung zeigt die tagesgenauen Kurse von 50 Aktien des FAZ-Indexes über einen Zeitraum von 20 Jahren.

Textvisualisierung

Strukturen, Inhalte und Gewichtung eines Textes können beispielsweise durch Größenverhältnisse von Hauptwörtern dargestellt werden, indem man sie nach Häufigkeiten sortiert.

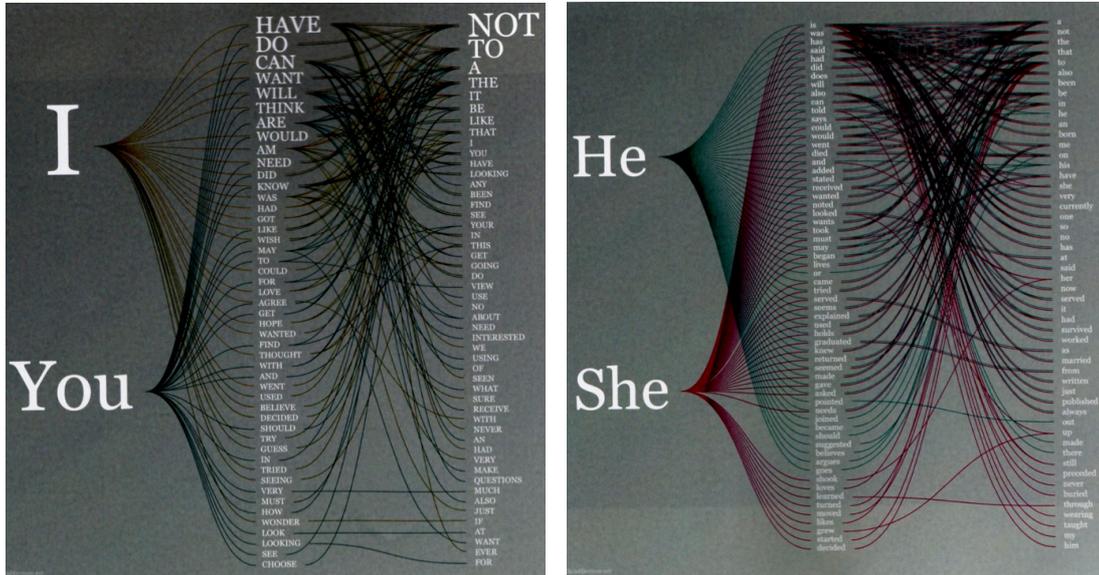


Abbildung 10 [Nix 2014; Chris Harrison 2006]: Web Trigramm

Die Grafik zeigt die Google-Fundstellen im Google-Index mit Häufigkeit absteigend nach Suchphrasen aus drei Wörtern bestehend. Es wird offenbart, dass „He“ nur in Verbindung mit „argues“ und „She“ nur mit „loves“ gesucht wurde.

Geografische Visualisierungen

Kartografie in Form von Land- oder Seefahrkarten, ob real in der Navigation oder abstrakt als U-Bahn-Plan oder Versorgungs-Netz dargestellt, sind früher wie heute allgegenwärtig und dienen als wichtige Orientierung.

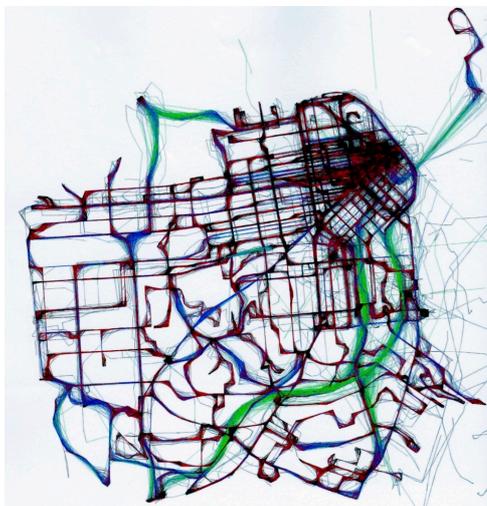


Abbildung 11 [Nix, 2014; Fischer 2010]: A Day of Muni.

