Interactive Visual Analysis of Heterogeneous Cohort Study Data

Paolo Angelelli, Steffen Oeltze, Judit Haász, Cagatay Turkay, Erlend Hodneland,
Arvid Lundervold, Astri J. Lundervold, Bernhard Preim and Helwig Hauser

Abstract— Cohort studies in medicine are conducted to enable the study of medical hypotheses in large samples. Often, a large amount of heterogeneous data is acquired from many subjects. The analysis is usually hypothesis-driven, i.e., a specific subset of such data is studied to confirm or reject specific hypotheses. In this paper, we demonstrate how we enable the interactive visual exploration and analysis of such data, helping with the generation of new hypotheses and contributing to the process of validating them. We propose a data-cube based model which handles partially overlapping data subsets during the interactive visualization. This model enables seamless integration of the heterogeneous data, as well as linking spatial and non-spatial views on these data. We implemented this model in an application prototype, and used it to analyze data acquired in the context of a cohort study on cognitive aging. We present case-study analyses of selected aspects of brain connectivity by using the prototype implementation of the presented model, to demonstrate its potential and flexibility. .

Index Terms—heterogeneous data, medical visualization, IVA

1 Introduction

Cohort studies in medicine become increasingly common, partly thanks to the availability and to the recent improvements in medical imaging technologies. Such studies are a type of observational study that follows one or more groups of people (samples), called cohorts, over time. They are used to evaluate medical hypotheses in samples sharing common characteristics, for example being healthy, or presenting specific risk factors, to gain a better understanding of the absolute risks of certain pathologies and of the pathology development. Cohort study data is often acquired over longer time periods, following strictly defined protocols, being therefore not trivial to set up. Because of that, they are often designed to deliver a larger variety of data than the focus of the initial study, which, later on, can be the basis for retrospective analyses, evaluating further sets of hypotheses.

There are means to evaluate specific hypotheses, based on such cohort study data, often involving accordingly designed data extraction, transformation, and fusion approaches. However, there is a lack of technology to support the flexible and open-ended exploration of such data, mostly because of its heterogeneity. This means collections of image and non-image (quantitative, often image-derived) data, which in turn can be categorical and numerical, and defined on domains that only partly overlap. Due to the complexities posed by the data heterogeneity, analysts often have to limit their attention to subsets of the data, making the analysis lose the overall relations within different modalities. Integrating all the available data within one visual analysis tool that allows to seamlessly combine them in an on demand fashion is expected to support the experts in the exploration of heterogeneous cohort study data and in the hypothesis generation and verification, and to accelerate their research workflow.

The exploration and analysis of heterogeneous cohort study data generates specific new challenges for visualization. The contribution of this article is therefore two-fold. First, in Section 2, we characterize these challenges, in relation to the substantial heterogeneity of the data, and in relation to the analysis tasks, goals, and typical analysis workflow in the specific context of a cohort study on cognitive aging. Second, in Section 4, we describe our solution, based on a new, general multi data-cube model to support heterogeneous data, and that can be also adapted to other situations of highly heterogeneous problems. Finally, in Section 5 we describe our prototype implementation of our model, that, in Section 6, we use to exemplify how our novel approach can enable the generation of new hypotheses, as well as the swift analysis of relations between otherwise unconnected data parts, thus improving the analysis and exploration process. In Section 6 we also provide an evaluation of our method by two domain experts from the medical and neuropsychological domain.

- *Paolo Angelelli and Helwig Hauser are with the department of Informatics at the University of Bergen. E-Mail: paolo.angelelli@uib.no .*
- *Cagatay Turkay is with giCentre at City University, London.*
- *Steffen Oeltze and Bernhard Preim are with the department of Informatics at the University of Magdeburg.*
- *Judit Haász, Erlend Hodneland and Arvid Lundervold are with the department of Biomedicine at the University of Bergen.*
- *Astri J. Lundervold is with the department of Biological and Medical Psychology at the University of Bergen.*

2 A Scenario of Heterogeneous Data in a Cohort Study

One major goal of this work is to create a solution to enable the explorative visualization and analysis of data that was acquired as part of a longitudinal study on cognitive aging. During this study, more than 100 healthy individuals (mean age 60.8 (7.8), 65% females at inclusion) were recruited through advertisements in local newspapers. At inclusion, all the subjects who responded were interviewed, to exclude those reporting previous or present neurological or psychiatric disorders, a history of substance abuse, or other significant medical conditions. The neuropsychological evaluation confirmed that the participants showed no symptoms indicating mild cognitive impairment (MCI) or dementia. Each participant was examined every three years, starting in year 2004/2005, and then in 2008. The participants were subjected to neuropsychological testing, genetic analysis (data not available for this work), and multimodal MR imaging. The result of each examination consisted of data on white matter fiber integrity, expressed by anisotropy measures computed from diffusion tensor imaging (DTI), cortical and subcortical gray matter measures, automatically calculated from structural MR images, and a number of neuropsychological tests, including the California Verbal Learning Test–Second Version (CVLT-II), the Color–Word Interference Test (CWIT), the Digit Symbol Substitution Task from WAIS-R, and the Mini Mental State Exam (MMSE). To summarize, each examination (per subject and year) consists of:

- white matter fiber bundles with anisotropy measures. Each individual fiber was divided into 100 segments of equal length for the derivation of associated measures.
- gray matter cortical and subcortical regions with quantitative measures for each region.
- scores from different neuropsychological tests.

For a detailed description of the study protocol and for previous selected analyses of this longitudinal study please refer to Ystad et

