Projektbericht zum Modul

Information Retrieval und Visualisierung

Wintersemester 2021/22

**Visualisierung**

**von Nährstoffen und Kalorien**

**in Lebensmitteln**

Name: Delia Storch

Matrikelnummer: 220233073

GitHub Repository:

20.12.2021

# Inhaltsverzeichnis

[Inhaltsverzeichnis II](#_Toc90400744)

[Abbildungsverzeichnis III](#_Toc90400745)

[1. Einleitung 1](#_Toc90400746)

[1.1 Anwendungshintergrund 1](#_Toc90400747)

[1.2 Zielgruppen 1](#_Toc90400748)

[1.3 Überblick und Beiträge 2](#_Toc90400749)

[2. Daten 3](#_Toc90400750)

[2.1 Technische Breitstellung der Daten 3](#_Toc90400751)

[2.2 Datenvorverarbeitung 4](#_Toc90400752)

[3. Visualisierung 5](#_Toc90400753)

[3.1 Analyse der Anwendungsaufgaben 5](#_Toc90400754)

[3.2 Anforderungen an die Visualisierungen 6](#_Toc90400755)

[3.3 Präsentation der Visualisierungen 7](#_Toc90400756)

[3.3.1 Visualisierung Eins 7](#_Toc90400757)

[3.3.2 Visualisierung Zwei 8](#_Toc90400758)

[3.3.3 Visualisierung Drei 9](#_Toc90400759)

[3.4 Interaktion 10](#_Toc90400760)

[4. Implementierung 11](#_Toc90400761)

[5. Anwendungsfälle 12](#_Toc90400762)

[5.1 Anwendung Visualisierung Eins 12](#_Toc90400763)

[5.2 Anwendung Visualisierung Zwei 13](#_Toc90400764)

[5.3 Anwendung Visualisierung Drei 14](#_Toc90400765)

[6. Verwandte Arbeiten 16](#_Toc90400766)

[7. Zusammenfassung und Ausblick 16](#_Toc90400767)

[Literaturverzeichnis i](#_Toc90400768)

[Anhang ii](#_Toc90400769)

[Git Historie ii](#_Toc90400770)

# Abbildungsverzeichnis

[Abbildung 1: Scatterplot 8](#_Toc90398123)

[Abbildung 2: Parallele Koordinaten 9](#_Toc90398124)

[Abbildung 3: Ausschnitt aus der Baumhierarchie 10](#_Toc90398125)

[Abbildung 4: Vergleich von Proteinen und Kohlenhydraten im Scatterplot 12](#_Toc90398126)

[Abbildung 5: Vergleich der Nährstoffgehalte in Parallelen Koordinaten 13](#_Toc90398127)

[Abbildung 6: Lebensmittelkategorie Getreideprodukte in der Baumhierarchie 14](#_Toc90398128)

[Abbildung 7: Lebensmittelkategorie Nüsse in der Baumhierarchie 15](#_Toc90398129)

[Abbildung 8: Lebensmittelkategorie Früchte in der Baumhierarchie 15](#_Toc90398130)

# Einleitung

Das Ziel dieses Projekts ist es, die Daten so aufzubereiten, dass Interessierte der Zielgruppe einen umfassenden Überblick über die ausgewählten Lebensmittel mit ihren entsprechenden Nährstoffen und Kalorien erhalten. Im Fokus steht dabei die Vergleichbarkeit der Nährwerte und Kalorien pro 100 Gramm des Produkts. Dafür sollte auf den ersten Blick verständlich sein, welche Beziehung zwischen zwei oder mehreren ausgewählten Nährstoffen der Lebensmittel besteht. Von besonderer Bedeutung ist außerdem die Darstellung verschiedener Lebensmittelkategorien mit ihren Produkten und einer Übersicht der entsprechenden Kalorien. Die Umsetzung des Projektziels erfolgt anhand drei ausgewählter Visualisierungstechniken mit der Programmiersprache Elm.

## Anwendungshintergrund

Die Umsetzung des Projekts erfolgt mithilfe von drei Visualisierungstechniken. Die ausgewählten Visualisierungstechniken sind Scatterplot, Parallele Koordinaten und Baumhierarchie.

Die Visualisierungstechniken ermöglichen die Darstellung der Daten aus dem zugrunde liegenden Datensatz. Der Datensatz enthält Informationen über Lebensmittel, einschließlich Lebensmittelname, Kategorie, Kalorien, Proteine, Fette, gesättigte Fette, Ballaststoffe und Kohlenhydrate.

Kategorie

Kalorien

Proteine

Fette

Gesättigte Fette

Ballaststoffe

Kohlenhydrate

Outro

## Zielgruppen

Für die vorliegenden Daten kommen drei potenzielle Zielgruppen in Frage: ernährungsbewusste Konsument:innen, Ernährungswissenschaftler:innen und Produktentwickler:innen aus der Lebensmittelindustrie. Im Folgenden werden die Zielgruppen der Visualisierungen analysiert.

Ernährungsbewusste Konsument:innen weisen ein besonders großes Interesse an den Lebensmitteln auf, die sie regelmäßig konsumieren. Sie achten auf eine ausgewogene Ernährung und verfolgen dabei Ziele wie die Reduktion oder Ergänzung spezieller Nährwerte und Kalorien oder den Vergleich verschiedener Nährwerte von Produkten aus gleichen Produktkategorien. Da sich die ernährungsbewussten Verbraucher:innen dieser Zielgruppe inhaltlich mit dem Thema Ernährung auseinandersetzen, ist davon auszugehen, dass sie über ein umfangreiches Wissen verfügen und in der Lage sind, Lebensmitteldaten intuitiv erfassen und einordnen zu können. Diese Zielgruppe kann ihr Vorwissen mithilfe der Visualisierungen um weitere Produkte mit ihren Nährstoffen und Kalorien erweitern, die Nährwerte ähnlicher Produkte vergleichen und darauf basierend Kaufentscheidungen für Lebensmittel zu ihrem gesundheitlichen Vorteil treffen.

Ernährungswissenschaftler:innen verfügen über ein ausgezeichnetes und umfangreiches Wissen zum Thema Ernährung. Von großer Relevanz ist für sie die Zusammensetzung von Lebensmitteln in Bezug auf ihre Nährstoffe, da sie Aufschluss über die Menge gibt, um den maximalen Tagesbedarf zu erreichen, beispielsweise bei der Höchstmenge von 50 Gramm Zucker pro Person pro Tag.

Produktentwickler:innen aus der Lebensmittelindustrie sind sich den Nährwerten von Produkten sicher bewusst. Unter Umständen verfügen sie allerdings nur über Nichenwissen über Lebensmittel aus ihrer spezifischen Branche. Um innovativ zu agieren, müssen Entwickler:innen stets neue Möglichkeiten in Betracht ziehen. So könnte die Substitution von Vorprodukten für eine neue Zusammensetzung des Endprodukts eine Möglichkeit darstellen, um gewisse Grenzwerte wie beispielsweise eine maximale Kalorienanzahl nicht zu überschreiten oder um besonders viel von einem gewissen Nährstofftypen im Endprodukt zu verarbeiten wie beispielsweise ein hoher Anteil an Proteinen in einem Produkt.

## Überblick und Beiträge

Das Interesse an bewusster Ernährung, speziellen Ernährungsformen oder der Zusammensetzung von Lebensmitteln steigt. Im Internet und in zahlreichen Büchern und Zeitschriften sind Informationen zum Thema Ernährung mit unterschiedlichsten Ansätzen zu finden. Um sich schnell und unkompliziert einen ersten Überblick zu verschaffen und einen Eindruck der Nährstoffgehalte und Kalorien in Lebensmitteln zu gewinnen, eignet sich eine Visualisierung entsprechender Daten. Diese soll möglichst so gestaltet sein, dass Anwender:innen keine Programmierkenntnisse benötigen, um mit den Darstellungen interagieren und Informationen erhalten zu können.

Die Visualisierungstechnik Scatterplot wurde gewählt, da eine solche Darstellung als allgemein bekannt und intuitiv zu bedienen eingeschätzt wird. Ein Scatterplot bietet die Möglichkeit, schnell gesuchte Informationen zu erhalten, indem gezielt zwei konkrete Nährstoffe ausgewählt und gegenübergestellt werden. Die Visualisierungstechnik Baumhierarchie stellt die Kalorienverteilung einzelner Lebensmittel, differenziert nach Lebensmittelkategorien, dar. In der Baumhierarchie liegen die Daten strukturiert und übersichtlich vor. Demzufolge ist davon auszugehen, dass Anwender:innen auch bei dieser Darstellung Informationen direkt erfassen können, selbst wenn diese über kaum oder kein Hintergrundwissen verfügen.

