Visualization and Visual Data Analysis

Nicole Cherches (01506832), Alexander Gelb (01268620),
Benjamin Neckam (01301917), Axinya Tokareva (01368965)

November 19, 2017

1 M2 - Lo-Fi Prototyping

1.1 Proposed visualization solution

The user interface is oriented on programs like "Tableau" or "Glue" because we think it is the most easiest and most intuitive way to work with data. Figure 1 shows a very rough prototype of it with just two main parts:
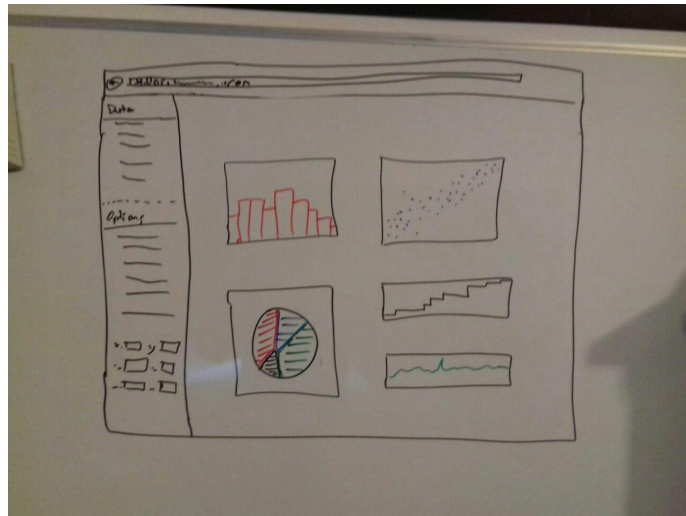
- Information view

- Plot view

Figure 1: User interface

1.1.1 Information view

At the moment the information view, figure 2, is divided in two parts, "Data" and "Options". Data provides some information about the dataset itself like name of the table, number of columns, rows and entries or column names. Options should give the possibility to add plots, interactions and other things to the plot view.

Figure 2: Information view

1.1.2 Plot view

The plot view is the area where, like the name tells us, all the plots appear and the interaction happens.

1.1.3 Graph proposals

Since we did not get a real specification of the customer what he would like to get visualize and just told us to try out whatever we want, we came up with a few ideas which might be interesting for astronomers. Unfortunately there are just 4 things, distance to sun, color, position and amount of stars which can be plotted in a meaningful manner, which made it really hard to find good plotting examples. At least we got six ideas so far and hope that the process of working with the data more intense we get new ideas for new plots.

- Scatterplot which shows the number of stars compared to the distance of the sun. (figure 3)

  Advantage is to get a good overview of how the stars are distributed in the area around the sun. An disadvantage will be the confusion if there

2

are to many stars and therefore no chance to find any patterns or other interesting things.
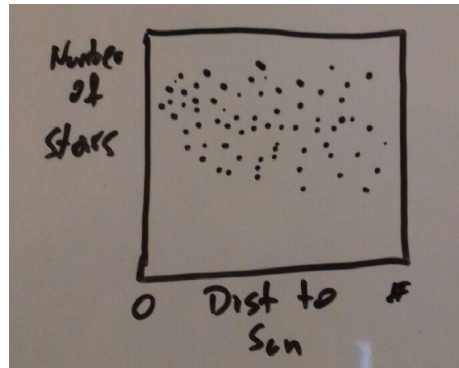

Figure 3: Scatterplot

- 3D representation of star clusters around the sun. (figure 4)

  Since the universe is a three dimensional space it is easier to see where specific star clusters are located but we are not sure if it is possible to create such a visualization with D3.
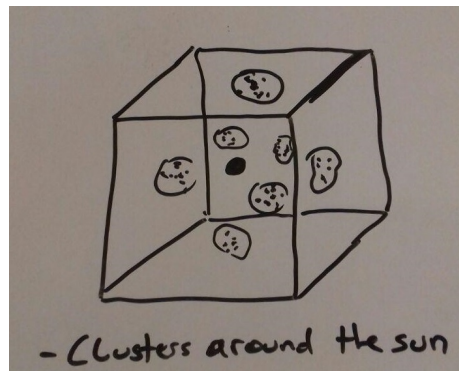

Figure 4: 3D visualization

- The ratio of hot and cold stars. (figure 5)

