Hochschule für Angewandte Wissenschaften Hamburg
Hamburg University of Applied Sciences

Ausarbeitung

Jessica Broscheit

Explorative Visualisierung von Wissen aus Datenbanken

Fakultät Technik und Informatik
Department Informatik

Faculty of Engineering and Computer Science
Department of Computer Science

Jessica Broscheit

Explorative Visualisierung von Wissen aus Datenbanken

Ausarbeitung
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
Eingereicht am 21.08.2015

Jessica Broscheit

Thema der Ausarbeitung
Explorative Visualisierung von Wissen aus Datenbanken

Stichworte
Explorative- und prädiktive Datenanalyse, Big Data, Knowledge Discovery in Databases,
Data-Mining, Machine Learning, Entscheidungsbaum, Clusteranalyse, Visualisierungsbibliothek

Kurzzusammenfassung
Diese Arbeit beschäftigt sich mit der Frage, wie Wissen aus großen Datenmengen sichtbar gemacht werden kann. Dabei wird im Folgenden der historische Hintergrund der explorativen Datenanalyse betrachtet, computergestützte Ansätze der Datengewinnung beschrieben und die Möglichkeiten für eine explorative Visualisierung im Browser gezeigt.

Inhaltsverzeichnis

1 Einleitung

»The greatest value of a picture is
when it forces us to notice what we never expected to see«
John W. Tukey

Herzschrittmacher, Barbiepuppe, Eierkarton, Auto, Laterne, Fahrrad, Telefon, Heizung, Ampel, Kochtopf, Abfalleimer, Schlüsselanhänger, Stereoanlage, Lichtschalter, Fotoapparat, Duschkopf, Kopfkissen, Himmel, Blume, Brille, Schuh, Tablette, Uhr... alltägliche Gegenstände werden „smart“ und erstellen immer genauere Informationen über uns und unser Verhalten.

Das Internet der Dinge entwickelt die Welt zu einem gut vernetzen System und lässt dadurch die Datenberge um ein weiteres wachsen. Big Data ist allgegenwärtig und verbreitet eine Goldgräberstimmung für noch ungeahntes Wissen, was nicht nur von Wirtschaft, Wissenschaft und Medien, sondern auch von der Gesellschaft genutzt werden möchte. Denn die Analyse von Daten verspricht nicht nur effizientere Abläufe, sondern auch ein verbessertes Leben. Doch weil die Daten nicht linear, sondern exponentiell wachsen, müssen neue technische Verfahren für die Datenanalyse benutzt werden, um an die neuen Erkenntnisse zu gelangen [vgl. Schmidt (2013), Schrader (2011)].

Diese Arbeit beschäftigt sich mit der Frage, wie Wissen aus großen Datenmengen sichtbar gemacht werden kann. Dabei wird im Folgenden der historische Hintergrund der explorativen Datenanalyse betrachtet, computergestützte Ansätze der Datengewinnung beschrieben und die Möglichkeiten für eine explorative Visualisierung im Browser gezeigt.

2 Grundlagen und Hintergrund

2.1 Explorative Datenanalyse

Definition: Die explorative Datenanalyse (engl. explorative data analysis, EDA) ist ein Teilgebiet der Statistik und dient dem Entdecken von Zusammenhängen zwischen verschiedenen Variablen.

In den 1970er Jahren führte der US-Amerikanische Statistiker John W. Tukey den Begriff ein und machte darauf aufmerksam, dass ein zu großer Schwerpunkt auf die Auswertung von gegeben Hypothesen gelegt wird. In seinem Buch EXPLORATORY DATA ANALYSIS [vgl. Tukey (1977)] stellte er neue Verfahren für hypothesen-generierende Statistiken vor, die mit grafischen Methoden, wie zum Beispiel Boxplot, Histogramm, QQ-Plot, Scatterplot, Mosaikplot dargestellt werden können und noch heute im Data-Mining verwendet werden.

Zu eines der ersten EDA-Beispielen gehört die Cholera Karte von Dr. John Snow aus dem Jahre 1854. Durch die Visualisierung von Pumpen und Cholera-Todesfällen entdeckte Snow eine kontaminierte Wasserstelle auf der Broad Street und ließ diese schließen. Dadurch kam die Epidemie zum Stillstand und seine These dass verunreinigtes Trinkwasser die Ursache für Cholera sein könnte wurde belegt [vgl. Gilbert (1958)].

