SHORT TERM SCIENTIFIC MISSION (STSM)
SCIENTIFIC REPORT

This report is submitted for approval by the STSM applicant to the STSM coordinator

Action number: BM1306 - 38709
STSM title: "Visualisation of big data in tinnitus"
STSM start and end date: 04/10/2017 to 16/10/2017
Grantee name: Christian Schneider

PURPOSE OF THE STSM

The main purpose of the STSM was to test and evaluate different visualisation strategies for large data sets in tinnitus research, with the aim to reveal patterns and underlying structures by means of visual analytics as opposed to traditional statistical methods. The secondary purpose was to implement a visualisation (prototype) for https://www.tinnitus-database.de using contemporary web technologies. The STSM addresses a contemporary fact: Enormous quantities of data go unused or underused, simply because the amount of data is overwhelming or it is difficult to find out where to start and what to do with data. The purpose of the STSM is to apply techniques from data visualisation to clinical data in tinnitus research. The amount of data continuously produced by science is enormous and therefore the data provided by the tinnitus data base is in its nature big data. State of the art visualisation techniques coupled with modern hardware applied to tinnitus databases could therefore add considerable value to tinnitus research.

DESCRIPTION OF WORK CARRIED OUT DURING THE STSMS

Kick off meeting with Dr. Winfried Schlee, Neuropsychologist, Dr. Patrick Neff, Neuropsychologist and Andreas Rein, Developer at https://www.tinnitus-database.de in order to decide what data to use during the STSM and setting the visualisation strategies to be applied to the data. The applicant talks about different visualisation strategies, shows prior work in order to create a common ground of understanding. Expectations for the outcome are defined.

Andreas Rein makes an introduction to the database and sets up required access rights for export and exploration of data. There were issues with the export which were resolved during the mission.

Data

The selected data sets consist of patient data from the University Clinic Regensburg (export from the Tinnitus Database, https://www.tinnitus-database.de) , Survey by the Tinnitus Talk Forum (tinnitustalk.com), and the TrackYourTinnitus Database (trackyourtinnitus.org)

Some data sets needed extensive filtering of invalid/missing data and restructuring before they could be visualised. Preliminary simple visualisations were programmed for each data set, allowing stakeholders to discuss the data, extract areas of interest, discuss ideas and choose an appropriate final visualisation.

13 visualisation experiments were undertaken during the mission, five finalised visualisations were produced and one visualisation will be permanently integrated into https://www.tinnitus-database.de . Three exemplary visualisations are presented in the upcoming main results section.

DESCRIPTION OF THE MAIN RESULTS OBTAINED

Visualisation: “Beeswarm 2D”

Data: Patient data from the University Clinic Regensburg (export from the Tinnitus Database, https://www.tinnitus-database.de): Patient ID, age, sex, THI score, TQ score, age of onset, tinnitus duration, tinnitus loudness and tinnitus pitch.

Data aggregations were avoided as much as possible in order to let intrinsic data patterns arise, if present. Each patient is displayed as a two dimensional point with variable area. A force directed layout including collision detection is used for placing the points on a 2D plane. An interactive tooltip provides detailed information about a patient. A menu allows to layout data according to different properties. Patients can be highlighted in order to observe their position in different configurations. K-means clustering is built into the visualisation, illustrating how clustering mechanisms could enhance the process of visual analysis.