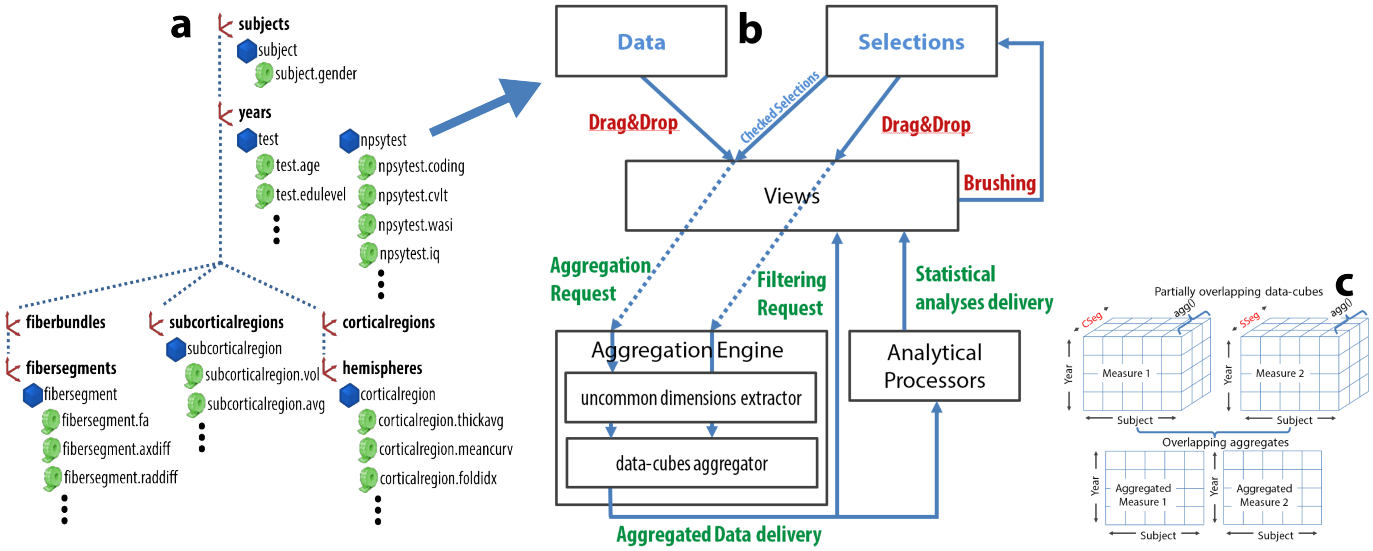

Fig. 1. **a**) Illustration of the dimensions (red), measures (green) and entities (blue) in the dataset of the cohort study on cognitive aging. The hierarchy in the figure is used only for presentation, as the presented model treats the dimensions independently. **b**) Simplified illustration of the proposed model. User interactions are colored in red, automatic operations, transparent to the user, are green, information sources are blue, and in black the components necessary to implement the model. Note that the selections require interaction to be used as filters, but are also automatically re-aggregated upon measure changes in views, or brush changes, and the result is automatically updated in the views. **c**) Illustration of the projection operation. The dimensions which are not common (in red) are processed using a statistical estimator (e.g., average). This operation can be steered by using a selection for each data-cube to filter the elements that are aggregated.

al. [14].

2.1 A heterogeneous dataset

Resulting from this study, a number of measures related to different aspects are available. One specific challenge with respect to the data exploration and analysis is that the measure's domains overlap only partially. Taking a scatterplot as an example, how should two heterogeneous measures be combined? In our case, these measures could be the *fractional white matter fiber anisotropy* (FA), that describes the degree of anisotropy of water diffusion along a fiber, defined for each segment of each fiber bundle, and the *thickness of the cortex*, available for each cortical region in both left and right brain hemisphere. This partial incompatibility of the data domains proved to be one if not *the* key challenge of this work. To overcome this challenge we developed the method presented in this article, able to seamlessy combine heterogeneous measures on the fly.

2.2 Abstract and physical data and their representation

In such studies certain measures, such as white matter FA or gray matter region volume, as well as others, are quantitative abstract measures that relate to physical (anatomical) entities. These, for the example, would be the white matter fibers or the gray matter regions. For these entities additional qualitative data is often also acquired, such as the bundles trajectories, or brain regions meshes or volumes. While analyses are often performed on the quantitative measures, it also becomes necessary to occasionally fetch and inspect the related anatomical data, to explain, for example, data outliers, or to see what effects certain conditions have on the anatomy. For these reasons domain experts would benefit from a system that can link different types of data, and bring up the appropriate sets on demand, e.g. in linked views.

In addition, when dealing with abstract views of measures related to physical entities, domain experts often need to relate groups of entities, such as selections, in abstract views to their physical location. To ease this process we propose to use a view with an illustrative physical model, or atlas, of the entities, which is linked to the other views. Through this atlas, the content of the selections is put in its physical context, to improve the understanding of such data. The definition of this model for the specific case described in this article, and its use, are described in Section 4.5.

3 Related Work

While the majority of visualization research –in particular also medical visualization– was (and still is) focused on the visualization of individual datasets, the visualization of data from population studies has not been a research topic until recently. One recent exception is the work of Bruckner et al. [1], presenting a system to retrieve and visualize anatomical brain data of *Drosophila*, covered in a large database of such flies' brains. This system enables a novel way to perform visual queries, combined with a volume rendering solution called Maximum Intensity Difference Accumulation (MIDA). Still in the biology domain, Jeanquartier and Holzinger presented a visual analytics approach for cell physiology to support the exploration and sensemaking process. [5]. Steenwijk et al. [10] also presented a novel visual analytics framework to query and visualize data from a cohort study , consisting of imaging and non-imaging data for each subject. Their approach was to preprocess and store the imaging and non-imaging data in a searchable relational database, to which a visual interface would perform dynamic queries. Still in the healthcare domain, Simonic et al. [9] presented a visualization system to improve prediction and treatment of patients based on longitudinal data.

More generally, few other visual analysis methods have been proposed for the analysis of higher-dimensional and heterogeneous data. One relevant related solution was presented by North et al. [7], who introduced visualization *schemas* to achieve the concurrent analysis of different sources of information in relational databases. Their system enables building coordinated visualizations in a similar fashion as when constructing relational data schemas. More recently, Weaver uses a method called cross-filtered views [13] to interactively drill down into multidimensional relations between multiple datasets. In his method, different variables are visualized in particular views and brushes in these multiple views are cross-filtered to discover complex relations in the data.

4 A Data-Cube Based Model To Enable Interactive Visual Analysis

The typical workflow approach to analyze the data coming from such studies is to manually extract the pieces of data to analyze from the dataset (e.g., using custom scripts or programs for each analysis), and

then process them using mathematical and statistical packages. Finally, plots of the results are generated either using custom scripts, or by importing the results into applications that can plot the data.

The first, and perhaps the biggest challenge in designing an interactive visualization system targeted at this problem is storing the data acquired with such studies in a way that allows fast and flexible access, retaining the meta-information expressing the relationships between the different pieces of data. Organizing the data in a relational database, similarly to Steenwijk et al. [10], is probably the first solution at hand, and possibly the easiest to design from scratch.

However, organizing data in a relational database is relatively inflexible: the database schema is bound to the specific structure of a particular study, together with the queries associated to it. Using a system designed in such a way to analyze a different dataset would require the redefinition of the database schema, as well as reprogramming the logic for data access. In addition, processing the queried data with mathematical or statistical methods that are not implemented in the database itself would require an additional application layer into which the data should be loaded, thus voiding the benefits of using a relational database. Finally, from a performance point of view, using a relational database to perform complex queries touching all the rows on a large amount of data becomes quickly a performance bottleneck in interactive operations, and this is even more problematic when item selection and measure filtering based on multiple attributes, requiring table joins, are used.

