Communicating large datasets through the use of dimensionality reduction, exploratory data visualization, and storytelling techniques

Ondrej Spetko and Anna Lunterova

Aalborg University Copenhagen

Abstract. This paper covers the topic and investigative experimental process of the combination of dimensionality reduction, data visualization, interactivity and storytelling techniques, with an aim of achieving effective (functional, informative, insightful) communication of a large dataset. The conducted analysis suggests this chosen combination can potentially strengthen and create effective communication. The interactive data visualization alone is considered as a new field, with only a little existing previous research. In this paper the area of interactive visualization is further expanded by implementing storytelling elements and data analysis theory. Food database consisting of 41 dimensions and over 8400 points was chosen. The platform or medium through which the experiment was chosen to be conducted and created was a website, that can be found here. As evaluation method, the communicativeness of the interface and its usability was measured through the use of self-reported pre and post-questionnaires. The analysis of 23 participant's answers shows that the research is rather inconclusive and lacking further iterations and improvements in both, the design, and in the evaluation method. However, this can serve as an initial exploration of this quite potential novel combination, and when taking into the account the complexity of the proposed design this solution can lead as a starting point to create a more specific, or larger scope research in the future.

Table of Contents

List of Figures

Introduction

Over the past decades great amounts of electronic data are being stored. These data hold values that are not always obvious and comprehensible at the first sight. Complexity of the data and quantity of dimensions per record are though obstacle in the way of understanding them and presenting them. In this project we are experimenting with the latest methodology and approaches to try to find an effective way of visually presenting a large dataset to common people.

Displaying and communicating a large high dimensional datasets is a challenging problem. Common visualization solutions deal only with the default types of simple visualizations of low-dimensionality data [**?**]. Unsupervised machine learning is an Artificial Intelligence technique used for finding similarities between data features of high-dimensionality unlabeled datasets. This allows to visualize data points of complex datasets into 2D maps and and cluster them by their similarities, aiding the communication process with a reader. Two of the most common dimensionality reduction algorithms, PCA and t-SNE, will be considered.

Additionally by adding storytelling and interactivity , elements to Data Visualization(DV) we complement the experience by enhancing and promoting user-interaction and further exploration of the presented information, turning the plain visualizations into a more effective and memorable narrative experience with context and direction for general audience [**?**,**?**,**?**,**?**,**?**]. The agenda of the paper is to firstly research each of the mentioned fields in the analysis, conclude requirements for designing of possible solution, implement the solution and finally, evaluate the communicativness of the solution.

Motivation

The aim is therefore as mentioned, to explore and understand each field of the combination of dimensionality reduction, data visualization, interactivity and storytelling techniques, to be able to design and implement our own interface for communicating a specific dataset. As inspiration, similiar concepts of visualizations were found. Only few similiar experiments were found. These served as a starting point of understanding possibilities of a potential solution.

Drum machine

Most related experiment done in this area that was found is “The infinite drum machine” experiment done by Google Creative Labs, uses unsupervised machine learning algortithm t-SNE to organize large dataset of thousands of points, each being a sample of an everyday sound [**?**]. Similiar sounds are placed together. The interface can be seen on the figure below (see figure 1). Map can be used to explore neighborhoods of similar sounds and create both random and also customized beats using the drum sequencer.

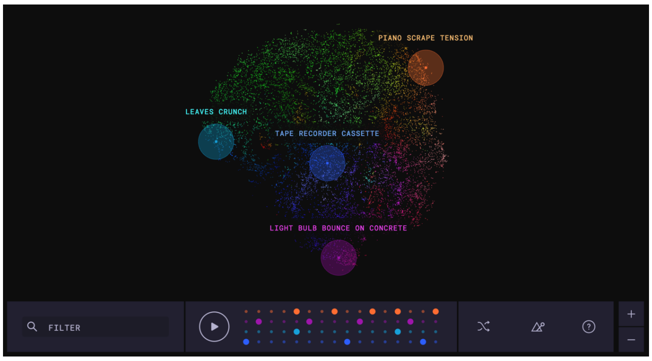


Fig. 1. The infinite drum machine [**?**]

Initial Problem statement

We formulate the following initial problem statement:

How can we communicate large datasets through the use of AI and exploratory data visualizations?

Analysis

Based on our motivation from the introduction that lead us to the Initial Problem Statement, the following chapter contains exploration of the fields of data analysis, data visualization, narratives, storytelling and their tools. Specifically, data analysis including data mining and dimensionality reduction, next the techniques and evaluation elements of data visualizations, and lastly the field of narratives and interactive storytelling in connection to data visualizations will be presented, altogether leading us to Final Problem Statement Formulation.

Data Analysis

Data analysis is process of extracting useful information out of data. To be able to deliver users comprehensible information we first have to analyse the dataset and get the useful information out of it. In this section we will go through general process of data analysis and data mining, introduce most famous tools and mention some examples relevant to our case.

Process The process of data analysis consist of many steps out of which not all are crucial for sufficient information extraction. Below is summary of the most important steps:

- Data collection

- Data processing/formatting
- Data cleaning
- Exploratory data analysis
- Communication/Visualization

Data collection is first fundamental step of gathering enough relevant data that we want to proceed with and get the information out of.
Data processing/formatting is next act in which the data is being shaped (columns, rows) to a form that suits our purpose.
Data cleaning is inevitable step in order to maintain data validity by removing duplicates and other potentially harmful and misleading features from the dataset.
Exploratory data analysis is process of applying mathematical models and functions to the dataset in order to uncover hidden relationships between data points.
Communication/Visualization is the final of the important steps and its purpose is to communicate the extracted information to the target group of the operation.

Data mining Data mining is considered to be particular technique of the data analysis that focuses specifically on modelling and knowledge discovery from the data for more predictive than descriptive purpose. Even though the data analysis methodology seems to cover the needs for this project the data mining is more involved with machine learning implementation by definition. This section will serve as brief introduction to the area of deeper data analysis known as data mining. Here we also introduce the machine learning concept and most famous algorithms useful for our case.

Machine learning is very old topic in theory but very young concept in practice. Is it only recently that the technology potential and electronic data amounts available found met and work together for purpose of uncovering hidden meanings and serve us as predictive and reactive tools. It is estimated machine learning algorithms will replace ˜25% of job positions across the world in the following decade. The most fundamental division of the machine learning algorithms has three groups[**?**]:

- Supervised
- Unsupervised
- Reinforced

Supervised learning algorithms are responsible for making predictions out of set of data samples and searches for pattern within value labels belonging to the data points.
Unsupervised learning algorithms do not work with labels as supervised learning does. Their purpose lays within discovery of relationships and simplifying of the understanding of complex data sets by grouping data points into clusters by their similarities.

Reinforced learning algorithms is the last group and aims for the most effective result being given goal and set of options. The process of learning is following try and compare cycle. These algorithms take a lot of time as their effectivity and level of evolution is purely dependant on the time and data provided but then are by far the superior executor of their goal.

