Bachelor thesis
Computer Science


Radboud University

# Expanding and evaluating a web-based multidimensional scatter plot tool

*Author:*
Thijs Werrij
s4305493

*First supervisor/assessor:*
Prof. dr. T.M. Heskes
t.heskes@science.ru.nl

January 26, 2017

**Abstract**

Scatter plots are a useful, but limited way of visualizing data. Can they be expanded to show more information in an image and does this also positively affect the users of these visualizations? To test this, this thesis describes the implementation of a web-based scatter plot visualization tool, which also has added new ways of adding variables to a plot. These ways are tested, along with some criteria of visualizations and tools, using expert reviews. The research finally concludes that expansion of scatter plots can be done, and in some cases may positively affect readability.

# Contents

# Chapter 1

# Introduction

Scatter plots are a well known and easy way to visualize data up to two or three dimensions. But what if you want to visualize a data set with twenty variables? Then you would need to generate multiple plots, or use another solution like dimension reduction. Would it not be easier to visualize multiple variables in one image? That is why, in this bachelor thesis, I focus on the following question: can scatter plots be expanded further and what influence does this have on the readability?

We will be looking at iDarkSurvey [1], a web-based visualization tool which can visualize a data set about dark matter in histograms and plots of two and three dimensions. This tool is limited, because it can only show three dimensions at once at maximum (plus extra color dimensions that cannot be chosen freely). The iDark team is also interested in the question whether you can expand a scatter plot, so next to the question if scatter plots can be expanded, my goal is to also make a new version of this tool, with added variables.

In the second part of the research, this tool and its expansions will be evaluated. A small group of experts will evaluate the tool based on visualization criteria, interactions with the tool and the functionality of the added variables.

# Chapter 2

# Preliminaries

A **scatter plot** is a multidimensional way of visualizing data where you can display data as points in a two or three-dimensional graph.

**Chernoff faces** were first described by Herman Chernoff as faces that could represent multiple variables by using their facial features [2]. For example, the mouth size and the curvature of the smile (happy, sad or somewhere in between) can represent different variables. For my version of the tool, I have used Chernoff faces with variable face, mouth and eyebrow sizes, and for the nose and eyes the width and the height (examples in Figure 2.1). This way, you can represent many variables of a point in just a single face.

Figure 2.1: Two examples of Chernoff faces, as used by my iDarkVis implementation. The variables associated with the eye height, mouth and face shape clearly differ, while other characteristics look alike. Generated with D3.

# Chapter 3

# Research

## 3.1 The tool

For my research, I will be implementing a scatter plot tool and gradually adding new ways to visualize variables, so that later on they can be tested. Because of this, I contacted iDark, as they already had created a tool that uses scatter plots to visualize large data sets, called iDarkSurvey. The research originally started with an assignment: they wanted to make a new visualization tool, so first they were looking how to improve on the old tool. Whether the tool could be expanded with extra variables was one of the most important questions, as they wanted it to be optimal for large data sets (initially around twenty variables).

They had a list of requirements for me:

- It needs to be a D3.js implementation with scatter plots that looks and works like the original iDarkSurvey and can be hosted online
- Needs to be expanded with multiple variables, so that more information can be seen in the visualization
- Adding compatibility for visualizing multiple data sets at once
- Being able to save the data set

These are the requirements that have been implemented. In the following parts, I will explain what these requirements mean and how they have been implemented.

### 3.1.1 D3

iDarkSurvey was made using Highcharts. This works well for the original tool, but for the new tool I had to use something different, as I had to make expansions to the tool that are not supported by Highcharts. It also had to

be hosted online, so it was not an option to choose existing programs like Matlab. This is why I chose D3.js [3]: there is more programming involved, but there is also a bit more freedom when you want to make changes to the charts themselves. While creating the tool, I have tried to stay as close to the original as possible. I did not really spend much time on the look of the tool, but the tools mostly work the same. For example, zooming works via dragging a box and you can choose the plotted variables yourself.