  It is very easy to understand but it can give a wrong picture of the data since not all stars has the state of their temperature.
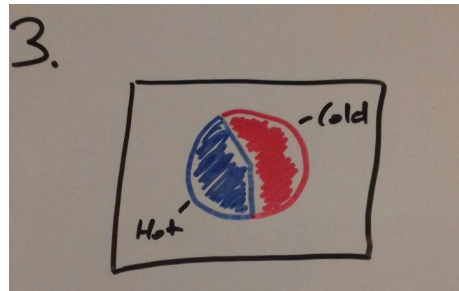
Figure 5: Pie chart

- Bar chart showing the size of specific star clusters. (figure 6)

  Like the pie it is very self explaining but if for example the scale of the y-axis is chosen wrong at can lead to false interpretation.
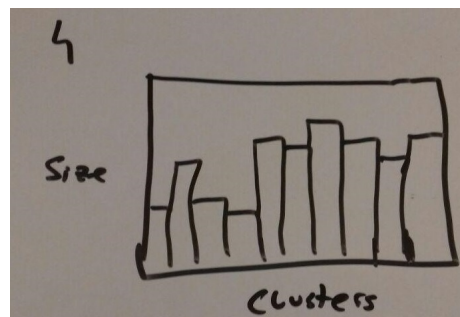

Figure 6: Bar diagram

- Line graph showing how the velocity behaves to the distance to the sun. (figure 7)

  An advantage of this view is that it shows in a good way of how the velocity changes with the distance to the sun but like with the bar chart choosing a good scale for the axis is important.
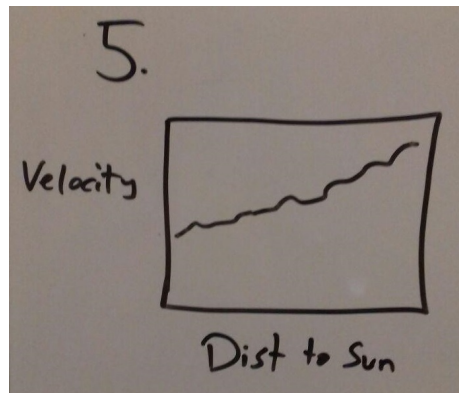
Figure 7: Line graph

- Shows the movement of a star in a certain time. (figure 8)

  Good could be to see if a star is moving and how much it is moving in terms of time. Disadvantage is that it is hard to compare many stars and how they are moving, because it just will show the motion of one star.

Figure 8: Dot plot

- Shows if there is a correlation in the parallax and the proper motion of the stars. (figure 9)

  The scatter plot could show the correlation if there is one but it is also possible that the amount of data makes it impossible to find out if a correlation exists.
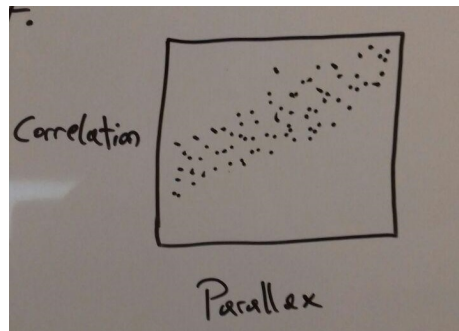
Figure 9: Scatter plot

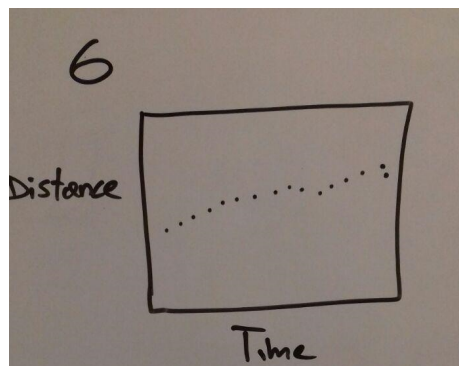- The Boxplot should show the minimum, maximum, average and median of errors of a specific star cluster. (figure 10)

  It can give a good overview of how the errors of measurement are distributed inside the star cluster. A disadvantage is that the data are potentially not meaningful because maybe some stars has no error measures and so the plot is corrupted.

