

InfoZoom – Analysing Formula One racing results with an interactive data mining and visualisation tool

Michael Spenke and Christian Beilken

GMD – German National Research Center for Information Technology

FIT – Institute for Applied Information Technology, <http://www.gmd.de/fit>

Schloss Birlinghoven, D-53754 Sankt Augustin

Michael.Spenke@GMD.de, <http://fit.gmd.de/hci/pages/michael.spenke.html>

Abstract

This paper describes the application of the data analysis tool InfoZoom to a database of Formula One racing results of the last 20 years. No automatic method for data mining is used. Instead, InfoZoom enables the user to interactively explore different visualisations of the data. In this way, the user gets a feeling of the data, detects interesting knowledge, and gains a deep understanding of the data set.

InfoZoom displays database relations in tables with attributes as rows and objects as columns. In our example, each column corresponds to the participation of a driver in a certain race. The table has about 8000 columns. The attributes include the name of the driver, his team, the date and location of the race, and the starting and final position. InfoZoom compresses this table by reducing the column width until all the 8000 columns fit on the screen. The column width is then about 0.1 pixels.

Special techniques are used to make such highly compressed tables readable. The most important is that neighbouring cells with identical values are combined into one larger cell. The width of each cell indicates the number of subsequent objects with this value. If a cell is too small to display a numeric value, a short horizontal line still indicates its relative height. In this way the value distribution of each attribute can be seen at a glance.

Correlations can be found by subsequent sorts by different attributes and by animated zooms into interesting areas of the table. Like the formula-cells in a spreadsheet program, derived summary attributes (like average, maximum etc.) can be defined which are automatically updated by InfoZoom when necessary.

InfoZoom was initially developed at GMD and is now extended and marketed by the spin-off company humanIT.

Basic concepts of InfoZoom

InfoZoom displays database relations in tables with attributes as rows and objects as columns. In Figure 1 our sample database of Formula One results is shown. Each column corresponds to the start of a driver in certain race. Among the attributes are the *Date* which uniquely denotes a race, the *Location* of the race, name and *Picture* of the driver, the *Team* of the driver, as well as *Start position* and the final *Result*. The attributes are hierarchically ordered like files in a directory.

8205 of 8205 Objects	Häkkinen M., 01.11.1998	Irvine E., 01.11.1998	Coulthard D., 01.11.1998	Hill D., 01.11.1998	Frentzen H., 01.11.1998	Villeneuve J., 01.11.1998	Alesi J., 01.11.1998
Start	Häkkinen M., 01.11.1998	Irvine E., 01.11.1998	Coulthard D., 01.11.1998	Hill D., 01.11.1998	Frentzen H., 01.11.1998	Villeneuve J., 01.11.1998	Alesi J., 01.11.1998
Season	98	98	98	98	98	98	98
Date	1998/11/1	1998/11/1	1998/11/1	1998/11/1	1998/11/1	1998/11/1	1998/11/1
Location	Suzuka	Suzuka	Suzuka	Suzuka	Suzuka	Suzuka	Suzuka
Picture							
Driver	Häkkinen M	Irvine E	Coulthard D	Hill D	Frentzen H	Villeneuve J	Alesi J
Statistics							
Team	McLaren- Mercedes	Ferrari	McLaren- Mercedes	Jordan-Muger Honda	Williams- Mecachrome	Williams- Mecachrome	Sauber- Petronas
Results							
Start position	2	4	3	8	5	6	12
Result	1	2	3	4	5	6	7
Points	10	6	4	3	2	1	0

Figure 1: Wide Table Mode

Clicking at the button left of an attribute name shows the list of possible values and their frequency. Selecting a value from the list restricts the table to the objects with this value. We can obtain the same result by double-clicking a value directly in the table. Suppose we double-click the *Season 98* and *Location Suzuka* in the table. After this, only 21 objects remain in the table, showing the race of 1998 in Suzuka. However, it is still difficult to get an overview because scrolling is necessary to see all drivers.