Im Gegensatz dazu ist die Visualisierung der Parallelen Koordinaten möglicherweise nicht vergleichbar intuitiv. Zwar ist eine Interaktion mit der Darstellung auch ohne Vorwissen aus dem Bereich der Wirtschaftsinformatik möglich. Jedoch ist es hilfreich, wenn eine gezielte Suche nach Lebensmitteln anhand vorher bekannter charakteristische Nährstoffeigenschaften durchgeführt wird, um aus den Daten den maximal möglichen Mehrwert zu ziehen. Dennoch wird diese Visualisierungstechnik im Vergleich zu potenziellen Alternativen als am besten geeignete Möglichkeit entsprechend ihrem Anwendungshintergrund eingeschätzt.

In Kapitel 2 folgen zunächst Informationen zur verwendeten Datengrundlage und Datenverarbeitung. Eine genaue Beschreibung der Umsetzungen der Visualisierungen erfolgt in Kapitel 3. Kapitel 4 beinhaltet die Implementierung. In Kapitel 5 werden die Anwendungsfälle zu den Visualisierungen aus Kapitel 3 ausgeführt. Verwandte Arbeiten in Kapitel 6 sowie Zusammenfassung und Ausblick in Kapitel 7 bilden den Abschluss dieser Arbeit.

# Daten

Der verwendete Datensatz, auf dem die Visualisierungen basieren, wurde von einer Nutzerin auf der Plattform *Kaggle* bereitgestellt. Die Daten dienen ursprünglich als Grundlage für ein Python-Projekt mit dem Titel „Nutritional facts for most common foods“. Die Datei beinhaltet im Original zehn Spalten mit insgesamt 329 Datensätzen. Bei der Verarbeitung der Daten und deren Anpassung für die Visualisierungen wurde die Datei zunächst um eine Spalte reduziert, anschließend um drei Spalten erweitert und beschädigte Datensätze wurden entfernt. Daraus resultiert eine Datei mit zwölf Spalten und 280 Datensätzen. Es wurden außerdem Anpassungen bei den Bezeichnungen der Lebensmittel und bei deren Kategorien vorgenommen und alle Werte der Kalorien, Proteine, Fette, gesättigten Fette, Ballaststoffe und Kohlenhydrate einheitlich umgerechnet, sodass sich der Wert auf 100 Gramm pro Produkt bezieht. Im Folgenden werden die Schritte der Datenvorverarbeitung konkreter erläutert.

Der originale Datensatz beginnt mit den Lebensmittelnamen in der Spalte *Food*, welche die Produkte bezeichnen. Die folgenden Mengenangaben unter *Measure* stellen eine durchschnittliche Portion dar. Die Angabe des Gewichts unter *Grams* gibt das Gewicht der entsprechenden Portionen an. Die Werte der Kalorien unter *Calories* sowie die Werte für Proteine unter *Proteins*, Fette unter *Fats*, gesättigte Fette unter *Sat.Fat*, Ballaststoffe unter *Fiber* und Kohlenhydrate unter *Carbs* beziehen sich auf das angegebene Gewicht pro Portion des Produktes. Unter *Category* ist für jedes Produkt eine aus insgesamt zwölf Kategorien erfasst.

Grundsätzlich sind die Daten gut geeignet, um genannte Fragestellungen beantworten zu können. Sie beinhalten die wesentlichen Informationen über in Lebensmitteln enthaltene Kalorien und Nährstoffe, die für die Zielgruppe interessant sein könnten. Allerdings enthält der Datensatz einige fehlerhafte Daten in Form von doppelten, nicht zuordenbaren oder unvollständigen Werten, welche entfernt werden müssen und erfordert darüber hinaus eine Erweiterung um drei Spalten, um ihn für die Visualisierungen, insbesondere für die Baumhierarchie, anzupassen.

Hier eventuell noch ergänzen!

## Technische Breitstellung der Daten

Die technische Bereitstellung der verwendeten originalen und verarbeiteten Daten erfolgt über ein öffentliches Github Repository. Die Dateien befinden sich im Ordner *Daten* in den Unterordnern *CSV*, *Excel* und *JSON*. Da die Originaldaten im CSV-Dateiformat vorliegen, werden sie zunächst in Excel eingefügt, bearbeitet und anschließend wieder in CSV überführt sowie in JSON umgewandelt.

Scatterplot und Parallele Koordinaten greifen auf die CSV-Datei mit den verarbeiteten Daten der Lebensmittel zu. Die modifizierte CSV-Datei mit der Bezeichnung *NutrientsFINAL.csv* befindet sich im Ordner *Daten/CSV*. Alle Werte innerhalb der CSV-Datei sind mit einem Komma getrennt und jedes Dezimaltrennzeichen ist ein Punkt. Die CSV-Datei beinhaltet alle IDs und Produkte, drei Kategorien, die einheitlichen Mengenangaben der Produkte in 100 Gramm, Kalorien, Proteine, Fette, gesättigte Fette, Ballaststoffe und Kohlenhydrate.

In der JSON-Datei werden die Beziehungen für die Baumhierarchie dargestellt. Dafür müssen die Daten umgewandelt werden. Die Baumhierarchie greift auf die JSON-Datei mit den verarbeiteten Daten der Lebensmittel zu. Die JSON-Datei hat die Bezeichnung *BaumhierarchieJSON.json* und befindet sich im Ordner *Daten/JSON*.

## Datenvorverarbeitung

Wie bereits in Kapitel 2 beschrieben, erfordert der Originaldatensatz einige Anpassungen. Da nicht alle Attribute für die Visualisierungen relevant sind und wiederum andere für die Visualisierungen gebraucht werden, wurde der Originaldatensatz in sieben Schritten verarbeitet, um die benötigte Ausgangsdatenlage zu erreichen.

Ziel der Datenvorverarbeitung ist außerdem die Aufbereitung in ein leicht zu verarbeitendes Format. Zunächst wurde die CSV-Datei mit den Originaldaten in eine Excel-Datei konvertiert. Diese ist im Ordner *Daten/Excel* mit der Bezeichnung *NutrientsRAW.xlsx* gespeichert.

Für alle Datensätze wurde eine eindeutige fünfstellige ID vergeben, um einen transparenten Datenvorverarbeitungsprozess zu ermöglichen und um in weiteren Verarbeitungsschritten eine eindeutige Zuordnung zu gewährleisten.

Anschließend wurden alle t- und 0-Werte entfernt und die Formatierung der Zellen überprüft und in eine einheitliche Form gebracht. Fehlerhafte Zellen wurden identifiziert und gegebenenfalls durch entsprechende Recherche ergänzt. Die Werte der Kalorien und Nährstoffe wurden auf eine Nachkommastelle formatiert. Letztendlich wurden fehlerhafte Datensätze aufgrund von doppelten, nicht zuordenbaren oder unvollständigen Informationen ausgeblendet.

Im folgenden Schritt wurde eine zweite Tabelle im gleichen Dokument angelegt. Aus Gründen der besseren Lesbarkeit wurden alle Spaltennamen sowie die Namen der einzelnen Lebensmittel und die Kategorienamen angepasst und vereinheitlicht. Es wird davon ausgegangen, dass die Zielgruppen in der Lage sind, englische Informationen zu erfassen. Deswegen wurde auf eine Übersetzung der Lebensmittelnamen verzichtet. Die angegeben Werte für Kalorien und Nährstoffe beziehen sich in der zweiten Tabelle auf 100 Gramm pro Produkt. Im Originaldatensatz hingegen sind die Lebensmittel in Gramm pro Portion angegeben. So richten sich beispielsweise die Kalorien und Nährwerte von Kuhmilch (*Cows‘ milk*) mit der ID *11000* nach der Portionsgröße von einem Quart (*1 qt.*), was umgerechnet einem knappen Liter Milch entspricht. Da sich die Angabe des Gewichts in Gramm (*Grams*) für die verschiedenen Lebensmittel nach den unterschiedlichen Portionsgrößen richtet, ist keine Vergleichbarkeit der einzelnen Lebensmittel möglich. Aufgrund dessen wurde die Portionsgröße aus allen für die Visualisierung relevanten Datensätzen entfernt und alle Werte für Kalorien und Nährstoffe wurden mithilfe eines Umrechnungsfaktors in der *solver*-Spalte auf 100 Gramm pro Produkt umgerechnet. Leere Zellen erhielten den Wert *0*.