The process of data mining is commonly divided into three steps[**?**]:

- Preprocessing
- Data mining
- Results validation

Preprocessing involves ensuring the dataset is big enough for algorithm to be able to form relationships between data points but compact enough to execute in desirable time limit. The step of preprocessing also includes acts of cleaning the data from noises and duplicates. Part of preprocessing is method of dimensionality reduction that focuses on reducing the number of features/dimensions of one data point to easier perceivable number usually two or three features/dimensions.
Data mining is the core of the process as it carries out the algorithm responsible for revealing the information from the dataset based on the purpose of the analysis. These purposes are commonly belonging to following groups [**?**]:

- Anomaly detection: search for outliers in the dataset
- Clustering: grouping data points by similarities of any kind
- Classification: generalizing known structure to apply to new data (e.g. mail spam filter)
- Regression: attempt of finding a function describing the data with lowest error

Results validation is step in which the result is analysed and validated to ensure its reproducibility and significance.

Dimensionality reduction To make sure that visually presented data is easily comprehensible we have to work with commonly perceivable range of dimensions or find a way to present the dimensions in visual features that indicate their value. In this section we will slightly elaborate on the state of the art of dimensional reduction methodologies.

Dimensionality reduction is a preprocessing stage, part of data analysis/data mining. By reducing the amount of features, the derived values become more informative, non-redundant, and leading to facilitating human interpretation. Two most common and most effective dimensionality reduction algorithms are principal component analysis (PCA) and T-distributed Stochastic Neighbor Embedding (t-SNE).

PCA The most common algorithm for reducing dimension of a dataset is principal component analysis (PCA) developed already back in 1933. PCA is linear algorithm that uses an orthogonal transformation to transform a set of observations of likely similar variables into a so called principal components, what is a set of values of linearly uncorrelated variables. Problem with linear dimension reduction algorithms is that the dissimilar data points are being placed far apart in lower dimensional representations. Linear algorithms won't be able to describe complex polynomial relations between features. Unlike other famous non-linear dimensionality reduction algorithm t-SNE that uses probability distribution with element of randomness instead to find the relations in the data. This allows t-SNE model distances between points in the low-dimensional map and a alteration of Kullback-Leibler divergence [?]. Figure below demonstrates PCA dimensionality reduction algorithm (see figure 2).

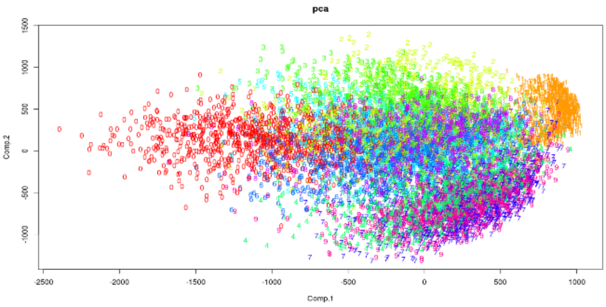


Fig. 2. PCA algorithm [?]

t-SNE t-SNE is a relatively new technique (2008) developed by Laurens van der Maaten that reduces amount of dimensions in high-dimensional data by assigning relative distance value based on the relative correlations to each datapoint resulting in a two or three-dimensional map [?]. The strength of t-SNE is in preserving the local distances of the high-dimensional data in mapping to low-dimensional data. Even though the t-SNE algorithm results in obvious clustering of the similar data points in the final map, the algorithm is not a clustering algorithm and is used only as exploratory or visualization tool. The reason for this is that original dimensions are mapped to lower dimensions without preserving any link to their original values. Because of that one can not make definite assumptions based only on t-SNE output. However, output of the t-SNE can be used in process of classification or clustering as input into further classification or clustering algorithms. t-SNE algorithm is quite heavy on the system resources because it compares the relations pairwise with goal of minimizing the sum of the difference of the probabilities in higher and lower dimensions. The visualizations produced by t-SNE are found to be significantly more accurate compared to famous dimensionality reduction algorithms like Principal Component Analysis (PCA), Sammon mapping, Isomap, Locally Linear Embedding and other. Figure below demonstrates the obvious clustering of the data points with clear separation of clusters (see figure 3).

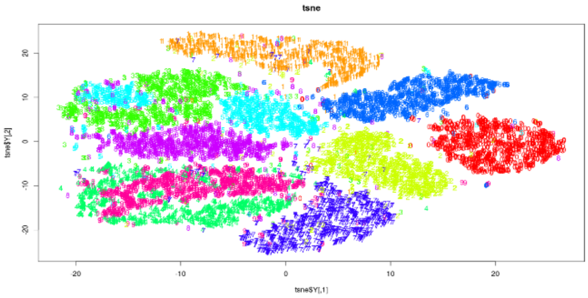
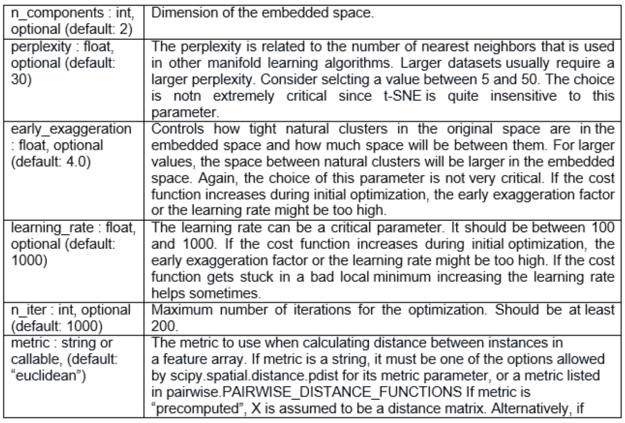


Fig. 3. t-SNE algorithm [**?**]

When implementing t-SNE it is important to be aware of the tuning parameters that are influencing the algorithm output. In the figure below is table of the parameters that are available when implementing t-SNE with Python programming language (see figure 4).



n_components : int, optional (default: 2)	Dimension of the embedded space.
perplexity : float, optional (default: 30)	The perplexity is related to the number of nearest neighbors that is used in other manifold learning algorithms. Larger datasets usually require a larger perplexity. Consider selcting a value between 5 and 50. The choice is notn extremely critical since t-SNE is quite insensitive to this parameter.
early_exaggeration : float, optional (default: 4.0)	Controls how tight natural clusters in the original space are in the embedded space and how much space will be between them. For larger values, the space between natural clusters will be larger in the embedded space. Again, the choice of this parameter is not very critical. If the cost function increases during initial optimization, the early exaggeration factor or the learning rate might be too high.
learning_rate : float, optional (default: 1000)	The learning rate can be a critical parameter. It should be between 100 and 1000. If the cost function increases during initial optimization, the early exaggeration factor or the learning rate might be too high. If the cost function gets stuck in a bad local minimum increasing the learning rate helps sometimes.
n_iter : int, optional (default: 1000)	Maximum number of iterations for the optimization. Should be at least 200.
metric : string or callable, (default: “euclidean”)	The metric to use when calculating distance between instances in a feature array. If metric is a string, it must be one of the options allowed by scipy.spatial.distance.pdist for its metric parameter, or a metric listed in pairwise.PAIRWISE_DISTANCE_FUNCTIONS If metric is “precomputed”, X is assumed to be a distance matrix. Alternatively, if

Fig. 4. t-SNE parameters using Python [**?**]

Examples of t-SNE applications In practical applications t-SNE is used as dimensionality reduction algorithm in challenging tasks like Facial Expression Recognition that suffers from high dimensional data or Identifying Tumor subpopulations (Medical Imaging) where t-SNE can uncover tumor subpopulations that are statistically linked to patient survival in gastric cancer and metastasis status in primary tumors of breast cancer.

