A Need for Exploratory Visual Analytics in Big Data Research and for Open Science

Yuzuru Tanaka, Jonas Sjöbergh, and Keisuke Takahashi
Hokkaido University
Sapporo, Japan
Email: {tanaka, js}@meme.hokudai.ac.jp, keisuke.takahashi@eng.hokudai.ac.jp

***Abstract*—We argue that exploratory visual analytics frameworks are needed for efficient big data research and data-driven research, and exemplify with experiences from our research. Such frameworks can be used for iterative hypothesis generation and hypothesis verification, and for exploratory creation of appropriate explanatory variables to use in data acquisition and analysis. We discuss how complex analysis tools, e.g. data mining tools, can be integrated with the coordinated multiple views framework and we briefly present a framework that can support such extended coordinated multiple views frameworks and that can be used for "open science", i.e. making scientific research, methods, data, etc. reusable and more accessible to everyone.**

***Keywords*-visualization; visual analytics; exploratory data visualization; coordinated multiple views; direct manipulation; big data; open science; Meme Media;**

I. Introduction

Recently, "big data" is a hot research topic. Big data research occurs in many varied research fields. We will argue that exploratory *visual analytics*[1] frameworks are needed in big data research and exemplify our points with experiences from our research. We discuss how complex analysis tools can be integrated with the coordinated multiple views[2] framework to support exploratory visual analytics and present Meme Media[3], a framework that can support exploratory visual analytics and open science.

II. Big gap between big data core technologies and big data applications

One problem in big data research today is the big gap between the core technology research and the big data application research. Through our involvement in several big data projects, including EU FP6 and FP7 projects on clinical cancer trials and Japanese government initiative projects on for example cyber-physical systems for optimizing infrastructure, we have experienced the difficulties in bridging the gap between the various available data analysis methods and the goals of finding new personalized medicine treatments or optimized resource scheduling for government services such as winter road management.

"Big data" is often characterized by "3Vs", i.e. the volume, velocity, and variety of the data, or by "4Vs", adding the veracity of data, or by "5Vs", also adding the value of the analysis result. The phrase "big data" in applications also symbolizes a paradigm shift from mission-driven research to data-driven research, where the volume of data may not always be the major property of the target data set. A shift to data-driven research gradually allows conducting scientific research studies completely in cyber worlds, after obtaining the required data sets or through access to real time data streams. This further allows easy sharing of not only data sets but also analysis and visualization tools and services, analysis scenarios, and meta knowledge about these. This trend in turn leads towards "open science", making scientific research, methods, data, etc. more accessible to everyone. "Open science" refers to a worldwide repository of documents, data sets, tools and services, analysis scenarios, and meta knowledge about these, for researchers from different disciplines to publish, share, exchange, and easily reedit and federate resources together to help in new contexts and for new purposes.

We believe the following are requirements both for efficient data-driven science and for successful open science:

- exploratory creation of appropriate explanatory variables to use in data analysis and for data acquisition, and
- exploratory visual analytics of acquired data sets.

The first bullet refers to the process to determine which explanatory variables need to be measured by instruments or to be computed in simulations, i.e. what data will be useful. Based on this, large sets of data are collected through measurements and/or simulations. The resulting data sets are then the subjects of analysis.

In data-driven science, it is common that the target systems are heterogeneous and that they are systems of systems, where more than one subsystem with different mechanisms interact with each other. They can in turn be heterogeneous systems, mixtures of different subsystems following different mathematical models or systems with the same model but different parameter values.

Due to this heterogeneous nature, visual exploration of data is helpful to determine what explanatory variables are important. It is usually not trivial to understand how different factors interact in these systems a priori. Thus, mutually connected interactive visualizations are useful for both bullets above. Likewise, improvisational federation of data sets ("data mash-up") and of visualization or analysis tools and services is also a powerful help to exploratorily determine the best complex analysis scenarios.

III. Two typical approaches in data-driven science

The first author has been working as research supervisor for the JST (Japanese agency for Science and Technology) large-scale research funding program (CREST program) on big data applications. This program includes nine big data application projects on such diverse topics as: (1) computational drug design, (2) meteorological disaster prediction, (3) epidemic prediction and control, (4) *omic*s approach to personalized medicine, (5) tsunami disaster prediction and prevention, (6) statistical cosmology, (7) e-agriculture, (8) developmental biology, and (9) knowledge extraction and discovery from documents. Supervising this wide range of cutting edge research projects led to recognizing two different typical approaches as the common denominator in the big data analyses:

- **Data assimilation approach** In applications where the target system can be mathematically modeled fully, the typical approach is assimilation using ensemble simulations, with multiple sets of parameter values, and observed data.
- **Machine learning approach** When the target system cannot be fully modeled, the typical approach is to design an appropriate set of explanatory variables and then represent objects as multidimensional vectors of these variables. Each explanatory variable may characterize some aspect of the target system or work as a parameter of a mathematical model describing some aspect of the system.

The simulated model may have some parameters that correspond to unobservable initial conditions or candidate design conditions. An ensemble simulation runs the same simulation program many times with a large number of different combinations of parameter values in parallel.

For dynamic target systems, multiple simulations with different initial conditions are executed simultaneously to predict a probabilistic distribution of the system status in a time step Δt from now. After Δt time, observed data from the actual system status are obtained. Data assimilation is applied to choose the simulation result that best approximates the observed system status and the simulation is continued with multiple combinations of possible parameter values for unobservable parameters. The ensemble simulation can then predict the future status of the target system Δt later.

For static target systems, the result of each simulation with a different combination of parameter values can be stored in a database. Later, the parameter values that lead to the result most similar to any requirements that may arise can be retrieved from the database. They can be used directly if the result is similar enough to what is needed or they can be used as the starting point for further simulations to find parameter values that result in an even better match with the requirements.

In the machine learning approach, success depends heavily on the quality of the explanatory variables used. If some information crucial to segmenting the data properly is not available in any of the variables the machine learning cannot solve the problem properly, and if there is noise or irrelevant information it makes the problem more difficult. To define good explanatory variables it is necessary to determine what aspects of the system describe its properties, which for complex real world systems is often not obvious, and to define the variables as parameters of the mathematical modeling of each of these aspects of the system. Aspect modeling is different from total-system modeling. It may be the case that a fairly simple model can explain one specific aspect of the system, while the system in total is difficult to model. Even if the aspect model is simple, the machine learning method may not be able to extract this information directly from the model parameters. The simulation results of the aspect modeling, simulating this one aspect that can be modeled, can then be added as new explanatory variables to improve the results.