The table can be **compressed** by clicking the second of the three mode buttons in the upper left corner of the table (above „8205 of 8205 objects“). Figure 2 shows the table after we have switched to *Compressed Mode*. Moreover, we have sorted the table by clicking on the arrow outline right of the attribute *Result*. We can now see the order in which the drivers crossed the finish line as well as their starting position.

Identical values in neighbouring cells are displayed only once. This is most obvious in the attributes *Season*, *Date*, and *Location*. Whenever the two drivers of a team finished next to each other, the name of the team is displayed only once. The 0 for attribute *Points* is also written only once. Numeric values are still represented by a horizontal line if the cell is too small to display a number.

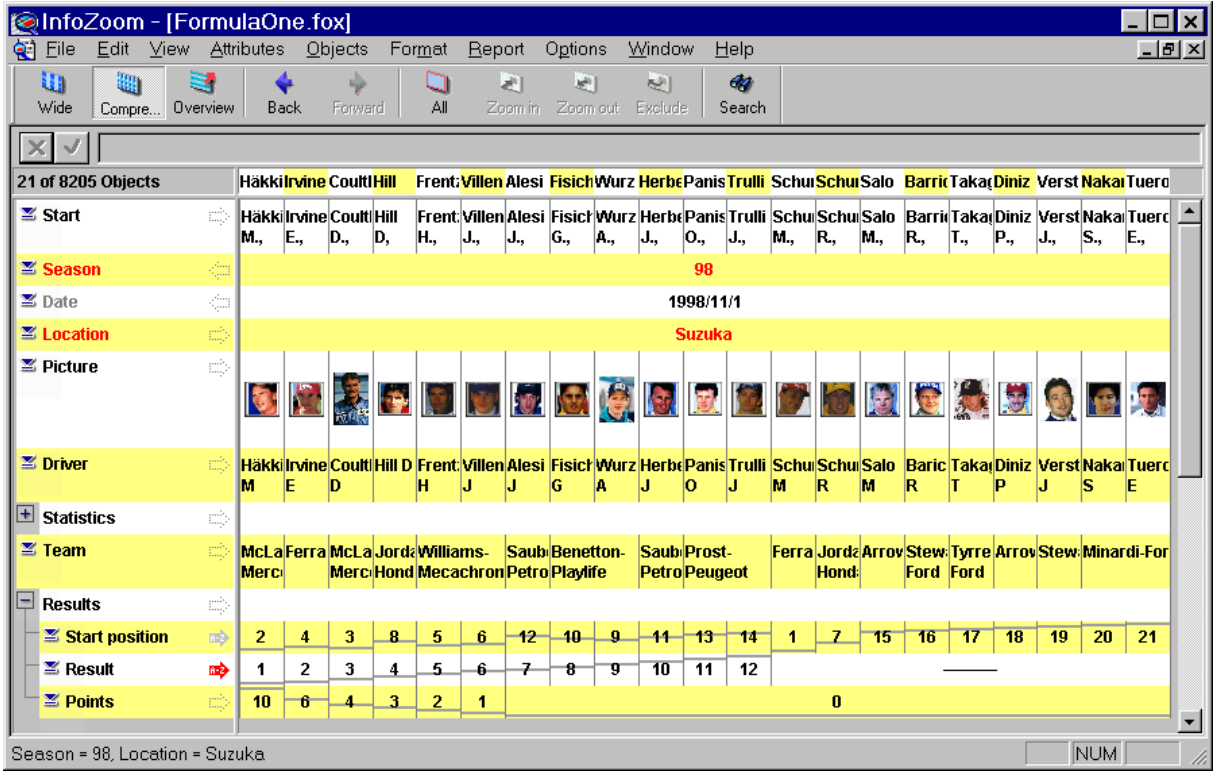


Figure 2: Compressed Table Mode

In *Compressed Mode*, the column width is always reduced until all the objects fit on the screen. In Figure 2 the column width is still about 30 pixels. In Figure 3, however, where the whole table is shown, each column only has about 0.1 pixels.