With *Polaris*, Stolte et al. [11] showed how visualization systems can also ground on data organized in a *n*-dimensional, possibly hierarchical, data-cube, which is also known as OLAP cube (for On-Line Analytical Processing) in the field of data warehousing. It has been reported that executing complex queries using OLAP cubes can perform about hundred times faster than doing the same on relational data [4]. A single, hierarchical, data-cube organization however, shows its limitations when the dataset, and its dimensionality, become heterogeneous.

4.1 Data-cubes: dimensions, entities and measures

In our model, data-cubes are constructed using categorical attributes as *dimensions*, while quantitative numerical values are stored as *measures* [11]. The dimensions and measures can be thought of as independent and dependent variables, and dimension coordinates are used to access the measures. Practically, after assigning an order to the dimensions of a cube, a data-cube can be implemented as an in-memory *n*-dimensional array. To make an example taken from the system presented in this paper, a measure for segments of white matter fiber bundles in our dataset, e.g., FA, is represented as a floating point n-dimensional array consisting of $n = 4$ dimensions: *subject*, *year*, *bundle*, and *segment*.

Compared to the model proposed for Polaris, we also introduce a third concept, called *entity*. An entity can be thought of as a row in a database table, and quantitative row fields would be the measures for that entity. In the example above, the measure *fibersegment.fa* (*fa* for fractional anisotropy) would be related to the entity *fibersegment*, being a measure of that entity. When, in our model, a data selection is defined, it also contains selection values for entities, which are then propagated to the measures related to it when it becomes necessary.

4.2 Multiple data-cubes and seamless dimension aggregation

A challenging feature of the data acquired in cohort studies is their heterogeneity. This means that mef white matter fiber segasures are collected for different entities, which do not share the same set of dimensions. In our specific case, when referring to entities, we can talk oments, grey matter subcortical regions and grey matter cortical regions, as well as neuropsychological tests. As shown in Figure 1a, the dimensions' sets of the measures are only partially overlapping, having all these entities in common only two dimensions, *subject* and *year*. The standard way to organize these data into a single data-cube would be to build a denormalized cube characterized by all the dimensions in the dataset, which would contain all the data. When the data is significantly heterogeneous, however, this strategy may lead to an explosion of the memory requirements caused by the denormalization.

In the model that we present here, the solution to this problem is twofold: on one side we store all the data in multiple, normalized data-cubes, to eliminate any kind of information redundancy and minimize the memory occupacy. Secondly, we propose runtime aggregation of the measures' data-cubes, when data which are held in data-cubes belonging to different entities have to be combined or cross-checked. Such aggregation operation is also referred to as the *projection* of a data-cube [11] (see Fig. 1c). Our model includes an engine to perform aggregation on-the-fly, for reducing the data-cubes' dimensions to their largest common subset, without having any embedded knowledge of the relations between measures. In contrast, this would be necessary when using a relational model for the data, as the system would need to incorporate knowledge about each specific database schema, together with logic for performing the operations.

In our model, when multiple measures are combined in a visualization (e.g., in a scatterplot, a parallel coordinate view, curve view, etc.), each measure is aggregated across those dimensions not belonging to the intersection. For the moment we can consider the mean as measure aggregator, but there are several other options, such as different statistic estimators which can be selected by the user.

In certain cases, it is also useful to change the level of detail. To allow this, we enable toggling which common dimension to keep during the aggregation. This is similar to a *roll-up* operation, with the difference that the dimensions' structure is treated as hierarchy-less.

Finally, even if some of the dimensions may embed a hierarchy, others are independent from each other. For example, it is easy to imagine that *subject* is independent from other dimensions, while *bundle* and *segment* are logically nested, as segments are part of a bundle. However, an imposed dimension hierarchy for all the dimensions would be useful to represent the data in a tree-like visualization, and let the user navigate the dataset (as shown in Fig. 1a). To compute such a hierarchy, we group entities recursively by the number of common dimensions, with each group reflecting dimensions occurring in the same number of entities. By letting the dimensions that occur in more entities floating higher in the tree hierarchy, and then proceeding recursively on subgroups, we can generate a complete hierarchy. Having defined such a hierarchy, it is possible to represent the measures in our cohort study data like in Fig. 1a.

4.3 Selections and selection-based filtering

In section 4.2 we explained how to create projections of a measure by aggregating it over entire dimensions. Obtaining an aggregate of a measure over a whole brain, however, may not always produce specific enough data to answer questions of interest. To enable a more focused analysis, selection techniques can be used in order to restrict the processed or visualized data to specific subsets under investigation. An example is the *Polaris specifications* [11], introduced for defining selections. Interactive visual analysis has introduced the related concept of brushing, a visual method to select items with certain characteristics (e.g., fitting certain ranges on specific measures), by defining a visual brush over a view on the data. These brushes normally contain a value for each data item, either binary or a percentage value, to express if or how much the data item is selected. Our model makes use of brushing to let the user define data selections. Using data-cubes, this brush should be transformed into a data-cube itself, where each item contains the tag information for the related entity. In our case, having several entities in the dataset generates an additional challenge: when tagging one entity, we must also propagate the selection to all those other entities in the dataset sharing at least one dimension with the tagged one. As a clarifying example, let us consider a selection of only those white matter fiber segments above a certain FA threshold. Such selection does not necessarily involve all the examinations, or even all the subjects. Let us say the user wants to cross only items in this selection with the cortical thickness. Then this selection has to be propagated to the entity *cortical region*, knowing that the shared dimensions between the entities *cortical region* and *fiber segment* are *subject* and *year*. This has to be done in an appropriate manner, so

