Visual Analytics

Grundseminar Ausarbeitung

Diana Topko

Hochschule für Angewandte Wissenschaften Hamburg
Fakultät Technik und Informatik/Informatik
diana.topko@haw-hamburg.de

Zusammenfassung. Wir leben in einem digitalen Zeitalter, in dem die Informationen immer wichtiger werden. Die ständig wachsenden Datenmengen aus heterogenen Quellen müssen erfasst und bearbeitet werden. Die sogenannten rohen Daten bringen einen geringen Nutzen, weil diese überwiegend unstrukturiert, unvollständig und zum großen Teil fehlerhaft sind. Die Analyse von großen Datenmengen befasst sich mit dem Extrahieren der wertvollen Informationen und Zusammenhängen aus den Daten und wird in mehreren Schritten mithilfe automatisierter Werkzeuge durchgeführt. Um die Effektivität und die Fehlerfreiheit dieses Prozesses zu sichern wird Human Computer Interaction in den Analyseprozess einbezogen. Ein wichtiger Aspekt solcher Analyse ist eine für die menschliche Wahrnehmung leicht verständliche Darstellungsform der Daten. In dieser Arbeit wird Visual Analytics näher betrachtet, ein interdisziplinäres Forschungsgebiet, welches sich mit Kombination der Visualisierung, der Datenanalyse und der Human Computer Interaction befasst.

Schlüsselwörter: Visual Analytics · Data Mining · Pharmakovigilanz

1 Einleitung

Heutzutage verschaffen die fortgeschrittenen Erfassungs- und Speicherungsmöglichkeiten einen relativ leichten Zugang zu Daten. Jedoch steht im Mittelpunkt die Auswahl der passenden Methoden und der passenden Modelle, um die zuverlässigen und beweisbaren Informationen daraus zu gewinnen.

1.1 Motivation

Bei der Datenanalyse reichen die vollständig automatisierte Werkzeuge wie die automatisierte Suche oder das automatisierte Filtern in den meisten Fällen nicht aus. Dafür müssen die zu lösenden Probleme klar definiert und verständlich sein, was in der reellen Welt selten der Fall ist. Oft ist der Pfad von Daten zu Benutzerentscheidungen über die Daten sehr komplex. Selbst die vollständig automatisierte Datenverarbeitungsmethoden repräsentieren nur das Wissen der Methodenerzeuger. Wenn die Ergebnisse solcher Methoden fehlerhafte Entscheidungen

produzieren, ist es besonders wichtig, die Vorgänge überprüfen zu können. An dieser Stelle wird die Bedeutung von Visual Analytics ersichtlich. Das Ziel von Visual Analytics ist die Verarbeitung der Daten und der Informationen transparent für eine analytische Betrachtung zu gestalten [5].

1.2 Definition

In [7] wird Visual Analytics definiert als Kombination automatisierter Analysetechniken mit interaktiver Visualisierung für ein effizientes Verstehen, Erläutern und Entscheiden vor dem Hintergrund sehr großer und komplexer Datenräume. Durch eine Visualisierung von Prozessen und Modellen entsteht die Möglichkeit die Zwischenergebnisse zu bewerten und darauf basierend die Entscheidungen über weitere Verarbeitung der Daten zu treffen. Der Mensch kann in den Analyseprozess eingreifen und beim Bedarf den Ablauf korrigieren oder verbessern.

Laut [6] besteht Visual Analytics aus drei Komponenten. Diese sind:

- Automatisierte Datenanalyse,
- Informationsvisualisierung,
- Interaktion mit dem Benutzer.

Automatisierte Analysetechniken wie Statistik und Data Mining entwickelten sich unabhängig von der Visualisierung. Später um von konfirmatorischer Datenanalyse auf explorative Analyse zu wechseln wurde Datenvisualisierung herangezogen. Daraus entwickelte sich das Visual Analytics Research Gebiet.

Dabei ist Visual Analytics nicht dasselbe wie die Informationsvisualisierung. Visual Analytics geht über die Datenvisualisierung hinaus indem es die Visualisierung mit Datenanalyse kombiniert und bei komplexen Problemen Human Computer Interaction einbezieht.

Weiterhin durch die wachsenden Datenmengen und komplexen Problemen entstand die Notwendigkeit den Benutzer in Knowledge Discovery Process einzubeziehen. Somit ist das Interactive Visual Analytics entstanden mit solchen Anwendungs-Richtungen wie Visual Data Exploration und Visual Data Mining.