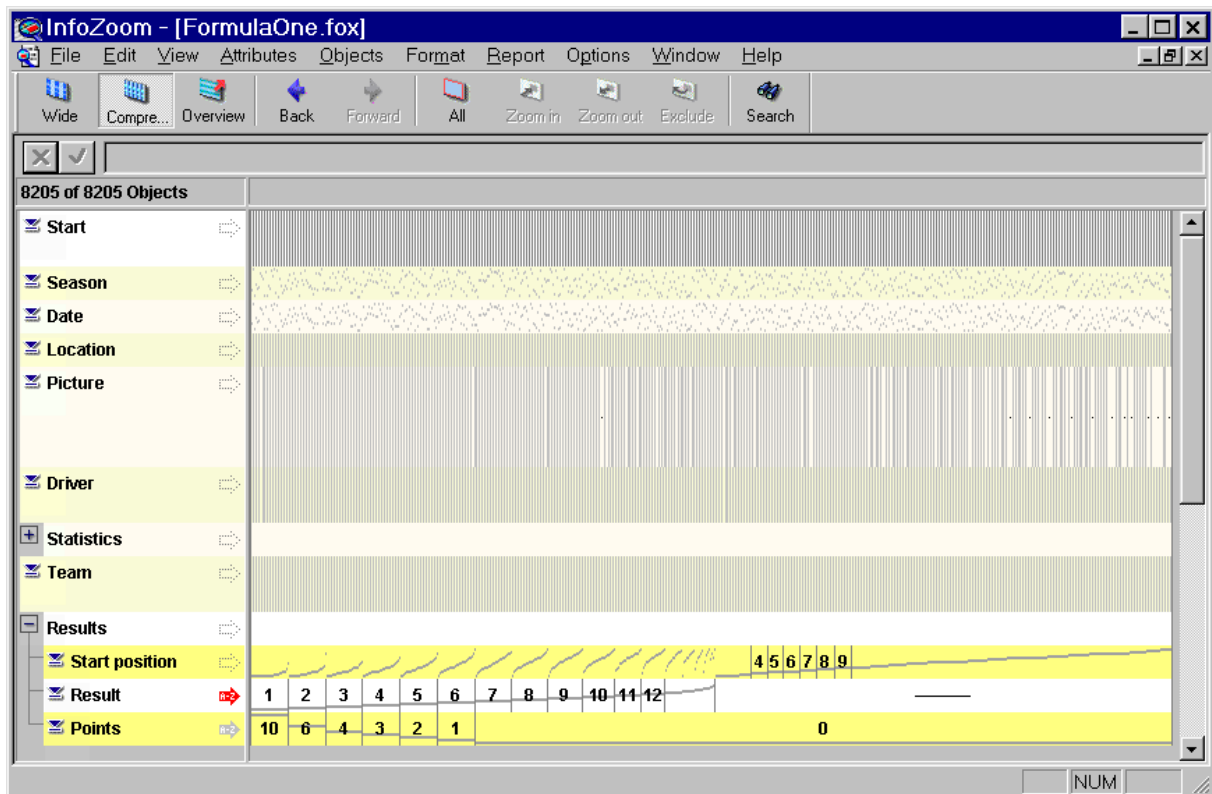


Figure 3: The full table in Compressed Mode

As a consequence, large parts of the table are unreadable. However, because the table is sorted by the attribute *Result*, most of the results are readable. The width of each cell indicates the number of subsequent objects with this value. In this way, it can be seen that higher results are less frequent and that about half of the drivers failed to finish.

It is also possible to observe correlations between the last three attributes: We can see how points are granted to the top six finishers who typically also had a good start position, as can be seen from the shape of the curve above each result.

If we double-click a value or a value-range in the compressed table, the clicked cells grow in a short animation, while the others shrink. Finally only the clicked values remain. This looks like **zooming** into the table.