Durch unterschiedlich eingefärbte Linien (Schwarz: bis 12 km/h, Rot: bis 30 km/h, Blau: bis 70 km/h, Grün: mehr als 70 km/h) werden mittels öffentlich verfügbarer GPS-Daten die unterschiedlichen Geschwindigkeiten der Municipal Transport Agency in San Francisco in einem 24-Stundenabschnitt analysiert.

3.5 Weiterführende Informationen

Konferenzen 2015/2016:

- **Visualized** 8./9. Oktober 2015 New York, USA
- **IEEE VIS** (Information Visualization Conference) bestehend aus den Haupt-Konferenzen **VAST** (Visual Analytics Science and Technology), **InfoVis** (Konferenz für Informationsvisualisierung) und **SciVis** (Konferenz für Wissenschaftsvisualisierung) 25.-30. Oktober 2015 Chicago, USA
- **VDA** (Conference on Visualization and Data Analysis): 12.-18. Februar 2016 San Francisco, USA
- **Eurographics**: 9.-13. Mai 2016 Lissabon, Portugal
- **EuroVis** 6.-10. Juni 2016 Groningen, Niederlande

Personen:

- Jörn Kohlhammer
- Daniel Keim
- Marco Maars
- Boris Müller

Internet:

- blog.zeit.de/open-data/category/datenvisualisierung
- www.informationisbeautiful.net
- www.oecdbetterlifeindex.org/de/
- www.netzpiloten.de/pbs-america-revealed-erzeugt-bilder-durch-datenvisualisierung/
- www.caida.org/tools/visualization/walrus
- www.nytimes.com/interactive/2008/02/23/movies/20080223_REVENUE_GRAPHIC.html
- infosthetics.com
- <http://pingmag.jp/2007/03/23/infosthetics-form-follows-data/>

4 Fazit/Ausblick

„Ein Bild sagt mehr als tausend Worte“

Visualisierungen veranschaulichen Inhalte und können sie trotz Komplexität oft auf den Punkt bringen, wo Texte und Tabellen scheitern. Damit ihre Verwendung sinnvoll ist, um eine erfolgreiche Aufmerksamkeit zu gewährleisten, sollten vier Hauptaspekte erfüllt sein: Visualisierungen müssen neu, nützlich, schnell verständlich und optisch ansprechend sein.

Der Bedarf an grafischen Darstellungen steigt, wie beispielsweise die als Startup gegründete Firma Statista zeigt, die Infografiken verkauft. Bereiche, in denen visuelle Datenexplorationen angewendet werden, nehmen stetig zu und finden ihre Anwendungen zunehmend bei der Polizei und Feuerwehr, Stadt, Medizin, Wissenschaft, Tourismus, Banken und Versicherungen, Dienstleistungen und IT, Handel und Marketing. Lagerbestände sollen zukünftig noch besser geplant und Kundenverhalten vorausgesagt werden, Kriminalität präventiv verhindert und Versicherungen individueller gestaltet werden.

Dabei geht es immer darum, die wachsenden Datenmengen zu beherrschen und schnell auswerten zu können. Zukünftig werden vermehrt leicht bedienbare Software-Anwendungen zur Verfügung gestellt werden, die statistische Methoden mit unterschiedlichen Visualisierungs-Methoden kombinieren, um daraus maschinelles Lernen zu generieren. Markus Nix stellte bezogen auf die Medien dazu fest: „In der Vergangenheit waren Infografiken häufig Stiefkinder des Journalismus, meist Lückenfüller, die das Geschriebene illustrieren...“ [Nix, 2014, S. 20]. Das ändert sich zunehmend, denn immer häufiger wird trotz noch hohem Aufwand der Mehrwert erkannt, den der Nutzer hat, wenn er beispielsweise auch selbst mit den Daten nach Interessenslage interagieren kann. Eine geeignete Darstellung von Fakten fasziniert den Betrachter, weckt das Interesse und zieht ihn in den Bann. Gefestigt wird dieser Trend auch durch die neue vorherrschende Erzählstruktur des Internets. Da dieses Medium schneller und individueller benutzt wird, versagen hier die üblichen Erzählstrukturen. Redaktionen scheitern oftmals mit langen Texten und suchen schon länger nach Alternativen zu üblichen Fotogalerien. In Zukunft werden Data-Mining-Tools immer wichtiger und wird die Visualisierung, als Teil des Data Mining Prozesses, weiterhin von wachsender Bedeutung sein. Viele Interessensgruppen sind nicht zuletzt aus wirtschaftlichen Gründen stark interessiert, Datensätze zu verstehen, Voraussagen treffen zu können sowie Daten schnellstens in Wissen umzuwandeln. Jede Software und jeder Anwender, der diesen Umgang mit Daten beherrscht, wird zukünftig gefragt sein. Business Analytics, Business Intelligence und Data Scientist sind nicht ohne Grund heute gefragte Berufsbezeichnungen.