Der Zusammenhang der drei Visual Analytics Komponenten ist in der **Abb. 1** dargestellt [6].

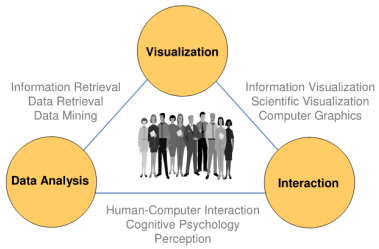

Abb. 1. Die drei Komponenten von Visual Analytics

2 Menschliche Wahrnehmung

Wie in dem **Kapitel 1.1** diskutiert wurde ist der Mensch die zentrale Instanz in dem Visual Analytics Prozess. Die Kenntnisse über die menschliche visuelle Wahrnehmung werden bei der Erstellung von visuellen Modellen eingesetzt damit der Benutzer die Informationen schnell versteht. So wird die Effizienz von Visual Analytics gesteigert. Um dies zu realisieren werden die Informationsdetails in bestimmte für das Wahrnehmungsmechanismus leicht erkennbare Informationstypen verschlüsselt.

A. Paivio betrachtet in [10] visuelle Wahrnehmung als ein paralleles Verarbeitungssystem, welches fähig ist, die räumlichen Matrizen mit Informationen zu empfangen und zu prozessieren.

Es wird zwischen der Sicht auf hoher und auf niedriger Ebene unterschieden. Auf der niedrigen Ebene werden aus der sichtbarer Umgebung physikalische Eigenschaften extrahiert wie beispielsweise die Tiefe, die Form, die Objektgrenzen und Materialeigenschaften der Oberfläche. Auf der hohen Ebene werden hingegen die Objekte erkannt und klassifiziert. Visual Analytics nutzt dabei die Forschungsergebnisse über die Sicht auf der niedrigen Ebene.

In [13] unterteilt Colin Ware den Prozess der visuellen Wahrnehmung in drei Ebenen.

- Auf der ersten Wahrnehmungs-Ebene werden Billionen von Neuronen im Auge parallel aktiv um die Details aus jeder Teil des Umfeldes zu extrahieren. Dabei werden einzelne Neuronen selektiv auf bestimmte Informationstypen eingestellt. Solche Informationstypen können beispielsweise Kantenrichtungen oder Lichtfarben sein.
- Auf der zweiten Ebene teilen schnelle aktive Prozesse im Gehirn die visuelle Felder in Regionen und in einfache Muster ein. Solche Regionen und Muster können beispielsweise durchgehende Konturen, oder die Regionen von gleicher Farbe oder von gleicher Textur sein.
- Auf der dritten Ebene, um eine externe Visualisierung zu verarbeiten, wird im Gehirn eine Sequenz der visuellen Queries konstruiert. Weiterhin, um die visuellen Queries zu beantworten, werden Objekte in visuelle Arbeitsspeicher des Gedächtnisses geladen. Das sind die Objekte, die aus in externer Visualisierung vorhandenen Muster konstruiert wurden.

3 Interaktive Visuelle Analyse

Visual Analytics ist angewiesen auf eine erfolgreiche Interaktion von dem Menschen und dem Computer. In diesem Prozess stellt der Mensch die Fähigkeit der visueller Informationsexploration, kognitive Fähigkeiten sowie das analytische Denken bereit. Der Computer hingegen übernimmt die Aufgabe der Informationsspeicherung, Informationsverarbeitung sowie die Rechenfähigkeit für die Machine Learning Algorithmen, für das Data Mining und für die statistische Berechnungen. [12]

Der Zweck von Visualisierungstechnologien ist es dem Benutzer zu ermöglichen Mustern oder Anomalien in den Daten zu erkennen, damit der Benutzer weitere Entscheidungen über die Aktionen mit den Daten treffen kann. Ohne Interaktivität sind die von Visualisierungssystemen erbrachten Informationen jedoch statisch und können somit nicht einen vollständigen Überblick verschaffen[3].

Interaktive Visuelle Analyse ist ein iterativer Prozess. Dieser Prozess fängt oft an mit einer einfachen Analyse um einen Übersicht über die Daten zu bekommen und um die ersten Hypothesen über die Zusammenhänge, Mustern oder Anomalien in den Daten zu bilden. Für fortgeschrittene Analyse, werden komplexe Methoden eingesetzt, welche die Erkenntnisse aus früheren Analysephasen kombinieren. In [12] wird der sogenannte Human in the Loop Konzept näher Betrachtet.

