VisualFood: A Visual Analytics Approach for Nutritional Trade-off Analysis and Product Comparison

Lorenzo Pizzi 2187091 (pizzi.2187091@studenti.uniroma1.it)

Introduction

The increasing complexity of processed foods has made it difficult for consumers and dietitians to assess the true nutritional value of products. A single product often presents conflicting attributes: high protein content might be offset by excessive saturated fats, or 'low-sugar' labels might hide high sodium levels. This represents a classic multi-objective optimization problem where the optimal choice depends on the user's specific health goals.

Traditional tools, such as static charts or database filters, lack the flexibility to explore these trade-offs dynamically. To address this, I developed **VisualFood**, an interactive Visual Analytics system. Stemming from the initial project proposal, the core design philosophy focuses on shifting the definition of healthy from a fixed universal metric to a user-defined parameter. The original concept envisioned a system driven by a weighted distance ranking, where users could manually assign importance weights to specific nutrients to generate a linear suitability score for each product. While this foundational goal of user-centric preference remains the system's backbone, the implementation strategy evolved from a simple ranked list to a multidimensional projection (Weighted PCA), allowing users to explore the structural landscape of nutritional similarity rather than just filtering a top-k list.

Related Work

Nutritional Visual Analytics has evolved from static food composition tables to interactive systems that support exploration, comparison, and decision-making in food choice contexts. To situate *VisualFood* within this landscape, I focus on related systems that explicitly target **food and nutrition data**, organized along three dimensions: **(i) consumer-centric nutrition/ingredient communication**, **(ii) preference- and similarity-based exploration of food spaces**, and **(iii) representations for nutritional trade-offs and profiling**.

Consumer-Centric and Ingredient Visualization

A first line of work aims to make nutrition labels and ingredient lists more comprehensible to non-expert users. **Nutrition Bytes** (He et al., 2017) presents a radial 'fingerprint' representation that juxtaposes nutrients and ingredients and uses small multiples to facilitate side-by-side comparison across a limited set of products. This design emphasizes item-level interpretability and label readability, enabling users to inspect the composition of individual foods in a compact view.

Commercial applications similarly target fast decisions in everyday contexts. FoodSwitch supports shopping-time substitution by combining barcode-based retrieval with simplified visual summaries and alternative suggestions (Dunford et al., 2014). Apps such as Yuka follow a related consumer-oriented approach by compressing information into a single score and color-coded evaluation, reducing decision effort at the cost of making multi-nutrient trade-offs less explicit (Yuka, n.d.).

Relation to VisualFood. While these systems prioritize clarity and speed for consumer decisions, *VisualFood* is designed for *comparative exploration* over a large dataset: instead of producing a single verdict, it exposes how a product stands relative to others under user-defined priorities.

Preference-Driven Exploration of Food Similarity

A second line of work explores food datasets as similarity spaces based on nutritional composition. Lunterova et al. propose an exploratory visualization that applies **t-SNE** to the USDA food database to reveal clusters and support awareness-oriented exploration through a macro (map) + micro (item detail) interface (Lunterova et al., 2019). In this framing, users can navigate the map and inspect items, but the similarity structure is largely defined by the fixed nutrient representation and the embedding procedure.

A more explicitly preference-driven approach is **NUT** (Dawson, 2013), which allows users to assign importance weights to nutrients and produces a ranked output, using dense visual encodings to handle large numbers of items. Here, user weights primarily drive scoring/ranking to support 'find best matches' workflows.

Relation to VisualFood. *VisualFood* combines these two perspectives: like NUT it supports user-defined priorities through weights, but like embedding-based systems it provides a global similarity overview. Instead of limiting the interaction outcome to a ranked list, *VisualFood* maps the weighted notion of similarity into a spatial overview to support exploratory tasks such as browsing clusters, identifying nearby alternatives, and inspecting why two foods are close or far under the current preference profile.