When preparing the target data set it is important to design an appropriate set of explanatory variables and then provide values for each of these for all the objects in the data set. Derived variables can be defined as functions of the original explanatory variables and need not be considered in the original design. The values of the basic explanatory variables are then obtained either through observations and measurements or through computer simulations. Derived variables can then be calculated from these.

In machine learning it is sometimes necessary to explicitly introduce some derived variables that describe some property of the target system. Depending on the machine learning method used, some types of derived variables are already implicitly considered. For example, the linear combinations of the original explanatory variables need not be explicitly introduced when using linear regression. On the other hand, a property such as x/y should be explicitly introduced as an additional explanatory variable if it plays an important role in the modeling of some aspect of the target system, since linear regression will not consider such interactions between the parameters.

As a simple example, consider the derived variable $\rho(t)$, the flow on a road where we directly measure the average velocity $v(t)$ and the number of cars $n(t)$ at time t. If we also know the static property l, the road length, the derived flow variable can be defined as $\rho(t) = n(t)v(t)/l$.

Next, consider a more complex example. In cancer treatment, "preop chemotherapy" means undergoing chemotherapy before the surgery to cut out a cancer tumor. For some cancers and in some treatment facilities, it is customary to do preop chemotherapy to reduce the size of the tumor so the surgery will be less invasive. In other places, surgery is done immediately. For some patients, the preop chemotherapy does not reduce the size of the tumor, and the tumor may

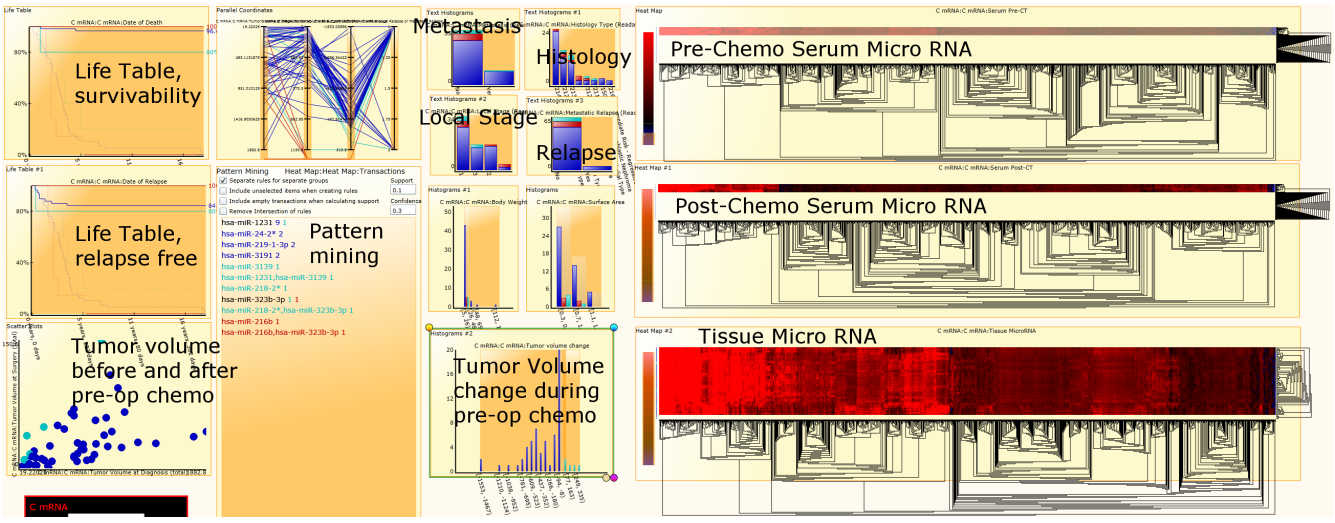

Figure 1. A framework supporting repeated hypothesis generation and hypothesis verification on clinical trial data. The system supports exploratory visual analytics through multiple coordinated views, also allowing coordination with data mining or other analysis components. It also allows interaction through direct manipulation of all visualization results.

even increase in size during these weeks. For such patients, it would be better to do the surgery immediately, and it would thus be useful to have some way, some explanatory variable, to determine if the patient will respond well to preop chemotherapy or not before making the decision.

Recently, it has been discovered that micro RNA in the serum (blood) at the time of diagnosis shows different expression patterns for patients that later respond well and patients that do not respond to preop chemotherapy[4]. Such differences in expression patterns could be mined by an appropriate pattern mining analysis of the micro RNA data together with the clinical data on the patients. Differences in expression patterns can then be used as biomarkers to identify patient groups where the preop chemotherapy is likely to be effective and groups where it likely is not effective. Such biomarkers can also be used for further segmentation of patients for deeper analysis.

This example shows that analysis results from complex analyses like pattern mining or clustering, e.g. pattern IDs or cluster IDs, may in turn become explanatory variables for further segmentation and analysis. We call such derived explanatory variables "marker variables" or simply "markers".

IV. Needs for Exploratory Visual Analytics

A. Clinical Trials for Personalized Medicine

In clinical trials, i.e. research to determine what treatments are effective for what medical problems, large amounts of patient diagnosis data, such as imaging data and genomic data, as well as treatment data is accumulated. The number of patients may range from hundreds to thousands. While the data volume is not extremely large, a large variety of data is collected and it is often not clear what factors influence the outcome in what way, i.e. the system is difficult to model and it is not straightforward to apply "planned-for" analysis scenarios.

The goal may be to find new "personalized medicine" treatments, in cases where different patients respond well to different treatments and there is no one treatment that all patients respond well to. When comparing the outcomes, e.g. the survival rate, of different treatments in the trial, there may be a need to segment patients in different ways and find specific groups of patients where one candidate treatment results in favorable outcomes more often than other treatments. Which candidate treatment is most suitable can be different for different groups of patients.