Figure 10: Boxplot

- To show how the measured error behaves in comparison to the velocity a line graph will be used. (figure 11)

  Advantage is it shows if the velocity has an influence on the error of the stars. Possible disadvantage is that the plot is not meaningful.
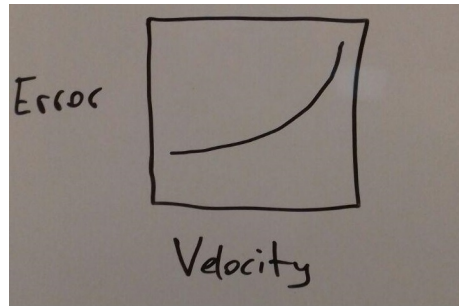
Figure 11: Linegraph

- The plot should compare the different types of weights, AC and AL (when writing this, we have not received an answer of the experts yet what it is exactly, but we thought it could be interesting), inside a star cluster. (figure 12)
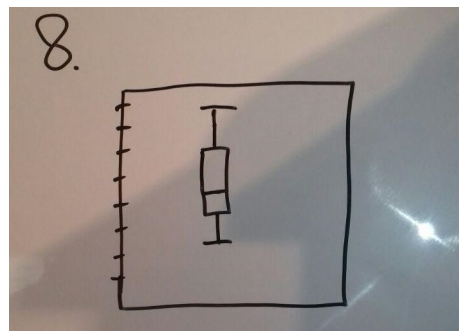
  Can give a good comparison of the two different weight types, but since we don't know what it exactly is yet we can not really say if it is a good and useful representation.

Figure 12: Bargraph

- Line graph showing the velocity of stars in a specific cluster. (figure 13) This view gives us the opportunity to compare the velocities of the stars in a specific cluster to make an assumption about the general movement of objects. It is important to choose a good scale for the axis, as in the previous illustration.

Figure 13: Linegraph

1.1.4 Interactions

1. **Dashboard 1**
   One dashboard could be to combine figure 4, 5, 6 and 12. While figure 4 shows the star clusters around the sun in a three dimensional context, figure 5 can give us an overview of the ratio of hot and cold stars in the data. To see how many stars are inside of a cl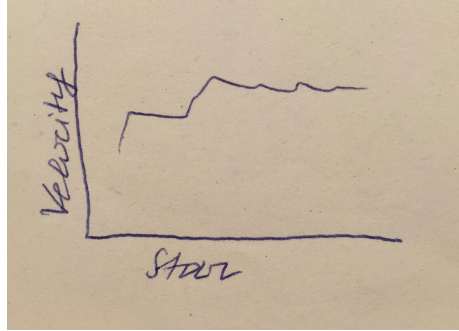uster, figure 5 can provide the information with a bar chart. Figure 12 contains the two types of weight of each cluster and compares the cumulated weights. The interactions could be to click on a specific cluster in the three dimensional representation and figure 5 and 6 updates their values corresponding to the selected cluster, while figure 12 highlights the bars of the selected cluster. Another interaction will be when you just want to see all the hot stars, you can select them in figure 5 and all other plots updates their view to show only the hot ones.

   Advantage
   With this dashboard it is very easy for the user to recognize patterns (e.g. amount of cold/warm stars in cluster, size of cluster, overall weight of stars in cluster) inside of to the clusters. The user can interact with this dashboard to gain more knowledge for each cluster in an easy and intuitive way.

   Disadvantage
   A big disadvantage of this dashboard is the generalization of the stars into clusters. The generalization leads to just see an overall information of the amount of stars and will not provide any information of a specific star.

2. **Dashboard 2**
   The second idea is to combine figure 4, 7 and 10. As mentioned before, figure 4 shows the clusters around the sun in a 3D scatterplot. Figure 7 plots a line graph sorted by the distance to sun for every star and compares it to the velocity of a star. The last figure (figure 10) plots a box

plot of the astronomic excess noise significance of all stars. If a cluster of figure 4 is selected, figure 7 and 10 will plot the line graph and the box plot just according to the stars contained in the chosen cluster.

Advantage

See if and how the distance to sun changes the velocity of the stars or star clusters and how big the astronomic excess noise significance is. It can be used to see if the error grows with distance and velocity.

Disadvantage

It will be very time consuming to process and sort the data every time if an interaction happens. Furthermore many information can be lost in the single views because not all single stars has all needed data and therefore the box plot is not completely true for example.