Ein Beispiel der Interaktion mit dem Benutzer ist die Brushing and Linking Technik. Das Brushing bedeutet das Selektieren einer Untermenge der Daten in einer graphischer Darstellung mit dem Input Device. Während das Linking bewirkt, dass die ausgewählten Daten in allen anderen graphischen Darstellungen die dieselben Daten darstellen, auch markiert werden. Um komplexere Anfragen zu ermöglichen können Brushes mit logischen Operatoren verknüpft werden. Darüber hinaus können on-demand Berechnungen durchgeführt werden.

In der **Abb. 2** ist ein Beispiel der Brushing and Linking Technik zu sehen. Anhand dieser Technik können die zu den trockenen Gebieten korrespondierende Daten gefunden werden. Dieses wird realisiert indem auf der linken Seite der Abbildung durch zwei Brushes die hohen Temperaturen und der geringe Niederschlag selektiert werden. Dementsprechend werden auf der rechten Seite der Abbildung die Daten gehighlighted die den trockenen Gebieten entsprechen [8].

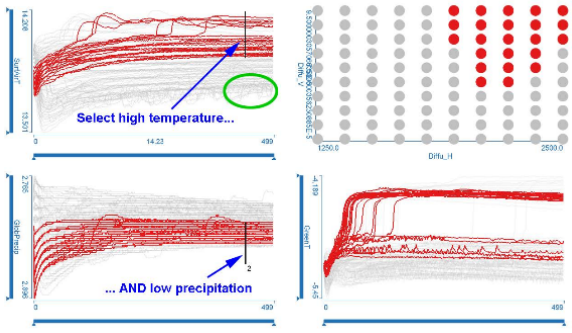

Abb. 2. Brushing and linking

Um die Daten besser zu verstehen werden häufig spezialisierte Interaktionstechniken benötigt. Solche Techniken verlangen das Wissen über die zukünftige

Analyse, welches in vielen Fällen nicht vorhanden ist. Somit entstehen Herausforderungen für die domänenspezifischen Anwendungen.

4 Domänenspezifische Anwendung

Visual Analytics ist ein interdisziplinäres Forschungsgebiet, unter anderem überschneidet es sich mit solchen Disziplinen wie Visualisierung, Data Mining, Data Management, Data Fusion, Statistik und Kognitionswissenschaft [5]. Demzufolge ist es besonders wichtig, die Visualisierung auf dem Stand der Technik zu halten. Dabei, wie in **Kapitel 3** diskutiert wurde, bilden die domänenspezifischen Anforderungen eine große Herausforderung.

4.1 Pharmakovigilanz

Laut [1] ist Pharmakovigilanz die Gesamtheit der Maßnahmen zur Entdeckung, Erfassung, Bewertung und Vorbeugung von Nebenwirkungen sowie anderen arzneimittelbezogenen Problemen, die bei der Anwendung von Arzneimitteln auftreten.

Die Erarbeitung von medizinischen Produkten ist ein hoch regulierter Prozess. Mehrere klinische Untersuchungen müssen durchgeführt werden bevor ein Produkt auf den Markt kommt. Dieser Prozess erfasst unter Anderem systematisches sammeln und analysieren der Daten über Nebenwirkungen und anderer sicherheitsgerichteten Daten (z.B. Elektrokardiogramme oder Labormessungen der Leberfunktion). Dennoch können die Sicherheitsmaßnamen solcher Untersuchungen nicht vollständig das Risiko von unerwünschten Ereignissen, welche das Produkt verursachen kann, abdecken. Die Gründe dafür sind:

- begrenzte Anzahl von Probanden,
- begrenzter Dauer der Untersuchungen (besonders kritisch bei Medikamenten, die langfristig angewendet werden),
- begrenzte Anzahl (oder kompletter Ausfall) von Untersuchenden mit einem erhöhtem Risiko Faktor (z.B. Kinder oder Patienten mit Organschwächen).

Diese Einschränkungen machen es notwendig das Sammeln und die Analyse von sicherheitsrelevanten Daten fortzusetzen auch nach dem das Produkt für den Markt produziert wurde [11]. Die Wirkung von Medikamenten, welche von einer bekannter Wirkung abweicht, kann durch sogenannte Sicherheitssignale festgestellt werden. Laut [4] sind Sicherheitssignale die aus einer oder aus mehreren Quellen stammenden (einschließlich Beobachtungen und Experimenten) Informationen, welche auf einen neuen potenziellen Kausalzusammenhang oder auf einen neuen Aspekt eines bekannten Zusammenhangs zwischen einer Intervention, einem Ereignis oder einer Reihe verwandter Ereignisse, entweder nachteilig oder vorteilhaft schließen lassen und welche als ausreichend wahrscheinlich angesehen werden können, um verifizierende Maßnahmen zu rechtfertigen.

