# Visualizing Missing Values

Jonas Sjöbergh and Yuzuru Tanaka
*Hokkaido University*
*Sapporo, Japan*
*Email: {js, tanaka}@meme.hokudai.ac.jp*

***Abstract*—Many real world data sets have data items with missing values. Values can be missing for many different reasons, such as sensor failure, respondents forgetting or refusing to answer a question in a survey, or a certain feature not being applicable to certain subsets of data. When visualizing data, some visualizations can easily handle missing values, while for others it is not obvious how to represent them without the resulting visualization being misleading. We give examples of different ways our system for interactive visual exploration of data handles missing data. These examples come from real world big data projects we took part in. Different ways to visualize missing values work well with different visualizations. Coordinated multiple views is a powerful way to visualize data with missing values, and having several views of the data helps explore the properties of the items with missing values.**

***Keywords*-missing values; visualization; visual analytics; coordinated multiple views; big data;**

## I. Introduction

With the availability of cheap sensors, network connectivity, and large storage devices, "big data" is now a hot topic in many research areas. We believe that exploratory visual analytics [1], interactively and visually exploring the data, is important for many applications of big data. It supports learning more about the data before the best modeling or analysis scenarios have been determined, as well as hypothesis generation and confirming hypotheses about the data.

Real world data sets both big and small often contain data items with missing values. Values can be missing for a variety of reasons, such as sensor failure or certain fields not being applicable to some data items. When visualizing data, missing values can be problematic. For some types of visualizations, just adding a "missing value" category is possible (e.g. in a bar chart with categorical data), but many times it is not trivial to find a way to represent missing values without the resulting visualization being misleading.

In this paper we give examples of missing data from our experiences in several big data projects. We also give examples of different ways missing values can be visualized in an interactive data exploration framework we have built.

## II. Example Visualization Framework

We have created a system that allows data visualization and interaction with all the visualizations [2]. The system is based on the coordinated multiple views [3] framework, which means all visualization components are linked to each other. Interacting with one visualization, e.g. selecting a subset of data in one plot, is automatically reflected in all the linked components, e.g. by highlighting that subset of data in the other plots.

Interaction is done by direct manipulation of the visualizations, i.e. clicking, dragging, etc., in the plots. The goal is for domain experts to be able to explore the data without being data analysis or visualization experts. Ideally, the system will still be useful to analysis experts too, though.

The system is built using a software component framework called Meme Media [4], which allows easy reuse of components from one system in another and also allows fast prototyping of new components. This makes it easy to add more visualization tools as a need arises, without the previously built components having to be adapted.

## III. Missing values

### A. Classifying Missing Values

Classifying and dealing with missing values in data is a large research area in itself. We will not go into details here, but for a good overview see for example [5].

Missing data can be grouped into three types: MCAR, MAR, and MNAR. MCAR stands for "Missing Completely at Random" and means that the reason a certain value is missing is completely independent of the data collected. One example is a test tube dropped on the floor in the lab during analysis. The likelihood of a certain sample being lost this way is independent of the measured properties of the samples.

MAR stands for "Missing at Random" and means that the reason the value is missing does not depend on the value that is missing but does depend on some other feature that is measured. The data feature that is missing has no effect on whether it is missing or not, but some other feature that is measured does. One example is a survey where older respondents are more likely to refuse to answer the survey than young respondents but the reason they refuse is unrelated to the answers they would have given if they did answer the survey.

MNAR stands for "Missing not at Random" and means that the reason a value is missing is related to the value itself. Examples include a temperature sensor failing more often in some temperature ranges or survey respondents with

low income being more likely to not divulge their income in a survey than respondents with higher incomes.

Data with missing values of type MCAR are the easiest to deal with. Somewhat simplistically viewed, they can be visualized by just ignoring the data items with missing values without being misleading, since the remaining data is still a fair random sample of the true data. Missing values of type MAR can often be treated as MCAR after correcting for the factor that the missing values depend on, if this mechanism is known. When missing values are of type MNAR it is important to visualize the data taking the missing values into account so as not to mislead the user. Whether the missing values are MNAR or MAR can often be difficult to know, though.

### B. Reasons for Missing Values

There are many reasons for missing values. Here we give some examples from projects we have been involved in.

Values can be missing because of sensor failures. In our projects we have examples of this from weather stations where the hardware at some stations break and there are no sensor readings from some stations some of the time.