3. **Dashboard 3**

2, 6, 7, 9 The last dashboard shows figure 4, 8, 9, and 11. Figure 4 plots all stars in a 3D scatterplot in which the user can zoom in and out to get a better sight of the stars. Figure 8 plots the movement of a star in a certain time compared to the distance of this star. For this plot there will be used a 2D scatterplot with brushing and linking to have a clearer view of the data. Figure 9 shows also a 2D scatterplot with brushing and linking with the distance to the sun of each star compared to the correlation of the distance to the sun and movement of a star. The last figure plots the error of velocity compared to velocity of each star as a simple line graph. One interaction will be the tooltip technique in figure 4. If a star is clicked in the 3D scatterplot, there will popup a tooltip with useful informations about this star. Furthermore the movement of the star will be plotted in figure 8. Another way to select a star for figure 8 will be to click a star in figure 7, which will also lead to zoom in to the chosen star in figure 4.

Advantage

The big advantage of this dashboard is, that the user is able to gain information and interesting patterns about specific stars and not only a whole cluster.

Disadvantage

The disadvantage can be that the user will be overwhelmed of the amount of data and is not able to extract useful information of for example figure 9.

4. **Dashboard 4**

This dashboard is similar to the first panel (the same disadvantages and advantages), except that it does not contain information about hot and cold stars. It also contains error information of a specific cluster (figure

10), and a linear graph representing the stars velocities in the selected cluster (figure 13). A more detailed description can be found in the scenario below.

1.1.5 VIS Techniques

- **Zoom**
  Because of the big amount of data, the zoom technique is very important for the stars around the sun and clusters around the sun 3D scatterplot. With the billions of stars in a dataset the plot has so many dots (stars) in it, that the user wouldn't be able to extract relevant information of the plot. The user should be able to zoom in and out of the Scatterplot and turn the plot around to get another sights of the distribution and information of stars and clusters.

- **Tooltip**
  Combined with the zoom function, the tooltip technique shows special information of a dot (e.g. name of the star, weight, distance to sun,...) if a star/cluster is clicked. This can be very useful if a dot attracts the attention of the user because of a strange behaviour (e.g. bunch of stars in a close area).

- **Brushing and Linking**
  The brushing and linking will be very helpful for the user to set limits to the min and max value of a scatterplot. For the 2D scatterplots there we have the same problem as with the 3D scatterplot. Because of the amount of data the user will be overwhelmed of the information and won't be able to extract useful information of the plot. With the brushing and linking technique the user can decide on his own what minimum and maximum is interesting to see for him and so can also just have a look on small parts of the data.

- **Filter**
  For all of the plots we will need filters to filter useless data which will have no values, null values or other values we could not process.

- **Dimension Stacking**
  The dimension stacking is used for the mean astronomic weight of the source in AL and AC direction compared to each cluster. The user has a fast overview of both weight variables for each cluster, which is triggers an interaction with other plots if clicked. This can help to determine eyecatching information of special clusters very fast.

1.2 The scenario of use

1.2.1 Fictitious user

Joao Alves is a professor for astrophysics at the university of Vienna. He is performing research in different topics of astronomy.

These are his research areas:

- The structure of the Interstellar Medium

- The origins of stars and planets

- Near-infrared instrumentation

- Life in the universe

1.2.2 Possibly Tasks

First thing we did, was meeting with Professor Alves. We explained our project to him and asked him to give us some ideas and information about tools for completing our task. He stated that astronomy researchers only engage in looking at parts of the universe most of the time, but what is interesting for Alves is the "big picture", so he said. He gave us the task to get to know the data and to present parts of it in a visualization that makes it easier to see coherences. The task was not so specific, but it is important that we look through the Gaia catalog and find interesting patterns and then visualize it in different plots. For him it is probably very helpful, as one of his research areas is "The structure of the Interstellar Medium". So, he can reference our graphics while working with the Catalogs from Gaia and quickly sees how a particular area he is looking at reacts with in a bigger range.

Based on the available data and information retrieval, we can assume that the following problems can be solved, or at least can be a little improved to their solution:

- The most important conceivably scientific task is to clarify the appearance and development of our Galaxy with the help of a survey of stars. Collected Gaia data allow astronomers to better understand how the stars arise and how they saturate the space around them when they die. Previously unattainable accuracy of parallax measurements, as well as its own and radial velocity for one billion stars will give astronomers a clearer picture of the development and structure of the Milky Way.

- The second task can be solved of the telescope is the discovery of exoplanets, asteroids and meteorites (for example with clustering).

- Why in the observable universe there is much more matter than antimatter?

- The possibility of the existence of supermassive black holes in the centers of galaxies can be checked by studying the motion of stars and interstellar matter around them.