In dieser Hausarbeit konnte nur ein kleiner Teil eines Data-Mining-Prozesses umrissen werden. Die Vielfältigkeit und einzeln zu vertiefenden Themen um Data Mining, explorativer Datenanalyse und Visualisierung durch neueste Technologien und Softwareanwendungen, weisen eigenständige große Forschungs- und Wissenschaftsthemen auf.

5 Literaturhinweis:

BITKOM: Big Data braucht eine moderne Datenpolitik, 17.06.2015, online im Internet:
<https://www.bitkom.org/Presse/Blog/Big-Data-braucht-eine-moderne-Datenpolitik.html>
Abgerufen am 20. Juli 2015

CHERNOFF, Hermann: The Use of Faces to Represent Points in k-Dimensional Space Graphically. Journal of American Statistical Association, Vol.68, 1973, S.361-368

CLEVE, Jürgen; LÄMMEL, Uwe: Datamining, Oldenbourg Wissenschaftsverlag, 2014

IDC, EMC Digital Universe with Research & Analysis, The Digital Universe of Opportunities: Rich Data and the Increasing Value of the Internet of Things, April 2014, online im Internet:
<http://germany.emc.com/about/news/press/2014/20140409-01.htm?fromGlobalSelector>
Abgerufen am 15. Mai 2015

IBM KOMMUNIKATION DE, 07.10.2014, online im Internet:
<https://plus.google.com/113660961925634388375/posts/iSWyMGbDLg2>
Abgerufen am 29. Juli 2015

GABLER WIRTSCHAFTSLEXIKON, Springer Gabler Verlag (Herausgeber),
Stichwort: Explorative Datenanalyse, online im Internet:
<http://wirtschaftslexikon.gabler.de/Archiv/3997/explorative-datenanalyse-v9.html>
Abgerufen am 29. Juli 2015

GABLER WIRTSCHAFTSLEXIKON, Springer Gabler Verlag (Herausgeber),
Stichwort: Big Data, online im Internet:
<http://wirtschaftslexikon.gabler.de/Archiv/-2046774198/big-data-v1.html>
Abgerufen am 18. Juli 2015

IT WISSEN, Das große Online-Lexikon für Informationstechnologie,
Stichwort: Datamining, online im Internet:
<http://www.itwissen.info/definition/lexikon/Data-Mining-data-mining.html>
Abgerufen am 15. Juli 2015

JAKUBETZ, Christian: DDJ: Als die Daten laufen lernten..., 17. Oktober 2013,
online im Internet:
<http://universal-code.de/2013/10/17/ddj-als-die-daten-laufen-lernten/>
Abgerufen am 20. Juli 2015

KEIM, Daniel: Datenvisualisierung und Data Mining, 4. Oktober 2002, online im Internet:
<http://fusion.cs.uni-magdeburg.de/pubs/spektrum.pdf>
Abgerufen am 12. Mai 2015

NIX, Markus: Visual Simplicity, Die Darstellung großer Datenmengen, Entwickler.press, 2014

SCHUMANN, Heidrun; MÜLLER, Wolfgang: Visualisierung: Grundlagen und allgemeine Methoden. Springer Verlag, 2000

SHNEIDERMAN, Ben: The Eyes Have It, A Task by Data Type Taxonomy for Information Visualizations. In *Proceedings of the IEEE Symposium on Visual Languages*, pages 336-343, Washington. IEEE Computer Society Press, 1996, online im Internet:
<http://citeseer.ist.psu.edu/409647.html>
Abgerufen am 15. Mai 2015

TURKEY, John W.: Exploratory Data Analysis. Addison und Wesley, 1977

TUFFTE, Edward R.: The Visual Display of Quantitative Information, Graphics Press, Cheshire, Connecticut, 1983