Similarly, values can be missing because not all sensors in a set of sensors sample at the same frequency. In another project, some sensors in a chemical plant give readings every second while other sensors give readings only every 30 seconds. Thus, for some (many) points in time when things are measured, there are no readings for some of the sensors.

Also quite similar is to not collect all types of data for all patients in a study on cancer patients that we participated in. Some types of diagnosis or analysis were not done for some of the patients, depending on things such as the equipment availability at the hospital a specific patient went to, or depending on previously collected data on the patient (i.e. a certain analysis may be deemed not to be necessary).

Missing values can also occur because of data transmission failures. One example is from a project on winter road management where traffic is collected by having the car navigation systems of lots of taxis report where the car is, how fast it is driving, etc., every minute. When the weather is really bad (very heavy snowfall), the radio transmissions from the cars often fail, leading to many missing values. In the cancer research example, data was collected on paper and later typed in manually. Values are missing because the typist mistakenly forgot to type in some of the data.

Values can also be missing because the coverage of the sensors is not good enough for complete coverage. One example is from the winter road management project where average speed data of the traffic on the streets in the city are collected from car navigation systems of cars driving in the city. Since there are many roads that have very little traffic, there are many roads and points in time where the average speed value is missing because no car passed that road during the sample period. Another example of this kind was weather radar data. How far the radar penetrates can depend on the weather and on things blocking the view, leading to areas with no coverage at certain times.

Values can also be missing because some features are not applicable in all cases. One simple example is the "time to death" data field in the cancer research, which is missing for all the patients that are still alive.

There are of course many other reasons for missing values than the examples from our real world projects. Some examples include respondents leaving something out in surveys (by mistake or because they do not want to answer that specific question), samples going missing, not enough material left in some samples for reanalysis later, a computer file disappearing or getting corrupted on the storage device, and much more.

### C. Dealing with Missing Values

How to deal with missing values is of course also a large research area. For more information, see for example [6] or [5]. Some examples of how to deal with missing values include deleting data items with missing values, replacing missing values with some default value, imputation (replacing missing values with substitute values based on the values that are not missing, for example through regression, interpolation, or replacing with values from the most similar other data item, etc.), or using an analysis method that can take the fact that some values are missing into account (though many analysis methods cannot do this).

Deleting data items with missing values is of course simple to implement, and in some cases this can be a good way to deal with missing values. If the missing values are randomly distributed (type MCAR) then removing items with missing values will not bias the analysis.

Replacing missing values with a default value is also easy, but is often misleading. In some cases, domain knowledge may indicate a reasonable value to use when values are missing. For example, the domain experts may know that the "tumor volume at surgery" value being missing always means that it was not measured because the patient did not undergo surgery because the tumor had disappeared from the previous treatment. Setting the missing value to 0 may then be reasonable in such cases (possibly depending on the analysis to be done).

If the system can be reasonably modeled, it may be possible to impute the missing values with good accuracy. One example from our projects was the chemical plant, where the values of the sensors with infrequent samples could be modeled very accurately based on sensor readings from other sensors sampling more frequently.

## IV. Visualizing Data with Missing Values

Our goal when building the visualization system has been to explore data sets that have missing values. Our goal has not been to explore the properties of the missing values,

though this can also be done with our system to some extent. Learning more about the missing values can be very useful when deciding how to impute values, learning why values are missing (if this is not already known), or trying to find a way to reduce the amount of missing values from data collected in the future, and is also an interesting goal.

Examples of visualizations that have been used specifically to study missing values but that we do not currently support in our system include such things as a "missing map", i.e. a 2D plot with the data items on one axis and the data features on the other axis, and coloring the cells depending on if the value for each feature is missing from each data item or not. Similarly, people use bar charts showing one bar for each data feature and the number of data items where the value for that feature is missing, see for example [7] and [8]. Other systems also use the coordinated multiple views [3] framework for exploring the properties of missing values. Showing what the distribution of the items with missing values look like in areas where the values are not missing can be very powerful to learn more about the missing values.

Our system does not support imputation of values in general, though there are some components that for instance do linear regression modeling to fill in missing values.

In visualization, some ways to handle missing values include: not handling missing values, by for instance prompting the user to fill in them missing values (often leading to the user just adding some default value) or delete the items with missing values, or in some systems even by the system crashing. Another way is to silently drop all items that have missing values, i.e. plotting the data as if the items with missing values were not in the data set. This can be a good way to handle missing values, but it can also lead to misleading results.