Tools There are many tools people can choose from when entering data science and analysis. However, the most famous go-to choices are following three [?]:

- SAS-Short for Statistical Analysis System
- R-R language
- Python

SAS is a licensed software. It was developed by NCSU back on 1970'. SAS is still often used. In fact most of the fortune 500 companies are using SAS.
R language is open source language designed for statistics and data analysis with plenty of libraries and big community. R runs predominantly within command line interface.
Python is by far the most famous solution thanks to its big community and plenty of learning sources. It si high level programming language that is simple to pick up and has plenty of data analysis libraries. Python is professional tool for people with career in the machine learning and artificial intelligence. Python's most useful data analysis libraries are Numpy (math, matrices and conversions), Pandas (datasets operations) and Matplotlib(plotting of the results).

Data Visualization

Data visualization is fundamental approach for presenting data to the audience. In this section we briefly take a closer look on the field breaking the concept into 3 steps: Types of visual representation, Tools of visual representation and Software for visual representation.

Definition

Being equivalent to visual communication, data visualizations (DV) stand for the representation and presentation of data to facilitate understanding [?,?]. More specifically characterized, it is an "information that has been abstracted in some schematic form, including attributes or variables for the units of informations" [?]. It is an effective way to transform data-driven information in its raw form (e.g. numbers), into more easily understandable and aesthetically pleasing form of a picture [?,?].

Types of visual representation

In order to create proper visualization for a given set of data it is important to reflect upon the properties of the data and align them with the purpose of the visualization. First factor on the way of choosing what type of visualization would fit best is number of variables that are to be shown on the chart. Another factor is number of items representing single data point. In the end it is needed to ask if the data is being displayed for a period of time, or is it being grouped? The figure below shows selection of the final visual representation based on these three factors (see figure 5).

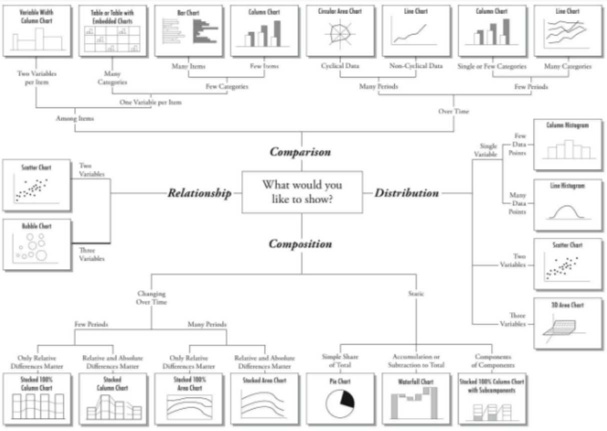


Fig. 5. Choice of visual representation scheme [?]

Visualization tools

DV are created by analyzing a data source and being visualized with specially designed software for representing or displaying the data. The simplest examples of such includes graphs, maps, 3D models, plots and others [?]. The process of creation a data visualization includes three main steps: import data, choose the type of visualization, and construct the visual aspects [?]. Several online tools and systems are being developed to make the creation of visualizations easier for non experts and people with low programming skills. (E.g.Many Eyes, Tableau, Power Map, Flourish, iNZight, RAWgraphs, QGIS, Gephi, NodeBox, etc.) However, being based on predefined templates, the shortcomings of such tools is little customization, visualizing only low dimensional data, and formed visualizations are mostly static with non or with very little interaction, making them almost entirely author-driven [?]. To overcome shortcomings of such softwares, building visualization from scratch is a solution. Custom built data visualizations are very popular and technologically well backed up approach. For custom built solution of data visualization people can choose from quite few programming languages and interfaces offering rich communities helping each other with their struggles along. One of the most progressive solutions for building custom built data visualizations is JavaScript based open source library D3.js published on GitHub. It allows creating any kind of customized visualization with JavaScript based on manipulation of HTML DOM elements [**?**].

Data visualization principles

Qualities of great visualizations are summarized into five, further below mentioned principles. This is based on Alberto Cairo's, recent book "*The Truthful Art*", and another globally recognized specialist in infographics and data visualizations Edward Tufte "The Visual Display of Quantitative Information" [**?**,**?**]. While the name of those principles slightly differs in between these two authors, the characteristics are the same. The order of these design principles is by their importance, from highest to lowest.

- Truthfulness, meaning data are based on thorough and honest research.
- Functionality, accurate depiction of data, that help the viewer to think about the shown information (rather than the design). Being truthful and functional is summarized as graphical integrity in Edward Tufte's book.
- Should be aesthetically pleasing, in the sense of being intriguing and attractive for its audience.
- Insightfulness, revealing otherwise harder to be revealed evidence.
- Enlightening in a way that if the audience grasps and understand the shown evidence, it changes their understanding or behaviour for better.

Those principles serves as a basis for evaluation of produced visualizations and are (or should be) commonly used in between data journalists, during visualizations contests, etc. [**?**].

Narratives and interactive storytelling

To enhance the comprehension of the data visualization and ensure user's engagement throughout the possible interaction we researched in the fields of narratives, storytelling and interactivity to gather some useful information about the tools we could use for our purpose of presenting dataset of nutrition values.

Narratives

Narratives are great feature to start and maintain user's engagement by pulling him into a plot or a story.

Storytelling and DV Storified visualizations, also called “visual data stories” , or “scientific storytelling” stand for DV that consists of story components “(structures, elements, and concepts)” and is contained within elements that mediates the telling of it “(people, tools, and channels)” [**?**,**?**] . The process of creating visual data stories in research “More Than Telling a Story: Transforming Data into Visually Shared Stories” by Bongshin Lee and Nathalie Henry Riche, three main phases are defined (see figure 6).