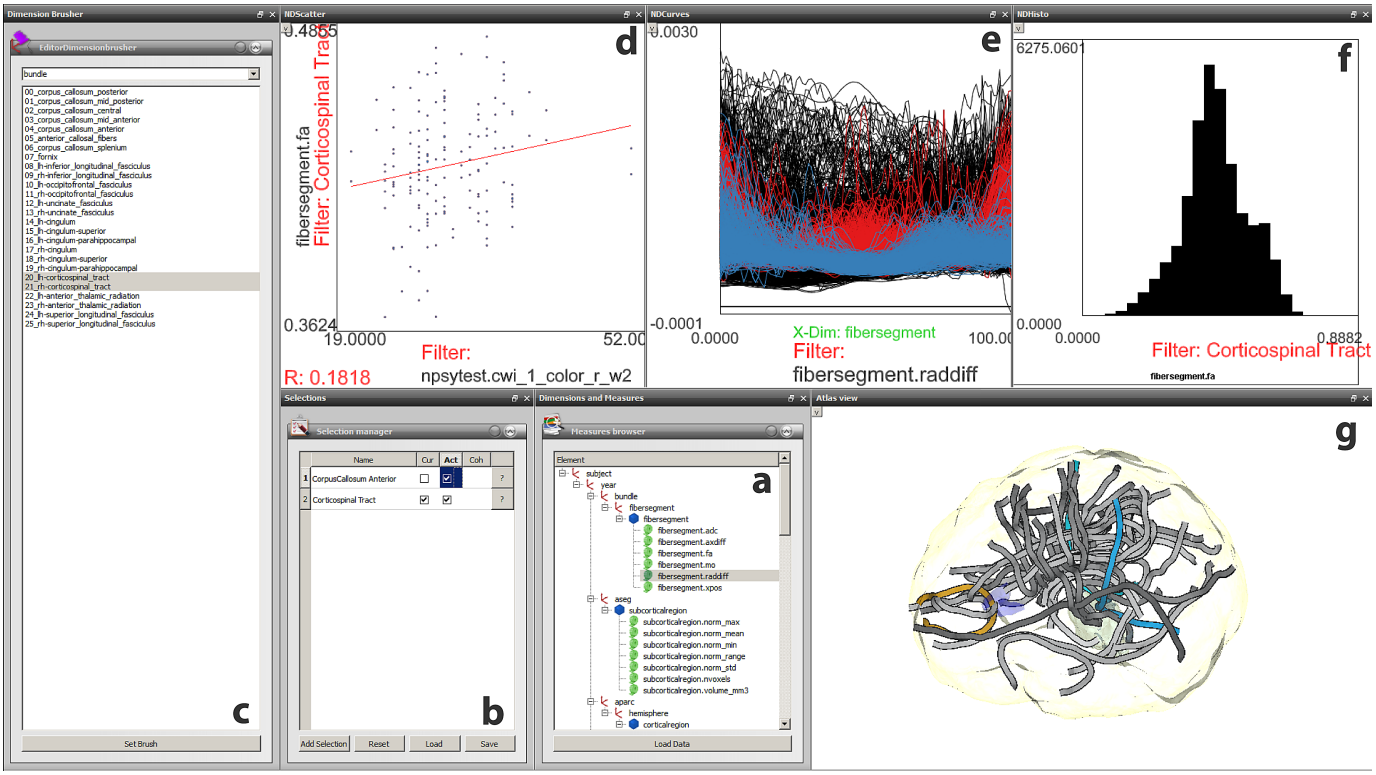

Fig. 2. Screen-shot of the prototype of the proposed model. The Measure Browser **(a)** lets the user drag desired measures into a view, and the Selection Manager **(b)** allows to add new selections, activate them, enable one of them for editing, and drag them into views, to be used as filters. The Dimension Brusher **(c)** enables slicing the data-cubes in the data collection, while the other views **(d,e,f)**, in this setup a scatterplot, a curve view and a histogram view, can be seen as projections of the data, and allow a more advanced definition of the selections, by means of brushing ranges of measures. In each view a drop down menu lets the user adjust the aggregation dimensions as well as the additional analyses to perform. Finally, the Atlas view **(g)** represents the selections in their anatomical context using a brain model. The two selections visualized contain, the first, both the fibers and the brain region of the Corpus Callosum anterior, and the second both the fibers of the corticospinal tract and the brainstem region (colors representing different bundles).

that only those (*subject*, *year*) pairs selected in one entity are selected also in the other one. In our model we propose a propagation scheme where a brush on one entity is propagated to all the other entities in the dataset that share dimensions with the brushed one. The propagation is done by first computing a projection of the brushed entity onto the common dimensions with all the other entities. Such projections of the brush are generated using the *max* operator, which produces, for each set of items being aggregated along one aggregation coordinate, the equivalent of a Boolean value indicating whether or not at least one item was selected. This scheme also allows multiple selections to be combined using Boolean logic, giving the user the necessary flexibility in building up expressive item selections.

Once a selection has been defined, it can be used in two manners. First, selections can be visually highlighted in the views, and thus compared with the whole dataset or with other selections. Second, since most of the views are built upon aggregated data-cubes, this aggregation can be steered, or *filtered*, using a selection (Fig. 2d). By setting a selection as aggregation filter, the aggregation is performed only using those items that are tagged in the selection. In this way, carefully selected information from the dataset can be cross-checked with other aspects, enabling the user to analyze virtually any aspect of the dataset.

4.4 Unrolling dimensions: a first step toward iterated visual analysis

Using a system implementing our model interactively is a flexible way to cross-analyze a wide variety of information in such heterogeneous datasets. In some cases, however, the analysis can benefit from automating certain steps, like repeating selected tests or analyses using a scheme defined by the user on different data, or with varying parameters or methods. This could be seen as extending a purely interactive visual analysis methaphor by using it as a analysis-setup tool for defining what type of actions to automate. The results of this extension could be thought as an *iterated visual analysis*. A clarifying example could be correlating age with subcortical region volume. The user could first define a selection, for example by filtering specific ages, or other parameters such as the IQ. This selection could then be used to filter the aggregation, which could conclude the interactive analysis step. Since it is also interesting to have details of how the volume of each specific subcortical region correlates with age, the user might want to combine his interactively specified selection with another one, selecting only a specific subcortical region, and repeat the process for every subcortical region. To ease this process, enabling at the same time to produce comparable results, we propose a method to automatically dissect and process the measures present in a specific view, by iteratively *slabbing* each measure's data-cube along those dimension that are specific to the data-cube (e.g., not common). In the example above, the only non-common dimension in a view containing only age and subcortical region volume is the *subcortical region*, as both the *year* and *subject* measures are common to both the entities (see Fig. 1a). The expression *unrolling a dimension* here means automatically generating a sequence of selections for an entity having such dimension, each selection containing only data items along one specific coordinate of that dimension at a time. The user can choose one or more of the non common dimensions in the view to unroll, and the automatically generated selection is combined with a user specified one, if present, before aggregation and further analysis take place. When performing dimension unrolling, however, a large amount of data is being generated, and we currently deal with it by outputting only the

analyses results, such as regression or correlation values. We make use of this technique in the case-studies illustrated in Section 6, and the results of the unrolling are shown in Fig. 4.

4.5 Visualizing aggregates in physical space

Sometimes it is of interest to link abstract information of physical entities to these entities in a spatial visualization of the data. Examples could be various, ranging from the analysis of mechanical components to various kinds of simulations. In the case of cohort studies, it can be useful to visualize the content of selections, as well as other parameters in the context of the brain's anatomy. A practical example would be visualizing where the parts of the white matter fiber bundles within a certain range of anisotropy, or having certain properties (e.g., sensitivity to aging) are located. To represent statistical information for a selection in physical space, we propose to use a physical atlas of what the data refer to, which in our case is a brain atlas (Fig. 2g). The selection aggregation is then performed on the dimensions present in the atlas.