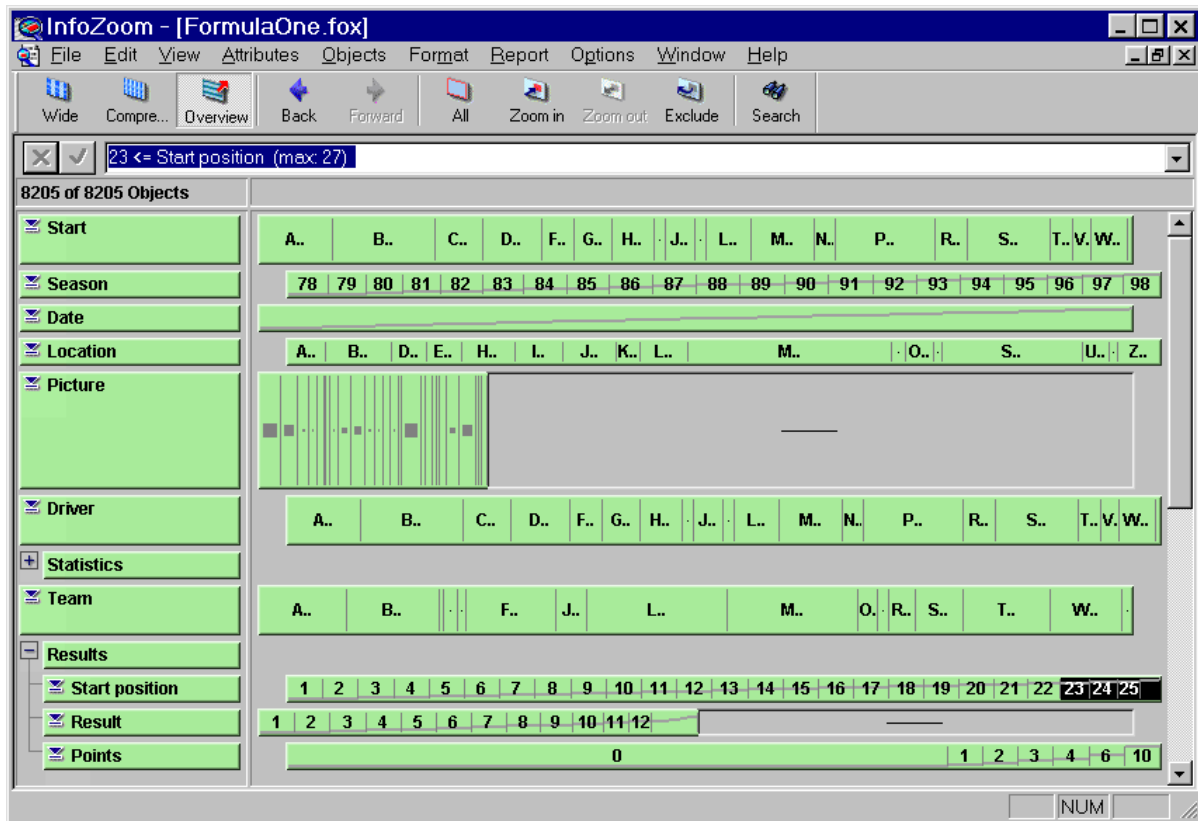


Figure 4: Overview Mode

In *Overview Mode* the values of each attribute are sorted independently (Figure 4). Thus most of the attributes are readable. It is important to understand, however, that this is not a table but something like a bar chart displaying the value distribution for each attribute. Correlations between attributes cannot be observed directly but only by watching the value distributions change during a zoom-operation.

Exploring the data set

In the *Overview Mode* (Figure 4) we can make the following observations:

- The database contains races of the seasons 1978 to 1998. There is roughly the same number of starts in each season.
- Pictures are available only for a part of the drivers.
- We can see starting positions 1 to 25. Higher values are only represented by horizontal lines because they do not fit into the cells. These values are less frequent. As the values from 23 to the right have been selected, the value range is shown above the table. There we can see that 27 is the maximum starting position. Zooming-in on the selected values (by a double-click) would show that only in 7 races there were 27 starters.
- Only rarely did more than 12 cars reach the finish line. About half of the cars failed to finish.
- Drivers can gain 1,2,3,4,6, or 10 points.

Next, we zoom-in on Michael Schumacher's races by first double-clicking *S..* to see all the drivers who's name starts with *S*, and then double-clicking *Schumacher M* (or his picture). As a consequence all value distributions change and now show only facts about Schumacher's career.