Such an analysis process is inherently iterative. There is a need for iterations of exploratory hypothesis generation, e.g. segmenting patients based on some criteria, and hypothesis verification through some analysis of the results for the resulting subsets of patients. An example of a framework for exploratory visual analytics supporting such hypothesis generation and hypothesis verification for clinical trial data[5] is shown in Figure 1. Focus on personalized medicine has increased since it became clear that the best treatment for one patient is not necessarily the best treatment for another patient who shows the same macroscopic diagnostic properties. This means that an analysis based on segmenting patients with respect to some of their macroscopic diagnostic properties may not work well.

B. Winter Road Management

Another example where macroscopic analysis may not work well comes from research on optimizing use of infrastructure, more specifically from winter road management in Sapporo. Sapporo has almost two million citizens and gets

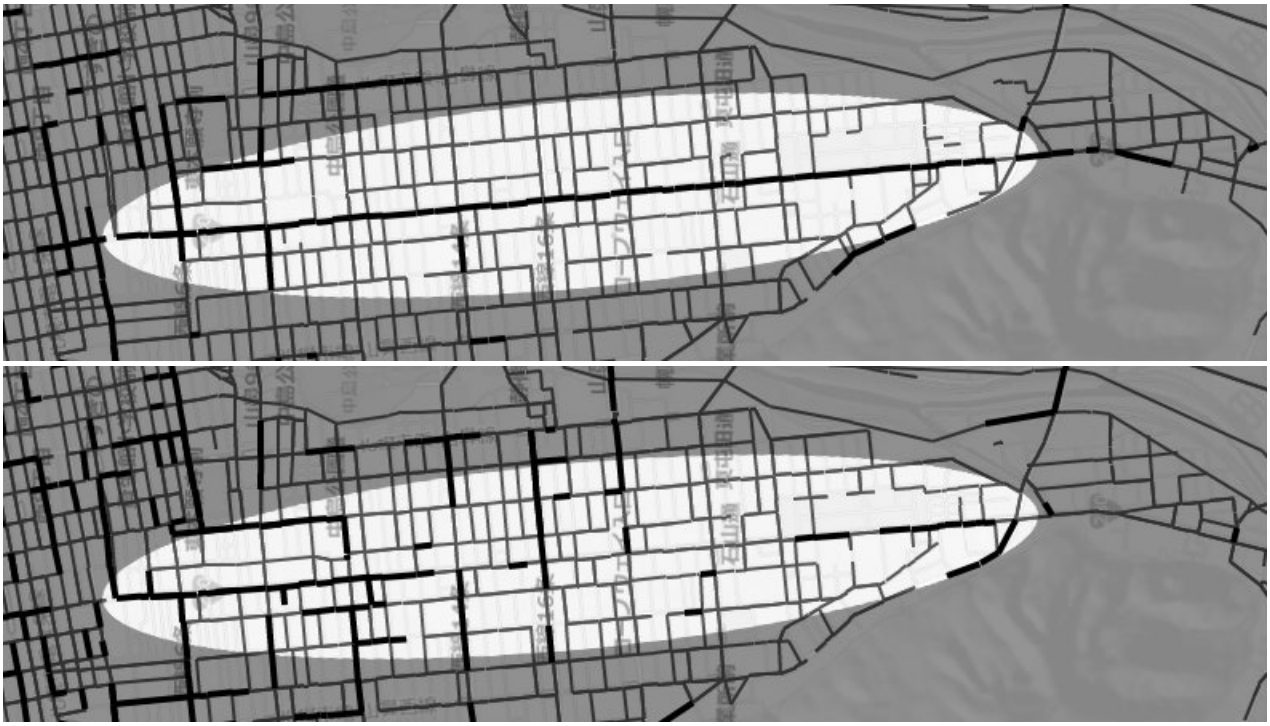

Figure 2. Clustering of road links represented by vectors of 288 dimensions: one average speed reading for each 5 minutes period during the 24 hour day. Top: Clustering result in summer. Bottom: Clustering result after snowfall in winter.

six meters of snow per year. This means snow removal and other winter road management to keep the city functioning during the winter months is a very big problem.

Some of the data to study how snowfall and snow removal operations influence traffic is collected for each road link, i.e. a stretch of road from one intersection to the next, in the city. This data includes average vehicle speed at different times and the number of vehicles passing the link. Snow removal data can also be collected per link. Other data, such as snowfall and temperature are collected in fewer locations, e.g. weather stations at different locations, and the data for a specific link has to be extrapolated from these locations, e.g. assuming that the amount of snowfall was the same as the snowfall at the closest weather station.

Average speed data is collected every five minutes, for over 100,000 road links, so the volume of data is large. Other relevant data sources that are also used in the analyses have smaller volumes. Weather data is for instance collected every 30 minutes at around 50 locations.

For a 24 hour interval, we can create a 288 dimensional vector with the average speed of each five minutes interval for each road link. Representing road links with these vectors, we can cluster the links in the city into groups of links with similar traffic patterns over the duration of the day. Vectors could of course also be constructed for longer periods than 24 hours, or for the number of cars or the traffic flow instead of for the average speed.

Consecutive links on the same road tend to cluster together, of course, since they will tend to have traffic at the same times of day and will tend to have similar average speed since the posted speed limit is the same and the posted speed limit will of course have a very strong influence on this clustering under normal conditions.

In Figure 2, the clustering result of the same clustering algorithm on the road links is shown for two different days. One day is a summer day and the other day is a winter day with snowfall, but otherwise the days are comparable, e.g. it is the same day of the week and neither day is a public holiday. The figure shows the result from only a small area but the clustering is actually done for the whole city.

One cluster of links is shown with broader lines and we can see that there is a thick line going from left to right more or less in the middle of the image. This is one of the main roads through the city, and the clustering has grouped the links on this road together in the same cluster, as expected. When clustering the data from the winter day, the links on this main road end up in many different clusters and the cluster that contains the most links from this road still does not contain very many of the links.

This indicates that the influence of snow on the traffic is not uniform even for road links on the same main road, and that any macro analysis method that assumes the same mathematical model for different links is unlikely to obtain meaningful knowledge about winter road management.