For some types of missing values, just treating a missing value as a special value is possible, and for visualizations of categorical data this is common, i.e. adding a "Value Missing" category and having a bar for that in a bar chart or adding a slice for the items with missing values in a pie chart. Similarly a heat map can have a special color, a color that is not used to indicate any of the normal values, for cells with missing values. For some other types of visualizations, not drawing anything where the missing value should have been visualized is understandable to the user. Examples include data in grid structures, where a cell in the grid being empty can indicate that the value for this cell is missing, or time series visualization where a gap in the time series because data are missing can be understood.

In user studies, users have indicated that they prefer to have missing values marked in the visualizations [9], preferably with an explanation of why the values are missing if this is known, over just leaving the missing data out of the visualization or filling the data with default values (which can be very misleading).

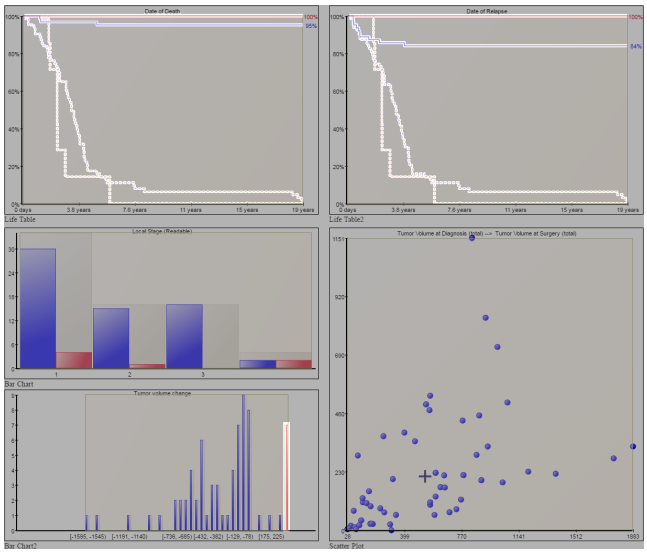

Figure 1. Cancer patient data visualized with life tables, bar charts, and a scatter plot. The bottom bar chart has a bar with a "[No Value]" category (the rightmost bar, colored red) for the missing values in the data it displays. The scatter plot also has missing values and these are simply not plotted. The values missing from the scatter plot are instead automatically colored in a different color and can easily be found in all the other visualizations.

Figure 1 illustrates three different ways of visualizing missing values using our system for visual exploration of data. The data comes from a project on kidney cancer in children. The figure shows our system set up with two life tables, the two top components, that show the number of patients (in percent) still alive (in the left plot) or still relapse free (the right plot) as a function of the time since the patient was diagnosed with cancer.

On the bottom right there is a scatter plot that plots the tumor volume at the time of diagnosis on the horizontal axis and the tumor volume after four weeks of chemotherapy on the vertical axis. In this study, patients were first given four weeks of chemotherapy to shrink the tumor, then had surgery to cut out the remaining tumor, and then either had chemotherapy again or had radiotherapy.

The bar charts show tumor "stage" (how far the cancer has progressed; upper bar chart), and the tumor volume change during the four weeks of chemotherapy (bottom chart).

The bottom bar chart showing the tumor volume change has missing values. These are handled by having a category "[No Value]" with its own bar, the rightmost bar in the plot (colored red). This illustrates a simple case of dealing with missing values, creating a separate category for the items with no value and showing this category together with the rest of the data.

The scatter plot also has missing values. The patients with values missing for either the tumor volume at diagnosis or after chemotherapy (or both) are simply not plotted. This illustrates the also very simple way of plotting the data as