TUFTE, Edward R.: Visual Explanations: Images and Quantities, Evidence and Narrative, Graphics Press, Cheshire, Connecticut, 1997

WIKIPEDIA, Eintrag John Snow, online im Internet:
https://de.wikipedia.org/wiki/John_Snow_%28Arzt%29
Abgerufen am 15. Juli 2015

WIRTSCHAFTSLEXIKON 24, Strichwort: Streupunktdiagramm, online im Internet:
<http://www.wirtschaftslexikon24.com/d/streupunktdiagramm/streupunktdiagramm.htm>
Abgerufen am 12. Mai 2015

6 Abbildungen

Abbildung 1: Cholera-Map

WIKIPEDIA, Eintrag John Snow, online im Internet:

[http://de.wikipedia.org/wiki/John_Snow_\(Arzt\)](http://de.wikipedia.org/wiki/John_Snow_(Arzt))

Abgerufen am 10.05. 2015

Abbildungen 2: Datentabelle vs. Grafik 1

TUFFTE, Edward R.: The Visual Display of Quantitative Information, Graphics Press, Cheshire, Connecticut, 1983, S. 13

Abbildungen 3: Datentabelle vs. Grafik 2

TUFFTE, Edward R.: The Visual Display of Quantitative Information, Graphics Press, Cheshire, Connecticut, 1983, S. 14

ANSCOMBE, Frank J.: Graphs in Statistical Analysis, The American Statistician, 1973, S. 17- 21

Abbildung 4: Parallel-Koordinaten zu Automodellen 1

KORSARA, Robert: Parallel Coordinates, 2010, online im Internet:

<https://eagereyes.org/techniques/parallel-coordinates>

Abgerufen am 15. August 2015

Abbildung 5: Parallel-Koordinaten zu Automodellen 2

KORSARA, Robert: Parallel Coordinates, 2010, online im Internet:

<https://eagereyes.org/techniques/parallel-coordinates>

Abgerufen am 15. August 2015

Abbildung 6: Die Wahlkreise von Grossbritanniens 1987

DORLING, Daniel: House price, non-voting and employments with voting colour 1987 constituency faces cartogram, 1991, online im Internet:

<http://www.dannydorling.org/books/visualisation/Graphics/Pages/Figures.html#156>

Abgerufen am 15. Mai 2015

Abbildung 7: Die Circle Segments Technik

KEIM, Daniel: Datenvisualisierung und Data Mining, 4. Oktober 2002, S. 13,

online im Internet: <http://fusion.cs.uni-magdeburg.de/pubs/spektrum.pdf>

Abgerufen am 12. Mai 2015

Abbildung 8: The Middle East

MCCANDLESS, David; UNIVERSLAB: Knowledge is Beautiful project, online im Internet:

<http://www.informationisbeautiful.net/visualizations/the-middle-east-key-players-notable-relationships/>

Abgerufen am 18. Mai 2015

Abbildung 9: OECD Better Life Index

FIEDLER, Steffen; OECD Studio NAND, online im Internet:

<http://www.oecdbetterlifeindex.org/de/>

Abgerufen am 11. Mai 2015

Abbildung 10: Web Triagramm

NIX, Markus: Visual Simplicity, Die Darstellung großer Datenmengen,
Entwickler.press, 2014, S. 113

HARRISON, Chris: Web Trigrams, Visualizing Google's Tri-Gram Data, 2006,

online im Internet: <http://chrisharrison.net/index.php/Visualizations/WebTrigrams>

Abgerufen am 19. August 2015

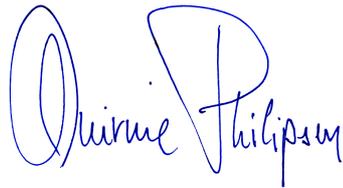
Abbildung 11: A Day of Muni

NIX, Markus: A Day of Muni, 2010 by Eric Fischer, Visual Simplicity,
Die Darstellung großer Datenmengen, Entwickler.press, 2014, S. 113

Versicherung über Selbstständigkeit

Hiermit versichere ich, dass ich die vorliegende Arbeit ohne fremde Hilfe selbstständig verfasst und nur die angegebenen Hilfsmittel benutzt habe.

Hamburg, den 23. August

A handwritten signature in blue ink, reading "Quirine Philipsem". The signature is written in a cursive style with large, rounded letters.