4.6 Performance and limitations of the data model

We compared the performance between our implementation of the data model described above and a relational database (SQLite) on simple queries involving aggregation. We found that, using the dataset introduced in Section 2 (approximately 50MB in SQLite form, including only the quantitative measures), operations on our data model were more than ten times faster than the corresponding operations on the relational database. For example, such operations on the largest table/cube in our database, consisting of approximately 500,000 rows, lasted 260ms using our model, compared to the 2700ms using SQLite. Data-cubes are, however, able to provide such performance only when the data fits the system memory, and our model is, at the present, not supporting out-of-core data. In case of datasets not fitting the system memory, a standard database is necessary, but with the awareness that new technology will be necessary to allow systems like this to perform interactively.

5 Prototype realization

The prototype implementation of our model has been specifically realized to explore and analyze selected aspects of the brain aging dataset produced by the cohort study described in Section 2. The prototype consists of a coordinated multiple view application implementing linking and brushing on top of the proposed aggregation engine (2). Measures can be visualized and cross-checked on demand and in different views by using drag-and-drop interaction from the measure browser window into the view of choice. Selections can be initialized and modified by means of brushes on the views. Using selections as filters is implemented via drag-and-drop: dragging a selection into a view opens a selection dialog for the measure to filter. Choosing the measure re-triggers the filtered aggregation process.

To present selections and statistical information in physical space, we employ a brain atlas onto which aggregated statistics can be mapped. For simplification purposes, we treat the brain of a representative subject S and the fine-granular parcellation of its cortical and sub-cortical white and gray matter as the atlas. A more sophisticated approach would require the averaging of brain regions across all subjects and the computation of average fiber tracts. Instead of displaying all fibers of S ($>$20000), we compute a representative fiber for each fiber bundle (Fig. 2g). This reduces visual clutter and facilitates the mapping of statistics, aggregated across all fibers of a bundle and all subjects. Previous work suggests choosing the longest fiber traveling though the densest parts of the bundle as representative fiber [8]. We apply this approach directly to homogenous bundles, i.e. all fibers following a similar course. In heterogeneous bundles, we first subdivide the bundle by grouping similar fibers, and then compute the representative of each group. For the grouping, we employ a spectral clustering technique [2]. The white matter measures (such as *FA*) in our data were extracted after subdividing each individual fiber into 100 segments of equal length, to allow tract analysis. Therefore, we also divide each representative fiber into 100 segments, allowing the system to map the measures to each segment individually. We assign a unique color and add halos to each of them to enhance the visual separation of the representatives. The aggregated values are then encoded, upon normalization, by modifying the color saturation of each fiber segment (high values resulting in high saturation). Segmentations and related measures for brain regions are also included in the study, and were extracted with Freesurfer (http://surfer.nmr.mgh.harvard.edu/). For displaying the measures, an isosurface is constructed per segmented region. The visual separation of brain regions is enhanced by assigning unique colors according to the Freesurfer's color look-up table. Mapping a measured value is then performed upon normalization by modifying the surface transparency (high values resulting in high opacity). Finally, a highly transparent outer surface of the brain is superimposed, to augment the overall atlas visualization (Fig. 3g).

6 Case-studies and Evaluation

We conducted a two-phase study with domain experts: a design requirement phase to understand the analysts' needs, and an evaluation phase to evaluate our method after the suggestions of the experts were included. In the first phase, we have been able to gather initial impressions from two neuropsychologists and one neurologist, as well as some feature requests. The prototype has been received with strong interest. Thanks to the flexibility and simplicity of performing data selections and cross-analyses, it has been seen as a practical alternative to the current way of analyzing data, consisting of extracting the values by various means into separate tables, and loading them into commercial statistical packages or tools. The experts also explicitly requested to be able to get a detailed description, save, and load the selections, and to be able to export filtered data.

After this first cycle, we included these additional functionalities, and performed the second evaluation phase with a neuropsychologist and with a neurologist. These two evaluation sessions were subdivided in three parts, organized as follows: first, a thorough explanation of the application of the underlying model was given. The explanation was followed by few questions about the model, in order to ensure the understanding of the model. This first part of the session was successful, and the domain experts could explain well the difference between our model and the table-based data model present in all the statistical analysis tools used in a standard analysis workflow, where the observations are the rows, and the variates are the columns. Neither of them was previously familiar with relational databases or OLAP cubes. During this first part of the sessions, the demonstration of the application functionalities was also well understood. The second part of the sessions had a dual aim: to verify that our model is capable of producing the same results obtainable with a standard analysis workflow, but in a faster way, and to prove that our model is capable of helping the generation of new hypotheses. For this second part two case-studies, one for each domain, were set up, and are described below. The third part of the sessions was used to gather an assessment of the proposed method by asking the domain experts specific questions, and details are given in Section 6.3.

6.1 Neurologic case-study

Jointly with a neurologist, we attempt to confirm or reject three hypotheses which were already statistically evaluated in previous work [12]:

- The increased age-related anisotropy decline in the anterior callosal fiber (CC-Anterior), as compared to the posterior portion of the corpus callosum, called splenium (CC-Splenium).
- The higher sensitivity to age-related anisotropy decline of superior fibers (Superior-LF), as compared to inferior fibers (Inferior-LF).
- The resistance of the cortico-spinal tract to age-related anisotropy decline.

To confirm the first hypothesis, we begin with selecting the fibers under investigation. Then, we use these selections as filters in scatterplots opposing the age of the subjects and the FA of the fiber segments

