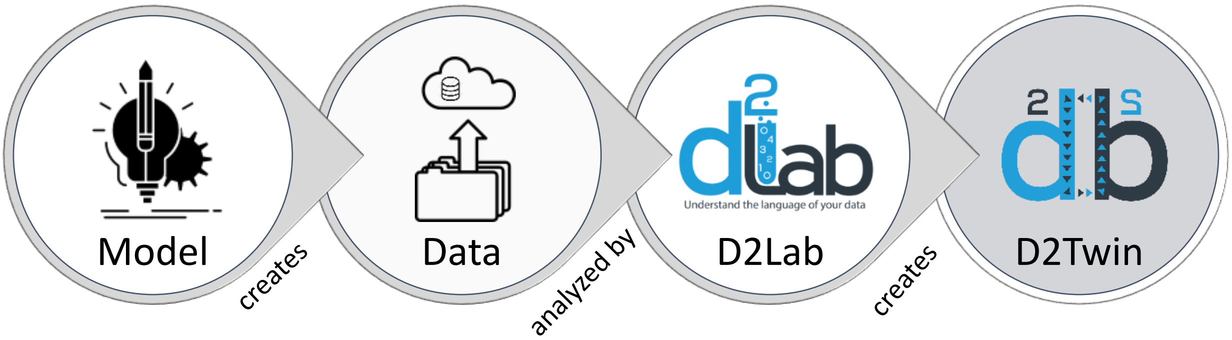
D2Twin

We are introducing Data Diagnostic Laboratory, the new generation of big data analytics platforms. It is designed as a service for creating values by processing and analyzing a huge amount of data.

D2Twin is a data-driven "replica" of an asset (product, system or process). It is based on the analysis of the past behavior of the asset. It is derived from past data using advanced data analytics techniques. D2Twin extends the concept of the Digital Twin from the point of view of how the asset has been used in the past. Analyzing the data generated by a concrete asset in the past operation, our approach derives the knowledge about the model of the asset. Traditionally, Digital Twin is the result of human expert-driven modeling, independently of its usage. Additionally, D2Twin is a digital replica of a concrete artifact and not the model of a type of assets (e.g. type of machine).



D2Twin supports all types of analytics:

Observe
Descriptive Analytics
- What did happen or is happening

Analyze
Diagnostic Analytics
- Why did it happen or is it happening?

Predict
Predictive Analytics
- What is likely to happen?

Influence
Prescriptive Analytics
- How can we make it happen and what are the consequences?

The key advantages of D2Twin include:

- Applicability - it can be employed in any domain where the data is available
- Scalability - model learning can be applied on high-dimensional process space
- Complexity - learning is unsupervised so that no human expert involvement is needed
- Up-to-date model - the new model is periodically generated so that the model is always up-to-date, compensating model drift

D2Twin Exploratory Analytics

D2Twin Exploratory Analytics is a tool made for exploring process behavior as it continuously analyses the manufacturing process and its parameters.

D2Twin Exploratory Analytics is divided into five main services:

- Behavior Analyzer
- Model Analyzer
- Parameters Analyzer
- Instances Analyzer
- Statistical Process Control (SPC) Analyzer

The dashboard of the D2Twin web service is given in Figure 1 with available sections.

Figure 1. Dashboard

Behavior Analyzer gives the visual representation of statistical summary for the generated model. The behavior of one system is explained through D2Twin model based on its data. Model is being created by analyzing the data set with the use of D2Twin analytics. It can be interpreted as the best possible grouping of similar process data from a given set which is analyzed. If the data set describes the process, groups are equivalents for process modes or variations in the process execution.

In the top-left corner, the user can choose a model by selecting the date. On the left part of the Behavior Analyzer section are statistics referring to the selected model. It includes generation time, number of clusters, number of instances, number of outliers, training configuration and parameters statistics (figure 2). Training configuration contains information about the number of iterations, iteration boundary, filter window size, maximum number of clusters and parameters used for the training. Parameters statistics shows the list of all parameters and details about a particular parameter by expanding the list item. Details for particular parameter include deviation, dispersion coefficient, mean and quartiles.