There are some changes which I will have to point out: there is no real 3D, like in the original. This has been replaced by a 'multiplot' mode, in which you can view multiple plots at the same time, so you can still automatically simulate a third variable. At first this was necessary because I had trouble with the implementation, but later I decided it would also be interesting to look at this in the research and see what difference it makes when compared to actual 3D.

Another change is that you cannot decide how many points there are in the plot. This was purposefully left out, because then you would have to work with data set algorithms that decide which points are important and which can be left out (e.g. k-nearest neighbors algorithm), on which my research does not focus. Mind that my tool is not a finished product, but built according to initial requirements; focusing on visualizations, because that is the scope of my research.

### 3.1.2 Variable expansion

The most important change to the original tool would be the extra variables. First, I had to decide what expansions I would try to implement. To simulate extra variables, I have used three visualizations: the color of the dot, the size, and the Chernoff face that goes with the dot.

I decided against implementing a form of discrete variable visualizations, as I mainly worked with continuous data sets. Cleveland et al. [4] talk about possible discrete variable implementations. I chose color and size because they were the simplest additions to implement, while still having a big visibility within the plot. Color was already implemented in the original tool, but was not used the way that I wanted to: a user should be able to freely choose a variable for the color scale.

Chernoff faces were chosen because of the high amount of variables that can be represented. I could also have chosen for another multidimensional visualization, like star plots. In a star plot, you represent variables as axes all coming from the same point, where the size of the axis indicates the value. The problem with this is that star plots using data sets with very

differing variables, where some variables can be in the millions and others in the tiniest fractions, could possibly confuse users, as the scaling would be the same for every variable. I chose Chernoff faces because every facial characteristic is different, which hopefully makes it easier to focus on just one aspect. Star plots would be an interesting suggestion for future work.

### Color

To add color to a dot in a scatter plot in D3, you have to add it to the 'fill' style of the dot itself. For this, I used color scales by chroma.js [5]. With this JavaScript library, you can simply choose a color range you like and then generate a color from the scale by giving a number between 0 and 1 as parameter. To do this, I had to normalize the numerical values of the data set to a scale of [0,1]. For this I used the formula:

$$x' = \frac{x - \min}{\max - \min}$$

where x is a value of the variable you want to normalize, and max and min are respectively the highest and lowest value of that variable in the data set. This is also known as feature scaling. For example, when using the array [2; 3; 4; 6], the formula will transform value 3 into $(3 - 2)(6 - 2) = 0.25$ (the array will become [0; 0.25; 0.5; 1]).

### Size

Size of a dot can be decided by changing the value of 'r' (radius). For this, I have also used the normalizing formula described above, but multiplied it by 4 and added 3, so the radius will always be between 3 and 7 pixels for any value.

### Chernoff faces

The Chernoff faces were added by using an existing D3 example [6]; the only change I made was the removal of the hair, because I wanted it to look more like the original Chernoff idea, and because it looked a bit silly. At first I wanted to see if it was possible to add Chernoff faces instead of dots as an alternative, but I decided that this was both too hard to accomplish and would be too confusing. I decided to show the Chernoff face next to the plot when you hover over a dot, accompanied by the variable values that define it (which can also be chosen freely). The Chernoff faces are initialized by values between 0 and 1 or -1 and 1, so here the normalizing formula had to be used again.

### 3.1.3 Other requirements

The other requirements include being able to load multiple data sets at the same time and being able to save the data set when clicking on a dot. These are not really relevant to this research, but as they have been implemented, they might be worth mentioning.

### 3.1.4 Domain

Before concluding the part about the tool and following up with the questionnaire, first something about the domain of the scatter plots. The data set I have used, is used in dark matter research. It contains observations of dark matter and the variables associated with it. The reason I have chosen to expand this tool is not because of the subject of the data set, but because it is numerical and has many variables. This is of course interesting when expanding and testing scatter plots with multiple variables. Understanding the domain itself is not really relevant to this paper, but it might be interesting for recreating or expanding this research.

### 3.1.5 GitHub

The code of the new tool can be found online on GitHub. See [7].