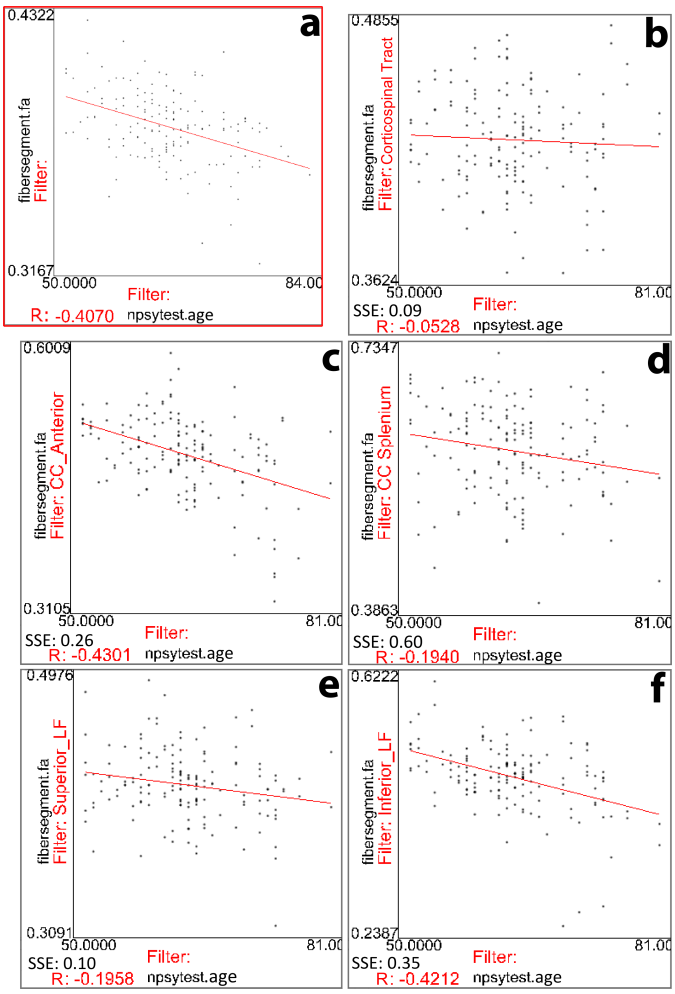

Fig. 3. **a**) Age opposed to FA for each examination (subject, year). Visualizing the linear regressor depicts the general declining trend, also summarized by the negative correlation r-value. **b**,**c**,**d**,**e**,**f**) Age opposed to FA aggregated (across segments and bundles) using a different filter in each plot, as labeled in the views. **c** and **f** show a stronger negative correlation, while **d**) and **e**) show a weaker negative correlation. **b**) shows almost no correlation between FA and age for the corticospinal fiber tracts, which confirms previously published studies, and can be used as control. In each plot: R is the correlation coefficient, SSE is the sum of squared residuals of the regression analysis.

in the subjects' brains. In these scatterplots, shown in Fig. 3, each point represents a single subject examination (*subject*, *year*), while the other dimensions are aggregated for each of the measures. In the case of FA, this aggregation is filtered using the selections above. The system automatically computes the Pearson's r value of the two measures (one aggregated using the filter), the p-value, which, in our case, is below 0.05 except for the corticospinal tract (that, therefore, does not show a correlation that is statistically significant) and the regression line. The regression analysis also provides the regression coefficient and the sum of squared residuals (SSE) as a metric for the goodness of fit. These plots confirm that the corticospinal tract is relatively insensitive to the age effect. They also show that the posterior portion of the corpus callosum is less prone to age effect compared to the frontal portion. But, in contrast to our expectations, superior fibers are less prone to age effect than inferior fibers. This could suggest the new hypothesis that language functions stay normal while the visual integration might decline. Such hypothesis, however, requires further investigations.

In the second part of this case study we decide to perform an explorative investigation of the relation between the anisotropy decline in the white matter fiber tracts and age. We do this by looking at the correlation coefficient, as well as the regression coefficient, between FA and age. The result of this analysis is shown in Fig. 3a, and we discover that there is a significant negative correlation (-0.406) between these two aspects. The regression coefficient (-0.001, not normalized) is small since the data has not been normalized, but the regressor (the red line) provides a better picture of the trend than the value alone. Once we discovered that these aspects are worth investigating, we use the unrolling mechanism described in Section 4.4 to evaluate this relation selectively for each fiber bundle. The system estimates these statistics for the chosen measure by iterating over a user specified dimension, in our case *fiberbundle*. These estimates are presented in two bar charts shown in Fig. 4a and 4b. It is easy to spot one fiber (fornix) that goes against the general declining trend, also showing a bad fitting (sum of squared residual, SSE). We decide to bring this fiber up for inspection in a scatterplot (Fig. 4c), by using the filtering capability of the system. So we manually create a specific selection, defined by *slabbing* the data-cube along the fornix coordinate of the fiberbundle dimension and use it to filter the aggregation. In the scatterplot we detect several zero values (Fig. 4c), probably due to missing data, which tells that the information for this fiber should be discarded or the missing data should be removed. We opt for cleaning the data, by performing a selection with a brush on the scatterplot that excludes the incorrect values. This leads to opposite results (Fig. 4d), in line with the overall declining trend (these results are sketched with a dashed line in the barcharts of Fig. 4a and 4b). We also notice that the corticospinal fiber tracts seem to be particularly insensitive to age decline, while other tracts have very strong decline (anterior callosal fibers and inferior longitudinal fasciculi). Finally, we notice two corresponding tracts, left and right occipitofrontal fasciculi, which are not homogeneous, with the right one showing a more pronounced anisotropy decline, even though they are anatomically symmetric to each other. This finding should be investigated further, to verify the fibers' geometrical path along which the measures have been sampled, in order to possibly formulate a new hypothesis on this phenomenon.

6.2 Neuropsychologic case-study

Jointly with a neuropsychologist, we attempt to verify the relation between the volume of the frontal regions of the cortex and the performance in the Stroop task. We focus on this task since a functional correlation between these brain regions and the Stroop effect has been discovered using functional imaging techniques [6]. Therefore we would expect that subjects with smaller frontal cortices would perform worse at the Stroop task.

We begin by opposing the cortical gray matter volume measure to the scores of the Stroop task (where higher is worse) in a scatterplot (Fig. 4f). A general declining trend (smaller cortical regions causing worse performance) is therefore expected, and it is also what we get. Then we create a selection with the frontal pole cortical area, and filter the aggregation accordingly. The resulting plot shows a minor increase in correlation and regressor slope (these values are outlined in red in Fig. 4e). Such result confirms a mild correlation between the volume of frontal cortical area and Stroop task performance, but the aggregated result tells us that there must be other cortical regions whose decline is even more correlated to the Stroop task than the frontal cortex. Therefore we decide to use the unrolling option to get an overview of how each single cortical region correlates with the Stroop task scores. The resulting statistics (Pearson's r and p-value, non-normalized regression coefficient and sum of the squared residuals) are displayed using a table widget, allowing us to sort the rows by a chosen column (Fig. 4e, sorted by r-value). Sorting these data by r value unveils an area, the parahippocampal cortical region, for which no know hypotheses have been formulated, but which shows the strongest correlation between its volume and the Stroop task scores. According to the neuropsychologist, this is a new finding, but something that requires further investigation because this analysis is done on the raw data, not corrected for basic skills.