Nutritional Trade-offs and Profiling Representations

Nutrition decisions often involve trade-offs across multiple nutrients, and several systems support this through profiling or multi-attribute comparison. **NutriMap** (Labouze et al., 2007) proposes a profiling method that positions foods based on nutritional assets and weaknesses within categories, enabling comparisons among foods of the same type and supporting nutrition-quality reasoning at the category level. Compared to consumer apps that output a single score, NutriMap emphasizes interpretable dimensions tied to nutritional evaluation within a defined framework.

Relation to VisualFood. Where profiling approaches summarize nutritional quality into compact representations, *VisualFood* keeps nutrient-level trade-offs visible during exploration and comparison. Its coordinated views support both overview (similarity landscape) and inspection (attribute-level comparison), while leaving the notion of 'better' contextual and user-dependent.

Summary of Design Differences

Table 1 summarizes design-focus differences between *VisualFood* and representative food/nutrition systems.

Feature	**Food/Nutrition Systems**	**VisualFood**
Primary output	Score / fingerprint / ranking / map	Similarity landscape + coordinated details
User intent	Quick decision or item inspection	Goal-dependent comparative exploration
Trade-offs	Summarized or item-level inspection	Explicit multi-nutrient comparison + exploration
Ingredients	Label-centric representation	Metadata to contextualize nutritional profiles

Table 1: Design-focus differences between *VisualFood* and related food/nutrition approaches.

Conclusion. Related work in food and nutrition spans label-centric consumer visualization and scoring (Dunford et al., 2014; He et al., 2017; Yuka, n.d.), embedding-based exploration of food landscapes (Lunterova et al., 2019), preference-driven ranking tools (Dawson, 2013), and profiling approaches for nutrition-quality reasoning (Labouze et al., 2007). *VisualFood* explores a complementary design point by coupling preference-sensitive similarity exploration with coordinated views for nutrient-level inspection.

Data and Preprocessing

Data quality is paramount for reliable Visual Analytics. My initial development phase utilized a popular nutrition dataset available on Kaggle. However, preliminary visual exploration using the scatterplot revealed inconsistencies: unexpected clusters and item-level inspection suggested that values were not consistently standardized (e.g., mixed serving-based vs. per-100g values) and in some cases implausible.

To ensure scientific accuracy and consistency, I discarded the initial dataset and engineered a custom dataset derived from **USDA FoodData Central (FDC)**. Specifically, I used the FDC *survey foods* release (FNDDS/WWEIA categories), which provides standardized nutrient profiles and meaningful food categories.

ETL Transformation Pipeline

The raw USDA data is distributed as a relational collection of CSV files (foods, nutrients, food_nutrients, survey categories, etc.). I developed a Python-based ETL pipeline to convert this structure into a flat, analyzable table:

1. **Relational merging and labeling:** I extracted food names from `food.csv` and joined them with WWEIA category metadata (via `survey_fndds_food.csv` and `wweia_food_category.csv`) to obtain a human-readable `category` for each food.
2. **Nutrient selection (per 100g):** I loaded `food_nutrient.csv` and filtered rows to a curated set of macro- and micro-nutrients by `nutrient_id`. Since FDC nutrient amounts are provided per 100g, the resulting values are directly comparable across items.
3. **Pivoting to wide format:** I pivoted the nutrient table from long format (one row per nutrient measurement) to wide format (one row per food), producing a dense nutrient matrix for visualization.
4. **Ingredient-derived metadata:** From `input_food.csv`, I derived two additional descriptors: `num_ingredients` (ingredient count proxy) and `dominant_share` (share of the most prevalent ingredient when gram weights are available), plus a compact string of the top ingredients.