Da mittels der vorliegenden Daten keine sinnvolle Visualisierung der Baumhierarchie umsetzbar ist, musste der Datensatz um die Attribute *supercategory* und *caloriescategory* erweitert werden. In der Überkategorie *supercategory* ist für alle Lebensmittel eine Zuordnung zu Nahrung (*Food*) oder Getränke (*Drinks*) erfasst. Die Kategorie der Kalorien *caloriescategory* beinhaltet kategorische Informationen über die Kalorien der Lebensmittel, wobei es zehn verschiedene Kategorien gibt. Dies ermöglicht die strukturierte Integration der Daten in eine JSON-Datei für die Baumhierarchie.

Darüber hinaus ist in der Excel-Datei auch eine Spalte für die Erstellung der JSON-Bezeichnungen enthalten. Die Bezeichnungen wurden bei der Erstellung der JSON-Datei aus der Excel-Datei entnommen und an der entsprechenden Stelle eingefügt. Die JSON-Datei wurde mit Visual Studio Code erstellt und ist im Ordner *Daten/JSON* mit der Bezeichnung *BaumhierarchieJSON.json* gespeichert.

Die zweite Tabelle mit den vollständigen Werten aller Attribute wurde aufgrund der Übersichtlichkeit in einem neuen Excel-Dokument zusammengefasst. Dieses Dokument mit der Version aller finaler Daten stellt die Grundlage für die zu erstellende CSV-Datei dar, die für die Visualisierung von Scatterplot und parallelen Koordinaten benötigt wird. Dabei stellte das Konvertieren der Excel-Datei in eine CSV-Datei eine besondere Herausforderung dar, da dabei zu beachten ist, dass das Dezimaltrennzeichen innerhalb der Werte ein Punkt ist und dass die Daten kommagetrennt sind. Die CSV-Datei ist im Ordner *Daten/CSV* mit der Bezeichnung *NutrientsFINAL.csv* gespeichert.

# Visualisierung

In diesem Kapitel werden die drei Visualisierungen beschrieben. Im Vordergrund stehen dabei die jeweiligen Einsatzmöglichkeiten der Techniken sowie die Anforderungen der Zielgruppe. Dazu gehört der Vergleich von zwei oder mehreren Attributen in Form von Nährstoffen von Lebensmitteln sowie eine geordnete Übersicht über die Kalorien einzelner Lebensmittelkategorien. Dafür werden die Visualisierungen vorgestellt und Interaktionsmöglichkeiten aufgezeigt. Dazu gehört, dass die Anwender selbst wählen können, welche beiden Nährstoffe der Lebensmittel miteinander verglichen werden und in welcher Reihenfolge mehrere Nährstoffe verglichen werden sollen. Da der Wert der Kalorien eines Produktes nicht zu den Nährstoffen zählt, der Vergleich für Interessierte aber trotzdem von Bedeutung ist, soll dieser durch die Umsetzung der Baumhierarchie dargestellt werden.

## Analyse der Anwendungsaufgaben

Durch die Visualisierungen soll eine übersichtliche Darstellung der Informationen zu Kalorien und Nährstoffen ausgewählter Lebensmittel ermöglicht werden. Nutzer:innen soll es möglich sein, die Daten schneller zu erfassen und zusätzliche interessante Erkenntnisse zu gewinnen.

Ziel der ersten Visualisierung ist es, zwei Attribute (Nährstoffe) unterschiedlicher Lebensmittel miteinander zu vergleichen. Dafür wurde die Visualisierungstechnik **Scatterplot** verwendet. Daraus lassen sich Erkenntnisse zwischen Zusammenhängen von zwei ausgewählten Nährstoffen in Lebensmitteln ableiten: *Welche Lebensmittel haben die meisten Kohlenhydrate? Welche Lebensmittel haben am wenigsten Fett? Wie ist das Verhältnis zwischen Ballaststoffen und Eiweißen in Lebensmitteln?* Die Punkte im Scatterplot kennzeichnen die Lebensmittel und zeigen bei der Auswahl mit der Maus den entsprechenden Lebensmittelnamen und die Namen und Werte der zwei ausgewählten Nährstoffe an. Da der Vergleich von zwei Faktoren mittels eines zweidimensionalen Koordinatensystems zu den gängigen Techniken bei der Darstellung von Statistiken gehört, sollten Anwender:innen die Bedeutung der Daten schnell einordnen können. Über einen Hinweis an den Buttons zur Zuweisung des Attributs auf der x- und y-Achse soll Anwender:innen die Benutzung erleichtert werden.

Ziel der zweiten Visualisierung ist es, mehrere Attribute (Nährstoffe) unterschiedlicher Lebensmittel miteinander zu vergleichen. Dafür wurde die Visualisierungstechnik **Parallele Koordinaten** verwendet. Daraus lassen sich grundsätzliche Zusammenhängen von mehreren Nährstoffen in Lebensmitteln ableiten: *Welche Lebensmittel haben umfassend schlechte Nährwerte, beispielweise einen hohen Wert an Kohlenhydraten und Fett, aber wenige Proteine und Ballaststoffe?* Anders als beim Scatterplot werden die Daten der Lebensmittel jetzt auf vier Achsen mit durch Linien verbundenen Attributen dargestellt. Bei der Auswahl der Linien mit der Maus wird der Lebensmittelname sowie Namen und Werte für vier ausgewählte Nährstoffe angezeigt. Auch hier gibt es einen Hinweis an den Buttons zur Zuweisung der Attribute. Die Anwendung ist etwas weniger intuitiv als die des Scatterplots und es könnte hilfreich sein, einen Hinweis zur ordnungsgemäßen Benutzung hinzuzufügen.

Ziel der dritten Visualisierung ist es, die in Nahrung oder Getränken enthaltenen Kalorien abzubilden und dabei die Lebensmittelkategorien sowie eine aufsteigende Sortierung der Produkte pro Kategorie nach Kalorien zu berücksichtigen. Dafür wurde die Visualisierungstechnik **Baumhierarchie**verwendet. Das Baum-Diagramm gibt Auskunft über die folgenden Fragestellungen: *In welchen Lebensmittelkategorien haben die Lebensmittel besonders viele Kalorien? Welche Lebensmittel haben im Einzelnen besonders wenige Kalorien?* Die Anforderungen an die Visualisierung unterscheiden sich grundsätzlich von den beiden vorher genannten Visualisierungen, da hierbei die Kalorien, nicht aber die Nährstoffe abgebildet werden. Zudem sollen die Lebensmittelkategorien stärker in die Visualisierung einbezogen werden. Die Baumdarstellung sollte für Anwender:innen mit unterschiedlichen Wissensständen leicht verständlich sein.

## Anforderungen an die Visualisierungen

Aus dem vorhergehenden Kapitel geht hervor, dass für jede der drei Visualisierungen eine spezifische Zielstellung besteht, aus der sich die zu erfüllenden Anforderungen ergeben.

Da im Rahmen des Projekts angenommen wird, dass die Zielgruppen nicht aus dem Fachbereich der Informatik stammen, ist die verständliche und übersichtliche Darstellung der Daten die bedeutendste Anforderung an die Visualisierungen. Dazu gehört eine sprachlich korrekte sowie optisch gut lesbare Beschriftung einiger Elemente wie Buttons, Achsen, Punkte und Linien. Auch die farblichen Darstellungen der Punkte im Scatterplot sowie der Linien in Parallele Koordinaten sollen zur besseren Übersicht beitragen.

Die erste Visualisierung soll das Verhältnis zwischen zwei Attributen (Nährstoffen) von Lebensmitteln in einem Scatterplot abbilden. Dafür müssen die Koordinatenachsen mit den ausgewählten Attributen beschriftet sein und eine Einteilung erhalten, die sich nach den Zahlenwerten dieser Attribute richtet. Die Zahlenwerte der Attribute sind immer in Gramm pro 100 Gramm des Lebensmittels angegeben. Deswegen wurde auf Einheiten verzichtet. Zusätzlich sollen Anwender:innen die Möglichkeit erhalten, für jede Koordinatenachse eines aus den verfügbaren Attributen auszuwählen. Die Umsetzung erfolgt durch Buttons, wobei es für die x- und die y-Achse die Option gibt, ein Attribut individuell auszuwählen bzw. das Attribut zu wechseln, indem der zugehörige Botton betätigt wird. Aus der Kombination der Zahlenwerte zweier Attribute ergibt sich der zugehörige Punkt im Scatterplot. Damit die Punkte unterschieden werden können, ist es notwendig, dass sie eine eindeutige Beschriftung erhalten, einschließlich Lebensmittelname, ausgewählte Nährstoffe und deren Wert pro 100 Gramm. Die Beschriftung soll allerdings nur angezeigt werden, wenn Anwender:innen die Punkte mit der Maus auswählen, damit sich die Beschriftungen nicht überschneiden.