## 3.2 The questionnaire

To answer the second part of my research question, I conducted expert reviews. In an expert review, you let a small group (maximum of five) interact with a tool, after which they can give their opinion about it. This can give good feedback on not only what is good or bad, but also on how to improve certain features. That is one of the reasons I chose to do expert reviews. The other reason is that this tool is still a prototype. Tory and Möller [8] recommend to first do expert reviews on prototypes, and later, in the final stages of the tool, also conduct quantitative research with a larger group of participants. Most of my results can not be quantified anyway, as I said earlier that I also need feedback on how to fix certain problems.

To assist the expert reviews, I designed a questionnaire with multiple open questions, following the procedure as described in [8]. The questionnaire consists of three parts (see appendix):

1. Visual representation criteria
2. Task list evaluating interaction criteria
3. Evaluating the changes between the tools

### 3.2.1 Visual representation criteria

The first part is based on visual representation criteria described by Freitas et al. [9], using the criteria that felt the most relevant to the tools. I will give descriptions of the extracted criteria.

- Data density

  Data density means how close data points are standing to each other and the amount of data points. This can be important to find out, as very dense data can make some points unreadable, as they are hidden behind other points.

- Number of dimensions

  This means the total amount of dimensions, or variables, the tool can visualize. The participant will need to judge if the tool has too many, which can be distracting or confusing, or just enough.

- Relevance of displayed information

  Is everything displayed in the tool relevant, or can something be left out?

- Logical object locations and visibility

  The participants will need to judge if objects are in the right places and visible.

- Loading times

  Loading times are important for a visualization tool, because a very long loading time can make a tool unusable.

- Clarity of relation between old and newly generated image after interactions

  After using, for example, zoom, is it still clear how the new image relates to the old one? Because if not, this would mean that function is either broken or useless.

### 3.2.2 Interaction criteria

The second part is based on the task list by Shneiderman [10]. It is a list of important interactions a user can have with a tool. Interaction is an important aspect of visualization tools, as it makes the difference between just a still image and a changeable, interactive tool. See the following list:

- Overview: Gain an overview of the entire collection.
- Zoom: Zoom in on items of interest.

- Filter: Filter out uninteresting items.
- Details-on-demand: Select an item or group and get details when needed.
- Relate: View relationships among items.
- History: Keep a history of actions to support undo, replay, and progressive refinement.
- Extract: Allow extraction of sub-collections and of the query parameters.

### 3.2.3 Tool differences

Lastly, the third part consists of questions I made to compare the features, ultimately the most important part of this research. I have simply made one question for each notable change between the tools, in this case the use of color in dots, the size of dots, the use of Chernoff faces. I also added a question about the difference between actual 3D in the original tool and the 'simulated' 3D in the new one, where three 2D plots simulate three variables.

## 3.3 The evaluation

For the actual evaluation, I had a question session with a few participants who either had experience with visualizations or the domain of the data set. As I did not want them to influence each other, I made them fill out the questionnaire themselves. Before starting, I showed them where they could find both tools, let them play with them for some time so they got an idea of how they work, and then gave them some time to answer the questionnaire. While this happened, I responded to their questions and explained the functionality of the tools.

# Chapter 4

# Results

In the following chapter, I will list the results of the questionnaires, categorized by each subject. Then I will discuss the results.

## 4.1 Visual representation criteria

### Data density

All participants agree that there is quite a bit of overlap in the data points in both tools, to the point that some points are unreadable. This is dependent on which variables you plot and which data set you are using. Participants suggest either using the zoom function to focus on certain parts of the data, or implementing a clustering algorithm.

### Number of dimensions

Four participants agree that the number of dimensions is not too high at the moment, as they are okay with the Chernoff faces, which is also optional. The fifth has no opinion. Two participants recommend not adding any further dimensions, as this would confuse users. Adding non-continuous (discrete) visualization options, like shapes, would be an option for data sets with those kind of variables.

### Relevance of displayed information

The design is minimalistic and no info should be left out. One participant suggests that some Chernoff characteristics are too subtle. Another notes that there is not enough explanation of functionality in the tool.