Preprocessing and Normalization

Once the merged dataset was generated, I applied additional preprocessing to improve reliability and support projection-based analysis:

- **Handling incomplete records:** Items missing key macronutrients (e.g., energy, protein, fat, carbohydrates) were removed to avoid undefined comparisons.
- **Plausibility filters:** I removed clear outliers (e.g., macronutrients exceeding 100g/100g; energy exceeding a conservative threshold) to reduce noise introduced by corrupted entries.
- **Z-score normalization:** Finally, all nutrient attributes were standardized using Z-scores (saved as `_norm` columns).

Final Dataset

The final dataset is exported as a single flat table (`cleaned_food.csv`) with 5,431 food items (rows) and 36 columns. Columns can be grouped into **categorical/text descriptors** (used for labeling, grouping, and filtering) and **numerical attributes** (used for projection, similarity, and quantitative inspection):

- **Categorical / text columns:** `food` (item name/description), `category` (WWEIA/FNDDS category label), and `ingredients` (compact text list of top ingredients). The column `id` is a unique identifier (numeric) used for linking records across views.
- **Numerical columns (original scale, per 100g):** ingredient-context metadata `num_ingredients` and `dominant_share`, plus 15 nutrient variables: `Caloric Value`, `Protein`, `Total Fat`, `Saturated Fats`, `Carbohydrates`, `Sugars`, `Dietary Fiber`,

`Cholesterol`, `Sodium`, `Water`, `Magnesium`, `Potassium`, `Iron`, `Calcium`, `Vitamin C`.

- **Numerical columns (standardized):** for each nutrient above, a Z-score normalized counterpart is provided (suffix `_norm`). These normalized columns are used for PCA-based projection and similarity computations, while the original per-100g values are preserved for interpretability in the details view.

System Interface

The interface was developed using D3.js for the frontend and Python (Flask) for backend data analytics. The system comprises a global control bar and three coordinated views, designed to support the *Overview → Zoom & Filter → Details-on-Demand* workflow.

Global Toolbar and Search Interface

Located at the top of the application, the Global Toolbar serves as the primary control center for data filtering and item retrieval. It includes two key components that operate transversely across all coordinated views:

Purity Filter (Dominance Slider). To manage the complexity of composite foods, a *Purity Score* (defined as the dominance of the primary ingredient, $D \in [0, 1]$) was introduced. An interactive slider allows users to filter the dataset based on this metric. Setting a high threshold (e.g., > 90%) isolates raw, unprocessed ingredients (e.g., Raw Apple), while a lower threshold reveals processed, multi-ingredient products (e.g., Apple Pie). This global filter immediately propagates to the Parallel Coordinates (hiding non-compliant polylines) and the Scatterplot (removing corresponding nodes), allowing the analysis to focus on specific levels of food processing.

Semantic Search. An auto-completion search bar supports queries for known items (*known-item search*). Users can type a product name to instantly locate it within the high-dimensional space. Upon selection, the system performs a *Focus+Context* operation: the product is highlighted across all views utilizing a unified color coding. The item appears as a bold polyline in the Parallel Coordinates (matching the distinct hue assigned in the selection queue) and as a pulsating node in the Scatterplot, allowing for an immediate inspection of its nutritional profile relative to the global distribution.

1. Preference Tuner (Parallel Coordinates)

The first view serves a dual role: it provides an overview of the multivariate nutrient distributions and acts as the main interaction surface to define user preferences. I implemented it as a Parallel Coordinates Plot (PCP), where each polyline represents one food item and each vertical axis encodes a nutrient dimension.

Semantic axis ordering. Axes are arranged to place conceptually related nutrients next to each other, so that trends and trade-offs can be read as local slopes (reducing cognitive load and line crossings). The adopted order is: *Energy (Kcal)* and *Water* first (to expose energy-density patterns), followed by grouped macronutrients and their subcomponents (*Fat → Sat. Fat → Cholesterol*, then *Carbs → Sugar → Fiber*, and *Protein*). The rightmost part collects *micronutrients and electrolytes* (*Sodium*, *Potassium*, *Calcium*, *Magnesium*, *Iron*, *Vitamin C*). This ordering supports common analytical questions such as 'high-protein but low-sodium' or 'high-carb but high-fiber (vs. sugar)' by keeping the relevant axes adjacent.