Abb. 1: Snow, John: Cholera Karte (1854)

3 Wissensentdeckung in Datenbanken

3.1 Big Data

»You can have data without information,
***but you cannot have information without data.* «**
Daniel Keys Moran

Definition: »Big Data beschreibt Datenbestände, die aufgrund ihres Umfangs, Unterschiedlichkeit oder ihrer Schnelllebigkeit nur begrenzt durch aktuelle Datenbanken und Daten-Management-Tools verarbeitet werden können« [Plattner (2013)]

Für viele der heutigen Analysen werden Datenmengen verwendet, die nicht mehr so überschaubar sind wie in dem Beispiel von John Snow (Siehe Abb. 1). Das täglich produzierte Datenvolumen setzt sich aus unstrukturierten Formaten, wie Fotos, Videos, Text, Geoinformationen, GPS-Signalen, Kauf-Transaktionen und vielen mehr zusammensetzen. Diese komplexe und agile Datenmenge wird mit dem Begriff Big Data (Massendaten) beschrieben und lässt sich durch die drei Dimensionen (Volume, Velocity, Variety, kurz: 3V) erklären [vgl. Laney (2001)]:

Volume (Volumen): Das Volumen besteht nicht nur aus der Verarbeitung von Textdaten, sondern auch aus unterschiedlichen Datensätze wie z.B. Videos, Bilder und Musik.

Velocity (Geschwindigkeit): Die Geschwindigkeit beschreibt die Häufigkeit, mit der Daten erzeugt, erfasst und gemeinsam genutzt werden.

Variety (Vielfalt): Die Vielfalt beschreibt polystrukturierte Daten, die aus unterschiedlichsten Formaten bestehen und dadurch nur schwer vergleichbar sind.

3.2 Der Prozess Knowledge Discovery in Databases

»Computers have promised us a fountain of wisdom
but delivered a flood of data«
W. Frawley, G. Piatetsky-Shapiro, and C.Matheus

Definition: Der Begriff Knowledge Discovery in Databases (kurz: KDD) wurde durch Fayyad, geprägt und wie folgt definiert: »Knowledge Discovery in Databases is the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data.« [Fayyad u.a. (1996)]

Knowledge Discovery in Databases beschreibt somit den Gesamtprozess der Wissensentdeckung in Datenbanken. Ziel dieser Prozedur ist es bislang unbekanntes Wissen aus Big Data zu gewinnen und neue Zusammenhänge zu entdecken.

Im KDD-Prozess werden folgende Schritte durchlaufen (siehe Abb. 2): Im ersten Schritt werden Rohdaten aus Datenbanken ausgewählt. Die Zieldaten durchlaufen dann den Prozess der Vorverarbeitung. Die aufbereiteten Daten werden anschließend umgewandelt und dem Schritt Data-Mining zur Verfügung gestellt. An dieser Stelle kann der Nutzer unterschiedliche Methoden des Data-Mining für die Analyse nutzen. Durch die Interpretation von Mustern entstehen neue Hypothesen, die im Anschluss überprüft werden und so neue Erkenntnisse hervorbringen.

Um den KDD-Prozess anzuwenden kann zum Beispiel die Software RapidMiner verwendet werden.

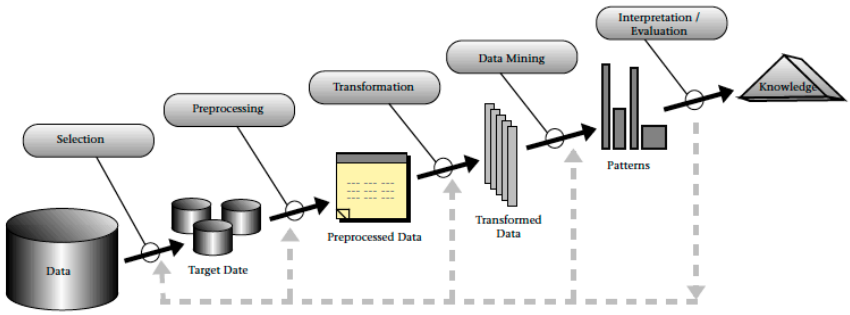

Abb 2: Der KDD-Prozess in einzelnen Schritten
Fayyad, Piatetsky-Shapiro, Smyth: From Data Mining to Knowledge Discovery in Databases (1996)