Die zweite Visualisierung soll den Zusammenhang zwischen mehreren Attributen (Nährstoffen) in Lebensmitteln mithilfe der Visualisierungstechnik Parallele Koordinaten abbilden. Die Einteilung der vier Spalten für die ausgewählten Attribute richtet sich wie beim Scatterplot nach den Zahlenwerten der Attribute in Gramm pro 100 Gramm des Lebensmittels. Die dynamische Anordnung der Spalten soll über Buttons erfolgen, wobei es für jede Spalte die Option gibt, eines aus den verfügbaren Attributen auszuwählen. Die Spalte soll automatisch den Namen entsprechend dem ausgewählten Button erhalten. Lebensmittelnamen, ausgewählte Attribute und deren Zahlenwerte sollen auf zwei verschiedene Arten visualisiert werden, um eine Unterscheidung der Daten zu gewährleisten. Zum einen sollen zusammengehörende Attributwerte durch Linien verbunden werden, die bei der Auswahl mit der Maus farbig angezeigt werden und zum anderen soll die zugehörige Beschriftung dieser Linie angezeigt werden.

Die dritte Visualisierung soll dazu dienen, die Lebensmittel aus dem Datensatz unter Berücksichtigung der Lebensmittelkategorie nach Kalorien zu sortieren und in Form eines Baumdiagramms abzubilden. Um herausfinden zu können, ob ein Produkt viele oder wenige Kalorien beinhaltet, soll die hierarchische Strukturierung so gewählt werden, dass zunächst aus ihr hervorgeht, zu welcher Lebensmittelkategorie die Produkte gehören. Auf der letzten Stufe sollen die Produkte unter der entsprechenden Kalorienkategorie aufgelistet werden. In dieser Visualisierung besteht keine Interaktionsmöglichkeit, durch vertikales oder horizontales Scrollen kann lediglich der Bildausschnitt verändert werden.

## Präsentation der Visualisierungen

Nachdem die Anforderungen an die Visualisierungen analysiert wurden, werden im folgenden Kapitel die drei Visualisierungstechniken vorgestellt. Darüber hinaus werden Interaktionsmöglichkeiten beschrieben.

### Visualisierung Eins

In der ersten Visualisierung ist der Vergleich von zwei Attributen der Lebensmittel in einem Scatterplot dargestellt. Die Koordinatenachsen sind mit den ausgewählten Attributen beschriftet und passen sich in ihrer Skalierung beim Wechsel der Attribute automatisch an. Die Auswahl der Attribute wird durch die Anwender:innen per Button festgelegt und kann für die x- und die y-Achse erfolgen. Jedes Lebensmittel wird im Scatterplot durch einen Punkt visualisiert. Bewegen Anwender:innen die Maus über einen Punkt, wird dieser farbig und ein Text erscheint. Dieser Text setzt sich aus Lebensmittelname und den beiden ausgewählten Attributen zusammen. Der Scatterplot ist in Abbildung 1 dargestellt.

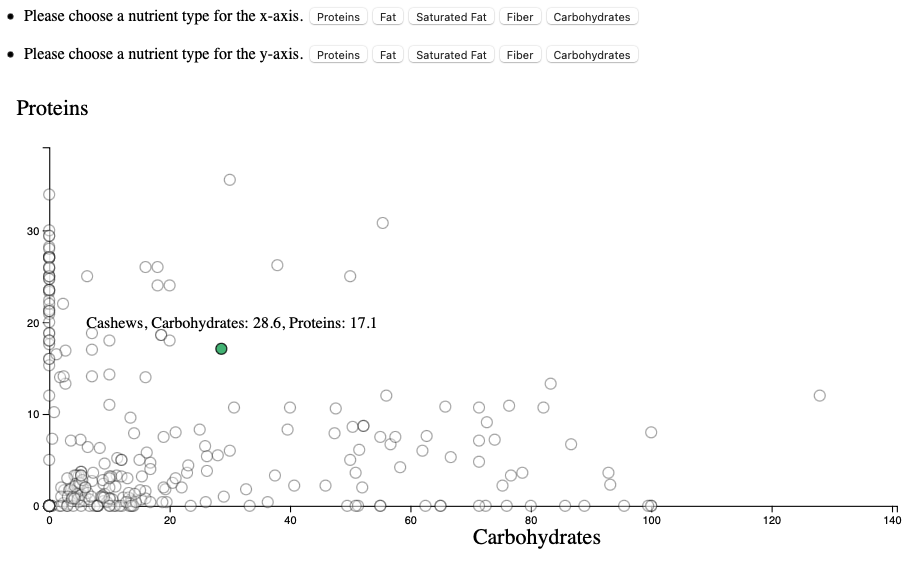


Abbildung 1: Scatterplot (Quelle: eigene Darstellung)

Die bestehenden Anforderungen an den Scatterplot konnten durch die oben beschriebenen Merkmale erfüllt werden, indem die Gegenüberstellung von zwei Nährstoffen für verschiedene Lebensmittel in übersichtlicher Weise visualisiert wurden. Darüber hinaus kann die Darstellung durch Anwender:innen dynamisch angepasst werden.

Der Scatterplot erweist sich im Vergleich zu alternativen Darstellungsmöglichkeiten zweidimensionaler Zusammenhänge als am besten geeignet. Um zwei Attribute ins Verhältnis zu setzen, ist eine Darstellung als Liniendiagramm denkbar, die im Fall der vorliegenden Daten allerdings nicht gewählt wurde. Dies begründet sich darin, dass die Punkte im Liniendiagramm durch Linien verbunden sind, was im Fall der dargestellten Daten als teilweise überlagerte Punkte unübersichtlich wäre. Durch den Scatterplot ist es möglich, die Datenpunkte voneinander zu differenzieren.

### Visualisierung Zwei

Für den Vergleich von mehr als zwei Attributen eignet sich sie Visualisierungstechnik Parallele Koordinaten. Dabei werden vier Attribute von jeweils einer Achse repräsentiert. Die Achsen verlaufen vertikal, liegen mit Abstand parallel nebeneinander und sind mit den ausgewählten Attributen beschriftet. Darüber hinaus erfolgt die Anpassung ihrer Skalierung beim Wechsel der Attribute automatisch. Anwender:innen können für jede der vier Achsen mithilfe von verschiedenen Buttons ein Attribut individuell auswählen. Die Lebensmittel werden durch Linien dargestellt, welche die zugehörigen Datenwerte auf den Achsen schneiden. Bewegen Anwender:innen ihre Maus über eine Linie, wird diese farbig und ein Text erscheint. Dieser Text zeigt den Lebensmittelnamen und die vier ausgewählten Attribute an. Die Parallelen Koordinaten sind in Abbildung 2 dargestellt.

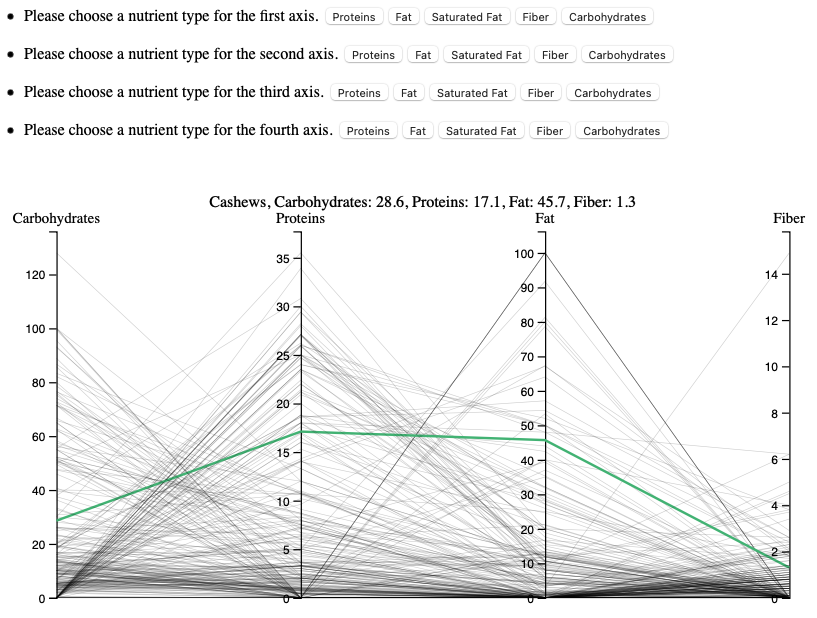


Abbildung 2: Parallele Koordinaten (Quelle: eigene Darstellung)

Da Parallele Koordinaten den Vergleich von mehreren Nährstoffen in Lebensmitteln ermöglichen, konnten die Anforderungen an die Visualisierung erfüllt werden. Dabei können die vier Attribute mithilfe von Buttons ausgewählt werden. Die Lebensmittel werden durch beschriftete und farbige Linien dargestellt, wenn sie mit der Maus ausgewählt werden.