Brushing for filtering. Users can brush any axis to specify a value range. Each brush defines a constraint on that nutrient, and multiple brushes combine through conjunction (AND), producing a filtered subset. Filtering is propagated to the other views: non-matching items are visually suppressed (context reduction), while the similarity scatterplot and the detail panel update to reflect only the currently eligible candidates. This interaction turns the PCP into an explicit query builder for nutritional constraints.

Weight handles for preference steering. Above each axis, an interactive handle (slider) lets users set an *importance weight* in the range $[0, 1]$. These weights do not merely restyle the PCP: they directly steer the similarity model by parameterizing the weighted PCA used for the 2D projection. To provide immediate feedback, the axis opacity is mapped to the current weight (higher weight → higher salience), making the active preference profile readable at a glance. When the user releases a handle, the system requests an updated projection from the backend and redraws the similarity landscape, enabling iterative 'what-if' exploration driven by user-defined similarity.

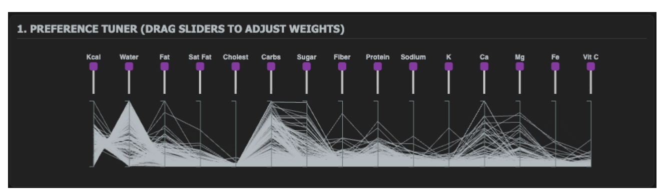

Figure 1: Parallel Coordinates for preference tuning

2. Similarity Landscape (Steerable Scatterplot)

While the Parallel Coordinates view focuses on individual dimensions, the Similarity Landscape provides a holistic summary of the dataset structure. It projects the high-dimensional nutritional space into a 2D plane, where Euclidean distance serves as a visual proxy for nutritional dissimilarity.

Steerable Weighted Projection. The core innovation of this view is its coupling with the feature weights defined in the Preference Tuner. Instead of a standard Principal Component

Analysis (PCA), the system performs a *Weighted PCA* on the backend. When a user manipulates a weight handle (e.g., increasing the importance of Sugars and Fiber), the updated weight vector adjusts the covariance matrix computation. This forces the projection to reorganize: the axes of maximum variance rotate to align with the user's prioritized attributes. This allows users to answer morphological questions such as 'How do products cluster if I strictly prioritize macronutrient balance over vitamins?'.

Categorical Context and Clusters. To support the interpretation of the emerging clusters, data points are color-coded according to six high-level macro-categories: *Plant-based*, *Animal Protein*, *Dairy and Cheese*, *Sweets and Snacks*, *Fats and Oils*, and *Prepared/Other*. As the original dataset comprises over 170 fine-grained categories (e.g., apples, beef products, legumes), assigning a distinct color to each would result in severe visual clutter. To address this issue, a semantic aggregation strategy maps individual food items to macro-groups via keyword-based rules. This encoding enables users to assess whether nutritional similarity aligns with biological taxonomy or reveals cross-category overlaps, such as high-fat plant products clustering with dairy items. Additionally, the interactive legend allows users to selectively hide or display categories, mitigating occlusion and enabling focused comparative analysis of specific data subsets.

Selection and Coordination. The scatterplot acts as the primary selection interface for the detailed comparison. Clicking a data point adds it to the *Comparison Set* (up to three items), highlighting it with a high-contrast neon stroke (Red, Green, Yellow) that persists across all linked views. This interaction supports the *Overview* → *Detail* mantra: users identify promising clusters or outliers in the projection and select them to drill down into their specific values in the Detail Panel. Furthermore, hovering over any point triggers a temporary linking effect, projecting a ghost line in the Parallel Coordinates to reveal that item's multidimensional profile instantly.