3.3 Data-Mining

»You can use all the quantitative data you can get,
but you still have to distrust it and
use your own intelligence and judgment. «
Alvin Toffler

Definition: Data-Mining beschreibt einen Teil im gesamten KDD-Prozesses und wurde durch Decker und Focardi folgendermaßen definiert: »Data Mining is a problem-solving methodology that finds a logical or mathematical description, eventually of a complex nature, of patterns and regularities in a set of data.« [Decker (1995)]

Data Mining ist ein interdisziplinäres Gebiet und setzt sich aus Statistiken, Maschinelles Lernen und Datenbank-Technologie zusammen. In Data-Mining wird die eigentliche Mustererkennung auf die bereits vorverarbeiteten Daten angewendet. Somit kann die Vergangenheit untersuchen und mit Hilfe von Modellierungstechniken neue Hypothesen entdeckt werden [vgl. Sayad (2011)]. »Die verschiedenen Data-Mining-Ziele können anhand unterschiedlicher Kriterien unterteilt werden. Nachfolgend wird die Einteilung in Klassifikation, Regression, Segmentierung und Abhängigkeits-analyse verwendet.« [Chamoni (2006)]:

Klassifikation: Elemente die keiner Klassifikation zugeordnet sind, werden Aufgrund ihrer Merkmale in bestehende Klassen einsortiert.

Regression: Durch die Modellierung von statistischen Zusammenhängen und unterschiedlichen Attributen wird eine Prognose von fehlenden Attributwerten ermöglicht.

Segmentierung: Durch die Ballung von ähnlichen Objekten können Gruppierungen in Daten erkannt werden.

Abhängigkeitsanalyse: Die Suche nach starken Regeln stellt Beziehungen zwischen zwei oder mehreren Objekten her. Sie wird häufig in der Warenkorbanalyse als Abhängigkeitsmodellierung angewendet. Zum Beispiel: Wenn Windeln -> dann Bier.

3.4 Maschinelles Lernen in der Datenanalyse

Definition: **»**Maschinelles Lernen (engl. Machine Learning) ist ein Oberbegriff für die „künstliche“ Generierung von Wissen aus Erfahrung: Ein künstliches System lernt aus Beispielen und kann nach Beendigung der Lernphase verallgemeinern**«** [Wikipedia].

Im Bereich der Datenanalyse kann Maschinelles Lernen zur Beschaffung und Auswertung von Daten genutzt werden. Hierzu werden Algorithmen mit Anweisungen geschrieben und lassen sich dabei grundsätzlich in Überwachtes Lernen und Unüberwachtes Lernen einordnen.

Überwachtes Lernen (engl. supervised learning). Mit dem Verfahren, wie dem Entscheidungsbaum (engl. Decision Trees) können einströmende Daten durch Befehle so klassifiziert werden, dass eine zielgerichtete Vorhersage möglich wird. Das Beispiel Playing Tennis von Tom M. Mitchell (Abb. 3) zeigt einen Entscheidungsbaum für die Vorhersage, bei welchem Wetter Tennis gespielt wird. Dabei wird die Datenmenge in immer kleinere Teilmengen gebrochen, bis am Ende ein Baum mit Entscheidungsknoten und Blattknoten entsteht. Hierzu werden Algorithmen mit Anweisungen geschrieben und durch Trainings- und Testdaten auf ihre Qualität geprüft werden. [vgl. Mitchell (1997)]