### Logical object locations and visibility

About the tools themselves, the participants said that the locations and visibility were fine. One participant suggested to reorder the Chernoff face

variables so that variable names are in front of the drop-down menus. One other remark was that some images in the plots were too dense, but this was also described in the question about data density.

#### Loading times

All participants agree that the loading times are fast, but two remark that this worsens for more dense data and could be very slow when using a data set with a million points.

#### Clarity of relation between old and newly generated image after interactions

From most participants, I got a clear 'yes' on the question about if it is clear how a new image relates to an old one after interaction (e.g. hovering, zooming). One participant says it could be improved by using animations. One participant says 'no', but without a clear explanation why.

## 4.2 Interaction criteria

For this part, the participants had to sum up which given interactions can be used in the tools and which of those are necessary or should be added. Apparently there was a bit of confusion about this question, as only four participants listed if the tools have those interactions implemented, and two named which were necessary. See table 4.1 below (the responses are formatted like 'is it implemented'/'is it necessary').

| **Participant** | One | Two | Three | Four |
|---|---|---|---|---|
| **Overview** | Yes / Yes | Yes / Yes | Yes | Yes |
| **Zoom** | Yes / Yes | Yes / Yes | Yes | Yes |
| **Filter** | No / No | No / * | Yes | No |
| **Details-on-demand** | Yes / Yes | Yes / Yes | Yes | No |
| **Relate** | No / No | No / * | Yes | No |
| **History** | No / Yes | No / Yes | No | No |
| **Extract** | Yes / Yes | Yes / Yes | Yes | No |

Table 4.1: 'Are the following tasks implemented/are they necessary?'

As you can see, participants one and two agree on which functions are implemented, but participant three apparently also saw filtering and relations. This was probably confusing because of the zooming and other functions.

Participant four could not find some tasks that were present. This might be because the tools themselves do not have much explanation.

Furthermore, the first two participants also gave feedback on what they found important. They said that the implemented tasks are all necessary, and both agree on that there should be some sort of history (one suggested an 'undo' function).

## 4.3 Tool differences

### Colored dots

One of the most important complaints about the color is that it can be confusing or too subtle to read. One participant says that when there is an outlier, the entire color scale adapts, which makes all non-outlier points look too much alike to see a difference. One user does like how dynamic the new way to control color is, but also suggests implementing something like in the old tool, where color indicates which data set the point belongs to. The new tool does contain some form of this kind of functionality, but it has not been tested as it was not a high priority functionality. Another user thinks the colors are unclear, but does think it could have advantages.

### Dot size

Because of the data density, dot size is hard to read. This can be fixed by using zoom to see the points better, but the actual problem could be solved another way (see part about data density). Two participants say that the point size does not work well, because it is confusing. Another user suggests using width and height for points, as it would be read easily and add an extra variable dimension.

### Chernoff faces

Four participants are positive about the faces themselves, but two suggest to make the changes between faces more visible. The other two do think that the changes and similarities between faces are shown well. The fifth participant does like the idea, but says it would be better to implement a way to compare two faces with each other, as the current way of using faces does not improve readability.

### 3D

The opinions are divided on this issue. One participant thinks 3D is necessary to show the data in other ways. Another thinks 3D is better, because it gives you a quick scan of the data. A third states that it could be useful,

but not necessary, as a too high data density could obscure the image. The last two participants do not think 3D is necessary. One thinks the Chernoff faces give a nice alternative to 3D. The second does not like 3D plots and prefers the multiple plots that simulate a third variable.

## 4.4 Discussion

It seems like the participants were mostly positive about the tool when looking at the visual representation criteria. The main complaint was the data density, as many data points are too close to each other and overlap, which makes those points hard to identify. As said earlier, this can be solved by the users themselves by zooming, or can be solved by a developer by using a clustering algorithm. The original tool already had the option to limit the amount of points in the plot, but this is something that I did not focus on in my new plotting tool, so this could be a suggestion for future tools.

Participants were also positive about the amount of dimensions in the tool. There were not too many dimensions, which would confuse users, but they also said that a limit has been reached and that you should be careful when adding more. So for future tools, it would be reasonable to test extensions as alternative functions. You could also test further to see when you actually reach the limit of expansions.

