A Visual Analytics Approach for the Diagnosis of Heterogeneous and Multidimensional Machine Maintenance Data

Xiaoyu Zhang* University of California, Davis
Takanori Fujiwara* University of California, Davis
Senthil Chandrasegaran† Delft University of Technology
Michael P. Brundage‡ National Institute of Standards and Technology
Thurston Sexton‡ National Institute of Standards and Technology
Alden Dima‡ National Institute of Standards and Technology
Kwan-Liu Ma* University of California, Davis

ABSTRACT

Analysis of large, high-dimensional, and heterogeneous datasets is challenging as no one technique is suitable for visualizing and clustering such data in order to make sense of the underlying information. For instance, heterogeneous logs detailing machine repair and maintenance in an organization often need to be analyzed to diagnose errors and identify abnormal patterns, formalize root-cause analyses, and plan preventive maintenance. Such real-world datasets are also beset by issues such as inconsistent and/or missing entries. To conduct an effective diagnosis, it is important to extract and understand patterns from the data with support from analytic algorithms (e.g., finding that certain kinds of machine complaints occur more in the summer) while involving the human-in-the-loop. To address these challenges, we adopt existing techniques for dimensionality reduction (DR) and clustering of numerical, categorical, and text data dimensions, and introduce a visual analytics approach that uses multiple coordinated views to connect DR + clustering results across each kind of the data dimension stated. To help analysts label the clusters, each clustering view is supplemented with techniques and visualizations that contrast a cluster of interest with the rest of the dataset. Our approach assists analysts to make sense of machine maintenance logs and their errors. Then the gained insights help them carry out preventive maintenance. We illustrate and evaluate our approach through use cases and expert studies respectively, and discuss generalization of the approach to other heterogeneous data.

Keywords: Visual analytics, heterogeneous data, high-dimensional data, machine learning, text analytics, maintenance logs.

1 INTRODUCTION

Making sense of large-scale, heterogeneous data is one of the main challenges faced by data science and visualization communities in real-world application scenarios. For instance, in large-scale manufacturing setups, human- and machine-created logs of operation and maintenance need to be analyzed to identify problem areas and prevent major failures before they occur [8]. These logs can easily number over hundreds of thousands of records and often include multiple types of data: numerical data (e.g., operating temperatures), categorical data (e.g., machine types), ordinal data (e.g., error severity), and text data (e.g., machine status description) [26]. In addition, such logs also feature manual entries—including natural-language descriptions—which are prone to inconsistencies, such as the same problem described differently at different times or by different people [44]. These factors make it difficult for managers and technicians—even with the help of data analysts—to analyze logs to identify patterns (e.g., common phenomena seen in some type of errors) and perform preventive maintenance. While such issues are common in maintenance analysis and prognostics, the challenge of heterogeneous and inconsistent data spans domains.

Machine learning (ML) assisted visual analytics has been developed to address the challenge in reviewing large, high-dimensional data [42, 52]. For instance, researchers have used dimensionality reduction (DR) to provide an overview of high-dimensional data in lower dimensions [28, 54] and clustering to summarize the information of large data into a small number of groups [3, 29]. Contrastive learning, which extracts salient patterns in one dataset relative to the other, is then used to help interpret the results of DR and clustering [20, 22]. Such approaches can help maintenance log analysts extract and explain important patterns specific to certain kinds of issues, while data inconsistency can be mitigated by keeping the human in the loop. However, these ML methods are designed to apply to a single datatype, such as numerical or categorical. Thus, when analyzing heterogeneous data, we need to consolidate different methods. In addition, existing contrastive learning methods are applicable only to either numerical or binary data. New methods for other datatypes (e.g., categorical and text) are needed.

In this paper, we present an approach to separate different variable types—numerical, categorical, and text—in a heterogeneous dataset and provide lower-dimensional, clustered visualizations for each type. We then use ccPCA [20]—contrasting clusters in Principal Component Analysis—as the contrastive learning method for *numerical* variables in the data. To provide a similar functionality for *categorical* variables, we introduce a method called contrasting clusters in Multiple Correspondence Analysis (ccMCA). ccMCA helps characterize a selected cluster (of categorical data) by comparing its attributes with those of the remaining data. For *text* variables, we first convert natural-language descriptions into high-dimensional vectors using word embeddings [48], and then perform DR and clustering. In place of contrastive learning, we plot text frequencies compare each cluster with the rest of the data.

Finally, we link the visualizations across all the views to help the analyst characterize clusters in the context of the other data dimensions. We illustrate our approach with use-case scenarios and expert reviews using a real-world dataset of maintenance and repair logs for heating, ventilation, and air-conditioning (HVAC) systems.

Our main contributions include: (1) integrating existing DR and clustering techniques to make sense of multidimensional, heterogeneous maintenance log data by introducing a visual analytics approach to coordinate views resulting from these techniques for each datatype, (2) introducing a new contrastive learning method called ccMCA to help the user characterize data clustered on the basis of categorical dimensions, and (3) illustrating the use of domain knowledge to characterize the clustered data.

2 RELATED WORK

While the proposed work falls under the application area of machine maintenance data analysis, our approach draws from and contributes to existing approaches in heterogeneous and high-dimensional data. We highlight representative research on these topics in this section.

*e-mail: {xybzhang, tfujiwara, klma}@ucdavis.edu
†e-mail: r.s.k.chandrasegaran@tudelft.nl
‡e-mail: {michael.brundage, thurston.sexton, alden.dima}@nist.gov

2.1 Machine Maintenance Log Analysis

With an increasing emphasis on smart manufacturing and reducing machine downtime, process monitoring, diagnostics, and prognostics have gained prevalence. This trend—coupled with cheaper and more accessible sensors and data storage solutions—has led to an increase in maintenance data [8]. Despite the potential benefits of high-volume maintenance data for better machine management, companies frequently struggle to adopt advanced manufacturing technologies and strategies due to cost and lack of technical expertise in data analysis [27]. Simple yet powerful solutions for data analysis are necessary to aid manufacturers in improving their practices. There has been an increasing focus on sensor data and predictive maintenance using AI techniques [11,51]. However, these works often neglect a large portion of maintenance data: natural language, short-text maintenance logs. Annotation methods for short-text maintenance work orders [34,43] have been the subject of recent research. For instance, Sexton et al. [45] developed Nestor (https://nist.gov/services-resources/software/nestor), an open-source tool that uses internal "importance" heuristics and seed data annotated with domain-relevant tags by experts. Nestor uses these to annotate maintenance logs with similar tags.

Visual analytics is another technique that has gained popularity in this domain in recent years. Notable work in visual analytics for machine log visualization and monitoring includes ViDX [56] for historical analysis and real-time monitoring of assembly lines, La VALSE [24] and MELA [46] for interactive event analysis logs, and ViBR [12] for vehicle fault diagnostics. These approaches are created for specific datatypes. On the other hand, we treat the data as high-dimensional, heterogeneous datasets that include unstructured text, making our approach usable across different domains.

2.2 Visualizing Heterogeneous Data

The challenges of visualizing heterogeneous data, i.e., data with mixed datatypes or variables, such as numerical, categorical, and text, were recognized early in visualization research. Almost 25 years ago, Zhou and Feiner [57] provided a systematic approach to design visualizations for heterogeneous data based on data characteristics and the tasks involved. The *size* of heterogeneous datasets poses additional challenges for visualization, such as requiring large screens and appropriate visual mappings. Different approaches were developed to address these challenges, such as developing automated specification algorithms to map data attributes to visual attributes [9], and high-resolution immersive visualization environments [40].

Visualizing heterogeneous data also provides a way for the user to establish *context*. For instance, coordinated timeline visualizations of audio, video, and text data of human-human or human-machine interactions can provide context to observations about movement, speech, and activity data [13, 19]. More recently, immersive visualizations of system activity overlaid on a spatial layout corresponding to the physical locations of said systems were used to provide contextual information in real-time network security analysis [35].

Unstructured text also forms an important datatype. Descriptive text about problems and repairs is often entered by operators and maintenance personnel who assume familiarity with the machines and related processes. The text thus tends to be terse and laden with jargon, and is often inconsistent across people. Developing a lexicon—a domain-specific vocabulary—is often necessary to interpret such text data in a semantically consistent way. The General Inquirer [50] is one of the earliest attempts to build a lexicon for content analysis of text. Categories such as Linguistic Inquiry and Word Count (LIWC) [39] focus on psychological relevance (such as moods) and general-purpose applications. Such models are trained on general text corpora such as news articles, online forums, and fiction. For application to large-scale technical text data, automated tagging needs to be balanced via manual sifting of the text.

Visual analytics has been used to achieve better-balanced tags, using a combination of high-dimensional data visualizations and user-steered analyses. For instance, ConceptVector [37] visualizes word-to-concept similarities to guide users to categorize text data given a specific domain, such as politics or finance. Similar vector space representations are used by Heimerl and Gleicher [25] to design visualizations that help users understand word vector embeddings. In addition, several tools such as the Exploratory Labeling Assistant [18] and AILA [14] use machine-learning based recommendations to help users characterize or label documents.