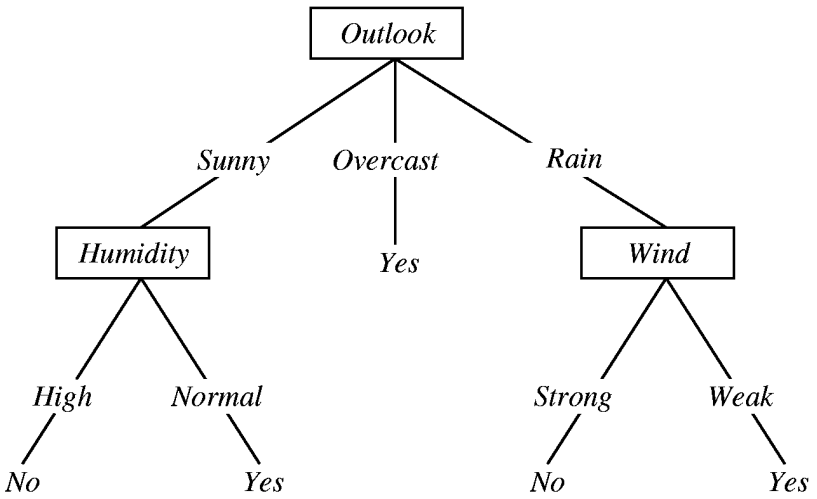

Abb. 3: Tom M. Mitchell: Playing Tennis, Machine Learning, McGraw Hill (1997)

Unüberwachtes Lernen (engl: unsupervised learning) kann in der Regel als eine Methode zur Modellierung der Wahrscheinlichkeitsdichte betrachtet werden. Hierbei wird versucht in den Eingabedaten Muster zu erkennen, die vom strukturlosen Rauschen abweichen [vgl. Hinton (1999)]. Dieser Ansatz wird zum Beispiel für die Clusteranalyse verwendet. Mit Hilfe des K-Means-Algorithmus, können Objekte automatisch erkannt und Gruppen zugeordnet werden. (Siehe Abb. 4)

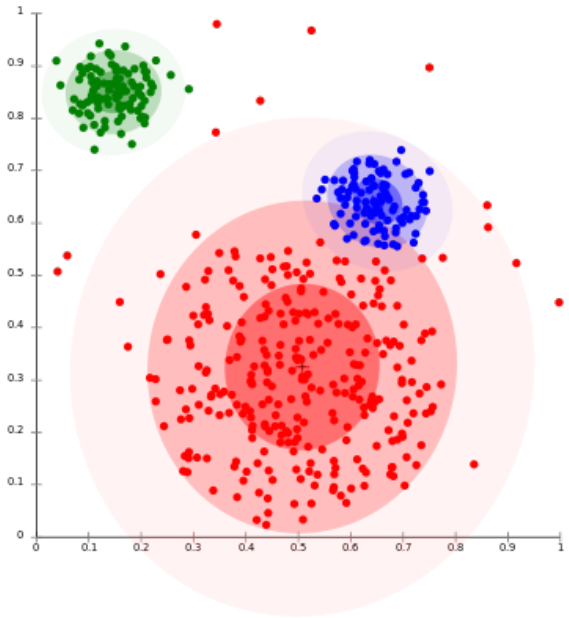

Abb. 4: Ergebnis einer Clusteranalyse

4 Visualisierung im Browser

4.1 Datenvisualisierung

»Every single pixel should testify directly to content.«
Edward Tufte

Die Visualisierung soll zu einem leichtern Verständnis der neu gewonnenen Erkenntnis führen. Dabei steht in erster Linie der Inhalt im Vordergrund und nicht das Design. Eine gelungene Visualisierung präsentiert somit Substanz, Statistik und Design und verfolgt die komplexe Idee von Klarheit, Präzision und Effizienz. Durch einen geringen Einsatz von Farbe auf kleinstem Raum, soll dem Betrachter in kürzester Zeit ein wahrheitsgetreues Ergebnis vermittelt werden [vgl. Tufte (1983)].

Damit die Visualisierungen auch mit der Geschwindigkeit der einströmenden Daten Schritt halten kann, macht es Sinn die Analysen dort zu präsentieren, wo ein ständiger Echtzeit Zugriff möglich ist. Durch eine Präsentation im Webbrowser hat der Betrachter nicht nur die Möglichkeit direkte Veränderungen zu beobachten sondern auch durch reagierendes (responsive) Design Ergebnisse auf unterschiedlichen Endgeräten auszugeben. Die Daten verhalten sich somit dynamisch, können immer aktuell gehalten und interaktiv betrachtet werden.