Die ersten Nebenwirkungsberichte erscheinen bereits während der klinischer Produktentwicklung, die weiteren Berichte treten ein in der post-marketing Phase. Solange die Anzahl der Berichte gering bleibt, können diese mit großer Aufmerksamkeit einzeln betrachtet werden. Die statistischen Methoden oder Algorithmen können medizinische und wissenschaftliche Auswertungen keinesfalls ersetzen, dennoch bringen die maschinellen Verfahren einen großen Nutzen, bei einer Vielzahl der Berichte. So können bei bestimmten Ereignissen die Mustern erkannt werden. Zum Beispiel ist es möglich, dass gleiche oder ähnliche Nebenwirkungsberichte aus einer klinischen Prüfstelle oder aus einem Region stammen und somit einen Cluster bilden [11].

4.2 OpenVigil

OpenVigil ist ein Framework für die Analyse von Pharmakovigilanz Daten. Es verarbeitet die Daten aus Berichten über spontane Nebenwirkungsereignisse. Das Framework gehört zu einem an der Universität Kiel betriebenem Projekt. Zurzeit gibt es in zwei Versionen, OpenVigil 1 und OpenVigil 2. Dabei ist OpenVigil 2 eine weiterentwickelte Version von OpenVigil 1. Die 2. Version arbeitet mit teilweise bereinigten und für die Berechnungen vorbereiteten Daten und bietet mehr Funktionalität.

Es werden die Daten sowohl aus amerikanischen(Food and Drug Administration (FDA) Adverse Event Reporting System) als auch aus internationalen (WHO Uppsala Monitoring Centre) Berichtswesen verwendet.

Ein Vorteil bei der Benutzung der FDA-Daten liegt in der Größe der Datenmenge. Die relevanten, anonymen Krankdaten werden öffentlich zur Verfügung gestellt. Die FDA und die WHO benutzen ähnliche Methoden wie OpenVigil in deren Berechnungen (zb Multi-Item-Gamma-Poisson-Shrinker(MGPS)), dennoch machen diese Organisationen ihre Implementierung nicht öffentlich. Damit lassen sich die Ergebnisse nicht verifizieren. Bei OpenVigil wird hingegen viel Wert auf Transparenz gesetzt und auf die Möglichkeit die Ergebnisse zu validieren [9].

Die Data Mining Funktionen von OpenVigil erfassen hoch-konfigurierbare Such- und Ausgabefilter. Analyse einschließt Unverhältnismäßigkeitsanalyse für Erkennung von Sicherheitssignalen wie Proportional Reporting Ratio (PRR) Kalkulationen und Multi-Item-Poisson-Shrinker (mit dem Einsatz von MGPS-Alorithmus). Weiterhin können die Ergebnisse entweder in einem Webbrowser veranschaulicht, geordnet wie auch gefiltert werden oder diese können für eine weitere Analyse in Statistik Software Paketen gespeichert werden [2].

4.3 Unverhältnismäßigkeitsanalyse in Pharmakovigilanz

In der Regel basieren sich die Zusammenhanganalysen zwischen der Einnahme des Medikamenten und den unerwünschten Ereignissen auf 2x2-dimensionalen Kontingenztabellen. Die **Tabelle 1** ist eine Kontingenztabelle mit folgenden

Bezeichnungen: D - Medikamenteneinnahme positiv (Drug), d - Medikamenteneinnahme negativ, E - unerwünschtes Ereignis positiv (Event), e - unerwünschtes Ereignis negativ.

	Medikamenten-einnahme	keine Medikamenten-einnahme	Summe
unerwünschtes Ereignis	DE	dE	E
kein unerwünschtes Ereignis	De	de	e
Summe	D	d	N

Tabelle 1. Kontingenztabelle

Die erste Wahl für die Analyse von Kontingenztabellen sind die Häufigkeitsmethoden der Unverhältnismäßigkeitsanalyse. Diese Methoden stützen sich auf beobachteten und auf erwarteten Werten. So zeigt das *Relativ Reporting Ratio (RRR)* ein Verhältnis zwischen den Wahrscheinlichkeiten, dass ein Ereignis auftreten wird:

- in einer Gruppe und
- in der Bevölkerung.