Drawing from this combination of statistical and manual approaches, we use word embeddings to translate short texts to high-dimensional vectors, and apply DR and clustering to find groups of semantically related short texts in a 2D space. We use similar DR and clustering representations for numerical and text data dimensions, which gives us consistent representations across datatypes.

2.3 Visualizing High-Dimensional Data

Most machine maintenance log data tend to be high-dimensional, as each breakdown or maintenance event is recorded with multiple fields relating to different personnel and/or departments [44]. While high dimensionality has its advantages, such as the ability to contextualize and correlate features of the data, it also makes the data less usable for sampling or statistical analysis [15]. Dimensionality reduction provides a lower-dimensional representation while preserving the essential information of the original data [54]. Nonlinear DR techniques, such as UMAP [36], are especially relevant for large-scale, high-dimensional data as they preserve local neighbor relationships, which can help identify subgroups in the data.

DR can be further exploited to cluster the data with higher speed and performance [47] or to produce an overview of the data [32,42]. During this process, visual analytics of the clustered data is often needed to help users determine *which* attributes contribute to the distinctness of each cluster [5]. Statistical charts (e.g., boxplots) [29] or density plots [49] of selected clusters from the DR result have been used for this purpose. However, showing one statistical chart for each attribute becomes visually overloaded as the number of attributes increases. A better approach would be to identify and visualize salient attributes that contribute to a selected cluster. For instance, Broeksema et al. [6] visualized the results of multiple correspondence analysis (MCA) [30]—a variant of principal component analysis (PCA) for categorical data—together with a colored Voronoi cell that represents a highly-related attribute to each data point. Similarly, Joia et al. [28] drew a convex hull around each cluster and filled the resulting polygon with a word cloud consisting of names of the attributes related to the cluster. Faust et al. [17] took a different approach, using local perturbations in the input data to represent how the higher dimensions are represented in the projected views. More recently, Fujiwara et al. [20] used contrastive learning to find attributes that contrast a selected cluster from the rest of the data. We incorporate this contrastive learning-based approach to analyze the numerical attributes of the maintenance log data while introducing an analogous approach for the categorical attributes.

3 Requirements

Typically, visual analysis of heterogeneous, multidimensional data is performed with the goal of identifying patterns within the data and extracting meaning from them [2, 55]. With our application area of machine maintenance data analysis in mind, we draw our requirements from existing work on maintaining and tagging machine performance, error, and maintenance log data.

Most of our requirements are based on prior work by Brundage et al. [7, 8] who generate a set of commonly-occurring data elements from their study of various maintenance work order datasets including temporal (e.g., time between failures, machine downtime, etc.), machine (machine type, location, etc.) human (operator/tech name, skill level, etc.), raw text (problem descriptions, solution, etc.), and tagged elements (items, actions, etc.). Broadly speaking,

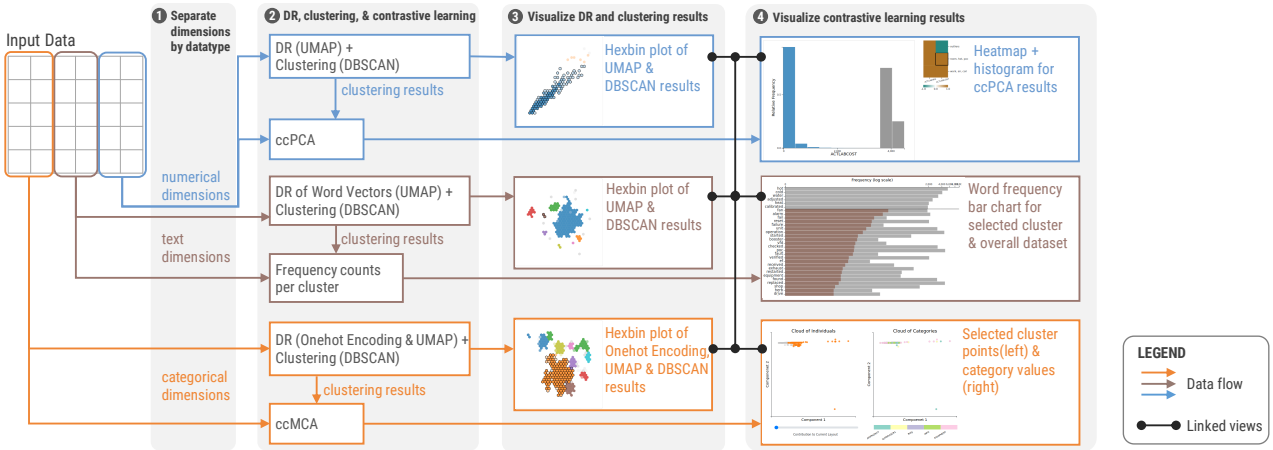

Figure 1: Data processing pipeline for individual views based on the category of data dimensions (categorical, text, and numerical). The figure also shows which views are linked via selection and filtering interactions.

these elements can be classified based on their datatype as numerical, categorical, and text. They also propose a maintenance management workflow with six steps: (1) analyzing the work order, (2) selecting and prioritizing work orders, (3) planning equipment, resources, and labor, (4) scheduling the tasks involved, (5) executing the tasks, and (6) completing and documenting the tasks performed. Our goal is to aid the user—assumed to be a planning engineer or an analyst—in the execution of Steps 1–3. Depending on the scenario, this may require accurate identification of the maintenance task involved, using maintenance logs to anticipate component failure, or correcting work orders with misdiagnosed problems or misidentified tasks.

We thus infer that a system that uses maintenance log data to aid maintenance planning and management needs to be robust to different datatypes, supports visual analysis of data at scale, and helps the user characterize and label parts of the data based on their domain knowledge. The system requirements are:

- **R1 Robustness to Datatype:** The system should accommodate all three types of data commonly required for the analysis of maintenance logs, i.e., numerical, categorical, and text data. Given the inherent difference between the datatypes, an appropriate analysis approach is needed for each.
- **R2 Scalability:** Maintenance log data in an organization can vary from a few thousand records to hundreds of thousands of records, depending on the organization size. With each record consisting of several dimensions of mixed datatypes, the system needs to be robust to different data scales.
- **R3 Data Subset Identification:** When visualizing large-scale data with heterogeneous dimensions, it is not optimal or practical to start by examining individual data points. It is more important and efficient to be able to identify subsets comprising data points that are closely related to each other. This may mean that all data points in an identified subset have common attributes, or that they may be related to each other based on their values along multiple dimensions. With different dimensions composed of different datatypes, the system should allow subset identification approaches suitable across datatypes.
- **R4 Data Subset Characterization:** Analyzing maintenance logs requires not only the identification of patterns/subsets within the data, but also their *characterization*, or what separates them. For instance, a problem common to a group of machines could be characterized by all machines being similar (e.g., lathes), or requiring replacement of the same component, or of components supplied by the same vendor. Identifying such common characteristics become more difficult as the relationship shared by a subset of maintenance logs becomes more complex. Thus, the system should provide effective analysis support to characterize the subsets from many dimensions.
- **R5 Extensibility:** Different organizations may choose to log information about their maintenance activity in different forms and granularities. The only aspect that may be common across these datasets is that they are likely to be multidimensional and heterogeneous. The system should be extensible to a different dataset with minimal effort, and not be overly dependent on any one specific dataset's attributes or format.

4 Data Processing & Visualization

Based on the requirements identified in Sect. 3, it is clear that the three types of data common to machine maintenance logs—numerical, categorical, and text—need to be processed appropriately and visualized using approaches that are robust to changes in the data scale. In this section, we describe the data processing approaches and visualization designs that address the identified requirements[1] .

4.1 Workflow

In Sect. 2, we see that visualizing heterogeneous data is advantageous as it allows the user to draw inferences based on context from different data dimensions. We also see that the issues of scale and dimensionality make it challenging for such observations and inferences to be drawn. Both issues are addressed by using clustering techniques to form subsets within the data (requirement **R3**). These can then be visually and interactively explored to understand the relationship between the data points that make up the subset.

To aggregate the techniques mentioned above, we model our data processing and visualization workflow as a pipeline with six steps: **Step 1**: grouping the data dimensions together based on their datatype (Fig. 1 stage 1); **Step 2**: performing DR for numerical, categorical, and text data separately and obtaining a 2D projection for each (Fig. 1 stage 2); **Step 3**: clustering the 2D data to form subsets (Fig. 1 stage 2); **Step 4**: visualizing the 2D projection and clustering results to provide scalable overviews of the dataset (Fig. 1 stage 3 and Fig. 2 A1, B1, C1); **Step 5**: characterizing the clusters separately for each datatype using contrastive learning or statistical methods (Fig. 1 stage 2); **Step 6**: cluster characterization for each datatype with an appropriate visualization (Fig. 1 stage 4 and Fig. 2 A2, B2, C2). Each step is detailed in the rest of this section.

4.2 Identifying Subsets in Heterogeneous Data

DR (step 2) and clustering (step 3) are two essential data processing steps to identify subsets in the data. Informed by our review in Sect. 2.3, we choose UMAP [36] to project the data to a lower-dimensional space. By using a nonlinear DR method such as UMAP, we can effectively extract similar records from high-dimensional

[1]The source code is available at https://github.com/Xiaoyu1993/Machine-Maintenance-Log-Analysis.