4.2 Open-Source Bibliothek

Mit Data-Driven Documents (kurz: D3.js) entwickelten Mike Bostock, Jeffrey Heer und Vadim Ogievetsky, unterstützt durch die D3 Open-Source Community, eine umfangreiche Bibliothek für Datenvisualisierungen. Durch das Einhalten von Web-Standards und mit Hilfe von HTML5, JavaScript, CSS und skalierbarer Vektorgrafik (kurz: SVG) ermöglicht D3 eine schnelle Verarbeitung von großen Datenmengen und unterstützt dynamisches Verhalten für Interaktion und Animation im Web.

4.3 Beispiel

Das Beispiel (Abb. 5) zeigt eine praktische Abwendung von explorativer Datenvisualisierung in einem Online Artikel der New York Times. Es zeigt einen ähnlichen räumlichen Ansatz, wie die Karte von John Snow (Abb. 1). Allerdings wird im Vergleich zu John Snow's Karte schnell deutlich, welche Datenmassen heutzutage verarbeitet werden.

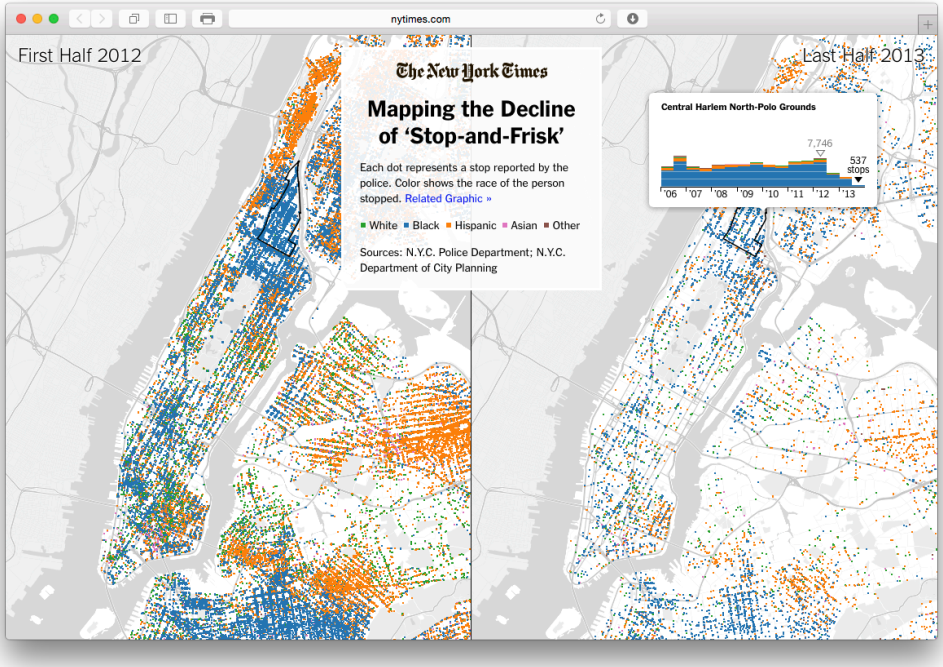

Abb. 5: Mike Bostock und Ford Fessenden:
'Stop-and-Frisk' Is All but Gone From New York" New York Times, 2014

5 Fazit

Ziel der vorliegenden Ausarbeitung war es, sich mit der explorativen Visualisierung von Wissen aus Datenbanken auseinander zusetzen. Dabei ergab sich ein breites und interdisziplinäres Feld, was nur teilweise betrachtet werden konnte, aber dennoch zu folgenden Erkenntnissen führte.

Die Grundidee der explorativen Datenanalyse bleibt im Prinzip gleich. Doch erst durch die Computer-Technologie können große Datenmengen ausgewertet, verarbeitet und in Echtzeit angezeigt werden. Dadurch gewinnt die Datenanalyse an Qualität und Transparenz. Denn durch die Interaktion im Browser erhält der Betrachter nicht nur die Möglichkeit Analysen aus unterschiedlichen Perspektiven zu untersuchen sondern den Prozess der Hypothese auch leichter nachzuvollziehen. Außerdem kann man sich vorstellen, dass Auswertungen den „klassischen“ Analyseraum am Monitor verlassen und durch das Internet der Dinge oder Virtual Reality weitere explorative Konzepte entstehen.