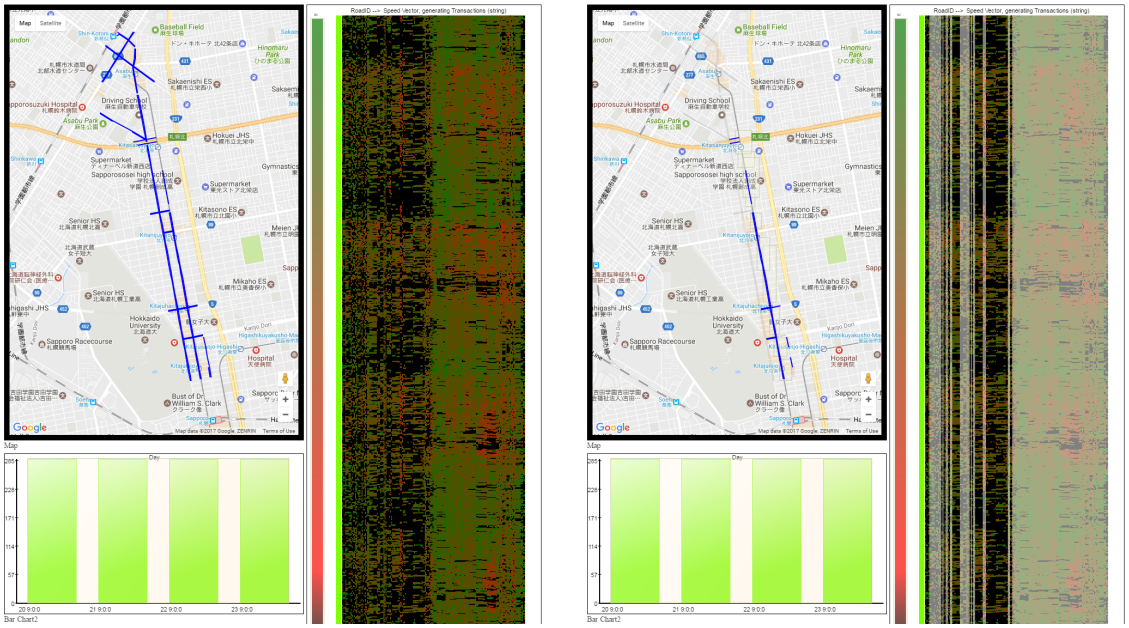

Figure 2. Traffic data collected from taxis visualized in a heat map, on a geographical map[2], and in a bar chart. Missing values in the heat map are colored in black. On the right, measuring locations on a street with very little taxi traffic have been selected using the map.

if the patients with missing values did not exist in the data.

The scatter plot component does however signal to the coordinating component of the coordinated multiple views framework that these patients had missing values, and these patients are colored in a different color (red instead of the blue used for the patients with no missing values in this case) in all other visualizations. The life tables thus have two groups of patients, one shown in red that does very well (no patient died and no patient had relapse) and one in blue. The bar charts also show the patients in different colors and we can see that as expected the patients with missing values for tumor volume change in the bottom bar chart are the same patients that had missing values for tumor volume at diagnosis or after chemotherapy (i.e. the "[No Value]" group is only colored in red and there is no red in any of the other bars). The multiple linked views idea used like this is a powerful help when exploring missing values.

For plots of numerical data like the scatter plots it is also possible to use a reserved location to show items with missing values. For a scatter plot, patients with only one value missing could be plotted below (or to the left of, if it is the vertical axis) the axis line of the axis for which they have a value. This would allow the user to see the one value that these patients have, and to interactively select these patients in different ways. Our component does not currently support this, but there are systems that allow this [10].

Finally, the life tables also have patients with missing values. Here, the fact that a patient is missing a value has clear meaning. In the case of the life table showing the patients that are still alive, it uses the date of diagnosis and the date of death of a patient. If a patient is missing a value for date of death, it normally means that the patient is still alive, and the visualization shows the result if we assume this is true. This illustrates how a missing value can have a clear meaning.

The date of death not having a value does however not mean the patient is still alive with 100% certainty. If a patient dies, the hospital or researchers normally receive this information but in rare cases it can happen that a patient dies and this information never gets reported back. The life table visualization shows this uncertainty by also plotting a dashed line based on the "date of last communication with patient" data field. If the patient was still alive at the last time data was collected for this patient, we can be sure that the patient lived at least this long.

The upper curve in the plot shows the optimistic estimate of the number of patients that are alive (or relapse free), assuming that if there is no information that the patient is dead, the patient is still alive (and this curve is usually very close to the truth). The lower dashed curve shows the most pessimistic estimate, showing the result if every patient died immediately after the last data was collected. The true value is somewhere between these two lines, and usually close to the optimistic estimate.