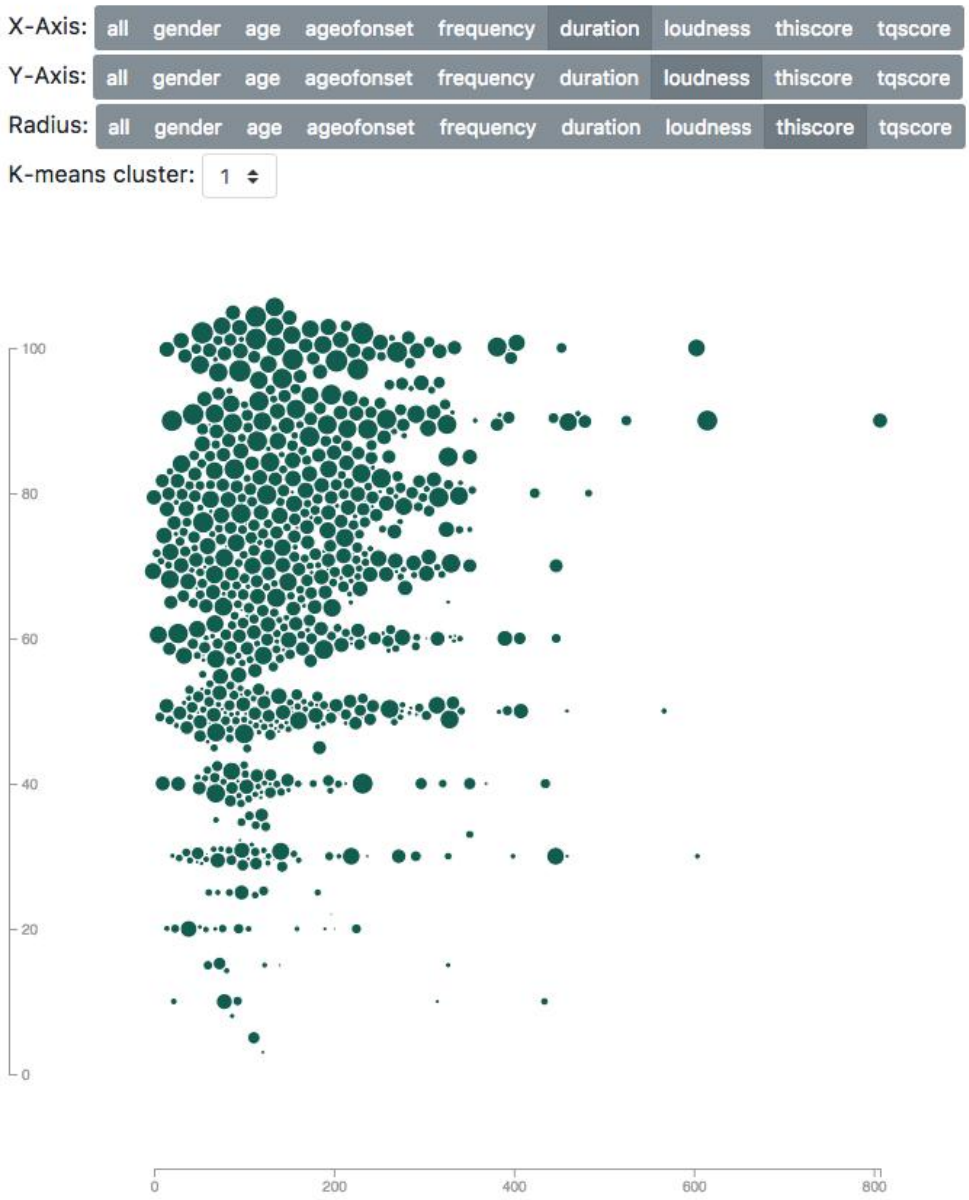

Figure 1: Beeswarm 2D visualisation of patient data of Regensburg (https://www.tinnitus-database.de)

Visualisation: Treatment networks

Data: Survey by the Tinnitus Talk Forum (https://www.tinnitustalk.com). The used data mainly consists of attempted methods to cure tinnitus.

Considerable mining and refinement was necessary to create the data set used for this visualisation, in which connections between treatments are shown. More precisely, it aims to show the likelihood of a person to apply a certain treatment when he/she was undergoing another treatment. So connections between different treatment techniques are shown. The added scientific value here is therefore the interactive nature which allows displaying connections of interest (figure 3) in contrast to the visualisation of all connections (figure 2), where it is hard to visually discern connections of interest.
For example, it becomes evident that data seems to cluster along connections with typical clinical treatment paths (e.g. psychologist/psychiatrist with medications) vs. more self-help oriented treatment paths (e.g. herbal therapy, acupuncture, self sound stimulation). On the other hand, patterns arise where these different therapy domains share certain treatments (e.g. CBT with self sound stimulation).

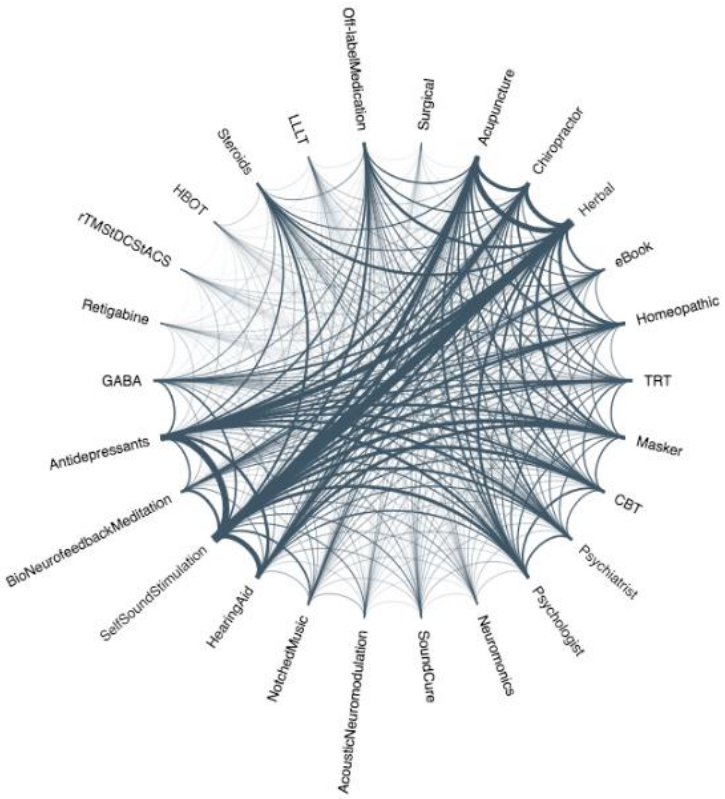

Figure 2: Treatment networks visualisation of survey data of www.tinnitustalk.com. All connections.