Als alternative Darstellungsformen zu Parallelen Koordinaten kommen Datentinte, Projektion oder Selektion in Frage, da sie sich zur Visualisierung mehrdimensionaler Zusammenhänge eignen. Allerdings sind diese Visualisierungstechniken vor allem bei großer Datenmenge unübersichtlich und im Vergleich zu den Parallelen Koordinaten weniger intuitiv, was zur Fehlinterpretation der Daten führen könnte.

### Visualisierung Drei

Die dritte Visualisierung erfolgt in Form eines Baumdiagramms. Neben den Nährstoffdaten gehen aus den Datensätzen noch weitere Informationen hervor, die in den beiden vorhergehenden Visualisierungen nicht benötigt wurden. Diese sollen in dieser Visualisierung Anwendung finden. Durch die Baumdarstellung wird eine hierarchische Beziehung der Lebensmittelkategorien mit ihren zugehörigen Lebensmitteln abgebildet, wobei der Fokus auf vorletzter und letzter Stufe insbesondere auf der Anzahl der Kalorien in den einzelnen Lebensmitteln liegt. Dafür sind beschriftete Kreise hierarchisch angeordnet und mit Linien verbunden. Ein Ausschnitt aus der Baumhierarchie ist in Abbildung 3 dargestellt.

Stufen erklären + Reihenfolge letzter Knoten



Abbildung 3: Ausschnitt aus der Baumhierarchie (Quelle: eigene Darstellung)

Die Anforderungen an das Baumdiagramm konnten nur teilweise erfüllt werden. Zwar lassen sich dadurch Lebensmittel in Kategorien einordnen und Anwender:innen erhalten einen groben Überblick darüber, welche Lebensmittel in welchen Lebensmittelkategorien eher kalorienarm oder kalorienreich sind. Da die Einteilung der Kategorien für die Kalorien sehr grob gewählt wurde, können keine konkreten Aussagen über die tatsächlichen Kalorienwerte getroffen werden. Es ist lediglich eine Sortierung erkennbar, da die Lebensmittel sich innerhalb eines Rahmens von 100 Kalorien befinden auf der letzten Stufe von links nach rechts aufsteigend sortiert sind. Auch die Gesamtübersicht geht in dieser Visualisierung verloren, da aktives Scrollen zur Seite nötig ist, um bestimmte Informationen zu finden oder alle Daten anzuschauen.

Andere Visualisierungstechniken, wie beispielsweise das Sunburst-Diagramm, eignen sich ebenfalls zur Darstellung von Hierarchien. Dieses bietet zwar den Vorteil einer ansprechenden Optik und zusätzlicher Informationen durch die Größe der Elemente, haben jedoch den Nachteil, dass die Kindelemente ab einer gewissen Stufe kaum noch lesbar sind. Dieses Problem besteht beim Baumdiagramm nicht.

## Interaktion

Wie im vorherigen Kapitel beschrieben, sind in den Visualisierungen verschiedene Interaktionsmöglichkeiten enthalten. Dadurch sollen Anwender:innen die Möglichkeit erhalten, die Darstellungen so zu verändern, dass sie die für sie relevanten Informationen schneller finden. Scatterplot und Parallele Koordinaten enthalten beispielweise Buttons, durch die die Attribute der Achsen ausgewählt werden. Dadurch ist es möglich, dass gezielt ausgewählte Informationen angezeigt werden, während andere aus der Darstellung ausgeblendet werden. Außerdem besteht für die Anwender:innen die Möglichkeit, die Maus über die eingezeichneten Punkte im Scatterplot zu bewegen und so Namen und ausgewählten Attribute anzuzeigen. WARUM? Die Interaktionsmöglichkeit für Parallele Koordinaten ist ähnlich, wobei in diesem Fall Linien die Namen und vier ausgewählte Attribute repräsentieren, wenn Anwender:innen die Maus über die Darstellung bewegen. WARUM? Die Baumhierarchie bietet keine Interaktionsmöglichkeit, da die Datengrundlage in Form einer statischen JSON-Datei vorliegt.

Eine übergreifende Interaktion zwischen den Visualisierungen wurde nicht realisiert, da sie separat codiert wurden. Außerdem unterscheiden sich die Zielstellungen, sodass eine Interaktion zwischen den Visualisierungen nicht erforderlich ist und darüber hinaus die Komplexität erhöhen und die Übersichtlichkeit reduzieren könnte.

# Implementierung

Die Funktionalität der Visualisierungen basiert auf den Codes der Übungen des Moduls „Information Retrieval und Visualisierung“. Dabei zählen insbesondere Übung 1 und Übung 3 als Grundlage für den Code des Scatterplots. Für die Parallelen Koordinaten wurde Übung 7 modifiziert. Weiterhin wurden Teile aus Übung 8 verwendet, um den CSV-Decoder für Scatterplot und Parallele Koordinaten zu erstellen. Als Grundgerüst für die Baumhierarchie diente die Übung 10.

Die Aufteilung der Codes wurde weitestgehend einheitlich strukturiert, damit man schneller einen Überblick gewinnen und sich besser zurechtfinden kann. Alle drei Codes beinhalten zunächst die benötigten importierten Elm-Module. Es folgt die Funktion *main* sowie die Definitionen von *type* und *type alias*. Weiterhin erfolgt die Initialisierung durch die Funktion *init* und die CSV-Daten werden geladen und mittels des CSV-Decoders *decodingNutrients* decodiert. Anschließend ermöglicht die Funktion *update* den dynamischen Wechsel der Attribute an den Achsen. Darüber hinaus schließen sich *subscriptions*, visualisierungsspezifische Designanpassungen und die Sektion *view* an.

Übung 1 diente als Grundlage zur Darstellung der Lebensmittel im Scatterplot. Zudem konnten die Funktionen zum Laden der CSV-Datei übernommen werden. Der CSV-Decoder erforderte allerdings einige Anpassungen, da möglichst viele Informationen aus dem Datensatz visualisiert werden sollten und wurde entsprechend auf mehrere Werte erweitert. Außerdem musste der *type Msg* angepasst werden. Die Funktion *update* wurde basierend auf Übung 3 und Übung 8 aufgebaut, wobei sie festlegt, welche Attribute zuerst ausgewählt sind und was bei der Bedienung eines Buttons passiert. Eine große Herausforderung der ersten Visualisierung bestand in der Verarbeitung der Daten, insbesondere beim Export von Excel zu CSV, da die exportierten Werte zunächst durch ein Semikolon getrennt sind. Für die erfolgreiche Weiterverarbeitung wird allerdings als Trennzeichen ein Komma benötigt. Auch die Einstellung des Dezimaltrennzeichens als Punkt anstelle eines Kommas ist zu beachten. Außerdem bestanden einige Schwierigkeiten bei der Verknüpfung der Daten mit den Buttons, welche jedoch durch den Einsatz genannter Übungen gelöst werden konnten. Der Code des Scatterplots wird über das Github Repository im Ordner *src/Develop* unter dem Namen *ELMScatterplot.elm* bereitgestellt.

Übung 7 wurde für die Visualisierung der Lebensmittel in Parallelen Koordinaten verwendet. Zudem konnte der Code zum Laden der CSV-Datei aus dem Scatterplot übernommen werden. Der Code der Parallelen Koordinaten wird über das Github Repository im Ordner *src/Develop* unter dem Namen *ElmParalleleKoordinaten.elm* bereitgestellt.