In Figure 2 our system is shown with data from a project on winter road management. Traffic data from taxis are shown on a map, showing the measuring locations (blue lines on the streets where data was collected), a bar chart showing

[2]Google, map data: Google, ZENRIN

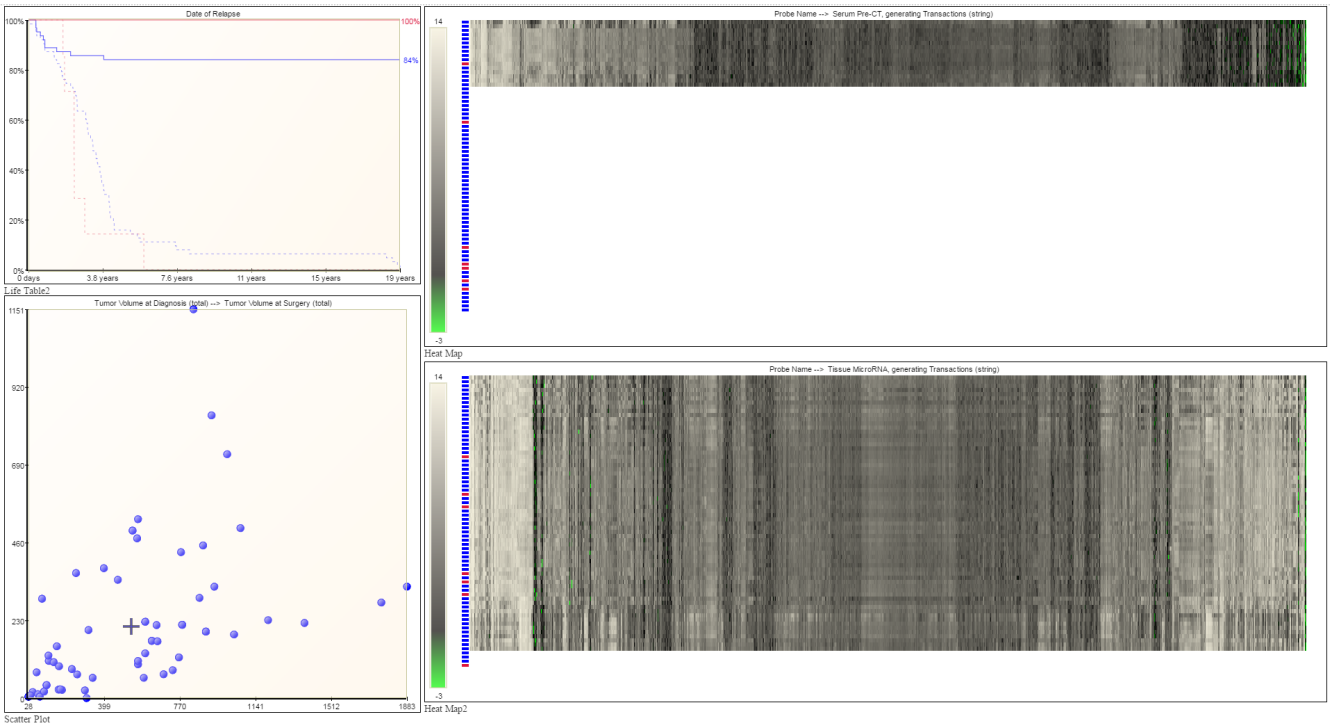


Figure 3. Cancer patient data visualized with a life table, a scatter plot and two heat maps, similar to Figure 1. The heat maps show microRNA data: rows are patients, columns are genes, and the color reflects the DNA probe response. Patients are colored in red if they have missing values for the scatter plot and blue otherwise. The heat maps also have missing data and rows corresponding to patients for which no microRNA data was collected are simply left blank and all patients with missing values are grouped together.

the amount of data for each of four days of data from this sample, and a heat map showing the average speed (the color of the heat map; green meaning fast, red meaning traffic jams, and black meaning missing values) of each measuring time (the rows) at each measuring location (the columns).

The heat map has many missing values, since there are many times during the day when some of the roads in the city do not have any taxis passing by. Here, a special value, the color black, has been used to visualize the times and locations with missing values. We can see that there seems to be two groups of measuring locations, one with very few missing values (columns with very little black) and one group with very many missing values (columns with a lot of black). This actually comes from two parallel streets, one which is heavily used by taxis (and thus has few missing values) and one that is mostly avoided by taxis (and thus has many missing values), and the heat map has grouped all the measuring locations from these two streets together.