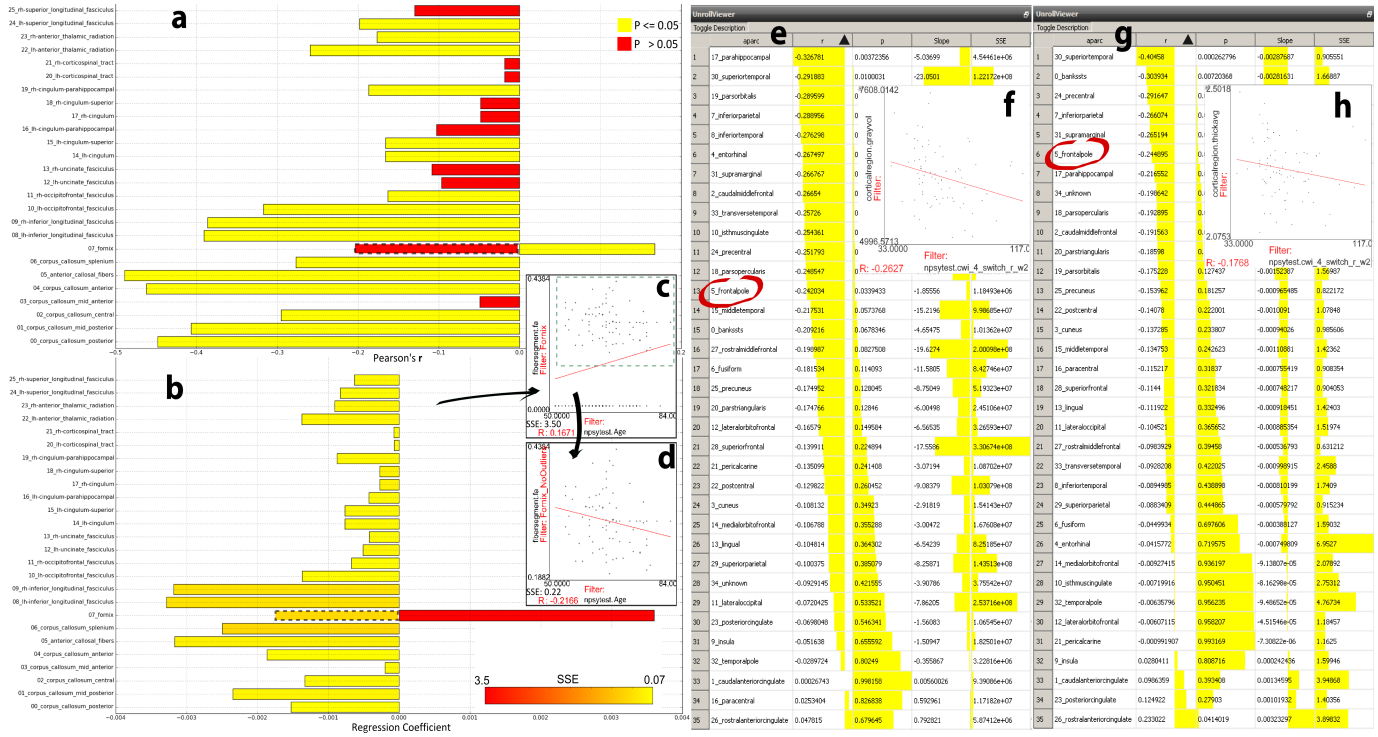

Fig. 4. **a**) The correlation coefficient between age and the FA of the fibers. The dimension *fiberbundle* for the measure FA is unrolled, meaning that FA is filtered by automatically iterating over a chosen dimension, in this case *fiberbundle*. Each bar represents the correlation for a specific coordinate in the *fiberbundle* dimension. **b**) The same type of visualization for the regression coefficient. **c**,**d**) Scatterplots related to the *fornix* fiber bundle, before and after excluding wrong values. **f**) Stroop task scores related to cortical region volume in a scatterplot. The cortical region volume is aggregated along the cortical region dimension (called *aparc*). **e**) The same analysis after unrolling the cortical region dimension. Each row in the table reports the correlation and regression results for the data filtered for a single cortical region, reported in the first column. **d**,**c**) The same analyses, but for Stroop task scores and cortical region thickness.

At this point we also wonder whether any relation between the Stroop task scores and the cortical thickness is present in the data, as thickness is another measure that has been shown to correlate with level of cognitive functions [3] We proceed as before, opposing the cortical thickness measure to the scores of the Stroop task in a scatterplot (Fig. 4h). A general declining trend is visible also in this case, but less strong than with the cortical volume. We then use the unrolling option to get an overview of how each single cortical region correlates with the Stroop task scores (Fig. 4g). In this case, the frontal pole cortical area shows a stronger correlation as compared to the overall cortical thickness. However, the task performance seems to be even more affected by other areas, most notably the superiortemporal cortical region. This is also a new finding, which however requires, as in the previous case, further investigation in order to formulate a new hypothesis on this phenomenon.

6.3 Assessment of the model

In the third part of the sessions we asked the domain experts the following questions: a) whether or not our prototype system was useful for data exploration tasks, b) whether or not such system was capable of answering specific questions, c) whether or not such system was useful to generate new hypotheses and d) whether or not such system could potentially replace their current tools.

Both scientists answered positively to question a), stating that such a tool, able to load and combine in a quick yet flexible way all the measures from such large studies, would be certainly helpful in data exploration tasks. This answer was supported by the fact that, in the current analysis workflow based on data in tabular form and commercial statistical analysis packages, all the work of data combination and selection has to be done manually for each question to analyze, which makes data exploration tasks especially cumbersome.

The scientists were also particularly positive regarding question c). The key aspects that were regarded as most useful in generating new hypotheses are: having the whole data at hand in one tool, the ease of use, and being able to fire queries in the tool. Moreover, and what impressed them the most, is to be able to automatically generate relevant selections in an iterated way while processing the data with a specific statistical method.

Concerning question b), which is also related to question d), our system proved to be effective in performing basic multivariate statistical analyses on the data. However, the domain experts stated to rarely use only basic multivariate statistics, but rather adding advanced techniques to assess relations between two or more measures. In addition, the neuropsychologist that was interviewed stated to rarely use the raw data alone, but often combine multiple measures into more advanced descriptors (e.g., correcting test results for the basic subject skills). However, the fast and flexible selection and filtering capabilities that the presented model offers were highly appreciated, since both the scientists stated to perform selections on the subjects to include in each analysis based on different parameters that vary from case to case. The conclusion for question b), and also for question d), was that an ideal tool would combine the presented model with more advanced data derivation and statistical analysis tools. This is a good lesson learnt, and a direction that, in some way, was already taken by having the R software environment embedded into our prototype system, even if not all of the requested methods are bound to the prototype yet.

7 Conclusion and Future Work

Medical cohort studies are an excellent starting point for exploratory data analysis, since most of the data acquisitions are standardized before specific hypotheses are formulated. Often, such studies are designed to provide enough data, of very heterogeneous character, such that a large set of possible hypotheses can be tested on them. Accordingly, hypothesis generation becomes an own challenge, when asso-

ciated with populations studies. In this work, we have demonstrated that an exploratory interface, which is capable of flexibly linking up different aspects of the data even if they are not given with respect to exactly the same domain, can help to swiftly identify new and possibly promising research hypotheses. We also showed, that the same approach is also capable of enabling a first quick analysis of the identified hypotheses, leading to an accelerated analysis methodology with respect to such highly rich and versatile data. The prototype system presented here, however, is still relatively limited in features, but such an application could potentially benefit from a broad spectrum of functionalities. In future, we plan to continue in this research direction, and extend the capabilities of this tool.