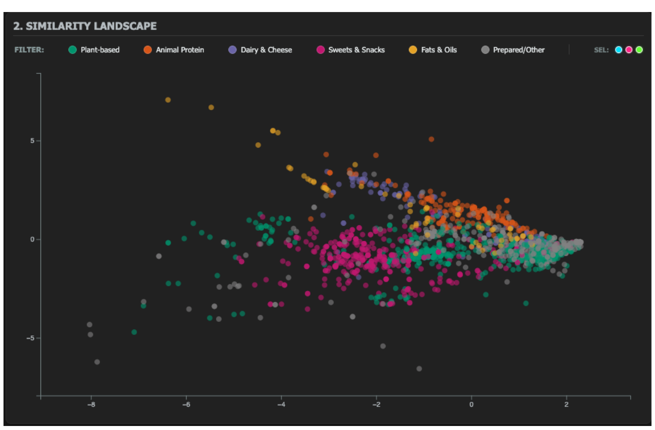

Figure 2: Scatterplot showing the PCA embedding computed with uniform weights across all nutrient attributes.

3. Detail Panel (Nutritional Comparator)

While the scatterplot reveals global patterns, the Detail Panel implements the 'Details-on-Demand' stage of the visual analytics mantra. It activates whenever a user selects items from the Similarity Landscape, supporting a direct side-by-side comparison of up to three products. This constraint was introduced to adhere to **cognitive load theory**, as comparative analysis of high-dimensional data (15+ nutritional attributes) becomes increasingly cognitively demanding beyond a small set of items. Furthermore, limiting the selection ensures optimal visual discriminability: it allows the use of a high-contrast qualitative color palette that remains distinct from the categorical encoding, facilitating rapid cross-view identification. Finally, this limit preserves the spatial integrity of the Detail View, ensuring that nutritional bar charts remain legible side-by-side without requiring horizontal scrolling, which would hinder immediate visual comparison.

Juxtaposed Micro-Bars for Comparison. To handle the heterogeneity of nutritional scales (ranging from grams for macronutrients to milligrams for minerals), the view avoids a shared absolute axis. Instead, it employs a list of *juxtaposed horizontal bar charts*. For each nutrient row, the bar lengths are normalized relative to the *local maximum* of the currently selected set ($L = v_i / \max(v_{set})$). This design choice allows users to instantly perceive ratios without reading the numerical labels, which are nonetheless provided for precision.

Semantic Coloring and Cognitive Conflict Resolution. A key design challenge was encoding the 'healthiness' of values without introducing cognitive conflicts. Standard visualizations often map 'High Value' to 'Positive Color', but for nutrients like Sodium or Sugar, a high value is undesirable. I implemented a *semantic coloring logic* that classifies attributes into:

- *Beneficial Nutrients* (e.g., Protein, Fiber): High values are rendered with a green bar and an upward arrow (▲).
- *Limit Nutrients* (e.g., Sugar, Saturated Fat): High values are rendered in red to signal caution. However, the 'winning' product (the one with the lowest amount) is highlighted with a green downward arrow (▼).
- *Context-Dependent Nutrients* (e.g., Calories, Carbohydrates, Total Fat): These are rendered in a **neutral blue**. Unlike the other categories, the system avoids assigning a judgment (color or directional arrow). This design choice acknowledges that the desirability of these metrics is subjective and goal-dependent, leaving the interpretation to the user.

This resolves the cognitive conflict: the color green consistently signals the *best choice*, regardless of whether 'best' means maximizing (protein) or minimizing (sugar) the quantity.

Adaptive Vertical Layout. To maximize screen utilization, the view abandons static row heights in favor of a *viewport-aware* rendering logic. A `ResizeObserver` monitors the container's dimensions in real-time. The system dynamically computes the optimal row height ($H_{row} = H_{total}/N_{attributes}$) to expand the visualization and fully occupy the available vertical space, effectively eliminating redundant whitespace on larger monitors. Vertical scrollbars are introduced *conditionally*: only if the computed row height falls below a minimum legibility threshold (e.g., 16px), the system enforces a fixed minimum height and enables scrolling.