Übung 10 konnte nahezu vollständig bei der Umsetzung der Baumhierarchie einfließen. Es ist lediglich eine Anpassung des Pfads der JSON-Datei notwendig. Die Datenbereitstellung für die Baumdarstellung ist allerdings deutlich aufwendiger. Die CSV-Datei musste dafür in ein JSON-Format umgewandelt werden, welches zusätzlich eine Hierarchie festlegt. Dies musste im Rahmen des Projekts manuell mit Visual Studio Code umgesetzt werden. Der Code der Baumhierarchie wird über das Github Repository im Ordner *src/Develop* unter dem Namen *ElmBaumhierarchie.elm* bereitgestellt.

Nach der Fertigstellung drei funktionsfähiger Codes erfolgten designtechnische Anpassungen in Bezug auf Diagramm-, Punkt-, Linien- und Schrifteigenschaften.

# Anwendungsfälle

Die Visualisierungen ermöglichen Erkenntnisse über Lebensmittel und deren Nährstoffe und Kalorien. Im folgenden Kapitel wird für die drei Visualisierungen jeweils ein konkretes Anwendungsszenario vorgestellt.

## Anwendung Visualisierung Eins

Die Visualisierung des Scatterplots in Abbildung 4 zeigt den Vergleich der beiden Attribute Proteine (*Proteins*) und Kohlenhydrate (*Carbohydrates*). Die Anwendung der Visualisierung wird anhand einer Lebensmittelkonsumentin dargestellt, die sich vegan ernährt.

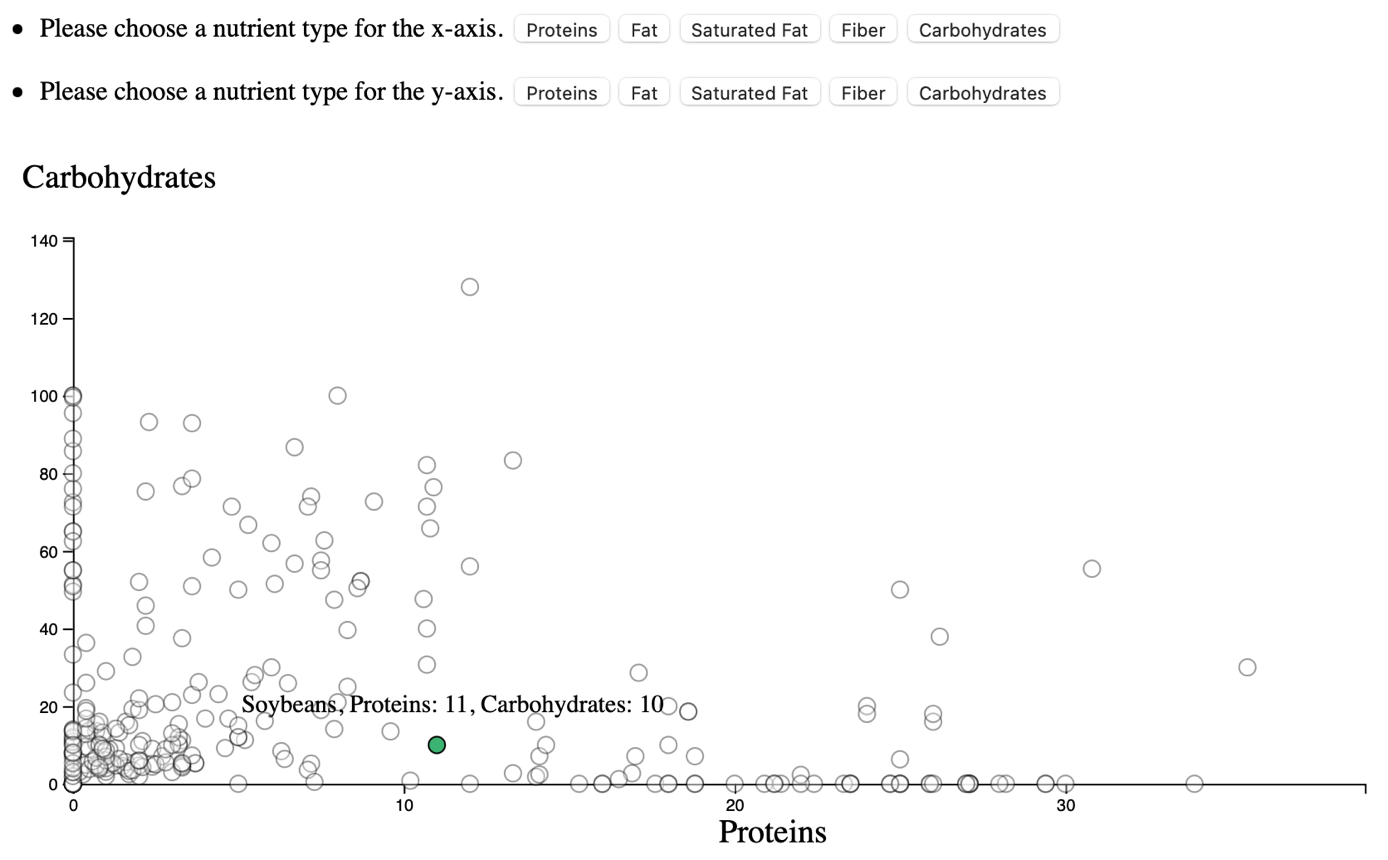


Abbildung 4: Vergleich von Proteinen und Kohlenhydraten im Scatterplot (Quelle: eigene Darstellung)

Die Konsumentin möchte fleischlose Burger zubereiten. Weil sie sich vegan ernährt, muss sie besonders darauf achten, dass sie ausreichend Proteine aufnimmt. Sie möchte aber auch, dass die Basis für ihren Burger nicht übermäßig viele Kohlenhydrate enthält. Sie wählt für die x-Achse das Attribut *Proteins* und für die y-Achse das Attribut *Carbohydrates*.

Bei der Betrachtung von Abbildung 4 fällt auf, dass einige Lebensmittel pro 100 Gramm sehr viele Proteine enthalten. Beim Prüfen dieser Punkte durch die Auswahl mit der Maus stellt die Lebensmittelkonsumentin fest, dass es sich bei Produkten mit sehr hohem Proteingehalt und niedrigem Kohlenhydratgehalt entweder um Fleisch oder Fisch handelt. Grundsätzlich ist daraus abzuleiten, dass tierische Produkte einen hohen Proteingehalt besitzen. Dem gegenüber steht ein sehr geringer Kohlenhydratgehalt – die meisten Produkte dieser Kategorie enthalten gar keine Kohlenhydrate. Bei ihrer Überprüfung weiterer Punkte mit vielen Proteinen lernt die Konsumentin außerdem, dass Nüsse im Allgemeinen sehr proteinreich sind. Diese eignen sich allerdings nicht für ihr Vorhaben. Schließlich entscheidet sie sich dafür, Sojabohnen (*Soybeans*) zu verwenden. Diese sind in Abbildung 4 an dem farbigen (grünen) Punkt mit Beschriftung erkennbar. Sojabohnen beinhalten pro 100 Gramm 11 Gramm Proteine (*Proteins: 11*) und 10 Gramm Kohlenhydrate (*Carbohydrates: 10*) und eignen sich entsprechend ihrer Anforderungen im Vergleich zu den Alternativen der umliegenden Punkte am besten.

## Anwendung Visualisierung Zwei

Die Visualisierung der Parallelen Koordinaten in Abbildung 5 zeigt den Vergleich der vier Attribute Kohlenhydrate (*Carbohydrates*), Proteine (*Proteins*), Fette (*Fat*) und Ballaststoffe (*Fiber*). Die Anwendung der Visualisierung wird anhand eines Ernährungswissenschaftlers dargestellt, der sich im Rahmen seiner Datenrecherche diese Visualisierung ansieht.