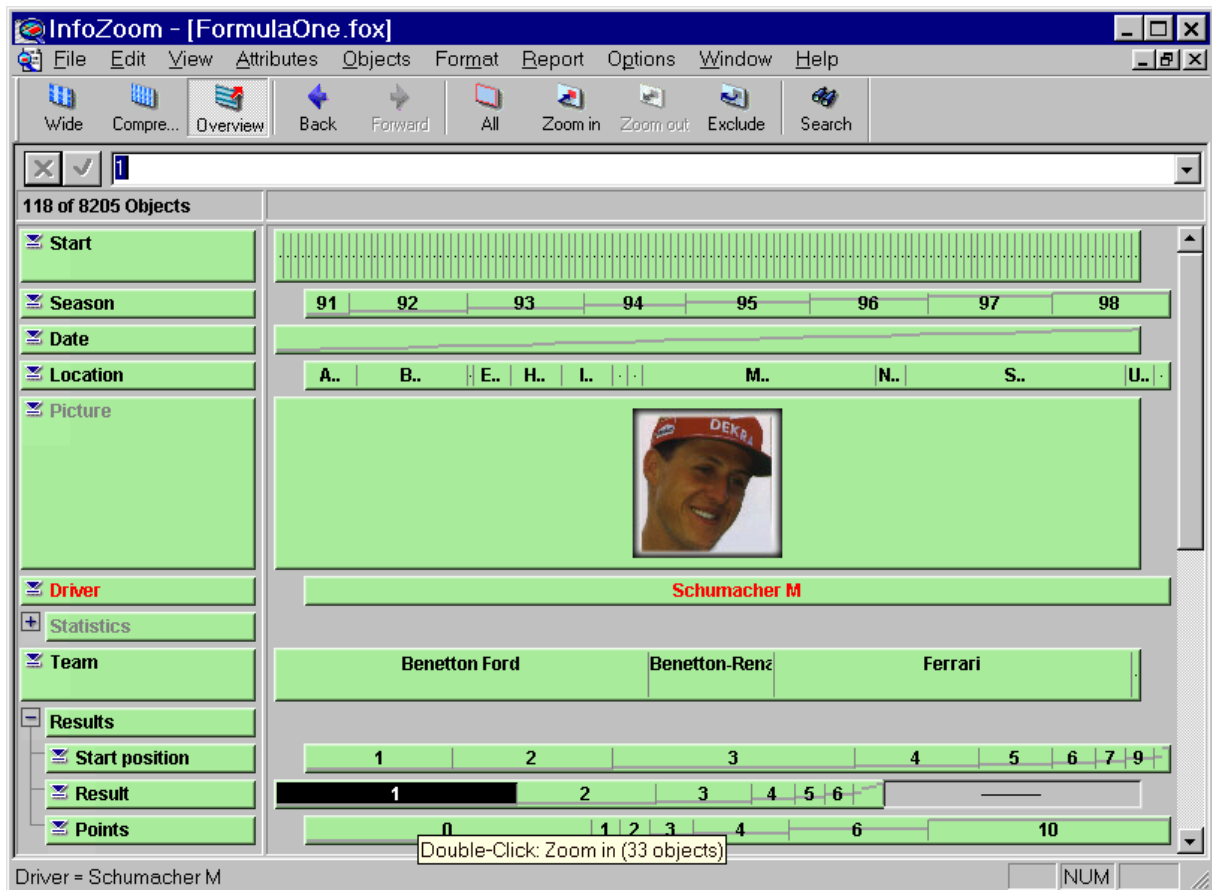


Figure 5: All races of Michael Schumacher

By inspecting Figure 5 we can make the following observations:

- Michael Schumacher participated in 118 races.
- He started his Formula One career in 1991 but had comparatively few races in that year.
- He drove for *Benetton-Ford*, *Benetton-Renault*, *Ferrari* and also one race for *Jordan Ford*. (*Jordan Ford* can only be seen by positioning the mouse over the small cell on the right.)
- We can see that 3 is Schumacher's most frequent starting position, and positions worse than 9 are extremely seldom.
- Schumacher won each third or fourth race and could not finish the race in about 30% of the cases.