- To make a detailed map of the distribution of stars in our Galaxy.

1.2.3   Specific problem and scenario

For the scenario were chosen the latest model (dashboard) and the problem of detecting black holes. As already described above: by observing the stars, it is possible to make assumptions about their existence. Of course, only these data are not enough for approval (we also need data from radiation and observations in the optical range, for example), but it is sufficient for the initial assumption and for further, more detailed study. In this problem, cluster analysis of stars can help us, which can tell us about quasars, globular or open clusters of stars or even new objects. It may also be useful to study the general motion of stars of a certain cluster. A potentially existing black hole in the center of the cluster acts as a cosmic "spoon" that mixes the stars, resulting in these stars moving at higher speeds and longer distances. Observing the behavior of stars gives a hint of the existence or absence of a black hole in the center of the cluster. So, the main method of searching for black holes at the present time is to study clusters (their density, for example), brightness distribution and velocity of stars.
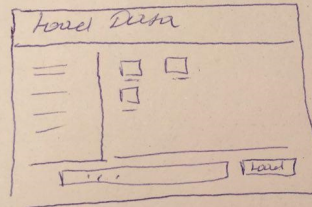
- Step-by-step:

  1. User creates a new file.
  2. Selects data for analysis.
  3. Specifies the name and required configurations
  4. He studies the first and third illustrations to identify a huge cluster of stars or new objects in the first plot as a secondary task.
  5. The user is a bit confused, because he uses the program for the first time, and does not know how to get a more detailed view of a particular cluster. He clicks on the help button and gets the information he needs.
  6. Clicks on the largest cluster for a more detailed study. (By clicking, the cluster is highlighted in the first illustration, the bars of the diagrams on the third and fifth graphs in accordance with the cluster, the remaining diagrams are updated).
  7. Explores the second graph to see if the data is significant.
  8. If not, then goes to a smaller cluster, or downloads / rechecks the data, changes the configurations.
  9. Explores the fourth and fifth graph.
  10. Identifies the general movement of the cluster, high velocities and masses, and hence high luminosities (It was found that the more weight, the higher the luminosity of the star).
  11. Makes an assumption about the existence of a black hole.
  12. Data is saved by clicking "Save" or "Save as" button.
  13. Clicks on the button "About us" to read about us.