As future work we also plan to import genotype data for the subjects, that at the time being was not readily available, and to integrate 2D/3D graph views for representing the brain connectivity information. We are also trying to obtain a more thorough evaluation of the system in terms of new requirements, in particular from a statistical and data-mining perspective. Finally, we plan to include the retrieval and visualization of patient-specific image data, to assess whether outliers originate from the image data, or whether they are the result of an erroneous derivation process.

ACKNOWLEDGMENTS

We wish to thank Stefan Bruckner for developing the VolumeShop framework, used to create the application prototype. The cohort study used in this work was financially supported by the Research Council of Norway, University of Bergen, MedViz, and Western Norway Health Authority (grant #911593 to A.L., and grants #911397 and #911683 to A.J.L.).

REFERENCES

[1] S. Bruckner, V. Solteszova, M. Groller, J. Hladuvka, K. Buhler, J. Yu, and B. Dickson. Braingazer-visual queries for neurobiology research. *Visualization and Computer Graphics, IEEE Transactions on*, 15(6):1497–1504, 2009.

[2] A. Brun, H. Knutsson, H.-J. Park, M. Shenton, and C.-F. Westin. Clustering fiber traces using normalized cuts. *Medical Image Computing and Computer-Assisted Intervention (MICCAI'04)*, pages 368–375, 2004.

[3] A. Z. Burzynska, I. E. Nagel, C. Preuschhof, S. Gluth, L. Bäckman, S.-C. Li, U. Lindenberger, and H. R. Heekeren. Cortical thickness is linked to executive functioning in adulthood and aging. *Human brain mapping*, 33(7):1607–1620, 2012.

[4] S. Chaudhuri and U. Dayal. An overview of data warehousing and olap technology. *ACM Sigmod record*, 26(1):65–74, 1997.

[5] F. Jeanquartier and A. Holzinger. On visual analytics and evaluation in cell physiology: A case study. In *Availability, Reliability, and Security in Information Systems and HCI*, pages 495–502. Springer, 2013.

[6] M. Milham, M. Banich, E. Claus, and N. Cohen. Practice-related effects demonstrate complementary roles of anterior cingulate and prefrontal cortices in attentional control. *Neuroimage*, 18(2):483–493, 2003.

[7] C. North, N. Conklin, K. Indukuri, and V. Saini. Visualization schemas and a web-based architecture for custom multiple-view visualization of multiple-table databases. *Information Visualization*, 1(3-4):211, 2002.

[8] L. O'Donnell, C. Westin, and A. Golby. Tract-based morphometry for white matter group analysis. *NeuroImage*, 45(3):832–844, 2009.

[9] K.-M. Simonic, A. Holzinger, M. Bloice, and J. Hermann. Optimizing long-term treatment of rheumatoid arthritis with systematic documentation. In *Pervasive Computing Technologies for Healthcare, 5th International Conference on*, pages 550–554. IEEE, 2011.

[10] M. D. Steenwijk, J. Milles, M. A. Buchem, J. H. Reiber, and C. P. Botha. Integrated Visual Analysis for Heterogeneous Datasets in Cohort Studies. In *IEEE VisWeek Workshop on Visual Analytics in Health Care*, 2010.

[11] C. Stolte, D. Tang, and P. Hanrahan. Polaris: A system for query, analysis, and visualization of multidimensional relational databases. *Visualization and Computer Graphics, IEEE Transactions on*, 8(1):52–65, 2002.

[12] A. Voineskos, T. Rajji, N. Lobaugh, D. Miranda, M. Shenton, J. Kennedy, B. Pollock, and B. Mulsant. Age-related decline in white matter tract integrity and cognitive performance: a dti tractography and structural equation modeling study. *Neurobiology of aging*, 2010.

[13] C. Weaver. Cross-filtered views for multidimensional visual analysis. *Visualization and Computer Graphics, IEEE Transactions on*, 16(2):192–204, 2010.

[14] M. Ystad, A. Lundervold, E. Wehling, T. Espeseth, H. Rootwelt, L. Westlye, M. Andersson, S. Adolfsdottir, J. Geitung, A. Fjell, et al. Hippocampal volumes are important predictors for memory function in elderly women. *BMC medical imaging*, 9(1):17, 2009.

Paolo Angelelli received his Ph.D. in Computer Science from the University of Bergen, Norway. He is currently a Post-Doctoral researcher at the University of Bergen. His research interests are in fields of medical visualization, heterogeneous data visualization and interactive visual analysis.

Steffen Oeltze received his Ph.D. in Computer Science from the University of Magdeburg, Germany. He is currently a Post-Doctoral researcher at the University of Magedburg. His research interests are the visual analysis of perfusion and functional data, the visual exploration of vasculature and biological data.

Judit Haász received her M.D. from the Semmelweis University, Budapest, Hungary. She is currently a Ph.D. candidate at the University of Bergen, Norway. Her research interests are in fields of functional and structural brain changes after strokes and correlation of cognition, brain function and genetic biomarkers.

Cagatay Turkay received his Ph.D. from University of Bergen, Norway. He is a faculty member and lecturer at the gi-Centre at City University, London, UK. His research interests are the integration of interactive visualizations, data analysis techniques and supporting exploratory knowledge of experts. with special focus on bioinformatics and biomolecular modeling.

Erlend Hodneland received his Ph.D. in Physics from the University of Bergen, Norway. He is currently a Post-Doctoral researcher at the University of Bergen. His research interests are in the fields of image processing and pattern recognition, functional imaging, image registration and quantification in medicine and biology.

Arvid Lundervold received his M.D. from the University of Oslo, Norway, and his Ph.D. in physics from the University of Bergen, Norway. He his currently a professor in medical information technology at the University of Bergen, and head of the Neuroinformatics and Image Analysis Laboratory in the Neural Networks Research Group. His research interests are in the fields of image processing and pattern recognition, functional imaging, image registration, quantification and visualization, and mathematical modeling.

Astri J. Lundervold received her Ph.D. in neuropsychology from the University of Oslo, Norway. She is currently a professor in neuropsychology at the University of Bergen, Norway. She leads the Clinical Cognitive Neuroscience group. Her main research interest is to characterize behavior associated with normal function and neuropsychiatric disorders across the life span.

Bernhard Preim received his Ph.D. in Computer Science from the University of Magdeburg, Germany. He is currently a professor for visualization at the University of Magdeburg. His research interests are in medical visualization and applications in diagnosis, surgical education and surgical planning.

Helwig Hauser received his Ph.D. in Computer Science from the Vienna University of Technology, Austria. He is currently a professor for visualization at the University of Bergen, Norway. His research interests are interactive visual analysis, illustrative visualization, the combination of scientific and information visualization and the application of visualization to various domains.