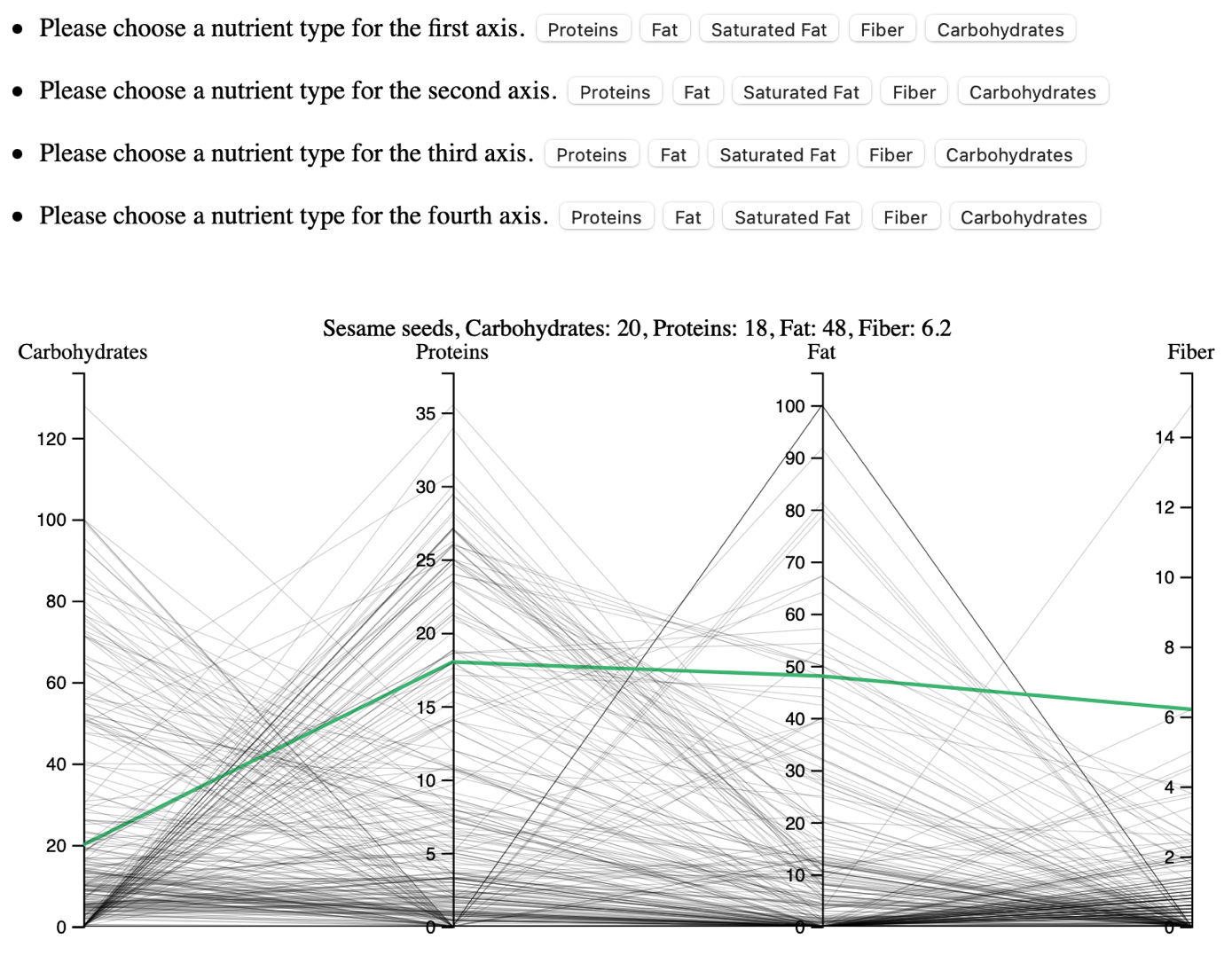


Abbildung 5: Vergleich der Nährstoffgehalte in Parallelen Koordinaten (Quelle: eigene Darstellung)

Der Ernährungswissenschaftler möchte sich über die Nährstoffgehalte von verschiedenen Nüssen informieren und diese vergleichen. Er entscheidet sich dafür, die Attribute Kohlenhydrate (*Carbohydrates*), Proteine (*Proteins*), Fette (*Fat*) und Ballaststoffe (*Fiber*) auszuwählen. Auf den ersten Blick sind keine Muster in den Daten erkennbar.

Da er weiß, dass Nüsse einen hohen Ballaststoffgehalt haben, sucht er entlang der Achse Ballaststoffe (*Fiber*) von oben nach unten. Schnell findet er heraus, dass Sesam (*Sesame seeds*) mit 6,2 Gramm (*Fiber: 6.2*) den höchsten Ballaststoffgehalt pro 100 Gramm hat. Darüber hinaus kann er auf den ersten Blick erkennen, dass Sesam 20 Gramm Kohlenhydrate (*Carbohydrates: 20*), 18 Gramm Proteine (*Proteins: 18*) und 48 Gramm Fett (*Fat: 48*) pro 100 Gramm enthält. Er vergleicht die verschiedenen Produkte und sieht für Nüsse allgemein bestätigt, dass diese neben einem hohen Ballaststoffgehalt auch einen vergleichsweise hohen Fettgehalt haben.

Der Ernährungswissenschaftler besitzt ein umfangreiches Hintergrundwissen und findet sich demzufolge schnell in der Visualisierung zurecht. Zwar ließ sich auf den ersten Blick kein Muster in der Gesamtbetrachtung erkennen, aber aus den Daten geht auch hervor, dass ähnlich Produkte ähnlich Nährstoffgehalte besitzen. So haben beispielsweise Öle und Fette einen hohen Fettgehalt, aber keine Kohlenhydrate, Proteine oder Ballaststoffe. Viele Proteine sind in Fleischprodukten enthalten, diese haben aber kaum Fett und keine Kohlenhydrate oder Ballaststoffe.

## Anwendung Visualisierung Drei

Die Visualisierung der Baumhierarchie in Abbildung 6 zeigt die Lebensmittel der Kategorie Getreideprodukte (*Cereal Products*), geordnet nach Kalorien pro 100 Gramm. Die Anwendung der Visualisierung wird anhand eines Produktentwicklers dargestellt, der sich für eine kalorienarme Zusammensetzung seines neuen Produktes interessiert.

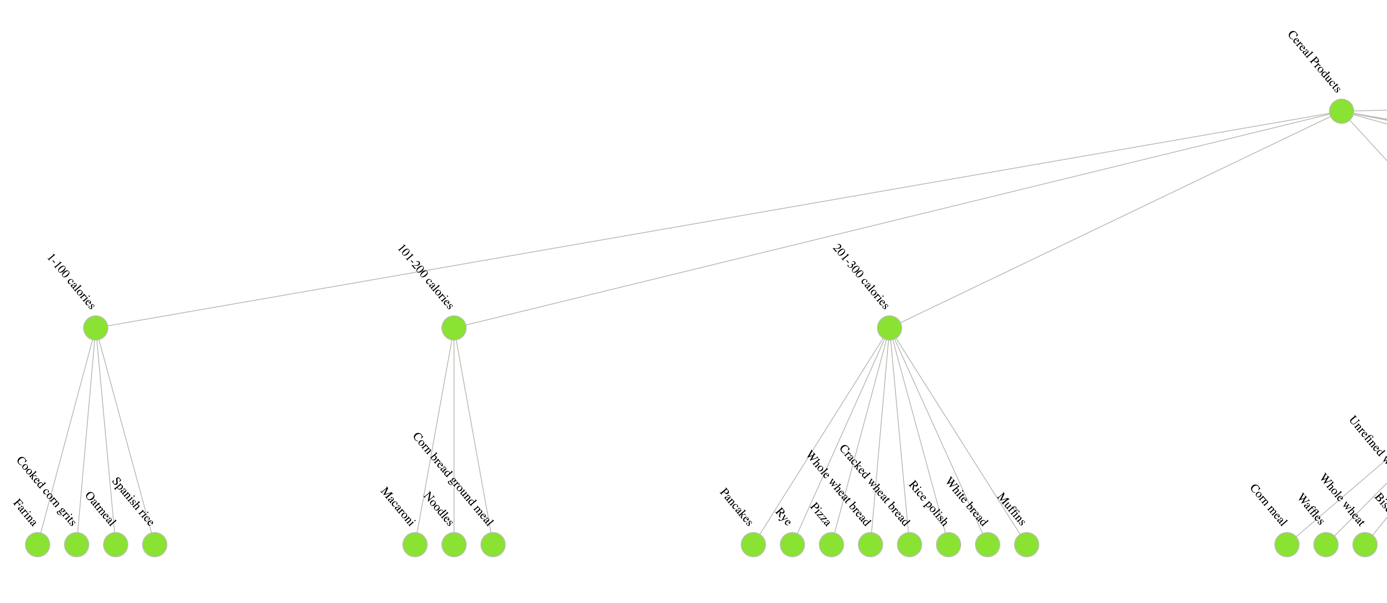


Abbildung 6: Lebensmittelkategorie Getreideprodukte in der Baumhierarchie (Quelle: eigene Darstellung)

Dabei sucht er zunächst eine Basis für einen Müsliriegel in der Kategorie Getreideprodukte (*Cereal Products*). Aus der Visualisierung entnimmt er, dass dafür Haferflocken (*Oatmeal*) in Frage kommen, da diese nur zwischen 1 und 100 Kalorien (*1-100 calories*) enthalten. Weiterhin möchte er eine geeignete Zutat aus der Kategorie Nüsse (*Nuts*) heraussuchen. In Abbildung 7 ist der Ausschnitt aus der Baumhierarchie mit dieser Kategorie dargestellt.