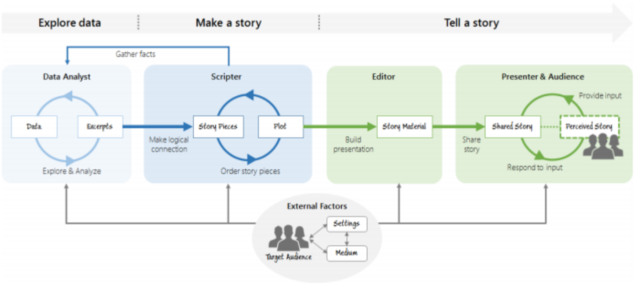


Fig. 6. Storytelling process in DV [**?**]

Firstly the chosen data is explored and analyzed. Secondly, after seeing patterns, trends, differences etc., the chosen facts, (*story nodes*) are connected (*story transitions*) in meaningful order (*plot*), to form a story. Lastly, the story is told using a chosen medium and narrative technique.

Those “story nodes” are visualized in forms of graphs, charts, maps etc., with the use of clarifying annotations, or narrations [**?**]. After having created story pieces, that is the core of the information to be communicated, they need to be transformed into a physical form of a design.

The layout of the visual data story, (defined during editor’s phase) characterized as “*design space dimensions*” in research “*Narrative visualization: Telling stories with data*”, where they analyzed 58 visualization, is composed based on three characteristics.

- Its genre. 7 main genres are magazine style, annotated chart, partitioned poster, flowchart, comic strip, slideshow, and video.
- Visual narrative tactics, such as “visual structuring, highlighting, and transition guidance”.
- Narrative structure tactics, “ordering, interactivity, and messaging” [**?**].

Those characteristics are chosen by the author, based on what kind of data or story nodes he has, and his rhetorical strategies (story, plot, narration) [**?**].

Those clear design practices give a more solid understanding of what does a visual data story consists from, how the layout of the design is shaped based on the three main characteristics, and together with the process of creating storified

DV they serve as author's basic toolbox for further recreation of visual data stories. However, the level of abstractedness or didascalicity that the stories have, should be intentional in order to relate to the audience better. For this purpose the goal of the story that is being created should be clear and support the goals of the system or the experience where the story is going to be presented. Author-Audience distance serves as a function of Narrative intelligibility and closure that the system achieve or aim to achieve [**?**]. As part of the communication strategy in relation to the problem of author-audience distance, Hulman & Diakopoulos and Segel & Heer suggest that in data visualizations there need to be certain balance between author-driven and reader-driven scenarios. This is described as an optimal interplay between unrestricted exploration from reader side and clear communication of the story from the author side. A balanced narrative should be a then a combination of persuasive, rhetorical strategies to transmit an intended information to users, and "exploratory, dialectic strategies" with aim to give the reader certain control over the insights, and allowing free interaction with the visualization. This is specifically called Martini glass structure with narrow, fully author driven beginning, and reader driven exploration at the end of the interface [**?**,**?**]. This introduces us to the concept of adding interactivity.

Interactive storytelling

An ability to interact with the visualization brings an exploratory aspect to the visualization, and makes the interaction more reader-driven, inviting the reader to be part of the story creation. Moreover, compared to static visualizations exploratory visualizations can also present a variety of perspectives on the same information [**?**,**?**].

Interactive storytelling is defined as user experience in the story being unique based on his interaction in the story world. That goes beyond simple interactions such as clicking, hovering and scrolling, beyond the typical linear structure of the story [**?**]. To clarify the concept, the figure 6 includes some examples of alternative structures (see figure 7).

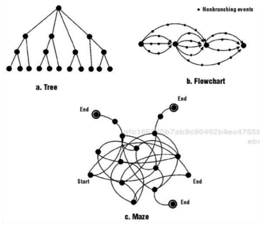


Fig. 7. Examples of interactive architectures affecting the story [**?**]

In the area of DV this would mean the possibility to explore the data without typical linear guidance, choosing the new story nodes in an unrestricted way while not going outside the structure of the overall story with beginning and ending.

FPS

After researching the fields of data analysis, data visualization, narratives and interactive storytelling we decided to narrow down the initial problem statement and form following final problem statement:

How can a big dataset be communicated through the use of dimensionality reduction, exploratory data visualization and storytelling techniques?

Methods

The goal derived from the final problem statement is to design a solution for communicating a large dataset using a dimensionality reduction algorithm. Further data visualization principles, storytelling elements and interactivity are to be applied. Aimed solution would provide narrative experience in form of interactive interface, with an emphasis on data visualizations and communication. In this section the fundamental methods for evaluating the solution are introduced.

Comunicativness and Usability

Methods for measurement and evaluation of the solution will be focused on the usability of the product, and how well the data were communicated. The communicativeness of data visualization will be evaluated by the identified DV principles of functionality/clarity of the presented information, insightfulness, and aestheticity, as mentioned in analysis. Usability will be evaluated by the easiness of navigation and use of the interface.

Questionnaire

Both qualitative and quantitative data collection method through the use of online questionnaire was chosen. Although there are withdrawals of this approach, it allows fast feedback from participants without being restricted to test one participant at a time, and shortens the necessary time. The questionnaire is described further in the following chapter of experimental design and can be seen in Appendix A.

Experimental design

The following chapter describes the design of the solution and evaluation process based on the requirements mostly originating in the analysis.

Requirements

The requirements for the design derived from the analysis are as follows:

- General topic of dataset to target general population: Food
- Large enough dataset dataset with many features: Nutrition alues
- Reduce dataset dimensionality using t-SNE to 2D representation: t-SNE
- Ideal platform for reaching general population is website: Website
- Visualize t-SNE output data points so the output can be explored and interacted with: D3.js
- Designing individual data visualizations, based on the principles of hierarchy of perception, and most common DV techniques: Visual elements
- Balance out author driven and reader driven approaches through the use of a Martiny structure narrative for the interface: Story
- Interactivity for user engagement: Interactivity

Platform chosen for presenting points was a website, as it can reach more people compared to stand-alone application, and is nowadays a common way of communicating infographics. As the main goal of the system is communication of information, the website further serves as a basis for evaluating the solution. The website encapsulates further requirements such as usability, and informativeness/insightfulness. Aim is to provide clarity of the represented information by having clear goal of the system and the narrative, enhancing the experience by adding context through storytelling and interaction.

Dataset

Chosen data set was open USDA National Nutrient database 2017 consisting of over 8400 points, with 41 dimensions. Each data is one food point, with dimensions consisting of the food category, amount of calories, proteins, carbohydrates, fats, fiber, amount of different minerals and vitamins, and recommended amount of those different minerals and vitamins per day. Multiple topics were considered, and at the end the reason for choosing this dataset over others was that the topic of food is very closely related to all humans and our well-being, and the dataset is easily accessible.

Dimensionality reduction with t-SNE

Then the dataset was analyzed, and visualized by using t-SNE algorithm in order to see the distribution of the points in the space while preserving local similarity structure (see figure 8). During the process of finding the optimal way of visualizing, the used algorithm parameters were chosen by comparing different outcomes of visualizations. The figure 7 below shows the comparison between principal component analysis visualization and final t-SNE maps of perplexity 10, 30 50, 90, and all have iterations of 1000. The final chosen was with parameter of perplexity 90 and 4000 iterations, and took around 45 minutes for the points in the two dimensions to be displayed.