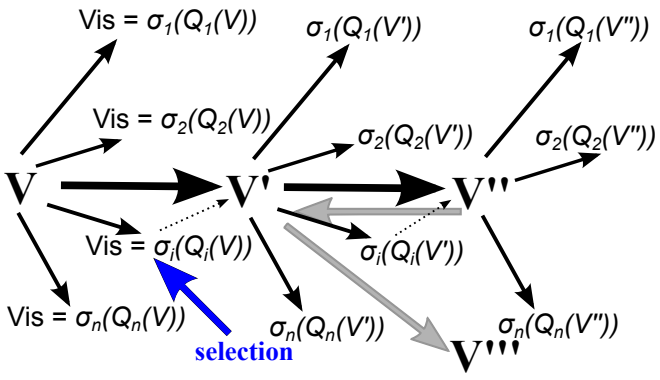

Figure 3. Schematic view of exploration of different database views using the coordinated multiple views framework.

C. Shortcomings of Macroscopic Analyses

Both examples above indicate that there are problems where macroscopic analysis without appropriate segmentation of the target data set may not work well, since it assumes that the same mathematical model will work for different objects when this may not be appropriate. It may also not be trivial to determine ways to segment the target data set to allow focus on meaningful homogeneous subsets of objects.

Data can be segmented by selecting clusters obtained by clustering methods or by specifying value ranges for some explanatory variables. The appropriate segmentation may require a new explanatory variable that is not one of the basic explanatory variables in the data set. It may be defined as a derived variable, or as a cluster ID or pattern ID obtained as the result of some analysis.

D. Exploration using Coordinated Multiple Views

Coordinated multiple views[2] is a well-known framework for exploratory visual analytics. Multiple visualizations of the same database are provided. These visualizations are linked, and changes in one, e.g. selection of a subset of data, are automatically reflected in the other linked views.

For a universal relation view V of a database Δ, the ith visualization issues a query Q_i, evaluates $Q_i(V)$, and visualizes this result using a visualization scheme σ_i to show $\sigma_i(Q_i(V))$. Each such visualization allows a user to directly specify a closed area to select visual objects in this area, which quantifies $Q_i(V)$, and hence also quantifies the underlying database view V to V'. This selection immediately changes each visualization $\sigma_j(Q_j(V))$ to $\sigma_j(Q_j(V'))$, updating all visualizations to take the new selection into account. Starting from V, you may quantify V by directly specifying a closed area of visual objects in the ith visualization of V, $\sigma_i(Q_i(V))$, to quantify V to V'. By also allowing the rollback operation, quantified views can easily be explored as shown schematically in Figure 3.

Conventional coordinated multiple views frameworks normally coordinate only visualization views, not complex analyses of the data. Frameworks often allow an analysis to be applied to each quantified database view V_i but specifying a closed area on the visualization of the analysis result for further exploration is normally not possible, e.g. selecting outliers, a specific cluster in the clustering result, or a set of mined patterns from the pattern mining result to further quantify the underlying database view is not supported. The required extension of the framework to allow directly quantifying a cluster or pattern in such an analysis result visualization is quite simple, though.

E. Extending the Coordinated Multiple Views Framework

First let us consider the integration of clustering tools and visualization of the clustering results into the coordinated multiple views framework. The result of any clustering applied to objects identified by some attribute A can be seen as a relation $Cluster(A, ClusterID)$. The values of A work as object IDs of the objects that are clustered and $ClusterID$ denotes the ID of each cluster. This relation can be visualized in various ways, for example with colored areas with sizes corresponding to the cluster sizes or with a dendrogram if the clustering is hierarchical.

Each visualization needs to provide a direct manipulation operation for selection of some objects or some clusters, for example clicking on areas representing clusters or clicking on links in a dendrogram to select a subtree. Such a selection corresponds to a quantification condition on the attribute A or $ClusterID$, which further quantifies the underlying database.

Figure 4 shows the Geospatial Digital Dashboard[6], a coordinated multiple views framework that also supports coordination with analysis tools, with the integration of a clustering tool. It has two clustering visualization views on the right-hand side. Each rectangle in the visualization represents one cluster. The size of the rectangle is proportional to the cluster size, i.e. the number of objects in the cluster. Each clustering result visualization also shows a phylogenetic tree of the clusters above the rectangles.

In this example, road links are represented by vectors with the average speed of cars passing the road link at different times of the day and by vectors of the number of cars passing. Road links are clustered based on the similarity of these vectors, i.e. they are clustered into roads with similar amounts of traffic during different times during the day (number of cars), in the top clustering view, and into roads with similar traffic speed during the day (average speed) in the bottom clustering view.

Segmenting the data based on the result of the clustering of vectors of average speeds, coloring one cluster in purple, one in green, and one in orange, we see that some clusters in the clustering based on traffic amount are also homogeneous, i.e. all objects come from the same speed based cluster,