In Figure 2, on the right, measuring locations from the street mostly avoided by taxis have been selected using the map. In the heat map the selected streets are shown in stronger colors and the unselected streets are whitewashed. The selected measuring locations as expected have a lot of black cells in the heat map (i.e. these columns have a lot of rows with missing values because no taxi passed the measuring location at these times) and selecting these locations covered most of the heat map cells with missing values.

In Figure 3 the same heat map component has been used with the kidney cancer data. The life table and scatter plot show the same data as in Figure 1, i.e. the life table shows patients that are relapse free as function of the time since they were diagnosed with cancer, and the scatter plot shows the tumor volume before and after four weeks of chemotherapy.

The heat maps show microRNA data. Each row in the heat map corresponds to one patient and the columns correspond to genes. The heat map intensity reflects the DNA probe response from the microRNA analysis. The upper heat map shows microRNA from blood serum samples, while the bottom heat map shows microRNA from the tumor tissue cut out during the surgery to remove the cancer. The upper heat map has a lot of missing values because this type of analysis is not done for all patients. The bottom heat map has some, but much fewer, missing values because some patients had such a good response to the four weeks of chemotherapy that the tumor disappeared and no surgery was necessary.

Here, the missing values are handled by simply not showing anything for the patients with missing values and by grouping all patients with missing values together at

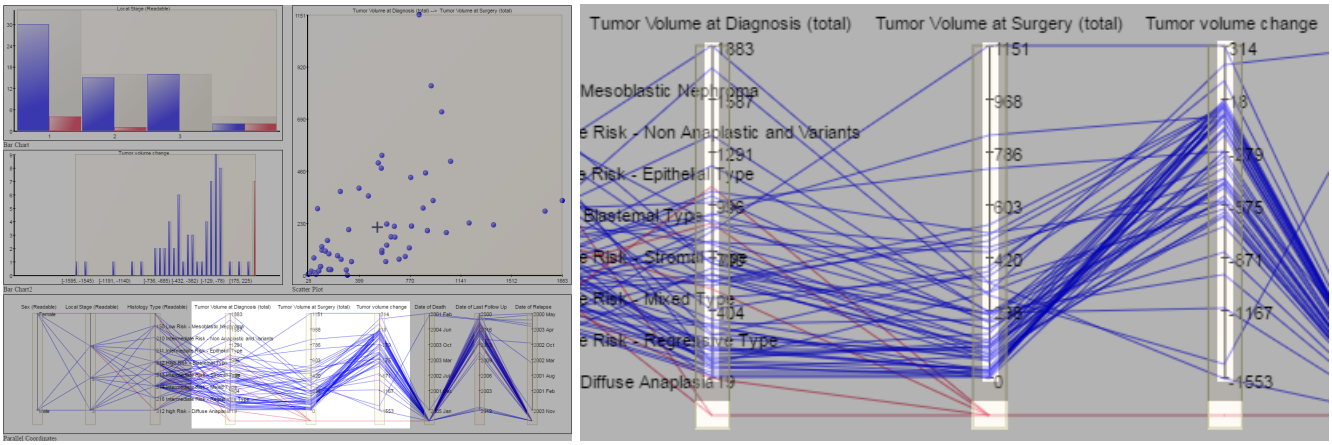


Figure 4. Adding parallel coordinate visualization to the cancer data visualization of Figure 1. Missing values are drawn outside the coordinate axes.

the bottom. Each patient has a colored marker to indicate that there is a patient and to show which group the patient belongs to. As in Figure 1 the patients have been colored red if the scatter plot indicated that the patient had missing values for the tumor volumes and blue for the other patients.

This illustrates displaying nothing at all to indicate missing values, in a grid structure where it is easy to see that something is missing, It is easy to see that there are many patients with missing data in the upper heat map since there is a huge white area where there are still red and blue dots in the left hand side column showing the colors of the patients. It is also easy to see that the upper heat map has more missing values than the bottom heat map, since the white area under the heat map is much larger.

Figure 4 also shows data from the cancer patients. Here a parallel coordinate [11] component has been added to show and allow interaction with the very high-dimensional data. In the parallel coordinate visualization, each patient becomes a polyline from one vertical bar to the next bar and then again to the next etc. Each vertical bar is one dimension, i.e. one feature in the data such as “tumor volume at diagnosis”, “date of death”, or “patient relapsed”, etc. The polyline for a patient goes from the position on the vertical bar corresponding to the value for that patient in that dimension, e.g. from the location corresponding to the tumor volume the patient had, to the location on the next vertical bar corresponding to the value for the patient in this dimension, e.g. to the position of the date of death of this patient.