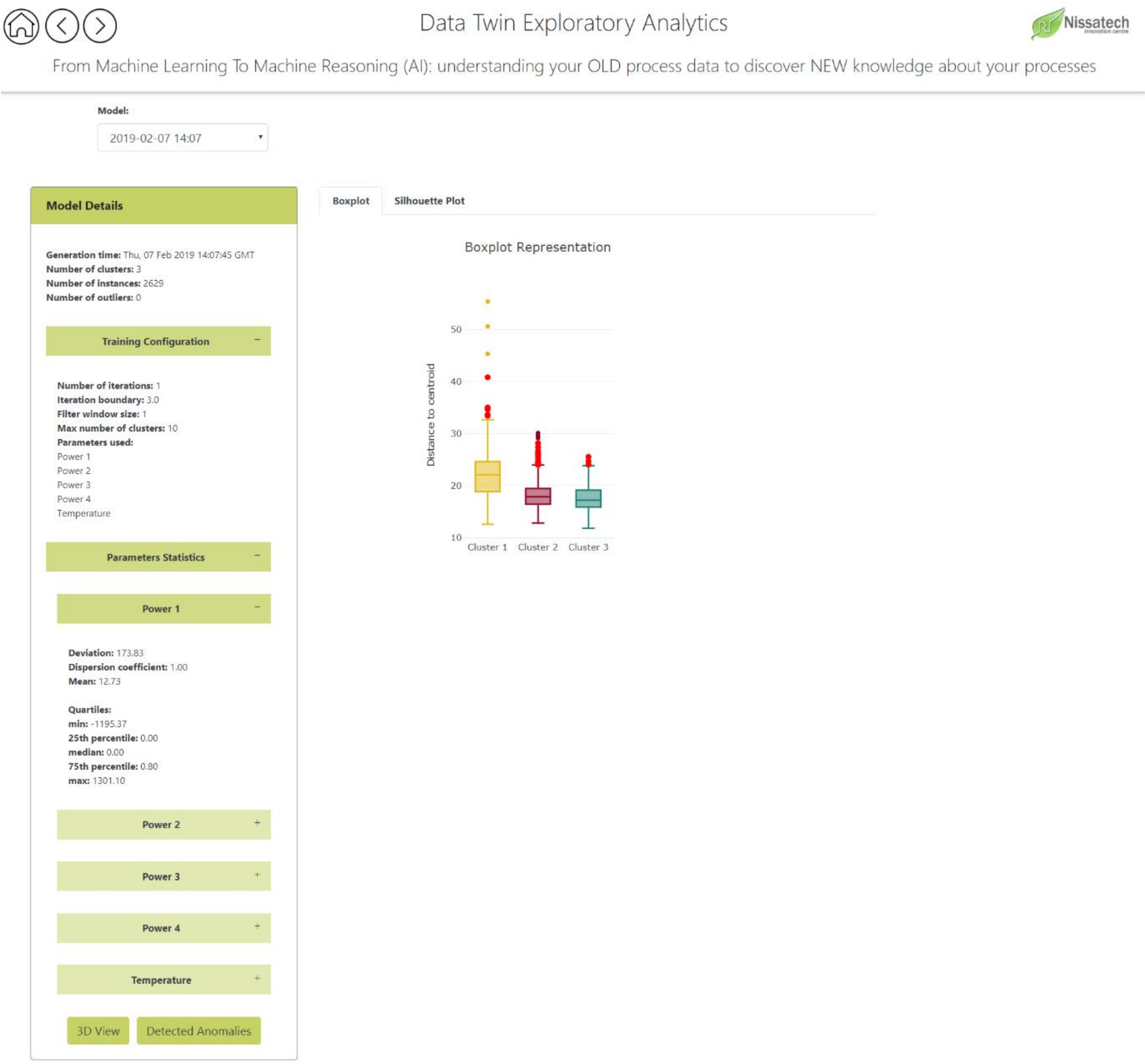


Figure 2. Behavior Analyzer

The quality of the created model and clusters can be checked with boxplots and silhouette plots.

Every boxplot is representing one cluster and displaying the distribution of data referring to distances from the centroid of the cluster. Boxplots are based on a five-number summary: minimum, first quartile (q1), median, third quartile (q3) and maximum, which are shown on boxplot hover. Boxplots can also tell about outliers and their values. They are marked as red dots on plots.

Silhouettes are another method of interpretation and validation of consistency within clusters. The silhouette ranges from -1 to +1, where a high value indicates that the instance is well matched to its cluster and far away from neighboring clusters.

Figures 3 and 4 are showing boxplots and silhouette plots respectively.

Figure 3. Boxplots

Figure 4. Silhouette plots

Model Analyzer is developed for exploring generated clusters. If there aren't any variations in the process, it consists of one cluster and can be considered a stable process. If more than one cluster is found, it is possible that these clusters describe different process execution modes, which is also normal. However, one of the clusters could represent anomalies, in which case we talk about an anomalous cluster.

In the center of the section is graphic which visualizes clusters in 3D representation (figure 5). It gives a 3D view on the data, where each dot represents an instance belonging to a particular cluster, represented with different colors. On dot hover is shown its id and date of its production, and by clicking on an instance, a user is redirected to the Instances Analyzer for a particular instance. Instances can be filtered based on their date of production by choosing a specific date from the calendar picker above the 3D chart.