14. The user proceeds to other methods of research to confirm or refute the hypothesis.
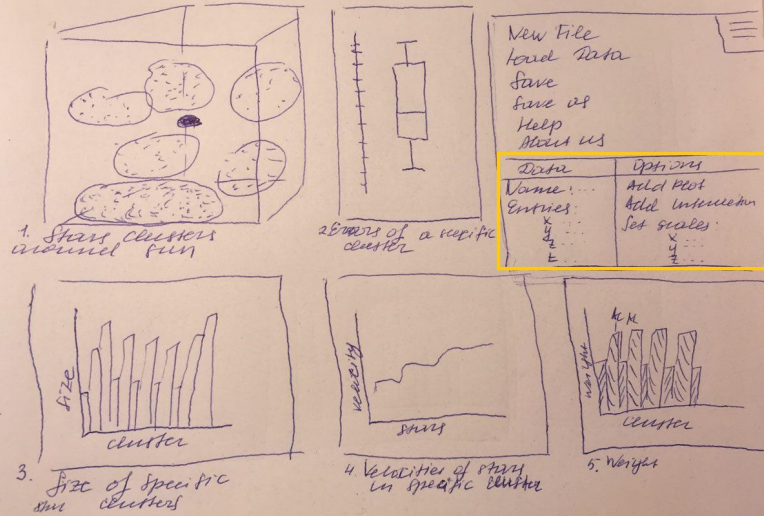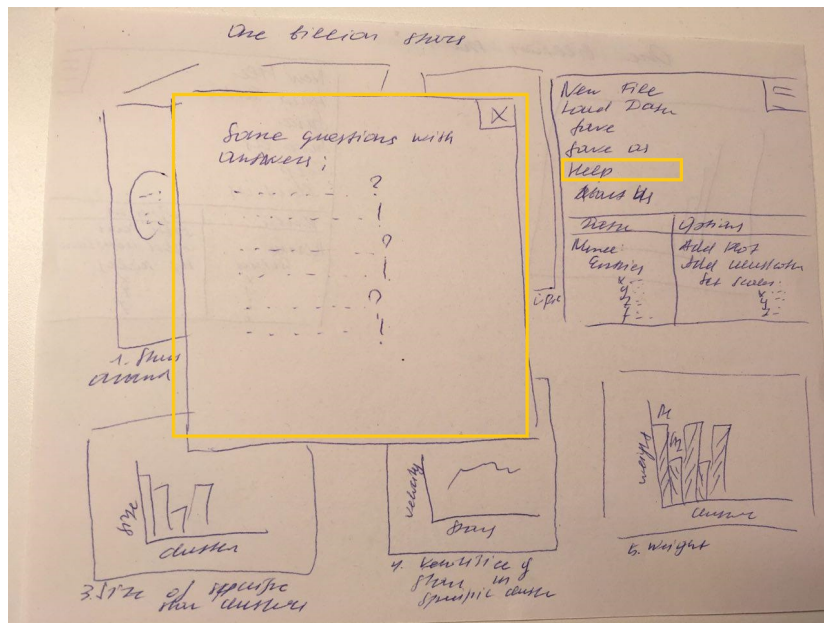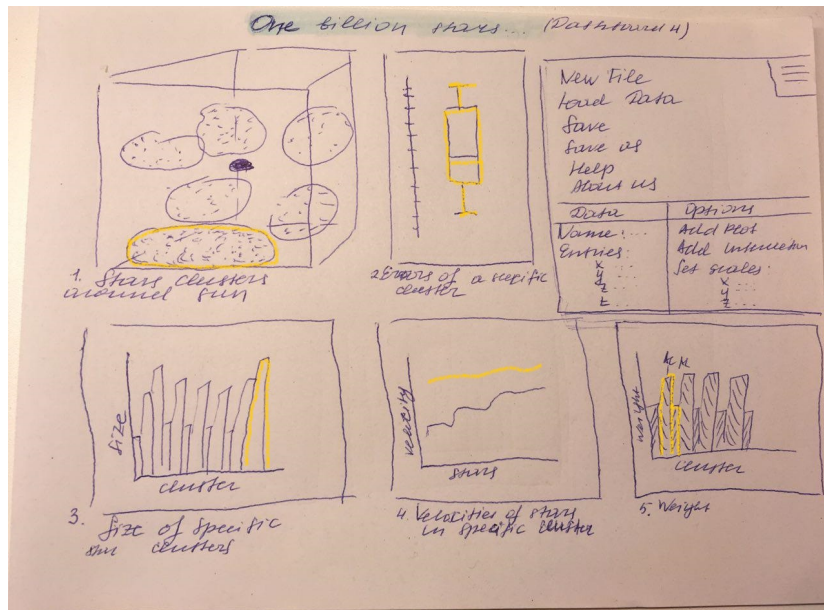
- Sketches
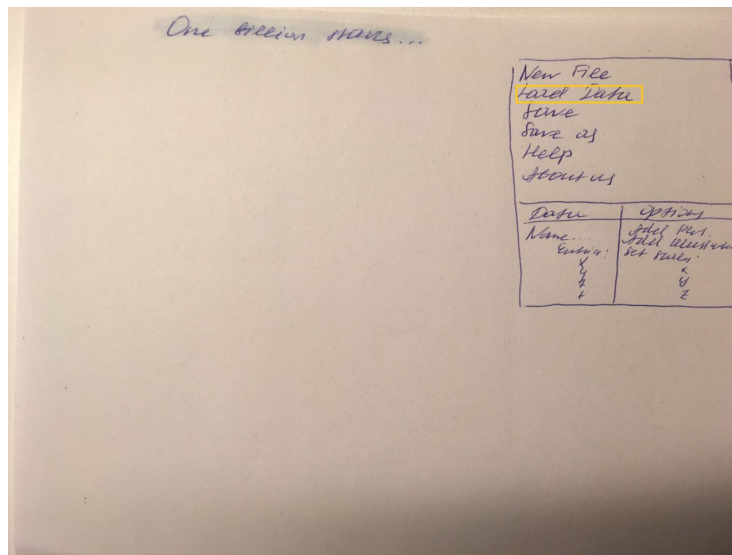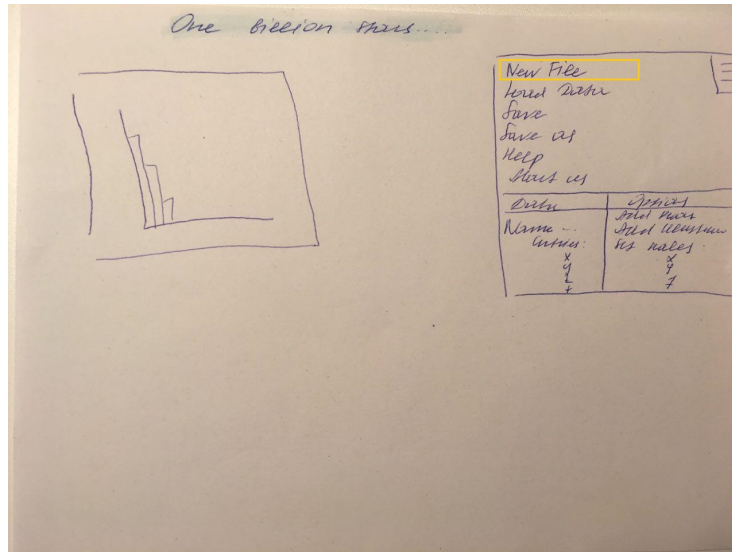
One billion stars... (Dashboard H)