We further zoom-in on the **victories** of Michael Schumacher by simply double-clicking the value 1 for the attribute *Result*. The result is shown in Figure 6.

We can observe the following:

- Michael Schumacher had 33 victories in his career.
- The first victory was in 1992.
- *Magny Cours* and *Spa-Francorchamps* are his favourite courses.
- He won the most races in 1994 and 1995 when he became World Champion.
- He was most successful when at *Ferrari*.
- He won most races from starting position 2. He once started on position 16 and still became winner.

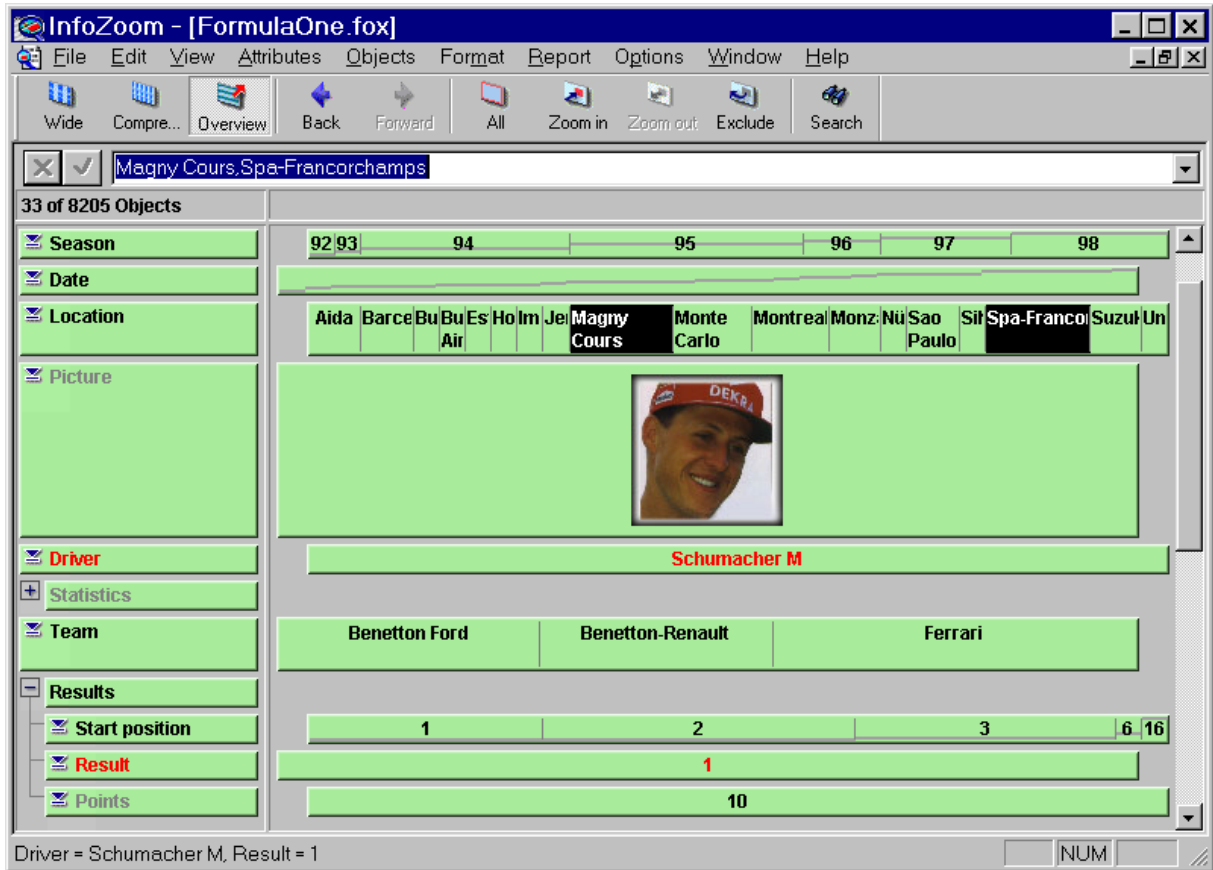


Figure 6: The victories of Michael Schumacher

Starting in the overview of the whole table again, we can display all victories (of all drivers) simply double-clicking the *1* for *Result*. As expected most winners started from the pole position or at least in the first rows of the starting grid.