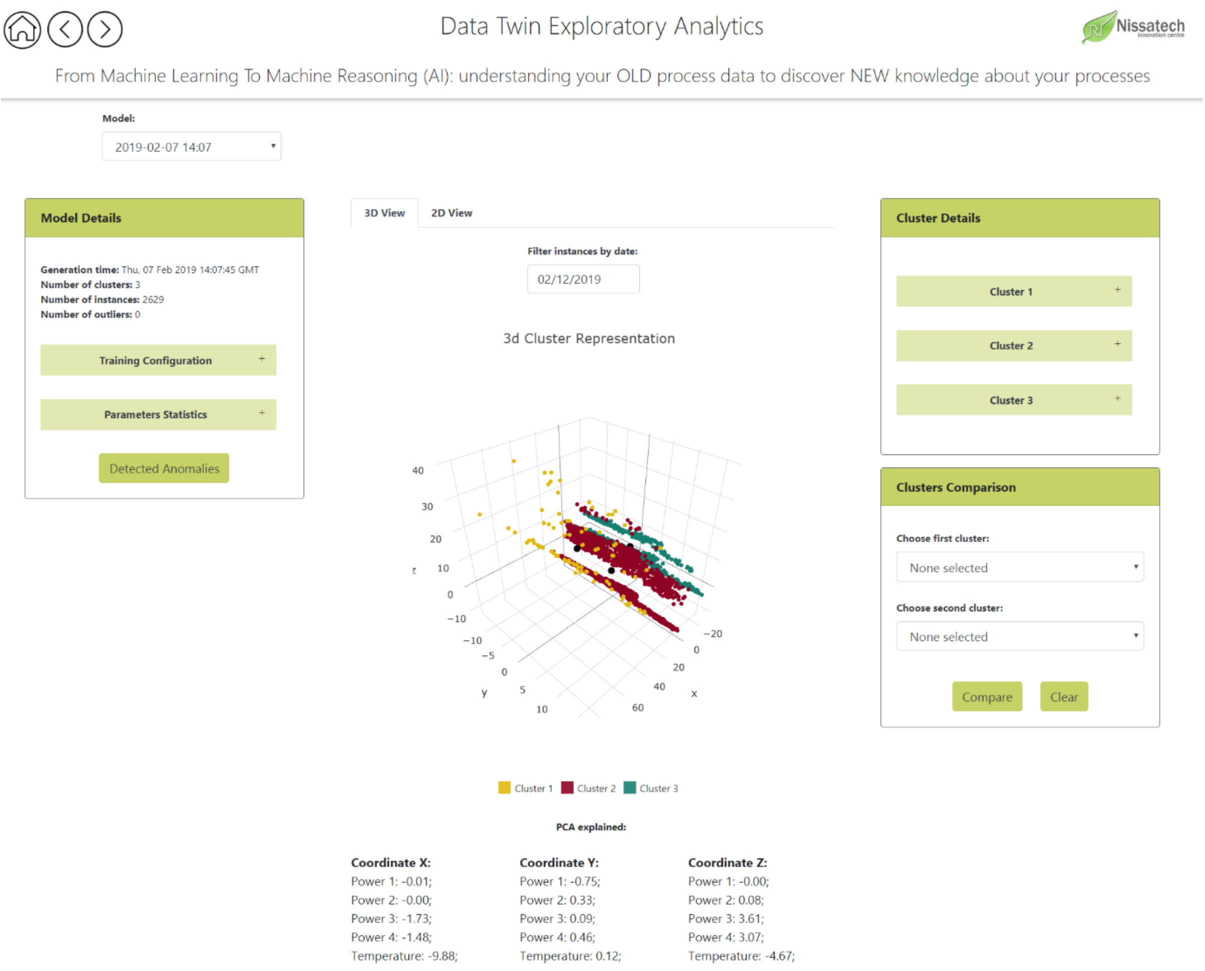


Figure 5. 3D cluster representation

Next to the tab for 3D representation is the tab which gives the 2D view of the data, where the x-axis represents time and y-axis cluster to whom particular instance belongs. On dot hover instances production date is shown. This time view visualizes the flow of the model in time (figure 6). Again, instances can be filtered based on their date of production by choosing a specific date range from the calendar picker above 2D chart.

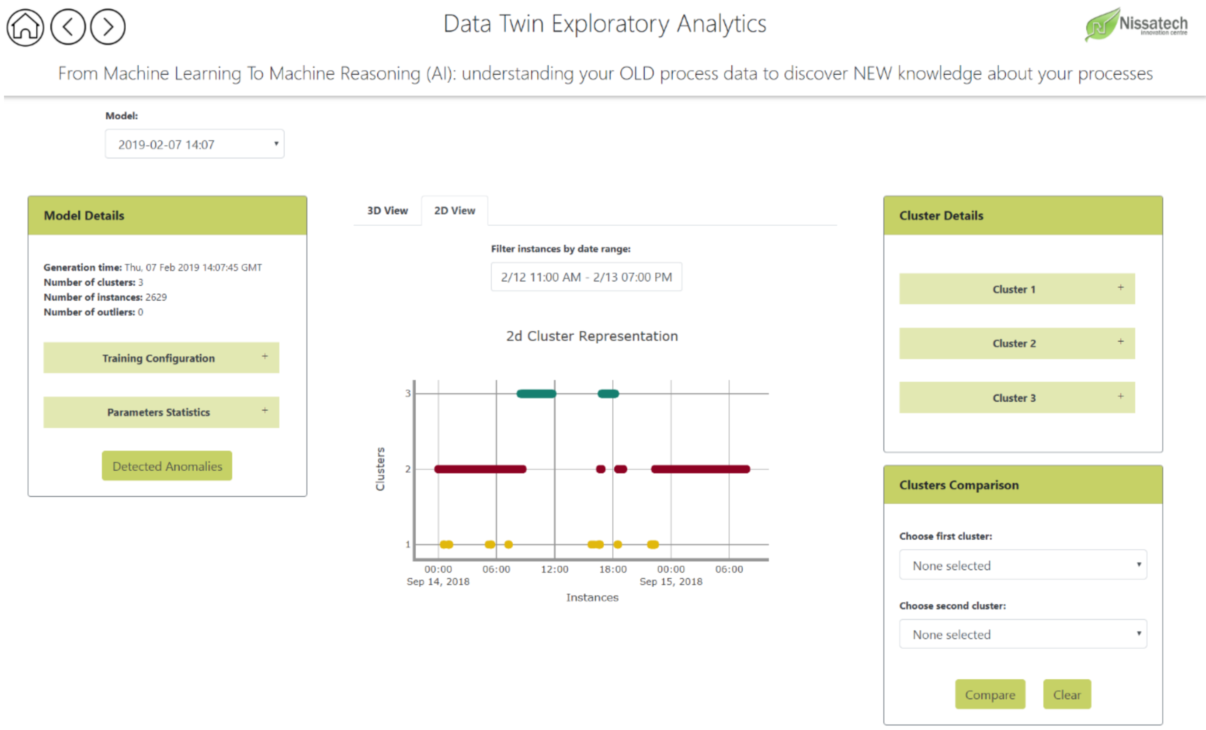


Figure 6. 2D cluster representation

On the left part are the same statistics as on the previous view, while in the right part are sections referring to cluster details and clusters comparison. Cluster details section contains information about the number of instances, coordinates of the centroid, total distance, average distance, deviation distance, cohesion, and subsections referring to stability and parameters statistics, which are given in Figure 7.