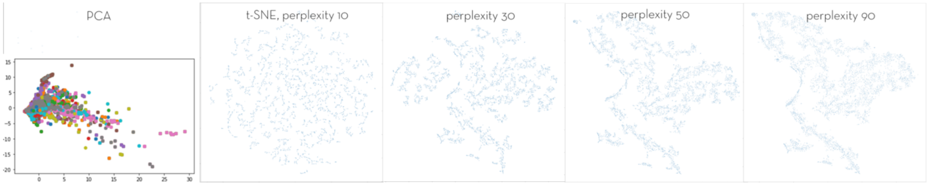


Fig. 8. Dimensionality reduction, visualizations comparison

The points were colored depending on what food group out of total 23 they belong to (see figure 9).

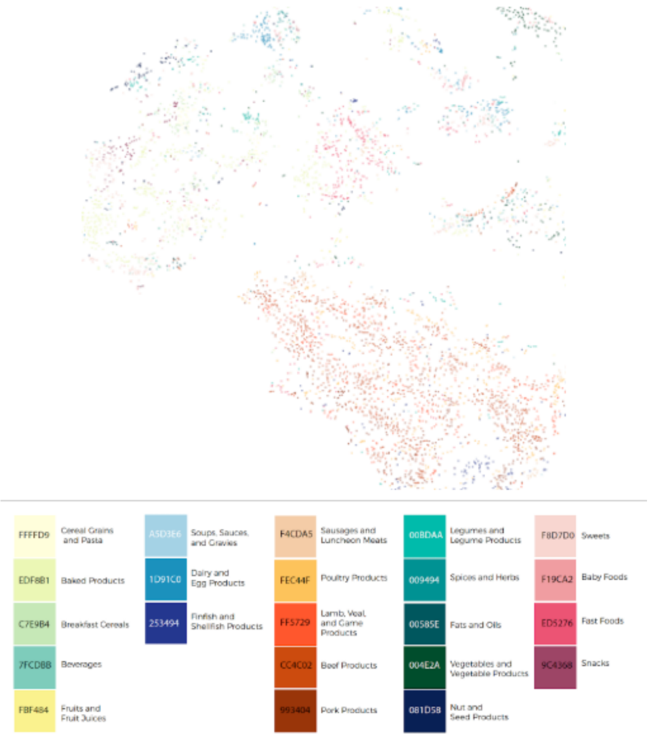


Fig. 9. Final used t-SNE map of food database and legend of categories

Visualization

Before continuing with the creation of visualization of each data point that would be accessible in the map, prioritization of what should be communicated and by what visual attributes was decided. This was based upon small research what matters to our body about food, and few discussions of the provided database and the intake from it with two medical students. For visually encoding data we have followed the hierarchy of elementary perceptual tasks as a rule for choosing the most appropriate form of visualising data by their importance [**?**]. Figure 9 shows the original visualization of a data point (see figure 10).

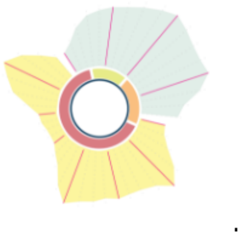
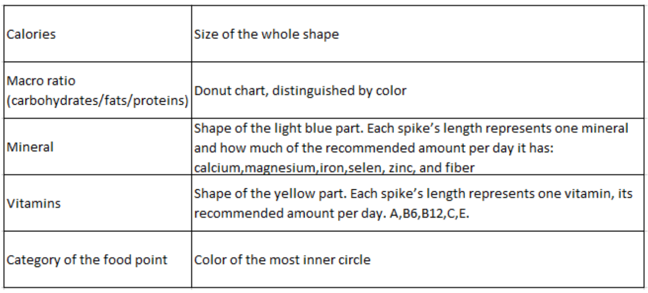


Fig. 10. Designed representation of each data point

The chosen characteristics for visualization and their visual attribute are ordered in the table below (see figure 11). The process of creating the visualization consisted of creating many various visualizations, funelling, merging, and choosing the final.



Calories	Size of the whole shape
Macro ratio (carbohydrates/fats/proteins)	Donut chart, distinguished by color
Mineral	Shape of the light blue part. Each spike's length represents one mineral and how much of the recommended amount per day it has: calcium,magnesium,iron,selen, zinc, and fiber
Vitamins	Shape of the yellow part. Each spike's length represents one vitamin, its recommended amount per day. A,B6,B12,C,E.
Category of the food point	Color of the most inner circle

Fig. 11. Visual representations of data point features

The intention was to communicate the differences and similarities between different foods. How different categories are distributed in space, similarities of foods within short distance even they are in different categories. The intention for each point visualizations was to create/strengthen a mind map of food representation, enhance the awareness of how our bodies perceive food rather then our taste buds. Donut charts are used for representing the amount in relation to whole, and as there are often used for representing macro ratio in food tracking applications, the representation was kept. Nutritions are shown in a form of length, affecting the overall size of are for the minerals and vitamins. Legend for each point in its final can be seen below (see figure 12).

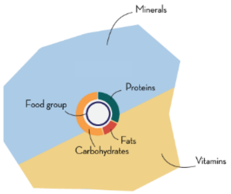


Fig. 12. Data point legend

Additionally to display the legend on the website, explained labels would appear on hovering together with fixed name of the food point. Comparison of different visualized points, after implementation, can be seen on figure below (see figure 13).

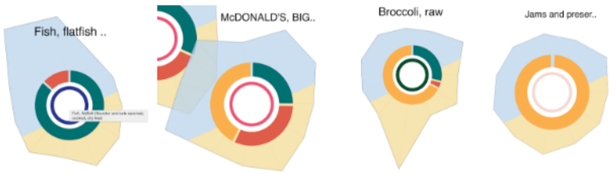
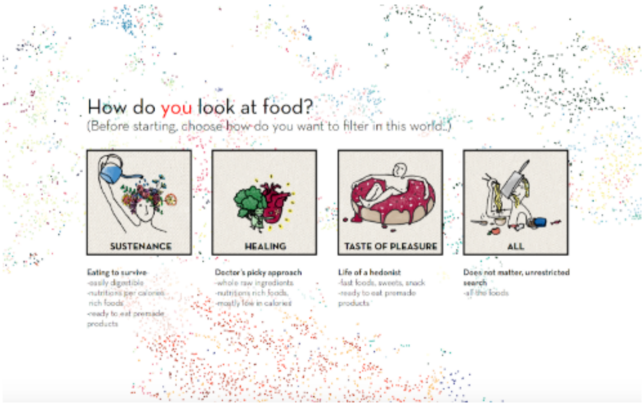


Fig. 13. Comparison of different food points

Storytelling and narratives

The goal of the narrative was to acquire narrative closure with the audience, and potentially intelligibility, with aim to question or reconsider the relationship with food user has, and the purpose of this relationship. The storyboard is attached in Appendix A. Story consists of three phases. Firstly, introduction to the topic of relationship with food, through a story reminding of the relationship with food from the time person was born until adolescence. This is done through an embedded slideshow and short animations, narrated through gradually appearing text.