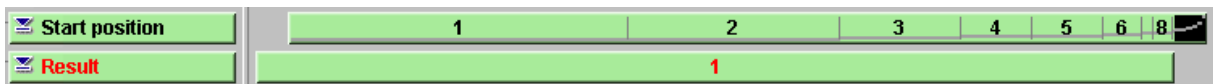


Figure 7: Starting positions of the later winners




Date	1990/6/24	1996/5/19	1995/8/27	1982/6/6	1983/3/27
Location	Mexico City	Monte Carlo	Spa-Francorchamps	Detroit	Long Beach
Picture				—	
Driver	Prost A	Panis O	Schumacher M	Watson J	
Team	Ferrari	Ligier-Mugen-Hond	Benetton-Renault	Mc Laren Ford	
Start position	13	14	16	17	22
Result			1		

Figure 8: Winners from the worst starting positions

In some rare cases drivers with a very bad starting position still won the race. To find these, we can simply select the rightmost starting positions (Figure 7) and double-click into the selection. This initiates a zoom-in which leads to the result shown in Figure 8. We can see that John Watson gained the most astonishing victory from starting position 22.

Derived attributes

Similar to the formulas in a spreadsheet program, derived attributes can be defined in InfoZoom. In our example database two derived attributes have been defined: *Sum(Points) per Driver* and *Count(Start) per Driver*. Such summary attributes correspond to queries using the *group by* operator in SQL. The two attributes can be easily used to zoom-in on the drivers who gained the most points in their careers or the drivers with the most starts.

Derived attributes are automatically recomputed by InfoZoom whenever a zoom is performed. Thus, if we display all victories as described above, the attribute *Count(Start) per Driver* is updated and shows the number of **victories** for each driver. Similarly we can display the number of pole positions, crashes or podium finishes of each driver.

Picture										
Driver	Schumacher	Fisichella G	Wurz A	Frentzen H	Hill D	Villeneuve	Irvine E	Coulthard D	Schumacher	Häkkinen M
Statistics										
Sum(Points) per Driver	14	16	17	20	21	47	56	86	100	
Count(Start) per Driver	16									
Team	Jordan-Mugen	Benetton-Playlife	Williams-Mecachron	Jordan-Mugen	Williams-Mecachron	Ferrari	McLaren-Mercedes	Ferrari	McLaren-Mercedes	McLaren-Mercedes

Figure 9: Final standings of season 1998

If we zoom-in on a certain year, only the points of this year are added for each driver. If we then sort by *Sum(Points) per Driver*, we see the final standing of the world championships for the selected year. In Figure 9 the ten best drivers of the season 1998 are shown.

In order to look for a correlation of the two derived attributes, we start from the complete table in *Compressed Mode* and subsequently sort by the two attributes.



Figure 10: Correlation of points and number of races per driver

We observe the general trend that drivers with more starts also have collected more points. But we can also observe exceptions: Some drivers have a relatively high number of points in comparison to other drivers with a similar number of races. These are selected in Figure 10. A double-click into the selection zooms-in on these five drivers. Figure 11 shows the result. As we can see, these are indeed the most successful drivers of the last years.






Picture					
Driver	Hill D	Schumacher M	Senna da Silva A	Mansell N	Prost A
Statistics					
Sum(Points) per Driver	353	526	648	501	847
Count(Start) per Driver	100	118	161	187	201

Figure 11: Drivers with many points in relatively few races

If we want do the same analysis more systematically, we can define a new derived attribute $Sum(Points) \text{ per Driver} / Count(Start) \text{ per Driver}$. It turns out that *Michael Schumacher* has the highest quotient of 4.46 followed by *Alain Prost* (4.21) and *Ayrton Senna* (4.02). We could also directly define an attribute $Average(Points) \text{ per Driver}$ and obtain the same results.