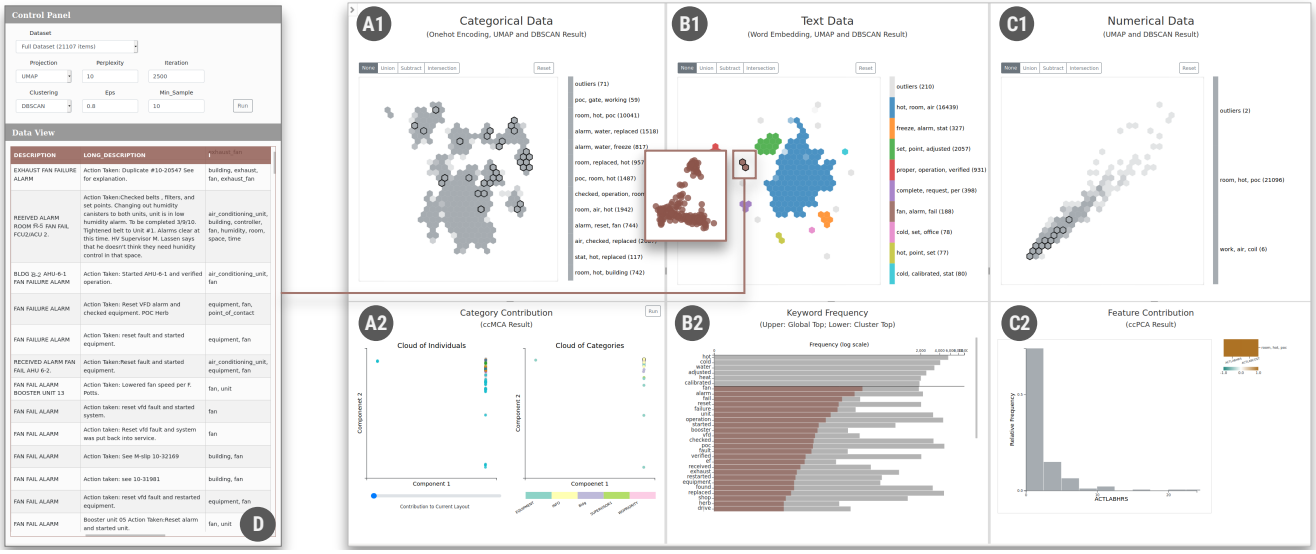

Figure 2: Dashboard interface showing projected views of categorical components (A1), text components (B1), and numerical components (C1) of the dataset using the DR algorithm UMAP. Each projected view is clustered using a chosen clustering algorithm (DBSCAN in the example above). Each projected view is supported by an additional view that is used to characterize a chosen cluster in that view. For the categorical data view, ccMCA (A2) is used to show the selected cluster's separation and the attribute values that contribute to it. A text frequency chart (B2) contrasts the text that occurs most frequently in the selected cluster against the overall text frequency in the dataset. Finally, ccPCA is used to display a heatmap of cluster vs. data dimensions and a histogram showing the value distribution of a selected numerical dimension against the rest of that data (C2). Raw data for any chosen cluster can be viewed using a slide-out tabular view (D). Linking across views A1, B1, and C1 shows the distribution of data clustered in the active view (in color) across the other two views (grayscale).

maintenance log data. High-dimensional representations are obtained for the categorical data with one-hot encoding [23], and for text with word embeddings (see Sect. 4.4.1) before the DR step. The 2D projection of the data can then be clustered using any approach.

We choose DBSCAN (Density-Based Spatial Clustering of Applications with Noise) [16] as it uses a density-based approach that is more suitable for data that may have outliers (e.g., an unusual machine breakdown or repair). Combined with our visualization approach, this technique is more suitable for our case as the analyst can probe into individual records in the case of outliers, and can also examine larger clusters using the linked views. By separating the data dimensions based on datatype, we ensure that our approach is robust to datasets with different dimensions with mixed datatypes (addressing requirement **R1**). This approach of dimension grouping by datatype, DR & clustering to find subsets, and characterizing based on contrastive learning and text frequency comparison makes our approach extensible to most heterogeneous datasets (**R5**).

4.3 Scalable Overview

To show an overview of the DR and clustering results (step 4), we use a hexbin plot [10] for each datatype in the dashboard visualization shown in Fig. 2, i.e., categorical (Fig. 2A1), text (B1), and numerical (C1) dimensions. The hexbin plot is robust to different data scales (requirement **R2**) in that its rendering speed is not significantly impacted by data size or screen resolution. Instead of using a linear color scale typical to hexbin plots, we use a different hue for each cluster and map data density to color intensity within every cluster.

We also preserve the conventional DR representation, i.e., a scatterplot with each data object shown by a dot. We adopt Lindstrom's [31] Level of Detail (LOD) rendering and allow users to switch between these two plots or change the granularity of hexagonal bins by simply zooming in or out of the area they are interested in. Thus, only a small part of the scatterplot needs to be rendered when the users zoom in close. Finally, users can choose to examine the data objects in detail by perusing the slide-out tabular view (Fig. 2D) or by hovering over the dots.

Note that at any point, only one of the three clustered views (A1, B1, or C1 in Fig. 2) can be active. The active view is indicated by its clusters highlighted with a categorical color palette. The remaining views are monochromatic/greyscale to prevent the user from mistakenly assuming that a cluster of one color (e.g., blue) in one view corresponds to a cluster of the same color in another view.

4.4 Characterizing clusters

Characterizing a cluster or subset in the data (requirement **R4**) requires the determination of how the cluster is different from the rest of the data. Different datatypes necessitate different contrastive analysis techniques. We discuss the techniques we use to characterize clusters for text, numerical, and categorical data in this subsection.

4.4.1 Text Dimensions

Detailed text descriptions of problems, symptoms, and solutions, form perhaps the richest component of maintenance log data. They are also rife with inconsistencies, typographical errors, or the use of non-standard shorthand that is endemic to that particular organization. Text descriptions are also often supplemented by "tags"—standardized phrases that label the descriptions to identify the problems, items, and solutions. These tags are typically assigned partly based on the knowledge of the user who tags the descriptive text, and partly using machine learning approaches [43, 45].

In order to group the data based on text dimensions, the *meaning* of the text needs to be considered instead of specific keywords that may vary across technical personnel. A more consistent semantic representation would focus on the meaning of the text rather than its form, such that synonyms and related terms are grouped closely. To achieve this, we use word embeddings, which are vector representations of words that take into account their semantic relationships [53]. Words such as "warm" and "hot" can thus be translated to vectors that are close to each other, but distant from a vector representing a word different in meaning, such as "telephone". We create high-dimensional vector representations for the descriptive text by summing and normalizing words in the text. We then use a suitable DR technique (UMAP) to obtain 2D projections of the vectors, and cluster them using DBSCAN (Fig. 3).

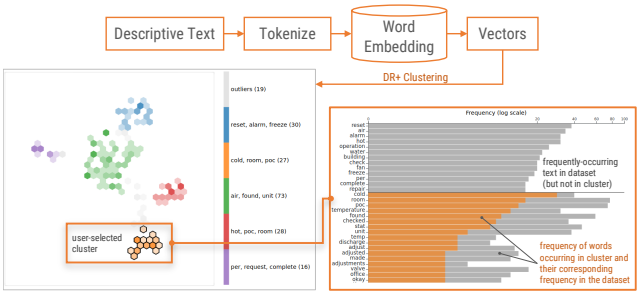

Figure 3: Text processing steps and visualization of the related information. The text frequency chart shows frequencies of text occurring in the selected cluster and contrasts it with both, the frequencies of the corresponding text in the overall dataset as well as the text that is most frequently-occurring in the dataset but not in the selected cluster.

Each cluster represents a collection of descriptions. To characterize a given cluster, we overlay a frequency plot of the most common terms occurring in the cluster on a frequency plot of terms occurring in the overall dataset (Fig. 3 right). Contrasting the most frequent terms of both plots helps the user identify defining characteristics of the cluster. For instance, by examining the frequency plots of Cluster 1 in Fig. 3, we can surmise that the cluster represents maintenance logs of ventilation systems related to lower room temperatures, commonly remedied by adjusting certain valves. The user can examine the raw data related to any cluster using the slide-out tabular view (Fig. 2D) to further gain insight into the cluster and characterize it (requirement **R4**). The cluster labels are editable. For instance, the user can replace the default cluster name with a more descriptive phrase "Lower temperature adjustment".

4.4.2 Numerical Dimensions

As Brundage et al. [7] illustrated with various maintenance key performance indicators (KPIs), measures such as the number of problems/breakdowns, time between failures, and time taken to repair can be used to quantify the role of other performance indicators, such as machine type, problem severity, and technician skill. Other parameters such as cost can be derived from these factors. To understand how those parameters contribute to the separation of clusters for numerical data, we adopt a method called ccPCA [20]. We briefly describe ccPCA and its application to our system. Notations used in the following sections are summarized in Table 1.