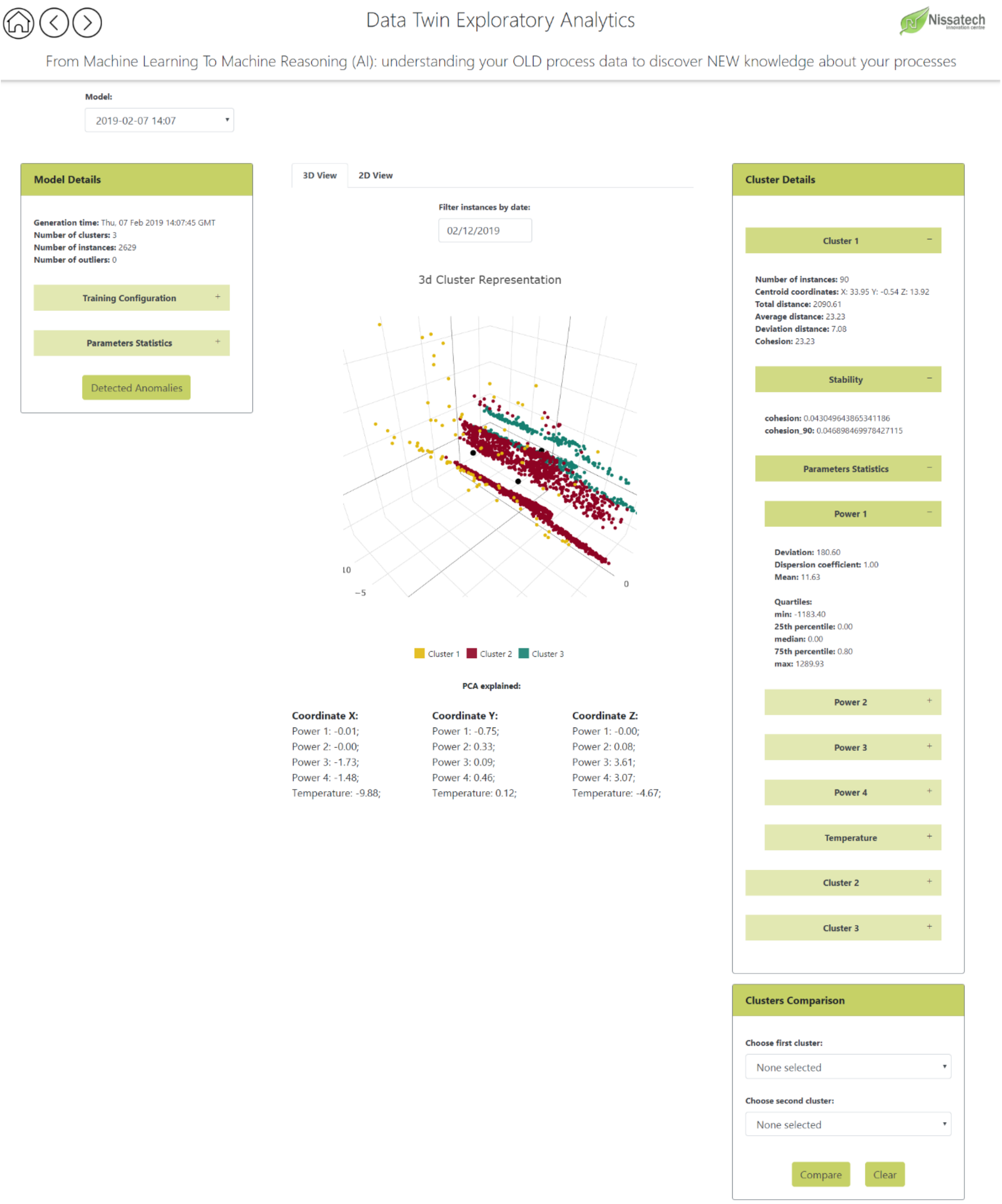


Figure 7. Cluster details

Clusters comparison section is given in Figure 8. By choosing the first and second cluster, information about quartile difference and overlap is given.