Das Proportional *Reporting Ratio (PRR)* hingegen zeigt das Verhältnis zwischen den Wahrscheinlichkeiten:

- in der Gruppe 1 und
- in der Gruppe 2

und kann in Kohortenstudien eingesetzt werden.

In Fall-Kontroll-Studien wird das *Reporting Odds Ratio (ROR)* verwendet. Dabei wird das Verhältnis zwischen den Ereignisauftrittswahrscheinlichkeiten gezeigt:

- in der Gruppe 1 und
- in der Gruppe 2.

Ein wichtiger Aspekt ist die Wechselwirkung der Arzneimittel. Für die Betrachtung solcher Fälle muss die Kontingenztabelle erweitert werden. Dabei sind zwei Szenarien möglich:

1. zwei Medikamente und ein unerwünschtes Ereignis (drug-drug interaction)
2. ein Medikament und zwei unerwünschte Ereignisse (syndrome with multiple symptoms)

In der **Tabelle 2** ist das erste Szenario abgebildet.

	Medikament 1+ Medikament 2+	Medikamenten 1+ Medikament 2-	Medikament 1- Medikament 2+	Medikament 1- Medikament 2-	Summe
E+	D1D2E	D1d2E	d1D2E	d1d2E	E
E-	D1D2e	D1d2e	d1D2e	d1d2e	e
Summe	D1D2	D1	D2	d1d2	N

Tabelle 2. Aufeinandereinwirken von Medikamenten

OpenVigil 2 ermöglicht die Berechnungen der Wechselwirkung sowohl bei einem als auch bei mehreren Medikamenten. Dabei wird bei der Berechnung der Wechselwirkung von einem Medikament mit allen anderen eingenommenen Medikamenten die PRR-Methode eingesetzt. Bei der Berechnung der Wechselwirkung von mehreren Medikamenten wurde mit [9] der MGPS-Algorithmus implementiert. Ein Beispiel der Ausgabe ist in der **Abb. 3** zu sehen.

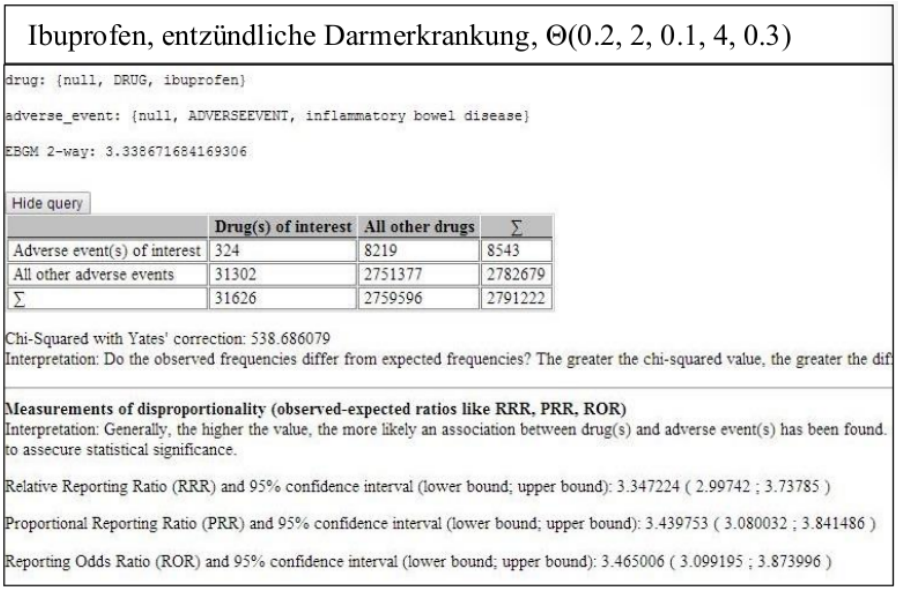

Abb. 3. Unverhältnismäßigkeitsanalyse

An dieser Stelle sollte der Einsatz von Visual Analytics die Übersicht von Ergebnissen wesentlich verbessern. Eine Möglichkeit die Ergebnisse verständlicher darzustellen ist die Form des Venn-Diagramms (**Abb. 4**).