Introduction to cPCA. cPCA aims to reveal enriched patterns in a target matrix $\mathbf{X}_T$ relative to a background matrix $\mathbf{X}_B$. To do so, cPCA finds directions (called contrastive principal components, cPCs) that maximally preserve the variation in $\mathbf{X}_T$ while simultaneously minimizing the variation in $\mathbf{X}_B$. This can be achieved by performing EVD on $(\mathbf{C}_T - \alpha\mathbf{C}_B)$ where $\mathbf{C}_T$ and $\mathbf{C}_B$ are covariance matrices of $\mathbf{X}_T$ and $\mathbf{X}_B$, respectively. α $(0 \leq \alpha \leq \infty)$ is a hyperparameter, called a contrast parameter, which controls the trade-off between having high target variance and low background variance. When $\alpha = 0$, the resultant cPCs only maximize the variance of $\mathbf{X}_T$ (i.e., the same with using ordinary PCA). As α increases, cPCs place greater emphasis on directions that reduce the variance of $\mathbf{X}_B$.

Introduction to ccPCA. In order to characterize clusters, ccPCA utilizes cPCA as its base. Let $\mathbf{X}_E$, $\mathbf{X}_K$, and $\mathbf{X}_R$ be matrices of the entire dataset, a target cluster selected from the entire dataset, and the rest of the data points, respectively. ccPCA enhances the original cPCA by using $\mathbf{X}_E$ as a target matrix and $\mathbf{X}_R$ as a background matrix, instead of using $\mathbf{X}_K$ and $\mathbf{X}_R$ as target and background matrices, respectively. With the automatic selection of a contrast parameter [20], ccPCA finds the directions that preserve both the variety and separation between a target cluster and others. These directions are difficult to find with the original cPCA (see the work by Fujiwara et al. [20] for details). By referring to feature contributions (called contrastive principal component loadings or cPC loadings) to the directions, we can obtain the information of which numerical features contribute to the uniqueness of a target cluster relative to others.

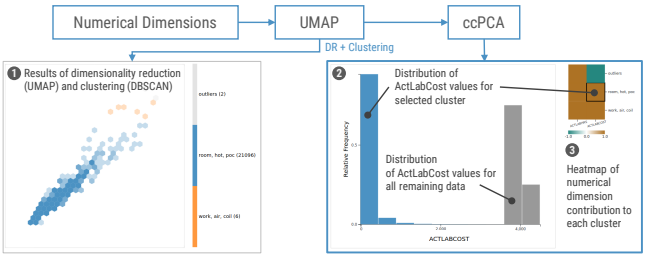

Figure 4: Projection and visualization of numerical data along with the ccPCA view show that the selected cluster has a lower labor cost than the rest of the data.

Table 1: Summary of notations.

$\mathbf{X}_T$, $\mathbf{X}_B$	target, background matrices
$\mathbf{X}_E$, $\mathbf{X}_K$, $\mathbf{X}_R$	matrices of the entire, target cluster, rest data points
$\mathbf{C}_T$, $\mathbf{C}_B$, $\mathbf{C}_E$, $\mathbf{C}_R$	covariance matrices of $\mathbf{X}_T$, $\mathbf{X}_B$, $\mathbf{X}_E$, $\mathbf{X}_R$
$\mathbf{G}_T$, $\mathbf{G}_B$, $\mathbf{G}_T$, $\mathbf{G}_R$	disjunctive matrices of $\mathbf{X}_T$, $\mathbf{X}_B$, $\mathbf{X}_E$, $\mathbf{X}_R$
$\mathbf{Z}_T$, $\mathbf{Z}_B$, $\mathbf{Z}_E$, $\mathbf{Z}_R$	probability matrices of $\mathbf{G}_T$, $\mathbf{G}_B$, $\mathbf{G}_T$, $\mathbf{G}_R$
$\mathbf{B}_T$, $\mathbf{B}_B$, $\mathbf{B}_E$, $\mathbf{B}_R$	Burt matrices of $\mathbf{X}_T$, $\mathbf{X}_B$, $\mathbf{X}_E$, $\mathbf{X}_R$
α	contrast parameter

Visualization. ccPCA provides how strongly each dimension contributes (positively or negatively) to each cluster's contrast with the rest of the data. This contribution is shown as a heatmap (Fig. 4(3)) that indicates the magnitude and direction of the contribution of the numerical dimensions to each cluster with a blue-green-to-brown diverging colormap. By selecting a cell in the heatmap, Fig. 4(2) shows histograms of the corresponding dimension's value distributions of the selected cluster and the rest of the data with the cluster color and gray color, respectively. Based on Fig. 4(2), we can infer that the numerical dimension "actual labor cost" (`ActLabCost`) contributes strongly to Cluster 0's contrast against the rest of the data, and the histograms show that the `ActLabCost` values for the selected cluster are much lower than the rest of the data. The user can further investigate this cluster by selecting it in the DR view (Fig. 4-1) to examine the corresponding data distribution in the text and categorical dimension views as described in Sect. 4.3, or examine the cluster in detail using the tabular view (Fig. 2(D)). Note that Fig. 4 shows only two numerical dimensions due to the dataset we analyze; however, as demonstrated in [20], the combination of using DR, clustering, and ccPCA is useful in identifying and characterizing subsets within high-dimensional numerical data.

4.4.3 Categorical Dimensions

We cannot use ccPCA—which requires numerical or binary data—to characterize categorial data (**R4**). Thus, we introduce a new contrastive learning method, called contrasting clusters in multiple correspondence analysis (ccMCA) by extending multiple correspondence analysis (MCA). Table 2 compares the related methods.

Multiple Correspondence Analysis (MCA) Here, we provide a brief introduction to MCA (refer to [30] for details). MCA can be considered as PCA for categorical data. That is, MCA learns a lower-dimensional representation from high-dimensional categorical data as it maximally preserves the variance of the data. The issue of PCA when applying to categorical data is that PCA handles each category in the data as a numerical value and, as a result, it unnecessarily ranks the categories (e.g., red: 0, green: 1, blue: 2).

To avoid this, MCA first converts an input matrix $\mathbf{X}_T$ of categorical data into a disjunctive matrix $\mathbf{G}_T$ (or disjunctive table) by applying one-hot encoding to each categorical dimension. For example, when $\mathbf{X}_T$ consists of two columns (or often called questions) of "color" and "shape" and each has categories (i.e., categorical answers) of {"red", "green", "blue"} and {"circle", "rectangle"}, $\mathbf{G}_T$ will have five columns of "red", "green", "blue", "circle", and "rectangle" and each of the matrix elements will be either 0 or 1. Afterward, by dividing each cell in $\mathbf{G}_T$ with a total of $\mathbf{G}_T$, we obtain

Table 2: Comparison of representation learning methods. ccMCA is a new method we introduce in this paper.

data type	method	purpose	solution
numerical, binary	PCA	preserving the variance of $\mathbf{X}_T$	EVD on $\mathbf{C}_T$
	cPCA	identifying enriched patterns in $\mathbf{X}_T$	EVD on $(\mathbf{C}_T - \alpha\mathbf{C}_B)$
	ccPCA	characterizing a cluster $\mathbf{X}_K$	EVD on $(\mathbf{C}_E - \alpha\mathbf{C}_R)$
categorical, binary	MCA	preserving the variance of $\mathbf{X}_T$	EVD on $\mathbf{B}_T$
	cMCA	identifying enriched patterns in $\mathbf{X}_T$	EVD on $(\mathbf{B}_T - \alpha\mathbf{B}_B)$
	ccMCA	characterizing a cluster $\mathbf{X}_K$	EVD on $(\mathbf{B}_E - \alpha\mathbf{B}_R)$

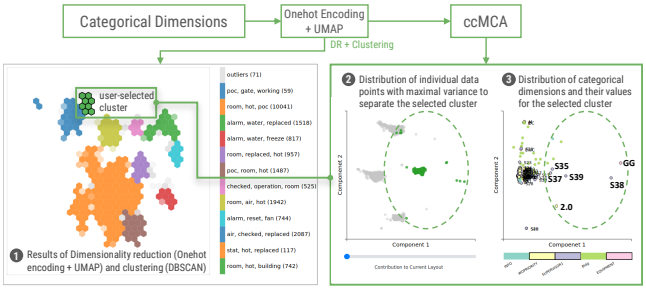

Figure 5: Projection of categorical data (1), with the ccMCA view showing the separation of the selected cluster (2), and its corresponding category distribution (3). The categories of "GG" in "EQUIPMENT", 2.0 in "WOPRIORITY" (work priority), and S35–S39 in "SUPERVISOR1" are most likely to be the characterization of this cluster.

a probability matrix (or correspondence matrix) $\mathbf{Z}_T$. This probability matrix corresponds to an input feature matrix for PCA. Similar to PCA, we apply normalization to $\mathbf{Z}_T$. With the normalized $\mathbf{Z}_T$, we can obtain a Burt matrix, $\mathbf{B}_T$, with $\mathbf{B}_T = \mathbf{Z}_T^\top \mathbf{Z}_T$. $\mathbf{B}_T$ corresponds to a covariance matrix used in PCA (note: in PCA, a covariance matrix of $\mathbf{X}_T$ can be obtained with $\mathbf{C}_T = \mathbf{X}_T^\top \mathbf{X}_T$). Thus, as PCA obtains principal components by performing eigenvalue decomposition (EVD) on $\mathbf{C}_T$, MCA obtains the principal directions by performing EVD on $\mathbf{B}_T$ to preserve the variance of $\mathbf{G}_T$.