Second phase transitions into introduction about the food world shown as points, presentation of it and of the navigation tools. In this phase the user can choose a character or a category as option for filtering from the total number of points, and to do that he needs to face questions connected to his perspective on food. Figure below shows the screenshot of the options (see figure 14).



How do you look at food?

(Before starting, choose how do you want to filter in this world..)

Eating to survive
-easily digestible
-nutritions per calories rich foods
-ready to eat premade products

Doctor's picky approach
-whole raw ingredients
-nutritions rich foods,
-mostly low in calories

Life of a hedonist
-fast foods, sweets, snack
-ready to eat premade products

Does not matter, unrestricted search
-all the foods

Fig. 14. Options of four categories before the phase of exploring the data stories

This leads the user to question how and why he chooses his food, to be reminded of what matters particularly to him about the meal. This should be guiding user towards the next phase with increased or reminded interest in the food topic. At the end of the second phase, inspirations for questions the user can try to answer by exploring the dataset are asked, and the explanation of navigation is explained.

The last, third phase then consists of purely reader driven exploration, with selective interaction. This is the conclusion of the story, and as the variety of presented information is large, the resolution is specific to the user depending on his insights. The points of possible inquiry is in form of a) data point stories representing the food nutrients, calories, and food group. User can choose specific point on the map, or write the name of the searched food in the navigation bar, b) comparison of that visualizations with other food points, or other food groups c) random recommendation by the system of the complementary food, that gets highlighted after choosing specific point in the map. Figure 14 shows the third phase, the starting point for exploration (see figure 15).

The plot of the complete story can be found in Appendix (see *Story plot* part). The exploration phase is intended as unrestricted exploration, fully reader

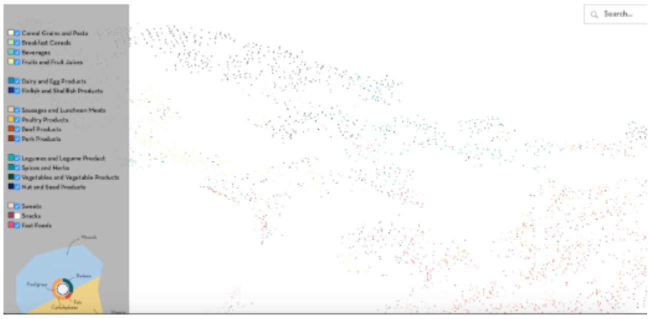


Fig. 15. Third, exploratory phase

driven for the user. User's specific investigation process should lead to his own conclusion of the presented story.

After the design was decided upon and communicated between members, prioritization of most important requirements for the system were followed during the process of implementation.

Experimental procedure

To collect self reported data regarding communicativness and usability of the solution, participants will be asked to fill out a pre and a post-questionnaire (see *Questionnaires* in the appendix), in a form of a survey, that will be given together with the website link. The questionnaires consist of demographic questions as age, occupation and gender, 8 Likert scale questions that ranged from 1-5, (1 for not at all and 5 for very much), and 3 open questions.

Pre-questionnaire serves as introduction of the participant to the topic, includes the demographic questions, and questions their own interest in the topic of food and reasons of their interest towards the topic. This is to compare if there will be any difference observed in their thinking when relating to the topic after the experiment.

After filling the pre-questionnaire, the participant gets a link to the website, and is being asked to return to the post-questionnaire section after finishing the experiment. Then in the post-questionnaire the Likert scale questions serve for self-evaluating the usability of the interface, for evaluating perceived attentiveness, and other questions to see if the user can relate to and understand the information presented. Additional questions for evaluation of the visualization on the informational (whether or not the user perceived it as insightful), aesthetic level, an investigation what kind of information was obtained, if any, and if what user was searching for was found. At the very end of the the post-questionnaire, two open questions serve for further reflection of the user's relatedness to the topic after the experiment, and optional space for feedback.

Implementation

For the implementation we decided to go with the tools and languages that are well supported and have fairly large community to ensure as fast problem solving as possible. We ended up with implementation of t-SNE algorithm reducing the dimensionality of the nutrition dataset for the purpose of demonstrating existing relations between the food nutrition features. The entire data analysis part of the solution up until visualization was written in Python language and it was executed in Jupyter Notebook environment (see *tsneFinal.pdf* and *OutputFormatting.pdf* in the appendix). Outputted format of the Python program was tsv file of t-SNE 2D position values (x,y) combined with initial data point's attributes (Food name, group and nutrition values) for the purpose of supplementing the visualization elements for better communication of the output to the users. For data visualization we chose website publishing of the rendered tsv file. Rendering was done using open source JavaScript library D3.js hosted on GitHub.

Tools

For the pre-processing, formatting, dimensional reduction, post-processing and outputting the following tools were used:

- Jupyter Notebooks (Python)

For the rendering of the outputted tsv file on website we chose following web development tools:

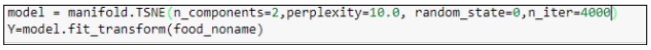
- JavaScript (with library D3.js)
- CSS
- HTML

Data analysis and processing

The process of data analysis of the data to visualizing the output can be described in following steps:

- t-SNE (see *tsneFinal.pdf* in the appendinx)
- Output Formatting (see *OutputFormatting.pdf* in the appendix)
- Visualization (see *Main.html*, *Index.html* in the appendix)

t-SNE Python implementation of t-SNE allows us to change lot of parameters as described in analysis chapter already. For the implementation we chose 2D representation (*n_components=2*) with perplexity of 10. The parameter *random_state=0* ensures the output is reproducible as the random generator within the algorithm use this parameter as seed. The most power consuming parameter is *n_iter* which is eventually responsible for accurate representation of the output which was set to *4000*. In the figure below the t-SNE configuration parameters using Python can be seen (see figure 16).



```
model = manifold.TSNE(n_components=2,perplexity=10.0, random_state=0,n_iter=4000)
Y=model.fit_transform(food_noname)
```

Fig. 16. t-SNE configuration using Python

Output Formatting Output of the t-SNE is a 2D array of XY coordinates that are used for displaying and plotting. With quiet few operations on the initial food dataset the following information was extracted:

- Food name
- Food category
- Food nutrition values (filtered)

The output of the t-SNE was then merged with the extracted information from initial food dataset to form final tsv file. This file is the only source of the data for the website application that is responsible for rendering of the data.

Visualization and story

As mentioned earlier the web development languages used for the visualization solution are HTML, CSS and JavaScript with library D3.js.