Figure 3: Detail and Trade-off panel

Case Studies and Insights

To demonstrate the analytical capabilities of the system, I present three case studies conducted on the dataset. These scenarios illustrate how the coordination between the *Preference Tuner*, the *Similarity Landscape*, and the *Detail Panel* allows users to move from general hypotheses to specific nutritional evidence.

Case 1: Identification of Hyper-Dense Outliers (Structural Analysis)

Goal: Upon launching the system, the user observes the global topology of the nutritional space to identify structural anomalies in the standard distribution (Weighted PCA with all weights set to 0.5).
Interaction:

1. *Observation (Scatterplot):* The user notices that while the majority of the 5,000+ products form a central dense cloud, a small group of three distinct outliers is isolated in the extreme bottom-left corner of the projection, far removed from any nutritional cluster.
2. *Selection (Linking):* The user clicks on these points to identify them. The system reveals them as: 'Cocoa Powder', 'Wheat Bran', and 'Nutritional Powder Mix (Slim Fast)'.
3. *Comparison (Detail Panel):* The user analyzes their composition. The comparator view highlights a striking similarity: all three possess near-zero *Water* content (empty bar) but extremely high values in distinct macronutrients (Fiber for Bran, Protein for Slim Fast, Fat/Fiber for Cocoa).

Insight: The projection successfully separated foods based on their *Physical State* (Hydrated vs. Anhydrous), effectively isolating the 'Energy Vertices' of the dataset. This geometric segregation is not accidental but validates the status of Cocoa and Wheat Bran as natural **superfoods**: they exhibit the highest nutrient-to-weight ratio in the entire dataset. The serendipitous clustering of an artificial diet product (*Slim Fast*) with these raw ingredients exposes the underlying engineering of meal replacements: they are designed to biologically mimic the extreme nutritional density of natural superfoods, creating a synthetic profile that occupies the same geometric niche as nature's most efficient ingredients.

Case 2: The 'Hidden Sugar' Trap in Dairy (Dietary Management)

Goal: A Dietitian needs to select dairy products for a **pre-diabetic patient**, aiming for calcium without insulin spikes (Low Sugar), while navigating the marketing noise of 'low-fat' labels.
Interaction:

1. *Visual Focus (Scatterplot):* The user utilizes the **Interactive Legend** to filter the view. By toggling the category labels, they hide irrelevant groups and strictly isolate the *Dairy & Cheese* cluster (Purple points) for focused analysis.
2. *Steering (PCP):* To simulate the patient's metabolic needs, the user steers the weight of *Sugars* to maximum (1.0).
3. *Visual Reaction:* The Purple cluster, initially compact, fractures visibly. A subset of points migrates rapidly towards the high-sugar region of the projection, separating from the main group.
4. *Comparison (Detail Panel):* The user selects two outliers from this divergent group for a side-by-side comparison: a standard 'Flavored Coffee Creamer' (Red selection) and a 'Fat Free' version (Green selection).

Insight (Trade-off Discovery): The system exposes the *Substitution Effect*. The Detail Panel reveals a paradox: while the 'Fat Free' product effectively minimizes lipids ($1.9g$ vs $21.5g$), it compensates with a drastic increase in sugar ($73.4g$ vs $58g$). The Weighted PCA successfully allowed the Dietitian to debunk the 'healthy halo' of the low-fat label, identifying a product that is metabolically worse for a pre-diabetic patient than the standard version.

Case 3: Sodium in Plant-Based Substitutes (Hypertension Management)

Goal: A consumer with **hypertension** is looking for meat substitutes in the *Plant-based* category, assuming they are inherently heart-healthy.
Interaction:

1. *Filtering & Context (Global State):* The user relaxes the *Purity* slider to include composite foods and uses the **In-**

teractive Legend to visualize only the *Animal Protein* and *Plant-based* categories, hiding the rest.