Using derived attributes many interesting results can be obtained. Here is only a short list:

- By defining $Maximum(Result) \text{ per Date}$ we can see that in 1996 in *Monte Carlo* only four drivers finished the race.
- $Average(Maximum(Result) \text{ per Date}) \text{ per Location}$ shows the courses where the least drivers finished the race on average.
- $Minimum(Result) \text{ per Driver}$ shows the best result for each driver.
- $Count(Team) \text{ per Driver}$ shows the drivers who started for the most teams.
- Zooming-in on starting positions 1 and 2, and then on the cases where $Count(Team) \text{ per Date}$ equals 1 shows which teams had the most double pole positions.

Team	McLaren Mercedes	Ferrari	Renault	McLaren Honda	Williams-Renault
Count(Date) per Team	9	13	17	34	46
Count(Team) per Date	1				
Results					
Start position	1 2	1 2	1 2	1 2	1 2

Figure 12: Teams with most double pole positions

Percentages

InfoZoom allows the definition of derived attributes which display percentages. In Figure 13 the results of all 100 starts of Damon Hill are displayed. For each result, not only the frequency has been counted but also the percentage of total number of starts has been computed.

Result		1	2	4	8	3	6	7	9
Count(Start) per Result	30	22	15	6	5	3	2	1	
Percent(Start) per Result	30.0%	22.0%	15.0%	6.0%	5.0%	2.0%	1.0%		

Figure 13: Most frequent results of Damon Hill

It is also possible to compute the percentage of objects currently displayed in comparison to the overall number of objects in the database. If we define the attribute $\%displayed(Start) \text{ per}$

Driver and all objects are displayed, *100%* is computed for each driver. In Figure 14 we have already zoomed-in on the victories only. Now, the derived attribute shows the percentage of victories for each driver.

Driver	Mansell N	Hill D	Villeneuve J	Prost A	Senna da Silva A	Schumacher M
% displayed(Start) per Driver	16.6%	22.0%	22.4%	25.4%	25.5%	28.0%

Figure 14: Drivers with highest percentage of victories

By simply zooming-in on other subsets we can determine the percentage of podium finishes, pole positions, or non-finished races of each driver.

Conclusion

We have shown the interactive techniques for visual data mining supported by InfoZoom. The goal of our approach is not a completely automatic algorithm that searches for interesting results. Instead, InfoZoom enables the user to interactively explore the data set and to get a feeling of the contained information. It introduces a novel visualisation technique which displays the whole data set on a single screen. Queries are simply performed by selecting parts of the displayed data. Derived attributes can be defined like in a spreadsheet program and are automatically updated when necessary. We are convinced that using InfoZoom is simple enough to be used by domain experts in order to understand their data and to detect the hidden knowledge.

Even if an automatic data mining algorithm is used, a tool like InfoZoom is important for getting a first impression of the data set, checking the value ranges, and finding typos, inconsistencies, and missing values. Moreover, defining the right derived attributes is a very important step before running the mining tool.

InfoZoom lets the user perform queries very easily and quickly, but a systematic search of all possible queries is usually too time consuming. Therefore, we plan to integrate a data mining algorithm into InfoZoom which leads the user to interesting states of the table.

InfoZoom was initially developed at GMD – the German National Research Center for Information Technology. It is now extended and marketed by the GMD spin-off company humanIT [1].

References

- [1] <http://www.humanIT.de> – The InfoZoom home page. A free test version of InfoZoom can be obtained. [2] is available online.
- [2] Spenke, M.; Beilken, Chr.; Berlage, Th., FOCUS: The Interactive Table for Product Comparison and Selection, Proceedings of the UIST 96 Ninth Annual Symposium on User Interface Software and Technology, Seattle, November 6 - 8, 1996. ACM 1996, pp. 41 – 50.

[3] <http://www.formel-eins.de>, <http://www.rtlnews.de> – The RTL Formula One online database.