Loading times are not really a problem right now, except for the 'multiplot' option. One of the participants was correct in stating that the loading times will become a problem when using bigger data sets. This could also be an argument to implement some sort of a clustering algorithm, because the data density and extreme loading times could make such a tool unusable.

There is not much to discuss about the interaction criteria. Most participants did have an idea about which interactions were implemented, but one participant also missed multiple. This is probably because of the missing explanation in the tool itself. It would be important for a future, finished tool to have explanations, as ordinary users could have more issues with understanding the tool. Adding a history function is a good idea, but was temporarily solved by a reset zoom button, as history did not have a high priority in development.

The best expansion of the tool were probably the faces, as the reaction to them was mostly positive. All expansions received some criticism that they needed improvement, especially color and size. The 3D vs multiple plots question does not have a clear answer yet, but now we do have a list of pros and cons.

# Chapter 5

# Related Work

There is a lot of information available about visualizing multi-dimensional data. Most sources only refer to scatter plots when talking about scatter plot matrices (similar to the alternative to 3D in my tool), for example Wong et al. [11]. Hao et al. [12] do use color as a third variable in their scatter plots, but only for data distribution and clustering. Cleveland et al. [4] actually suggest ways to expand scatter plots, but these are mostly for discrete variables, which could not be used in my research because of continuous data. Friendly et al. [13] talk about several enhancements, like scatter plot matrices, but also glyph plots (also known as star plots). In development, I decided between star plots and Chernoff faces [2], but chose the latter (for explanation, see 3.1.2).

Articles that talk about the use of expert reviews are [8] and [14]. Expert reviews are useful when evaluators need special knowledge or skill to work with a tool. Tory and Möller [8] explain that, even though expert reviews are usually used in a usability context, it could also be useful for evaluating visualizations. They also conclude that expert reviews should not be used exclusively, but could complement formal user studies. They do, however, say that 'One possibility is to have experts evaluate early prototypes (formative evaluation), and then end users evaluate a refined version (summative evaluation).'

For the questionnaire, I looked at existing evaluation methods of visualizations. Freitas et al. [9] talk about visualization criteria; for my questionnaire, I chose those criteria that felt most relevant to the tools and based questions on them. Shneiderman [10] proposes a task list of possible interactions with a visualization interface. I have based a part of my questionnaire on this, mostly because I was interested in the opinions of the experts on which of these tasks are (un)important for the tools.

# Chapter 6

# Conclusions

It seems to be possible to expand scatter plots to include more dimensions in just one image. I have been able to develop a web-based tool capable of doing this. The added readability of expansions differs, but it looks like a promising subject to do more research in. The Chernoff faces were, surprisingly, the best addition to the tool. The color and size had potential, but are limited by their current implementation, as sometimes they can be too hard to read. The tool was overall fine, as there were not many complaints about the tool itself. There is still the issue of data density and the loading times, especially when you might want to use bigger data sets in the future.

For further work, I would suggest several things. Firstly, add and test more expansions. I have done only a few, but there must be other interesting things that could be added to a tool like this. One suggestion could be discrete variables, where certain values could be represented as, for example, squares, circles and triangles. Secondly, solve the data density problem. When dealing with bigger data sets, there really should be some form of clustering algorithm, so I suggest working on that. Finally, finish the tool and do quantitative testing, as I explained in Related work that expert reviews should not be used exclusively, and there should be some form of quantitative testing in later development.

# Acknowledgements

I want to sincerely thank my supervisor Tom Heskes for all the help during my project, and Peter Achten and Janos Sarbo, who supervised the bachelor thesis course.

I want to thank the iDark team, including Sascha Caron, Faruk Diblen, Luc Hendriks and Bob Stienen, for providing requirements and help during development, and all experts who participated in the evaluation.

And finally I want to thank my parents, Frans and Thea, my sister, Annemarie, and Diede Heitink for proofreading and giving their support.

# Bibliography


[1] iDarkSurvey. `http://www.idarksurvey.com`.