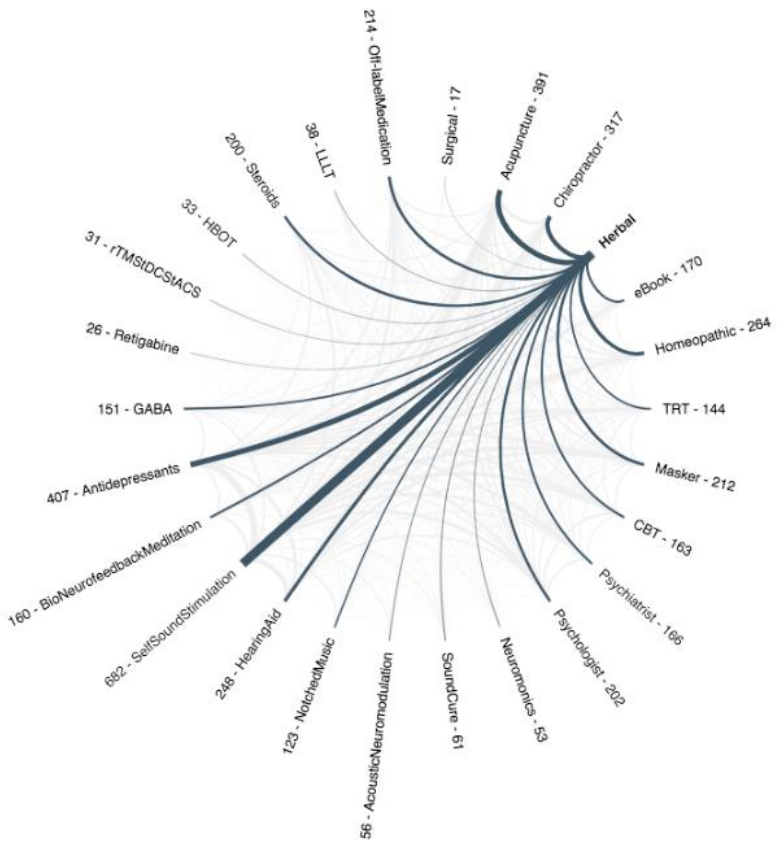

Figure 3: Treatment networks visualisation of survey data of www.tinnitustalk.com. Connection of interest (here: herbal therapies).

Visualisation: TrackYourTinnitus raw data visualisation

Data: TrackYourTinnitus Database (trackyourtinnitus.org)

“TrackYourTinnitus” data (TYT) consists of a data set with high dimensionality (NxM = 40’000x64). No aggregations were performed in order to visualise the data in its most raw form. This allows first of all for detecting outliers, invalid entries and anomalies at a first glance. Second, its allows for detecting interesting aspects for further exploration. For most of the visualisations, the data was normalised and categorical data was vectorised in order to allow direct comparison between all dimensions.

Each row is displayed as a multidimensional point projected onto 2D. Dimensionality reduction techniques such as PCA were applied in the data exploration process.