1. Stars clusters around sun
2. Errors of a specific cluster

New File
Load Data
Save
Save as
Help
About us

Data	Options
Name:	Add Plot
Entries:	Add Interaction
X	Set scales
Y	
Z	

3. Size of specific star clusters
4. Velocities of stars in specific cluster
5. Weight

One billion stars

Some questions with answers:

New File
Load Data
Save
Save as
Help
About Us

Data	Options
Name:	Add Plot
Entries:	Add Interaction
X	Set scales
Y	
Z	

1. Stars around
3. Size of specific star clusters
4. Velocities of stars in specific cluster
5. Weight
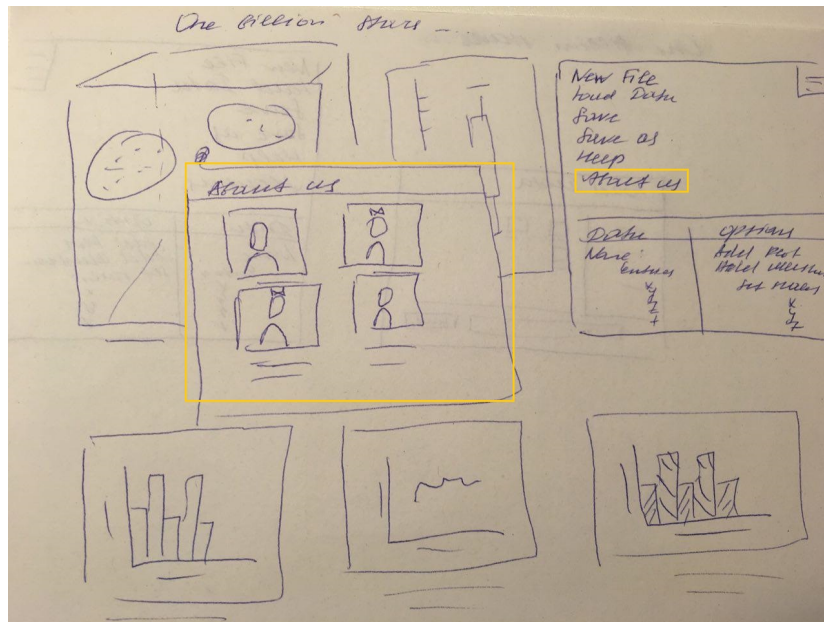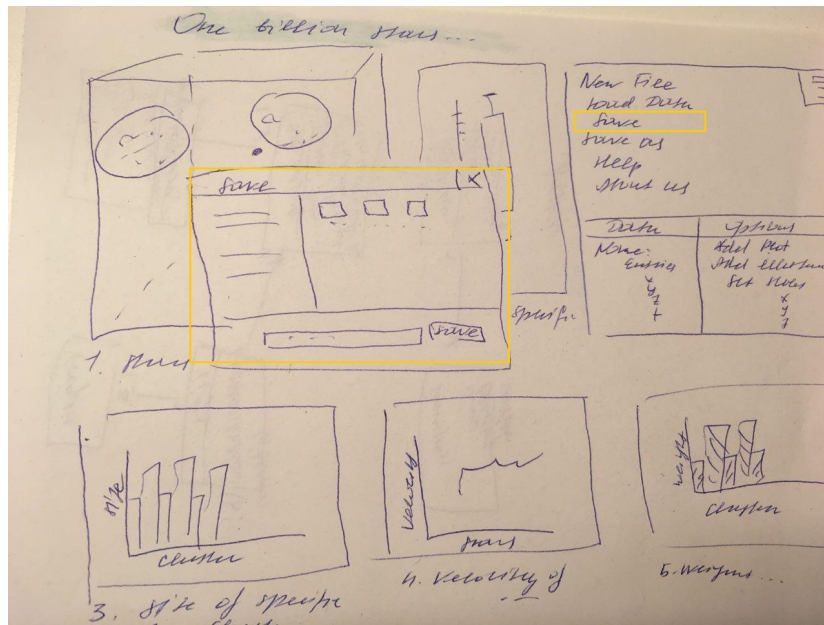
1.3 Implementation details