2. *Brushing (PCP):* To find viable protein sources, the user applies a **brush filter** on the *Protein* axis of the Parallel Coordinates, isolating only high-protein items.
3. *Cluster Analysis (Scatterplot):* The user observes two main clusters: a dense group of animal products and a more dispersed plant-based group. Crucially, visual inspection reveals a subset of plant-based points positioned in close proximity to the animal cluster.
4. *Selection (Drill-down):* Investigating this overlap, the user selects two proximal neighbors for comparison: 'Breakfast link, meatless' (Vegetarian) and 'Cheesburger, NFS' (Animal).

Insight: The Detail Panel reveals the hidden trade-off of this spatial proximity. To achieve a texture and protein profile similar to the *Cheesburger*, the *Meatless Link* is ultra-processed, containing **888mg** of Sodium. This is drastically higher than the real meat counterpart, debunking the assumption that 'Plant-based' equals 'Low Sodium' and highlighting the system's ability to expose nutritional mimicry through geometric clustering.

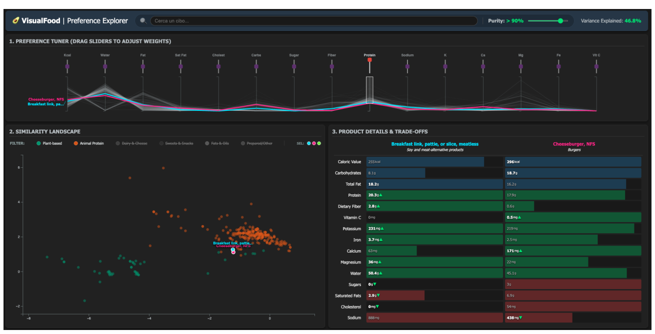

Figure 4: Full overview of the system that figure out this third insight

Conclusion

The final system preserves the original design intent (three coordinated views and user-defined nutritional similarity), while the implementation evolved in two aspects: (i) the dataset was replaced with USDA FoodData Central to ensure per-100g standardization and reliability, and (ii) the similarity computation shifted from a weighted-distance ranking to a weighted PCA projection, enabling an interactive similarity landscape rather than only a ranked list.

References

Dawson, J. (2013). *Nut: Visualizing the nutritional content of what we eat* (tech. rep.) (Course project report (CPSC 533C)). University of British Columbia. https://www.cs.ubc.ca/~jqdawson/publications/533c_NUT.pdf

Dunford, E., Trevena, H., Goodsell, C., Ng, K. W., Webster, J., Millis, A., Goldstein, S., Hugueniot, S., & Neal, B. (2014). Foodswitch: A mobile phone app to enable consumers to make healthier food choices. *JMIR mHealth and uHealth*, *2*(3), e37. https://doi.org/10.2196/mhealth.3230

He, S., Kerrigan, D., & Metoyer, R. (2017). Nutrition bytes: Visualizing food content [Poster extended abstract]. *IEEE VIS Posters (Poster Extended Abstract)*. https://dankerrigan.me/papers/NutritionBytesFinal.pdf

Labouze, E., Goffi, C., Moulay, L., & Azaïs-Braesco, V. (2007). A multipurpose tool to evaluate the nutritional quality of individual foods: Nutrimap. *Public Health Nutrition*, *10*, 690–700. https://doi.org/10.1017/S1368980007382505

Lunterova, A., Spetko, O., & Palamas, G. (2019). Explorative visualization of food data to raise awareness of nutritional value. In C. Stephanidis (Ed.), *Hci international 2019 – late breaking papers* (pp. 180–191, Vol. 11786). Springer. https://doi.org/10.1007/978-3-030-30033-3_14

Yuka. (n.d.). How are food products rated? [Explains the score composition (nutrition/additives/organic).]. Retrieved January 3, 2026, from https://help.yuka.io/l/en/article/ijzgfvi1jq-how-are-food-products-scored