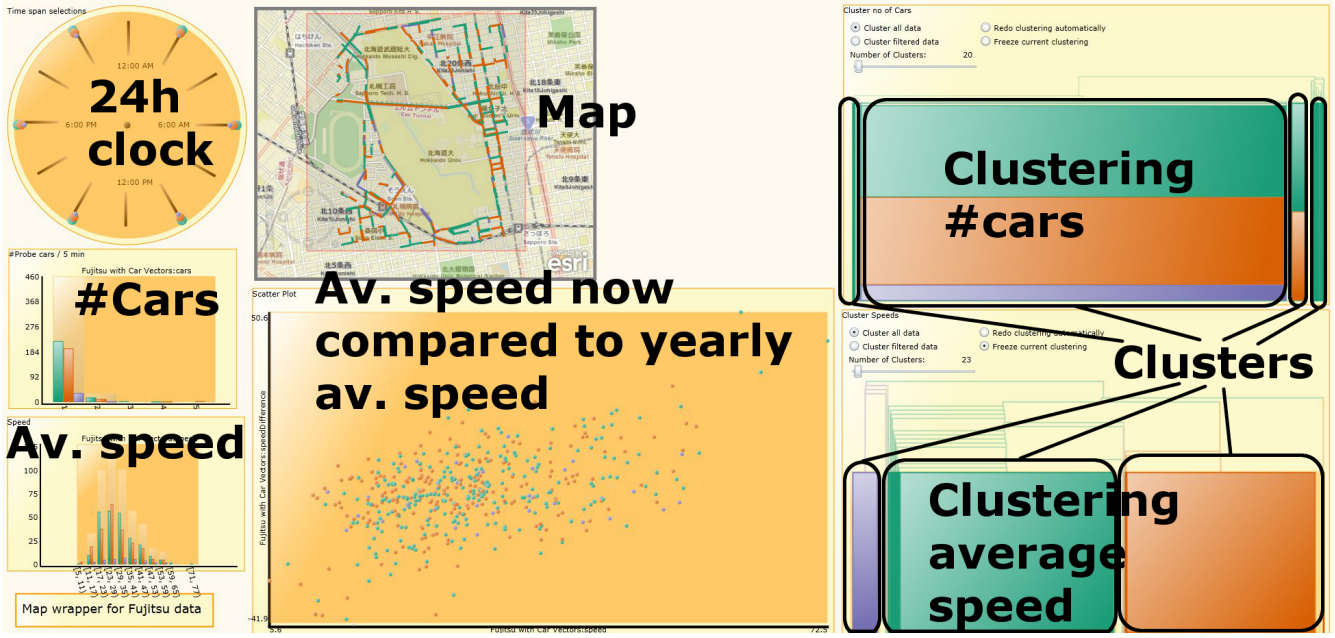

Figure 4. Geospatial Digital Dashboard with the integration of a clustering tool. Traffic data is shown in various ways. The time of day, the average speed, the location of the measurements, etc. are shown for stretches of road. The roads have also been clustered based on the amount of cars during different hours (top right component) and the average speeds during the day (bottom right component). The data has then been grouped based on the average speed clustering result.

for example the rightmost cluster which is all green. This is natural since all the big roads in the city have similar posted speed limits, so for instance road links in clusters for roads with very large amounts of traffic tend to have similar distributions of average speed.

Next, let us consider the integration of frequent pattern mining tools and visualization of the mining results into the coordinated multiple views framework. Pattern mining results can be represented by two relations, $Mining(PatternID, Support, Confidence)$ and $Matches(A, PatternID)$. The first relation lists the frequent patterns with their support and confidence values. The second relation tells which objects, identified by the attribute value of A, each of the mined patterns are applicable to.

The first relation can be visualized for instance by showing distributions of the number of patterns for different confidence and support levels, or by simply showing sorted lists of patterns with support and confidence scores. Selections can be done by selecting thresholds for support and confidence to remove patterns with values below the thresholds, or by directly clicking on listed patterns to select them. Using the second relation, the selection of some patterns is converted to the corresponding quantification condition on attribute A, which further quantifies the underlying database and changes other coordinated visualization views.

Figure 5 shows the Geospatial Digital Dashboard[6] with the integration of a frequent pattern mining tool. The heat map on the right shows traffic speed. Each column represents a road link and each row represents a five minutes interval of time, with a total of four days of data. The color of a cell represents the average speed at that road link at that time, with black being 0 km/h, and red being high speeds. Blue color means that there was no data from that link at that time, i.e. not even one car passed at that time.

The map next to the heat map shows the geo-locations of the road links on a map. Next to the geographical map a frequent pattern mining tool displays a list of patterns resulting from pattern mining together with support and confidence values for each pattern, as well as support and confidence thresholds. There are also other linked visualizations, to allow easy selection of for instance each day in the four day data sample.

On the left-hand side of the heat map component, there is a color gradient bar showing how different colors correspond to different traffic speeds. This bar allows selection of heat map cells based on the heat map intensity, i.e. the speed in this case. In Figure 5, cells with speeds below 10 km/h, i.e. roads and times where there were traffic jams, have been selected in this way.

Based on cell selections, the heat map component generates transactions for each row, i.e. one transaction for each five minutes interval of time. Each item in each transaction has a binary value, here corresponding to 1 for roads with traffic jams at this specific time and 0 when there is no traffic jam. Pattern mining can be done on these transactions, and in the example patterns of roads with co-occurrent traffic