There are several coordinates in the parallel coordinate visualization that have missing values, for example the tumor volume coordinates (since we know there are missing values for these from seeing the scatter plot before) and the coordinate for date of death (since not all patients died). For such cases, it is possible to simply not draw the line segment starting or ending on the coordinate with the missing value for such patients. It would however be difficult to notice if there are values missing by doing this, since there are many crossing line segments when there are many patients.

This component handles missing values by drawing lines to a position completely outside the parallel coordinate axis, as can be seen in the right part of Figure 4 where a few coordinates with missing values have been enlarged. This is a common way to deal with missing values when using parallel coordinates, used in other similar systems too. Similar things can be done for other plots such as the scatter plot too.

As before, the patients have been colored red if they had missing values for the scatter plot and blue otherwise, and we can see that the red line segments go to positions outside the coordinate axes, indicating missing values, as expected. Visualizing the missing values like this also makes it easy to select patients with missing values for specific coordinates to highlight them on other parallel coordinates or in other visualization components.

Figure 5 shows data from a chemical plant. There are sensors measuring temperature, flow, pressure, etc., in various parts of the plant. Some sensors have a high sampling rate, while some sensors have a much lower sampling rate. Since at some parts of the chemical process there are very large changes under a fairly short time, it would be good to have an estimate of the values for these sensors too during the times when no actual sensor reading is available.

In Figure 5, the leftmost component shows the layout of the plant, the locations of the sensors, and the status (working or not) of the sensors. In the middle there are time series plots for a selection of the sensors. On the right there is a plot showing a sensor with lower sampling rate and thus many times with missing values (no actual sensor data available). Here, the actual sensor readings are shown as large red circles. There are also smaller blue dots showing estimates of the values at times with missing values.

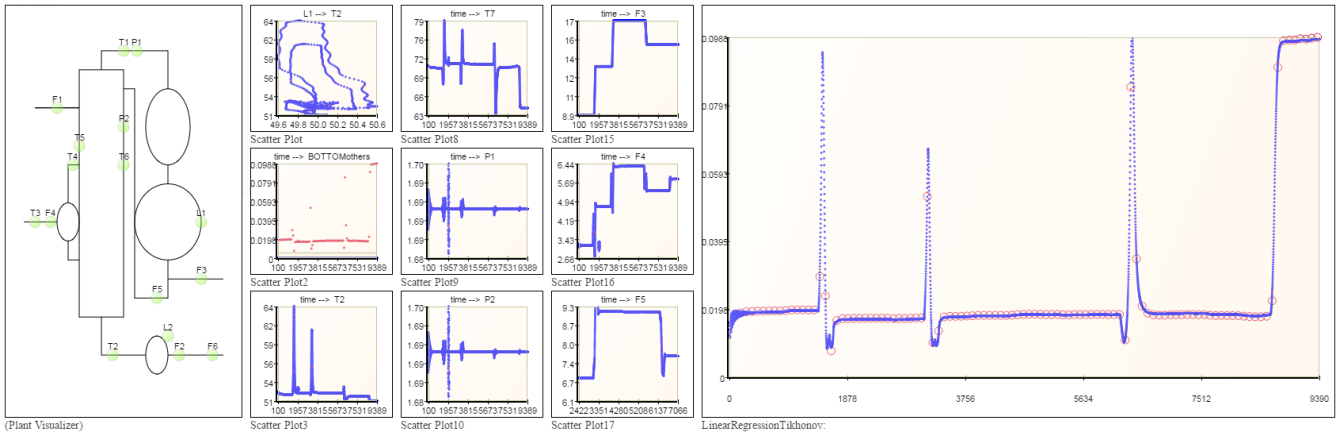


Figure 5. A chemical plant with various sensors. The values from a selection of sensors are shown in the time series plots in the middle. On the right the actual readings (in red) and predicted values (in blue) from a sensor location with a lower sample frequency, i.e. many times with missing values, are shown.

These are based on a linear regression model using the data from the sensors with higher sampling rates and training the model using the times where there is sensor data available for this sensor too. This illustrates how missing values can be dealt with by imputing values based on modeling the measured property.