Allerdings führte die Untersuchung auch zu klärungsbedarf bei den Punkten Datenrecht und Datenschutz: Durch die Möglichkeit automatische Roboter nach Daten suchen zu lassen (engl: Scraped Data), stellt sich die Frage, welche Daten urheberrechtlich geschützt sind und wem die Daten überhaupt gehören. Zudem wird der Mensch durch die Auswertung von Big Data in Muster zugeordnet, egal ob man Informationen von sich Preis gibt oder nicht. Wird sich dadurch unsere Gesellschaft zu einer homogenen Masse verformen, die nur noch effizient handelt?

6 Abbildungsverzeichnis

Literaturverzeichnis

[Schmidt (2013)] Schmidt, Eric & Cohen, Jared:
„Die Vernetzung der Welt: Ein Blick in unsere Zukunft“ Rowohlt, 2013

[Schrader (2011)] Schrader, Christopher: „Explosion des Cyberspace“ (2011)
URL: sueddeutsche.de/digital/datenwachstum-der-digitalisierten-welt-explosion-des-cyberspace-1.1058394 (17.08.2015)

[Tukey (1977)] Tukey, John W.: „Explroatory Data Analysis“ Pearson, 1977

[Gilbert (1958)] Gilbert, E.W.: „Pioneer Maps of Health and Disease in England"
Geographical Journal, 124 (1958), 172-183.

[Plattner (2013)] Plattner, Prof. Dr. Hasso: „Big Data“ (2013)
URL: enzyklopaedie-der-wirtschaftsinformatik.de (17.08.2015)

[Laney (2001)] Laney, Doug: „3D Data management: Controlling Data Volume, Velocity and Variety“
URL: blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf (17.08.2015)

[Fayyad u.a. (1996)] Usama Fayyad, Gregory Piatestsky-Shapiro, Padhraic Smyth:
From Data Mining to Knowledge Discovery in Databases (AAAI, 1996)
URL: aaai.org/ojs/index.php/aimagazine/article/viewArticle/1230 (03.08.2015)

[Decker (1995)] Decker, K.; Focardi, S.: Technology overview: a report on data mining. Swiss Federal Institute of Technology (ETH Zurich) Technical Report CSCS TR-95-02, Zürich, 1995.

[Sayad (2011)] Sayad, Dr. Saed: An Introduction to Data-Mining
URL: saedsayad.com/data_mining.htm (17.082015)

[Chamoni (2006)] Chamoni, Prof. Dr. Peter: „Data Mining“ (2013)
URL: enzyklopaedie-der-wirtschaftsinformatik.de/wi-enzyklopaedie/lexikon/daten-wissen/Business-Intelligence/Analytische-Informationssysteme--Methoden-der-/Data-Mining/index.html/?searchterm=Knowledge%20Discovery%20in%20Databases (17.082015)

[Wikipedia] URL: wikipedia.org/w/index.php?title=Maschinelles_Lernen&redirect=no (17.082015)

[Hinton (1999)] Hinton, Geoffrey & Senjnowski, Terrence J.:
„Unsupervised Learning: Foundations of Neural Comutation“ MIT Press, (1999)

[Tufte (1983)] Tufte, Edward „The Visual Display of Quantitative Information“ Graphics Press, 1983

[Offbook (2013)] Offbook „The Art of Data Visualization“ 2013
URL: youtube.com/watch?v=AdSZJzb-aX8 (17.082013)

[Mitchell (1997)] Tom M. Mitchell: Decision Tree Learning
URL: jmvidal.cse.sc.edu/talks/decisiontrees/allslides.html
(03.08.2015)

R2D3, A visual introduction to machine learning.
URL: r2d3.us/visual-intro-to-machine-learning-part-1
(03.08.2015)

WIRED: Meet the man Google hired to make AI a reality,
URL: www.wired.com/2014/01/geoffrey-hinton-deep-learning/
(08.08.2015)

Hrsg.: Kurbel, Becker, Gronau; Sinz, Stuhl: Enzyklopädie der Wirtschaftsinformatik
URL: enzyklopaedie-der-wirtschaftsinformatik.de
(11.08.2015)

Versicherung über Selbstständigkeit

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

Hamburg, 21.08.2015

Ort, Datum	Unterschrift