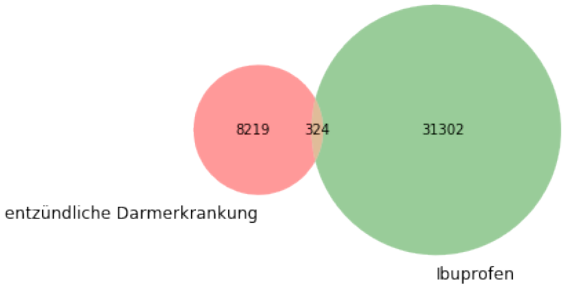

Abb. 4. Venn-Diagramm

5 Ausblick

Visual Analytics verfolgt das Ziel den Benutzer in die automatisierte Datenanalyse einzubeziehen und somit die Abläufe flexibler zu gestalten sowie die möglichen Fehler in den frühen Phasen zu entdecken und zu beheben. Um das zu ermöglichen müssen die Daten in einer für den Benutzer verständlicher Form dargestellt werden. Zu den Aufgaben von Visual Analytics gehört unter anderem das Entwickeln von vielseitigen Tools, welche in domänenspezifischen Anwendungen eingesetzt werden.

Ein großes Problem in Visual Analytics ist nicht ausreichender Wissenstransfer zwischen den Domänen. Der Kommunikationsmangel führt zu der Benutzung veralteter Techniken und verlangsamt die Weiterentwicklung von den Tools. Häufig werden die Tools von den Fachexperten entwickelt die über das nötige Fachwissen verfügen, jedoch keine ausreichende Informatik Kenntnisse haben. Wenn die Visual Analytics Tools jedoch von Informatikern entwickelt werden, ist es wichtig, dass die Verbindung zu der Domäne nicht verloren geht.

Im Grundprojekt werden durch die gewonnenen Kenntnisse aus Visual Analytics visuelle Komponenten für OpenVigil Framework entwickelt.

Literaturverzeichnis

[1] Richtlinie 2001/83/eg.

[2] URL `http://openvigil.sourceforge.net/`. Abgerufen am 31.10.2018.

[3] R. Fernandez and N. Fetais. Survey of information visualization techniques for enhancing visual analysis. In *2017 International Conference on Computer and Applications (ICCA)*, pages 360–363, September 2017. https://doi.org/10.1109/COMAPP.2017.8079755.

[4] M. Hauben and J. K. Aronson. Defining 'signal' and its subtypes in pharmacovigilance based on a systematic review of previous definitions. *Drug Safety*, vol. 32(no. 2):pages 99–110, 2009.

[5] D. Keim, G. Andrienko, and J. Kohlhammer G. Melançon J. D. Fekete, C. Görg. Visual analytics: Definition, process, and challenges. vol 4950. Springer, Berlin, Heidelberg, 2008.

[6] D. Keim, F. Mansmann, A. Stoffel, and Hartmut Ziegler. *Visual Analytics*, pages 3341–3346. Springer US, Boston, MA, 2009. ISBN 978-0-387-39940-9. https://doi.org/10.1007/978-0-387-39940-9$_1$122. URL `https://doi.org/10.1007/978-0-387-39940-9_1122`.

[7] D. Keim, J. Kohlhammer, G. Ellis, and F. Mansmann. *Mastering the Information Age Solving Problems with Visual Analytics*. 2010.

[8] Z. Konyha, A. Lež, K. Matković, M. Jelović, and H. Hauser. Interactive visual analysis of families of curves using data aggregation and derivation. In *Proceedings of the 12th International Conference on Knowledge Management and Knowledge Technologies*, i-KNOW '12, pages 24:1–24:8, New York, NY, USA, 2012. ACM. ISBN 978-1-4503-1242-4. https://doi.org/10.1145/2362456.2362487. URL `http://doi.acm.org/10.1145/2362456.2362487`.

[9] A. Kühl. Anwendung des mgps-algorithmus (multi-item-poisson-shrinker) für die analyse von pharmakovigilanzdaten. Christian-Albrechts-Universität zu Kiel Institut für Informatik, Bachelorarbeit, 2015.

[10] A. Paivio. *Imagery and Verbal Processes*. pages 33-34.

[11] A. Shibata and M. Hauben. Pharmacovigilance, signal detection and signal intelligence overview. In *14th International Conference on Information Fusion*, pages 1–7, July 2011.

[12] M. Tropmann-Frick and J. Sm. Andersen. Visual data science - an exploration from visualization to a vision. Hamburg University of Applied Sciences, Department of Computer Science, 2019.

[13] C. Ware. *Information Visualization: Perception for Design*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 3 edition, 2012. ISBN 9780123814647, 9780123814654.