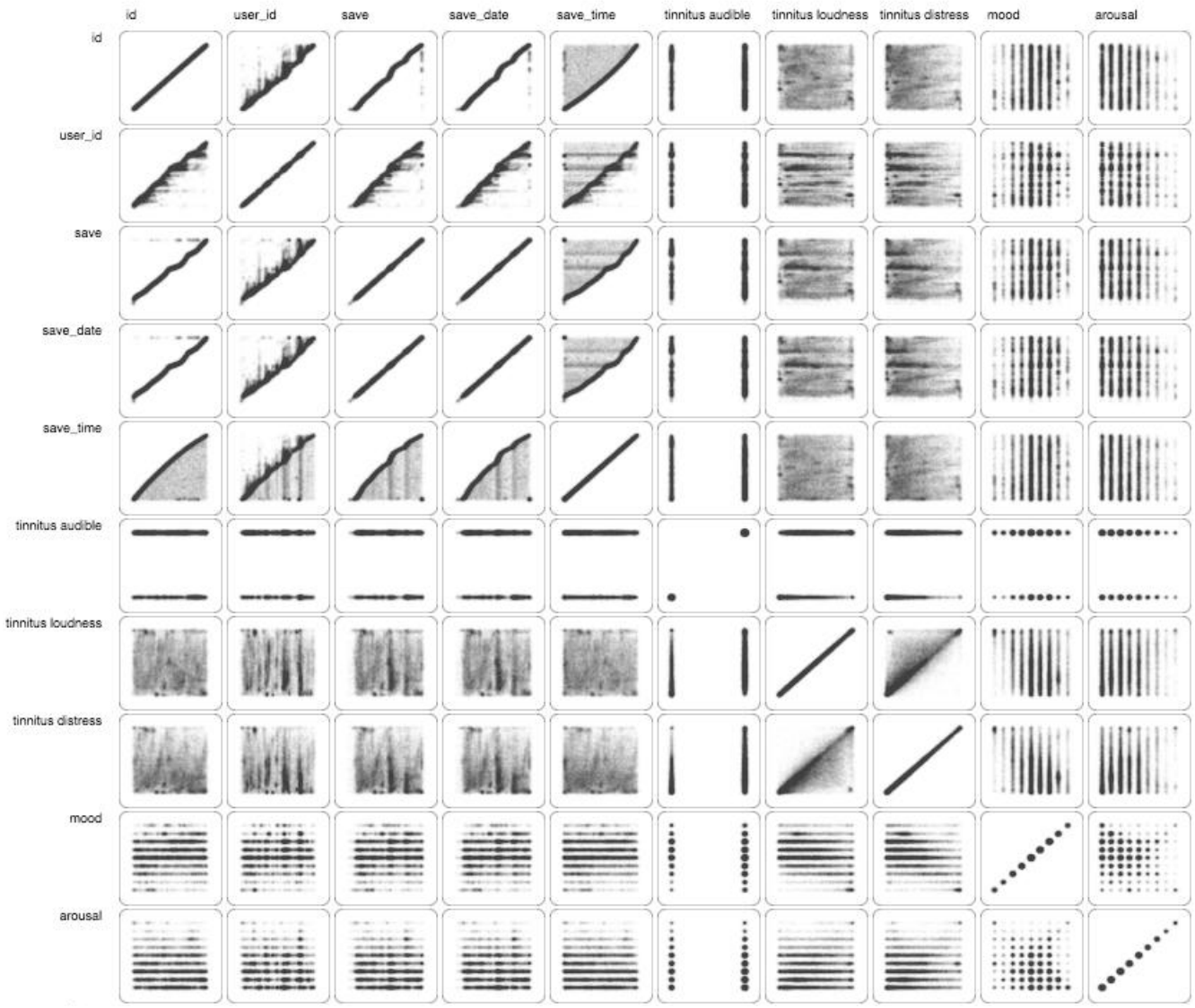

Figure 4: All combinations of the 64 variables in the TrackYourTinnitus Data (a subset of 10x10 variables is shown in the image)

General Discussion

Besides already known patterns, the evaluation by the host and the affiliated neuropsychologists indeed produced novel insights or inspired research questions. On the other hand the amount of insights and inspiration was heavily dependent on a) how much variables and data was available and b) (naturally) how much these datasets have been studied beforehand. Therefore the feasibility of these visualisation approaches to probe novel big data sets with a large amount of variables can be considered as the main outcome of the work performed during the mission.

Second to that and of daily routine clinical and scientific relevance is the useful implementation of these visualisations in the online databases. A clinician could therefore quickly get an impression of certain distributions and relations of tinnitus data without performing statistics. Furthermore, any scientist planning a study or checking on his ideas could also profit from these visually appealing and exhaustive interactive displays.

Much potential can be seen in the visualisation of the TYT Data, which can be considered coming close to real “big data”. The high dimensionality of the data set makes it further interesting. The first visualisation experiments show potentials of revealing patterns in the data but also unearthing limitations to the app technology . Looking at the timeline plots of the major variables, one is even able to see certain accumulations of the app’s usage after e.g. press releases (visualisation not shown here).

The technologies used during the STSM (mainly web technologies) quickly reached their limits. More powerful technologies with easy support for hardware accelerations should be used in a next step (e.g., C++, OpenGL, GPGPU). All the participants agree that this is a really interesting data set to be explored further with the means of data visualisation.

FUTURE COLLABORATIONS

Given the productive and insightful mission, further collaboration is envisioned. To start this collaboration it is planned to further explore TYT data (Kickoff Dec 5, 2017). Also the Beeswarm 2D visualisation will be enhanced and refactored and built into the tinnitus database, beginning of 2018. As a final concrete output, it is planned to display interactive data visualisation at the upcoming TRI/TINNET conference next spring.