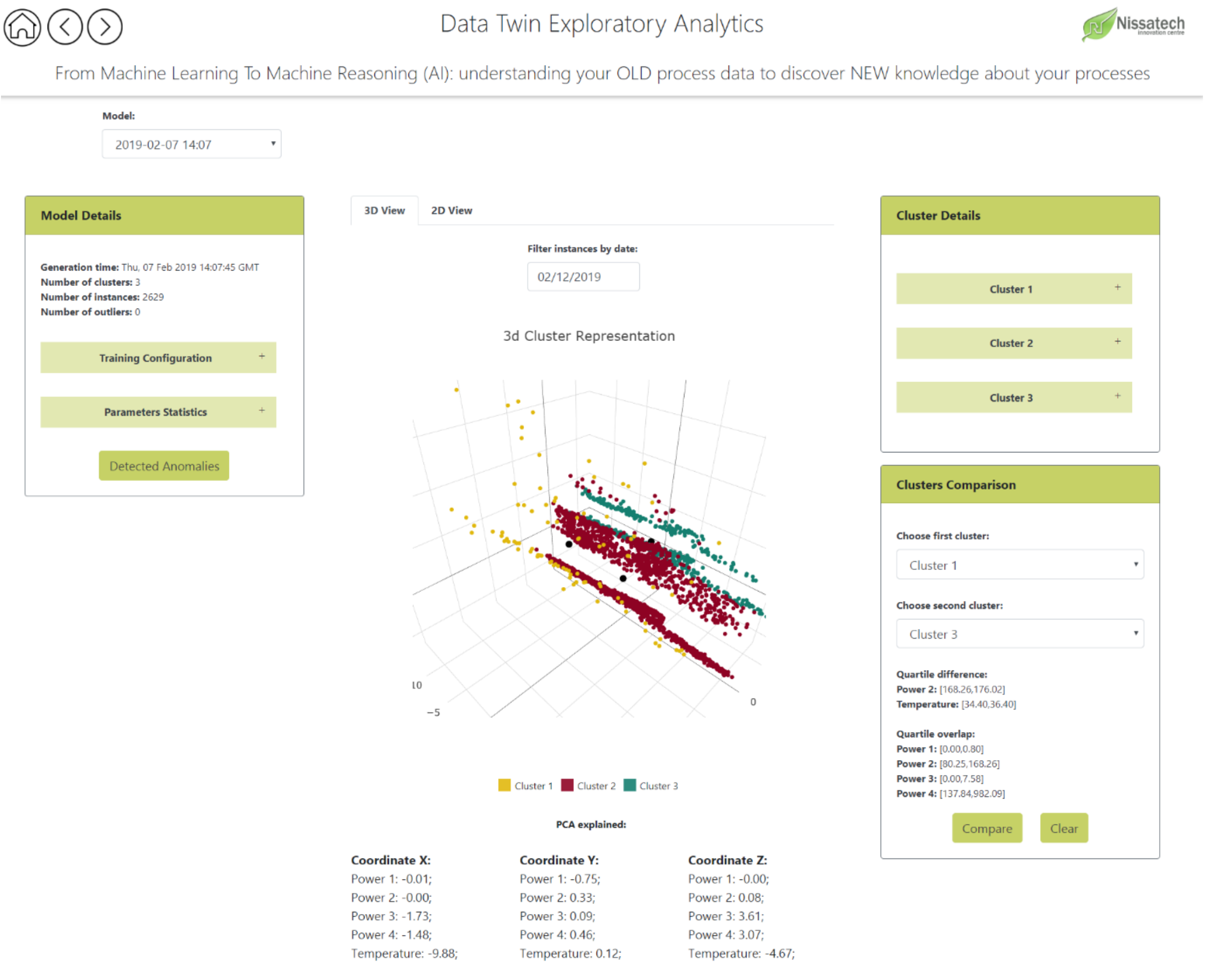


Figure 8. Clusters comparison

Parameters Analyzer is used for getting insight into parameters distribution, variations and relations between them. One process can have a lot of different parameters. They are defined as characteristics of the process which are continuously measured. This view includes four subsections, shown in separated tabs: boxplots, difference variations, parameters pairing and parameter through time.

Boxplots tab gives the overview of the distribution of parameters through differences between clusters represented with boxplots for each parameter. Standard five-number summary is shown on boxplot hover. Boxplots which are displayed can be filtered by selecting parameters in the combo box list above the boxplots (figure 9).

Figure 9. Boxplots

Difference variations tab shows, based on the two parameters chosen from the dropdown lists, differences between their standardized values for every instance belonging to the selected model. This is represented with the line chart, where x-axis corresponds to time and y-axis to subtracted values (figure 10). Each dot hover gives the pair of time and difference value.

Figure 10. Difference variations

Parameters pairing gives an overview of how synchronized are their values and to what extent specific parameters influence clusterization. It can help to determine relations between the two selected parameters, how uniformly they are related or whether it is a more heterogeneous relationship (figure 11). Each dot hover shows the production date of an instance.

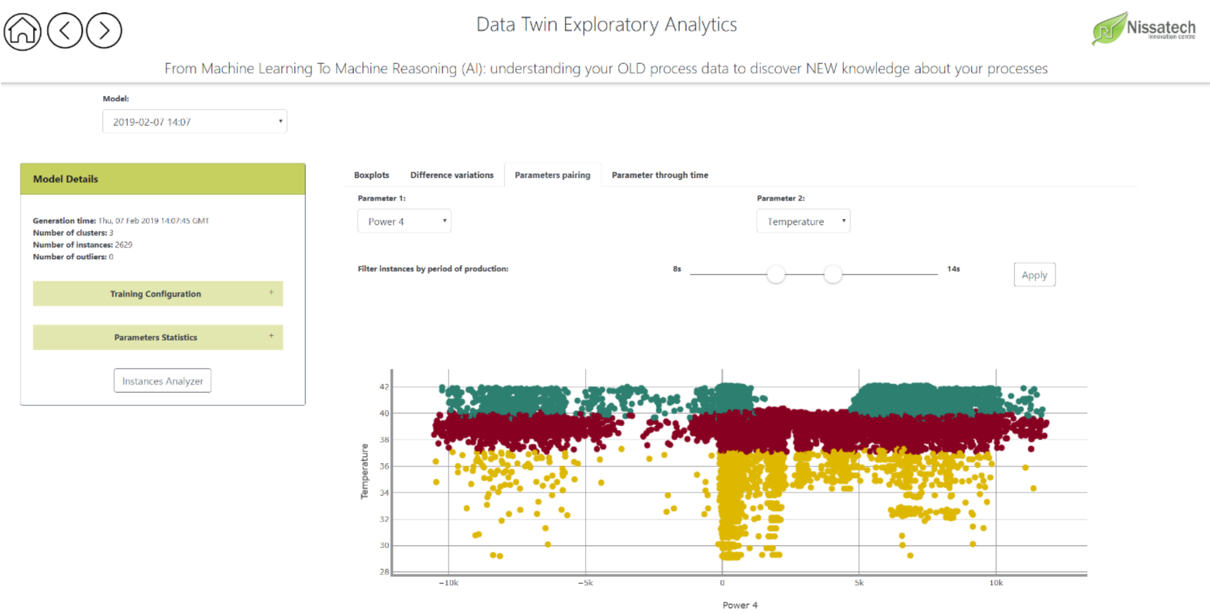


Figure 11. Parameters pairing

Parameter through time tab shows how a selected parameter has changed over time, represented with the line chart, where each dot's color is based on the matching cluster (figure 12). Again, each dot hover gives the pair of time and parameter value.