Contrastive MCA (cMCA) Now, we introduce contrastive version of MCA (cMCA) [21] and enhance cMCA to ccMCA in the next subsection. As described above, MCA and PCA fundamentally share the same idea of finding the best directions to preserve the variance by using EVD on a covariance matrix. Therefore, we can extend MCA to cMCA by employing the same idea with cPCA.

Extension from MCA to cMCA. As described in Sect. 4.4.2, the only difference between PCA and cPCA is that while PCA directly performs EVD on a target covariance matrix $\mathbf{C}_T$, cPCA takes a subtraction of target and background covariance matrices with a contrast parameter (i.e., $\mathbf{C}_T - \alpha\mathbf{C}_B$) and then performs EVD on it. To reveal enriched patterns in a target matrix of categorical values, we can use the same idea that we use with cPCA and apply it to MCA. As stated in Sect. 4.4.3, in MCA, a Burt matrix $\mathbf{B}_T$ contains similar information with a covariance matrix $\mathbf{C}_T$ in PCA. Therefore, we can obtain contrastive directions by computing $\mathbf{B}_T - \alpha\mathbf{B}_B$, where $\mathbf{B}_T$ and $\mathbf{B}_B$ are target and background Burt matrices, and then performing EVD on $(\mathbf{B}_T - \alpha\mathbf{B}_B)$. Here, α $(0 \leq \alpha \leq \infty)$ is also a contrast parameter and has the same role as cPCA.

Contrasting Clusters in MCA (ccMCA) For the cluster characterization, we enhance cMCA to ccMCA. Here, we apply the similar idea of the extension from cPCA to ccPCA.

Extension from cMCA to ccMCA. cMCA can be enhanced to ccMCA by using $\mathbf{X}_E$ and $\mathbf{X}_R$ as input target and background matrices. Since the directions identified by ccMCA differ based on the contrast parameter α, we also provide the automatic selection method of α by employing the same method introduced by Fujiwara et al. [20], which utilizes the histogram intersection for its optimization. Fig. 5(2) shows the ccMCA result when selecting the green cluster from Fig. 5(1) as a target cluster. The green points are clearly separated from others while keeping a high variance.

One ccMCA's major and different challenge from ccPCA is that how we inform the feature contributions. ccMCA also provides

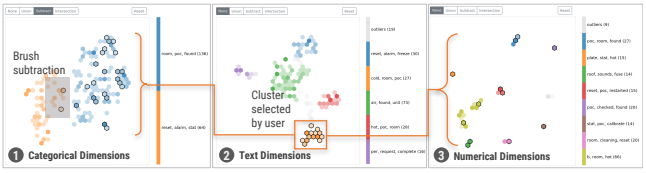

Figure 6: Linking among the projected views of categorical, text, and numerical data allows the user to explore the data clusters from the perspective of datatypes. For instance, selecting Cluster "cold, room, poc" from the projected and clustered view of the text dimensions (2) highlights the distribution of the same points in the other two views (1) & (3). We can see some correlation between the selected cluster and Cluster "room, poc, found" in the categorical dimension view. The brush and Boolean subtraction tools can be used to refine the selection and further reveal the correlation between the two clusters.

contributions (or loadings) of each dimension (i.e., category) of $\mathbf{G}_T$ with $\mathbf{w}_i = \sqrt{\lambda_i}\mathbf{v}_i$ where $\mathbf{w}_i$ is feature contributions to the i-th principal direction, λ_i is the i-th top eigenvalue generated via EVD, and $\mathbf{v}_i$ is the corresponding eigenvector. Because EVD is performed on Burt matrices of $\mathbf{G}_T$ and $\mathbf{G}_B$, which are obtained by applying one-hot encoding to $\mathbf{X}_T$ and $\mathbf{X}_B$, $\mathbf{w}_i$ shows a contribution for each category (e.g., "red", "green", and "blue") but not for each question (e.g., "color"). Therefore, the number of dimensions of $\mathbf{w}_i$ can be easily overwhelmed. For example, when there are 6 questions and 5 categories for each question, the number of dimensions in $\mathbf{w}_i$ becomes 30. Also, as each data point's position in a ccMCA projection (e.g., Fig. 5(2)) reflects a compound of contributions, looking at each contribution may not be sufficient to understand the association between the projection and contributions. For instance, even when one category may have a strong contribution to the positive direction of the first axis (x-axis in Fig. 5(2)), this does not ensure that data points with large positive x-coordinates have answered the corresponding category because, at the same time, many other categories may have a weak contribution to the positive direction.

To address this issue, similar to MCA, we provide the *principal cloud of categories* (or *column principal coordinates*), as shown in Fig. 5(3). In MCA, the principal cloud of categories (PCC) is used to grasp which categories each data point likely has answered by comparing the positions of data points in an MCA projection (or the *principal cloud of individuals*, PCI) and categories in PCC. When a data point in PCI is placed at a close position with certain categories in PCC, this data point tends to have these categories as its answers. We can also perform the same analysis above for ccMCA.

In MCA, PCC $\mathbf{Y}_T^{\text{col}}$ is usually obtained by taking a product of a diagonal matrix $\mathbf{D}_T$ of the sum for each column of $\mathbf{G}_T$ and the top-k eigenvectors $\mathbf{W}_T$ obtained by EVD (i.e., $\mathbf{Y}_T^{\text{col}} = \mathbf{D}_T\mathbf{W}_T^\top$). However, because ccMCA performs EVD on $(\mathbf{B}_T - \alpha\mathbf{B}_B)$ and the result is influenced by $\mathbf{X}_B$ as well, we cannot compute PCC in the above manner. Instead, we use MCA's *translation formula* from PCI to PCC [30]. The translation from PCI to PCC can be performed with:

$$\mathbf{Y}_{\text{col}} = \mathbf{D}_T^{-1}\mathbf{Z}_T^\top \mathbf{Y}_T^{\text{row}} \text{diag}(\boldsymbol{\lambda})^{-1/2} \quad (1)$$

where $\mathbf{Y}_T^{\text{row}}$ is PCI of a target matrix and $\boldsymbol{\lambda}$ is a vector of the top-k eigenvalues. An example of the resultant PCC is shown in Fig. 5(3). By referring to Fig. 5(2) and (3), the analyst can characterize a selected cluster by understanding which categories are highly associated with the uniqueness of the cluster.

4.5 Linking and Interactions

The visualizations across all six panels of the dashboard and tabular view are fully linked, and support brushing and direct selection (of a bin/cluster). Users can select, say, a cluster of interest in one of the projected 2D views (A1, B1, or C1 in Fig. 2) and observe the distribution of the cluster in the remaining two views. Each projected view is supported by a cluster characterization view (A2, B2, and C2 in Fig. 2). When a cluster is selected from one of the projected

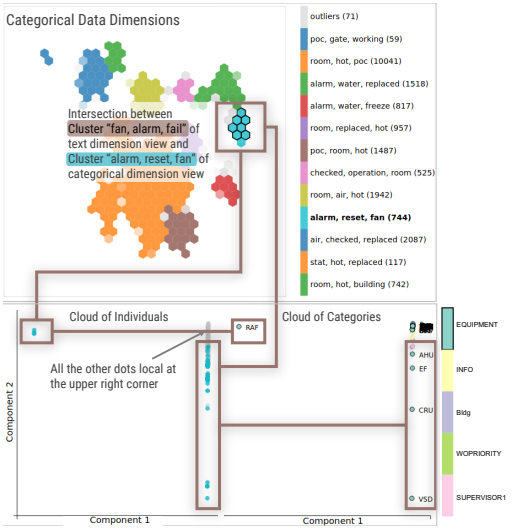

Figure 7: Characterizing a customized cluster (see Sect. 6-Cluster Characterization). Dots in the cloud of categories that share similar locations to the colored dots in the cloud of individuals reveal equipment types that contribute more to the separation of this cluster.

views, all three characterization views update to show the results of that cluster's characterization analysis based on categorical (A2), text (B2), and numerical (C2) data dimensions. The tabular view also updates to show the attributes of the data in the selected cluster.

The linked views update in a similar manner even if—instead of selecting a cluster—the user selects, say, a single hexbin, or brushes across multiple hexbins. Boolean operations such as the union, intersection, and difference are also supported for more sophisticated selections of data across the three projected views. For instance, the user can intersect multiple clusters across different views to find points common across clusters, or combine the clusters by the union.

Fig. 6 shows an example of interactive linking. The user selects the cluster labeled "cold, room, poc" in panel 2 (projected view of text). This highlights hexbins in the other two views that correspond to this cluster. In the example shown, most of the data points overlap with Cluster "room, poc, found" in panel 1 (categorical dimensions), indicating a correlation between these two clusters. To better observe the overlapping points, the user subtracts the two outliers in panel 1 by brushing them out, and checks the supplementary views. Panel 3 shows the points distributed across clusters, indicating no correlation between the selected clusters along their numerical dimensions.

5 Implementation

The dashboard visualization is implemented as a web framework with a Flask server at the back end. The separation of numerical, categorical, and text dimensions is currently performed manually. We compute dimensionality reduction and clustering at the back end for each of these three groups of dimensions and visualize the results by creating an interactive web-based dashboard application. We use HTML/JavaScript for the front end using Bootstrap and React libraries, and D3 [4] to create interactive visualizations.