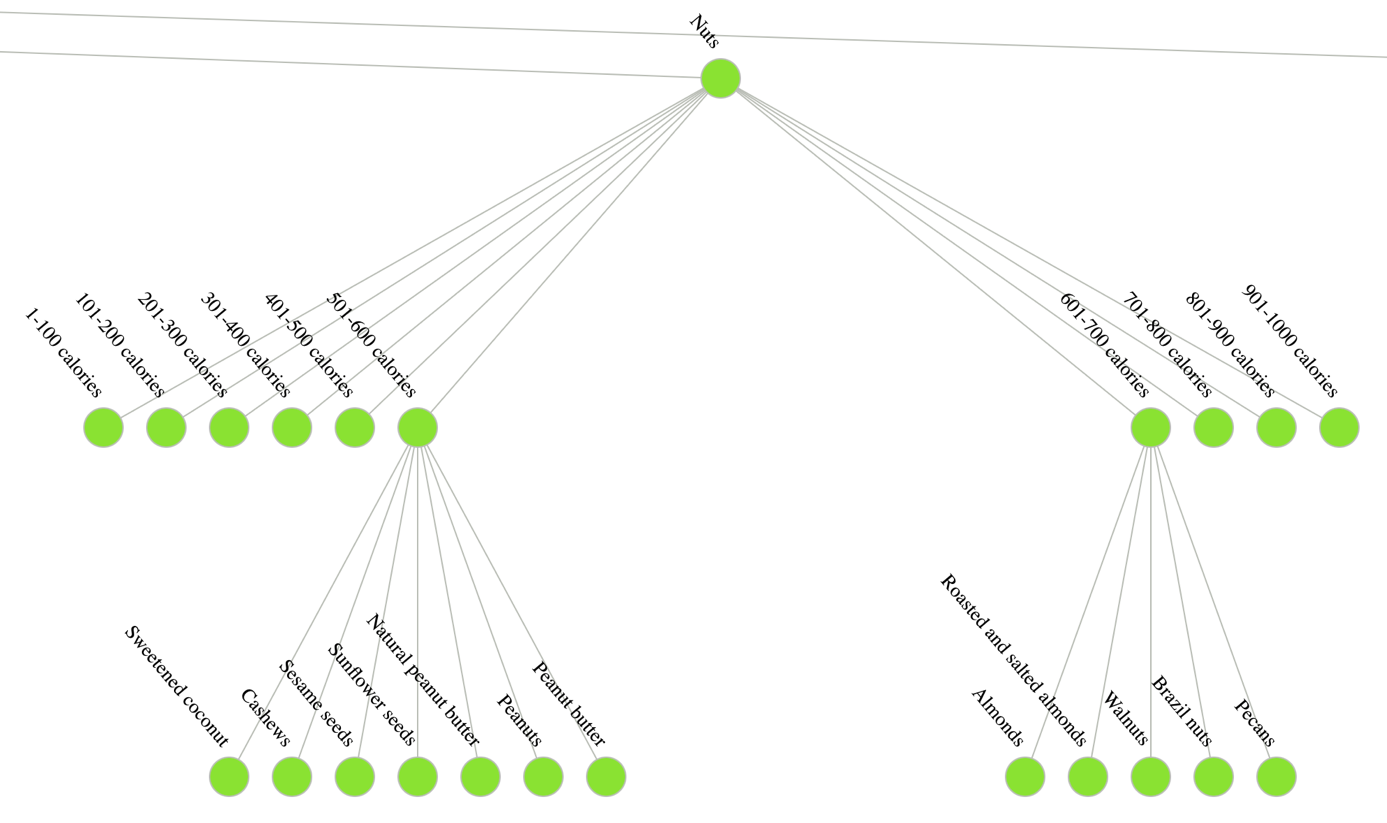


Abbildung 7: Lebensmittelkategorie Nüsse in der Baumhierarchie (Quelle: eigene Darstellung)

Zunächst fällt auf, dass Nüsse allgemein verhältnismäßig viele Kalorien pro 100 Gram enthalten. Aufgrund seiner Anforderung, dass der Müsliriegel möglichst kalorienarm ist, entscheidet er sich für Cashewnüsse (*Cashews*). Außerdem möchte er gesüßte Kokosnuss (*Sweetened Coconut*) verarbeiten. Beide Zutaten beinhalten mit 501-600 Kalorien (*501-600 calories*) die geringste Kalorienzahl. Anschließend schaut er sich die Kategorie Früchte (*Fruits*) an, um für seinen Müsliriegel die geeigneten Früchte auszuwählen. Diese Kategorie ist in Abbildung 8 dargestellt.

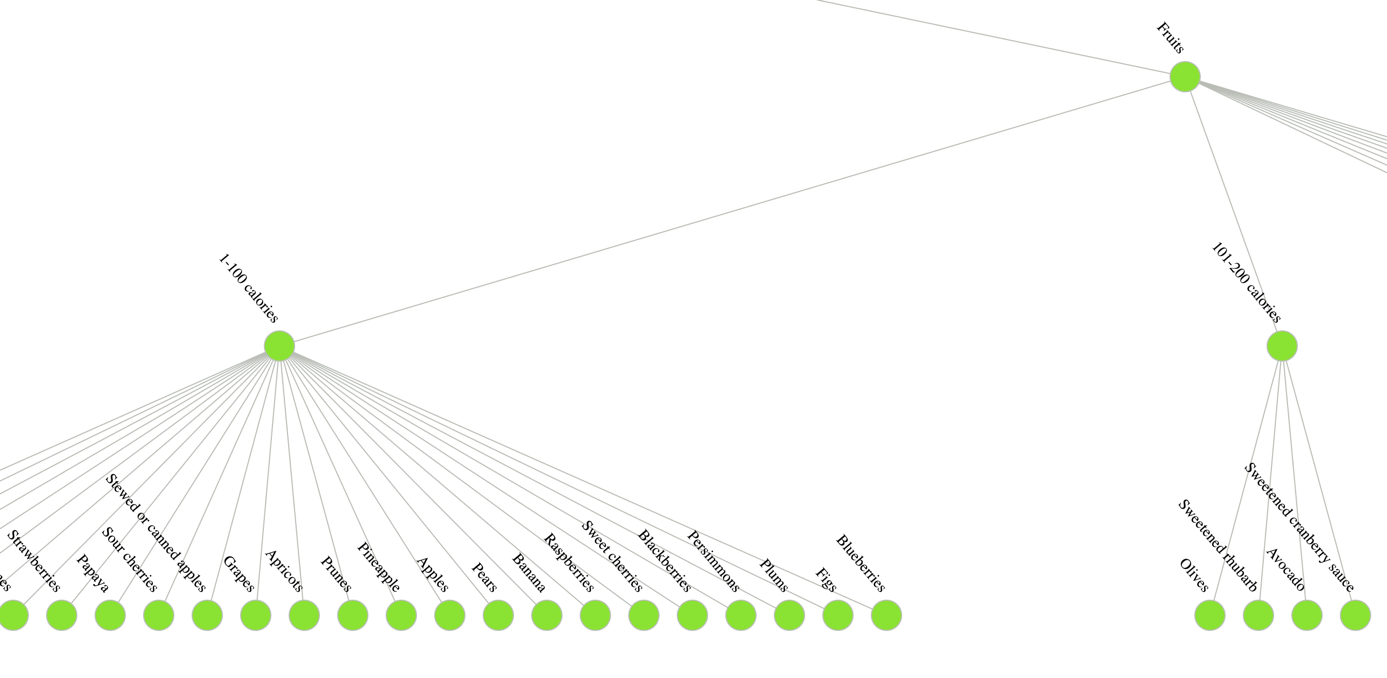


Abbildung 8: Lebensmittelkategorie Früchte in der Baumhierarchie (Quelle: eigene Darstellung)

Einige kalorienarme Früchte kommen für ihn als Geschmacksrichtung nicht in Frage. Folglich könnten für in Bananen und Himbeeren in Frage kommen, da diese mit 1-100 Kalorien (*1-100 calories*) noch im unteren Bereich liegen.

# Verwandte Arbeiten

# Zusammenfassung und Ausblick

Literaturverzeichnis

# Anhang

## Git Historie