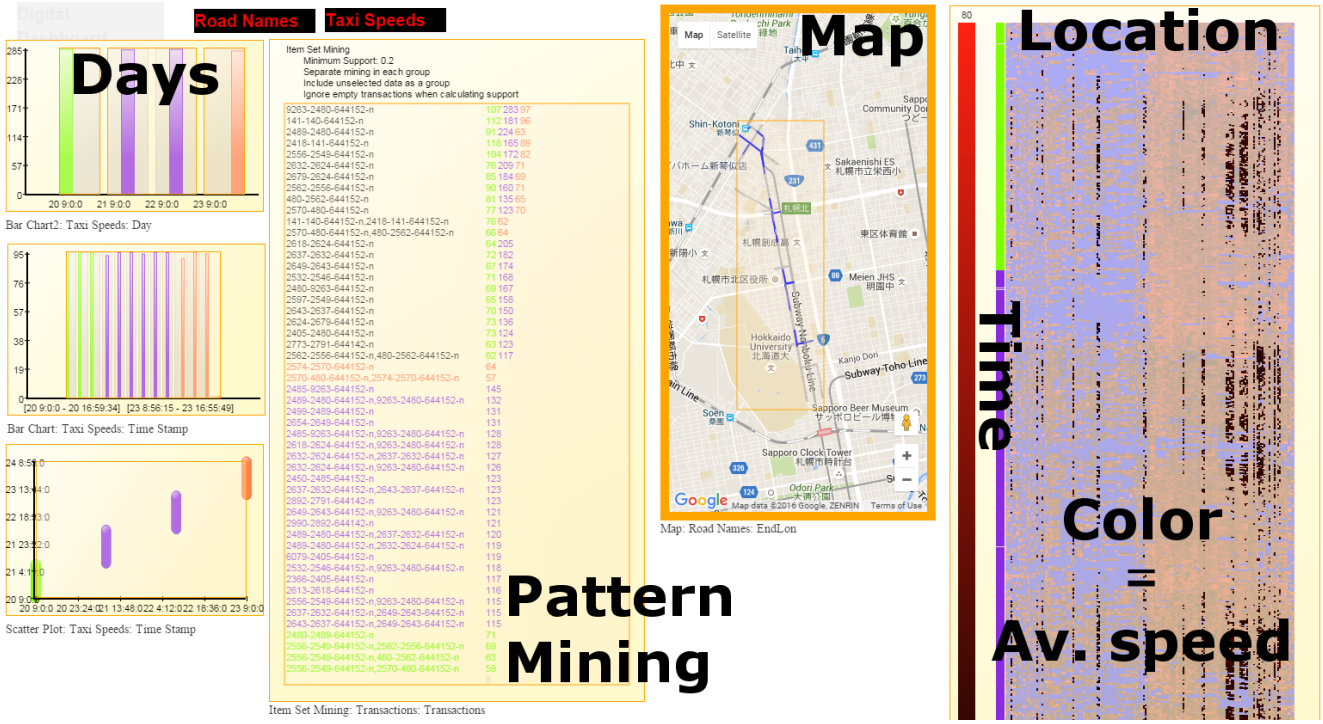

Figure 5. Geospatial Digital Dashboard with the integration of a frequent pattern mining tool. Traffic data is shown with a map showing the measuring locations, a heat map showing the speed (color) at each location (columns) at each five minutes interval (rows) for four days. A pattern mining component has been connected and used to mine patterns of locations with traffic problems. The patterns have been grouped by days with different weather using a connected bar chart.

jams have been mined and locations that match the resulting patterns, i.e. roads that often have traffic jams, are shown on the geographic map.

The pattern mining component allows specification of threshold values for support and confidence (and other pattern mining parameters that may be relevant), to select what mined patterns to display and thus also to quantify the database view shown in any other linked views. It also allows direct selection of patterns displayed in the list of mined patterns matching the threshold values, to further quantify the database view.

Since all views are linked in both directions, in the example the patterns have also been grouped into patterns matching different selections made using other visualizations, more specifically the visualization of the different days in the four day data sample has been used to group the data by day and patterns specific to each day are shown separately, as well as patterns that occur on every day etc. Visualizing the traffic problems specific to one day or the areas that have problems every day is easy, by just selecting patterns in the list of mined patterns.

V. Exploratory Visual Creation of Appropriate Explanatory Variables

In this section we use an example from materials science. Development of supercomputers and mathematical algorithms has opened up a new era of material science where material properties can be calculated on the basis of first principle calculations, i.e. not using empirical models but using only a model based on the physics of the interactions between the electrons and nuclei of atoms. While the physics is not so complex, the size of the numerical formulation (i.e. many particles need to be simulated) makes the calculations difficult.

Development of supercomputers and of computational techniques has made such simulations possible and it is now possible to create “big data” based on such simulations. The rise of big data in materials science has received much attention as material properties can potentially be predicted from data or simulations, reducing experiments or focusing them on promising areas[7], [8]. This has led to the development of the field of materials informatics.

Recent work has shown that material properties can be predicted from data using machine learning[9]. However, naive application of machine learning to materials data has some issues where some material properties become difficult to predict accurately. Material properties such as lattice constant and magnetic moment can be accurately predicted from simple descriptors using basic machine learning methods and straightforward data features[9]. However, in the experiments, machine learning did not work well to predict

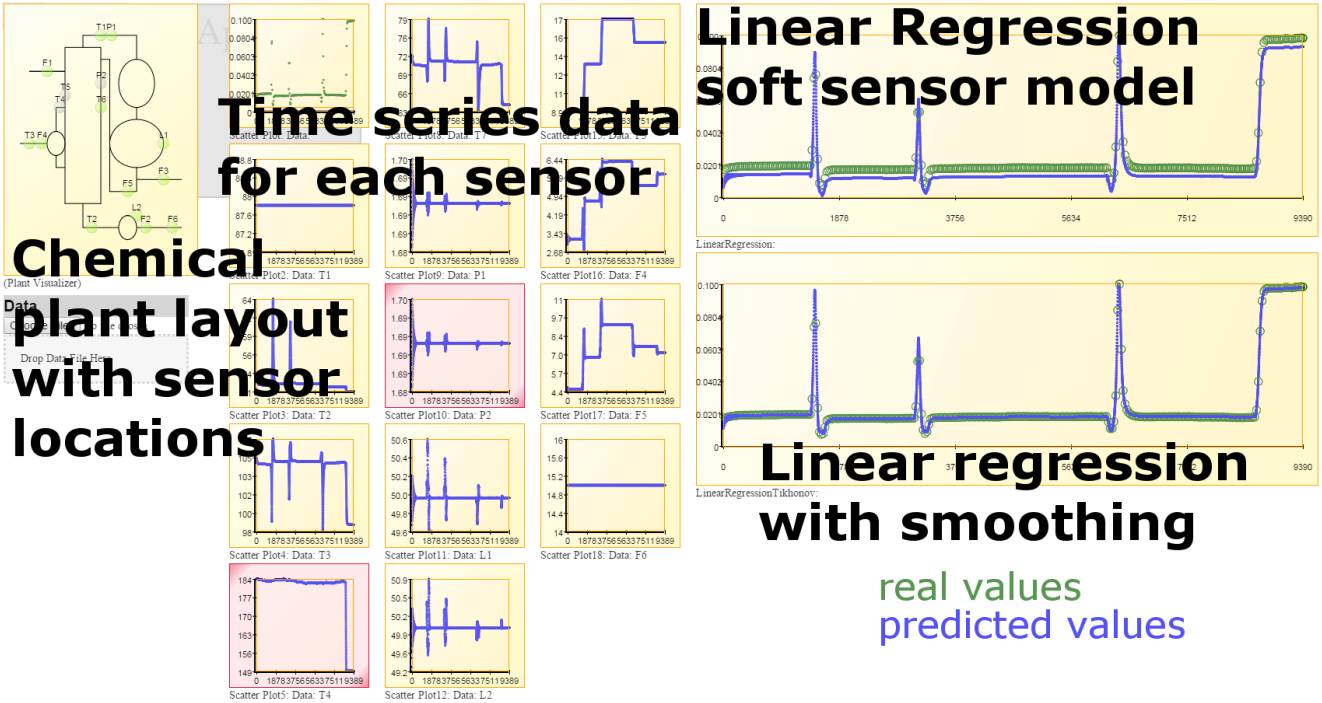

Figure 6. Interactive coordinated visualizations and soft sensor modeling components. The leftmost component shows an overview of a chemical plant and sensor locations in the plant. Time series visualizations for each sensor are shown, and on the right there are two visualizations of different soft sensor modeling results with different modeling methods showing the actual data (large green dots) and the data predicted by the model (small blue dots). Changing modeling options, input data, or turning sensors on or off updates the visualizations in real time.

the material bulk modulus (the resistance to compression of the material).