Narrative introduction When users lunch the application they will go through the narrative introduction of the application. This part ensures the users are getting somewhat idea of what the application is going to be about. After short story they will be shown brief instructions including demonstration footage. The transition between the narrative part of the application and the "core" visualization begins when users choose one of the four options in the end of the narrative part of the application. These options represent 4 different combinations of food groups. After choosing one of the options the food groups included in the selected option will be filtered out and displayed in the "core" visualization application. The narrative, introduction part of the application can be found in the appendix (see *Index.html* in the appendix) as well as the core application file (see *Main.html* in the appendix).

Loading First event in the application is loading of the tsv file and it is triggered when the user choose one of the options in the end of the narrative introduction part of the application. The process of loading the tsv file output from Python is done in separate JavaScript file that is responsible for loading of the tsv file into JavaScript array structure. This JavaScript file can be found in appendix (see *load-csv.js* in the appendix).

Initial render and Phase 1 After the loading script is done it calls a method responsible for creating initial DOM structure of SVG environment that will be used for the rendering and interaction with the data visualization. In this step the data points coordinates XY are being scaled to real coordinates used for displaying them on the screen. Data points are then rendered as simple circles with fill color of the respective category they belong to, creating so called Phase 1 layer. This phase is being active and visible to the users while the zoom exists within a threshold (scale 1-40). The figure below shows demonstration of the rendering of the Phase 1 (see figure 17).

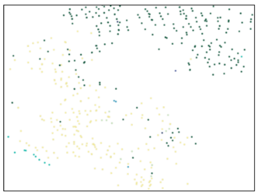


Fig. 17. Phase 1 rendering demonstration

Phase2 If users or application zooms over the zoom threshold the Phase 1 layer is hidden and Phase 2 is revealed and rendered. Phase 2 contains more detailed visualization of the data points but still respects the XY position of the data points the same way as Phase 1 does. The visualization of a Phase 2 data point consist of following features:

- Size of the entire element: Energy level
- Donut chart: Amounts of Carbs, Fat and Protein
- Polygon shape: Amounts of Minerals and Vitamins
- Circle: Food group/category

The visualization of the Phase 2 data point element can be seen in figure below (see figure 18).

Since there is quite few visual representations being rendered for single data point, the rendering of the Phase 2 data points is restricted and only those located within certain range from the point of entering Phase 2 are being rendered to reduce the system load dramatically. Demonstration of the rendering of the Phase 2 can be seen in the figure below (see figure 19).

UI and functionality Users can interact with the applications in various way. The manual controls allow following actions:

- Mouse: Zoom, Drag, Click, Double Click

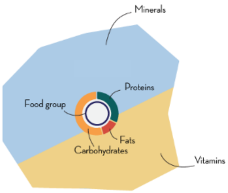


Fig. 18. Phase 2 element features

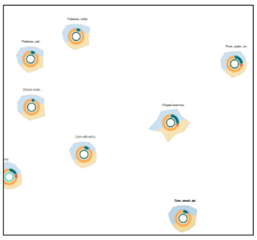


Fig. 19. Phase 2 rendering demonstration

- Keyboard: Esc

Users can navigate around the visualization freely with just use of a mouse input. In Phase 1 users can click on a data point (circle) to activate function that searches and highlights two nearest data points (foods) that are of a unique category. The Esc key is serving as reset button and resets the view zoom and position to initial state.

The UI provides more functionality to the users designed to improve the filtering and searching process. UI consist of two elements:

- Search bar
- Food categories - Checkbox panel

Search bar is located in the top right corner and allows users to search for a food item by typing in a keyword. When users type in the keyword, pressing Enter button confirms the input and searches for the keyword in the array of the data. If a single record has been found the application will execute smooth transition from current view to the target location and zoom in to Phase 2. If multiple data points containing the keyword were found the list of all these data points will be shown to the user as scrollable list. Clicking on any of the data points from the list will execute the transition bringing the view to the data point same way as mentioned earlier. The figure below shows demonstration of search bar with multiple matches found (see figure 20).

Last UI element available to the users is the food group checkbox panel for filtering food groups. Here users can simply check and uncheck food groups to be

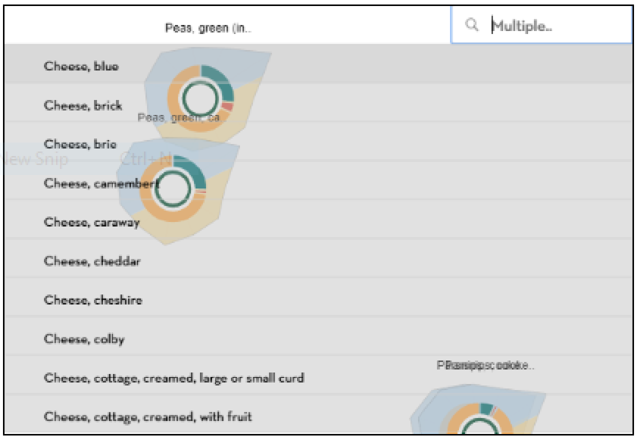
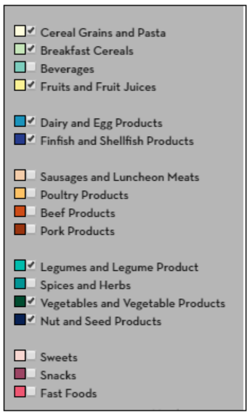


Fig. 20. Demonstration of search bar with multiple matches found

displayed in both Phase 1 and Phase 2. Elements that are not being displayed (unchecked) will not be intractable with and therefore searching the item in the search bar in top right won't trigger transition to it. Figure below shows demonstration of the checkbox menu (see figure 21).


☑ Cereal Grains and Pasta
☑ Breakfast Cereals
☐ Beverages
☑ Fruits and Fruit Juices

☑ Dairy and Egg Products
☑ Finfish and Shellfish Products

☐ Sausages and Luncheon Meats
☐ Poultry Products
☐ Beef Products
☐ Pork Products

☑ Legumes and Legume Product
☐ Spices and Herbs
☑ Vegetables and Vegetable Products
☑ Nut and Seed Products

☐ Sweets
☐ Snacks
☐ Fast Foods

Fig. 21. Demonstration of checkbox menu with food groups

Evaluation

Experiment procedure.

After internal testing in between team members and pre-tests by observing the interaction of two close friends, we chose to start the experiment by sending out the website link between friends and posting in facebook groups. Total of 23 participant's answers were collected in span of two days. The experiment took approximately 5 to 10 minutes, with identical questionnaire and procedure for all participants.

Results from experiment

There were 14 male participants and 9 females, aged from around 18-54 years old, but mostly students in between 18-35.

Pre-questionnaire Pre-questionnaire analysis shows that from the likert scale questions the mean for the question "how much do you care about food in general" was 4.1. That means, most of the participants considered themselves as having general interest in the food. From analyzing the occurrence of keywords for the question of what exactly people care about food, the most common answers go from taste, price, nutritions, quality and lastly the ecological impact.