Figure 12. Parameter through time

Instances Analyzer is intended for comparing an instance or a group of selected instances with clusters representatives. The model is based on instances of a system or process. An instance is defined as a set of consecutive data points which have a semantic meaning, i.e. an instance's data is related to the process parameters during production.

By choosing a model and a specific date on the left part of the view, all instances produced on that day represented with corresponding ids are shown in the list view below. If an instance belongs to a particular cluster, its list item contains also production date and background color representing a matching cluster (figure 13).

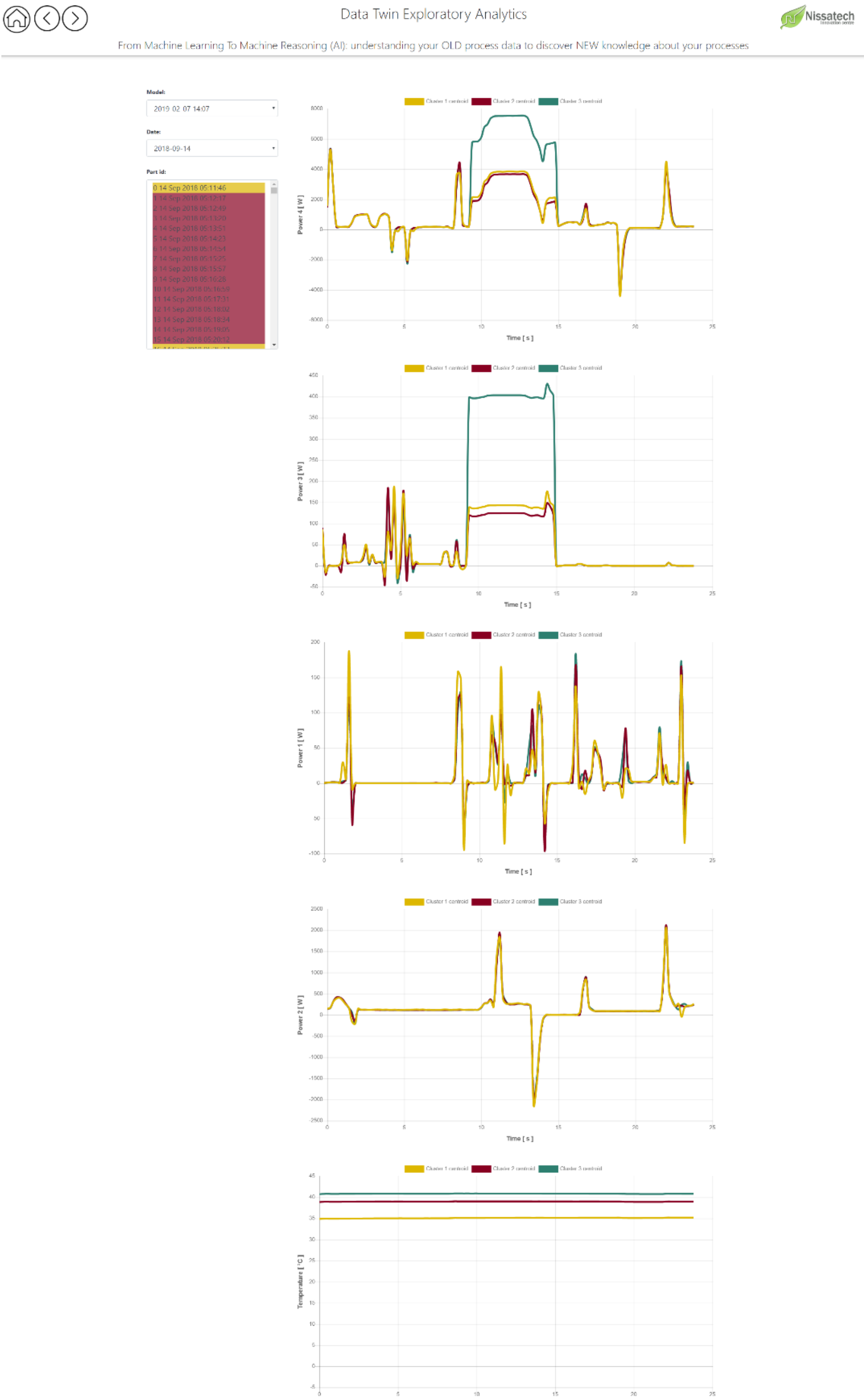


Figure 13. Instances Analyzer

By clicking on an instance from the list or selecting a group of instances, a group of graphs is shown, where the number of graphs corresponds to the number of parameters and each graph represents parameters values for chosen subset of instances and clusters centroids overtime, giving the option to compare a specific instance with clusters representatives (figure 14).