For protyping we will use Tableau, because we think it is a great tool for creating some quick views and planning how the end project should look like. We also tried softwares recommended by Joao Alves and his PhD-Student. Topcat for example helps us to go through the data and filter out the important parts. Glue is another program that helps us to gain insight for the data we are dealing with. For the end project, we are will use d3.

1.4 Milestones

- 19th-26th of November: Evaluation study, ask a few colleagues and friends, show them the Lo-Fi prototypes, every group member should talk to at least two people. Collect results and discuss pros and cons together in the group, doing a first implementation prototype in Tableau.

- 27th of November-10th of December (M3 due): Implementing the prototype in d3.

- 11th-14th of December: Preparing M3 presentation, meeting and going through slides together, discussing who sais what.

- 14th-22nd of December: Discussing, how the project went and pros and cons of implementation, meeting with Joao Alves, getting feedback.

- 22th of December-7th of January: Apply changes in implementation according to Alves' feedback

- 8th-21st (M4 due): Meeting Joao Alves again and proposing our end solution, writing report.

At this point, we cannot tell, who will do what for the project because there are many big tasks which we want to distribute fairly among us four. There are not really tasks that can be completed by one person only, so we will try to communicate and work together on all of them as good as possible.

1.5 The work distribution

Expand on your proposed visualization solution. **Benjamin + Alexander + Axinya (only figure 13 and dashboard 4)**

- Provide detailed illustrations of what the interface will look like and how the user will interact with it (e.g. a storyboard). You may either draw these by hand or use a mockup tool of your choice.

- To help with coming up with novel designs, in order to get full marks, you need to provide at least 10 different possible single graphs that show different aspects of the data. In addition, you need to provide at least three different ways of combining different views into an interactive application (dashboard) and demonstrate what type of interactions are possible.

- In order to ensure you are exploring the visualization design space, we require you to, reason and design for at least 5 different vis techniques. These methods can be in terms of vis encodings: aggregation over all variables except for one, aggregation in general, tree-maps, heat-maps, hierarchical exploration, correlation of two variables, correlation over multiple var, filter, scented widgets, etc.; for interactions: tooltip, zoom, table lens, focus-and-context (fisheye or similar), small multiples, brushing linking, etc. You should reason about the advantages and disadvantages of each of the different views, dashboards, and interactions.

Present a scenario of use, using sketches and text to demonstrate how the user accomplishes a specific task with your tool.

- The first step here is to make your tasks much more detailed and specific. Create a fictitious user and describe a specific problem they have. **Axinya + Nicole**

- A scenario then spells out what a user would have to do and what he or she would see step-by-step in performing a task using a given system. The key distinction between a scenario and a task is that a scenario is design-specific, in that it shows how a task would be performed if you adopt a particular design, while the task itself is design-independent: it's something the user wants to do regardless of what design is chosen. **Axinya**

Implementation details. **Nicole**
Milestones that break down the work into smaller parts. **Nicole**
Website. **Axinya + Nicole**

1.6   References

1. Our website: http://wwwlab.cs.univie.ac.at/∼a1368965/vis17/

2. https://medienportal.univie.ac.at/uniview/professuren/cv/artikel/univ-prof-dr-joao-alves/

3. vis lectures

4. http://gaia.ari.uni-heidelberg.de/tap/tables

5. http://sci.esa.int/gaia/

6. http://sci.esa.int/gaia/58275-data-release-1/

7. https://en.wikipedia.org/wiki/

8. http://www.mattboldt.com/demos/typed-js/