We use the Scikit-Learn [38] machine learning library for most of the dimensionality reduction and clustering algorithms, except for UMAP and ccPCA, for which we use implementations by McInnes et al. [36] and Fujiwara et al. [20] respectively. We use our own implementations of MCA and ccMCA for DR and contrastive learning for categorical dimensions. For the text dimensions, we use the Natural Language Toolkit (NLTK) [33] for the text processing, ConceptNet Numberbatch [48] as the word embedding to vectorize the text, and Gensim [41] to perform the word-vector lookup.

We tokenize the descriptive text and tags, and remove stop words. Vector representations of words in the remaining text are retrieved

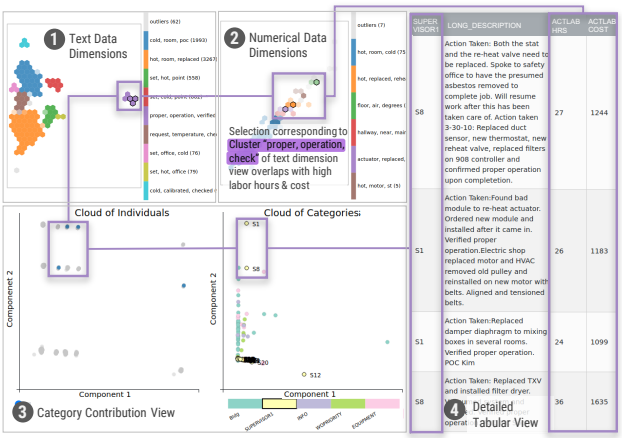

Figure 8: Examining a subset of the original data characterized by temperature-related complaints. The selected purple cluster is—using the category contribution view (3)—shown to be related to high costs associated with two supervisors (see Sect. 6).

using the word embedding, normalized, and added to obtain a single vector representing the unstructured text component of each data point. While ConceptNet Numberbatch contains a fairly large vocabulary of over 500,000 words, there may be domain- or organization-specific terms used in maintenance logs that are not present in the word embedding. In our current implementation, we discard these terms on the assumption that enough of the meaning is captured in the rest of the text for clustering data. However, in future iterations, we plan to update word embeddings using vocabulary from technical manuals and organizational documentation.

6 Use Case Scenario

We illustrate the use of our system using maintenance log data of HVAC systems used in multiple office buildings of an organization. We focus on tasks where the user would analyze the data for patterns and trends to better allocate resources, assign technicians, and schedule future work. Such tasks are an important part of the maintenance cycle, as described in Brundage et al. [8]. The maintenance logs consist of over 21,000 records collected over ten years and contain multiple dimensions of categorical, text, and numerical data. For the purpose of this use-case scenario, we select dimensions of the data that have the least number of missing values. The dataset is grouped by the following sets of dimensions.

The first group involves the categorical dimensions of (1) the `building number` where a complaint on the HVAC system was recorded, (2) `equipment type` of the HVAC subsystem or machine, (3) work order `priority`, (4) system/complaint `location` (building number + floor + room), (5) the index of the `supervisor` in charge of the systems at the time of logging the problem/solution.

Most of the numerical dimensions in the data involve dates and times of logging and are not accurate or consistent enough to compute meaningful timespans. The second group thus involves the two remaining numerical dimensions of (1) `actual labor hours` incurred, and (2) `actual labor cost` incurred.

The third group consists of the text dimensions of (1) `long description` or a description of the problem or complaint that needed addressing, (2) `description` or a small set of keywords highlighting the important aspects of the problem, and (3) a set of multiple `tags` assigned to each maintenance record. The text fields were cleaned up to remove extraneous characters (e.g., HTML tags, symbols, URLs, etc.), remove punctuation, normalize whitespace sequences, and correct typographical and Unicode errors.

Our scenario involves Alice, a maintenance supervisor responsible for the smooth running of HVAC systems across the organization. Alice uses our prototype to examine the dataset and identify patterns

in the logs to identify potential issues and plan preventive maintenance. One of her initiatives has been to try and allocate manpower for recurring or preventable maintenance problems.

Overview. Alice loads all three data groups discussed above into the prototype to get an overview of the data after DR and clustering. She looks over the default tags assigned to each cluster and notices such commonly-occurring terms as *"room"*, *"air"*, *"hot"*, and *"cold"*. The largest cluster in panel B1 (Fig. 2) showing text dimensions appears to contain complaints related to room temperature, with the tags *"too hot"* and *"too cold"* being the most common. From experience, she figures these represent the most typical complaints about HVAC systems in offices. The top keywords in the frequency plot (B2 in Fig. 2) confirm her hunch.

Looking over at the numerical data projection (C1 in Fig. 2), Alice notices that it appears to be linearly correlated. Examining the heatmap in the feature contribution panel (C2 in Fig. 2) confirms this observation as she finds that the *"actual labor hours"* do indeed correlate with *"actual labor cost"*. She makes a mental note to refer to the correlation to filter the data by time or cost in her analysis.

Cluster Characterization. Apart from the clusters related to the temperature problems, Alice notices a unique brown cluster in the text dimension views tagged as "*fan, alarm, fail*" (B1 in Fig. 2) and decides to take a closer look at it. From panel B2, she finds that the top keywords in this cluster—*fan, alarm, fail, reset, repair*—are significantly different from those in the rest of the dataset. She also finds that the cluster overlaps with all clusters in the categorical view (A1 in Fig. 2) that have the above three terms as one of their main tags. The highest overlap in the categorical data view is with a cyan cluster (see Fig. 7) tagged with *"alarm"*, *"reset"*, and *"fan"*.

She uses the Boolean operator to separate the intersection between these two clusters. From the category contribution panel (Fig. 7), she notices that several types of equipment including *"RAF"* (Return Air Fan), *"EF"* (Exhaust Fan), *"VSD"* (Variable Speed Drive), *"CRU"* (Customer Replaceable Unit), and *"AHU"* (Air Handling Unit) contribute the most to this cluster. From her experience, she knows that the above equipment has always had relatively unreliable fans. Cross-checking with the numerical data panel, she realizes that the labor cost is relatively low for these problems, so she makes a note to have regular preventive maintenance done on the equipment.

Projection Interpretation. Now Alice decides to have the "too hot/cold" issue looked into further, and calls in an engineer to filter the dataset by these two tags and examine this filtered dataset separately. After loading the subset into the system, she notices a symmetry in the layout pattern of the text dimension view, about a horizontal axis. The clusters in the upper half of the projection all contain the keyword *"cold"* while those on the lower half contain *"hot"*. She infers that the vertical direction in the projected space relates to temperature, and becomes interested in the clusters located in the middle, especially the solitary purple cluster with tags *"proper, operation, verified"* (Fig. 8-1). She notices a significant variation of labor cost and hours in this cluster (Fig. 8-2). Selecting all points with a higher labor cost and hours, she learns from the updated keyword frequency plot that they correspond to the action *"replaced"*. From the category contribution panel, she finds that this part of the data is highly related to two supervisors *"S1"* and *"S8"* (Fig. 8-3). She confirms this observation by checking the tabular view (Fig. 8-4). She believes that the high cost may either a clerical mistake or an issue with the vendor supplying the parts. She decides to talk with these two supervisors to get to the bottom of the issue.

7 Expert Review

Our prototype was reviewed by three experts in machine maintenance analysis to determine its usefulness and throw light on the kinds of patterns or insights it might reveal to domain practitioners. The first expert (E1) was a data scientist from the industry who had developed approaches for extracting actionable information from maintenance data for over six years. The second expert (E2) was an industrial engineer specializing in model-based systems engineering methodologies. Finally, the third expert (E3) was a computer scientist from academia who worked on algorithm development and natural language processing for 25 years.

We conducted two pilot studies with co-authors of this work who are also domain experts in maintenance log analysis. This helped us simulate the remote setup and related logistics planned for the study, determine needs for—and forms of—tutorials and examples, and identify tasks that could be performed within existing constraints. Based on these, we designed a semi-structured, open-ended expert review. We followed the "pair analytics" paradigm [1] with one of the authors as the experimenter and the participant as the subject matter expert (SME). Pair analytics has been shown to be optimal for open-ended studies with SMEs, since it stimulates dialogue between the experimenter and the participant and explicates the participant's thought process, and reduces the burden of fluency with the application from the SME who is usually not a visual analytics expert [1]. The additional constraint of the pandemic, requiring a remote setup, further informed the decision to employ this paradigm.

The study used a video conference setup where the experimenter controlled the tool while the expert observed the visualizations and suggested filters, queries, and interactions via screen sharing. The experts perused a document explaining the views and functions of the prototype prior to the study, and were shown a 20-minute tutorial demonstrating the prototype at the start of the study. They were then asked to explore two datasets (30 mins each), one the HVAC dataset described in Sect. 6, and the other a subset of 17,000 records from the HVAC dataset involving temperature-related complaints.

We categorize our observations on the domain experts' remarks during the exploratory tasks and their feedback on the prototype into *functionality* of the prototype, *visual encoding*, and *interaction*.