After adding new explanatory variables such as bond type, energy difference in compression and expansion, and density, to the machine learning data set, bulk modulus could be predicted. These properties can be calculated from previously used properties and the new explanatory variables are thus derived variables. Thus, no new simulations are needed but without adding the new derived explanatory variables the machine learning did not work well.

This shows that it is important to determine what key descriptors or explanatory variables are important in representing each material property that is to be predicted, but it is not trivial to determine this beforehand. Iterative refinement by adding new derived explanatory variables to try to improve the result for properties that are not accurately predicted, and then verifying which new variables actually improve the results, is important to get good predictions.

This type of iterative refinement of the machine learning by creating appropriate explanatory variables is another example of when exploratory visual analytics is helpful. Automatically showing the changes in the machine learning results when adding a new explanatory variable is helpful when trying out candidate explanatory variables. Such iterative processes to determine appropriate explanatory variables are of course not limited to only machine learning.

VI. The Hands-on Portal and Exploratory Visual Creation of Explanatory Variables

Soft sensor (or virtual sensor) means the processing of several measurements and the calculation of new properties from the measurements. A common example of using soft sensors is in the process industry where prediction of process variables which can only be measured at low sampling rates is used to get values for these variables at more frequent times. More frequent values are calculated based on variables that can be measured at higher sampling rates, and the values actually measured at the low sampling rate are used to create the model used to calculate the predictions and to adjust the model if the predictions do not match the measured values well later.

Soft sensors can be seen as a form of explanatory variables. In a project we are involved in, the *Hands-on Portals for the dissemination of technologies and analysis scenarios on CREST Big Data Applications*, we have built tools for interactively determining such explanatory variables. By showing different soft sensor modeling results visually and allowing change of input data, sampling rates, model parameters, turning sensors on or off, etc., different modeling approaches can be contrasted and how well the models fit the data etc. can be seen easily. A simple example setup is shown in Figure 6.

It is possible to choose what to model, what modeling method to use, and what to base the model on, and then see the modeling result on different data sets. Interactively changing model parameters or data sets immediately updates the visualizations. It is thus possible to visually explore different options for explanatory variables to create and to see what the different ways to create such variables, i.e. soft sensors, result in. It is thus hopefully possible to easily find the most appropriate explanatory variable, for example by seeing which modeling method and what modeling parameters seem to best fit the data.

VII. The Meme Media architecture for exploratory visual analytics and open science

Meme Media[3] is an architecture aimed to uniformly represent documents, data sets, tools and services, analysis scenarios, and meta knowledge about these as primitive and composite Meme Media objects. It also aims to provide a worldwide repository of such objects for researchers of different disciplines to publish, share, and exchange them, and to easily reedit and federate objects together for new purposes and new contexts. The most recently developed version of Meme Media is a Webtop version called Webble World 3.0[10]. It uses the Web as such a worldwide repository and the objects can run in any computer, tablet, smartphone, etc. that has a Web browser.

Meme Media components can easily be federated with each other, so it is quick and easy to connect visualization components and data mining components and then allow them to communicate. They can be federated at runtime, so if a need for one more visualization component is encountered during an analysis of some data, the needed component, if one exists somewhere on the Web, can be loaded immediately and connected to the already running components without having to restart the analysis.

Meme Media also allows wrapping of external software, i.e. software that was not built to be Meme Media, with a Meme Media wrapping layer. Once such a wrapper has been created, the wrapped software can be used together with other Meme Media objects and the other objects do not need to know that they are connected to external software.

Meme Media objects can be registered in repositories to make it easy for other users to find and reuse them but they can also be used without having been registered. For registered components there is also a trust framework, so a user can specify that they only want to run applications that use components from sources that the user has specified that they trust.

The Meme Media architecture allows each user to decompose any composite object, both static content and applications and services, published by another user and then reuse some of these components for composition of a new composite Meme Media object.

If an exploratory visual analytics environment for a specific discipline is developed in the Webble World, it can easily be embedded in a Web page and published together with the page. Any user in any discipline can then access this page and make their own copy of some components in the embedded exploratory visual analytics environment and use these components in their own context. Thus, it is easy to reuse some visualization tool that may be useful in another context. It is also easy to try some other researcher's machine learning tool on your own data by simply replacing one tool in your analysis framework or pipeline with a copy of their tool. This makes Meme Media very suitable for open science, where the goal is a worldwide repository of documents, data sets, tools, etc., that are open to anyone to share and reuse.

We have built a framework for exploratory visual analytics[11] in Webble World. It has various visualization and data mining components that can be combined in different ways. New components can easily be added if some needed functionality is missing. Since the framework is built in Webble World, it is freely available for anyone to use.

VIII. Related systems

Here we briefly mention some existing systems for visualization and exploration of data that support coordinated multiple views[2] or visual analytics[1] in general.

Some systems for visual analysis of data use graphical interfaces such as flow charts to set up pipelines of preprocessing, analysis, data mining, visualization, etc. These pipelines can fork and lead to multiple visualizations. Changing the flow chart style design automatically updates the visualizations, but interaction with the visualization results is limited or not possible. *RapidMiner*[12] and *DEVise*[13] are examples of such systems.

There are also systems that use coordinated multiple views and allow interaction with the visualization results. Interacting with one view, e.g. grouping the data or selecting a subset of data, automatically updates the other views or show details about the focused subset in other views, etc. Creating new explanatory variables using complex analysis tools and using them for further exploration is not possible, though. *SpotFire*[14], *Tioga-2*[15] (now *Tioga DataSplash*), and *Snap-Together Visualization*[16] are examples of such systems.

KNIME[17] is a system that does a lot of what we believe is needed for exploratory visual analytics. A graphical flow chart is used to set up how data should be processed and visualized. It is possible to add new user built components. Multiple coordinated views are used for visualization and selecting subsets of data in one view highlights these in other views. Such selections do not trigger recalculation of related data mining results etc., though.