[2] Herman Chernoff. The use of faces to represent points in k-dimensional space graphically. *Journal of the American Statistical Association*, 68(342):361–368, 1973.

[3] D3.js. `https://d3js.org/`.

[4] William S. Cleveland and Robert McGill. The many faces of a scatterplot. *Journal of the American Statistical Association*, 79(388):807–822, 1984.

[5] chroma.js. `http://gka.github.io/chroma.js/`.

[6] D3 plugins (Chernoff face). `https://github.com/d3/d3-plugins`.

[7] iDarkVis prototype. `https://github.com/idark-project/idarkvis`.

[8] Melanie Tory and Torsten Moller. Evaluating visualizations: Do expert reviews work? *IEEE Comput. Graph. Appl.*, 25(5):8–11, September 2005.

[9] Carla MDS Freitas, Paulo RG Luzzardi, Ricardo A Cava, Marco Winckler, Marcelo S Pimenta, and Luciana P Nedel. On evaluating information visualization techniques. In *Proceedings of the working conference on Advanced Visual Interfaces*, pages 373–374. ACM, 2002.

[10] Ben Shneiderman. The eyes have it: A task by data type taxonomy for information visualizations. In *Proceedings of the 1996 IEEE Symposium on Visual Languages*, VL '96, pages 336–, Washington, DC, USA, 1996. IEEE Computer Society.

[11] Pak Chung Wong and R. Daniel Bergeron. 30 years of multidimensional multivariate visualization. In *Scientific Visualization, Overviews, Methodologies, and Techniques*, pages 3–33, Washington, DC, USA, 1997. IEEE Computer Society.

[12] Ming C. Hao, Umeshwar Dayal, Ratnesh K. Sharma, Daniel A. Keim, and Halldór Janetzko. Visual analytics of large multidimensional data using variable binned scatter plots. volume 7530, pages 753006–753006–11, 2010.

[13] Michael Friendly and Daniel Denis. The early origins and development of the scatterplot. *Journal of the History of the Behavioral Sciences*, 41(2):103–130, 2005.

[14] Niklas Elmqvist and Ji Soo Yi. Patterns for visualization evaluation. In *Proceedings of the 2012 BELIV Workshop: Beyond Time and Errors - Novel Evaluation Methods for Visualization*, BELIV '12, pages 12:1–12:8, New York, NY, USA, 2012. ACM.

# Appendix A

# Questionnaire

### User descriptions

- A researcher who wants to visualize dark matter data, but thinks the original tool is too limited and is looking for alternatives.
- A student who works with iDarkVis and thinks it is already good enough. What could the new tool add/take away? Changes between old and new tool

## Questions

### Visual representation criteria

Answer the following questions for both visualization tools (if there are notable differences, write them down):

1. What do you think of the data density (are you able to read data without too much overlap)?
2. What do you think of the number of dimensions and where would you draw the line as what is useful and what is distracting?
3. Is all the displayed information relevant and what could be left out?
4. Are objects in logical locations and not occluded by others?
5. Are loading times for zooming okay? And what about the loading times of Chernoff faces when hovering over different points?
6. When the image changes after interaction with the tool (hovering, zooming), is it still clear how it relates to the old image?

### Task list

For the following tasks, answer the questions: have these tasks been implemented, and are they necessary for the tool? If they are necessary, but not implemented, how could they be added?

- Overview: Gain an overview of the entire collection.
- Zoom: Zoom in on items of interest.
- Filter: Filter out uninteresting items.
- Details-on-demand: Select an item or group and get details when needed.
- Relate: View relationships among items.
- History: Keep a history of actions to support undo, replay, and progressive refinement.
- Extract: Allow extraction of sub-collections and of the query parameters.

### Changes between old and new tool

1. What do you think about the use of color in the old and new tool?
2. Do you think adding an extra dimension using data point size helps with readability? Why (not)?
3. Do you think adding Chernoff faces helps with readability? Why (not)?
4. What about the missing 3D: would it be necessary to implement it anyway or are the multiple plots with the added z-variable a good enough alternative?