Functionality. The tutorial and demonstration at the start of the study involved use cases and observations such as the one presented in Sect. 6. At the end of the demonstration, all three experts found our workflow to be *"highly reasonable"* (E2) and found the cases compelling. Yet, during the exploratory part of the study, they found it difficult to pin down the questions they could ask and answer of the data. For instance, E2 asked, *"What key question am I to answer?"* Based on the experts' questions about the visualizations, filters, and interactions during the exploratory study, we infer that the difficulty encountered by the experts was partly due to the relatively short time they spent with the interface and their unfamiliarity with the data.

Visual Encoding. Domain experts found the linked views to be intuitive and useful. E3 remarked, *"I like the ways that the panels are automatically updated with respect to the selections that (are) made. And being able to see the three types of data all together is good. Definitely a good idea to have them combined"*. However, they found it too abstract to separate the data dimensions into categorical, numerical, and text. As E1 explained, *"Looking at categorical, text and numerical data makes sense from a data perspective, but it's not necessarily the functional break down that makes sense."* Instead, they reported that they would have preferred a way of representing the data that allowed them to see the problems in a **functional** way, e.g., wherein the building, or wherein the machine a problem occurred, or what temporal patterns were observable in the data. E1 and E3 also found it a little confusing that the default cluster labels in the numerical and categorical panels still used keywords from the text component of the data. On the other hand, while they were able to characterize at least one of the clusters, none of them re-labeled the cluster(s). All three experts also found the characterization view for the categorical data difficult to understand. E1 said that they had *"a really hard time understanding this visualization"*. E3 noted that they had *"never seen the information displayed in this way with two side by side panels of the cloud of individuals and categories... it's a little non-specific as far as whether the dots that show up in the (cloud of) categories are close enough to the dots in the (cloud of) individuals and how relevant it is."*

Interactions. All three experts found the brushing and linking to be highly useful, though the hexbin plots were a little confusing for E1 and E2, who took a highlighted hexbin in one view to indicate that all the points in that bin were linked to the cluster selected in another view. E1 suggested providing *"a measurement of how much the correlation or lack of correlation is."* E3 initially found the Boolean operations to be less intuitive, but after asking for and seeing examples of how they were used, deemed the operations to be highly useful. Finally, E3 suggested the addition of numerical filters using which they could identify maintenance costs higher than a certain threshold, while E2 suggested filtering out data associated with commonly-occurring tags to help examine less-common problems.

Overall, the experts found the datatype-based separation less intuitive but considered the coordinated views and Boolean operations across the views to be of value. They recommended more tangible ways of grounding the data in the domain familiar to them by using locations of machines in buildings, locations of components in machines, and filtering by cost, dates, and keywords.

8 Discussion

The use case scenario and the expert review illustrate the importance of interactive visual analysis in the maintenance workflow. For instance, the overview visualizations were seen to provide useful groupings for analysts to explore and interpret using their domain knowledge. Additionally, the expert review illustrates the usefulness of interaction and filtering in helping interpret unfamiliar visual abstractions, and highlights a need to ground the representations in a way that is familiar to the domain experts.

In the use case scenario, we saw how the overview visualizations can help identify common patterns across the dataset (e.g., the *"hot/cold"* cluster) and help small but closely related clusters stand out (e.g., maintenance of equipment involving fans). The ccMCA views (Fig. 7) allowed the user to not only verify common traits—such as the presence of unreliable fans, or replacements ordered by a small subset of supervisors—across a problem group but also identify which equipment (in the case of fans) and supervisors (in the case of replacements) had the common traits.

The expert review highlighted both the advantages and disadvantages of our approach. When the domain experts were demonstrated scenarios, such as that described in Sect. 6, they were convinced and impressed by the capability of the prototype. Their validation of the workflow used to create the projected views and characterization views (Fig. 1) also verified that our approach was well-motivated. On the other hand, the experts found the data separation and visualization too abstract to pick up in a single session. They preferred a more tangible means of viewing the data, based on the location of the machines, locations of the components in the machines, and based on cost. However, representing high-dimensional data based on only one or two characteristics may not reveal important insights. In addition, one of the main advantages of our approach—its generalizability to other domains—will be lost by grounding it too much in one domain. However, there may be a middle ground wherein the user is able to add an additional "custom" view based on familiar data characteristics. We will explore this in future iterations.

In spite of the difficulty the experts faced with the abstract representations, they found the coordination or linking across views to be a useful feature that helped them understand the data better. As with the representations, they did express a preference for more tangible filters (e.g., based on specific cost ranges). However, at least one expert (E3) had started to appreciate the sophisticated filtering possible through the coordinated views and Boolean operations. The expert feedback suggested that some of what they found difficult about the interface was more due to the short duration of the sessions rather than the data abstractions themselves. A longitudinal study—though unrealistic at this time with restrictions on data sharing and the current constraint of remote sessions—would help address some of the familiarity issues that the domain experts currently face.

9 Conclusion

In this paper, we present a design that couples machine learning with interactive visualization for analyzing large, heterogeneous, multidimensional maintenance log data. A key approach is to separate numerical, categorical, and text dimensions of the data, and use lower-dimensional, clustered views that reveal groups in the dataset by each dimension type. We apply existing techniques such as ccPCA and word embeddings with frequency plots to characterize the dataset based on its numerical and text dimensions. Notably, a unique capability is provided with our new contrastive learning method, ccMCA, to characterize a dataset with its categorical dimensions. We present these approaches of clustering and characterization in the form of a dashboard with linked views, and illustrate its utility through a use-case scenario and an expert review. These scenarios allow us to highlight the use of ccMCA in identifying categorical dimensions and their values that contribute to a cluster. The expert review highlighted the usefulness of linked views to characterize clusters across different dimension types. We also identify the need for more grounded, domain-specific representations of data to scaffold the experts' understanding of the system.

NIST Disclaimer

The use of any products described in this paper does not imply recommendation or endorsement by the National Institute of Standards and Technology, nor does it imply that products are necessarily the best available for the purpose.

Acknowledgement

We are grateful to the domain experts who volunteered with our hour-long study during these challenging times. We also thank the anonymous reviewers for their suggestions that helped improve this paper. This research is sponsored in part by the financial assistance award 70NANB20H197 from the U.S. Department of Commerce, National Institute of Standards and Technology.

References

[1] R. Arias-Hernandez, L. T. Kaastra, T. M. Green, and B. Fisher. Pair analytics: Capturing reasoning processes in collaborative visual analytics. In *Proc. HICSS*, pages 1–10. IEEE, 2011.

[2] A. Arleo, C. Tsigkanos, C. Jia, R. A. Leite, I. Murturi, et al. Sabrina: Modeling and visualization of financial data over time with incremental domain knowledge. In *Proc. VIS*, pages 51–55. IEEE, 2019.

[3] J. Bae, T. Helldin, M. Riveiro, S. Nowaczyk, M.-R. Bouguelia, and G. Falkman. Interactive clustering: A comprehensive review. *ACM Computing Surveys*, 53(1):1–39, 2020.

[4] M. Bostock, V. Ogievetsky, and J. Heer. D^3: Data-driven documents. *IEEE Trans. on Visualization and Computer Graphics*, 17(12):2301–2309, 2011.

[5] M. Brehmer, M. Sedlmair, S. Ingram, and T. Munzner. Visualizing dimensionally-reduced data: Interviews with analysts and a characterization of task sequences. In *Proc. BELIV*, pages 1–8, 2014.

[6] B. Broeksema, A. C. Telea, and T. Baudel. Visual analysis of multi-dimensional categorical data sets. *Computer Graphics Forum*, 32(8):158–169, 2013.

[7] M. P. Brundage, K. Morris, T. Sexton, S. Moccozet, and M. Hoffman. Developing maintenance key performance indicators from maintenance work order data. In *Proc. MSEC*. ASME, 2018.

[8] M. P. Brundage, T. Sexton, M. Hodkiewicz, K. C. Morris, J. Arinez, et al. Where do we start? guidance for technology implementation in maintenance management for manufacturing. *J. of Manufacturing Science and Engineering*, 141(9):091005, 2019.

[9] M. Cammarano, X. Dong, B. Chan, J. Klingner, J. Talbot, et al. Visualization of heterogeneous data. *IEEE Trans. on Visualization and Computer Graphics*, 13(6):1200–1207, 2007.

[10] D. B. Carr, R. J. Littlefield, W. Nicholson, and J. Littlefield. Scatterplot matrix techniques for large N. *J. of the American Statistical Association*, 82(398):424–436, 1987.

[11] T. P. Carvalho, F. A. Soares, R. Vita, R. d. P. Francisco, J. P. Basto, and S. G. Alcalá. A systematic literature review of machine learning

methods applied to predictive maintenance. *Computers & Industrial Engineering*, 137:106024, 2019.

[12] G. Y.-Y. Chan, P. Xu, Z. Dai, and L. Ren. ViBR: Visualizing bipartite relations at scale with the minimum description length principle. *IEEE Trans. on Visualization and Computer Graphics*, 25(1):321–330, 2019.

[13] S. Chandrasegaran, S. K. Badam, L. Kisselburgh, K. Peppler, N. Elmqvist, and K. Ramani. VizScribe: A visual analytics approach to understand designer behavior. *International J. of Human-Computer Studies*, 100:66–80, 2017.