Another system with many of the features we would like to see is *Orange*[18]. Here too, a flow chart is used

to set up data flows, with both visualization and analysis components. User built components can be added. Selections in a visualization component can trigger recalculation in data mining components etc. Two components cannot feed back into each other, though, so selections in one component can affect the other, but selections in the other component cannot be reflected back to the first.

CONCLUSIONS

We argued our belief that exploratory visual analytics frameworks are needed for efficient big data research and data-driven research in cutting edge engineering and science. Such frameworks can be used for the important processes of iterative hypothesis generation and hypothesis verification, and for exploratory creation of appropriate explanatory variables to use in data acquisition and analysis.

We also showed how complex analysis tools such as clustering and pattern mining can be integrated with coordinated multiple views frameworks in a theoretically sound way.

We briefly presented Webble World, a framework that can support extended coordinated multiple views frameworks and that can be used for "open science", i.e. it can be used as a worldwide repository of documents, data sets, tools and services, analysis scenarios, and meta knowledge about these, for researchers from different disciplines to publish, share, exchange, and easily reedit and federate resources together to use in new contexts or for new purposes.

ACKNOWLEDGEMENTS

This research was funded by the CPS-IIP Project in the research promotion program for national level challenges by the Ministry of Education, Culture, Sports, Science and Technology (MEXT), Japan, and by the Hands-on Portals for the dissemination of technologies and analysis scenarios on JST (Japanese agency for Science and Technology) CREST Big Data Applications project.

REFERENCES

[1] D. A. Keim, F. Mansmann, A. Stoffel, and H. Ziegler, "Visual analytics," in *Encyclopedia of Database Systems*, L. Liu and M. T. Özsu, Eds. Boston, MA: Springer US, 2009, pp. 3341–3346.

[2] J. C. Roberts, "State of the art: Coordinated & multiple views in exploratory visualization," in *Proc. of CMV '07*, Washington, DC, USA, 2007, pp. 61–71.

[3] Y. Tanaka, *Meme Media and Meme Market Architecture*. Piscataway; NJ; USA: IEEE Press, 2003.

[4] N. Ludwig, N. Nourkami-Tutdibi, C. Backes, H.-P. Lenhof, N. Graf, A. Keller, and E. Meese, "Circulating serum miRNAs as potential biomarkers for Nephroblastoma," *Pediatric Blood & Cancer*, vol. 62, no. 8, pp. 1360–1367, 2015.

[5] J. Sjöbergh, M. Kuwahara, and Y. Tanaka, "Visualizing clinical trial data using pluggable components," in *Proceedings of the 16th International Conference on Information Visualisation IV'2012*, Montpellier, France, 2012, pp. 291–296.

[6] J. Sjöbergh and Y. Tanaka, "Visual data exploration using Webbles," in *Proceedings of the Webble World Summit 2013*, ser. Springer CCIS, vol. 372, Erfurt, Germany, 2013, pp. 119–128.

[7] X. Gonze, J.-M. Beuken, R. Caracas, F. Detraux, M. Fuchs, G.-M. Rignanese, L. Sindic, M. Verstraete, G. Zerah, F. Jollet, M. Torrent, A. Roy, M. Mikami, P. Ghosez, J.-Y. Raty, and D. Allan, "First-principles computation of materials properties: The ABINIT software project," *Computational Materials Science*, vol. 25, no. 3, pp. 478–492, 2002.

[8] A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder, and K. A. Persson, "Commentary: The Materials Project: A materials genome approach to accelerating materials innovation," *APL Materials*, vol. 1, no. 1:011002, 2013.

[9] K. Takahashi and Y. Tanaka, "Material synthesis and design from first principle calculations and machine learning," *Computational Materials Science*, vol. 112, pp. 364–367, 2016.

[10] M. Kuwahara, "The power of Webble World and how to utilize it," in *Proceedings of the Webble World Summit 2013*, ser. Springer CCIS, vol. 372, Erfurt, Germany, 2013, pp. 31–55.

[11] J. Sjöbergh and Y. Tanaka, "From multiple linked views to multiple linked analyses: The Meme Media Digital Dashboard," in *Proceedings of the 18th International Conference on Information Visualisation IV'2014*, Paris, France, 2014, pp. 170–175.

[12] I. Mierswa, M. Wurst, R. Klinkenberg, M. Scholz, and T. Euler, "YALE: Rapid prototyping for complex data mining tasks," in *KDD'06: Proceedings of the 12th ACM SIGKDD*, Philadelphia, PA, USA, 2006, pp. 935–940.

[13] M. Livny, R. Ramakrishnan, K. Beyer, G. Chen, D. Donjerkovic, S. Lawande, J. Myllymaki, and K. Wenger, "DEVise: Integrated querying and visual exploration of large datasets," in *Proceedings of SIGMOD'97*, Tucson, AZ, USA, 1997, pp. 301–312.

[14] C. Ahlberg, "Spotfire: An information exploration environment," *SIGMOD Rec.*, vol. 25, no. 4, pp. 25–29, Dec. 1996.

[15] A. Aiken, J. Chen, M. Stonebraker, and A. Woodruff, "Tioga-2: a direct manipulation database visualization environment," in *Proc. of ICDE'96*, New Orleans, USA, 1996, pp. 208–217.

[16] C. North and B. Shneiderman, "Snap-together visualization: a user interface for coordinating visualizations via relational schemata," in *Proceedings of AVI'00*, Palermo, Italy, 2000, pp. 128–135.

[17] M. R. Berthold, N. Cebron, F. Dill, T. R. Gabriel, T. Kötter, T. Meinl, P. Ohl, K. Thiel, and B. Wiswedel, "KNIME - the Konstanz information miner: Version 2.0 and beyond," *SIGKDD Explorations Newsletter*, vol. 11, pp. 26–31, 2009.

[18] J. Demšar, T. Curk, A. Erjavec, Č. Gorup, T. Hočevar, M. Milutinovič, M. Možina, M. Polajnar, M. Toplak, A. Starič, M. Štajdohar, L. Umek, L. Žagar, J. Žbontar, M. Žitnik, and B. Zupan, "Orange: Data mining toolbox in Python," *Journal of Machine Learning Research*, vol. 14, pp. 2349–2353, 2013.