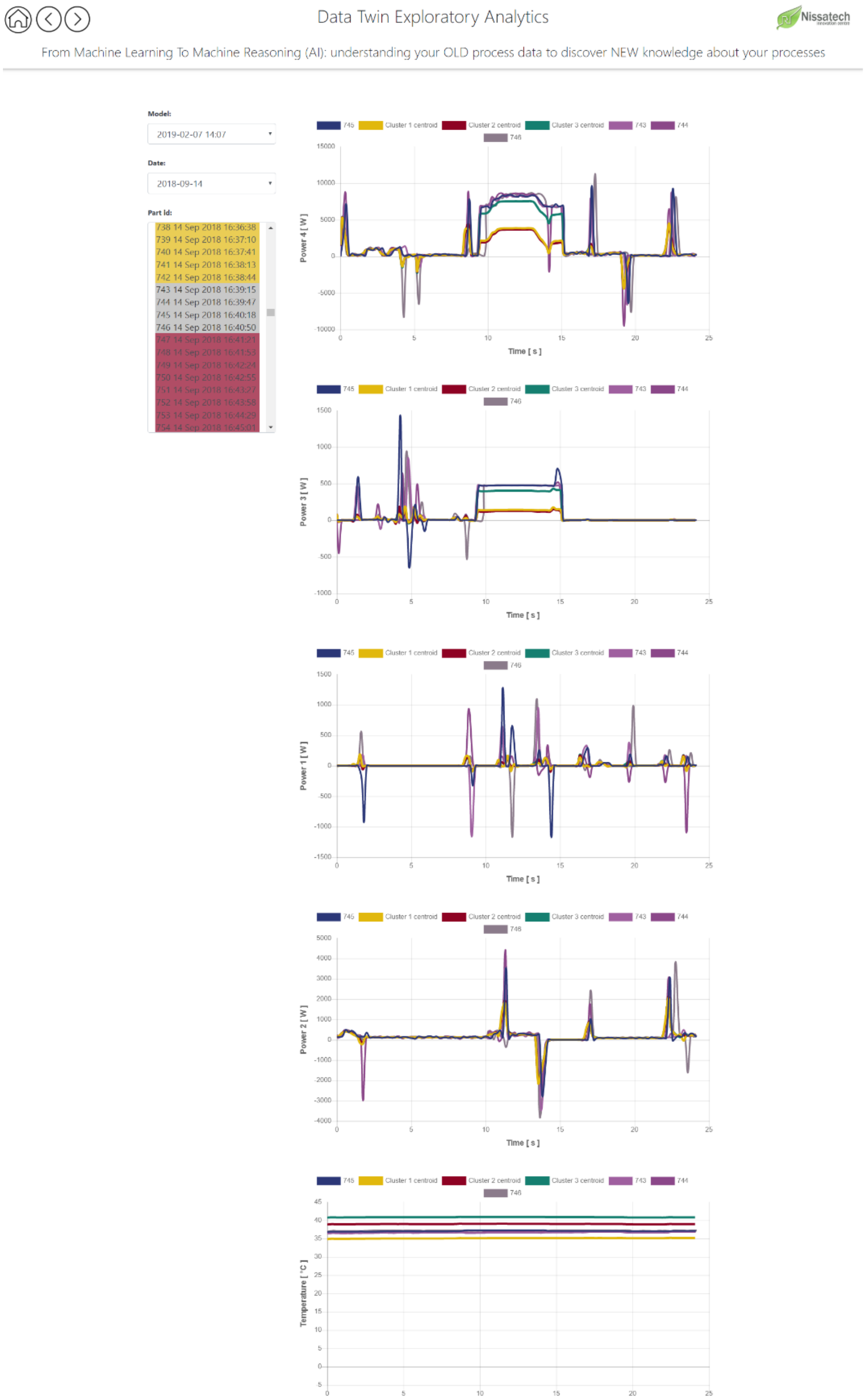


Figure 14. Instances Analyzer for a group of selected instances

This view also supports exploration of instances detected as anomalies. They are marked as ‘anomaly!’ in the list and visualized the same way as other instances (figure 15).