[14] M. Choi, C. Park, S. Yang, Y. Kim, J. Choo, and S. R. Hong. Aila: Attentive interactive labeling assistant for document classification through attention-based deep neural networks. In *Proc. CHI*, pages 1–12, 2019.

[15] D. L. Donoho. High-dimensional data analysis: The curses and blessings of dimensionality. In *Proc. AMS Conf. on Math Challenges of the 21st Century*, 2000.

[16] M. Ester, H.-P. Kriegel, J. Sander, X. Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proc. KDD*, pages 226–231, 1996.

[17] R. Faust, D. Glickenstein, and C. Scheidegger. DimReader: Axis lines that explain non-linear projections. *IEEE Trans. on Visualization and Computer Graphics*, 25(1):481–490, 2018.

[18] C. Felix, A. Dasgupta, and E. Bertini. The exploratory labeling assistant: Mixed-initiative label curation with large document collections. In *Proc. UIST*, pages 153–164, 2018.

[19] A. Fouse, N. Weibel, E. Hutchins, and J. D. Hollan. ChronoViz: a system for supporting navigation of time-coded data. In *Proc. ACM Extended Abstracts on CHI*, pages 299–304, 2011.

[20] T. Fujiwara, O.-H. Kwon, and K.-L. Ma. Supporting analysis of dimensionality reduction results with contrastive learning. *IEEE Trans. on Visualization and Computer Graphics*, 26(1):45–55, 2020.

[21] T. Fujiwara and T.-P. Liu. Contrastive multiple correspondence analysis (cMCA): Using contrastive learning to identify latent subgroups in political parties. *arXiv:2007.04540*, 2020.

[22] T. Fujiwara, Shilpika, N. Sakamoto, J. Nonaka, K. Yamamoto, and K.-L. Ma. A visual analytics framework for reviewing multivariate time-series data with dimensionality reduction. *IEEE Trans. on Visualization and Computer Graphics*, 27(2):1601–1611, 2021.

[23] C. Guo and F. Berkhahn. Entity embeddings of categorical variables. *arXiv:1604.06737*, 2016.

[24] H. Guo, S. Di, R. Gupta, T. Peterka, and F. Cappello. La VALSE: Scalable log visualization for fault characterization in supercomputers. In *Proc. EGPGV*, pages 91–100. Eurographics, 2018.

[25] F. Heimerl and M. Gleicher. Interactive analysis of word vector embeddings. *Computer Graphics Forum*, 37(3):253–265, 2018.

[26] M. Hodkiewicz and M. T.-W. Ho. Cleaning historical maintenance work order data for reliability analysis. *J. of Quality in Maintenance Engineering*, 22(2):146–163, 2016.

[27] X. Jin, B. A. Weiss, D. Siegel, and J. Lee. Present status and future growth of advanced maintenance technology and strategy in us manufacturing. *International J. of Prognostics and Health Management*, 7(Special Issue on Smart Manufacturing PHM), 2016.

[28] P. Joia, F. Petronetto, and L. G. Nonato. Uncovering representative groups in multidimensional projections. *Computer Graphics Forum*, 34(3):281–290, 2015.

[29] B. C. Kwon, B. Eysenbach, J. Verma, K. Ng, C. De Filippi, et al. Clustervision: Visual supervision of unsupervised clustering. *IEEE Trans. on Visualization and Computer Graphics*, 24(1):142–151, 2017.

[30] B. Le Roux and H. Rouanet. *Quantitative Applications in the Social Sciences: Multiple Correspondence Analysis*. SAGE Publications, 2010.

[31] P. Lindstrom, D. Koller, W. Ribarsky, L. F. Hodges, N. Faust, and G. A. Turner. Real-time, continuous level of detail rendering of height fields. In *Proc. SIGGRAPH*, pages 109–118, 1996.

[32] S. Liu, D. Maljovec, B. Wang, P.-T. Bremer, and V. Pascucci. Visualizing high-dimensional data: Advances in the past decade. *IEEE Trans. on Visualization and Computer Graphics*, 23(3):1249–1268, 2017.

[33] E. Loper and S. Bird. NLTK: The natural language toolkit. In *Proc. ETMTNLP*, pages 63–70. ACL, 2002.

[34] S. Lukens, M. Naik, K. Saetia, and X. Hu. Best practices framework for improving maintenance data quality to enable asset performance analytics. *Annual Conference of the PHM Society*, 11(1), 2019.

[35] E. Mahfoud, K. Wegba, Y. Li, H. Han, and A. Lu. Immersive visualization for abnormal detection in heterogeneous data for on-site decision making. In *Proc. HICSS*, 2018.

[36] L. McInnes, J. Healy, and J. Melville. UMAP: Uniform manifold approximation and projection for dimension reduction. *arXiv:1802.03426*, 2018.

[37] D. Park, S. Kim, J. Lee, J. Choo, N. Diakopoulos, and N. Elmqvist. ConceptVector: Text visual analytics via interactive lexicon building using word embedding. *IEEE Trans. on Visualization and Computer Graphics*, 24(1):361–370, 2017.

[38] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, et al. Scikit-learn: Machine learning in python. *J. of Machine Learning Research*, 12:2825–2830, 2011.

[39] J. W. Pennebaker, C. K. Chung, M. Ireland, A. Gonzales, and R. J. Booth. The development and psychometric properties of LIWC2007. LIWC Manual, `https://www.liwc.net/LIWC2007LanguageManual.pdf`. Accessed: 2020-4-30.

[40] K. Reda, A. Febretti, A. Knoll, J. Aurisano, J. Leigh, et al. Visualizing large, heterogeneous data in hybrid-reality environments. *IEEE Computer Graphics and Applications*, 33(4):38–48, 2013.

[41] R. Řehůřek and P. Sojka. Software framework for topic modelling with large corpora. In *Proc. LREC Workshop on New Challenges for NLP Frameworks*, pages 45–50, 2010.

[42] D. Sacha, L. Zhang, M. Sedlmair, J. A. Lee, et al. Visual interaction with dimensionality reduction: A structured literature analysis. *IEEE Trans. on Visualization and Computer Graphics*, 23(1):241–250, 2017.

[43] T. Sexton, M. P. Brundage, M. Hoffman, and K. C. Morris. Hybrid datafication of maintenance logs from ai-assisted human tags. In *Proc. IEEE Big Data*, pages 1769–1777, 2017.

[44] T. Sexton, M. Hodkiewicz, and M. P. Brundage. Categorization errors for data entry in maintenance work-orders. *Proc. PHM*, 11(1), 2019.

[45] T. B. Sexton and M. P. Brundage. Nestor: A tool for natural language annotation of short texts. *J. of Research of National Institute of Standards and Technology*, 124:124029, 2019.

[46] Shilpika, B. Lusch, M. Emani, V. Vishwanath, M. E. Papka, and K.-L. Ma. MELA: A visual analytics tool for studying multifidelity hpc system logs. In *Proc. DAAC*, pages 13–18. IEEE, 2019.

[47] A. S. Shirkhorshidi, S. Aghabozorgi, T. Y. Wah, and T. Herawan. Big data clustering: A review. In *Proc. ICCSA*, pages 707–720, 2014.

[48] R. Speer, J. Chin, and C. Havasi. ConceptNet 5.5: An open multilingual graph of general knowledge. In *Proc. AAAI*, pages 4444–4451, 2017.

[49] J. Stahnke, M. Dörk, B. Müller, and A. Thom. Probing projections: Interaction techniques for interpreting arrangements and errors of dimensionality reductions. *IEEE Trans. on Visualization and Computer Graphics*, 22(1):629–638, 2016.

[50] P. J. Stone, D. C. Dunphy, and M. S. Smith. *The general inquirer: A computer approach to content analysis.* MIT press, 1966.

[51] G. A. Susto, A. Schirru, S. Pampuri, S. McLoone, and A. Beghi. Machine learning for predictive maintenance: A multiple classifier approach. *IEEE Transactions on Industrial Informatics*, 11(3):812–820, 2014.

[52] G. K. Tam, V. Kothari, and M. Chen. An analysis of machine-and human-analytics in classification. *IEEE Trans. on Visualization and Computer Graphics*, 23(1):71–80, 2017.

[53] J. Turian, L. Ratinov, and Y. Bengio. Word representations: A simple and general method for semi-supervised learning. In *Proc. ACL*, pages 384–394, 2010.

[54] L. van der Maaten, E. Postma, and J. van den Herik. Dimensionality reduction: A comparative review. *J. of Machine Learning Research*, 10:66–71, 2009.

[55] J. Xia, W. Chen, Y. Hou, W. Hu, X. Huang, and D. S. Ebertk. DimScanner: A relation-based visual exploration approach towards data dimension inspection. In *Proc. VAST*, pages 81–90. IEEE, 2016.

[56] P. Xu, H. Mei, L. Ren, and W. Chen. ViDX: Visual diagnostics of assembly line performance in smart factories. *IEEE Trans. on Visualization and Computer Graphics*, 23(1):291–300, 2016.

[57] M. X. Zhou and S. K. Feiner. Data characterization for automatically visualizing heterogeneous information. In *Proc. InfoVis*, pages 13–20, 1996.