Post-questionnaire Post-questionnaire analysis shows that most of the people considered themselves as somewhat attentive to the content, with a mean from the scale being 3.57. The interface felt easy to navigate also somewhat, with a mean of 3.3. Clearness and understandability was 3.3, "somewhat." Visualization felt "much" aesthetically pleasing, with mean of 4. Visualization sparked insight, somewhat, with a mean 3.4. From the question of multiple checkboxes, it shows the interface was mainly used for exploring around, 15 people, comparing nutritional value, 8 people, and around 5 people marked that they have used visualization for comparing macro ratio or as searching information. Optional answers included quotes as "looked into what sweets contain" or "enjoyed looking at the transitions." Visualization was helping to find the searched information only into somewhat extent, with mean of exact 3. In the question if the perception of food has changed at any level, after this experiment, 12 said no, 8 said somewhat, one yes, and one marked other option. From the optional question about what does a healthy relationship means to you, the 14 obtained answers varied in being filled with words mostly just "eating healthy" and to only few elaborated answers on what does this "health" means, and wrote more specifically as "eating less junk food," "eating less," "stop overeating," "going for energy instead of pleasure," "eating consciously," "choosing loving kind relationship," "controlled relationship" etc.

Optional feedback Optional feedback about the experiment was positive overall about the presented concept. Yet further potential improvements regarding usability were mentioned a lot. One of the quotes that summarizes the overall feedback quite well: “The app is not very user friendly; I had a lot of trouble understanding the purpose of the app. As a user, the first 10 seconds of using a new app is critical for me and more often than not it decides whether or not I like the app and stick with it. To this moment I am not entirely sure if I had learned anything the app is trying to show me. The points visible on the website strike me as chaotic, the food labels are incomplete and the whole presentation of the data is not intuitive. I would definitely recommend a revamp of the front-end side of the project. However, the idea is original and very compelling, and once the presentation side of things is fixed I am sure there will be many users who will find the app helpful and appreciate the knowledge it holds.”

Summary

From the pre-questionnaire, commonness of prioritizing taste when it comes to eating can be recognized as main drive for choosing the food. However, from analyzing and comparing the answers after the experiment, there can be seen a pattern of the answers being more concerned about the health side of the food, even though perhaps definition of “healthy food” is not clear for most of the participants. This suggest that the goal of the narrative was partly fulfilled, through the connection with the participants, and reconsideration of the relationship with food participants have. For properly evaluating the usability of the interface and the communicativeness of the visualization, more data are needed to confirm the results.

However, the results suggests the presented interface and information was overall only “somewhat” communicated and usable, leaving space for a lot of potential further improvement that will be more elaborated in the discussion chapter.

Discussion

The analysis provided us the core techniques and principles of the process of creating data visualizations. That was the main focus and motivation of the research. The emphasis was to connect various quite wide areas, such as machine learning, storytelling, interactivity, data visualizations, and information communication to improve the communicativness of the information (food nutrition similarities). The time schedule was poorly designed and the width of the research field became a burden. Time span for designing and implementation was short. Ideally more user testing would greatly improve the final solution as the feedback is the essential variable in process of the software development. Interface navigation, was designed based on common types of interaction, but rather intuitively. Aesthetics of visualization, narrative and the connection with the audience would also have been improved with iteration during the designing phase.

There were many possible confound factors regarding evaluation. Evaluation consisted only of self reported measures, without any data source triangulation. Even though the questions for questionnaire measuring usability and communicativeness were well researched and tested, half-structured interviews and observations would help for further clarification to determine the user's experience from more objective point of view.

However, altogether this research managed to connect various disciplines and deepen our understanding about the field. Results are only suggestive and can be considered favorable regarding the overall idea, but additional, more elaborated evaluation with more design iterations is necessary. Other points regarding the future perspective, and what can be improved in the process is reflected upon in the chapter below.

Future works

We believe the field of combining AI for dimensionality reduction with the elements of storytelling and DV is feasible and will continue growing in the near future. For further improvements in the future research multiple points can be taken.

Advanced data analysis From the analysis we know that t-SNE and PCA are mainly dimensionality reduction algorithms that does not necessarily communicate accurate relations in the dataset. For more accurate representation of these relations further research into the machine learning and specifically clustering algorithms should be done.

Generative art Generative art is fascinating field that is directly connected with communicating and perceiving of the information. This field should be researched in the future to ensure simpler but more impactful visual representation solution of the problem.

Time management Developing proper time management would greatly improve the validity of the solution.

Evaluation For main evaluation, comparison of different communication strategies between control and multiple experimental groups would help to clarify and evaluate if adding all the areas included in this research makes the communication more effective, and to see the difference. Additional proposed measure would be testing user's engagement and depth of interaction. More participants and gathering and triangulating data through semi-structured interviews, observations, web-analytics, and using statistics for analyzing the results would increase both the validity and the reliability of the product.

User testing and feedback From the perspective of implementation, interviews with the users should be done regularly during the experimental deign phase including re-evaluation of the methods for choosing specific narrative introducing the concept. Users' feedback would drive the solution the right way.

Separate usability and intuitivness evaluation The usability and intuitiveness of the interface could be rather as separate testing to identify possible problems before testing the way the information is communicated.

Conclusion

The aim was to communicate a large dataset, consisting of over 8 000 data points, by bridging multiple areas as machine learning for data dimensionality reduction, exploratory data visualization and storytelling techniques. With the use of web-page development tools by following the principles from analysis, we implemented an interface providing a narrative interactive experience. Firstly the dimensionality of originally 41 was reduced into a 2D map including all the data points by using t-SNE machine learning algorithm. Then a data visualization was created for each point and visualized using a JavaScript D3 library. The map was made explorative with features of choosing what food categories to visualize, exploring through zooming in and out of individual point visualizations, getting random complementary food inspiration after clicking, or finding specific foods with the search button. Short presentation of this map including tutorial was added, together with a story consisting of introducing and relating the user to the overall topic. Whole experiment was shaped into a Martini glass narrative structure. The core of the experiment was to evaluate the effectiveness of communicating this dataset through the created website. As measurements, the communicativeness of the interface and its usability was tested through analyzing self-reported pre and post questionnaires collected through surveys distributed together with the link to the hosted website. 23 answers were collected and analyzed. The design requirements can not be considered as met and there is no conclusion that can be done from the obtained data, and further evaluation is necessary. This is due to mainly lack of further design iterations before final testing, and necessity of collecting other data sources than self-reported measures. However, from the obtained answers the goal of the narrative is considered as partly fulfilled. From the perspective of the potential effectiveness of the whole proposed interface, and taking into the account the complexity of the proposed design, its effectiveness from this research is rather inconclusive. We believe further iterations, bigger and more randomized population, a better tool for measuring and multiple sources of data collection, and more elaborated experiments divided into more groups are great opportunities for exploring this idea further, leading into a project of a bigger scope.