## V. Discussion

In the previous section, we showed several different ways to visualize missing values:

- a "missing value" category (the bar chart, Figure 1),
- using a special value (the heat map with traffic data, Figure 2),
- plotting in a special location (drawing to a position outside the coordinate axis on the parallel coordinates, Figure 4),
- drawing nothing for the items with missing values when the fact that something is missing is easy to see (the heat map with microRNA data, Figure 3),
- not drawing anything in the plot itself for items with missing values but highlighting them in all linked views where they do have values (the scatter plot when linked to other visualizations, Figure 1),
- visualizing with a method that takes the missing values into account when the missing values have a clear meaning (the life tables, Figure 1),
- and imputing values (the chemical plant visualization, Figure 5).

Different visualizations work well with different ways to visualize missing values. Some ways of dealing with missing values can be misleading in one type of visualization but not in others, for example leaving things out (silently dropping) from a bar chart can be misleading while leaving things out from a heat map is less so (since the grid structure makes it clear that something is missing).

The coordinated multiple views framework is powerful for visualizing data with missing values. Even if one particular visualization cannot show the missing values in a good way, the ability to have items with missing values automatically highlighted in other views where they do have values gives a lot of information about the properties of the items with missing values. It also makes it easy to see that some items are missing (e.g. there are no items in red in the scatter plot but there are items in red in the other visualizations). This powerful functionality comes "for free" with the coordinated multiple views. Looking at the data in several different ways can to some extent mitigate any visualization that is misleading because of how the missing values are handled.

## Conclusions

In real world data it is common that values are missing. When visualizing data with missing values, it is important to take care to not make the visualizations misleading. Different ways of visualizing missing values are appropriate for different visualizations; one way of dealing with missing values can work well with one type of visualization but be misleading with another.

Having more than one view of the data helps mitigate problems with ways of visualizing missing values that lead to misleading results. Coordinated multiple views is thus a powerful framework when visualizing data with missing values. Having data that is missing from one visualization automatically highlighted in other visualizations is a powerful help when trying to understand the properties of the items with missing values. It also helps make it clear that something is missing from one visualization and can thus help reduce the misleading effects of some ways of dealing with missing values.

## 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. Sjöbergh, X. Li, R. Goebel, and Y. Tanaka, "A visualization–analytics–interaction workflow framework for exploratory and explanatory search on geo-located search data using the Meme Media Digital Dashboard," in *Proceedings of IV'2015*, Barcelona, Spain, 2015, pp. 300–309. [Online]. Available: http://dr-hato.se/research/IV2015.pdf

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

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

[5] R. Little and D. Rubin, *Statistical Analysis with Missing Data*, 2nd ed. New York, NY: John Wiley & Sons, 2002.

[6] J. Schafer and J. Graham, "Missing data: Our view of the state of the art," *Psychological Methods*, vol. 7, no. 2, pp. 147–177, 2002. [Online]. Available: http://dx.doi.org/10.1037/1082-989X.7.2.147

[7] X. Cheng, D. Cook, and H. Hofmann, "Visually exploring missing values in multivariable data using a graphical user interface," *Journal of Statistical Software*, vol. 68, no. 1, pp. 1–23, 2015. [Online]. Available: https://www.jstatsoft.org/index.php/jss/article/view/v068i06

[8] Z. Zhang, "Missing data exploration: Highlighting graphical presentation of missing pattern," *Annals of Translational Medicine*, vol. 3, no. 22, 2015. [Online]. Available: http://atm.amegroups.com/article/view/8666

[9] C. Eaton, C. Plaisant, and T. Drizd, "Visualizing missing data: Classification and empirical study," in *Proceedings of IFIP International Conference on Human-Computer Interaction*, Rome, Italy, 2005, pp. 861–872.

[10] M. Templ, A. Alfons, and P. Filzmoser, "Exploring incomplete data using visualization techniques," *Advances in Data Analysis and Classification*, vol. 6, no. 1, pp. 29–47, 2012. [Online]. Available: http://dx.doi.org/10.1007/s11634-011-0102-y

[11] A. Inselberg and B. Dimsdale, "Parallel coordinates: a tool for visualizing multi-dimensional geometry," in *VIS'90: Proceedings of the 1st conference on Visualization '90*, Los Alamitos, CA, USA, 1990, pp. 361–378.