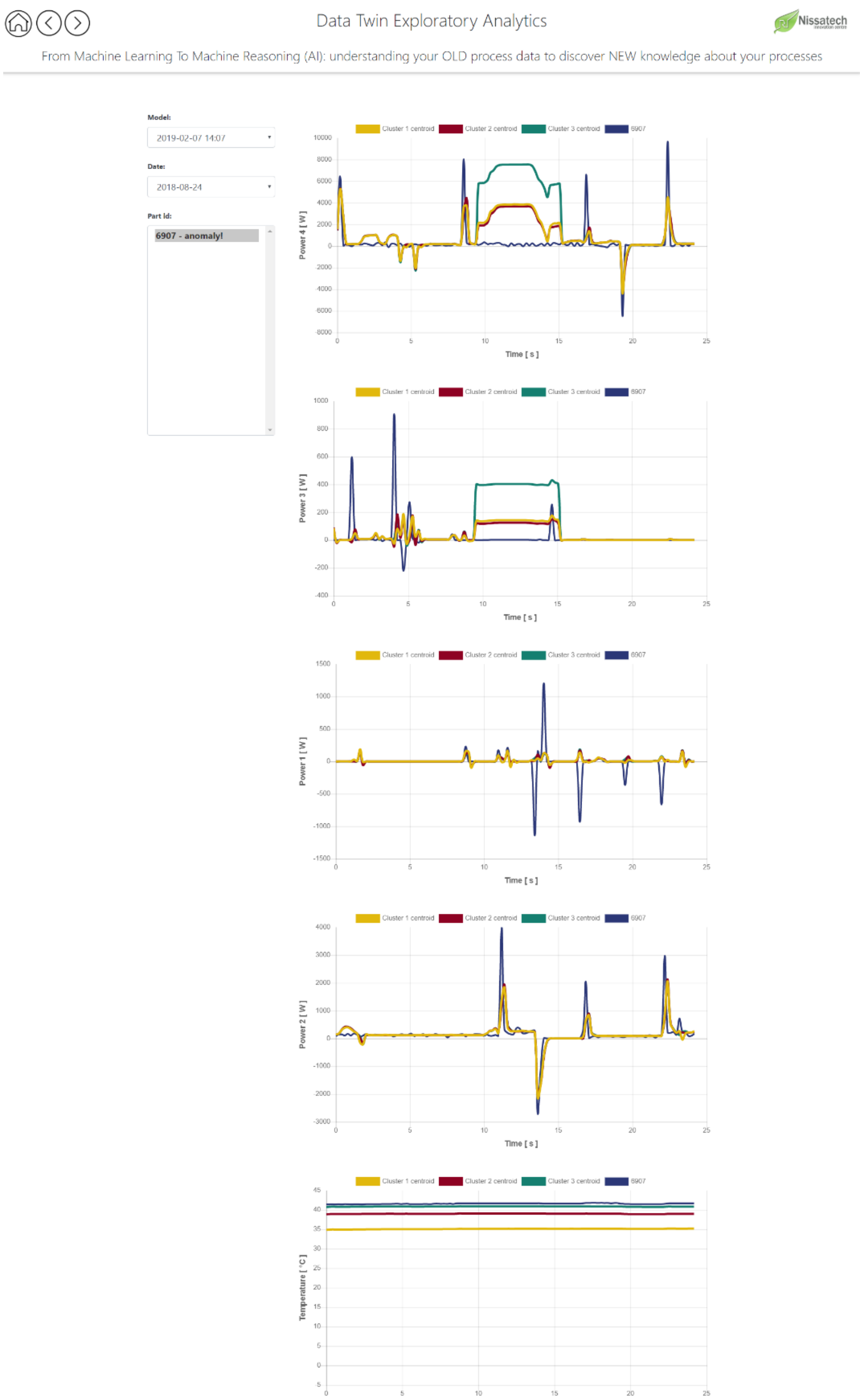


Figure 15. Instance Analyzer for detected anomalies

SPC Analyzer is a method for quality control which employs statistical methods to monitor process stability and control. This view includes four subsections for the selected parameter and date or time range: x-bar and range control charts, histogram and statistics for control and specification measurements (figure 16).

Both x-bar and range control charts are represented as time series graphs, with the central line (grand mean) as a visual reference for detecting shifts or trends in the process and control limits computed from available data and placed equidistant from the central line. Because control limits are calculated from process data, they are independent of customer expectations or specification limits. Determining control limits can be used to monitor whether the process is out-of-control. Instances which are outside these established control limits are marked on charts as violation points. The x-bar control chart is used to monitor the mean of parameter values based on samples taken from the process. The consistency of process averages is evaluated with the x-bar chart, whereas range control chart evaluates the consistency of process variations.

The histogram gives the information about the distribution of parameter values, represented as a bar chart where the height of each bar represents the number of instances falling within a range of parameter values.

Statistics for control and specification measurements includes upper and lower specification limits and its target value, upper and lower control limits for both x-bar and range and their average values and measurements describing the capability of the process.

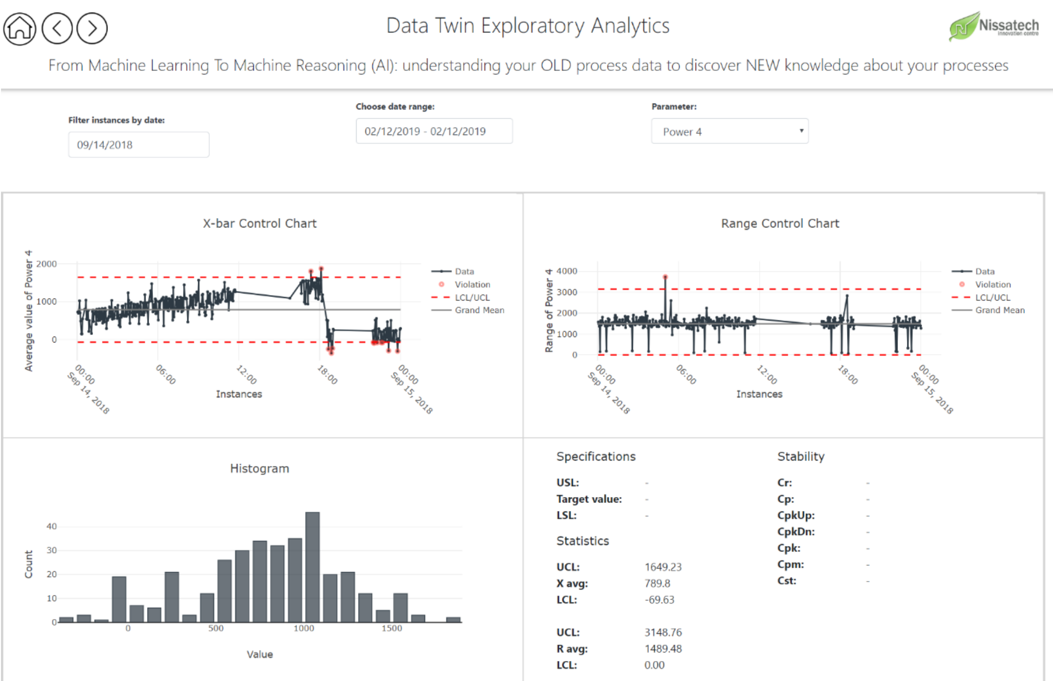


Figure 16. SPC Analyzer